# Optimal Stopping in Latent Diffusion Models

**Anonymous authors**
Paper under double-blind review

## Abstract

We identify and analyze a surprising phenomenon of *Latent* Diffusion Models (LDMs) where the final steps of the diffusion can *degrade* sample quality. In contrast to conventional arguments that justify early stopping for numerical stability, this phenomenon is intrinsic to the dimensionality reduction in LDMs. We provide a principled explanation by analyzing the interaction between latent dimension and stopping time. Under a Gaussian framework with linear autoencoders, we characterize the conditions under which early stopping is needed to minimize the distance between generated and target distributions. More precisely, we show that lower-dimensional representations benefit from earlier termination, whereas higher-dimensional latent spaces require later stopping time. We further establish that the latent dimension interplays with other hyperparameters of the problem such as constraints in the parameters of score matching. Experiments on synthetic and real datasets illustrate these properties, underlining that early stopping can improve generative quality. Together, our results offer a theoretical foundation for understanding how the latent dimension influences the sample quality, and highlight stopping time as a key hyperparameter in LDMs.

## 1 Introduction

A pivotal advancement in the evolution of diffusion models is the introduction of the Latent Diffusion Model (LDM, Rombach et al., 2022). Instead of performing the computationally intensive diffusion process in the high-dimensional pixel space, LDMs first compress the data into a lower-dimensional latent space using a pretrained autoencoder (AE, Kingma & Welling, 2013). The diffusion steps then occur within this more manageable latent representation, significantly reducing computational requirements and training time without a meaningful loss of quality. Once the generative process is complete, a decoder maps the resulting latent vector back into a full-resolution image.

Recent research has shown that classical diffusion models excel at learning the geometric structure of low-dimensional data (see, for example, Li & Yan, 2024; Azangulov et al., 2024). In addition, prior work shows that the final steps of the diffusion are essential to contract the probability mass onto the data manifold (George et al., 2025). However, the fidelity of generated samples critically depends on the numerical stability of the stochastic differential equation (SDE) solver used in the backward diffusion process. One well-documented challenge in this method is the onset of numerical instability as the timestep $t$ approaches $0$ (Song et al., 2021; Yang et al., 2023; Xie & Agam, 2025). This phenomenon is primarily caused by a vanishing signal-to-noise ratio, which makes the corresponding score function difficult to accurately estimate and causes the SDE to become stiff. To avoid this, in practice both the training objective and the inference-time integration are restricted to the interval $[0, T - \delta]$ (Vahdat et al., 2021) for some small, non-zero stopping time, $\delta > 0$.

By contrast, this suggests a key benefit of LDMs, which to our knowledge has not been explored in the literature so far: by relying on the autoencoder to reduce the dimensionality, the LDM provides an alternative mean to learn low-dimensional manifolds without relying on the last few steps. This intuition motivates the following hypothesis:

*In latent diffusion models, the last diffusion steps do not improve, or even degrade, sample quality.*

We find empirical evidence of this hypothesis by comparing samples from a LDM with those of a standard diffusion model directly trained in the pixel space, both trained on the dataset CelebA. In the case of an LDM, degradation in the last sampling steps is evidenced by a rising FID score, as illustrated in Figure 1 and 2. In addition, we observe that, with an LDM achieving same level of performance as the pixel diffusion, there is a benefit in early-stopping. Furthermore, by early-stopping, the LDM with smaller latent dimension may achieve a similar FID score than the one operating on a larger latent dimension. In contrast, this phenomenon, which happens much earlier in the diffusion process than potential numerical instabilities close to $T$, is absent in standard diffusion models. Visual inspection of the associated images confirms that their quality does not improve in the last steps of the LDM, contrarily to standard diffusion (see Figure 3). Our main contribution in this work is to provide a theoretical justification of this observation. To this aim, we analyze the phenomenon using Gaussian data and a linear autoencoder. This choice is deliberate, as this simplified setting already exhibits phenomena similar to the larger-scale evidence, while being analytically tractable, allowing us to rigorously demonstrate the effect of early stopping and dimension reduction.

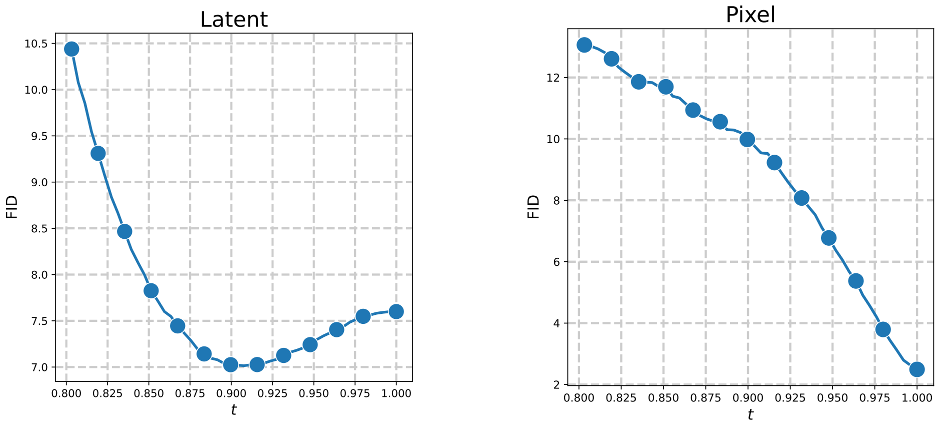

Figure 1: (left) FID-30k score of latent diffusion model on CelebA-HQ, with latent shape $64 \times 64 \times 3$. (right) FID-30k score of standard diffusion model (trained in pixel space) on CelebA64 ($64 \times 64 \times 3$).

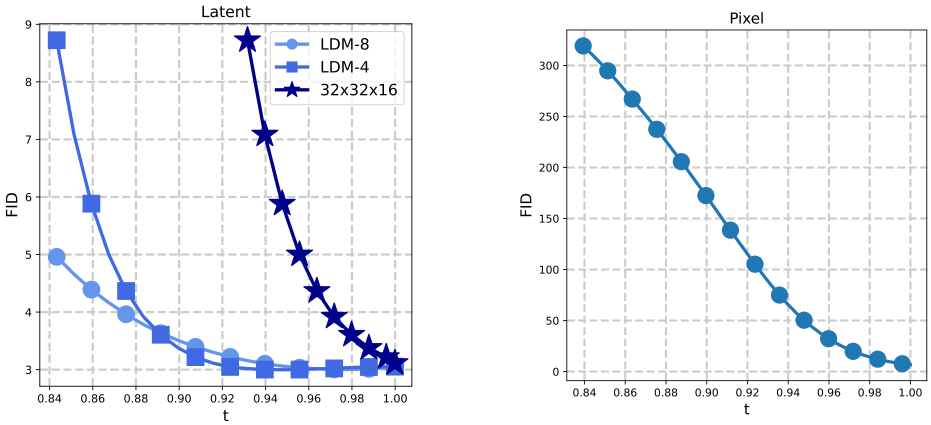

Figure 2: We train LDMs with latent dimension $32 \times 32 \times 4$ and $64 \times 64 \times 3$ on ImageNet-256, and we compare it to the diffusion process of a DM trained on ImageNet-64. We also measured the image quality using MMD (Jayasumana et al., 2024) and sliced Wasserstein, and report generated samples, see Appendix D. See also the Appendix for a zoom of the left figure that highlights the non-monotonicity of the curves.

**Contributions and organization.** Our contributions are as follows:

- We propose a theoretical framework to analyze the impact of the last steps of the LDM by studying the evolution, along the backward diffusion, of the Wasserstein-2 distance between the data distribution and the generated distribution for Gaussian data (Section 3).

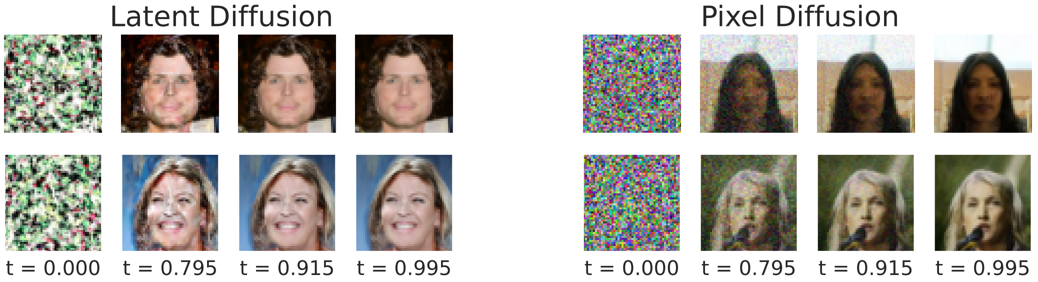

Figure 3: Samples generated with a latent diffusion model (LDM) and a pixel-space diffusion. In the LDM, the before-last sample is nearly denoised and indistinguishable from the final one, whereas in the pixel-space model stronger noise remains at that timestep. See Appendix D for more examples.

- We identify the optimal latent dimension as a function of the covariance matrix and the stopping time, for the case of a diagonal data covariance matrix. In particular, when the data lies on a linear subspace, we prove that the optimal strategy is to project the diffusion onto this subspace, and we determine the corresponding optimal stopping time. This reveals a time-dependent trade-off: distributions generated in the early stages of the backward process are best approximated in lower-dimensional spaces, while higher dimensions are required to faithfully reconstruct the data in the final sampling steps (Section 4).
- When the score is learned by a restricted class of parametrized models (i.e., when the weights of the model are capped), we further establish the existence of an optimal latent projection that optimizes the backward diffusion, and we investigate how its dimension depends on the model class constraint and the data covariance (Section 5).
- We extend our analysis to general covariance matrices, showing that the interaction between early stopping and latent dimensionality persists beyond the independent setting (Section 6).

## 2 Related work

**Learning low-dimensional data with diffusion models.** Riemannian Diffusion Models, introduced by Huang et al. (2022); De Bortoli et al. (2022), generalize the diffusion process to operate on Riemannian manifolds and preserve a known geometric structure by design. Subsequent theoretical work has analyzed the behavior of standard Denoising Diffusion Probabilistic Models (DDPMs) under the manifold hypothesis, demonstrating that they can implicitly adapt to the data's intrinsic dimension without explicit knowledge of the manifold (Tang & Yang, 2024; George et al., 2025).

Further improvements in computational and memory efficiency were introduced by LDMs (Rombach et al., 2022) by first training a compression model to transform images into a lower-dimensional latent space, from which the original data can be reconstructed at high fidelity. In practice, this approach is implemented with a regularized VAE (Esser et al., 2021). The LDM is then trained in the latent space. Building on this core concept, LDMs have been extended to new domains, such as the generation of high-resolution videos (Blattmann et al., 2023). Furthermore, extensive research has focused on improving LDM's sampling quality, including methods like aligning encoded images with DINOv2 representations (Yu et al., 2024), and enhancing the robustness of the latent space through explicit or implicit equivariance constraints (Kouzelis et al., 2025; Skorokhodov et al., 2025; Zhou et al., 2025). In contrast to standard diffusion models, theoretical properties of LDMs have been little studied; in this work, we investigate the connection of the latent dimension with diffusion stopping time and score matching regularization.

**Optimal stopping time of diffusion models.** Focusing on a theoretical analysis of this phenomenon, Achilli et al. (2025) investigate the optimal stopping time for diffusion models under the assumption that the data is concentrated on a low-dimensional manifold, a concept formalized by the Hidden Manifold Model (Goldt et al., 2020). Closer to our contribution is the work of Hurault et al. (2025). They also investigate the scenario where the true data distribution is Gaussian. Their analysis focuses on learning the score function using SGD, and allows them to determine an optimal stopping time.

However, the study of these authors is limited to the diffusion model and did not consider the two-stage architecture of LDMs. Furthermore, the relationship between the data dimension and the derived optimal stopping time remained unexplored in their findings. In contrast, our work directly investigates the influence of the latent dimension on the optimal stopping time by incorporating an autoencoder into the diffusion model framework. We also demonstrate the need of early stopping without discretization of the backward diffusion process.

## 3 Notations and problem setup

This section introduces the mathematical formalism of diffusion models in the considered setting.

**Latent Diffusion Models.** Let $p_0$ be an unknown distribution in $\mathbb{R}^D$. With a slight abuse of notation, we use in the following the same notation for a distribution and its density function. The goal of diffusion models is to generate new observations following $p_0$, given an i.i.d. sample $(X_1, \ldots, X_n)$ drawn from $p_0$. The mechanism is as follows. Given a final diffusion time $T > 0$, a latent dimension $d \leq D$, an orthogonal matrix $P \in \mathbb{R}^{d \times D}$, and a scalar function $w : [0, T] \to \mathbb{R}$, the latent forward variance-preserving (VP) SDE (Song et al., 2020) is defined by

$$dP\overrightarrow{X}_t = -w_t^2 P\overrightarrow{X}_t dt + \sqrt{2w_t^2} dP\overrightarrow{W}_t, \quad P\overrightarrow{X}_0 \sim P_{\#}p_0, \tag{1}$$

where $\overrightarrow{W}_t$ is a standard $D$-dimensional Brownian motion. The role of the matrix $P$ is to perform linear dimension reduction. Two special cases are of interest: first, if $d = D$ and $P$ is the identity matrix, we recover the standard formulation of diffusion models. Second, if $P$ projects on the first few principal components of the sample covariance matrix, this amounts to performing principal component analysis (PCA, Jolliffe, 2002). This projection is equivalent to linear autoencoders (Plaut, 2018), and there exists a pseudo-inverse $P^+ \in \mathbb{R}^{D \times d}$ which allows us to map sample back to $\mathbb{R}^D$.

Letting $s_P$ be the score function of $P\overrightarrow{X}_t$, i.e., $s_P(x, t) = \nabla \log p_P(x, t)$ where $p_P(\cdot, t)$ is the density function of $P\overrightarrow{X}_t$, the forward diffusion can be reversed in time using the backward process

$$dP\overleftarrow{X}_t = (w_{T-t}^2 P\overleftarrow{X}_t + 2w_{T-t}^2 s_P(P\overleftarrow{X}_t, T-t))dt + \sqrt{2w_{T-t}^2} dP\overleftarrow{W}_t, \quad P\overleftarrow{X}_0 \sim P_{\#}p_T, \tag{2}$$

where $\overleftarrow{W}_t$ is a standard $D$-dimensional Brownian motion. This means that the marginal distribution of $P\overleftarrow{X}_{T-t}$ matches the marginal distribution of $P\overrightarrow{X}_t$ (Anderson, 1982). Hence running the backward diffusion allows to generate a sample from $\overleftarrow{X}_T \sim P_{\#}p_0$, and then the pseudo-inverse $P^+$ can be used to map the generated sample back to $\mathbb{R}^D$. Importantly, this procedure requires knowledge of $s_P$, which can be estimated using the training sample.

**Problem setup.** In the following, we assume that $p_0$ is a $D$-dimensional centered Gaussian distribution with independent components, i.e.,

$$p_0 = \mathcal{N}(0, \Sigma), \quad \text{and} \quad \Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_D^2), \tag{3}$$

where $\sigma_1 \geq \ldots \geq \sigma_D > 0$. This specific setting simplifies our study but still provides important insights for more general distributions.

We consider a hierarchy of latent spaces with increasing dimension $d$ from 1 to $D$. This corresponds to taking the matrix $P$ in the VP-SDE (1) as the orthogonal projection $P_d$ onto the first $d$ dimensions. When $p_0$ is a Gaussian distribution, $P_d\overleftarrow{X}_{T-t}$ and $P_d\overrightarrow{X}_t$ also follow Gaussian distributions

$$P_d\overleftarrow{X}_{T-t} \overset{\mathcal{D}}{=} P_d\overrightarrow{X}_t \sim \mathcal{N}(0, P_d(a_t^2 I_D + b_t^2\Sigma)P_d^\top) = \mathcal{N}(0, a_t^2 I_d + b_t^2 P_d\Sigma P_d^\top), \tag{4}$$

where the covariance matrix of $P_d\overleftarrow{X}_{T-t}$ effectively zeros out the last $D - d$ dimensions and $a_t = \sqrt{1 - b_t^2}$ while $b_t = e^{-\int_0^t w_t^2 dt}$. A typical choice is the Ornstein-Uhlenbeck process, where $w_t \equiv 1$, which implies $a_t = \sqrt{1 - e^{-2t}}$ and $b_t = e^{-t}$.

When $p_0$ is a Gaussian distribution, the score function $\nabla \log p_t$ is completely determined by the covariance matrix $\Sigma$. Indeed, in (4) we have

$$s_{P_d}(x, t) = -(a_t^2 I_d + b_t^2 P_d\Sigma P_d^\top)^{-1}x, \quad x \in \mathbb{R}^d.$$

Therefore, we may simplify the learning of the score function to the task of covariance matrix estimation. Since we assumed that the components of $p_0$ are independent, we consider a class of estimators consisting of diagonal matrices $\hat{\Sigma} = \mathrm{diag}(\hat{\sigma}_1^2, \dots, \hat{\sigma}_D^2)$, where its diagonal elements are the estimated variances given by $\frac{1}{n}\sum_{i=1}^{n} X_{id}^2$ for all $d \in \{1, \dots, D\}$ where $X_i = (X_{i1}, \dots, X_{iD}) \in \mathbb{R}^D$. Furthermore, we assume $\hat{\sigma}_1 \geq \dots \geq \hat{\sigma}_D > 0$, which is satisfied with high probability when $n$ is large enough (up to permutations of the indices in case of equality of some of the variances).

Plugging the estimated covariance matrix in the final distribution of $P_d\overrightarrow{X_T}$ and in the score function $s_{P_d}$, we can define the estimated sampling procedure

$$dP_d\overleftarrow{\hat{X}}_t = (w_{T-t}^2 P_d\overleftarrow{\hat{X}}_t + 2w_{T-t}^2 \hat{s}_{P_d}(P_d\overleftarrow{\hat{X}}_t, T-t))dt + \sqrt{2w_{T-t}^2}\, dP_d\overleftarrow{W_t}, \quad P_d\overleftarrow{\hat{X}}_0 \sim (P_d)_{\#}\hat{p}_T, \tag{5}$$

where

$$\hat{s}_{P_d}(x,t) = -(a_t^2 I_d + b_t^2 P_d \hat{\Sigma} P_d^\top)^{-1} x \quad \text{and} \quad (P_d)_{\#}\hat{p}_T \sim \mathcal{N}(0, a_T^2 I_d + b_T^2 P_d \hat{\Sigma} P_d^\top).$$

By the same derivations as above, we have the following identity

$$P_d\overleftarrow{\hat{X}}_t \sim \mathcal{N}(0, a_t^2 I_d + b_t^2 P_d \hat{\Sigma} P_d^\top).$$

In practical applications, some numerical scheme is used to solve the backward SDE (5). Typical choices include replacing the initial distribution $\hat{p}_T$ in the SDE (5) by a standard Gaussian distribution $\mathcal{N}(0, I_D)$, which is a valid approximation when $T$ is large. In addition, for numerical stability, the backward diffusion is early stopped at a preset time $T - \delta$ (see, e.g., Yang et al., 2023).

We quantify the distance between distributions by the Wasserstein-2 distance (Villani, 2008), which is equivalent in the Gaussian case to the Fréchet distance (Heusel et al., 2017):

$$d_F^2(\mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}(\mu_2, \Sigma_2)) = \|\mu_1 - \mu_2\|_2^2 + \mathrm{tr}(\Sigma_1 + \Sigma_2 - 2(\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2})^{1/2}). \tag{6}$$

With a small abuse of notation, we use $d_F(X, Y)$ to denote the distance between the distributions of the random variables $X$ and $Y$. This distance is the de facto standard to evaluate generative models. Other distances such as Kullback–Leibler (KL) divergence are not appropriate in our setup since the generated distribution is degenerated, and we expect Sliced-Wasserstein (SW) and Maximum Mean Discrepancy (MMD) distances to have a similar behavior as the Fréchet distance—see Appendix D for experimental results with these two metrics.

## 4 Optimal dimension reduction and stopping time

In this section, we address the important question of how dimensionality reduction affects the diffusion process with respect to the intrinsic geometric structure of the data. Our analysis focuses on selecting the rank $d$ of the projection matrix $P_d$ and the stopping time of the diffusions (2) and (5).

### 4.1 An analysis of non-monotonic behavior of Fréchet distance

This subsection examines the non-monotonic behavior of the Fréchet distance as a function of diffusion timesteps, challenging the intuitive expectation of monotonic evolution. The common belief of monotonicity (Jayasumana et al., 2024) implies that a stopping time closer to $T$ consistently yields a smaller Fréchet distance. First, we derive a necessary and sufficient condition for this non-monotonicity to occur in the scenario where the target distribution is Gaussian, as in (3). The proof of this result, as well as those of the subsequent ones, can be found in the Appendix.

**Proposition 1.** *Let $P_d\overleftarrow{X}_t$ and $P_d\overleftarrow{\hat{X}}_t$ be given as in (2) and (5), respectively. For $d \in \{1, \dots, D\}$, the Fréchet distance $d_F(P_d^\top P_d\overleftarrow{X}_t, \overrightarrow{X}_0)$ is non-increasing with respect to $t$. On the other hand, $d_F(P_d^\top P_d\overleftarrow{\hat{X}}_t, \overrightarrow{X}_0)$ is non-increasing if and only if*

$$\sum_{d'=1}^{d} (1 - \frac{\sigma_{d'}}{\hat{\sigma}_{d'}})(1 - \hat{\sigma}_{d'}^2) \geq 0. \tag{7}$$

Roughly speaking, the variance of each backward diffusion component $P_d\overleftarrow{\hat{X}}_t$ scales monotonically from an initial value close to 1 ($a_T^2 + b_T^2\hat{\sigma}_{d'}^2 \approx 1$) to the estimated variance $\hat{\sigma}_{d'}^2$ or any values satisfying (7), e.g., when $\hat{\sigma}_{d'}^2$ are given by an oracle. The distance $d_F(P_d^\top P_d\overleftarrow{\hat{X}}_t, \overrightarrow{X_0})$ is therefore minimized when the process variance is closest to the true set of variances $(\sigma_{d'}^2)_{1\leq d'\leq d}$, which happens before time $T$ under condition (7).

For a clearer understanding, consider the scenario where $p_0$ is a distribution lying in a linear subspace that is isomorphically equivalent to $\mathbb{R}^{d_0}$. In other words, suppose that $\sigma_D = \ldots = \sigma_{d_0+1} = 0$, and also $\hat{\sigma}_D = \ldots = \hat{\sigma}_{d_0+1} = 0$. Let us first consider the case where there is no projection, i.e., $d = D$. Then, the left-hand side of (7) can be rewritten

$$\sum_{d'=1}^{D}(1-\frac{\sigma_{d'}}{\hat{\sigma}_{d'}})(1-\hat{\sigma}_{d'}^2) = \sum_{d'=1}^{d_0}(1-\frac{\sigma_{d'}}{\hat{\sigma}_{d'}})(1-\hat{\sigma}_{d'}^2) + D - d_0.$$

For large enough $n$ and with high probability, $|\sigma_{d'} - \hat{\sigma}_{d'}| \leq \hat{\sigma}_{d'}$ for every $d' \in \{1,\ldots,d_0\}$, thus

$$\sum_{d'=1}^{D}(1-\frac{\sigma_{d'}}{\hat{\sigma}_{d'}})(1-\hat{\sigma}_{d'}^2) \geq \sum_{d'=1}^{d_0}\big(-1-\max_{d'\in\{1,\ldots,d_0\}}\hat{\sigma}_{d'}^2\big) + D - d_0 = D - \big(2+\max_{d'\in\{1,\ldots,d_0\}}\hat{\sigma}_{d'}^2\big)d_0.$$

The last term is positive as long as the ambient dimension $D$ is large enough. Therefore, in this context, $d_F(\overleftarrow{\hat{X}}_t, \overrightarrow{X}_0)$ is non-increasing. However, if projecting the diffusion onto the $d_0$-dimensional linear subspace in which the data distribution lies, the $D - d_0$ term in the computation above vanishes, and we are left with the sum up to $d_0$. Then the behavior of the Fréchet distance is linked to how the model estimates the variances of the data. If, for most $d'$, the sign of $1 - \sigma_{d'}/\hat{\sigma}_{d'}$ matches the sign of $1 - \hat{\sigma}_{d'}^2$, the Fréchet distance exhibits monotonic behavior. Conversely, if most of the signs differ, the Fréchet distance is non-monotonic. In addition, for a given estimation error, non-monotonicity is more likely to occur when the latent dimension is small.

This insight suggests that early stopping can improve the backward diffusion process, bringing the generated distribution closer to the data distribution. We next ask the reverse question: given a stopping time $t$, what is the optimal latent dimension?

### 4.2 Optimal projection at time $t$

In this subsection, we continue the study of the interaction of the dimension of projection and the stopping time. In contrast to the previous subsection, we show that for each fixed time $t$, there exists an optimal projection $P_d$. We still consider Gaussian data with independent components (3). Recall that $a$ is defined in (4) and that it is an increasing map from $[0, T]$ to $[0, a_T]$. We then let $\bar{a}^{-2} : \mathbb{R}\cup\{\infty\} \to [0, T]$ be the extended inverse function of $a^2$ (see plot in Figure 4), meaning that

$$\bar{a}^{-2}(x) = \begin{cases} 0, & \text{for } x < 0, \\ a^{-2}(x), & \text{for } x \in [0, a_T^2], \\ T, & \text{for } x \in (a_T^2, \infty]. \end{cases} \tag{8}$$

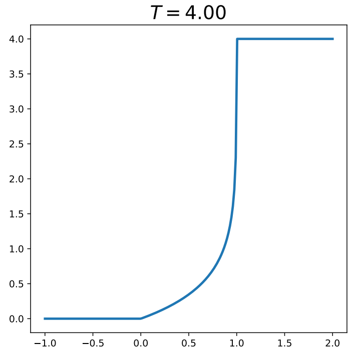

Figure 4: $\bar{a}^{-2}$ in the Ornstein-Uhlenbeck process.

In particular, for $t \in [0, T]$, $\bar{a}^{-2}(a_t^2) = t$. For $d \in \{2,\ldots,D\}$, we then let

$$t_d = T - \bar{a}^{-2}\Big(\frac{3\sigma_d^2}{(1-\sigma_d^2)_+}\Big) \quad \text{and} \quad \hat{t}_d = T - \bar{a}^{-2}\Big(\frac{4\sigma_d^2 - \hat{\sigma}_d^2}{(1-\hat{\sigma}_d^2)_+}\Big).$$

By convention, we let $\hat{t}_1 = t_1 = 0$ and $\hat{t}_{D+1} = t_{D+1} = T$. Observe that the times $t_d$ are in increasing order and between 0 and $T$. Given these time partitions, we can characterize the optimal projection dimension, both for the exact backward process and the one incorporating score estimation, with the aim of minimizing the distance between the generated and target distributions.

**Proposition 2.** *Assume that* $0 < \sigma_D < \cdots < \sigma_1$. *Then, for* $d \in \{1,\ldots,D\}$ *and* $t \in [t_d, t_{d+1})$,

$$d_F(P_d^\top P_d\overleftarrow{X}_t, \overrightarrow{X_0}) = \min_{d'\in\{1,\ldots,D\}} d_F(P_{d'}^\top P_{d'}\overleftarrow{X}_t, \overrightarrow{X_0}).$$

*Furthermore, with high probability, the $\hat{\sigma}_d$ and the $\hat{t}_d$ are well-ordered. In this case, for* $t \in [\hat{t}_d, \hat{t}_{d+1})$

$$d_F(P_d^\top P_d \overleftarrow{\hat{X}}_t, \overrightarrow{X_0}) = \min_{d' \in \{1,\dots,D\}} d_F(P_{d'}^\top P_{d'} \overleftarrow{\hat{X}}_t, \overrightarrow{X_0}).$$

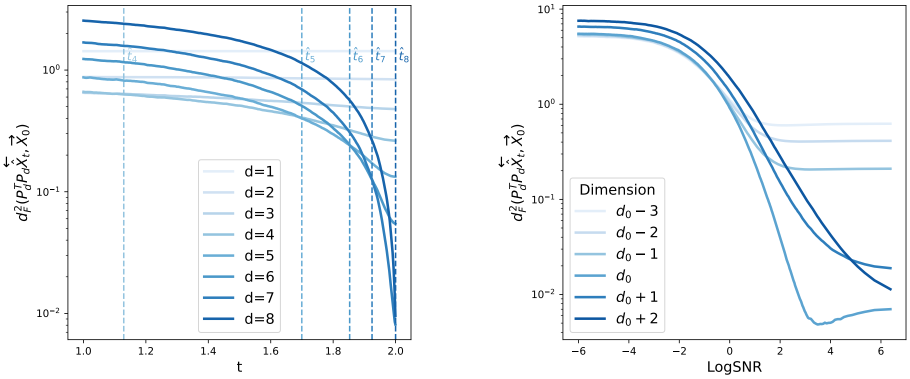

Figure 5: Plots of $d_F^2(P_d^\top P_d \overleftarrow{\hat{X}}_t, \overrightarrow{X_0})$ as a function of the diffusion time $t$, for two sets of variances $(\sigma_1, \dots, \sigma_D)$. (left) All the $\sigma_i$ are nonzero. As expected from Proposition 2, the $d$-dimensional projection is optimal in $[t_d, t_{d+1})$. (right) The data is supported on a linear subspace of dimension $d_0 = 4$ with $D = 6$. As expected from Proposition 3, we observe that the minimum distance is achieved in dimension $d_0$ and with early stopping. LogSNR in the $x$-axis is a remapping of time $t$, defined as $\log(b_t^2/a_t^2)$, which we use to increase readability. Experimental details are in Appendix D.

Proposition 2 quantifies a direct link between early stopping in the backward diffusion process and dimensionality reduction. It reveals a time-dependent trade-off: distributions at early stages of the backward process are best approximated in lower-dimensional spaces, while higher dimensions become necessary to faithfully reconstruct the data as $t \to T$, as illustrated in Figure 5 (left). In other words, at an early time step, projecting onto an unnecessarily high-dimensional space can introduce more noise than signal, making a lower-dimensional representation more accurate. Notice that when $4\sigma_d^2 \geq 1$, both $t_d$ and $\hat{t}_d$ are equal to 0. This implies that a component whose variance is sufficiently large should always be included in the projection, aligning with the intuition that major components are essential for representation. These results hold for backward processes using scores based on either the true or empirical variances.

We next characterize the behavior of the optimal latent dimension and stopping time when data lies on a $d_0$-dimensional subspace, providing a similar result to Proposition 2. This analysis allows us to precisely determine these two key parameters, as shown next.

**Proposition 3.** *Assume that* $\Sigma = \mathrm{diag}(\sigma^2, \dots, \sigma^2, 0, \dots, 0)$ *with the last* $D - d_0$ *entries equal to* $0$. *Let* $\varepsilon \in (0, 1)$. *Then, there exists* $\hat{\delta}_{d_0} \in [0, T]$ *such that with probability* $1 - 2d_0 e^{-\frac{n}{8}}$,

$$d_F(P_{d_0}^\top P_{d_0} \overleftarrow{\hat{X}}_{T-\hat{\delta}_{d_0}}, \overrightarrow{X_0}) = \min_{\substack{t \in [0,T] \\ d' \in \{1,\dots,D\}}} d_F(P_{d'}^\top P_{d'} \overleftarrow{\hat{X}}_t, \overrightarrow{X_0}).$$

The proposition shows that the optimal generation strategy for data with a low-rank structure involves both early stopping and projection (see Figure 5 (right) for an illustration). The proof indicates that, under the non-monotonicity condition of Proposition 1, the optimal early stopping time $T - \hat{\delta}_{d_0}$ is strictly before $T$. Beyond preventing numerical instability as $t \to T$ (e.g., Yang et al., 2023), Proposition 3 thus offers a new justification for early stopping. In other words, stopping at a specific time $\hat{\delta}_{d_0}$ is not merely a practical fix, but an optimal strategy to improve generation quality by minimizing the distance between the generated and true data distributions.

Furthermore, this result confirms the intuition that confining the generative process to the dataset's intrinsic dimensionality is the most effective approach for low-rank data. This strategy is not only computationally more efficient than running the diffusion in the ambient space, but also enhances generation quality by avoiding the noise introduced by superfluous dimensions.

# 5 Performance of the score matching ERM

In the previous section, we analyzed the properties of diffusion processes with a score tailored to independent Gaussian distributions involving either exact or plugged-in estimated variances. In practice, the score is rather *learned* by solving a regression problem called score matching. Specifically, given a training sample $(X_1, \dots, X_n)$ independently drawn from the data distribution $p_0$, the empirical score matching objective writes

$$\mathcal{R}(s) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{t \sim \mathcal{T}, \varepsilon \sim \mathcal{N}(0, I_D)} \left\| s(b_t X_i + a_t \varepsilon, t) + \frac{\varepsilon}{a_t} \right\|^2, \tag{9}$$

for some absolutely continuous distribution $\mathcal{T}$ with positive mass over $[0, T]$, and where the predictor $s : \mathbb{R}^D \times \mathbb{R} \to \mathbb{R}$ belongs to some hypothesis class $\mathcal{F}_C$, typically a neural network architecture. In our context, recall that the score function of a Gaussian distribution with diagonal covariance $\Sigma$ is

$$\nabla \log p_t(x) = -(a_t^2 I_D + b_t^2 \Sigma)^{-1} x, \tag{10}$$

which takes the form of a time-dependent diagonal matrix multiplied by $x$. Thus a natural choice of hypothesis class given the form of the true score function (10) is, for $C > 1$,

$$\begin{aligned} \mathcal{F}_C = \big\{ s_M : \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}^d : s_M(x, t) = -M(t)x, \qquad & M(t) = \operatorname{diag}(m_1(t), \dots, m_D(t)), \\ & m_i \in \mathcal{L}_2(\mathbb{R}_+, \mathbb{R}), \|m_i\|_\infty < C \big\}. \end{aligned}$$

The assumption of $C > 1$ is essential since we start the backward diffusion from a standard Gaussian distribution whose score function is the identity function. We introduce the norm constraint on the weights to account for two phenomena. First, the norm of the true score function (10) blows up for times close to 0 (in particular if the covariance matrix is singular or close to singular), which is known to create numerical instabilities (Lu et al., 2023; Yang et al., 2023). This is mitigated in practice for instance by early stopping the diffusion. Here, we implement this mitigation by capping the weight norm. Second, it is known that gradient descent has an implicit bias towards learning low-norm solutions. Although quantifying this effect is beyond the scope of this paper, the explicit weight constraint provides an analytically tractable analogue. More precisely, one can easily derive the following explicit formula for the minimizer of the score matching over $\mathcal{F}_C$.

**Proposition 4.** *Let $\hat{\sigma}_d^2 = \frac{1}{n} \sum_{i=1}^n X_{id}^2$ be the empirical variance for the $d$-th component of the training data. Then the minimizer of the score matching objective* (9) *over $\mathcal{F}_C$ is given by $\hat{M}(t) = \operatorname{diag}(\hat{m}_1(t), \dots, \hat{m}_D(t))$ where, for $d \in \{1, \dots, D\}$,*

$$\hat{m}_d(t) = \min\left( C, \frac{1}{a_t^2 + b_t^2 \hat{\sigma}_d^2} \right).$$

Our goal in the following is to characterize the optimal latent dimension when using the score defined by Proposition 4. For this purpose, as before, we quantify the distance between the data distribution and the distribution generated by the backward process for $d \in \{1, \dots, D\}$. We do not consider early stopping here, to focus on the influence of the regularization parameter $C$ on the choice of the latent dimensionality. For simplicity, we keep the data distribution $p_0$ to be a Gaussian distribution with independent component, and specialize to the Ornstein-Uhlenbeck process. In this case, the sample $\overleftarrow{\tilde{X}}_t$ are generated by the backward SDE for $t \in [0, T]$

$$d\overleftarrow{\tilde{X}}_t = (\overleftarrow{\tilde{X}}_t + 2 s_{\hat{M}}(\overleftarrow{\tilde{X}}_t, T - t))dt + \sqrt{2} d\overleftarrow{W}_t, \quad \overleftarrow{\tilde{X}}_0 \sim \mathcal{N}(0, I_D).$$

Note that we consider the standard setting in which the backward process starts from a standard Gaussian. We can then characterize the optimal projection for the latent diffusion, as shown next.

**Proposition 5.** *Define $1 \leq d_1 \leq d_2 \leq D$ as follows:*

$$d_1 = \max\{d' \in \{1, \dots, D\} : 1/C \leq \hat{\sigma}_{d'}^2\} \text{ and } d_2 = \min\Big\{ d' \in \{1, \dots, D\} : \frac{1}{2C - 1} \geq 4\sigma_{d'}^2 \Big\}.$$

*(If the corresponding set in their definition is empty, we let $d_1 = 1$ and $d_2 = D$, respectively.) Then, with high probability, there exists an optimal projection dimension $d_1 \leq d_{\min} \leq d_2$ such that*

$$d_F(P_{d_{\min}}^\top P_{d_{\min}} \overleftarrow{\tilde{X}}_T, \overrightarrow{X_0}) = \min_{d' \in \{1, \dots, D\}} \big\{ d_F(P_{d'}^\top P_{d'} \overleftarrow{\tilde{X}}_T, \overrightarrow{X_0}) \big\}.$$

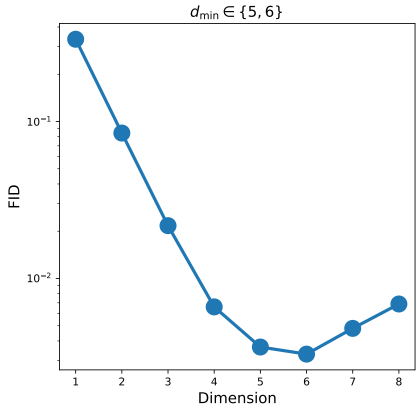

Figure 6: Fréchet distance of the sample generated by a diffusion model, where the data follows a 10-dimensional Gaussian distribution with exponentially decaying eigenvalues, while the score is a linear function trained with sample projected onto a lower dimension. Corollary 1 correctly predicts the optimal projection dimension, here $d_{\min} \in \{5, 6\}$.

To gain intuition into Proposition 5, let us consider some illustrative cases depending on the weight constraint $C$ (for the full derivation of these cases, see Appendix A.6). First, when $C = \infty$ and the data covariance is non-singular, we get $d_1 = d_2 = D$, thus Proposition 5 suggests to take the projection matrix $I_D$, which is expected since $\mathcal{F}_C$ is then large enough to contain the true empirical score function. Second, consider the scenario when the data distribution lies on a linear subspace of dimension $d_0$. If $C$ is large enough, we obtain $d_1 = d_2 = d_0$, meaning that the projection onto the data subspace is the optimal sampling strategy, which is in line with Proposition 3. Finally, the optimal projection can also be made explicit for exponentially-decaying covariance spectrum.

**Corollary 1.** *Let $\lambda > 16$. Assume that $\Sigma = \operatorname{diag}(\lambda^{-1}, \dots, \lambda^{-D})$ and $\lambda \leq C \leq \lambda^D$. Let $d \in \{1, \dots, D\}$ be such that $\hat{\sigma}_{d+1}^2 \leq 1/C \leq \hat{\sigma}_d^2$. Then, with $n$ large enough and high probability,*

$$d_{\min} \in \{d, d+1\}.$$

Interestingly, when the covariance structure decays exponentially, the capacity of the score-function class—captured here by the parameter $C$—is directly tied to the optimal projection dimension. The latter scales logarithmically in the parameter $C$ since $\hat{\sigma}_{d_{\min}}^2 = \lambda^{-d_{\min}} \approx 1/C$. The result is illustrated in Figure 6. An analogous result can be derived for covariance structures with power-law decay.

## 6 Generalization to arbitrary Gaussian distributions

We now explain how to generalize some of our preceding analysis from Gaussian distributions with diagonal covariance matrices to the more general case $p_0 = \mathcal{N}(0, \Sigma)$ for arbitrary $\Sigma$, and the backward processes $\overleftarrow{X}_t$ and $\overleftarrow{\hat{X}}_t$ given as in (2) and (5) with the new general data distribution $p_0$. Our goal is to establish a result analogous to Proposition 2, that is, to characterize the optimal latent dimension given a stopping time of the diffusion process.

To this end, let $\Sigma = O \Lambda O^\top$ be the eigen decomposition of $\Sigma$, where $O$ is an orthogonal matrix and $\Lambda$ is the diagonal matrix of eigenvalues, which we assume are distinct and ordered $\sigma_1^2 > \ldots > \sigma_D^2 > 0$. As in Section 4, we define a time partition by setting $t_1 = 0$ and $t_{D+1} = T$, and defining the intermediate timesteps for $d \in \{2, \dots, D\}$ as:

$$t_d = T - \bar{a}^{-2} \left( \frac{3\sigma_d^2}{(1 - \sigma_d^2)_+} \right),$$

where $\bar{a}^{-2}$ is given in (8). This definition, combined with the ordering of the eigenvalues, yields a sequence $0 = t_1 \leq t_2 \leq \cdots \leq t_D \leq t_{D+1} = T$. We show next that for this general Gaussian case, PCA projection onto $d$ components is optimal precisely within the interval $[t_d, t_{d+1})$.

**Proposition 6.** *For $2 \leq d \leq D$ and $t \in [t_d, t_{d+1})$, we have*

$$d_F(OP_d^\top P_d O^\top \overleftarrow{X}_t, \overrightarrow{X_0}) = \min_{d' \in \{1,\ldots,D\}} d_F(OP_{d'}^\top P_{d'} O^\top \overleftarrow{X}_t, \overrightarrow{X_0}).$$

However, in practical applications, one rarely has access to the true underlying covariance matrix $\Sigma$ or its eigenbasis $O$. Instead, one must rely on estimations derived from observed data, where PCA is commonly used. Denote $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$ to be the empirical covariance matrix. Applying a spectral decomposition yields $\hat{\Sigma} = \hat{O}\text{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_D^2)\hat{O}^\top$, where $\hat{O}$ contains the orthonormal eigenvectors and $\hat{\sigma}_D^2 < \ldots < \hat{\sigma}_1^2$ are the corresponding eigenvalues. Denote $S(\Sigma) = \sum_{d'=1}^D \max(\sigma_d, \sigma_d^2)$. For $u \geq 0$ and $d \in \{2, \ldots, D\}$, we let $\hat{T}_d(u)$ and $\hat{t}_d(u)$ be

$$\hat{T}_d(u) = T - \bar{a}^{-2}\left(\frac{\hat{\sigma}_d^2 - 4S(\Sigma)\varepsilon_u + 2\hat{\sigma}_d\sqrt{\hat{\sigma}_d^2 - 4S(\Sigma)\varepsilon_u}}{(1-\hat{\sigma}_d^2)_+}\right),$$

$$\hat{t}_d(u) = T - \bar{a}^{-2}\left(\frac{\hat{\sigma}_d^2 + 4S(\Sigma)\varepsilon_u + 2\hat{\sigma}_d\sqrt{\hat{\sigma}_d^2 + 4S(\Sigma)\varepsilon_u}}{(1-\hat{\sigma}_d^2)_+}\right),$$

where $\varepsilon_u = \frac{8C}{3}(\sqrt{\frac{D+u}{n}} + \frac{D+u}{n})$. We assume that $\varepsilon_u$ is sufficiently small (i.e., $n$ large enough) so that the square root in the definition above is well-defined and the argument of $\bar{a}^{-2}$ is positive. By convention, we set $\hat{T}_1(u) = 0$ and $\hat{t}_{D+1}(u) = T$. Thus, for small $\varepsilon_u$, these timesteps are ordered as

$$0 = \hat{T}_1(u) < \hat{t}_2(u) < \hat{T}_2(u) < \cdots < \hat{t}_D(u) < \hat{T}_D(u) < \hat{t}_{D+1}(u) = T.$$

We are now in a position to describe the optimal projection strategy at each stopping time.

**Proposition 7.** *For $d \in \{1, \ldots, D\}$ and any $t \in [\hat{T}_d(u), \hat{t}_{d+1}(u)]$, with probability $1 - 2e^{-u}$,*

$$d_F(\hat{O}P_d^\top P_d \hat{O}^\top \overleftarrow{\hat{X}}_t, \overrightarrow{X_0}) = \min_{d' \in \{1,\ldots,D\}} d_F(\hat{O}P_{d'}^\top P_{d'} \hat{O}^\top \overleftarrow{\hat{X}}_t, \overrightarrow{X_0}).$$

This proposition generalizes the result of Proposition 2 to the case of a general Gaussian data distribution. The analysis reveals that, for any latent dimension $d$, there exists a time interval where a $d$-dimensional projected diffusion process minimizes the distance to the target distribution with high probability. Notably, this result is consistent with our previous conclusions. In the idealized scenario where the variance estimation error is zero (i.e., $\varepsilon_u = 0$, implying $\hat{\Sigma} = \Sigma$,) the formula for the optimal time $\hat{t}_d = \hat{T}_d$ simplifies precisely to the one derived in Proposition 6.

## 7 Conclusion

This paper provides a theoretical analysis of optimal stopping time in latent diffusion models, showing its critical dependence on latent space dimensionality and its interaction with other hyperparameters of the diffusion process, such as weight regularization in the score matching phase. Our results focus on Gaussian distributions, given their tractability and prominence in prior theoretical works (Pierret & Galerne, 2024; Hurault et al., 2025). Taken together, these insights open compelling research directions, for deepening the theoretical properties of latent diffusion models and assessing when they can match or surpass the sampling quality of standard diffusion models. We highlight that a future research direction is to extend the analysis of this phenomenon to more modern architectures such as EDMs where the model enforces unit-variance inputs and outputs at each noise level.

## References

Beatrice Achilli, Luca Ambrogioni, Carlo Lucibello, Marc Mézard, and Enrico Ventura. Memorization and generalization in generative diffusion under the manifold hypothesis. *arXiv:2502.09578*, 2025.

Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12:313–326, 1982.

Iskander Azangulov, Georgios Deligiannidis, and Judith Rousseau. Convergence of diffusion models under the manifold hypothesis in high-dimensions. *arXiv:2402.16439*, 2024.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22563–22575, 2023.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/google/jax`.

Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 2406–2422. Curran Associates, Inc., 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12873–12883, 2021.

Antony J. George, Ricardo Veiga, and Nicolas Macris. Analysis of diffusion models for manifold data. *arXiv:2502.04339*, 2025.

Malay Ghosh. Exponential tail bounds for chisquared random variables. *Journal of Statistical Theory and Practice*, 15:35, 2021.

Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, 10: 041044, 2020.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Roger A Horn and Charles R Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 2012.

Chin-Wei Huang, Milad Aghajohari, Joey Bose, Prakash Panangaden, and Aaron C Courville. Riemannian diffusion models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 2750–2761. Curran Associates, Inc., 2022.

Samuel Hurault, Matthieu Terris, Thomas Moreau, and Gabriel Peyré. From score matching to diffusion: A fine-grained error analysis in the Gaussian setting. *arXiv:2503.11615*, 2025.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9307–9315, 2024.

Ian T Jolliffe. *Principal Component Analysis*. Springer, New York, 2002.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv:1312.6114*, 2013.

Theodoros Kouzelis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. EQ-VAE: Equivariance regularized latent space for improved generative image modeling. *arXiv:2502.09509*, 2025.

Gen Li and Yuling Yan. Adapting to unknown low-dimensional structures in score-based diffusion models. *arXiv:2405.14861*, 2024.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv:2209.03003*, 2022.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Yubin Lu, Zhongjian Wang, and Guillaume Bal. Mathematical analysis of singularities in the diffusion model under the submanifold assumption. *arXiv:2301.07882*, 2023.

Emile Pierret and Bruno Galerne. Diffusion models for Gaussian distributions: Exact solutions and Wasserstein errors. *arXiv:2405.14250*, 2024.

Elad Plaut. From principal subspaces to principal components with linear autoencoders. *arXiv:1804.10253*, 2018.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241. Springer, 2015.

Simo Särkkä and Arno Solin. *Applied Stochastic Differential Equations*, volume 10. Cambridge University Press, Cambridge, 2019.

Ivan Skorokhodov, Sharath Girish, Benran Hu, Willi Menapace, Yanyu Li, Rameen Abdal, Sergey Tulyakov, and Aliaksandr Siarohin. Improving the diffusability of autoencoders. *arXiv:2502.14831*, 2025.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv:2011.13456*, 2020.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 1415–1428. Curran Associates, Inc., 2021.

Ruoqi Tang and Yun Yang. Adaptivity of diffusion models to manifold structures. In Yingzhen Li Sanjoy Dasgupta, Stephan Mandt (ed.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238, pp. 3169–3177. PMLR, 2024.

Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 11287–11302. Curran Associates, Inc., 2021.

Roman Vershynin. *High-Dimensional Probability: An introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018.

Cédric Villani. *Optimal Transport: Old and New*. Springer, New York, 2008.

Bin Xie and Gady Agam. Rethinking timesteps samplers and prediction types. *arXiv:2502.01990*, 2025.

Zhantao Yang, Ruili Feng, Han Zhang, Yujun Shen, Kai Zhu, Lianghua Huang, Yifei Zhang, Yu Liu, Deli Zhao, Jingren Zhou, et al. Lipschitz singularities in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv:2410.06940*, 2024.

Yifan Zhou, Zeqi Xiao, Shuai Yang, and Xingang Pan. Alias-free latent diffusion models: Improving fractional shift equivariance of diffusion latent space. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 34–44, 2025.

# Appendix

## A Proofs of results

### A.1 Proof of Proposition 1

We show the equivalent statement: $t \in [0, T] \mapsto d_F^2(P_d^\top P_d \overleftarrow{\hat{X}}_{T-t}, \overrightarrow{X_0})$ is non-decreasing if and only if (7) holds. We start by calculating the Fréchet distance $d_F^2(P_d^\top P_d \overleftarrow{\hat{X}}_{T-t}, \overrightarrow{X_0})$ by using (6):

$$\begin{aligned} d_F^2(P_d^\top P_d \overleftarrow{\hat{X}}_{T-t}, \overrightarrow{X}_0) &= \sum_{d'=d+1}^{D} \sigma_{d'}^2 + \sum_{d'=1}^{d} \left( b_t^2 \hat{\sigma}_{d'}^2 + a_t^2 + \sigma_{d'}^2 - 2\sigma_{d'} \sqrt{a_t^2 + b_t^2 \hat{\sigma}_{d'}^2} \right) \\ &= \sum_{d'=d+1}^{D} \sigma_{d'}^2 + \sum_{d'=1}^{d} \left( \sqrt{a_t^2 + (1 - a_t^2)\hat{\sigma}_{d'}^2} - \sigma_{d'} \right)^2. \end{aligned}$$

Since $t \mapsto a_t^2$ is strictly increasing with $a_0 = 0$, the monotonicity of $d_F^2(P_d^\top P_d \overleftarrow{\hat{X}}_{T-t}, \overrightarrow{X_0})$ with respect to $t$ is equivalent to the monotonicity with respect to $a_t^2$. By considering the function $f : [0, a_T^2] \to \mathbb{R}$ defined by

$$f(x) = \sum_{d'=1}^{d} \left( \sqrt{x + (1 - x)\hat{\sigma}_{d'}^2} - \sigma_{d'} \right)^2,$$

we see that $d_F^2(P_d^\top P_d \overleftarrow{\hat{X}}_{T-t}, \overrightarrow{X_0})$ is non-decreasing if and only if $f$ is non-decreasing. Additionally,

$$f'(x) = \sum_{d'=1}^{d} (\sqrt{x + (1 - x)\hat{\sigma}_{d'}^2} - \sigma_{d'}) \frac{1 - \hat{\sigma}_j^2}{\sqrt{x + (1 - x)\hat{\sigma}_{d'}^2}} = \sum_{d'=1}^{d} \Big( 1 - \frac{\sigma_{d'}}{\sqrt{x + (1 - x)\hat{\sigma}_{d'}^2}} \Big)(1 - \hat{\sigma}_{d'}^2),$$

and

$$f''(x) = \sum_{d'=1}^{d} \frac{\sigma_{d'}(1 - \hat{\sigma}_{d'}^2)^2}{2(x + (1 - x)\hat{\sigma}_{d'}^2)^{3/2}} > 0.$$

Hence, $f$ is convex so it is non-decreasing if and only if $f'(0) \geq 0$. Therefore,

$$d_F^2(P_d^\top P_d \overleftarrow{\hat{X}}_{T-t}, \overrightarrow{X_0}) = f(a_t^2) + \sum_{d'=d+1}^{D} \sigma_{d'}^2$$

is non-decreasing if and only if $f'(0) \geq 0$, i.e., if and only if $\sum_{d'=1}^{d} (1 - \frac{\sigma_{d'}}{\hat{\sigma}_{d'}})(1 - \hat{\sigma}_{d'}^2) \geq 0$. This shows the second statement of the proposition. The monotonicity of $d_F(P_d^\top P_d \overleftarrow{X}_{T-t}, \overrightarrow{X_0})$ can be shown by replacing $\hat{\sigma}_{d'}$ with $\sigma_{d'}$ in the derivative $f'$, which is 0 when $a_t = 0$.

### A.2 Proof of Proposition 2

The first part of Proposition 2 concerns the minimization of $d_F(P_d^\top P_d \overleftarrow{X}_t, \overrightarrow{X_0})$. Recall that $t_d = T - \bar{a}^{-2}\left(\frac{3\sigma_d^2}{1-\sigma_d^2}\right)$. To prove that $P_d \overleftarrow{X}_t$ achieves the minimal distance to the target for $t \in [t_d, t_{d+1})$ (where the time interval is fixed), we will demonstrate how the distance $d_F(P_d^\top P_d \overleftarrow{X}_t, \overrightarrow{X_0})$ behaves as a function of the projection dimension $d$. Specifically, we aim to show that, for any $d \in \{2, \ldots, D\}$,

$$d_F(P_d^\top P_d \overleftarrow{X}_t, \overrightarrow{X_0}) \leq d_F(P_{d-1}^\top P_{d-1} \overleftarrow{X}_t, \overrightarrow{X_0}) \quad \text{iff } t \geq t_d. \tag{11}$$

This inequality in turn implies that for a given $t$ in a fixed interval $[t_d, t_{d+1})$, the minimum distance $d_F(P_d^\top P_d \overrightarrow{X_t}, \overrightarrow{X_0})$ is attained by the projected process $P_d \overrightarrow{X_t}$ in dimension $d$.

To establish them, we first explicitly compute the Fréchet distance $d_F(P_d^\top P_d \overrightarrow{X_t}, \overrightarrow{X_0})$. Recall that the Fréchet distance between two zero-mean Gaussian distributions $\mathcal{N}(0, \Sigma_1)$ and $\mathcal{N}(0, \Sigma_2)$ is given by $\mathrm{Tr}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2})$, and that the covariance matrix of $P_d\overleftarrow{X}_t$ is equal to $P_d(a_{T-t}^2 I_d + b_{T-t}^2\Sigma)P_d$. Therefore, it is possible to calculate the Fréchet distance to the target for the projected processes directly, as, for any $d \in \{1, \dots, D\}$,

$$d_F^2(P_d^\top P_d \overleftarrow{X}_t, \overrightarrow{X}_0) = \sum_{j=1}^{D} \sigma_j^2 + \sum_{j=1}^{d}(a_{T-t}^2 + b_{T-t}^2\sigma_j^2) - 2\sum_{j=1}^{d}\sigma_j\sqrt{a_{T-t}^2 + b_{T-t}^2\sigma_j^2},$$

so that

$$\begin{aligned}
\Delta_{d,t} &:= d_F^2(P_d^\top P_d \overleftarrow{X}_t, \overrightarrow{X}_0) - d_F^2(P_{d-1}^\top P_{d-1} \overleftarrow{X}_t, \overrightarrow{X}_0) \\
&= b_{T-t}^2\sigma_d^2 + a_{T-t}^2 - 2\sigma_d\sqrt{a_{T-t}^2 + b_{T-t}^2\sigma_d^2}, \\
&= \sqrt{b_{T-t}^2\sigma_d^2 + a_{T-t}^2}\Big(\sqrt{b_{T-t}^2\sigma_d^2 + a_{T-t}^2} - 2\sigma_d\Big), \\
&= \sqrt{(1-a_{T-t}^2)\sigma_d^2 + a_{T-t}^2}\Big(\sqrt{(1-a_{T-t}^2)\sigma_d^2 + a_{T-t}^2} - 2\sigma_d\Big), \\
&= \sqrt{\sigma_d^2 + a_{T-t}^2(1-\sigma_d^2)}\Big(\sqrt{a_{T-t}^2(1-\sigma_d^2) + \sigma_d^2} - 2\sigma_d\Big).
\end{aligned}$$

We see that $\Delta_{d,t}$ has the same sign as the term in the parenthesis on the last line, which itself has the same sign as $a_{T-t}^2(1-\sigma_d^2) - 3\sigma_d^2$. Then,

- if $\sigma_d \geq 1$ or $\frac{3\sigma_d^2}{1-\sigma_d^2} \geq a_T^2$, $\Delta_{d,t}$ is non-positive for all $t \in [0, T]$, while $t_d = 0$ by definition;
- otherwise, $\Delta_{d,t}$ is non-positive if and only if $a_{T-t}^2 \leq \frac{3\sigma_d^2}{1-\sigma_d^2}$ which is equivalent to

$$T - t \leq a^{-2}\Big(\frac{3\sigma_d^2}{1-\sigma_d^2}\Big) = T - t_d.$$

Putting things together, we obtain that $d_F^2(P_d^\top P_d \overleftarrow{X}_t, \overrightarrow{X}_0) - d_F^2(P_{d-1}^\top P_{d-1} \overleftarrow{X}_t, \overrightarrow{X}_0)$ is non-positive iff $t \geq t_d$, which is exactly (11).

The proof in the case of estimated variances can be derived in a similar fashion as long as the estimated variances $\hat{\sigma}_i$ and times $\hat{t}_i$ are well-ordered, which happens with high probability for a sufficiently large sample.

### A.3 Proof of Proposition 3

We first state the full proposition.

**Proposition 8.** *Assume that $\Sigma = \mathrm{diag}(\sigma^2, \dots, \sigma^2, 0, \dots, 0)$ with the last $D - d_0$ entries equal to $0$, and the estimated variances are ordered as $\hat{\sigma}_1^2 \geq \hat{\sigma}_2^2 \geq \ldots \geq \hat{\sigma}_{d_0}^2$. Let $\varepsilon \in (0, 1)$. For*

$$t \in \Big[T - \bar{a}^{-2}\Big(\frac{3-\varepsilon}{1+\varepsilon}\frac{\hat{\sigma}_1^2}{1-\hat{\sigma}_1^2}\Big), T\Big),$$

*with probability $1 - 2d_0 e^{-\frac{\varepsilon^2 n}{8}}$, we have*

$$d_F(P_{d_0}^\top P_{d_0}\overleftarrow{\hat{X}}_t, \overrightarrow{X_0}) = \min_{d' \in \{1,\dots,D\}} d_F(P_{d'}^\top P_{d'}\overleftarrow{\hat{X}}_t, \overrightarrow{X_0}).$$

*If, in addition,*

$$\sum_{d'=1}^{d_0}(1 - \frac{\sigma}{\hat{\sigma}_{d'}})(1 - \hat{\sigma}_{d'}^2) < 0, \tag{12}$$

*then*

$$\sum_{d'=1}^{d_0}(1 - \frac{\sigma}{\sqrt{\hat{\sigma}_{d'}^2 + (1-\hat{\sigma}_{d'}^2)a_t^2}})(1 - \hat{\sigma}_{d'}^2) = 0,$$

*has a unique solution which we denote by $\hat{\delta}_{d_0}$. By convention, if the condition* (12) *is not satisfied, we set $\hat{\delta}_{d_0} = 0$. Then, with probability* $1 - 2d_0 e^{-\frac{\varepsilon^2 n}{8}}$,

$$d_F(P_{d_0}^\top P_{d_0} \overleftarrow{\hat{X}}_{T-\hat{\delta}_{d_0}}, \overrightarrow{X_0}) = \min_{\substack{t\in[0,T]\\ d'\in\{1,\ldots,D\}}} d_F(P_{d'}^\top P_{d'} \overleftarrow{\hat{X}}_t, \overrightarrow{X_0}).$$

Let $\varepsilon \in (0,1)$. We first note that according to Proposition 9, by the union bound, with probability $1 - 2d_0 e^{-\frac{\epsilon^2 n}{4(1+\epsilon)}} \geq 1 - 2d_0 e^{-\frac{\epsilon^2 n}{8}}$ we have $|\sigma^2 - \hat{\sigma}_d^2| \leq \varepsilon\sigma^2$ for all $d \in \{1, \ldots, d_0\}$. We work under this event in the remainder of the proof. In particular, for all $d \in \{1, \ldots, d_0\}$, $\sigma_d^2 = \sigma^2 \geq \hat{\sigma}_1^2/(1+\varepsilon)$. Thus, by separating cases depending on whether $4\sigma_d^2 \leq 1$, a short calculation gives that

$$\min\Big(1, \frac{\frac{4}{1+\varepsilon}\hat{\sigma}_1^2 - \hat{\sigma}_1^2}{1 - \hat{\sigma}_1^2}\Big) \leq \min\Big(1, \frac{4\sigma_d^2 - \hat{\sigma}_1^2}{1 - \hat{\sigma}_1^2}\Big) \leq \frac{4\sigma_d^2 - \hat{\sigma}_d^2}{1 - \hat{\sigma}_d^2}.$$

The last inequality is derived as follows: if $4\sigma_d^2 \leq 1$, then we use the fact that $x \mapsto \frac{a-x}{1-x}$ is non-increasing if $a < 1$. On the other hand, if $4\sigma_d^2 \geq 1$, then $\frac{4\sigma_d^2 - \hat{\sigma}_d^2}{1-\hat{\sigma}_d^2} \geq 1$. Hence, by the monotonic increase of $\bar{a}^{-2}$,

$$\begin{aligned}
\hat{t}_d &= T - \bar{a}^{-2}\Big(\frac{4\sigma_d^2 - \hat{\sigma}_d^2}{1 - \hat{\sigma}_d^2}\Big) \\
&= T - \bar{a}^{-2}\Big(\min\Big(1, \frac{\frac{4}{1+\varepsilon}\hat{\sigma}_1^2 - \hat{\sigma}_1^2}{1 - \hat{\sigma}_1^2}\Big)\Big) \\
&\geq T - \min\Big(T, \bar{a}^{-2}\Big(\frac{3-\varepsilon}{1+\varepsilon}\frac{\hat{\sigma}_1^2}{1 - \hat{\sigma}_1^2}\Big)\Big) \\
&= \max\Big(0, T - \bar{a}^{-2}\Big(\frac{3-\varepsilon}{1+\varepsilon}\frac{\hat{\sigma}_1^2}{1 - \hat{\sigma}_1^2}\Big)\Big) \\
&= T - \bar{a}^{-2}\Big(\frac{3-\varepsilon}{1+\varepsilon}\frac{\hat{\sigma}_1^2}{1 - \hat{\sigma}_1^2}\Big).
\end{aligned}$$

Thus, $t \geq T - \bar{a}^{-2}\Big(\frac{3-\varepsilon}{1+\varepsilon}\frac{\hat{\sigma}_1^2}{1-\hat{\sigma}_1^2}\Big)$ implies $t \geq \hat{t}_d$ for every $d \in \{1, \ldots, d_0\}$. On the other hand, $t < T = \hat{t}_d$ for all $d \in \{d_0+1, \ldots, D\}$ since $\sigma_d = \hat{\sigma}_d = 0$. From here we deduce the desired result applying Proposition 2.

In this second part, we study under the event where $|\sigma^2 - \hat{\sigma}_d^2| \leq \sigma^2$ for every $d \in \{1, \ldots, d_0\}$, which holds with probability $1 - 2d_0 e^{-n/8}$ by Proposition 9. To prove the desired result, we first show that the minimum of the distance $d_F(P_{d_0}^\top P_{d_0} \overleftarrow{\hat{X}}_t, \overrightarrow{X_0})$ is attained at $t = T - \hat{\delta}_{d_0}$, as per its definition. We consider two cases depending on whether condition (12) is satisfied. First, if condition (12) holds, the proof of Proposition 1 establishes that $a^2_{T-\hat{\delta}_{d_0}}$ is the unique zero of the derivative $\frac{d}{da_t^2} d_F^2(P_{d_0}^\top P_{d_0} \overleftarrow{\hat{X}}_t, \overrightarrow{X_0})$. This confirms that $T - \hat{\delta}_{d_0}$ is the unique minimizer of the distance. Conversely, if condition (12) is not satisfied, then $\hat{\delta}_{d_0} = 0$. In this scenario, the squared distance $d_F^2(P_{d_0}^\top P_{d_0} \overleftarrow{\hat{X}}_t, \overrightarrow{X_0})$ is a non-increasing function of $t$ and thus attains its minimum at the endpoint $t = T$. This result is consistent, as $t = T = T - \hat{\delta}_{d_0}$.

We remark by Proposition 2 that, since $\hat{t}_d = T$ for every $d \in \{d_0+1, \ldots, D\}$, for every $t \in [0, T]$,

$$d_F(P_{d_0}^\top P_{d_0} \overleftarrow{\hat{X}}_{T-\hat{\delta}_{d_0}}, \overrightarrow{X_0}) \leq d_F(P_{d_0}^\top P_{d_0} \overleftarrow{\hat{X}}_t, \overrightarrow{X_0}) \leq d_F(P_d^\top P_d \overleftarrow{\hat{X}}_t, \overrightarrow{X_0}).$$

Observe that $\hat{t}_1 = \max_{d\in\{1,\ldots,d_0\}} \hat{t}_d$, which is in the same order of $\hat{\sigma}_d$. This is due to the fact that $x \mapsto \frac{a-x}{1-x}$ is non-increasing if $a < 1$. Then from the proof of Proposition 2 we deduce that, for $t \geq \hat{t}_1$ and $d \in \{1, \ldots, d_0\}$, that

$$d_F(P_{d_0}^\top P_{d_0} \overleftarrow{\hat{X}}_{T-\hat{\delta}_{d_0}}, \overrightarrow{X_0}) \leq d_F(P_{d_0}^\top P_{d_0} \overleftarrow{\hat{X}}_t, \overrightarrow{X_0}) < d_F(P_d^\top P_d \overleftarrow{\hat{X}}_t, \overrightarrow{X_0}).$$

If $\hat{t}_1 = 0$, the proof is finished. Note that this is the case if $\sigma^2 \geq 1/4$, since $\frac{4\sigma^2 - \hat{\sigma}_1^2}{(1-\hat{\sigma}_1^2)_+} \geq 1$. We study from now the case where $\hat{t}_1 > 0$ and $\sigma^2 \leq 1/4$, with $t \leq \hat{t}_1$ and $d \in \{1, \dots, d_0\}$. We do this by showing for every dimension $d \in \{1, \dots, d_0\}$, $d_F(P_d^\top P_d \overleftarrow{\hat{X}}_t, \overrightarrow{X_0})$ is non-increasing on $[0, \hat{t}_1]$. This implies

$$d_F(P_{d_0}^\top P_{d_0} \overleftarrow{\hat{X}}_{T-\hat{\delta}_{d_0}}, \overrightarrow{X_0}) \leq d_F(P_{d_0}^\top P_{d_0} \overleftarrow{\hat{X}}_{\hat{t}_{d_0}}, \overrightarrow{X_0}) \leq d_F(P_d^\top P_d \overleftarrow{\hat{X}}_{\hat{t}_{d_0}}, \overrightarrow{X_0}) \leq d_F(P_d^\top P_d \overleftarrow{\hat{X}}_t, \overrightarrow{X_0}).$$

In the remainder of the proof, we show, for $d \in \{1, \dots, d_0\}$, that $d_F(P_d^\top P_d \overleftarrow{\hat{X}}_t, \overrightarrow{X_0})$ is non-increasing on $[0, \hat{t}_1]$. This is equivalent to proving that $d_F(P_d^\top P_d \overleftarrow{\hat{X}}_{T-t}, \overrightarrow{X_0})$ is non-decreasing on $[T - \hat{t}_1, T]$. Recall that, as in the proof as Proposition 1,

$$d_F^2(P_d^\top P_d \overleftarrow{\hat{X}}_{T-t}, \overrightarrow{X_0}) = \sum_{d'=d+1}^{d_0} \sigma^2 + \sum_{d'=1}^{d} \left(\sqrt{a_t^2 + (1-a_t^2)\hat{\sigma}_{d'}^2} - \sigma\right)^2.$$

Consider $f_d$ given by

$$f_d(x) = \sum_{d'=1}^{d} \left(\sqrt{x + (1-x)\hat{\sigma}_{d'}^2} - \sigma\right)^2.$$

What we want to show is equivalent to $f$ being non-decreasing on $[a_{T-\hat{t}_1}^2, a_T^2]$. Since $f_d$ is convex as proven in Proposition 1, it is sufficient to show that $f'$ is positive at $a_{T-\hat{t}_1}^2$. All in all, since the derivative of $f_d$ is

$$f_d'(x) = \sum_{d'=1}^{d} \Big(1 - \frac{\sigma}{\sqrt{\hat{\sigma}_{d'}^2 + (1-\hat{\sigma}_{d'}^2)x}}\Big)(1 - \hat{\sigma}_{d'}^2),$$

if we are able to show that for any $d' \leq d_0$,

$$(1 - \frac{\sigma}{\sqrt{\hat{\sigma}_{d'}^2 + (1-\hat{\sigma}_{d'}^2)a_{T-\hat{t}_1}^2}})(1 - \hat{\sigma}_{d'}^2) \geq 0, \tag{13}$$

then

$$f_d'(a_{T-\hat{t}}^2) = \sum_{d'=1}^{d} \Big(1 - \frac{\sigma}{\sqrt{\hat{\sigma}_{d'}^2 + (1-\hat{\sigma}_{d'}^2)a_{T-\hat{t}_1}^2}}\Big)(1 - \hat{\sigma}_{d'}^2) \geq 0.$$

The result above is twofold. First, we get that $d_F(P_d^\top P_d \overleftarrow{\hat{X}}_{T-t}, \overrightarrow{X_0})$ is increasing on the interval of interest. This also interestingly shows that the minimum of the Frobenius distance $t \mapsto d_F(P_{d_0}^\top P_{d_0} \overleftarrow{\hat{X}}_t, \overrightarrow{X_0})$ is reached after $\hat{t}_1$. Since by definition the minium is reached at $T - \hat{\delta}_{d_0}$, we get that $T - \hat{\delta}_{d_0} \geq \hat{t}$.

The only thing remaining is to show (13). Recall that $\hat{t}_1 = T - \bar{a}^{-2}\Big(\frac{4\sigma^2 - \hat{\sigma}_1^2}{1-\hat{\sigma}_1^2}\Big)$, and that we assumed $\hat{t}_1 > 0$, which implies $\frac{4\sigma^2 - \hat{\sigma}_1^2}{1-\hat{\sigma}_1^2} < a_T^2$. On the other hand, recall that we work under the event that $|\sigma^2 - \hat{\sigma}_1^2| \leq \sigma^2$. Hence, $\hat{\sigma}_1^2 \leq 2\sigma^2 < 1$ and $\frac{4\sigma^2 - \hat{\sigma}_1^2}{1-\hat{\sigma}_1^2} > 0$. Therefore, by definition of $\hat{t}_1$, we have

$$a_{T-\hat{t}_1}^2 = \frac{4\sigma^2 - \hat{\sigma}_1^2}{1 - \hat{\sigma}_1^2}.$$

From here, we prove (13). We rewrite (13) as

$$1 - \frac{\sigma}{\sqrt{\hat{\sigma}_{d'}^2 + (1-\hat{\sigma}_{d'}^2)a_{T-\hat{t}}^2}} \geq 0.$$

$$\Leftrightarrow \quad \sigma^2 \leq \hat{\sigma}_{d'}^2 + (1 - \hat{\sigma}_{d'}^2)\frac{4\sigma^2 - \hat{\sigma}_1^2}{1 - \hat{\sigma}_1^2}$$

$$\begin{aligned}
&\Leftrightarrow \quad \sigma^2 \le \hat{\sigma}_{d'}^2 + (1-\hat{\sigma}_{d'}^2)\Big(1 - \frac{1-4\sigma^2}{1-\hat{\sigma}_1^2}\Big) \\
&\Leftrightarrow \quad \sigma^2 \le 1 - (1-\hat{\sigma}_{d'}^2)\frac{1-4\sigma^2}{1-\hat{\sigma}_1^2} \\
&\Leftrightarrow \quad (1-\hat{\sigma}_{d'}^2)\frac{1-4\sigma^2}{1-\hat{\sigma}_1^2} \le 1-\sigma^2.
\end{aligned}$$

Therefore, since $\hat{\sigma}_{d'} < 1$,we deduce that (13) is equivalent to showing:

$$\frac{1-4\sigma^2}{1-\hat{\sigma}_1^2} \le \frac{1-\sigma^2}{1-\hat{\sigma}_{d'}^2}.$$

To show this, recall the bound $\hat{\sigma}_1^2 \le 2\sigma^2 < 1$. Thus,

$$\frac{1-4\sigma^2}{1-\hat{\sigma}_1^2} \le \frac{1-4\sigma^2}{1-2\sigma^2} = 1 - \sigma^2\frac{2}{1-2\sigma^2} \le 1-\sigma^2 \le \frac{1-\sigma^2}{1-\hat{\sigma}_{d'}^2},$$

which derives the desired inequality and we conclude the proof.

### A.4 Proof of Proposition 4

We begin by rewriting the expression of the score matching objective in the following form:

$$\begin{aligned}
\mathcal{R}(s_M) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{t\sim\mathcal{T},\varepsilon\sim\mathcal{N}(0,I_D)} \left\| s_M(b_t X_i + a_t\varepsilon, t) + \frac{\varepsilon}{a_t}\right\|^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{t\sim\mathcal{T},\varepsilon} \left\| -M(t)(b_t X_i + a_t\varepsilon) + \frac{\varepsilon}{a_t}\right\|^2 \quad \text{(since } s_M(x,t) = -M(t)x\text{)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{d=1}^{D} \mathbb{E}_{t\sim\mathcal{T},\varepsilon_d}\left[\Big(m_d(t)(b_t X_{ik} + a_t\varepsilon_d) - \frac{\varepsilon_d}{a_t}\Big)^2\right].
\end{aligned}$$

To find the optimal $M(t)$, we note that the objective and the constraint are separable across the time interval $[0,T]$. The objective is also separable across the dimensions $d \in \{1,\ldots,D\}$. Hence it suffices to minimize the quantity

$$r(m_d(t)) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\varepsilon_d}\left[\Big(m_d(t)(b_t X_{ik} + a_t\varepsilon_d) - \frac{\varepsilon_d}{a_t}\Big)^2\right]$$

separately over $m_d(t) \in [-C, C]$ for each $t \in [0,T]$ and $d \in \{1,\ldots,D\}$. Observe that the function $r : [-C,C] \to \mathbb{R}$ is a quadratic function. Its derivative is

$$\begin{aligned}
r'(m) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\varepsilon_d}\left[2\left(m(b_t X_{id} + a_t\varepsilon_d) - \frac{\varepsilon_d}{a_t}\right)(b_t X_{id} + a_t\varepsilon_d)\right], \\
&= \frac{2}{n}\sum_{i=1}^{n}\Big(m(b_t^2 X_{id}^2 + a_t^2) - 1\Big) \quad \text{(since } \mathbb{E}[\varepsilon_d] = 0, \mathbb{E}[\varepsilon_d^2] = 1\text{)}, \\
&= m\left(a_t^2 + b_t^2\frac{1}{n}\sum_{i=1}^{n} X_{id}^2\right) - 1.
\end{aligned}$$

Therefore, the minimum of $r$ over $[-C,C]$ is attained at

$$\hat{m}_d(t) = \min\left(C, \frac{1}{a_t^2 + b_t^2\hat{\sigma}_d^2}\right),$$

which concludes the proof.

### A.5 Proof of Proposition 5

Since the Fréchet distance is determined by the variance for centered random variables, the first step of the proof is to deduce the variance of $P_d\overleftarrow{X}_0$ for all $d$. Denote for $1 \leq d \leq D$, the variance of $\overleftarrow{X}_{t,d}$ by $V_{t,dd}$. It is known (see, for instance, Särkkä & Solin, 2019, Section 5.5) that $V_{t,dd}$ follows the following ODE:

$$\frac{dV_{t,dd}}{dt} = 2(1 - 2\hat{m}_d(T-t))V_{t,dd} + 2, \quad V_{0,dd} = 1. \tag{14}$$

An important intermediate step in this proof is to show the following:

$$(\sqrt{V_{t,dd}} - \sigma_d)^2 \leq \sigma_d^2 \quad \text{, if} \quad d \leq d_1, \tag{15}$$

$$(\sqrt{V_{t,dd}} - \sigma_d)^2 \geq \sigma_d^2 \quad \text{, if} \quad d \geq d_2. \tag{16}$$

To do so, we first develop an explicit expression for $V_{t,dd}$:

$$V_{t,dd} = \exp\left(\int_0^t 2(1 - 2\hat{m}_d(T-\tau))d\tau\right) + 2\int_0^t \exp\left(\int_s^t 2(1 - 2\hat{m}_d(T-\tau))d\tau\right)ds. \tag{17}$$

If $C < \frac{1}{a_0^2+b_0^2\hat{\sigma}_d^2} = \frac{1}{\hat{\sigma}_d^2}$, let $t'_d$ be the unique solution in $[0,T]$ of the equation $C = \hat{m}_d(T-t'_d) = \frac{1}{a^2_{T-t'_d}+b^2_{T-t'_d}\hat{\sigma}_d^2}$. Otherwise, we set $t'_d = T$, which is always the case for $d \leq d_1$. Remark that if $\hat{\sigma}_d \geq 1$, then $\frac{1}{a_t^2+b_t^2\hat{\sigma}_d^2} \leq 1 \leq C$. Thus, for such dimension $d$, we always have $t'_d = T$ and $d \leq d_1$.

We derive an explicit expression for the term $V_{T,dd}$. We first calculate the first part by plugging in the exact form of $\hat{m}_d$. To do so, we recall that $a_t = \sqrt{1-e^{-2t}}$ and $b_t = e^{-t}$. Also note that $\hat{m}_d$ is decreasing on $[0,T]$, more precisely it is equal to $C$ on $[0, T-t'_d]$ and equal to $1/(a_t^2 + b_t^2\hat{\sigma}_d^2)$ for $t \in [T-t'_d, T]$. With these keys facts in mind, we begin by calculating the following integrand, which for $s = 0$ gives the first term in (17) and is the integrand of the second term.

$$\begin{aligned}
&\exp\left(\int_s^T 2(1-2\hat{m}_d(T-\tau))d\tau\right) \\
&\quad = \exp\left(\int_0^{T-s} 2(1-2\hat{m}_d(\tau))d\tau\right) \\
&\quad = e^{2(T-s)}\exp\left(-4\int_0^{T-s\vee t'_d} \hat{m}_d(\tau)d\tau\right)\exp\left(-4\int_{T-s\vee t'_d}^{T-s} \hat{m}_d(\tau)d\tau\right) \\
&\quad = e^{2(T-s)}e^{-4C(T-s\vee t'_d)}e^{-4(s\vee t'_d-s)}\left(\frac{1-(1-\hat{\sigma}_d^2)e^{-2(T-s\vee t'_d)}}{1-(1-\hat{\sigma}_d^2)e^{-2(T-s)}}\right)^2,
\end{aligned} \tag{18}$$

where, in the last line, we use the following:

$$\int \hat{m}_d(\tau)d\tau = \int\left(1 + \frac{e^{-2\tau}(1-\hat{\sigma}_d^2)}{1-(1-\hat{\sigma}_d^2)e^{-2\tau}}\right)d\tau = \tau + \frac{1}{2}\log\left(1-(1-\hat{\sigma}_d^2)e^{-2\tau}\right).$$

By substituting $s = 0$, we see that the first term in (17) is equal to

$$\exp\left(\int_0^T 2(1-2\hat{m}_d(T-\tau))d\tau\right) = e^{-2T}e^{-4(C-1)(T-t'_d)}\left(\frac{1-(1-\hat{\sigma}_d^2)e^{-2(T-t'_d)}}{1-(1-\hat{\sigma}_d^2)e^{-2T}}\right)^2. \tag{19}$$

Next, we focus on deriving an explicit expression of the second term in (17). We plug in the term (18) and deduce that

$$2\int_0^T \exp\left(\int_s^T 2(1-2\hat{m}_d(T-\tau))d\tau\right)ds$$

$$
\begin{aligned}
&= 2(\int_{t'_d}^{T} + \int_0^{t'_d}) e^{2(T-s)} e^{-4C(T-s\vee t'_d)} e^{-4(s\vee t'_d - s)} \left( \frac{1-(1-\hat{\sigma}_d^2)e^{-2(T-s\vee t'_d)}}{1-(1-\hat{\sigma}_d^2)e^{-2(T-s)}} \right)^2 ds \\
&= 2\int_{t'_d}^{T} e^{2(T-s)} e^{-4C(T-s)} ds \\
&\quad + 2\int_0^{t'_d} e^{2(T-s)} e^{-4C(T-t'_d)} e^{-4(t'_d-s)} \left( \frac{1-(1-\hat{\sigma}_d^2)e^{-2(T-t'_d)}}{1-(1-\hat{\sigma}_d^2)e^{-2(T-s)}} \right)^2 ds \\
&= \frac{1}{2C-1}(1-e^{(2-4C)(T-t'_d)}) \\
&\quad + (1-(1-\hat{\sigma}_d^2)e^{-2(T-t'_d)})^2 e^{-4(C-1)(T-t'_d)} \int_0^{t'_d} \frac{2e^{-2(T-s)}}{(1-(1-\hat{\sigma}_d^2)e^{-2(T-s)})^2} ds \\
&= \frac{1}{2C-1}(1-e^{(2-4C)(T-t'_d)}) \\
&\quad + \frac{(1-(1-\hat{\sigma}_d^2)e^{-2(T-t'_d)})^2 e^{-4(C-1)(T-t'_d)}}{1-\hat{\sigma}_d^2} \left[ \frac{1}{1-(1-\hat{\sigma}_d^2)e^{-2(T-s)}} \right]_0^{t'_d}.
\end{aligned}
$$

We see that the last term can be rewritten in the following form

$$
\begin{aligned}
\left[ \frac{1}{1-(1-\hat{\sigma}_d^2)e^{-2(T-s)}} \right]_0^{t'_d} &= \frac{1}{1-(1-\hat{\sigma}_d^2)e^{-2(T-t'_d)}} - \frac{1}{1-(1-\hat{\sigma}_d^2)e^{-2T}} \\
&= \frac{(1-\hat{\sigma}_d^2)(e^{-2(T-t'_d)} - e^{-2T})}{(1-(1-\hat{\sigma}_d^2)e^{-2T})(1-(1-\hat{\sigma}_d^2)e^{-2(T-t'_d)})}.
\end{aligned}
$$

Thus, we derive that

$$
\begin{aligned}
&2\int_0^T \exp\left( \int_s^T 2(1-2\hat{m}_d(T-\tau))d\tau \right) ds \\
&\qquad = \frac{1}{2C-1}(1-e^{(2-4C)(T-t'_d)}) \\
&\qquad\quad + \frac{(1-(1-\hat{\sigma}_d^2)e^{-2(T-t'_d)})e^{-4(C-1)(T-t'_d)}}{1-(1-\hat{\sigma}_d^2)e^{-2T}} (e^{-2(T-t'_d)} - e^{-2T}). \qquad (20)
\end{aligned}
$$

Therefore, by summing up the two terms (19) and (20), we deduce that

$$
\begin{aligned}
V_{T,dd} &= \frac{1}{2C-1}(1-e^{(2-4C)(T-t'_d)}) \\
&\quad + \frac{(1-(1-\hat{\sigma}_d^2)e^{-2(T-t'_d)})e^{-4(C-1)(T-t'_d)}}{1-(1-\hat{\sigma}_d^2)e^{-2T}} (e^{-2(T-t'_d)} - e^{-2T}) \\
&\quad + e^{-2T} e^{-4(C-1)(T-t'_d)} \left( \frac{1-(1-\hat{\sigma}_d^2)e^{-2(T-t'_d)}}{1-(1-\hat{\sigma}_d^2)e^{-2T}} \right)^2 \\
&= \frac{1}{2C-1}(1-e^{(2-4C)(T-t'_d)}) \\
&\quad + e^{-4(C-1)(T-t'_d)} \frac{1-(1-\hat{\sigma}_d^2)e^{-2(T-t'_d)}}{1-(1-\hat{\sigma}_d^2)e^{-2T}} \\
&\quad \times \left( e^{-2(T-t'_d)} - e^{-2T} + e^{-2T} \frac{1-(1-\hat{\sigma}_d^2)e^{-2(T-t'_d)}}{1-(1-\hat{\sigma}_d^2)e^{-2T}} \right) \\
&= \frac{1}{2C-1}(1-e^{(2-4C)(T-t'_d)}) \\
&\quad + e^{-4(C-1)(T-t'_d)} \frac{1-(1-\hat{\sigma}_d^2)e^{-2(T-t'_d)}}{1-(1-\hat{\sigma}_d^2)e^{-2T}}
\end{aligned}
$$

$$\begin{aligned}
&\times \frac{e^{-2(T-t'_d)} - 2(1-\hat{\sigma}_d^2)e^{-2(2T-t'_d)} + (1-\hat{\sigma}_d^2)e^{-4T}}{1-(1-\hat{\sigma}_d^2)e^{-2T}} \\
&= \frac{1}{2C-1}(1-e^{(2-4C)(T-t'_d)}) \\
&\quad + e^{(2-4C)(T-t'_d)} \frac{1-(1-\hat{\sigma}_d^2)e^{-2(T-t'_d)}}{1-(1-\hat{\sigma}_d^2)e^{-2T}} \frac{1-2(1-\hat{\sigma}_d^2)e^{-2T}+(1-\hat{\sigma}_d^2)e^{-2(T+t'_d)}}{1-(1-\hat{\sigma}_d^2)e^{-2T}}.
\end{aligned}$$

We remark that, for $d \le d_1$ we have $t'_d = T$ and we may simplify the expression of $V_{T,dd}$ to

$$V_{T,dd} = \hat{\sigma}_d^2 \frac{1-2(1-\hat{\sigma}_d^2)e^{-2T}+(1-\hat{\sigma}_d^2)e^{-4T}}{1-2(1-\hat{\sigma}_d^2)e^{-2T}+(1-\hat{\sigma}_d^2)^2e^{-4T}}. \tag{21}$$

Before we prove (15) and (16), we categorize the behavior of $V_{t,dd}$ according to the value of $\hat{\sigma}_d$ and we summarize the result in the following lemma, the proof of which we delay to the end of this proof.

**Lemma 1.** *For $d \in \{1,\dots,D\}$. If $\hat{\sigma}_d \ge 1$, then $V_{t,dd} \ge 1$ for every $t \in [0,T]$. If $\hat{\sigma}_d \le 1$, then $V_{t,dd} \le 1$ for every $t \in [0,T]$.*

Let us deduce from (21), for $d \le d_1$, $(\sqrt{V_{T,dd}} - \sigma_d)^2 \le \sigma_d^2$. We work under the high probability event that $|\sigma_d^2 - \hat{\sigma}_d^2| \le \sigma_d^2$ for every $d \in \{1,\dots,D\}$, we split the proof into three cases:

- If $\sigma_d > 1$ then $\hat{\sigma}_d \ge 1$, from (21), we see that $V_{T,dd} < \hat{\sigma}_d^2 \le 4\sigma_d^2$, with high probability. Thus, $(\sqrt{V_{T,dd}} - \sigma_d)^2 \le \max((0-\sigma_d)^2, (2\sigma_d - \sigma_d)^2) \le \sigma_d^2$.
- If $\sigma_d \in [\frac{1}{2}, 1)$ which implies that $\hat{\sigma}_d < 1$, then $V_{T,dd} \le 1$, we then have $(\sqrt{V_{T,dd}} - \sigma_d)^2 \le \max((0-\sigma_d)^2, (1-\sigma_d)^2) \le \sigma_d^2$.
- If $\sigma_d = 1$, we again split cases depending on whether $\hat{\sigma}_d \ge 1$. We get the same bounds as in the two previous cases.
- Finally, if $\sigma_d \le \frac{1}{2}$, with high probability, $|\sigma_d^2 - \hat{\sigma}_d^2| \le \sigma_d^2$. Hence, $\hat{\sigma}_d^2 \le 2\sigma_d^2 \le \frac{1}{2}$. Observing that the fraction in (21) is bounded by $1/(1-\hat{\sigma}_d^2)$, we deduce that

  $$V_{T,dd} \le \frac{\hat{\sigma}_d^2}{1-\hat{\sigma}_d^2} \le 2\hat{\sigma}_d^2 \le 4\sigma_d^2, \quad \forall d \le d_1,$$

  which gives the desired bound.

Next, for $d \ge d_2$, remark by definition of $t'_d$ that $\frac{1}{1-(1-\hat{\sigma}_d^2)e^{-2(T-t'_d)}} = C$. Also note that the definition of $d_2$ and the fact that $C > 1$ implies that $\hat{\sigma}_d^2 < 1$. Hence,

$$\begin{aligned}
V_{T,dd} &= \frac{1}{2C-1} + e^{(2-4C)(T-t'_d)} \left( \frac{(1-2(1-\hat{\sigma}_d^2)e^{-2T}+(1-\hat{\sigma}_d^2)e^{-2(T+t'_d)})}{C(1-(1-\hat{\sigma}_d^2)e^{-2T})^2} - \frac{1}{2C-1} \right) \\
&\ge \frac{1}{2C-1} + e^{(2-4C)(T-t'_d)} \left( \frac{1}{C} - \frac{1}{2C-1} \right) \\
&\ge \frac{1}{2C-1}.
\end{aligned}$$

Therefore, for $d \ge d_2$ we deduce that

$$V_{T,dd} \ge \frac{1}{2C-1} \ge 4\sigma_d^2.$$

To summarize, we derived the following bounds

$$(\sqrt{V_{T,dd}} - \sigma_d)^2 \le \sigma_d^2, \quad \forall d \le d_1,$$

and

$$(\sqrt{V_{T,dd}} - \sigma_d)^2 \ge \sigma_d^2, \quad \forall d > d_2.$$

By definition of the Fréchet distance, we have

$$d_F^2(P_d^\top P_d \overleftarrow{\tilde{X}}_T, \overrightarrow{X_0}) = \sum_{j=1}^{d} (\sqrt{V_{T,jj}} - \sigma_j)^2 + \sum_{j=d+1}^{D} \sigma_j^2,$$

we deduce that, for any $d < d_1 \leq d_2 < d'$,

$$
\begin{aligned}
d_F(P_{d_1}^\top P_{d_1} \overleftarrow{\tilde{X}}_T, \overrightarrow{X_0}) &= \sum_{j=1}^{d_1} (\sqrt{V_{T,jj}} - \sigma_j)^2 + \sum_{j=d_1+1}^{D} \sigma_j^2 \\
&= \sum_{j=1}^{d} (\sqrt{V_{T,jj}} - \sigma_j)^2 + \sum_{j=d+1}^{d_1} (\sqrt{V_{T,jj}} - \sigma_j)^2 + \sum_{j=d_1+1}^{D} \sigma_j^2 \\
&\leq \sum_{j=1}^{d} (\sqrt{V_{T,jj}} - \sigma_j)^2 + \sum_{j=d+1}^{D} \sigma_j^2 \\
&= d_F(P_d^\top P_d \overleftarrow{\tilde{X}}_T, \overrightarrow{X_0}),
\end{aligned}
$$

and

$$
\begin{aligned}
d_F(P_{d_2}^\top P_{d_2} \overleftarrow{\tilde{X}}_T, \overrightarrow{X_0}) &= \sum_{j=1}^{d_2} (\sqrt{V_{T,jj}} - \sigma_j)^2 + \sum_{j=d_2+1}^{D} \sigma_j^2 \\
&= \sum_{j=1}^{d_2} (\sqrt{V_{T,jj}} - \sigma_j)^2 + \sum_{j=d_2+1}^{d'} \sigma_j^2 + \sum_{j=d'+1}^{D} \sigma_j^2 \\
&\leq \sum_{j=1}^{d'} (\sqrt{V_{T,jj}} - \sigma_j)^2 + \sum_{j=d'+1}^{D} \sigma_j^2 \\
&= d_F(P_{d'}^\top P_{d'} \overleftarrow{\tilde{X}}_T, \overrightarrow{X_0}).
\end{aligned}
$$

Therefore, the minimum of $d_F(P_d^\top P_d \overleftarrow{\tilde{X}}_T, \overrightarrow{X_0})$ must occur between $d_1$ and $d_2$.

**Proof of Lemma 1.** Recall that $V_{t,dd}$ satisfies the ODE (14)

$$
\frac{dV_{t,dd}}{dt} = 2(1 - 2\hat{m}_d(T-t))V_{t,dd} + 2, \quad V_{0,dd} = 1.
$$

Assume that $\hat{\sigma}_d > 1$ and by contradiction that $V_{t,dd} < 1$ for some $t \in [0,T]$. Let $t_0 = \inf\{t : V_{t,dd} < 1\}$, by continuity, we have $V_{t_0,dd} = 1$. Then we have

$$
\Big[\frac{dV_{t,dd}}{dt}\Big]_{t=t_0} = 2(1 - 2\hat{m}_d(T-t_0))V_{t_0,dd} + 2 = 2\Big(1 - \frac{2}{1 - e^{-2t} + e^{-2t}\hat{\sigma}_d^2}\Big) + 2,
$$

where we use the fact that $V_{t_0,dd} = 1$. The last term can be rewritten as

$$
\frac{4(\hat{\sigma}_d^2 - 1)e^{-2t}}{1 - e^{-2t} + e^{-2t}\hat{\sigma}_d^2}.
$$

Hence we have $[\frac{dV_{t,dd}}{dt}]_{t=t_0} > 0$ which contradicts the definition of $t_0$. Hence $V_{t,dd} \geq 1$ for all $t \in [0,T]$. The case for $\hat{\sigma}_d < 1$ can be derived similarly.

### A.6 Derivation of special cases of Proposition 5

First, consider the scenario where the learning capacity is unconstrained, effectively setting $C = \infty$, while the data covariance matrix is nonsingular. In this case, the condition on $d_1$ becomes $0 \leq \hat{\sigma}_d^2$, which is trivially satisfied for all $d \in \{1, \ldots, D\}$, implying $d_1 = D$. The condition for $d_2$ becomes $0 > 4\sigma_d^2$, which holds for none of $d$, thus implying $d_2 = D$. Therefore, when $C = \infty$, Proposition 5 entails that $d_{\min} = D$. This result is somewhat expected: if the score function is learned perfectly, the diffusion process can be reversed in the full ambient space, enabling sampling from the target distribution without any need for dimensionality reduction.

Second, consider the scenario addressed in Proposition 3 where the true data distribution lies within a $d_0$-dimensional linear subspace, i.e., $\sigma_{d_0+1} = \cdots = \sigma_D = 0$ and $\sigma_1 = \cdots = \sigma_{d_0} = \sigma$. Assume that

$C$ is sufficiently large to ensure that $1/C \leq \min(\sigma^2, \min_{d' \in \{1,\ldots,d_0\}} \hat{\sigma}^2_{d'})$. Therefore, for $d \leq d_0$, one has $\frac{1}{C} \leq \hat{\sigma}_d^2$ (which is not satisfied anymore for $d$ beyond $d_0$), leading to $d_1 = d_0$. On the other hand, for $d > d_0$ we have $\frac{1}{2C-1} \geq 0 = 4\sigma_d$. Hence $d_0 = d_1 \leq d_2 \leq d_0$, which implies $d_2 = d_0$. Thus, Proposition 5 predicts $d_{\min} = d_0$. This suggests that the projection onto the subspace in which the data distribution lies is the optimal sampling strategy, which is in line with the recommendation of Proposition 3.

**Proof of Corollary 1.** By the definition of $d$, we have $d_1 = d$. It remains to prove that $d_2 \leq d + 1$. With $n$ large enough and high probability, we have $\hat{\sigma}_{d+1} \geq \sigma_{d+1}^2/2$. Therefore,

$$\frac{1}{4(2C-1)} \geq \frac{1}{8C} \geq \frac{\hat{\sigma}^2_{d+1}}{8} \geq \frac{\sigma^2_{d+1}}{16} = \frac{\lambda^{-(d+1)}}{16} \geq \lambda^{-(d+2)},$$

where we use the fact that $\lambda \geq 16$. This shows that $d_2 < d + 2$. Hence $d_2 \leq d + 1$.

### A.7 Proof of Proposition 6

The proof follows by observing that the covariance matrix of $OP_dO^\top \overleftarrow{X}_t$ is given by

$$\mathrm{cov}[OP_d^\top P_d O^\top \overleftarrow{X}_t] = O\mathrm{diag}(a_{T-t}^2 + b_{T-t}^2\sigma_1^2, \ldots, a_{T-t}^2 + b_{T-t}^2\sigma_d^2, 0, \ldots, 0)O^\top.$$

Therefore, we have the following explicit form of the Fréchet distance between $OP_d^\top P_d O^\top \overleftarrow{X}_t$ and $\overrightarrow{X_0}$:

$$d_F(OP_d^\top P_d O^\top \overleftarrow{X}_t, \overrightarrow{X_0}) = \sum_{j=1}^{D} \sigma_j^2 + \sum_{j=1}^{d}(a_{T-t}^2 + b_{T-t}^2\sigma_j^2) - 2\sum_{j=1}^{d}\sigma_j\sqrt{a_{T-t}^2 + b_{T-t}^2\sigma_j^2}.$$

The proof is concluded by using the same argument as in the proof of Proposition 2.

### A.8 Proof of Proposition 7

Recall that $\hat{\Lambda} = \mathrm{diag}(\hat{\sigma}_1^2, \ldots \hat{\sigma}_D^2)$ the matrix of eigenvalues of the estimated covariance matrix $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$. We first remark that

$$\begin{aligned}\mathrm{cov}[\hat{O}P_d^\top P_d \hat{O}^\top \overleftarrow{\hat{X}}_t] &= \hat{O}\mathrm{diag}(a_{T-t}^2 + b_{T-t}^2\hat{\sigma}_1^2, \ldots, a_{T-t}^2 + b_{T-t}^2\hat{\sigma}_d^2, 0, \ldots, 0)\hat{O}^\top \\ &= \hat{O}(a_{T-t}^2 P_d^\top P_d + b_{T-t}^2 P_d^\top P_d \hat{\Lambda})\hat{O}^\top.\end{aligned}$$

Denote the covariance matrix of $\hat{O}P_d^\top P_d \hat{O}^\top \overleftarrow{X}_t$ by $\hat{\Sigma}_d(t)$. Recall that the Fréchet distance between two centered Gaussian distributions is

$$d_F^2(\mathcal{N}(0, \Sigma_1), \mathcal{N}(0, \Sigma_2)) = \mathrm{tr}(\Sigma_1 + \Sigma_2 - 2(\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2})^{1/2}).$$

In the case of interest for us, we get

$$d_F^2(\hat{O}P_d^\top P_d \hat{O}^\top \overleftarrow{\hat{X}}_t, \overrightarrow{X_t}) = \sum_{d'=1}^{D} \sigma_{d'}^2 + \sum_{d'=1}^{d}(a_{T-t}^2 + b_{T-t}^2\hat{\sigma}_{d'}^2) - 2\mathrm{tr}((\hat{\Sigma}_d^{1/2}(t)\Sigma\hat{\Sigma}_d^{1/2}(t))^{1/2}).$$

We now argue that $\mathrm{tr}((\hat{\Sigma}_d^{1/2}(t)\Sigma\hat{\Sigma}_d^{1/2}(t))^{1/2})$ is approximately $\sum_{d'=1}^{d}\hat{\sigma}_{d'}\sqrt{a_{T-t}^2 + b_{T-t}^2\hat{\sigma}_{d'}^2}$. Observe that the two quantities are equal when $\Sigma$ and $\hat{\Sigma}$ commute, which was the case in the previous sections where we assumed that both matrices were diagonal. By Proposition 10, with probability $1 - 2e^{-u}$, we have $\Sigma \preceq \frac{1}{1-\varepsilon_u}\hat{\Sigma}$, where $\preceq$ denotes the Loewner order (see, for instance, Horn & Johnson, 2012, Definition 7.7.1). Hence,

$$\hat{\Sigma}_d^{1/2}(t)\Sigma\hat{\Sigma}_d^{1/2}(t) \preceq \frac{1}{1-\varepsilon_u}\hat{\Sigma}_d^{1/2}(t)\hat{\Sigma}\hat{\Sigma}_d^{1/2}(t),$$

by Lemma 2 (i). Since square root is a matrix monotonic function (see Lemma 2 (ii)), we derive that

$$\mathrm{tr}((\hat{\Sigma}_d^{1/2}(t)\Sigma\hat{\Sigma}_d^{1/2}(t))^{1/2}) \leq \sqrt{\frac{1}{1-\varepsilon_u}}\mathrm{tr}((\hat{\Sigma}_d^{1/2}(t)\hat{\Sigma}\hat{\Sigma}_d^{1/2}(t))^{1/2})$$

$$\leq (1+\varepsilon_u)\mathrm{tr}((\hat{\Sigma}_d^{1/2}(t)\hat{\Sigma}\hat{\Sigma}_d^{1/2}(t))^{1/2}),$$

where we use $\varepsilon_u \leq 1/2$ in the last inequality. Then, by the commutativity of $\hat{\Sigma}_d(t)$ and $\hat{\Sigma}$,

$$\begin{aligned}
&\mathrm{tr}((\hat{\Sigma}_d^{1/2}(t)\hat{\Sigma}\hat{\Sigma}_d^{1/2}(t))^{1/2}) \\
&\qquad = \mathrm{tr}(\hat{O}(a_{T-t}^2 P_d^\top P_d + b_{T-t}^2 P_d^\top P_d \hat{\Lambda})^{1/4}\hat{\Lambda}^{1/2}(a_{T-t}^2 P_d^\top P_d + b_{T-t}^2 P_d^\top P_d \hat{\Lambda})^{1/4}\hat{O}^\top) \\
&\qquad = \mathrm{tr}(\hat{O}\hat{\Lambda}^{1/2}(a_{T-t}^2 P_d^\top P_d + b_{T-t}^2 P_d^\top P_d \hat{\Lambda})^{1/2}\hat{O}^\top) \\
&\qquad = \sum_{d'=1}^{d} \hat{\sigma}_{d'}\sqrt{a_{T-t}^2 + b_{T-t}^2\hat{\sigma}_{d'}^2}.
\end{aligned}$$

By combining the results, we obtain

$$\mathrm{tr}((\hat{\Sigma}_d^{1/2}(t)\Sigma\hat{\Sigma}_d^{1/2}(t))^{1/2}) \leq (1+\varepsilon_u)\sum_{d'=1}^{d}\hat{\sigma}_{d'}\sqrt{a_{T-t}^2 + b_{T-t}^2\hat{\sigma}_{d'}^2}.$$

We may use the same argument to derive a similar lower bound, and thus deduce that

$$\left|\mathrm{tr}((\hat{\Sigma}_d^{1/2}(t)\Sigma\hat{\Sigma}_d^{1/2}(t))^{1/2})^{1/2}) - \sum_{d'=1}^{d}\hat{\sigma}_{d'}\sqrt{a_{T-t}^2 + b_{T-t}^2\hat{\sigma}_{d'}^2}\right| \leq \varepsilon_u \sum_{d'=1}^{d}\hat{\sigma}_{d'}\sqrt{a_{T-t}^2 + b_{T-t}^2\hat{\sigma}_{d'}^2}.$$

Note that if $\hat{\sigma}_{d'} \geq 1$, then $\sqrt{a_{T-t}^2 + b_{T-t}^2\hat{\sigma}_{d'}^2} \leq \hat{\sigma}_{d'}$. Hence $\hat{\sigma}_{d'}\sqrt{a_{T-t}^2 + b_{T-t}^2\hat{\sigma}_{d'}^2} \leq \hat{\sigma}_{d'}^2$. On the other hand, if $\hat{\sigma}_d < 1$, then $\sqrt{a_{T-t}^2 + b_{T-t}^2\hat{\sigma}_{d'}^2} \leq 1$ and $\hat{\sigma}_{d'}\sqrt{a_{T-t}^2 + b_{T-t}^2\hat{\sigma}_{d'}^2} \leq \hat{\sigma}_{d'}$. Therefore, by recalling that $S(\Sigma) = \sum_{d'=1}^{D}\max(\hat{\sigma}_{d'}, \hat{\sigma}_{d'}^2)$, we deduce that

$$\left|\mathrm{tr}((\hat{\Sigma}_d^{1/2}(t)\Sigma\hat{\Sigma}_d^{1/2}(t))^{1/2})^{1/2}) - \sum_{d'=1}^{d}\hat{\sigma}_{d'}\sqrt{a_{T-t}^2 + b_{T-t}^2\hat{\sigma}_{d'}^2}\right| \leq S(\Sigma)\varepsilon_u.$$

The Fréchet distance $d_F(\hat{O}P_d^\top P_d\hat{O}^\top\overleftarrow{\hat{X}}_t, \overrightarrow{X_t})$ may now be bounded by

$$\begin{aligned}
\left|d_F^2(\hat{O}P_d^\top P_d\hat{O}^\top\overleftarrow{\hat{X}}_t, \overrightarrow{X_t}) - \Big(\sum_{d'=1}^{D}\sigma_{d'}^2 + \sum_{d'=1}^{d}(a_{T-t}^2 + b_{T-t}^2\hat{\sigma}_{d'}^2) - 2\sum_{d'=1}^{d}\hat{\sigma}_{d'}\sqrt{a_{T-t}^2 + b_{T-t}^2\hat{\sigma}_{d'}^2}\Big)\right| \\
\leq 2S(\Sigma)\varepsilon_u.
\end{aligned}$$

Hence, for $d \in \{2, \ldots, D\}$,

$$\begin{aligned}
\Big|d_F^2(\hat{O}P_d^\top P_d\hat{O}^\top\overleftarrow{\hat{X}}_t, \overrightarrow{X_t}) - d_F^2(\hat{O}P_{d-1}^\top P_{d-1}\hat{O}^\top\overleftarrow{\hat{X}}_t, \overrightarrow{X_t}) \\
- \sqrt{a_{T-t}^2 + b_{T-t}^2\hat{\sigma}_d^2}(\sqrt{a_{T-t}^2 + b_{T-t}^2\hat{\sigma}_d^2} - 2\hat{\sigma}_d)\Big| \leq 4S(\Sigma)\varepsilon_u.
\end{aligned}$$

We show in the following that if $t \geq \hat{T}_d(u)$, then

$$d_F^2(\hat{O}P_d^\top P_d\hat{O}^\top\overleftarrow{\hat{X}}_t, \overrightarrow{X_t}) \leq d_F^2(\hat{O}P_{d-1}^\top P_{d-1}\hat{O}^\top\overleftarrow{\hat{X}}_t, \overrightarrow{X_t}).$$

Observe that,

$$\begin{aligned}
d_F^2(\hat{O}P_d^\top P_d\hat{O}^\top\overleftarrow{\hat{X}}_t, \overrightarrow{X_t}) &\leq d_F^2(\hat{O}P_{d-1}^\top P_{d-1}\hat{O}^\top\overleftarrow{\hat{X}}_t, \overrightarrow{X_t}) \\
&\qquad + b_{T-t}^2\hat{\sigma}_d^2 + a_{T-t}^2 - 2\hat{\sigma}_d\sqrt{b_{T-t}^2\hat{\sigma}_d^2 + a_{T-t}^2} + 4S(\Sigma)\varepsilon_u \\
&= d_F^2(\hat{O}P_{d-1}^\top P_{d-1}\hat{O}^\top\overleftarrow{\hat{X}}_t, \overrightarrow{X_t}) \\
&\qquad + (\sqrt{a_{T-t}^2(1-\hat{\sigma}_d^2) + \hat{\sigma}_d^2} - \hat{\sigma}_d)^2 - \hat{\sigma}_d^2 + 4S(\Sigma)\varepsilon_u. \qquad (22)
\end{aligned}$$

Hence, for $t$ such that the last term (22) is non-positive, we have $d_F^2(\hat{O}P_d^\top P_d\hat{O}^\top \overleftarrow{\hat{X}}_t, \overrightarrow{X_t}) \leq d_F^2(\hat{O}P_{d-1}^\top P_{d-1}\hat{O}^\top \overleftarrow{\hat{X}}_t, \overrightarrow{X_t})$. We now show that this is true when $t \geq \hat{T}_d(u)$. To do so, we split our argument in two cases. We first consider the scenario where $\hat{\sigma}_d \geq 1$. In this case, by definition, $\hat{T}_d(u) = 0$ and therefore we prove the result holds for all $t \in [0, T]$. Observe that $\sqrt{a_{T-t}^2(1-\hat{\sigma}_d^2) + \hat{\sigma}_d^2} \in [1, \hat{\sigma}_d]$, therefore

$$(22) \leq (1-\hat{\sigma}_d)^2 - \hat{\sigma}_d^2 + 4S(\Sigma)\varepsilon_u = 1 - 2\hat{\sigma}_d + 4S(\Sigma)\varepsilon_u \leq 1 - 2\hat{\sigma}_d + \hat{\sigma}_d \leq 0,$$

where the last inequality holds for sufficiently small $\varepsilon_u$.

Now we consider the case where $\hat{\sigma}_d < 1$, and hence $\sqrt{a_{T-t}^2(1-\hat{\sigma}_d^2) + \hat{\sigma}_d^2} \geq \hat{\sigma}_d$. Therefore,

$$(22) \leq 0 \Leftrightarrow \sqrt{a_{T-t}^2(1-\hat{\sigma}_d^2) + \hat{\sigma}_d^2} \leq \hat{\sigma}_d + \sqrt{\hat{\sigma}_d^2 - 4S(\Sigma)\varepsilon_u}.$$

By squaring both sides and rearranging the terms, we deduce that

$$(22) \leq 0 \Leftrightarrow a_{T-t}^2 \leq \frac{\hat{\sigma}_d^2 - 4S(\Sigma)\varepsilon_u + 2\hat{\sigma}_d\sqrt{\hat{\sigma}_d^2 - 4S(\Sigma)\varepsilon_u}}{1-\hat{\sigma}_d^2},$$

and we conclude by observing that the last inequality is equivalent to $t \geq \hat{T}_d(u)$. We derive with a similar argument that if $t \leq \hat{t}_d(u)$ then

$$d_F^2(\hat{O}P_d^\top P_d\hat{O}^\top \overleftarrow{\hat{X}}_t, \overrightarrow{X_t}) \geq d_F^2(\hat{O}P_{d-1}^\top P_{d-1}\hat{O}^\top \overleftarrow{\hat{X}}_t, \overrightarrow{X_t}),$$

and we conclude the proof.

**Remark.** We can again in this case consider the same scenario as in Proposition 3 where the eigenvalues of the covariance matrix are equal. This can be an interesting direction for future work, as to generalize the previous results to this more general setup.

## B Bounds on Gaussian estimation

In this section, we give some bounds for the estimation error for Gaussian distributions.

**Proposition 9.** *Let $(X_1, \ldots, X_n)$ be sample drawn independently from $\mathcal{N}(0, \sigma^2)$. Then, for $\varepsilon > 0$, we have*

$$\mathbb{P}\Big[\big|\frac{1}{n}\sum_{i=1}^n X_i^2 - \sigma_d^2\big| \leq \varepsilon\sigma_d^2\Big] \geq 1 - 2\exp\Big(-\frac{\varepsilon^2 n}{4(\varepsilon+1)}\Big).$$

*Proof.* By Ghosh (2021), if $Z \sim \chi^2(p)$ and $u > 0$,

$$\mathbb{P}[|Z - p| \geq u] \leq 2\exp\Big(-\frac{u^2}{4(p+u)}\Big).$$

The result then unfolds from standard manipulations after observing that $\frac{1}{\sigma^2}\sum_{i=1}^n X_i^2$ follows a $\chi^2(n)$. □

**Proposition 10.** *Let $\Sigma$ be a semi-definite positive $D \times D$ matrix, and assume the sample $(X_1, \ldots, X_n)$ is drawn independently from $\mathcal{N}(0, \Sigma)$. Then, there is a universal constant $C$ such that, with probability $1 - 2e^{-u}$, the empirical covariance matrix $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$ satisfies:*

$$-\frac{8C}{3}(\sqrt{\frac{D+u}{n}} + \frac{D+u}{n})\Sigma \preceq \hat{\Sigma} - \Sigma \preceq \frac{8C}{3}\Big(\sqrt{\frac{D+u}{n}} + \frac{D+u}{n}\Big)\Sigma,$$

*where $\preceq$ denotes the Loewner order.*

*Proof.* It is shown in Vershynin (2018, Theorem 4.6.1) that, with probability $1 - 2e^{-u}$,

$$\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I_D\|_{op} \le K^2 C(\sqrt{\frac{D+u}{n}} + \frac{D+u}{n}),$$

where $\|\cdot\|_{op}$ denotes the operator norm and $K$ is a constant satisfying

$$\|X^\top x\|_{\psi_2} \le K\|X^\top x\|_{L_2}, \forall x \in \mathbb{R}^D,$$

where $\|X\|_{\psi_2} = \inf\{K > 0 : \mathbb{E}[e^{X^2/K^2}] \le 2\}$. It is shown in Vershynin (2018, Section 2.6.1) that, if $X$ follows a centered Gaussian distribution with standard deviation $\sigma$, then $\|X\|_{\psi_2} = \sigma\sqrt{8/3}$ and $\|X\|_{L_2} = \sigma$. Hence, $K = \sqrt{8/3}$ in our case and we have

$$-\frac{8C}{3}(\sqrt{\frac{D+u}{n}} + \frac{D+u}{n})I_D \preceq \Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I_d \preceq \frac{8C}{3}(\sqrt{\frac{D+u}{n}} + \frac{D+u}{n})I_D.$$

By multiplying $\Sigma^{1/2}$ from left and right for both side, we derive that

$$-\frac{8C}{3}(\sqrt{\frac{D+u}{n}} + \frac{D+u}{n})\Sigma \preceq \hat{\Sigma} - \Sigma \preceq \frac{8C}{3}(\sqrt{\frac{D+u}{n}} + \frac{D+u}{n})\Sigma.$$

□

## C Useful Lemma

In this section we provide some lemma that will be useful throughout the whole paper (see also, Horn & Johnson, 2012, Section 7.7).

**Lemma 2.** *Let $A, B$ be two symmetric $D \times D$ real matrices, and $S$ be an arbitrary $D \times D$ real matrix. The following statements hold:*

*(i) If $A \preceq B$, then $S^\top AS \preceq S^\top BS$.*

*(ii) If $A^2 \preceq B^2$, then $A \preceq B$. In particular if $A$ and $B$ are semi-definite positive, then $A \preceq B \Rightarrow \sqrt{A} \preceq \sqrt{B}$.*

## D Experiment details

### D.1 Natural image experiment

**Common details.** We use the dataset CelebA (Liu et al., 2015) and ImageNet (Deng et al., 2009). We use a U-Net model (Ronneberger et al., 2015) and an Adam optimizer (Kingma, 2014). The diffusion model uses rectified flow noise schedule (Liu et al., 2022). The code was implemented in JAX (Bradbury et al., 2018).

**Training of AE on CelebA-HQ.** We train an VQ-VAE using the VQ-GAN loss (Esser et al., 2021) for 1.95 million step on 20 TPUv2. The VQ-VAE encodes the images to a latent space of shape $64 \times 64 \times 3$ and reaches an 2k-rFID score of 2.44. Other hyperparameters for training is summarized in Table 1.

**Training of AE on ImageNet-256.** We train an VQ-VAE using the VQ-GAN loss (Esser et al., 2021) for 715 thousand (resp. 1.81 million) steps on 20 TPUv2 for a latent shape of $32 \times 32 \times 4$ (resp. $64 \times 64 \times 3$). The VQ-VAE achieves an rFID score of 3.6 (resp. 4.2). Other hyperparameters for training is summarized in Table 2.

**Training of LDM on CelebA-HQ.** We train an LDM on the images encoded by the AE we described above. We train for 5.25 million steps on 8 TPUv6. We summarize the hyperparameters used in Table 3.

| Name | Value |
|---|---|
| Coefficient of the adversarial loss | 0.1 |
| Coefficient of the generator loss | 100 |
| Coefficient of the LPIPS loss | 1.0 |
| Coefficient of the discriminator loss | 0.01 |
| Number of embeddings of the vector quantizer | 8192 |
| Optimizer | Adam with standard hyperparameters |
| EMA decay | 0.9999 |
| Learning rate | $10^{-5}$ |
| Batch size | 16 |

Table 1: Hyperparameters for training VQ-VAE on CelebA-HQ.

| Name | Value |
|---|---|
| Coefficient of the adversarial loss | 0.1 |
| Coefficient of the generator loss | 100 |
| Coefficient of the LPIPS loss | 1.0 |
| Coefficient of the discriminator loss | 0.01 |
| Number of embeddings of the vector quantizer | 8192 |
| Optimizer | Adam with standard hyperparameters |
| EMA decay | 0.999 |
| Learning rate | $10^{-5}$ |
| Batch size | 16 |

Table 2: Hyperparameters for training VQ-VAE on ImageNet-256.

**Training of LDM on ImageNet-256.** We train two LDMs on the images encoded by the AEs we described above. We train for 1 million steps on 8 TPUv6 for both LDMs. We summarize the hyperparameters used in Table 5.

**Training pixel diffusion model on CelebA and ImageNet-64.** We train diffusion models directly on CelebA and ImageNet-64. We train both models for 1 million steps on 12 TPUv2. We summarize the hyperparameters in Table 4.

**Results.** We previously introduced some results in Section 1. Here, we present additional evidence regarding the quality of the generated images. We observe (Figure 11) that in the final few steps, the sample of LDM does not change visibly. On the contrary, the images generated in pixel space (Figure 12) are still denoised even in the last steps.

**Synthetic Gaussian data.** In the experiment of Figure 5, we generate data using Gaussian distribution with covariance matrices equal to $\mathrm{diag}(1, 0.6, 0.6^2, \dots, 0.6^6, 10^{-10}, 10^{-10})$ (left) and $\mathrm{diag}(10, 0.2, 0.2, 0.2, 0, 0)$ (right). We then generate sample by first estimating the variances with the data with 1k sample, then solving the SDE (5) separately for each projection. We generate new sample using the Ornstein-Uhlenbeck process with $T = 2$ and 1000 discretization steps.

**Synthetic Gaussian data with exponentially decaying eigenvalues.** In the experiment of Figure 6, we generate data using Gaussian distribution with covariance matrices equal to $\mathrm{diag}(1, 1/4, 1/4^2, \dots, 1/4^9)$. We then train a linear model with score matching using 10 thousand samples, and we clip the parameters of the model. We generate new sample using the Ornstein-Uhlenbeck process with $T = 2$ and 1000 discretization steps.

**Linear AE and U-Net LDM on MNIST (Deng, 2012).** We train linaer AE on MNIST with dimension 64, 256 and 400 for 5k steps each using 4 TPUv2. Then we train U-Net diffusion models paired with each AE for 10k steps each using 4 TPUv2. We summarize other hyperparameters in Table 6 and 7.

| Name | Value |
|---|---|
| Noise schedule | Rectified Flow |
| Number of sampling steps | 250 |
| Optimizer | Adam with standard hyperparameters |
| EMA decay | 0.9999 |
| Learning rate | $10^{-4}$ |
| Batch size | 16 |

Table 3: Hyperparameters for training LDM on encoded images of CelebA-HQ.

| Name | Value |
|---|---|
| Noise schedule | Rectified Flow |
| Number of sampling steps | 250 |
| Optimizer | Adam with standard hyperparameters |
| EMA decay | 0.9999 |
| Learning rate | $10^{-4}$ |
| Batch size | 128 |

Table 4: Hyperparameters for training diffusion model on CelebA and ImageNet-64.

| Name | Value |
|---|---|
| Noise schedule | Rectified Flow |
| Number of sampling steps | 250 |
| Optimizer | Adam with standard hyperparameters |
| EMA decay | 0.999 |
| Learning rate | $10^{-4}$ |
| Batch size | 1024 |

Table 5: Hyperparameters for training LDM on encoded images of ImageNet-256.

| Name | Value |
|---|---|
| Optimizer | Adam with standard hyperparameters |
| Learning rate | 0.003 |
| Batch size | 256 |

Table 6: Hyperparameters for training Linear AE on MNIST.

| Name | Value |
|---|---|
| Noise schedule | Rectified Flow |
| Number of sampling steps | 250 |
| Optimizer | Adam with standard hyperparameters |
| EMA decay | 0.999 |
| Learning rate | $10^{-4}$ |
| Batch size | 256 |

Table 7: Hyperparameters for training LDM on encoded images of MNIST.

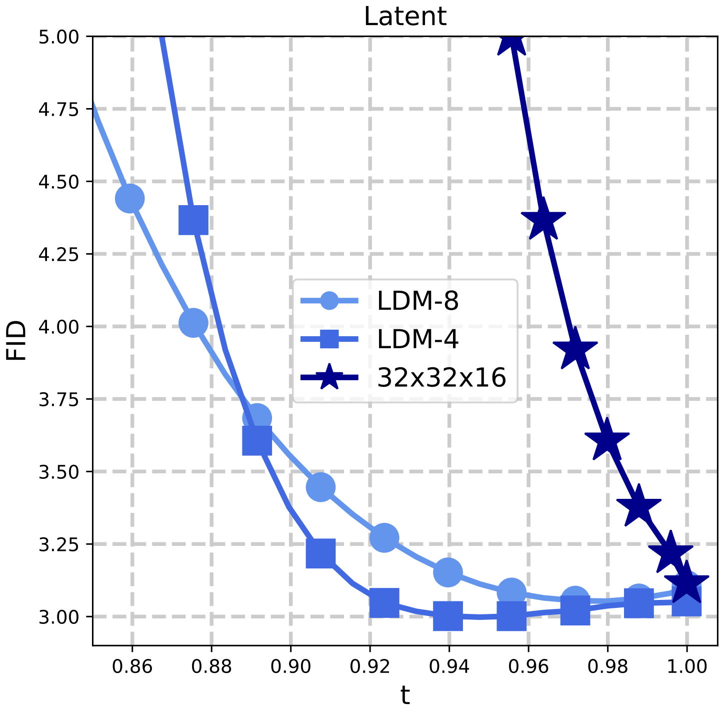

Figure 7: Zoom in of Figure 2.

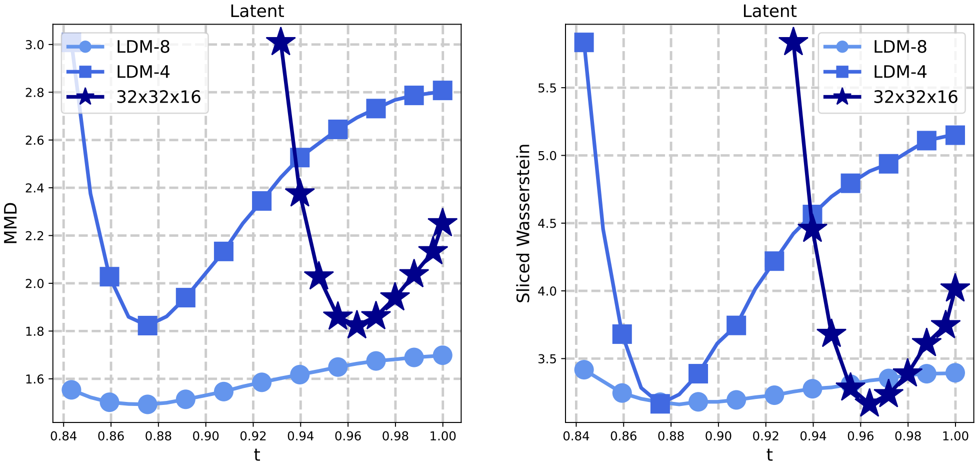

Figure 8: We measure the image quality of different LDMs on ImageNet-256 by MMD and Sliced Wasserstein distance.

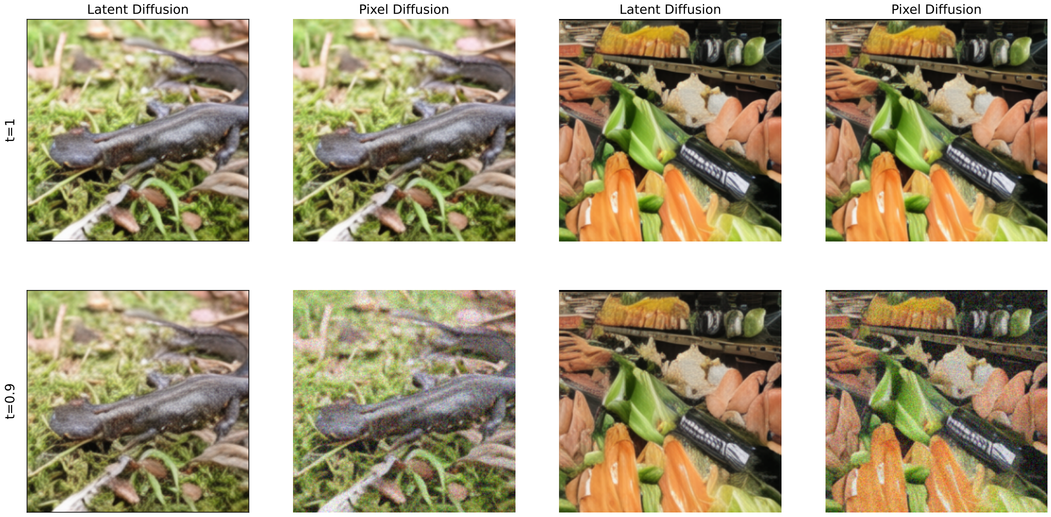

Figure 9: We compare qualitative results of image generated with LDM and pixel diffusion. The noisy pixel diffusion images are sampled by directing adding Gaussian noise to the LDM generated images.

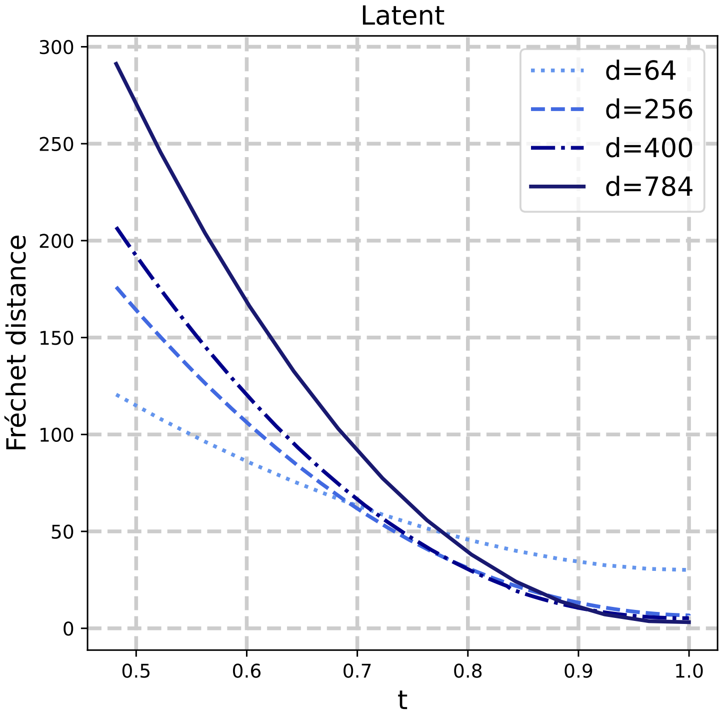

Figure 10: We train linear autoencoders on MNIST, then corresponding U-Net LDMs. The experiment shows a behavior similar to Proposition 2 and Figure 5: for each LDM there exists a time interval such that the LDM is optimal, and for later diffusion times, a larger latent dimension is better.

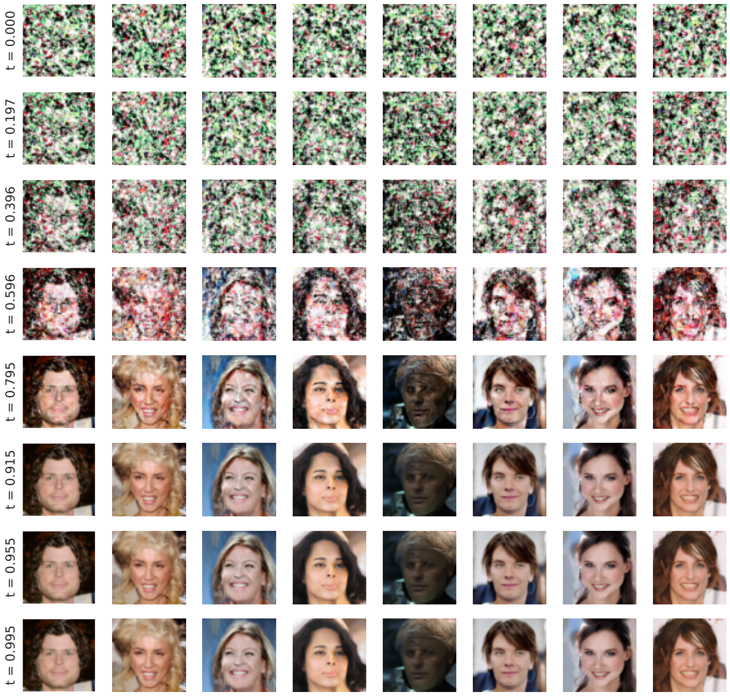

Figure 11: The final steps of LDM do not improve image quality.

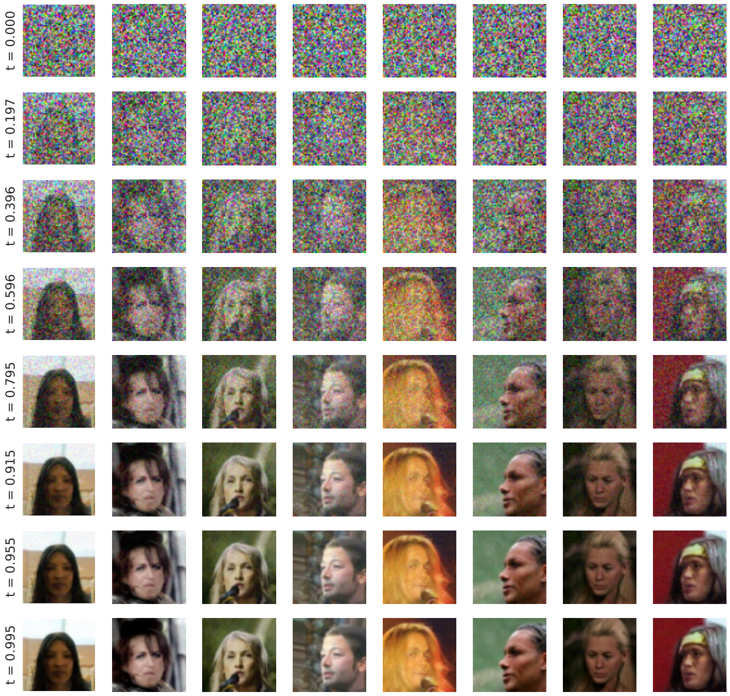

Figure 12: The quality of sample in diffusion on pixel space is still increasing in the final few steps.