Optimal Stopping in Latent Diffusion Models

Yu-Han Wu[*]
LPSM, Sorbonne Université
Google DeepMind

Quentin Berthet
Google DeepMind

Gérard Biau
LPSM, Sorbonne Université
Institut Universitaire de France
Paris, France

Claire Boyer
LMO, Université Paris-Saclay
Institut Universitaire de France
Orsay, France

Romuald Elie
Google DeepMind

Pierre Marion[*†]
Inria, École Normale Supérieure,
PSL Research University

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 and 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.

One well-documented challenge in diffusion model is the onset of numerical instability as the timestep t approaches 0 [Song et al., 2021]. 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

[*]Address correspondence to yhwu@google.com and pierre.marion@inria.fr.
[†]Part of this work was done while the author was at the Institute of Mathematics, EPFL.

EurIPS 2025 Workshop on Principles of Generative Modeling (PriGM).

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 2. 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 4 and Figure 5). 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.

2 Notations and Problem Setup

We consider a diffusion process where the initial distribution, p_0, is a D-dimensional centered Gaussian with independent components and ordered variances:

$$p_0 = \mathcal{N}(0, \Sigma), \text{ where } \Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_D^2) \text{ and } \sigma_1 \geq \dots \geq \sigma_D > 0. \tag{1}$$

The forward diffusion $\overrightarrow{X_t}$ and corresponding backward diffusion $\overleftarrow{X}_t$ (driven by the score function s) are defined as:

$$\begin{aligned} d\overrightarrow{X_t} &= -w_t^2 \overrightarrow{X_t} dt + \sqrt{2w_t^2} d\overrightarrow{W_t}, \quad \overrightarrow{X_0} \sim p_0, \\ d\overleftarrow{X}_t &= (w_{T-t}^2 \overleftarrow{X}_t + 2w_{T-t}^2 s(\overleftarrow{X}_t, T-t))dt + \sqrt{2w_{T-t}^2} dP\overleftarrow{W_t}, \quad \overleftarrow{X}_0 \sim p_T, \end{aligned}$$

where p_t is the distribution of $\overrightarrow{X_t}$. We analyze this system across a hierarchy of latent spaces by projecting the processes onto their first d components using P_d for $d \in \{1, \dots, D\}$. The true projected backward process $P_d \overleftarrow{X}_{T-t}$ and the estimated one $P_d \overleftarrow{\hat{X}}_t$ are Gaussian

$$P_d \overleftarrow{X}_{T-t} \sim \mathcal{N}(0, a_t^2 I_d + b_t^2 P_d \Sigma P_d^\top) \quad \text{and} \quad P_d \overleftarrow{\hat{X}}_{T-t} = \mathcal{N}(0, a_t^2 I_d + b_t^2 P_d \hat{\Sigma} P_d^\top) \tag{2}$$

where $a_t = \sqrt{1 - b_t^2}$, $b_t = e^{-\int_0^t w_i^2 dt}$, and $\hat{\Sigma}$ is the estimated variance matrix (learning the score function reduces to estimating Σ).

The distance between the true and estimated distributions is quantified using the Fréchet distance (d_F) [Heusel et al., 2017], which is equivalent to the Wasserstein-2 distance [Villani, 2008] for Gaussian distributions:

$$d_F^2(\mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}(\mu_2, \Sigma_2)) = \|\mu_1 - \mu_2\|_2^2 + \text{tr}(\Sigma_1 + \Sigma_2 - 2(\Sigma_2^{1/2} \Sigma_1 \Sigma_2^{1/2})^{1/2}). \tag{3}$$

We use $d_F(X, Y)$ to denote the distance between the distributions of random variables X and Y.

3 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. In addition, we assume that the eigenvectors of the true covariance matrix is known, this is made explicit by assuming that the true and estimated covariance matrices are both diagonal.

3.1 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 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)*, 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*

$$\sum_{d'=1}^{d}(1-\frac{\sigma_{d'}}{\hat{\sigma}_{d'}})(1-\hat{\sigma}_{d'}^2) \geq 0. \tag{4}$$

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?

3.2 Optimal projection at time t

This subsection continues our study on the interaction between the projection dimension and the stopping time of the backward process. In contrast to the previous section, we demonstrate that for any fixed time t, an optimal projection dimension P_d exists. We maintain the assumption of Gaussian data with independent components (1). The optimal dimension is characterized by specific time partitions $0 = t_1 \leq t_2 \leq \ldots \leq t_D \leq t_{D+1} = T$ and $0 = \hat{t}_1 \leq \hat{t}_2 \leq \ldots \leq \hat{t}_D \leq \hat{t}_{D+1} = T$ (given by (7) in the Appendix) derived from the true and estimated variances, respectively. This framework allows us to minimize the distance between the generated and target distributions.

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

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

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

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

This proposition shows a time-dependent trade-off: early stages of the backward process are best approximated in lower-dimensional spaces (d increases as $t \to T$). Intuitively, at early times, a lower-dimensional projection avoids introducing more noise than signal. This effect holds for both the true and estimated scores, though a component with sufficiently large variance ($4\sigma_d^2 \geq 1$) should always be included ($t_d = 0$).

We now extend this analysis to data that inherently possesses a low-rank structure. This allows us to precisely determine the two key generation parameters—dimension and stopping time—simultaneously.

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 $1 - 2d_0e^{-\frac{n}{8}}$,*

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

This result demonstrates that the optimal generation strategy for low-rank data requires both early stopping and projection. The optimal stopping time $T - \hat{\delta}_{d_0}$ is shown to be strictly before T (under the non-monotonicity condition of Proposition 1), providing a strong justification for early stopping: it is not merely a practical solution to prevent numerical instability [Yang et al., 2023] but an optimal strategy to minimize the distance between the generated and true data distributions.

4 Performance of the score matching ERM

The previous analysis assumed properties derived from the exact or estimated score of an independent Gaussian distribution. In practice, the score function is learned through a regression problem known

as score matching. Given a training sample $(X_1, \ldots, X_n) \sim p_0$, the empirical score matching objective for learning a predictor s is:

$$\mathcal{R}(s) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{t\sim\mathcal{T},\varepsilon\sim\mathcal{N}(0,I_D)} \left\| s(b_t X_i + a_t \varepsilon, t) + \frac{\varepsilon}{a_t} \right\|^2, \tag{5}$$

where $\mathcal{T}$ is an absolutely continuous distribution over $[0, T]$.

We restrict the predictor s to a hypothesis class $\mathcal{F}_C$ of linear functions with bounded diagonal weights:

$$\mathcal{F}_C = \{s_M(x,t) = -M(t)x : M(t) = \text{diag}(m_1(t), \ldots, m_D(t)), \\ m_i \in \mathcal{L}_2(\mathbb{R}_+, \mathbb{R}), \|m_i\|_\infty < C\}.$$

The condition $C > 1$ ensures compatibility since the backward process begins from a standard Gaussian whose score is the identity function. Let $\hat{M}$ be the optimal score minimizing $\mathcal{R}$. Specializing to the Ornstein-Uhlenbeck process ($w_t \equiv 1$), the generated backward sample $\overleftarrow{\tilde{X}}_t$ follows the SDE:

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

We then characterize the optimal projection dimension for this latent diffusion using the Fréchet distance between the final generated sample $P_d \overleftarrow{\tilde{X}}_T$ and the initial data distribution $\overrightarrow{X_0}$.

Proposition 4. *Define $1 \le d_1 \le d_2 \le D$ as follows:*

$$d_1 = \max\{d' \in \{1, \ldots, D\} : 1/C \le \hat{\sigma}^2_{d'}\} \text{ and } d_2 = \min\left\{d' \in \{1, \ldots, D\} : \frac{1}{2C-1} \ge 4\sigma^2_{d'}\right\}.$$

(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 $d_1 \le d_{\min} \le d_2$ such that

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

Interestingly, the optimal projection can be made explicit for exponentially-decaying covariance spectrum.

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

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

5 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 Σ, and the backward processes $\overleftarrow{X}_t$ and $\overleftarrow{\hat{X}}_t$ given as in (2) with the new general data distribution p_0. To this end, let $\Sigma = O\Lambda O^\top$ be the eigen decomposition of Σ, where O is an orthogonal matrix and Λ is the diagonal matrix of eigenvalues, which we assume are distinct and ordered $\sigma_1^2 > \ldots > \sigma_D^2 > 0$. Then, there are timesteps $0 = t_1 \le t_2 \le \ldots \le t_D \le t_{D+1} = T$ given by (8) in the Appendix. 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 5. *For $2 \le d \le D$ and $t \in [t_d, t_{d+1})$, we have*

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

However, in practical applications, 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. Given $u > 0$, define timesteps $\hat{T}_d(u)$ and $\hat{t}_d(u)$ as in (9) and (10) in the Appendix. We are now in a position to describe the optimal projection strategy at each stopping time.

Proposition 6. *For $d \in \{1, \dots, D\}$ and any $t \in [\hat{T}_d(u), \hat{t}_{d+1}(u)]$, with probability $1 - 2e^{-u}$,*

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

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.

6 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 and 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.

References

J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/google/jax`.

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

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

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

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

S. Hurault, M. Terris, T. Moreau, and G. Peyré. From score matching to diffusion: A fine-grained error analysis in the Gaussian setting. *arXiv:2503.11615*, 2025.

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

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

X. Liu, C. Gong, and Q. Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv:2209.03003*, 2022.

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

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

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

O. Ronneberger, P. Fischer, and T. 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*, pages 234–241. Springer, 2015.

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

Y. Song, C. Durkan, I. Murray, and S. Ermon. Maximum likelihood training of score-based diffusion models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 1415–1428. Curran Associates, Inc., 2021.

A. Vahdat, K. Kreis, and J. Kautz. Score-based generative modeling in latent space. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 11287–11302. Curran Associates, Inc., 2021.

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

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

Z. Yang, R. Feng, H. Zhang, Y. Shen, K. Zhu, L. Huang, Y. Zhang, Y. Liu, D. Zhao, J. Zhou, et al. Lipschitz singularities in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023.

Appendix

A Timesteps for optimal projections

Independent Gaussian. Recall that a is defined in (2) 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 1), meaning that

$$\bar{a}^{-2}(x) = \begin{cases} 0, & \text{for } x < 0, \\ a^{-2}(x), & \text{for } x \in [0, a_T^2], \\ T, & \text{for } x \in (a_T^2, \infty]. \end{cases} \tag{6}$$

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

$$t_d = T - \bar{a}^{-2}\Big(\frac{3\sigma_d^2}{(1-\sigma_d^2)_+}\Big) \quad \text{and} \quad \hat{t}_d = T - \bar{a}^{-2}\Big(\frac{4\sigma_d^2 - \hat{\sigma}_d^2}{(1-\hat{\sigma}_d^2)_+}\Big). \tag{7}$$

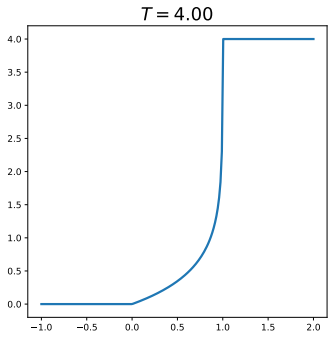

Figure 1: $\bar{a}^{-2}$ in the Ornstein-Uhlenbeck process.

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.

General Gaussian. As in Section 3, 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:

$$t_d = T - \bar{a}^{-2}\left(\frac{3\sigma_d^2}{(1-\sigma_d^2)_+}\right), \tag{8}$$

where $\bar{a}^{-2}$ is given in (6). This definition, combined with the ordering of the eigenvalues, yields a sequence $0 = t_1 \le t_2 \le \cdots \le t_D \le t_{D+1} = T$.

Denote $S(\Sigma) = \sum_{d'=1}^{D} \max(\sigma_d, \sigma_d^2)$. For $u \ge 0$ and $d \in \{2, \dots, D\}$, we let $\hat{T}_d(u)$ and $\hat{t}_d(u)$ be

$$\hat{T}_d(u) = T - \bar{a}^{-2}\left(\frac{\hat{\sigma}_d^2 - 4S(\Sigma)\varepsilon_u + 2\hat{\sigma}_d\sqrt{\hat{\sigma}_d^2 - 4S(\Sigma)\varepsilon_u}}{(1-\hat{\sigma}_d^2)_+}\right), \tag{9}$$

$$\hat{t}_d(u) = T - \bar{a}^{-2}\left(\frac{\hat{\sigma}_d^2 + 4S(\Sigma)\varepsilon_u + 2\hat{\sigma}_d\sqrt{\hat{\sigma}_d^2 + 4S(\Sigma)\varepsilon_u}}{(1-\hat{\sigma}_d^2)_+}\right), \tag{10}$$

where $\varepsilon_u = \frac{8C}{3}(\sqrt{\frac{D+u}{n}} + \frac{D+u}{n})$. We assume that ε_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 ε_u, these timesteps are ordered as

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

B Proofs of results

B.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 (4) 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 (3):

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

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,

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

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 $a_t = 0$.

B.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 $d \in \{1, \ldots, D\}$,

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

so that

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

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.

B.3 Proof of Proposition 3

We first state the full proposition.

Proposition 7. *Assume that $\Sigma = \mathrm{diag}(\sigma^2, \ldots, \sigma^2, 0, \ldots, 0)$ 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*

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

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

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

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 8, by the union bound, with probability $1 - 2d_0 e^{-\frac{\varepsilon^2 n}{4(1+\varepsilon)}} \geq 1 - 2d_0 e^{-\frac{\varepsilon^2 n}{8}}$ 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 $\bar{a}^{-2}$,

$$\hat{t}_d = T - \bar{a}^{-2}\Big(\frac{4\sigma_d^2 - \hat{\sigma}_d^2}{1-\hat{\sigma}_d^2}\Big)$$

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

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 $d \in \{1, \dots, d_0\}$. On the other hand, $t < T = \hat{t}_d$ for all $d \in \{d_0+1, \dots, D\}$ 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 $d \in \{1, \dots, d_0\}$, which holds with probability $1 - 2d_0 e^{-n/8}$ by Proposition 8. 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 $d \in \{d_0+1, \dots, D\}$, 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 $\hat{t}_1 = \max_{d \in \{1,\dots,d_0\}} \hat{t}_d$, 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 $d \in \{1, \dots, d_0\}$, 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,

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

Consider f_d given by

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

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

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

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

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

then

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

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

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

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

$$\begin{aligned}
& 1 - \frac{\sigma}{\sqrt{\hat{\sigma}^2_{d'} + (1-\hat{\sigma}^2_{d'})a^2_{T-\hat{t}}}} \geq 0. \\
\Leftrightarrow \quad & \sigma^2 \leq \hat{\sigma}^2_{d'} + (1-\hat{\sigma}^2_{d'})\frac{4\sigma^2 - \hat{\sigma}^2_1}{1-\hat{\sigma}^2_1} \\
\Leftrightarrow \quad & \sigma^2 \leq \hat{\sigma}^2_{d'} + (1-\hat{\sigma}^2_{d'})\Big(1 - \frac{1-4\sigma^2}{1-\hat{\sigma}^2_1}\Big) \\
\Leftrightarrow \quad & \sigma^2 \leq 1 - (1-\hat{\sigma}^2_{d'})\frac{1-4\sigma^2}{1-\hat{\sigma}^2_1} \\
\Leftrightarrow \quad & (1-\hat{\sigma}^2_{d'})\frac{1-4\sigma^2}{1-\hat{\sigma}^2_1} \leq 1 - \sigma^2.
\end{aligned}$$

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

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

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

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

which derives the desired inequality and we conclude the proof.

B.4 Derivation of the optimal score function in $\mathcal{F}_C$

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

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

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 $d \in \{1,\dots,D\}$. Hence it suffices to minimize the quantity

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

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

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

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.

B.5 Proof of Proposition 4

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 $P_d \overleftarrow{\tilde{X}}_0$ for all d. Denote for $1 \le d \le D$, the variance of $\overleftarrow{\tilde{X}}_{t,d}$ by $V_{t,dd}$. It is known [see, for instance, Särkkä and 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:

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

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

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 $C = \hat{m}_d(T - t'_d) = \frac{1}{a_{T-t'_d}^2 + b_{T-t'_d}^2\hat{\sigma}_d^2}$. Otherwise, we set $t'_d = T$, which is always the case for $d \le d_1$. Remark that if $\hat{\sigma}_d \ge 1$, then $\frac{1}{a_t^2 + b_t^2\hat{\sigma}_d^2} \le 1 \le C$. Thus, for such dimension d, we always have $t'_d = T$ and $d \le d_1$.

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.

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

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

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

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

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

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

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

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

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

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

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

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 $\hat{\sigma}_d \geq 1$. We get the same bounds as in the two previous cases.
- Finally, if $\sigma_d \leq \frac{1}{2}$, with high probability, $|\sigma_d^2 - \hat{\sigma}_d^2| \leq \sigma_d^2$. Hence, $\hat{\sigma}_d^2 \leq 2\sigma_d^2 \leq \frac{1}{2}$. Observing that the fraction in (21) is bounded by $1/(1-\hat{\sigma}_d^2)$, we deduce that

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

which gives the desired bound.

Next, for $d \geq d_2$, 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,

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

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

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

To summarize, we derived the following bounds

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

and

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

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.

B.6 Derivation of special cases of Proposition 4

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 4 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 $1/C \leq \min(\sigma^2, \min_{d' \in \{1,\ldots,d_0\}} \hat{\sigma}_{d'}^2)$. 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 4 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,

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

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

B.7 Proof of Proposition 5

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.

B.8 Proof of Proposition 6

Recall that $\hat{\Lambda} = \text{diag}(\hat{\sigma}_1^2, \dots \hat{\sigma}_D^2)$ 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

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

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

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

In the case of interest for us, we get

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

We now argue that $\text{tr}((\hat{\Sigma}_d^{1/2}(t)\Sigma\hat{\Sigma}_d^{1/2}(t))^{1/2})$ 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 Σ and $\hat{\Sigma}$ commute, which was the case in the previous sections where we assumed that both matrices were diagonal. By Proposition 9, 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 and 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

$$\begin{aligned}\text{tr}((\hat{\Sigma}_d^{1/2}(t)\Sigma\hat{\Sigma}_d^{1/2}(t))^{1/2}) &\le \sqrt{\frac{1}{1-\varepsilon_u}}\text{tr}((\hat{\Sigma}_d^{1/2}(t)\hat{\Sigma}\hat{\Sigma}_d^{1/2}(t))^{1/2}) \\ &\le (1+\varepsilon_u)\text{tr}((\hat{\Sigma}_d^{1/2}(t)\hat{\Sigma}\hat{\Sigma}_d^{1/2}(t))^{1/2}),\end{aligned}$$

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

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

By combining the results, we obtain

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

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

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

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

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

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

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

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

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

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,

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

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 ε_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.

C Bounds on Gaussian estimation

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

Proposition 8. *Let $(X_1, \dots, X_n)$ be sample drawn independently from $\mathcal{N}(0, \sigma^2)$. Then, for $\varepsilon > 0$, we have*

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

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

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

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 9. *Let Σ be a semi-definite positive $D \times D$ matrix, and assume the sample $(X_1, \dots, X_n)$ 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 σ, 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

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

□

D Useful Lemma

In this section we provide some lemma that will be useful throughout the whole paper [see also, Horn and 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 $A \preceq B \Rightarrow \sqrt{A} \preceq \sqrt{B}$.

E Experiment details

E.1 Natural image experiment

Common details. We use the dataset CelebA and CelebA-HQ [Liu et al., 2015]. 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. 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.

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.

Training of LDM. 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 2.

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 2: 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 3: Hyperparameters for training diffusion model on CelebA.

Training pixel diffusion model on CelebA. We train a diffusion model on CelebA. We train for 1 million steps on 12 TPUv2. We summarize the hyperparameters in Table 3.

Results. We previously introduced some results in Section 1. Here, we present additional evidence regarding the quality of the generated images. We observe (Figure 4) that in the final few steps, the sample of LDM does not change visibly. On the contrary, the images generated in pixel space (Figure 5) are still denoised even in the last steps.

Synthetic Gaussian data. In the experiment of Figure 3, 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 backward SDE separately for each projection. We generate new sample using the Ornstein-Uhlenbeck process with $T = 2$ and 1000 discretization steps.

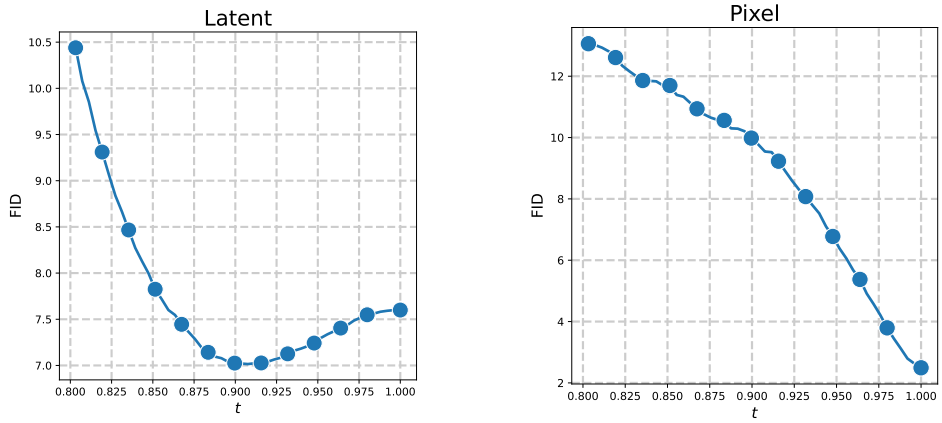

Figure 2: (left) FID-30k score of latent diffusion model on CelebA-HQ, with latent shape $64 \times 64 \times 3$. (right) FID-30k score of standard diffusion model (trained in pixel space) on CelebA64 ($64 \times 64 \times 3$).

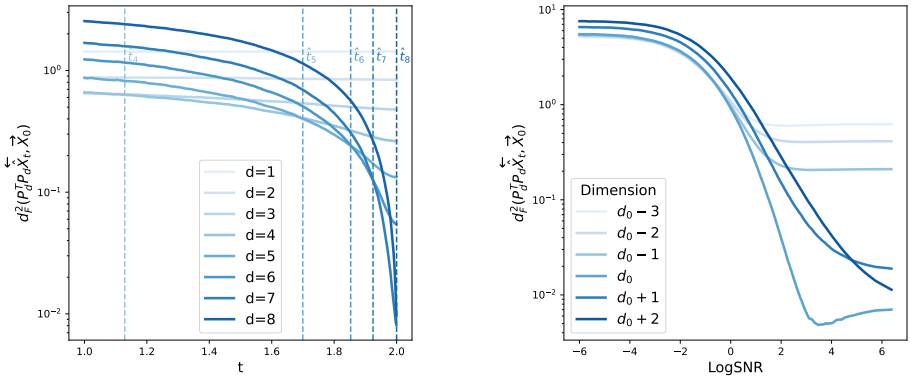

Figure 3: Plots of $d_F^2(P_d^\top P_d \overleftarrow{\hat{X}}_t, \overrightarrow{X_0})$ as a function of the diffusion time t, for two sets of variances $(\sigma_1, \dots, \sigma_D)$. (left) All the σ_i are nonzero. As expected from Proposition 2, the d-dimensional projection is optimal in $[t_d, t_{d+1})$. (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 E.

Figure 4: The final steps of LDM do not improve image quality.

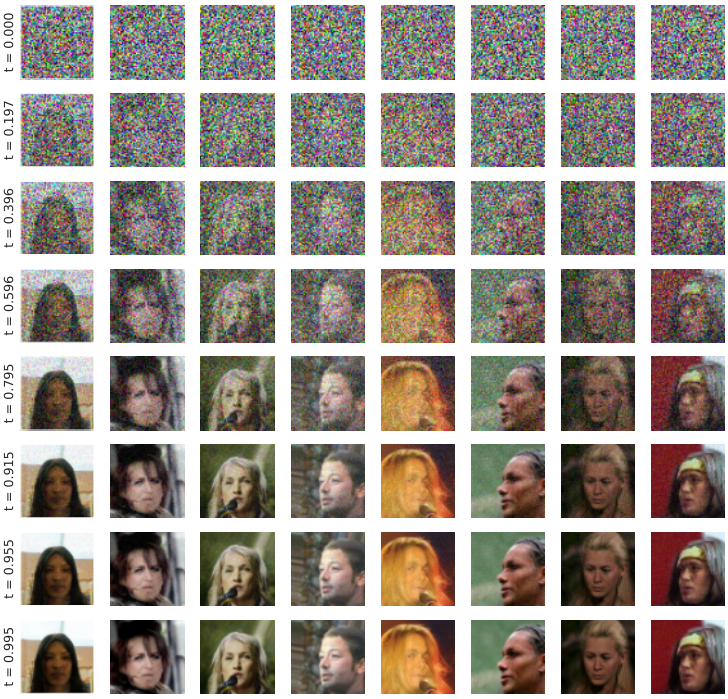

Figure 5: The quality of sample in diffusion on pixel space is still increasing in the final few steps.