Recursively Trained Diffusion Models: Limiting Collapse Distribution and Spectral Characterization

Nail B. Khelifa
University of Cambridge
nbk24@cam.ac.uk

Richard E. Turner
University of Cambridge
ret26@cam.ac.uk

Ramji Venkataramanan
University of Cambridge
rv285@cam.ac.uk

Abstract

Recursive training of generative models on their own outputs can lead to model collapse, a compounding drift away from the true data distribution. Existing theoretical works bound finite-round error accumulation in the context of diffusion models, but two questions remain open: what distribution does the recursion converge to, and how fast? We answer both, isolating a mechanism distinct from imperfect learning: even with perfect score estimation and exact sampling, the early stopping of the reverse diffusion (required for numerical stability) drives a progressive drift away from the data distribution. We prove that this recursion converges geometrically to a unique limiting distribution, which admits a closed-form characterization as an infinite mixture of increasingly Gaussian-smoothed versions of the data distribution. A Hermite spectral decomposition of this limit reveals that recursive training acts as a low-pass filter: higher-order modes, which encode fine non-Gaussian structure, are attenuated much more strongly than coarse modes. This spectral picture motivates annealed truncation schedules that progressively shrink truncation times across retraining rounds; we prove that any schedule converging to 0 asymptotically eliminates recursive compounding. Finally, we show our idealized characterization is robust: in the presence of discretization and score estimation errors, the learned distribution remains in a Wasserstein-2 ball around the ideal limit, with mode-dependent contraction rates that contract high-order errors faster than low-order ones. We validate the theory on synthetic Gaussian mixtures and CIFAR-10.

1 Introduction

Following the success of generative modeling, the use of synthetic data for training has attracted considerable attention. Such generated data may be used deliberately to improve or fine-tune existing models [33, 2, 71], or may enter training sets inadvertently when data are polluted with outputs of generative models [13]. It has been found that repeatedly retraining models on their own outputs causes systematic drift away from the target distribution [1, 59], a phenomenon referred to as *model collapse*. Subsequent work characterized collapse as a loss of the tails of the target distribution and a decrease in the diversity of generated samples [25, 57]. In settings that allow for a combination of synthetic and real data in each retraining iteration, different behaviors have been observed: in some cases, a small proportion of synthetic data can significantly degrade performance [24], while in others, it has been shown that collapse can be mitigated [32, 23, 22, 74, 31] or completely avoided [9]. In some settings, synthetic data could even be beneficial when used judiciously [40, 25].

In the context of diffusion models [60, 62, 64, 37], recent theory has quantified finite-round error propagation under population-level, finite-sample, or model-specific assumptions [28, 18, 45, 74]. However, these works do not characterize the asymptotic behavior of iteratively retrained diffusion models, i.e., whether the model tends toward a collapsed regime in which further retraining no longer changes the learned distribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Setting Let p_{data} denote the true target data distribution on $\mathbb{R}^d$. We consider a recursive procedure which, in each training round (generation), mixes fresh samples from p_{data} with synthetic samples generated by the current model. At generation $i \geq 0$, a proportion $\alpha \in (0, 1)$ of the training data consists of fresh samples from p_{data}, while the remaining proportion $(1 - \alpha)$ consists of synthetic samples drawn from the learned model at the previous generation, denoted $\hat{p}^i$. To capture a realistic setting where training is agnostic to how samples were generated, we assume samples are not labeled as fresh or synthetic. As a result, the effective training distribution at generation i is the mixture

$$q_i := \alpha\, p_{\text{data}} + (1 - \alpha)\, \hat{p}^i. \tag{1}$$

A score network is then trained on samples from q_i, and the resulting diffusion model defines the next synthetic distribution $\hat{p}^{i+1}$. Starting from $\hat{p}^0 = p_{\text{data}}$, this defines the recursion

$$\hat{p}^i \xrightarrow{\text{mix with } p_{\text{data}}} q_i \xrightarrow{\text{train model}} \hat{p}^{i+1}. \tag{2}$$

This framework matches that of existing works [45], and may be generalized to a setting where each past generation $k \leq i$ is sampled from in proportion α_k [28]. The goal of this work is to answer the following questions:

In this recursive setting, does the sequence $(\hat{p}^i)_{i\geq 0}$ converge to a specific distribution as i grows? If so, can it be characterized explicitly, and at what rate does the recursion converge to it?

Understanding this asymptotic behavior is fundamental because it identifies the regime towards which a model is driven by recursive training. Analyzing the collapsed regime highlights the role of each source of error and the mechanisms by which collapse occurs.

Sources of error Training and sampling a diffusion model in this recursive setting introduce four distinct sources of error. The first, and the main focus of this paper, is *truncation*: in practice, the reverse SDE is only integrated on $[t_0, T]$ for some $t_0 > 0$, because score estimation becomes numerically unstable as $t \downarrow 0$ [63, 46] (formalized in Appendix E). While modern samplers introduce various corrections (e.g. Tweedie-style denoising [26], x_0-prediction [44], and other post-truncation refinements [73]), the requirement of stopping the reverse SDE at $t_0 > 0$ is standard. As a result, even with perfect score estimation, exact initialization, and exact numerical sampling, the generated distribution is not q_i itself but a Gaussian-smoothed version of q_i. The second source of error is the score estimation error, caused by finite data and function-approximation constraints. The third is the discretization error, introduced when the reverse SDE is solved numerically. The fourth is the initialization error: the reverse SDE is started from $\mathcal{N}(0, \mathbf{I}_d)$ rather than from $q_{i,T}$, which induces an error that decays exponentially in T. There is a rich body of work providing non-asymptotic guarantees for diffusion-based models, accounting for score estimation errors under different discretization schemes [16, 10, 20, 14, 72, 29, 17, 48].

Our analysis first isolates truncation and shows that it alone induces a progressive drift away from p_{data} with recursive training. Errors due to imperfect score estimation, time discretization and initialization are then treated as perturbations of this idealized mechanism.

Contributions This paper develops an operator-theoretic framework for recursively trained diffusion models and characterizes their limiting behavior. Our main contributions are:

- **Explicit collapse distribution.** We identify truncation as a primary driver of model collapse and show that, in the ideal case with no discretization or score estimation errors, the recursive dynamics (2) converge to a unique limiting distribution $p^\star_\infty$. We derive a closed-form expression for $p^\star_\infty$ as an infinite mixture of progressively Gaussian-smoothed versions of p_{data}, and establish a geometric rate of convergence to that limit in the error-free regime (Theorem 3.1).
- **Spectral structure of collapse.** Decomposing the limiting distribution $p^\star_\infty$ in the Hermite polynomial basis, we provide a spectral characterization of model collapse (Proposition 3.4), which shows that recursive training acts as a low-pass filter on p_{data}, suppressing high-order components more strongly than low-order ones, with explicit mode-dependent attenuation controlled by α and the stopping time t_0.
- **Annealed truncation schedules.** Having identified fixed truncation as a driver of model collapse, we introduce *annealed* truncation schedules with generation-adaptive truncation times $t_0^{(N)}$. We prove that any schedule converging to zero asymptotically eliminates recursive

compounding in the error-free regime; schedules converging to a positive limit attenuate but do not remove compounding (Theorem 4.1).

- **Robustness and rate of collapse under imperfect training.** We extend the analysis beyond the ideal setting to account for discretization and score estimation errors. We show that the recursive dynamics remain stable, converging at an exponential rate to a Wasserstein-2 ball around the ideal limiting distribution $p^\star_\infty$ (Theorem 5.2 and Proposition 5.3). Thus, the limiting collapse regime becomes relevant even with a moderate number of rounds of recursive training.

Organization Section 2 introduces the mathematical background and the assumptions underlying our analysis. Section 3 shows that, even in an idealized regime (perfect score estimation and exact sampling), model collapse occurs because of truncation of the reverse SDE as $t \downarrow 0$. Section 4 shows that allowing the truncation time to vary across generations changes the asymptotic behavior of the recursion. Section 5 provides robustness results for the idealized framework in the presence of errors.

Notation Bold symbols (e.g. $\mathbf{X}_t$, $\mathbf{Y}_t$) denote $\mathbb{R}^d$-valued random variables or processes, and $\mathcal{P}_2(\mathbb{R}^d)$ denotes the space of Borel probability measures on $\mathbb{R}^d$ with finite second moment. The space of continuous functions from $[0,T]$ to $\mathbb{R}^d$ is denoted by $C([0,T],\mathbb{R}^d)$. For random variables $\mathbf{X}_1, \mathbf{X}_2$ defined on the same probability space, we write $\mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2$ when $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent. For a stochastic process $(\mathbf{Z}_t)_{t\in[0,T]}$ defined on $C([0,T],\mathbb{R}^d)$, we write $\mathrm{Law}((\mathbf{Z}_t)_t)$ for its law on path space and $\mathrm{Law}(\mathbf{Z}_t)$ for the law of its marginal at time t. For two measures μ and ν, we write $\mu \ll \nu$ to denote that μ is absolutely continuous with respect to ν, meaning that every ν-null set is also μ-null. All probability measures $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ are assumed absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$ and, with some abuse of notation, we also use μ to refer to their density. Calligraphic symbols (e.g. $\mathcal{T}$, $\mathcal{U}$, $\mathcal{S}$) refer to operators on probability distributions i.e., mapping $\mathcal{P}_2(\mathbb{R}^d) \to \mathcal{P}_2(\mathbb{R}^d)$, and Id is the identity operator ($\mathrm{Id}(\mu) = \mu$ for $\mu \in \mathcal{P}_2(\mathbb{R}^d)$). For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, the Wasserstein-2 distance between μ and ν, denoted $\mathcal{W}_2(\mu,\nu)$ is defined as

$$\mathcal{W}_2(\mu,\nu) := \left(\inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|\mathbf{x}-\mathbf{y}\|^2 \,\pi(\mathrm{d}\mathbf{x},\mathrm{d}\mathbf{y})\right)^{1/2}, \tag{3}$$

where $\Pi(\mu,\nu)$ denotes the set of all couplings of μ and ν, i.e., probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν, respectively.

2 Background

Forward diffusion The i-th generation diffusion model is defined by a variance-preserving Ornstein-Uhlenbeck (OU) process [43, 64] on the interval $[0,T]$, initialized at the mixture q_i:

$$\mathrm{d}\mathbf{X}^i_t = -\tfrac{1}{2}\mathbf{X}^i_t\,\mathrm{d}t + \mathrm{d}\mathbf{B}_t, \qquad \mathbf{X}^i_0 \sim q_i, \tag{4}$$

where $(\mathbf{B}_t)_{t\in[0,T]}$ is a standard d-dimensional Brownian motion. For each time $t \in [0,T]$, the OU process (4) defines a sampling operator $\mathcal{U}_t : \mathcal{P}_2(\mathbb{R}^d) \to \mathcal{P}_2(\mathbb{R}^d)$ [27, 52, 7, 19] that propagates the initial distribution q_i through the diffusion as follows:

$$\mathcal{U}_t q_i := \mathrm{Law}(e^{-t/2}\mathbf{X}_0 + \sqrt{1-e^{-t}}\mathbf{Z}), \quad \mathbf{X}_0 \sim q_i \perp\!\!\!\perp \mathbf{Z} \sim \mathcal{N}(0,\mathbf{I}_d). \tag{5}$$

Therefore, we denote the time-t marginal of q_i in (4) by

$$q_{i,t} := \mathrm{Law}(\mathbf{X}^i_t) = \mathcal{U}_t q_i.$$

The sampling operator $\mathcal{U}_t$ is central to our analysis, and we provide extended background on its properties in Appendix B.

Reverse-time generation and truncation Under standard regularity conditions [3, 35], the time reversal of (4) solves the reverse-time SDE integrated backward from $s = T$ to $s = 0$,

$$\mathrm{d}\mathbf{Y}^{i,\star}_s = \big[-\tfrac{1}{2}\mathbf{Y}^{i,\star}_s - \nabla_{\mathbf{x}} \log q_{i,s}(\mathbf{Y}^{i,\star}_s)\big]\,\mathrm{d}s + \mathrm{d}\bar{\mathbf{B}}_s, \tag{6}$$

where $(\bar{\mathbf{B}}_s)_{s\in[0,T]}$ is another Brownian motion on the same probability space. If we run the reverse-time SDE starting from $\mathbf{Y}^{i,\star}_T \sim q_{i,T}$, then $\mathbf{Y}^{i,\star}_0 \sim q_{i,0}$ [3, 35]. The true scores $\nabla \log q_{i,t}$ are unknown

and are approximated by a learned time-dependent network $\mathbf{s}_{\theta_i}$ trained on samples from q_i in (1). However, since q_i may be non-smooth or singular, the score $\nabla \log q_{i,t}$ can become unstable as $t \downarrow 0$ [15, 73]. Therefore, as is common in the literature [45, 10, 16], we work on a truncated interval $[t_0, T]$ with $t_0 > 0$, where the diffusion has already regularized the law: under suitable nondegeneracy and regularity assumptions [4, 5], $q_{i,t}$ admits a smooth density for every $t > 0$, so that the score is well defined and better behaved away from the singular endpoint. Substituting the learned score $\mathbf{s}_{\theta_i}$ into (6) yields

$$\mathrm{d}\hat{\mathbf{Y}}^i_s = \big[-\tfrac{1}{2}\hat{\mathbf{Y}}^i_s - \mathbf{s}_{\theta_i}(\hat{\mathbf{Y}}^i_s, s)\big]\mathrm{d}s + \mathrm{d}\bar{\mathbf{B}}_s, \tag{7}$$

with initialization $\hat{\mathbf{Y}}^i_T \sim \mathcal{N}(0, \mathbf{I}_d)$. The next generation model is therefore $\hat{p}^{i+1} := \mathrm{Law}(\hat{\mathbf{Y}}^i_{t_0})$.

Sampling operators The reverse SDEs (6) and (7) define two sampling procedures. In the absence of score estimation and discretization errors, the reverse sampling operator for (6) is exactly $\mathcal{U}_{t_0}$ and $\hat{p}^{i+1} = \mathcal{U}_{t_0}(q_i)$. Accounting for these two errors, the *learned sampling operator* at generation i, denoted $\hat{\mathcal{S}}_i$ is defined as $\hat{\mathcal{S}}_i(q_i) = \hat{p}^{i+1}$. These operators formally define the one-step recursion in (2).

Regularity conditions Let $\mathcal{P}_{\mathrm{reg}}(\mathbb{R}^d)$ be the set of probability distributions μ satisfying the following assumptions:

- **(A1)** μ admits a strictly positive smooth density with respect to the Lebesgue measure: $\mu(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathbb{R}^d$.
- **(A2)** Finite second moment: $\int_{\mathbb{R}^d} \|\mathbf{x}\|^2 \mu(\mathbf{x})\,\mathrm{d}\mathbf{x} < \infty$.
- **(A3)** The score $\nabla \log \mu(\mathbf{x})$ is square-integrable under μ.

Throughout, we assume that $p_{\mathrm{data}} \in \mathcal{P}_{\mathrm{reg}}(\mathbb{R}^d)$, and observe that assumptions A1–A3 are preserved under the forward diffusion (4) and under convex combinations.

3 Limiting Collapse Distribution in the Error-free Regime

In this section, we isolate the effect of truncation by studying an idealized error-free regime in which score estimation and discretization errors are ignored and the reverse SDE is initialized exactly at $q_{i,T}$. This setting shows that collapse does not require imperfect learning: truncation of the reverse SDE (at $t_0 > 0$) alone is sufficient to drive the recursive dynamics toward a nontrivial limiting distribution.

Time reversal and one-step dynamics In the error-free regime, the transition from one generation to the next is defined by the true reverse SDE (6) whose marginals, by time-reversal [3, 35], exactly match those of the forward SDE (4). For each $i \geq 0$, the model $\hat{p}^{i+1}$ can therefore be expressed using the OU operator (5) applied at the truncation time t_0: $\hat{p}^{i+1} = \mathcal{U}_{t_0}(q_i) = \mathrm{Law}(e^{-t_0/2}\mathbf{X}^i_0 + \sqrt{1-e^{-t_0}}\mathbf{Z})$. Thus truncation alone introduces a mismatch between $\hat{p}^{i+1}$ and q_i, the target distribution at generation i, and it corresponds to a small amount of residual Gaussian noise. Since this smoothing is repeated recursively, Gaussian-convolved residuals accumulate over generations, and truncation alone leads to progressive drift away from the true data distribution p_{data}.

Collapse distribution Since $\hat{p}^{i+1} = \mathcal{U}_{t_0}(\alpha p_{\mathrm{data}} + (1-\alpha)\hat{p}^i)$, the transition from the i-th to the $(i+1)$-th generation is generation-independent. Therefore, the limiting behavior of the recursion amounts to the fixed-point problem of

$$\mu \mapsto \mathcal{U}_{t_0}\big(\alpha p_{\mathrm{data}} + (1-\alpha)\mu\big). \tag{8}$$

The following theorem shows that this recursion admits a unique fixed point and expresses it as a Neumann series (proof in Appendix D.1).

Theorem 3.1 (Collapse distribution: existence, uniqueness, and explicit form)**.** *Under A1–A3, there exists a unique collapse distribution $p^\star_\infty \in \mathcal{P}_2(\mathbb{R}^d)$ that is a fixed point of (8). It is given by,*

$$p^\star_\infty \;=\; \alpha \sum_{k=0}^{\infty} (1-\alpha)^k\, \mathcal{U}_{(k+1)t_0}(p_{\mathrm{data}}). \tag{9}$$

Moreover, in the absence of score estimation, discretization and initialization errors, the sequence $(\hat{p}^N)_{N\geq 1}$ converges geometrically fast to $p^\star_\infty$: $\mathcal{W}_2(\hat{p}^N, p^\star_\infty) \;\leq\; \kappa^N\, \mathcal{W}_2(p_{\mathrm{data}}, p^\star_\infty)$ where $\kappa = \sqrt{1-\alpha}\, e^{-t_0/2}$.

Theorem 3.1 shows that the collapse distribution is a mixture, with geometric weights, of increasingly smoothed copies of the true data distribution. The kth component of the mixture, $\mathcal{U}_{(k+1)t_0}(p_{\text{data}})$, is the data distribution smoothed by $(k+1)t_0$ units of OU evolution, and it is weighted by $\alpha(1-\alpha)^k$. The rate of convergence depends on both the fresh-data proportion α and the truncation time t_0 through $\kappa = \sqrt{1-\alpha}\, e^{-t_0/2}$. The larger the t_0, the larger the Gaussian noise injection at each generation, and the faster the convergence; the larger the α, the faster the convergence as well, since the fresh component of q_i is identical across generations and only the synthetic component, of weight $(1-\alpha)$, carries memory of the past. The two parameters act differently on the *limit*, however: a larger α brings $p^\star_\infty$ closer to the one-step smoothed distribution $\mathcal{U}_{t_0}(p_{\text{data}})$, while a smaller t_0 brings it closer to p_{data} (Corollary 3.2, proof in Appendix D.2).

Corollary 3.2 (Limiting behaviors)**.** *The exact collapse distribution* (9) *satisfies the following:*

(i) $\alpha \to 1^-$*:* $\mathcal{W}_2(p^\star_\infty,\ \mathcal{U}_{t_0}(p_{\text{data}})) \to 0$.

(ii) $\alpha \to 0^+$*:* $\mathcal{W}_2(p^\star_\infty,\ \mathcal{N}(0, \mathbf{I}_d)) \to 0$.

(iii) $t_0 \to 0^+$*:* $\mathcal{W}_2(p^\star_\infty,\ p_{\text{data}}) \to 0$.

(iv) For any $\alpha \in (0,1)$*,* $t_0 > 0$*, and* $p_{\text{data}} \not\equiv \mathcal{N}(0, \mathbf{I}_d)$*:* $\mathcal{W}_2(p^\star_\infty,\ p_{\text{data}}) > 0$.

These limits have simple interpretations. When $\alpha \to 1$, the recursion effectively disappears and only the one-step truncation bias remains. When $\alpha \to 0$, the fraction of fresh data injected in each retraining generation vanishes, letting Gaussian smoothing compound without attenuation; this ultimately drives the recursion towards the standard Gaussian (the invariant measure of the OU semigroup). When $t_0 \to 0^+$, truncation vanishes and the collapse distribution approaches p_{data}.

Example 3.3 (Gaussian mixture)**.** An interesting example to understand $p^\star_\infty$ is the case where p_{data} is a Gaussian mixture, i.e. $p_{\text{data}} = \sum_{j=1}^J w_j \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, with non-negative weights $(w_j)_{j\in[J]}$ summing to 1, means $\boldsymbol{\mu}_j \in \mathbb{R}^d$ and positive definite covariance matrices $\boldsymbol{\Sigma}_j \in \mathbb{R}^{d\times d}$ for $j \in [J]$. Since $\mathcal{U}_t\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}\big(e^{-t/2}\boldsymbol{\mu},\ e^{-t}\boldsymbol{\Sigma} + (1-e^{-t})\mathbf{I}_d\big)$, from (9) the collapse distribution is

$$p^\star_\infty = \sum_{k=0}^{\infty} \alpha(1-\alpha)^k \sum_{j=1}^{J} w_j\, \mathcal{N}\Big(\lambda_k^{1/2}\boldsymbol{\mu}_j,\ \lambda_k\boldsymbol{\Sigma}_j + (1-\lambda_k)\mathbf{I}_d\Big), \qquad \lambda_k := e^{-(k+1)t_0}. \tag{10}$$

Each component of the outer sum is a Gaussian mixture, with the kth term representing the effect of training samples that have gone through $k+1$ rounds of truncated sampling. As k increases the means are contracted toward 0 and the covariances pulled toward $\mathbf{I}_d$. Two consequences follow. First, modes disappear progressively over retraining rounds rather than all at once. Indeed, for a bimodal mixture $p_{\text{data}} = \frac{1}{2}\mathcal{N}(-a, s^2) + \frac{1}{2}\mathcal{N}(a, s^2)$ on $\mathbb{R}$ with $a > s$, kth block is bimodal if and only if $(k+1)t_0 < \log(1 + a^2 - s^2)$, so only the most recent blocks still resolve both modes. Second, $p^\star_\infty$ is not Gaussian even when p_{data} is: for $p_{\text{data}} = \mathcal{N}(0, s^2)$ with $s^2 \neq 1$, (10) is a scale mixture of Gaussians with strictly positive excess kurtosis.

3.1 Spectral decomposition of the collapse distribution

While Theorem 3.1 gives a closed-form description of the collapse distribution, it does not directly identify how different features of the data distribution are affected by recursive training. In particular, the representation (9) does not distinguish between low-order structure (e.g. mean, covariance) and finer non-Gaussian features (e.g. tails, oscillations, multimodality). We now refine this description by studying collapse mode by mode. Specifically, we analyze $p^\star_\infty$ in the Hermite polynomial basis, the spectral coordinates naturally associated with the Ornstein-Uhlenbeck dynamics. The functional analytic details for this section are collected in Appendix B.

Let $\gamma = \mathcal{N}(0, \mathbf{I}_d)$, and let $L^2(\gamma)$ be the space of square-integrable functions with respect to γ. The inner product $\langle\cdot,\cdot\rangle_{L^2(\gamma)}$ on this space is defined as $\langle f, g\rangle_{L^2(\gamma)} = \int f(\mathbf{x})g(\mathbf{x})\gamma(\mathrm{d}\mathbf{x})$.

Decomposition of $\mathcal{U}_t$ in Hermite Polynomial Basis For any multi-index $\mathbf{n} = (n_1, \ldots, n_d) \in \mathbb{N}^d$, the multivariate normalized Hermite polynomial [12] $H_{\mathbf{n}}$ is defined as $H_{\mathbf{n}}(\mathbf{x}) := \frac{1}{\sqrt{\mathbf{n}!}} \prod_{j=1}^d H_{n_j}(x_j)$, where H_k denotes the k-th univariate Hermite polynomial (defined in (26)). The degree of $H_{\mathbf{n}}$ is

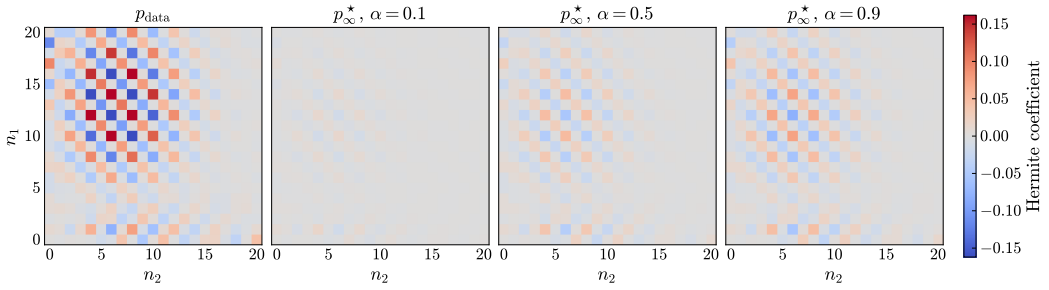


Figure 1: **Hermite-mode attenuation of $p^\star_\infty$ for varying α.** Each heatmap displays the coefficient $\langle f, H_{n_1,n_2}\rangle_{L^2(\gamma)}$ of the density ratio $f = \mathrm{d}p/\mathrm{d}\gamma$ in the two-dimensional Hermite basis, with n_1 and n_2 denoting the polynomial degrees along the two coordinates. *Left*: decomposition of p_{data}, chosen as an oscillatory distribution with substantial energy in medium- and high-degree Hermite modes. *Remaining panels*: similar spectral decomposition of the fixed-point distribution $p^\star_\infty$ for $\alpha \in \{0.1, 0.5, 0.9\}$. As predicted by Proposition 3.4, the attenuation is stronger for smaller α. Code is available at GitHub.

denoted by $|\mathbf{n}| = n_1 + \cdots + n_d$. The family $\{H_\mathbf{n}\}_{\mathbf{n}\in\mathbb{N}^d}$ forms an orthonormal basis of $L^2(\gamma)$. That is, any $g \in L^2(\gamma)$ can be decomposed as [12, 41]:

$$g = \sum_{\mathbf{n}} \langle g, H_\mathbf{n}\rangle_{L^2(\gamma)} H_\mathbf{n},$$

Now, for any probability measure $\mu \ll \gamma$, let $f = \mathrm{d}\mu/\mathrm{d}\gamma$ be its Radon-Nikodym derivative (density ratio w.r.t. γ), assumed to be in $L^2(\gamma)$. Then, writing $f = \sum_\mathbf{n} a_\mathbf{n} H_\mathbf{n}$, we have (Proposition B.4),

$$\frac{\mathrm{d}(\mathcal{U}_t\mu)}{\mathrm{d}\gamma} = \sum_{\mathbf{n}\in\mathbb{N}^d} e^{-|\mathbf{n}|t/2} a_\mathbf{n} H_\mathbf{n}. \tag{11}$$

That is, OU smoothing attenuates higher-degree Hermite modes of the density ratio more strongly than lower-degree ones. Applying the decomposition (11) to the representation (9) of $p^\star_\infty$ requires that (i) $p_{\text{data}} \ll \gamma$, which is guaranteed by A1–A3, and (ii) $\frac{\mathrm{d}p_{\text{data}}}{\mathrm{d}\gamma} \in L^2(\gamma)$, which we assume.

(A4) The data distribution p_{data} is such that $f_{\text{data}} := \frac{\mathrm{d}p_{\text{data}}}{\mathrm{d}\gamma} \in L^2(\gamma)$.

Assumption A4 fails when p_{data} has heavy tails relative to γ, and when it has no density with respect to γ (e.g. empirical measures or data supported on low-dimensional sets). For image datasets, pixel values are bounded, so p_{data} is compactly supported and $\mathrm{d}\mathcal{U}_t p_{\text{data}}/\mathrm{d}\gamma \in L^2(\gamma)$ for every $t > 0$. Under A4, $p^\star_\infty$ is also absolutely continuous with respect to γ, with density denoted by $f^\star_\infty := \frac{\mathrm{d}p^\star_\infty}{\mathrm{d}\gamma}$. The following result shows that the collapse dynamics act diagonally in the Hermite basis, with an explicit mode-wise attenuation factor (proof in Appendix D.4).

Proposition 3.4 (Spectral representation of the collapse distribution)**.** *Under Assumptions A1, A2, A3 and A4, the density $f^\star_\infty = \frac{\mathrm{d}p^\star_\infty}{\mathrm{d}\gamma}$ of the error-free collapse distribution $p^\star_\infty$ admits the expansion*

$$f^\star_\infty = \sum_{\mathbf{n}\in\mathbb{N}^d} m_\mathbf{n}(\alpha, t_0)\left\langle f_{\text{data}}, H_\mathbf{n}\right\rangle_{L^2(\gamma)} H_\mathbf{n}, \qquad m_\mathbf{n}(\alpha, t_0) = \frac{\alpha\, e^{-|\mathbf{n}|t_0/2}}{1-(1-\alpha)e^{-|\mathbf{n}|t_0/2}}. \tag{12}$$

Equivalently, for every $\mathbf{n} \in \mathbb{N}^d$, we have $\left\langle f^\star_\infty, H_\mathbf{n}\right\rangle_{L^2(\gamma)} = m_\mathbf{n}(\alpha, t_0)\left\langle f_{\text{data}}, H_\mathbf{n}\right\rangle_{L^2(\gamma)}$.

Equation (12) shows that the collapse distribution is obtained from p_{data} by a mode-wise attenuation in the Hermite basis relative to the Gaussian reference measure γ. The attenuation factor $m_\mathbf{n}$ is monotone decreasing in the total degree $|\mathbf{n}|$, meaning that high-order polynomials, which encode increasingly fine non-Gaussian features such as tails, oscillations, and multimodality, are suppressed more strongly than low-order ones, which encode Gaussian-like structure of a distribution relative to γ. We stress that this is a low-pass filter in the Hermite basis of $L^2(\gamma)$, not in spatial frequency:

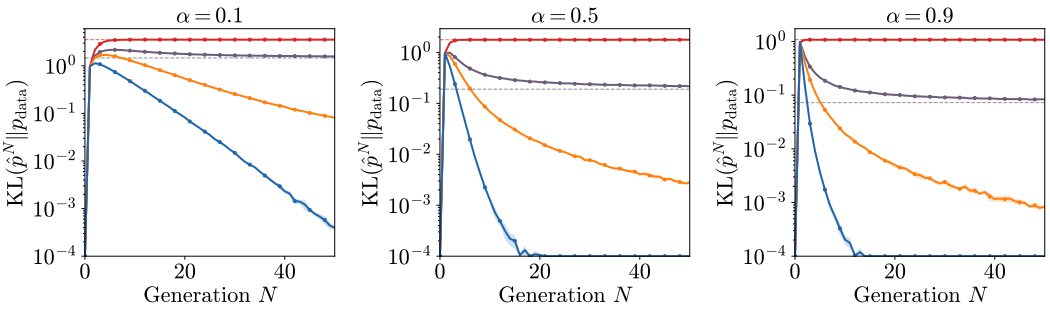


Figure 2: **Annealed truncation schedules eliminate recursive compounding (Theorem 4.1).** Error-free recursion on a 2D Gaussian mixture, $\mathrm{KL}(\hat{p}^N \,\|\, p_{\mathrm{data}})$ vs. generation N for different fresh-data proportions $\alpha \in \{0.1, 0.5, 0.9\}$. Each plot compares four schedules: fixed at $t_0 = 0.5$ (red), decreasing and converging to $t_\infty = 0.2$ (purple), and annealed converging to 0 given by $t_0/(1+i)^\beta$, with $\beta = 1$ (orange) and $\beta = 2$ (blue). The annealed schedules drive KL toward 0 at rates increasing with β, while the fixed and positive-limit schedules plateau at $\mathrm{KL}(p^\star_\infty, p_{\mathrm{data}})$ and $\mathrm{KL}(p^{(t_\infty)}_\infty, p_{\mathrm{data}})$, respectively. Experiment details in Appendix F.

attenuating Hermite modes pulls the distribution toward γ, whose samples are white noise, so in pixel space truncation *adds* high-frequency power (Figure 5, right). This observation matches previous results in which model collapse has, for example, been associated with a degradation of tails [59, 1, 58]. Moreover, the larger the proportion of fresh data injected in each generation of retraining, the weaker the attenuation of each modes of p_{data} affected by the recursion, consistent with similar observations in [11, 32, 45, 28]. Proposition 3.4 thus completes Theorem 3.1 by quantifying the mode-wise effect of *truncation* on p_{data}. Figure 1 illustrates the attenuation of high-degree Hermite modes of $p^\star_\infty$ on a two-dimensional distribution for varying α, and how this evolves with the proportion of fresh data.

Remark 3.5. Assumption A4 is used only to expand f_{data} in the Hermite basis of $L^2(\gamma)$ and is not required for the attenuation mechanism itself. For each $\mathbf{n} \in \mathbb{N}^d$, the corresponding Hermite coefficient is a moment $\langle f_{\mathrm{data}}, H_{\mathbf{n}} \rangle_{L^2(\gamma)} = \mathbb{E}_{p_{\mathrm{data}}}[H_{\mathbf{n}}(\mathbf{X})]$ which exists whenever $\mathbb{E}_{p_{\mathrm{data}}}[\|\mathbf{X}\|^{|\mathbf{n}|}] < \infty$. Under A1–A3, for every such $\mathbf{n}$,

$$\mathbb{E}_{p^\star_\infty}\big[H_{\mathbf{n}}(\mathbf{X})\big] \;=\; m_{\mathbf{n}}(\alpha, t_0)\, \mathbb{E}_{p_{\mathrm{data}}}\big[H_{\mathbf{n}}(\mathbf{X})\big]. \tag{13}$$

(Appendix D.5). Heavy tails thus shrink the set of finite Hermite moments (for which (13) is defined), with the same attenuation factors as in Proposition 3.4. Assumption A4 is only required for combining the coefficients to define $f^\star_\infty \in L^2(\gamma)$.

4 Mitigating Model Collapse with Annealed Truncation Schedules

Prior studies of model collapse for recursively trained general generative models identified the fresh-data proportion α as a key parameter to stabilize or attenuate recursive degradation [11, 32, 45, 28]. However, for diffusion models, Theorem 3.1 and Proposition 3.4 show that for any fresh-data proportion $\alpha < 1$, there is a smoothing bias created by the fixed truncation time $t_0 > 0$, which implies $\mathcal{W}_2(p^\star_\infty,\ p_{\mathrm{data}}) > 0$. Figure 5 (Appendix F) illustrates the effect of t_0 on collapse when varying α on CIFAR-10 [47], showing that while a higher fresh-data injection rate reduces the severity of collapse (the limit $p^\star_\infty$ moves closer to $\mathcal{U}_{t_0} p_{\mathrm{data}}$), it does not prevent it altogether. This suggests a complementary intervention on truncation: modifying the sampling operator across generations by progressively decreasing the truncation times. The resulting annealed schedule targets the specific source of bias identified above, namely the residual OU smoothing due to early stopping.

Generation-dependent truncation We consider a sequence of truncation times $(t_0^{(N)})_{N \geq 0}$, so that, in the error-free regime, the recursion becomes

$$\hat{p}^{N+1} = \mathcal{U}_{t_0^{(N)}}\big(\alpha\, p_{\mathrm{data}} + (1-\alpha)\hat{p}^N\big), \qquad N \geq 0. \tag{14}$$

The following result (proof in Appendix D.6) gives an exact representation of the iterates $\hat{p}^N$ obtained by the adaptive recursion (14), characterizes their limit when the schedule converges and shows that p_{data} can be exactly recovered.

Theorem 4.1 (Asymptotic effect of generation-dependent truncation)**.** *Let $(t_0^{(i)})_{i\geq 0}$ be a sequence of positive truncation times, and consider the error-free recursion (14). Then, for every $N \geq 1$,*

$$\hat{p}^N = \alpha \sum_{m=0}^{N-1} (1-\alpha)^m \, \mathcal{U}_{\sigma_{m,N}}(p_{\text{data}}) + (1-\alpha)^N \mathcal{U}_{s_{0,N}}(p_{\text{data}}), \tag{15}$$

where $\sigma_{m,N} := \sum_{\ell=N-1-m}^{N-1} t_0^{(\ell)}$ and $s_{0,N} := \sum_{\ell=0}^{N-1} t_0^{(\ell)}$. Furthermore, if $t_0^{(N)} \to t_\infty \in [0,\infty)$ as $N \to \infty$, then

$$\hat{p}^N \xrightarrow[N\to\infty]{\mathcal{W}_2} p_\infty^{(t_\infty)} := \alpha \sum_{m=0}^{\infty} (1-\alpha)^m \, \mathcal{U}_{(m+1)t_\infty} p_{\text{data}}.$$

In particular, if $t_\infty = 0$, then $p_\infty^{(t_\infty)} = p_{\text{data}}$, and the recursive compounding effect disappears asymptotically.

Theorem 4.1 shows that the long-run effect of a varying truncation schedule is controlled by its *asymptotic level* t_∞: because fresh data are re-injected at every generation, recent truncation times dominate the long-run dynamics. If the schedule converges to a positive limit, collapse persists as in the fixed-truncation regime with $t_0 = t_\infty$, but with a smaller bias than with fixed truncation at $t_0^{(0)}$ if the sequence $(t_0^{(i)})_{i\geq 0}$ is decreasing. By contrast, if $t_0^{(i)} \to 0$, then the error-free recursion converges back to p_{data}; asymptotically, no nontrivial collapse distribution remains. This result holds for any sequence of positive truncation times, and in particular, allows for adaptive truncation chosen online at each retraining round. For simplicity, our experiments use the predetermined β-annealed schedules defined below.

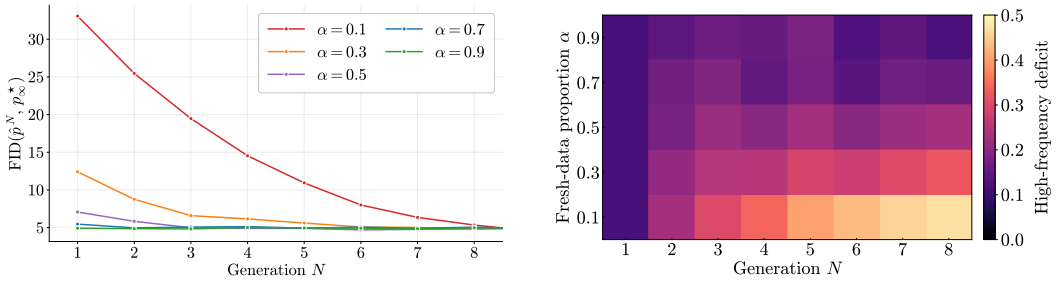
β-annealed schedules For any $\beta > 0$, the schedule $t_0^{(i)}(\beta) = t_0/(1+i)^\beta$ converges to 0, implying $\mathcal{W}_2(\hat{p}^N, p_{\text{data}}) \to 0$ as $N \to \infty$ in the error-free regime. However, different values of β induce different *finite-generation* distortions: larger β drives the truncation time to zero faster, therefore suppressing recursive attenuation more aggressively over a fixed number of generations. Figure 2 compares four schedules over 50 generations at $\alpha \in \{0.1, 0.5, 0.9\}$: fixed at $t_0 = 0.5$, decreasing to $t_\infty = 0.2$, and β-annealed schedules $t_0/(1+i)^\beta$ with $\beta \in \{1, 2\}$. The fixed and shifted schedules plateau at the predicted floors $\text{KL}(p_\infty^\star, p_{\text{data}})$ and $\text{KL}(p_\infty^{(t_\infty)}, p_{\text{data}})$, while the annealed schedules drive KL toward 0 at rates increasing with β, confirming that recursive compounding is eliminated whenever $t_\infty = 0$. Convergence slows with smaller α, as expected from $\kappa = \sqrt{1-\alpha}\, e^{-t_0/2}$.

Comparing schedules at finite N For a suitable non-decreasing map $h : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ (details in Appendix D.7), the representation (15) yields, for every $N \geq 1$,

$$\mathcal{W}_2^2(\hat{p}^N, p_{\text{data}}) \;\leq\; \alpha \sum_{m=0}^{N-1} (1-\alpha)^m \, h(\sigma_{m,N}) \;+\; (1-\alpha)^N \, h(s_{0,N}).$$

Since h is nondecreasing and $\sigma_{m,N}$, $s_{0,N}$ are nondecreasing in every $t_0^{(\ell)}$, a pointwise smaller schedule has a smaller bound at every N, consistent with the ordering in Figure 2.

Self-regularization regime The above finite time bound holds in the error-free regime. With errors, how fast the schedule may decay is limited, since the only data-independent control on the Fisher information of the training mixture, $J(q_{i,t}) \leq d/(1-e^{-t})$ (Proposition E.1(ii)), degenerates as $t \downarrow 0$. The recursive structure (2) provides a *partial* self-regularization mechanism: while the part of the synthetic component of q_i that was injected many generations ago has accumulated Gaussian smoothing $\sum_{j<i} t_0^{(j)}$ due to truncation, fresh data are re-injected at every generation, so the most recently injected components have been smoothed by only the last few truncation times, which vanish along the schedule.



(a) $\mathrm{FID}(\hat{p}^N, p^\star_\infty)$ vs. generation N.

(b) High-frequency energy ratio $\mathrm{HF}(\hat{p}^N)/\mathrm{HF}(p_{\mathrm{data}})$.

Figure 3: **Convergence and spectral drift under recursive training (CIFAR-10).** Recursive training for 8 generations at $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. **(a)** The iterates $\hat{p}^N$ contract toward $p^\star_\infty$ at a rate increasing with α, consistent with the contraction factor $\kappa(\alpha) = \sqrt{1-\alpha}\, e^{-t_0/2}$, and settle at a similar floor consistent with stability around $p^\star_\infty$ (Theorem 5.2). **(b)** High-frequency deficit $1 - \mathrm{HF}(\hat{p}^N)/\mathrm{HF}(p_{\mathrm{data}})$ over recursive generations and fresh-data proportions α. Larger values indicate a stronger loss of high-frequency content relative to the data distribution. The deficit grows fastest when α is small and N increases. This loss is not caused by truncation, which adds high-frequency power in pixel space (Figure 5, right); it reflects smoothing by the learned sampler (the one-step errors δ_i), which the recursion accumulates. Details in Appendix F.

5 Robustness of Limiting Distribution in the Presence of Errors

The idealized analysis in the last two sections isolated the effect of truncation, but in practice, recursive training is also affected by score estimation and discretization errors and by the initialization of the reverse SDE at $\mathcal{N}(0, \mathbf{I}_d)$. In this section we provide robustness results quantifying how perturbations of the learned sampling operator, induced by errors in score estimation and discretization, affect the divergence between $\hat{p}^N$ and the ideal limit $p^\star_\infty$. We return to the setting of a fixed truncation time $t_0 > 0$.

Score Estimation and Discretization Errors Comparing the ideal and reverse SDEs (6) and (7), the score estimation error at generation i is $\mathbf{e}_i(\mathbf{x}, s) = \mathbf{s}_{\theta_i}(\mathbf{x}, s) - \nabla_{\mathbf{x}} \log q_{i,s}(\mathbf{x})$. For a given score network, $\mathbf{e}_i$ depends, through the target q_i, on the mixing factor α, the ambient dimension d and the i-th generation training sample size n_i. The learned reverse SDE (7) is further discretized via a numerical scheme on a partition $\{t_k\}_{k=1}^{K_i}$ of $[t_0, T]$. The i-th generation *one-step sampling error* δ_i captures the combined effect of score estimation, discretization and initialization at $\mathcal{N}(0, \mathbf{I}_d)$ instead of $q_{i,T}$. We define it as the $\mathcal{W}_2$ distance between the output of the imperfect sampler $\hat{\mathcal{S}}_i$ and that of the ideal sampler $\mathcal{U}_{t_0}$:

$$\delta_i \equiv \delta_i(d, n_i, \{t_k\}_{k=1}^{K_i}, T, \alpha) \;=\; \mathcal{W}_2\big(\hat{\mathcal{S}}_i(q_i),\, \mathcal{U}_{t_0}(q_i)\big). \tag{16}$$

Remark 5.1 (Concrete instantiations of δ_i)**.** Several works make the dependence between δ_i and $(d, n_i, \{t_k\}_{k=1}^{K_i}, \alpha)$ explicit via upper bounds. For the Euler–Maruyama discretization, Arsenyan et al. [6] establish optimal $\mathcal{W}_2$ bounds, and for the exponential integrator, Chen et al. [16] provide TV bounds that can be transferred to $\mathcal{W}_2$ under moment assumptions. Higher-order schemes such as midpoint randomization [70], stochastic Runge–Kutta methods [69], and accelerated DDIM-type samplers [49] have been shown to have improved dependence on the step size or on the number of function evaluations. Deterministic samplers based on the probability flow ODE [30] admit analogous $\mathcal{W}_2$ guarantees under log-concavity. Our results are general and applicable to any of these schemes. These bounds also contain an initialization term that decays exponentially in T.

Concentration around $p^\star_\infty$ We now show that the imperfect iterates $\hat{p}^N$ track the ideal collapse distribution $p^\star_\infty$ at a geometric rate, perturbed by the one-step errors δ_i (proof in Appendix D.8).

Theorem 5.2 (Geometric convergence and stability)**.** *Assume* A1–A3. *Then for each* $N \geq 1$,

$$\mathcal{W}_2(\hat{p}^N, p^\star_\infty) \;\leq\; \kappa^N\, \mathcal{W}_2(p_{\mathrm{data}}, p^\star_\infty) \;+\; \sum_{i=0}^{N-1} \kappa^{N-1-i}\, \delta_i, \tag{17}$$

where $\kappa = \sqrt{(1-\alpha)}e^{-t_0/2} < 1$. *In particular:*

(i) if $\sum_{i\geq 0}\delta_i < \infty$, *then* $\mathcal{W}_2(\hat{p}^N, p^\star_\infty) \to 0$;

(ii) if $\sup_i \delta_i < \delta$, *then* $\limsup_{N\to\infty} \mathcal{W}_2(\hat{p}^N, p^\star_\infty) \leq \dfrac{\delta}{1-\kappa}$.

Theorem 5.2 establishes that $\hat{p}^N$ remains within a $\mathcal{W}_2$-ball of radius $\delta/(1-\kappa)$ around $p^\star_\infty$, and Figure 3 illustrates this robustness to error in a real image dataset (CIFAR-10) across varying levels of α. Proposition D.3 (Appendix D.9) complements Theorem 5.2 by providing an explicit characterization of the perturbation $(\hat{p}^N - p^\star_\infty)$.

Spectral Decomposition in the Learned Regime The spectral viewpoint on the error-free collapse distribution $p^\star_\infty$ can be extended to the general setting.

The following result (proof in Appendix D.10) complements Theorem 5.2 with a mode-by-mode description of the error: mode $\mathbf{n}$ contracts with geometric ratio $\kappa_{\mathbf{n}} = (1-\alpha)e^{-|\mathbf{n}|t_0/2}$, which decreases with $|\mathbf{n}|$ (substantially so once $|\mathbf{n}|t_0 \gtrsim 1$). Transient errors in high-order modes are therefore forgotten faster than errors in low-order modes, and persistent errors in mode $\mathbf{n}$ are amplified by at most $1/(1-\kappa_{\mathbf{n}}) \leq 1/\alpha$, a factor that tends to 1 as $|\mathbf{n}| \to \infty$.

Proposition 5.3 (Mode-dependent error propagation)**.** *Under A1–A4, further assume that for all* $i \geq 0$, *the learned distribution in the imperfect regime satisfies* $\mathrm{d}\hat{p}^i/\mathrm{d}\gamma \in L^2(\gamma)$. *Then, for* $N \geq 1$, $\mathrm{d}(\hat{p}^N - p^\star_\infty)/\mathrm{d}\gamma \in L^2(\gamma)$ *and, decomposing mode by mode, for each multi-index* $\mathbf{n} \in \mathbb{N}^d$,

$$\langle \hat{p}^N - p^\star_\infty,\, H_{\mathbf{n}} \rangle \;=\; \kappa_{\mathbf{n}}^N \,\langle p_{\mathrm{data}} - p^\star_\infty,\, H_{\mathbf{n}} \rangle \;+\; \sum_{i=0}^{N-1} \kappa_{\mathbf{n}}^{N-1-i}\, \langle \hat{p}^{i+1} - \mathcal{U}_{t_0}(q_i),\, H_{\mathbf{n}} \rangle, \tag{18}$$

where $\kappa_{\mathbf{n}} := (1-\alpha)\, e^{-|\mathbf{n}|t_0/2}$ *is the mode-dependent contraction rate.*

6 Discussion and Future Work

This work provides an asymptotic characterization of model collapse in recursively trained diffusion models. First, in an idealized regime without discretization and score estimation errors, we show that collapse still occurs because of the truncation of the reverse diffusion near $t = 0$, which is standard in practical sampling. Truncation induces a residual Gaussian smoothing in each generation, which compounds and drives the recursive dynamics to a unique limiting distribution $p^\star_\infty$, a geometric mixture of progressively smoothed copies of p_{data}. The spectral representation of $p^\star_\infty$ further shows that this limiting distribution corresponds to a mode-wise attenuation of the data distribution in the Hermite basis, making precise the sense in which recursive diffusion training acts as a low-pass filter.

Our results also highlight that in diffusion models, model collapse is not solely driven by the proportion of fresh data α. Increasing α attenuates recursive degradation, but it does not eliminate the smoothing effect induced by a fixed truncation time $t_0 > 0$. This observation motivates a diffusion-specific route to mitigating collapse, complementary to fresh data injection: by decreasing the truncation time across recursive generations (annealed truncation schedules), one can eliminate the asymptotic compounding effect in the error-free regime.

Future directions and limitations Several directions remain open. First, our analysis treats score estimation and discretization errors through abstract one-step perturbations. A natural next step is to combine the present fixed-point analysis with finite-sample score-learning bounds, thereby obtaining explicit end-to-end guarantees in terms of model class, sample size, dimension, and numerical solver. Second, the annealed truncation schedules studied here are prescribed, rather than learned or adapted to the current generation. Designing practical adaptive rules that balance numerical stability against recursive bias is an important direction. Third, the spectral characterization is developed for the variance-preserving OU diffusion and the Hermite basis associated with its invariant Gaussian measure. Extending this perspective to other diffusion parameterizations, alternative noising processes, or non-Gaussian reference measures may reveal more general spectral mechanisms underlying collapse.

Acknowledgements

NBK is supported by a G-Research Trinity College Studentship, and RET is supported by the EPSRC Probabilistic AI Hub (EP/Y028783/1).

References

[1] Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard Baraniuk. Self-consuming generative models go MAD. In *International Conference on Learning Representations (ICLR)*, 2024.

[2] Sina Alemohammad, Ahmed Imtiaz Humayun, Shruti Agarwal, John Collomosse, and Richard Baraniuk. Self-improving diffusion models with synthetic data. arXiv:2408.16333, 2024.

[3] Brian D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

[4] Donald G. Aronson. Bounds for the fundamental solution of a parabolic equation. *Bulletin of the American Mathematical Society*, 73(6):890–896, 1967.

[5] Donald G. Aronson. Non-negative solutions of linear parabolic equations. *Annali della Scuola Normale Superiore di Pisa, Classe di Scienze (3)*, 22(4):607–694, 1968.

[6] Vahan Arsenyan, Elen Vardanyan, and Arnak S. Dalalyan. Assessing the quality of denoising diffusion models in Wasserstein distance: Noisy score and optimal bounds. In *Advances in Neural Information Processing Systems*, volume 38, pages 19548–19591, 2025.

[7] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*, volume 348 of *Grundlehren der mathematischen Wissenschaften*. Springer, Cham, 2014.

[8] Stefan Banach. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta Mathematicae*, 3(1):133–181, 1922.

[9] Daniel Barzilai and Ohad Shamir. When models don't collapse: On the consistency of iterative MLE. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[10] Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *International Conference on Learning Representations (ICLR)*, 2024.

[11] Quentin Bertrand, Avishek Joey Bose, Alexandre Duplessis, Marco Jiralerspong, and Gauthier Gidel. On the stability of iterative retraining of generative models on their own data. In *International Conference on Learning Representations (ICLR)*, 2024.

[12] Vladimir I. Bogachev. *Gaussian Measures*, volume 62 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1998.

[13] Nicholas Carlini, Matthew Jagielski, Christopher A. Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical. In *IEEE Symposium on Security and Privacy (SP)*, pages 407–425, 2024.

[14] Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[15] Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

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

[17] Giovanni Conforti, Alain Durmus, and Marta Gentiloni Silveri. KL convergence guarantees for score diffusion models under minimal data assumptions. *SIAM Journal on Mathematics of Data Science*, 7(1):86–109, 2025.

[18] Hugo Cui, Cengiz Pehlevan, and Yue M. Lu. A solvable model of learning generative diffusion: Theory and insights. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[19] Giuseppe Da Prato and Jerzy Zabczyk. *Stochastic Equations in Infinite Dimensions*, volume 152 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, second edition, 2014.

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

[21] Amir Dembo, Thomas M. Cover, and Joy A. Thomas. Information theoretic inequalities. *IEEE Transactions on Information Theory*, 37(6):1501–1518, 1991.

[22] Apratim Dey and David Donoho. Universality of the $\pi^2/6$ pathway in avoiding model collapse. arXiv preprint arXiv:2410.22812, 2024.

[23] Elvis Dohmatob, Yunzhen Feng, and Julia Kempe. Model collapse demystified: The case of regression. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[24] Elvis Dohmatob, Yunzhen Feng, Arjun Subramonian, and Julia Kempe. Strong model collapse. In *International Conference on Learning Representations (ICLR)*, 2025.

[25] Elvis Dohmatob, Yunzhen Feng, Pu Yang, François Charton, and Julia Kempe. A tale of tails: Model collapse as a change of scaling laws. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

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

[27] Stewart N. Ethier and Thomas G. Kurtz. *Markov Processes: Characterization and Convergence*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, 1986.

[28] Shi Fu, Sen Zhang, Yingjie Wang, Xinmei Tian, and Dacheng Tao. Towards theoretical understandings of self-consuming generative models. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[29] Xuefeng Gao, Hoang M. Nguyen, and Lingjiong Zhu. Wasserstein convergence guarantees for a general class of score-based generative models. *Journal of Machine Learning Research*, 26(43):1–54, 2025.

[30] Xuefeng Gao and Lingjiong Zhu. Convergence analysis for general probability flow ODEs of diffusion models in Wasserstein distances. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, 2025.

[31] Anvit Garg, Sohom Bhattacharya, and Pragya Sur. Preventing model collapse under over-parametrization: Optimal mixing ratios for interpolation learning and ridge regression. In *International Conference on Learning Representations (ICLR)*, 2026.

[32] Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data. arXiv:2404.01413, 2024.

[33] Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. Reinforced self-training (ReST) for language modeling. arXiv:2308.08998, 2023.

[34] Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via Min-SNR weighting strategy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7441–7451, 2023.

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

[36] 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 *Advances in Neural Information Processing Systems*, volume 30, 2017.

[37] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[38] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

[39] Aapo Hyvärinen. Some extensions of score matching. *Computational Statistics & Data Analysis*, 51(5):2499–2512, 2007.

[40] Ayush Jain, Andrea Montanari, and Eren Sasoglu. Scaling laws for learning with real and surrogate data. In *Advances in Neural Information Processing Systems*, volume 37, pages 110246–110289, 2024.

[41] Svante Janson. *Gaussian Hilbert Spaces*, volume 129 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1997.

[42] Oliver Johnson and Andrew Barron. Fisher information inequalities and the central limit theorem. *Probability Theory and Related Fields*, 129(3):391–409, 2004.

[43] Ioannis Karatzas and Steven E. Shreve. *Brownian Motion and Stochastic Calculus*, volume 113 of *Graduate Texts in Mathematics*. Springer, New York, second edition, 1991.

[44] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, pages 26565–26577, 2022.

[45] Naïl B. Khelifa, Richard E. Turner, and Ramji Venkataramanan. Quantifying error propagation and model collapse in diffusion models. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

[46] Dongjun Kim, Seungjae Shin, Kyungwoo Song, Wanmo Kang, and Il-Chul Moon. Soft truncation: A universal training technique of score-based diffusion model for high precision score estimation. In *In International Conference on Machine Learning*, 2022.

[47] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[48] Rebecca M. Lewis, Oliver Y. Feng, Henry W. J. Reeve, Min Xu, and Richard J. Samworth. Learning the score under shape constraints. arXiv:2512.14624, 2025.

[49] Gen Li, Yu Huang, Timofey Efimov, Yuting Wei, Yuejie Chi, and Yuxin Chen. Accelerating convergence of score-based diffusion models, provably. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 27942–27954. PMLR, 2024.

[50] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

[51] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 22(4):730–751, 2025.

[52] Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Universitext. Springer, Berlin, sixth edition, 2003.

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

[54] Michael Reed and Barry Simon. *Methods of Modern Mathematical Physics. IV: Analysis of Operators*. Academic Press, New York, 1978.

[55] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351, pages 234–241, 2015.

[56] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, volume 87 of *Progress in Nonlinear Differential Equations and Their Applications*. Birkhäuser, Cham, 2015.

[57] Lianghe Shi, Meng Wu, Huijie Zhang, Zekai Zhang, Molei Tao, and Qing Qu. A closer look at model collapse: From a generalization-to-memorization perspective. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[58] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. arXiv:2305.17493, 2023.

[59] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

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

[61] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[62] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[63] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

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

[65] A. J. Stam. Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Information and Control*, 2(2):101–112, 1959.

[66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[67] Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, 2009.

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

[69] Yuchen Wu, Yuxin Chen, and Yuting Wei. Stochastic Runge–Kutta methods: Provable acceleration of diffusion models. arXiv:2410.04760, 2024.

[70] Yifeng Yu and Lu Yu. Advancing Wasserstein convergence analysis of score-based models: Insights from discretization and second-order acceleration. In *Advances in Neural Information Processing Systems*, 2025.

[71] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pages 15476–15488, 2022.

[72] Kaihong Zhang, Heqi Yin, Feng Liang, and Jingbo Liu. Minimax optimality of score-based diffusion models: Beyond the density lower bound assumptions. In *In International Conference on Machine Learning*, 2024.

[73] Pengze Zhang, Hubery Yin, Chen Li, and Xiaohua Xie. Tackling the singularities at the endpoints of time intervals in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6945–6954, 2024.

[74] Huminhao Zhu, Fangyikang Wang, Tianyu Ding, Qing Qu, and Zhihui Zhu. Analyzing and mitigating model collapse in reflow methods. In *Conference on Parsimony and Learning (CPAL)*, 2026.

A Notation

In the rest of this appendix, we use the following notation:

- We denote by $\gamma = \mathcal{N}(0, \mathbf{I}_d)$ the standard centered and unit variance Gaussian measure on $\mathbb{R}^d$. Explicitly, for any $\mathbf{x} \in \mathbb{R}^d$,

$$\gamma(\mathbf{x}) = (2\pi)^{-d/2} \exp\left(-\tfrac{1}{2}\|\mathbf{x}\|^2\right).$$

- The space $L^2(\gamma)$ is defined as,

$$L^2(\gamma) := \left\{ f : \mathbb{R}^d \to \mathbb{R} \;\middle|\; \int_{\mathbb{R}^d} \|f(\mathbf{x})\|_2^2 \, \gamma(\mathrm{d}\mathbf{x}) < \infty \right\},$$

  where functions are identified up to equality γ-almost everywhere.

- We equip $L^2(\gamma)$ with an inner product defined as,

$$\langle f, g \rangle_{L^2(\gamma)} := \int_{\mathbb{R}^d} f(\mathbf{x})\, g(\mathbf{x})\, \gamma(\mathrm{d}\mathbf{x}), \tag{19}$$

  making it a Hilbert space $\left(L^2(\gamma), \langle \cdot, \cdot \rangle_{L^2(\gamma)}\right)$. The $L^2(\gamma)$ inner product induces the following $L^2(\gamma)$-norm,

$$\|f\|_{L^2(\gamma)} = \left(\int_{\mathbb{R}^d} \|f(\mathbf{x})\|_2^2 \, \gamma(\mathrm{d}\mathbf{x}) \right)^{1/2}. \tag{20}$$

- For any sequence of distributions $(\mu_t)_{t \in \mathbb{R}_+} \in (\mathcal{P}_2(\mathbb{R}^d))^{\mathbb{R}_+}$ and any fixed $\mu^\star \in \mathcal{P}_2(\mathbb{R}^d)$, we say that $(\mu_t)_t$ converges to $\mu^\star$ in $\mathcal{W}_2$ and denote $\mu_t \xrightarrow[t\to\infty]{\mathcal{W}_2} \mu^\star$ if,

$$\mathcal{W}_2(\mu_t, \mu^\star) \xrightarrow[t\to\infty]{} 0.$$

B Extended Background and Preliminary Results

In this appendix, we provide a detailed background of the mathematical tools and the classical results underlying our proofs.

B.1 Diffusion, Ornstein–Uhlenbeck Process and Semigroup

Recall that, for each generation i, the forward process $(\mathbf{X}_t^i)_{t\in[0,T]}$ is defined as a solution of the following variance-preserving Ornstein-Uhlenbeck diffusion initialized at the mixture distribution q_i:

$$\mathrm{d}\mathbf{X}_t^i = -\tfrac{1}{2}\mathbf{X}_t^i \, \mathrm{d}t + \mathrm{d}\mathbf{B}_t, \qquad \mathbf{X}_0^i \sim q_i. \tag{21}$$

Then, integrating (21) yields

$$\mathbf{X}_t^i = e^{-t/2}\mathbf{X}_0^i + \sqrt{1 - e^{-t}}\mathbf{Z}, \qquad \mathbf{Z} \sim \mathcal{N}(0, \mathbf{I}_d).$$

The corresponding transition kernel can thus be written as,

$$\mathcal{K}_t(\mathbf{x}, \mathrm{d}\mathbf{y}) = \mathcal{N}\Big(e^{-t/2}\mathbf{x}, (1 - e^{-t})\mathbf{I}_d\Big)(\mathrm{d}\mathbf{y}). \tag{22}$$

Markov semigroup. The transition kernel $\mathcal{K}_t$ defines a Markov semigroup $(P_t)_{t\geq 0}$ acting on test functions $\varphi : \mathbb{R}^d \to \mathbb{R}$ by

$$(P_t\varphi)(\mathbf{x}) := \int_{\mathbb{R}^d} \varphi(\mathbf{y}) \, \mathcal{K}_t(\mathbf{x}, \mathrm{d}\mathbf{y}) = \mathbb{E}\big[\varphi(\mathbf{X}_t^i) \mid \mathbf{X}_0^i = \mathbf{x}\big]. \tag{23}$$

The family $(P_t)_{t\geq 0}$ satisfies the semigroup property $P_{t+s} = P_t P_s$ and $P_0 = \mathrm{Id}$ [27].

Sampling operator on probability measures. The adjoint of P_t acts on test functions. On the other hand, the OU diffusion (21) defines sampling operators $(\mathcal{U}_t)_{t\in[t_0,T]}$, acting on probability measures $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and defined by,

$$\mathcal{U}_t\mu(A) := \int_{\mathbb{R}^d} \mathcal{K}_t(\mathbf{x}, A)\, \mu(\mathrm{d}\mathbf{x}), \qquad A \subset \mathbb{R}^d. \tag{24}$$

Equivalently,

$$\mathcal{U}_t\mu = \mathrm{Law}\big(e^{-t/2}\mathbf{X}_0 + \sqrt{1-e^{-t}}\,\mathbf{Z}\big), \qquad \mathbf{X}_0 \sim \mu,\ \mathbf{Z} \sim \mathcal{N}(0, \mathbf{I}_d),\ \mathbf{X}_0 \perp\!\!\!\perp \mathbf{Z}. \tag{25}$$

The operator $\mathcal{U}_t$ is the sampling operator used throughout the paper. Given its central role in this work, we collect its properties in the following proposition.

Proposition B.1 (Basic properties of the OU sampling operator)**.** *The Ornstein–Uhlenbeck sampling operator $(\mathcal{U}_t)_{t\geq 0}$ as defined in (24), satisfies the following properties.*

(i) ***Linearity.*** *For any two probability measures μ, ν and scalars $a, b \in \mathbb{R}$,*

$$\mathcal{U}_t(a\mu + b\nu) = a\,\mathcal{U}_t\mu + b\,\mathcal{U}_t\nu.$$

(ii) ***Semigroup property.*** *For all $s, t \geq 0$,*

$$\mathcal{U}_{t+s} = \mathcal{U}_t\mathcal{U}_s, \qquad \mathcal{U}_0 = \mathbf{I}_d.$$

(iii) ***Weak duality identity.*** *For any bounded measurable test function φ and every finite measure μ,*

$$\int_{\mathbb{R}^d} \varphi(\mathbf{y})\,(\mathcal{U}_t\mu)(\mathrm{d}\mathbf{y}) = \int_{\mathbb{R}^d} (P_t\varphi)(\mathbf{x})\,\mu(\mathrm{d}\mathbf{x}),$$

(iv) ***Moment preservation.*** *Let $k \geq 1$, $\mathbf{Z} \sim \mathcal{N}(0, \mathbf{I}_d)$ and $\mu \in \mathcal{P}(\mathbb{R}^d)$ with $M_k(\mu) := \int_{\mathbb{R}^d} \|\mathbf{x}\|^k\, \mu(\mathrm{d}\mathbf{x}) < \infty$. Then, for every $t \geq 0$,*

$$\Big(\int_{\mathbb{R}^d} \|\mathbf{x}\|^k\, \mathcal{U}_t\mu(\mathrm{d}\mathbf{x})\Big)^{1/k} \;\leq\; M_k(\mu)^{1/k} + \big(\mathbb{E}\|\mathbf{Z}\|^k\big)^{1/k},$$

so $\sup_{t\geq 0} M_k(\mathcal{U}_t\mu) < \infty$. In particular $\mathcal{U}_t$ maps $\mathcal{P}_2(\mathbb{R}^d)$ into itself.

(v) ***Uniqueness and Invariance of the Gaussian measure.*** *Let $\gamma = \mathcal{N}(0, \mathbf{I}_d)$. Then γ is the unique measure satisfying $\mathcal{U}_t\gamma = \gamma$.*

(vi) ***Action in $L^2(\gamma)$.*** *For any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, writing $f = \mathrm{d}\mu/\mathrm{d}\gamma$,*

$$\frac{\mathrm{d}(\mathcal{U}_t\mu)}{\mathrm{d}\gamma} = P_t f.$$

(vii) ***Continuity in time.*** *For a fixed $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and a fixed $s \in [0, t]$, one has that,*

$$\mathcal{U}_t\mu \xrightarrow[t\to s]{\mathcal{W}_2} \mathcal{U}_s\mu.$$

In particular,

$$\mathcal{U}_t\mu \xrightarrow[t\to 0]{\mathcal{W}_2} \mu.$$

Proof. (i) Linearity. Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ be two probability distributions on $\mathbb{R}^d$ and $a, b \in \mathbb{R}$ two scalars. Let $A \subseteq \mathbb{R}^d$ be any measurable set in $\mathbb{R}^d$. Then using (24),

$$\mathcal{U}_t(a\mu + b\nu)(A) = \int_{\mathbb{R}^d} \mathcal{K}_t(\mathbf{x}, A)\,(a\mu + b\nu)(\mathrm{d}\mathbf{x}).$$

Looking at (22), and by linearity of the integral in the measure,

$$\mathcal{U}_t(a\mu + b\nu)(A) = a\int_{\mathbb{R}^d} \mathcal{K}_t(\mathbf{x}, A)\,\mu(\mathrm{d}\mathbf{x}) + b\int_{\mathbb{R}^d} \mathcal{K}_t(\mathbf{x}, A)\,\nu(\mathrm{d}\mathbf{x}) = a\,\mathcal{U}_t\mu(A) + b\,\mathcal{U}_t\nu(A).$$

(ii) Semigroup property. Let $\mu \in \mathcal{P}(\mathbb{R}^d)$ and consider $\mathbf{X} \sim \mu$. In addition, fix two independently distributed Gaussian random variables $Z_1, Z_2 \sim \mathcal{N}(0, \mathbf{I}_d)$ and fix $s, t \in [0, T]$. Applying successively $\mathcal{U}_s$ and $\mathcal{U}_t$,

$$\begin{aligned}\mathcal{U}_t\mathcal{U}_s\mu &= e^{-t/2}\left(e^{-s/2}\mathbf{X} + \sqrt{1-e^{-s}}\,\mathbf{Z}_1\right) + \sqrt{1-e^{-t}}\,\mathbf{Z}_2 \\ &= e^{-(t+s)/2}\mathbf{X} + \left(e^{-t/2}\sqrt{1-e^{-s}}\,\mathbf{Z}_1 + \sqrt{1-e^{-t}}\,\mathbf{Z}_2\right).\end{aligned}$$

Since Z_1, Z_2 are independent standard Gaussians, the variable $\left(e^{-t/2}\sqrt{1-e^{-s}}\,\mathbf{Z}_1 + \sqrt{1-e^{-t}}\,\mathbf{Z}_2\right)$ is a centered Gaussian with covariance

$$e^{-t}(1-e^{-s})\mathbf{I}_d + (1-e^{-t})\mathbf{I}_d = (1-e^{-(t+s)})\mathbf{I}_d.$$

Denoting $\mathbf{Z} \sim \mathcal{N}(0, \mathbf{I}_d)$, the resulting law is

$$\mathcal{U}_t\mathcal{U}_s\mu = \mathrm{Law}\left(e^{-(t+s)/2}\mathbf{X} + \sqrt{1-e^{-(t+s)}}\,\mathbf{Z}\right) = \mathcal{U}_{t+s}\mu.$$

This being true for any μ, one has that $\mathcal{U}_t\mathcal{U}_s = \mathcal{U}_{t+s}$. The identity $\mathcal{U}_0 = \mathbf{I}_d$ follows directly from the definition.

(iii) Weak duality identity. By the definitions of $\mathcal{U}_t$ in (24) and P_t in (23),

$$\int_{\mathbb{R}^d} \varphi(\mathbf{y})\,(\mathcal{U}_t\mu)(\mathrm{d}\mathbf{y}) = \int_{\mathbb{R}^d} \underbrace{\left(\int_{\mathbb{R}^d} \varphi(\mathbf{y})\,\mathcal{K}_t(\mathbf{x}, \mathrm{d}\mathbf{y})\right)}_{:=P_t\varphi(\mathbf{x})} \mu(\mathrm{d}\mathbf{x}) = \int_{\mathbb{R}^d} (P_t\varphi)(\mathbf{x})\,\mu(\mathrm{d}\mathbf{x}).$$

Thus $\mathcal{U}_t = P_t^*$ in the usual measure-function duality.

(iv) Moment preservation. Fix μ with $M_k(\mu) < \infty$ and let $\mathbf{X} \sim \mu \perp\!\!\!\perp \mathbf{Z} \sim \mathcal{N}(0, \mathbf{I}_d)$. By (25), $\mathbf{Y} = e^{-t/2}\mathbf{X} + \sqrt{1-e^{-t}}\,\mathbf{Z} \sim \mathcal{U}_t\mu$. Since $\|\mathbf{Y}\| \le e^{-t/2}\|\mathbf{X}\| + \sqrt{1-e^{-t}}\,\|\mathbf{Z}\|$, Minkowski's inequality in L^k gives

$$\left(\mathbb{E}\|\mathbf{Y}\|^k\right)^{1/k} \le e^{-t/2}\left(\mathbb{E}\|\mathbf{X}\|^k\right)^{1/k} + \sqrt{1-e^{-t}}\,\left(\mathbb{E}\|\mathbf{Z}\|^k\right)^{1/k},$$

which is finite because Gaussian vectors have moments of all orders. The second inequality uses $e^{-t/2}, \sqrt{1-e^{-t}} \le 1$. For $k = 2$, expanding the square and using that $\mathbf{Z}$ is independent of $\mathbf{X}$ with $\mathbb{E}\mathbf{Z} = 0$ and $\mathbb{E}\|\mathbf{Z}\|^2 = d$,

$$\mathbb{E}\|\mathbf{Y}\|^2 = e^{-t}\,\mathbb{E}\|\mathbf{X}\|^2 + d\,(1-e^{-t}).$$

(v) Gaussian invariance. Let $\mathbf{X} \sim \gamma = \mathcal{N}(0, \mathbf{I}_d)$ and $\mathbf{Z} \sim \mathcal{N}(0, \mathbf{I}_d)$ be independent Gaussians, then $\mathcal{U}_t\gamma = \mathrm{Law}(e^{-t/2}\mathbf{X} + \sqrt{1-e^{-t}}\,\mathbf{Z})$ is a sum of centered independent Gaussians, with covariance $e^{-t}\mathbf{I}_d + (1-e^{-t})\mathbf{I}_d = \mathbf{I}_d$. Hence, $\mathcal{U}_t\gamma = \gamma$. Let p_0 be another distribution such that $\mathcal{U}_t p_0 = p_0$ for all t. Then $p_0 = \mathcal{U}_{nt}p_0$ and, by Proposition B.5, $\mathcal{W}_2(p_0, \gamma) = \mathcal{W}_2(\mathcal{U}_{nt}p_0, \mathcal{U}_{nt}\gamma) \le e^{-nt/2}\mathcal{W}_2(p_0, \gamma) \to 0$ as $n \to \infty$, so $p_0 = \gamma$. Hence the uniqueness.

(vi) Action on densities relative to γ. Let $\mu \in L^2(\gamma)$ and let $f = \mathrm{d}\mu/\mathrm{d}\gamma$, then for any bounded measurable test function φ, by duality,

$$\int_{\mathbb{R}^d} \varphi\, d(\mathcal{U}_t\mu) = \int_{\mathbb{R}^d} P_t\varphi\, d\mu = \int_{\mathbb{R}^d} P_t\varphi(\mathbf{x}) f(\mathbf{x})\,\gamma(\mathrm{d}\mathbf{x}).$$

The OU semigroup is reversible with respect to γ, hence P_t is self-adjoint in $L^2(\gamma)$ [7]:

$$\int_{\mathbb{R}^d} P_t\varphi(\mathbf{x}) f(\mathbf{x})\,\gamma(\mathrm{d}\mathbf{x}) = \int_{\mathbb{R}^d} \varphi(\mathbf{x}) P_t f(\mathbf{x})\,\gamma(\mathrm{d}\mathbf{x}).$$

Therefore,

$$\int_{\mathbb{R}^d} \varphi\, d(\mathcal{U}_t\mu) = \int_{\mathbb{R}^d} \varphi(\mathbf{x}) P_t f(\mathbf{x})\,\gamma(\mathrm{d}\mathbf{x}).$$

Since this holds for all bounded measurable φ, we conclude that $\frac{d(\mathcal{U}_t\mu)}{d\gamma} = P_t f$.

(vii) Continuity in time. Fix $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $s \in [0,T]$. Let $\mathbf{X} \sim \mu$, $\perp\!\!\!\perp$ $\mathbf{Z} \sim \mathcal{N}(0, \mathbf{I}_d)$ and $\mathbf{Y}_t = e^{-t/2}\mathbf{X} + \sqrt{1-e^{-t}}\mathbf{Z}$. Then, $\mathbf{Y}_t \sim \mathcal{U}_t$ and coupling $\mathbf{Y}_t$ with $\mathbf{X}$ gives,

$$\begin{aligned}\mathcal{W}_2^2(\mathcal{U}_t\mu, \mu) &\leq \mathbb{E}\big[\|\mathbf{Y}_t - \mathbf{X}\|_2^2\big] \\ &= (e^{-t/2}-1)^2\mathbb{E}\big[\|\mathbf{X}\|_2^2\big] + (1-e^{-t})\mathbb{E}\big[\|\mathbf{Z}\|_2^2\big] \\ &= (e^{-t/2}-1)^2\mathbb{E}\big[\|\mathbf{X}\|_2^2\big] + d(1-e^{-t}).\end{aligned}$$

Hence,

$$\mathcal{W}_2^2(\mathcal{U}_t\mu, \mu) \leq (e^{-t/2}-1)^2\mathbb{E}\big[\|\mathbf{X}\|_2^2\big] + d(1-e^{-t}) \underset{t\to 0}{\longrightarrow} 0.$$

For the more general case, observe that for $s, t \geq 0$, using the semigroup property (proved in *(ii)*) and the contraction of the OU semigroup in $\mathcal{W}_2$ (Proposition B.5), one has that,

$$\mathcal{W}_2(\mathcal{U}_t\mu, \mathcal{U}_s\mu) = \mathcal{W}_2(\mathcal{U}_{\min(s,t)}\mathcal{U}_{|t-s|}\mu, \mathcal{U}_{\min(s,t)}\mu) \leq \mathcal{W}_2(\mathcal{U}_{|t-s|}\mu, \mu)$$

which tends to zero as $|t-s| \to 0$ by the previous argument. □

B.2 Role of the Gaussian distribution and Decomposition in the Hermite Polynomial Basis

We refer the reader to [12, 41] for a detailed review of the properties that we recall in the following. In our context, $\gamma = \mathcal{N}(0, \mathbf{I}_d)$ plays a special role because, as established in Proposition B.1, it is the *unique invariant measure* of the Ornstein–Uhlenbeck process, i.e. the unique distribution such that

$$\mathcal{U}_t(\gamma) = \gamma.$$

Hermite polynomials. Hermite polynomials are a family of polynomials that play a key role in the analysis of Gaussian Hilbert spaces, and hence, because of the central role of Gaussians in diffusion processes, in the theory of Markov operators. For $k \in \mathbb{N}$, the degree-k (probabilists') univariate Hermite polynomial is defined for all $x \in \mathbb{R}$ by

$$H_k(x) = (-1)^k e^{x^2/2}\frac{\mathrm{d}^k}{\mathrm{d}x^k}e^{-x^2/2}, \tag{26}$$

and satisfies $\int_{\mathbb{R}} H_k(x)H_\ell(x)\,\gamma_1(\mathrm{d}x) = k!\,\delta_{k\ell}$ with $\gamma_1 = \mathcal{N}(0,1)$. For a multi-index $\mathbf{n} = (n_1, \ldots, n_d) \in \mathbb{N}^d$ we write $\mathbf{n}! := \prod_{j=1}^d n_j!$ and $|\mathbf{n}| := n_1 + \cdots + n_d$, and define the *normalized* multivariate Hermite polynomials, as in Section 3, by

$$H_{\mathbf{n}}(\mathbf{x}) = \frac{1}{\sqrt{\mathbf{n}!}}\prod_{j=1}^d H_{n_j}(x_j), \qquad \mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d. \tag{27}$$

With this normalization, $\langle H_{\mathbf{n}}, H_{\mathbf{m}}\rangle_{L^2(\gamma)} = \delta_{\mathbf{nm}}$.

We now state two key properties that make these polynomials important analytic tools, the proofs of which can be find in [12, 41]. The first one is that they form an orthogonal basis of $L^2(\gamma)$.

Proposition B.2 (Orthonormal basis of $L^2(\gamma)$)**.** *The family $\{H_{\mathbf{n}}\}_{\mathbf{n}\in\mathbb{N}^d}$ defined in (27) is an orthonormal basis of $L^2(\gamma)$: for any $f \in L^2(\gamma)$,*

$$f = \sum_{\mathbf{n}\in\mathbb{N}^d}\langle f, H_{\mathbf{n}}\rangle_{L^2(\gamma)}H_{\mathbf{n}} \qquad \textit{and} \qquad \|f\|^2_{L^2(\gamma)} = \sum_{\mathbf{n}\in\mathbb{N}^d}\langle f, H_{\mathbf{n}}\rangle^2_{L^2(\gamma)}, \tag{28}$$

with convergence of the series in $L^2(\gamma)$.

The second key property is that Hermite polynomials are eigenfunctions of the OU semigroup $(P_t)_t$, defined in (23).

Proposition B.3 (Diagonalization of the OU semigroup)**.** *For each multi-index $\mathbf{n} \in \mathbb{N}^d$, denote the degree of a multivariate Hermite polynomial $H_{\mathbf{n}}$ as $|\mathbf{n}| = n_1 + \cdots + n_d$. Then, for any $t \in [0,T]$,*

$$P_t H_{\mathbf{n}} = e^{-|\mathbf{n}|t/2}H_{\mathbf{n}}.$$

In other words, multivariate Hermite polynomials are the eigenfunctions of the semigroup $(P_t)_t$, associated with the eigenvalue $e^{-|\mathbf{n}|t/2}$ for polynomials of degree $|\mathbf{n}|$.

In the context of this work, we are mainly interested in the effect of Hermite polynomials on the sampling operators $(\mathcal{U}_t)_t$. By combining Proposition B.3 and Proposition B.1, one has that, for any $f \in L^2(\gamma)$,

$$f = \sum_{\mathbf{n}} \langle f, H_{\mathbf{n}} \rangle_{L^2(\gamma)} H_{\mathbf{n}} \quad \Longrightarrow \quad P_t f = \sum_{\mathbf{n}} e^{-|\mathbf{n}|t/2} \langle f, H_{\mathbf{n}} \rangle_{L^2(\gamma)} H_{\mathbf{n}}. \tag{29}$$

Thus, combining this result with *(vi)* in Proposition B.1, we obtain a spectral representation of the action of the sampling operator $\mathcal{U}_t$ on probability measures.

Proposition B.4. *Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ such that $\mu \ll \gamma$ and $f = \frac{\mathrm{d}\mu}{\mathrm{d}\gamma} \in L^2(\gamma)$, then,*

$$\frac{\mathrm{d}(\mathcal{U}_t \mu)}{\mathrm{d}\gamma} = \sum_{\mathbf{n} \in \mathbb{N}^d} e^{-|\mathbf{n}|t/2} \langle f, H_{\mathbf{n}} \rangle_{L^2(\gamma)} H_{\mathbf{n}}. \tag{30}$$

This provides a complete spectral characterization of the OU dynamics.

B.3 Wasserstein-2 Geometry

The space of probability measures on $\mathbb{R}^d$ with finite second moment, denoted $\mathcal{P}_2(\mathbb{R}^d)$, is equipped with the Wasserstein-2 distance defined for all $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ as,

$$\mathcal{W}_2^2(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\mathbf{x} - \mathbf{y}\|^2 \, \pi(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y}), \tag{31}$$

where $\Pi(\mu, \nu)$ is the set of couplings of μ and ν. Altogether, the space $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$ is a complete separable metric space [56, 53, 67].

A key property used in this work is the contraction of the OU transition operator $(\mathcal{U}_t)_{t \in [0,T]}$ in the above Wasserstein-2 geometry.

Proposition B.5 (Wasserstein contraction of OU semigroup [67, 7])**.** *Let $(\mathcal{U}_t)_{t \geq 0}$ denote the Ornstein–Uhlenbeck sampling operator defined in (24). Then for all $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$\mathcal{W}_2\big(\mathcal{U}_t \mu, \mathcal{U}_t \nu\big) \leq e^{-t/2} \mathcal{W}_2(\mu, \nu). \tag{32}$$

Proof. Let $(\mathbf{X}, \mathbf{Y})$ be an optimal coupling of (μ, ν), i.e. a joint distribution on $\mathbb{R}^d \times \mathbb{R}^d$ such that the marginals satisfy $\mathbf{X} \sim \mu$ and $\mathbf{Y} \sim \nu$. Additionally, let $\mathbf{Z} \sim \mathcal{N}(0, I_d)$ be an independent Gaussian noise. Furthermore, define the time-t diffused marginals of $\mathbf{X}$ and $\mathbf{Y}$ as,

$$\mathbf{X}_t = e^{-t/2}\mathbf{X} + \sqrt{1 - e^{-t}}\mathbf{Z}, \qquad \mathbf{Y}_t = e^{-t/2}\mathbf{Y} + \sqrt{1 - e^{-t}}\mathbf{Z}.$$

Then, by definition of $\mathcal{U}_t$, $\mathbf{X}_t \sim \mathcal{U}_t \mu$ and $\mathbf{Y}_t \sim \mathcal{U}_t \nu$, hence

$$\mathcal{W}_2^2(\mathcal{U}_t \mu, \mathcal{U}_t \nu) \leq \mathbb{E}_{(\mathbf{X},\mathbf{Y})}\big[\|\mathbf{X}_t - \mathbf{Y}_t\|^2\big] = e^{-t}\mathbb{E}_{(\mathbf{X},\mathbf{Y})}\big[\|\mathbf{X} - \mathbf{Y}\|^2\big] = e^{-t}\mathcal{W}_2^2(\mu, \nu).$$

Taking the square-root yields the result. □

We now state a property of $\mathcal{W}_2$ that will prove very useful when working on the limiting distribution $p_\infty^\star$ which can be expressed as a Neumann series.

Proposition B.6 (Convexity of $\mathcal{W}_2^2$ under mixtures)**.** *Let $(\lambda_k)_{k \geq 0} \in (\mathbb{R}_+)^{\mathbb{N}}$ be nonnegative weights such that $\sum_{k=0}^{\infty} \lambda_k = 1$. Let $(\mu_k)_{k \geq 0}$ and $(\nu_k)_{k \geq 0}$ be probability measures in $\mathcal{P}_2(\mathbb{R}^d)$ such that the mixtures*

$$\mu := \sum_{k=0}^{\infty} \lambda_k \mu_k, \qquad \nu := \sum_{k=0}^{\infty} \lambda_k \nu_k$$

belong to $\mathcal{P}_2(\mathbb{R}^d)$. Then

$$\mathcal{W}_2^2(\mu, \nu) \leq \sum_{k=0}^{\infty} \lambda_k \mathcal{W}_2^2(\mu_k, \nu_k). \tag{33}$$

Proof. For each $k \geq 0$, let $\pi_k \in \Gamma(\mu_k, \nu_k)$ be an optimal coupling, so that

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \|\mathbf{x} - \mathbf{y}\|^2 \, \pi_k(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y}) = \mathcal{W}_2^2(\mu_k, \nu_k).$$

Define $\pi := \sum_{k=0}^{\infty} \lambda_k \pi_k$, then π is a probability measure on $\mathbb{R}^d \times \mathbb{R}^d$. Moreover,

$$\sum_{k=0}^{\infty} \lambda_k \mu_k = \mu, \qquad \sum_{k=0}^{\infty} \lambda_k \nu_k = \nu,$$

meaning $\pi \in \Gamma(\mu, \nu)$. Therefore,

$$\mathcal{W}_2^2(\mu, \nu) \leq \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\mathbf{x} - \mathbf{y}\|^2 \, \pi(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y}).$$

Using the definition of π,

$$\int \|\mathbf{x} - \mathbf{y}\|^2 \, \pi(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y}) = \sum_{k=0}^{\infty} \lambda_k \int \|\mathbf{x} - \mathbf{y}\|^2 \, \pi_k(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y}).$$

Thus,

$$\mathcal{W}_2^2(\mu, \nu) \leq \sum_{k=0}^{\infty} \lambda_k \mathcal{W}_2^2(\mu_k, \nu_k),$$

which proves the claim. □

B.4 Neumann Series and Fixed-Point Operators

This section introduces the following fundamental proposition linking Neumann Series, and linear operators on a Banach space.

Proposition B.7 (Neumann Series [54])**.** *Let $\mathcal{T}$ be a linear operator on a Banach space with $\|\mathcal{T}\| < 1$. Then the Neumann series*

$$(I - \mathcal{T})^{-1} = \sum_{k=0}^{\infty} \mathcal{T}^k \tag{34}$$

converges in operator norm.

In the setting described in Section 3, the recursion

$$\mu \mapsto \mathcal{U}_{t_0}\big(\alpha p_{\mathrm{data}} + (1 - \alpha)\mu\big) \tag{35}$$

can be rewritten as a linear fixed-point equation whose solution is expressed as a Neumann series in the operator $(1 - \alpha)\mathcal{U}_{t_0}$, viewed as a bounded linear operator on the Banach space of finite signed measures on $\mathbb{R}^d$ with the total-variation norm. Since $\mathcal{U}_{t_0}$ is given by a Markov kernel, $\|\mathcal{U}_{t_0}\mu\|_{\mathrm{TV}} \leq \|\mu\|_{\mathrm{TV}}$, so that $\|(1 - \alpha)\mathcal{U}_{t_0}\| \leq 1 - \alpha < 1$ and Proposition B.7 applies.

C Intermediate Results

In this section, we summarize two key results that are consistently invoked in our proofs. These are specific to our setting, but fundamentally rely on the classical results stated in Section B.

The first central result is the contraction property of the one-step ideal sampling operator in the error-free regime described in Section 3.

Proposition C.1 (Contraction of the one-step ideal sampling operator)**.** *Let $\mathcal{T} : \mathcal{P}_2(\mathbb{R}^d) \to \mathcal{P}_2(\mathbb{R}^d)$ denote the operator that maps, to each $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ the end-point marginal of the ideal reverse diffusion (truncated at $t_0 > 0$) , i.e.,*

$$\mathcal{T}\mu = \mathcal{U}_{t_0}\big(\alpha p_{\mathrm{data}} + (1 - \alpha)\mu\big).$$

Then, $\mathcal{T}$ is a contraction on $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$, with constant $\kappa = \sqrt{1 - \alpha}e^{-t_0/2}$, i.e.,

$$\mathcal{W}_2(\mathcal{T}\mu, \mathcal{T}\nu) \leq \sqrt{1 - \alpha}e^{-t_0/2}\mathcal{W}_2(\mu, \nu). \tag{36}$$

Proof. Consider two probability distributions $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, and denote

$$q_\mu := \alpha p_{\text{data}} + (1-\alpha)\mu, \qquad q_\nu := \alpha p_{\text{data}} + (1-\alpha)\nu.$$

Then, by contractivity of $\mathcal{W}_2$ on $\mathcal{U}_{t_0}$ (Proposition B.5),

$$\begin{aligned}\mathcal{W}_2(\mathcal{T}\mu, \mathcal{T}\nu) &= \mathcal{W}_2\Big(\mathcal{U}_{t_0}\big(\alpha p_{\text{data}} + (1-\alpha)\mu\big), \mathcal{U}_{t_0}\big(\alpha p_{\text{data}} + (1-\alpha)\nu\big)\Big) \\ &\leq\ e^{-t_0/2}\mathcal{W}_2(\alpha p_{\text{data}} + (1-\alpha)\mu,\ \alpha p_{\text{data}} + (1-\alpha)\nu) \\ &= e^{-t_0/2}\mathcal{W}_2(q_\mu, q_\nu). \end{aligned} \tag{37}$$

It remains to bound $\mathcal{W}_2(q_\mu, q_\nu)$. To that end, consider an optimal coupling π of μ and ν and let λ be the diagonal coupling of p_{data} with itself, namely the law of $(\mathbf{X}, \mathbf{X})$ with $\mathbf{X} \sim p_{\text{data}}$. Define the coupling,

$$\tilde{\pi} = \alpha\lambda + (1-\alpha)\pi.$$

Then $\tilde{\pi}$ has first marginal q_μ and second marginal q_ν. Therefore,

$$\mathcal{W}_2^2(q_\mu, q_\nu) \leq \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\mathbf{x} - \mathbf{y}\|_2^2 \tilde{\pi}(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y}).$$

But, by definition of $\tilde{\pi}$,

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \|\mathbf{x} - \mathbf{y}\|_2^2 \tilde{\pi}(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y}) = \alpha \int \|\mathbf{x} - \mathbf{y}\|_2^2 \lambda(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y}) + (1-\alpha) \int \|\mathbf{x} - \mathbf{y}\|_2^2 \pi(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y}).$$

The first term is zero, because λ is supported on the diagonal, while the second term equals $(1-\alpha)\mathcal{W}_2^2(\mu, \nu)$ since π is an optimal coupling between μ and ν. Hence,

$$\mathcal{W}_2^2(q_\mu, q_\nu) \leq (1-\alpha)\mathcal{W}_2^2(\mu, \nu),$$

and, equivalently, taking the square root,

$$\mathcal{W}_2(q_\mu, q_\nu) \leq \sqrt{1-\alpha}\,\mathcal{W}_2(\mu, \nu),$$

Combining with the OU contraction (37) gives the final result in (36). □

Since $\mathcal{T}$ is a contraction, the next natural question is to ask for a characterization of its fixed points.

Proposition C.2 (Fixed point of the one-step ideal sampling operator)**.** *Let $\mathcal{T} : \mathcal{P}_2(\mathbb{R}^d) \to \mathcal{P}_2(\mathbb{R}^d)$ denote the operator that maps, to each $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ the end-point marginal of the ideal reverse diffusion (truncated at $t_0 > 0$), i.e.,*

$$\mathcal{T}\mu = \mathcal{U}_{t_0}\big(\alpha p_{\text{data}} + (1-\alpha)\mu\big).$$

Then, any fixed point $\mu^\star$ of $\mathcal{T}$ satisfies,

$$(\mathrm{Id} - (1-\alpha)\mathcal{U}_{t_0})\mu^\star = \alpha\mathcal{U}_{t_0}(p_{\text{data}}).$$

Proof. Assume $\mu^\star$ is a fixed point of $\mathcal{T}$. Then,

$$\begin{aligned}\mathcal{T}\mu^\star = \mu^\star &\iff \mathcal{U}_{t_0}\big(\alpha p_{\text{data}} + (1-\alpha)\mu^\star\big) = \mu^\star \\ &\iff \alpha\mathcal{U}_{t_0}\big(p_{\text{data}}\big) + (1-\alpha)\mathcal{U}_{t_0}(\mu^\star) = \mu^\star \\ &\iff (\mathrm{Id} - (1-\alpha)\mathcal{U}_{t_0})\mu^\star = \alpha\mathcal{U}_{t_0}(p_{\text{data}}),\end{aligned}$$

where the second and third lines use the linearity of the sampling operator $\mathcal{U}_{t_0}$ (Proposition B.1). □

D Proofs

D.1 Proof of Theorem 3.1

Proof. (i) Limiting distribution expression

Recall that $\mathcal{T} : \mu \mapsto \mathcal{U}_{t_0}(\alpha p_{\text{data}} + (1-\alpha)\mu)$, where the operator $\mathcal{U}_{t_0}$ is defined in (5), is the one-step ideal sampling operator in the error-free regime. By Proposition C.1, $\mathcal{T}$ is a contraction

on $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$ with constant $\kappa = \sqrt{(1-\alpha)}\, e^{-t_0/2} < 1$, and Banach's fixed point theorem [8] gives existence and uniqueness.

It remains to identify $p^\star_\infty$ with the series (9). Define

$$\nu := \alpha \sum_{k=0}^{\infty} (1-\alpha)^k \, \mathcal{U}_{(k+1)t_0} p_{\text{data}}.$$

Since the weights $\alpha(1-\alpha)^k$ are nonnegative and sum to one, ν is a probability measure; moreover, by the identity $\mathbb{E}\|\mathbf{Y}\|^2 = e^{-t}\mathbb{E}\|\mathbf{X}\|^2 + d(1-e^{-t})$ established in the proof of Proposition B.1*(iv)*, each $\mathcal{U}_{(k+1)t_0} p_{\text{data}}$ has second moment at most $\int \|\mathbf{x}\|^2 p_{\text{data}}(\mathrm{d}\mathbf{x}) + d$ uniformly in k, so $\nu \in \mathcal{P}_2(\mathbb{R}^d)$. By linearity of $\mathcal{U}_{t_0}$ (Proposition B.1*(i)*) and the semigroup property *(ii)*,

$$\begin{aligned}
\mathcal{T}\nu &= \alpha \, \mathcal{U}_{t_0} p_{\text{data}} + (1-\alpha)\, \alpha \sum_{k=0}^{\infty} (1-\alpha)^k \, \mathcal{U}_{t_0} \mathcal{U}_{(k+1)t_0} p_{\text{data}} \\
&= \alpha \, \mathcal{U}_{t_0} p_{\text{data}} + \alpha \sum_{k=0}^{\infty} (1-\alpha)^{k+1} \, \mathcal{U}_{(k+2)t_0} p_{\text{data}} = \alpha \sum_{j=0}^{\infty} (1-\alpha)^j \, \mathcal{U}_{(j+1)t_0} p_{\text{data}} = \nu.
\end{aligned}$$

Hence ν is a fixed point of $\mathcal{T}$ and, by uniqueness, $p^\star_\infty = \nu$, which is (9). Equivalently, $p^\star_\infty$ is the unique solution of the linear equation $(\mathrm{Id} - (1-\alpha)\mathcal{U}_{t_0})\mu = \alpha \, \mathcal{U}_{t_0} p_{\text{data}}$ of Proposition C.2, and (9) is its Neumann-series solution; this can also be justified via Proposition B.7 on the Banach space of finite signed measures equipped with the total-variation norm, on which $\|\mathcal{U}_{t_0}\| \le 1$ (it is given by a Markov kernel) and therefore $\|(1-\alpha)\mathcal{U}_{t_0}\| \le 1 - \alpha < 1$.

(ii) Geometric rate of convergence

Define $D_N := \mathcal{W}_2(\hat{p}^N, p^\star_\infty)$, for $N \ge 0$. We claim that $D_{N+1} \le \kappa \, D_N$. By the triangle inequality,

$$D_{N+1} = \mathcal{W}_2(\hat{p}^{N+1}, p^\star_\infty) \le \mathcal{W}_2(\hat{p}^{N+1}, \mathcal{T}\hat{p}^N) + \mathcal{W}_2(\mathcal{T}\hat{p}^N, p^\star_\infty)$$

First term. Since we are working in the error-free regime $\hat{p}^{N+1} = \mathcal{T}\hat{p}^N$ (perfect sampler) thus $\mathcal{W}_2(\hat{p}^{N+1}, \mathcal{T}\hat{p}^N) = 0$.

Second term. Since $p^\star_\infty = \mathcal{T}p^\star_\infty$, we need to bound $\mathcal{W}_2(\mathcal{T}\hat{p}^N, \mathcal{T}p^\star_\infty)$. Once again, the contraction property of $\mathcal{T}$ (Proposition C.1) yields,

$$\mathcal{W}_2(\mathcal{T}\hat{p}^N, p^\star_\infty) \le \kappa \mathcal{W}_2(\hat{p}^N, p^\star_\infty)$$

where $\kappa := \sqrt{1-\alpha}\, e^{-t_0/2} < 1$. Combining the two bounds:

$$D_{N+1} \le \kappa \, D_N \qquad N \ge 0, \tag{38}$$

and iterating (38) yields the geometric decay rate,

$$D_N \le \kappa^N \, D_0 \qquad N \ge 0.$$

□

D.2 Proof of Corollary 3.2

Proof. In this proof, we use the Neumann series representation (Theorem 3.1)

$$p^\star_\infty = \sum_{k=0}^{\infty} \pi_k(\alpha) \mathcal{U}_{(k+1)t_0}(p_{\text{data}}), \tag{39}$$

with geometric weights $\pi_k(\alpha) = \alpha(1-\alpha)^k$. We also denote $D := \mathcal{W}_2(p_{\text{data}}, \gamma) < \infty$. The proofs mainly rely on the two following facts:

1. *Convexity of $\mathcal{W}_2$ under mixtures.* By Proposition B.6, for any target distribution $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, since $p^\star_\infty$ given in (39) is a convex combination with $\sum_k \pi_k(\alpha) = 1$,

$$\mathcal{W}_2^2(p^\star_\infty, \nu) \le \sum_{k=0}^{\infty} \pi_k(\alpha) \mathcal{W}_2^2(\mathcal{U}_{(k+1)t_0} p_{\text{data}}, \nu).$$

2. *Contraction of $\mathcal{W}_2$ on the OU sampling operator*. By Proposition B.5, for any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, the OU sampling operator satisfies

$$\mathcal{W}_2(\mathcal{U}_t \mu, \gamma) = \mathcal{W}_2(\mathcal{U}_t \mu, \mathcal{U}_t \gamma) \le e^{-t/2} \mathcal{W}_2(\mu, \gamma), \tag{40}$$

since γ is invariant under $\mathcal{U}_t$.

(i) Limit $\alpha \to 1^-$. Fix $t_0 > 0$ and let $\mu_k := \mathcal{U}_{(k+1)t_0}(p_{\text{data}})$. Then $\mu_0 = \mathcal{U}_{t_0}(p_{\text{data}})$ and

$$p^\star_\infty = \alpha \mu_0 + \sum_{k=1}^{\infty} \pi_k(\alpha) \mu_k.$$

We can use $\{\pi_k(\alpha)\}_{k \ge 0}$ to define a natural coupling between $\mathbf{X} \sim \mu_0$ and $\mathbf{Y} \sim p^\star_\infty$, where $\mathbf{X} = \mathbf{Y}$ with probability $\pi_0(\alpha) = \alpha$. Then, by Proposition B.6 with target μ_0,

$$\mathcal{W}_2^2(p^\star_\infty, \mu_0) \le \sum_{k=1}^{\infty} \pi_k(\alpha) \mathcal{W}_2^2(\mu_k, \mu_0).$$

Moreover, by the triangle inequality and (40),

$$\mathcal{W}_2(\mu_k, \mu_0) \le \mathcal{W}_2(\mu_k, \gamma) + \mathcal{W}_2(\mu_0, \gamma) \le e^{-(k+1)t_0/2} D + e^{-t_0/2} D \le 2D.$$

Therefore

$$\mathcal{W}_2^2(p^\star_\infty, \mathcal{U}_{t_0} p_{\text{data}}) \le 4D^2 \sum_{k=1}^{\infty} \pi_k(\alpha) = 4D^2 \sum_{k=1}^{\infty} \alpha(1-\alpha)^k = 4D^2(1-\alpha),$$

which converges to zero as $\alpha \to 1^-$.

(ii) Limit $\alpha \to 0^+$. Using (33) with target γ and then (40),

$$\mathcal{W}_2^2(p^\star_\infty, \gamma) \le \sum_{k=0}^{\infty} \pi_k(\alpha) \mathcal{W}_2^2\big(\mathcal{U}_{(k+1)t_0} p_{\text{data}}, \gamma\big) \tag{41}$$

$$\le D^2 \sum_{k=0}^{\infty} \pi_k(\alpha) e^{-(k+1)t_0}. \tag{42}$$

The geometric sum is

$$\sum_{k=0}^{\infty} \pi_k(\alpha) e^{-(k+1)t_0} = \sum_{k=0}^{\infty} \alpha(1-\alpha)^k e^{-(k+1)t_0} = \frac{\alpha e^{-t_0}}{1 - (1-\alpha)e^{-t_0}}.$$

Since $t_0 > 0$, the denominator converges to $1 - e^{-t_0} > 0$ as $\alpha \to 0^+$. Hence

$$\mathcal{W}_2^2(p^\star_\infty, \gamma) \le D^2 \frac{\alpha e^{-t_0}}{1 - (1-\alpha)e^{-t_0}} \longrightarrow 0.$$

(iii) Limit $t_0 \to 0^+$. Fix $\alpha \in (0, 1]$. By Proposition B.1 *(vii)*, for each fixed k,

$$\mathcal{U}_{(k+1)t_0}(p_{\text{data}}) \xrightarrow[t_0 \to 0^+]{\mathcal{W}_2} p_{\text{data}}. \tag{43}$$

Now applying Proposition B.6 with p_{data} yields,

$$\mathcal{W}_2^2(p^\star_\infty, p_{\text{data}}) \le \sum_{k=0}^{\infty} \alpha(1-\alpha)^k \mathcal{W}_2^2\big(\mathcal{U}_{(k+1)t_0} p_{\text{data}}, p_{\text{data}}\big). \tag{44}$$

But by (43), for each fixed k, $\mathcal{W}_2^2\big(\mathcal{U}_{(k+1)t_0}p_{\mathrm{data}}, p_{\mathrm{data}}\big) \xrightarrow[t_0\to 0^+]{} 0$. Moreover, uniformly in k and $t_0 \geq 0$,

$$\mathcal{W}_2\big(\mathcal{U}_{(k+1)t_0}p_{\mathrm{data}}, p_{\mathrm{data}}\big) \leq \mathcal{W}_2\big(\mathcal{U}_{(k+1)t_0}p_{\mathrm{data}}, \gamma\big) + \mathcal{W}_2(\gamma, p_{\mathrm{data}}) \leq 2D.$$

Thus each summand in (44) is bounded by $4D^2\alpha(1-\alpha)^k$, which is summable in k. Therefore, by dominated convergence for series, we have

$$\begin{aligned}\lim_{t_0\to 0} \mathcal{W}_2^2(p_\infty^\star, p_{\mathrm{data}}) &\leq \lim_{t_0\to 0} \sum_{k=0}^{\infty} \alpha(1-\alpha)^k \mathcal{W}_2^2\big(\mathcal{U}_{(k+1)t_0}p_{\mathrm{data}}, p_{\mathrm{data}}\big) \\ &= \sum_{k=0}^{\infty} \alpha(1-\alpha)^k \lim_{t_0\to 0} \mathcal{W}_2^2\big(\mathcal{U}_{(k+1)t_0}p_{\mathrm{data}}, p_{\mathrm{data}}\big) = 0.\end{aligned}$$

(iv) Strict positivity when $p_{\mathrm{data}} \neq \gamma$. If one assumes by contradiction that $\mathcal{W}_2(p_\infty^\star, p_{\mathrm{data}}) = 0$, then $p_\infty^\star = p_{\mathrm{data}}$ but since $p_\infty^\star$ is characterized as a fixed point of $\mu \mapsto \mathcal{U}_{t_0}\big(\alpha p_{\mathrm{data}} + (1-\alpha)\mu\big)$, we obtain

$$p_{\mathrm{data}} = \mathcal{U}_{t_0}\big(\alpha p_{\mathrm{data}} + (1-\alpha)p_{\mathrm{data}}\big) = \mathcal{U}_{t_0}p_{\mathrm{data}}.$$

Thus p_{data} is invariant under the OU transition at time t_0. But, the OU semigroup has the unique invariant probability measure $\gamma = \mathcal{N}(0, \mathbf{I}_d)$ (Proposition B.1 *(v)*). Therefore $p_{\mathrm{data}} = \gamma$, and this contradicts the initial assumption. Thus, $p_{\mathrm{data}} \neq \gamma$ and $\mathcal{W}_2(p_\infty^\star, p_{\mathrm{data}}) > 0$. □

D.3 Moments of limiting distribution

Looking at the first two moments of the collapse distribution $p_\infty^\star$ hints at what collapse implies in terms of concentration.

Proposition D.1 (Moments). *Let $\boldsymbol{\mu} := \mathbb{E}_{p_{\mathrm{data}}}[\mathbf{X}]$ and define*

$$c_1 := \frac{\alpha\, e^{-t_0/2}}{1-(1-\alpha)e^{-t_0/2}}, \qquad c_2 := \frac{\alpha\, e^{-t_0}}{1-(1-\alpha)e^{-t_0}}.$$

The mean, second-moment matrix and covariance of $p_\infty^\star$ are

$$\mathbb{E}_{p_\infty^\star}[\mathbf{X}] = c_1\, \boldsymbol{\mu}, \tag{45}$$

$$\mathbb{E}_{p_\infty^\star}[\mathbf{X}\mathbf{X}^\top] = c_2\, \mathbb{E}_{p_{\mathrm{data}}}[\mathbf{X}\mathbf{X}^\top] + (1-c_2)\, \mathbf{I}_d, \tag{46}$$

$$\mathrm{Cov}_{p_\infty^\star}(\mathbf{X}) = c_2\, \mathrm{Cov}_{p_{\mathrm{data}}}(\mathbf{X}) + (1-c_2)\, \mathbf{I}_d + (c_2 - c_1^2)\, \boldsymbol{\mu}\boldsymbol{\mu}^\top. \tag{47}$$

Remark D.2. Here $0 < c_1, c_2 < 1$ and $c_2 \geq c_1^2$, with equality only if $t_0 = 0$. Thus the mean is attenuated relative to p_{data} for any $\alpha < 1$, and when p_{data} is centered the covariance is pulled toward $\mathbf{I}_d$; for non-centered p_{data} an additional rank-one term along $\boldsymbol{\mu}$ appears, reflecting the spread of the block means $e^{-(k+1)t_0/2}\boldsymbol{\mu}$ in (9). Note that $c_1 = m_{\mathbf{n}}(\alpha, t_0)$ for $|\mathbf{n}| = 1$ and $c_2 = m_{\mathbf{n}}(\alpha, t_0)$ for $|\mathbf{n}| = 2$, consistently with Proposition 3.4.

Proof. Write $\lambda_k := e^{-(k+1)t_0}$ and $\pi_k := \alpha(1-\alpha)^k$, so that $p_\infty^\star = \sum_{k\geq 0} \pi_k\, \mathcal{U}_{(k+1)t_0}p_{\mathrm{data}}$ by (9). From (5), if $\mathbf{Y} \sim \mathcal{U}_t\mu$ with $\mathbf{X} \sim \mu$ and $\mathbf{Z} \sim \mathcal{N}(0, \mathbf{I}_d)$ independent, then $\mathbf{Y} = e^{-t/2}\mathbf{X} + \sqrt{1-e^{-t}}\,\mathbf{Z}$ and hence

$$\mathbb{E}[\mathbf{Y}] = e^{-t/2}\,\mathbb{E}[\mathbf{X}], \qquad \mathbb{E}[\mathbf{Y}\mathbf{Y}^\top] = e^{-t}\,\mathbb{E}[\mathbf{X}\mathbf{X}^\top] + (1-e^{-t})\,\mathbf{I}_d.$$

The mean and the second-moment matrix are linear in the underlying measure, so they commute with the mixture:

$$\begin{aligned}\mathbb{E}_{p_\infty^\star}[\mathbf{X}] &= \sum_{k\geq 0} \pi_k\, \lambda_k^{1/2}\, \boldsymbol{\mu} = \alpha e^{-t_0/2} \sum_{k\geq 0} \big[(1-\alpha)e^{-t_0/2}\big]^k\, \boldsymbol{\mu} = c_1\, \boldsymbol{\mu}, \\ \mathbb{E}_{p_\infty^\star}[\mathbf{X}\mathbf{X}^\top] &= \sum_{k\geq 0} \pi_k \big(\lambda_k\, \mathbb{E}_{p_{\mathrm{data}}}[\mathbf{X}\mathbf{X}^\top] + (1-\lambda_k)\, \mathbf{I}_d\big) = c_2\, \mathbb{E}_{p_{\mathrm{data}}}[\mathbf{X}\mathbf{X}^\top] + (1-c_2)\, \mathbf{I}_d,\end{aligned}$$

using $\sum_k \pi_k \lambda_k = c_2$ and $\sum_k \pi_k = 1$. For the covariance, we use the above expressions,

$$\begin{aligned}\mathrm{Cov}_{p^\star_\infty}(\mathbf{X}) &= \mathbb{E}_{p^\star_\infty}[\mathbf{X}\mathbf{X}^\top] - \mathbb{E}_{p^\star_\infty}[\mathbf{X}]\,\mathbb{E}_{p^\star_\infty}[\mathbf{X}]^\top \\ &= c_2\big(\mathrm{Cov}_{p_{\mathrm{data}}}(\mathbf{X}) + \boldsymbol{\mu}\boldsymbol{\mu}^\top\big) + (1-c_2)\,\mathbf{I}_d - c_1^2\,\boldsymbol{\mu}\boldsymbol{\mu}^\top \\ &= c_2\,\mathrm{Cov}_{p_{\mathrm{data}}}(\mathbf{X}) + (1-c_2)\,\mathbf{I}_d + (c_2 - c_1^2)\,\boldsymbol{\mu}\boldsymbol{\mu}^\top.\end{aligned}$$

Finally, $c_1 = \sum_k \pi_k \lambda_k^{1/2}$ and $c_2 = \sum_k \pi_k \lambda_k$ are the first and second moments of $\lambda_K^{1/2}$ for K geometric with $\mathbb{P}(K=k) = \pi_k$, so $c_2 - c_1^2 = \mathrm{Var}(\lambda_K^{1/2}) \geq 0$, with equality iff λ_k does not depend on k, i.e. iff $t_0 = 0$. □

D.4 Proof of Proposition 3.4

Proof. Under Assumption A4, the data distribution admits a density $f_{\mathrm{data}} := \frac{\mathrm{d}p_{\mathrm{data}}}{\mathrm{d}\gamma} \in L^2(\gamma)$, where $\gamma = \mathcal{N}(0, \mathbf{I}_d)$. Since the OU sampling operator $\mathcal{U}_t$ preserves absolute continuity with respect to γ, the collapse distribution $p^\star_\infty$ also admits a density $f^\star_\infty := \frac{\mathrm{d}p^\star_\infty}{\mathrm{d}\gamma}$.

From Theorem 3.1, we have the Neumann-series representation

$$p^\star_\infty = \alpha \sum_{k=0}^{\infty} (1-\alpha)^k \mathcal{U}_{(k+1)t_0} p_{\mathrm{data}}. \tag{48}$$

Recall $(P_t)_{t\geq 0}$ denotes the OU semigroup acting on $L^2(\gamma)$ (see Section B). By Proposition B.1 *(vi)*),

$$\frac{\mathrm{d}(\mathcal{U}_t \mu)}{\mathrm{d}\gamma} = P_t\left(\frac{\mathrm{d}\mu}{\mathrm{d}\gamma}\right).$$

Applying this to (48) and using the linearity of $\mathcal{U}_t$, we obtain

$$f^\star_\infty = \alpha \sum_{k=0}^{\infty} (1-\alpha)^k P_{(k+1)t_0} f_{\mathrm{data}}, \tag{49}$$

with convergence in $L^2(\gamma)$.

Since $f_{\mathrm{data}} \in L^2(\gamma)$, by Proposition B.2

$$f_{\mathrm{data}} = \sum_{\mathbf{n}\in\mathbb{N}^d} \langle f_{\mathrm{data}}, H_{\mathbf{n}}\rangle_{L^2(\gamma)} H_{\mathbf{n}}, \tag{50}$$

with convergence in $L^2(\gamma)$. But, by Proposition B.3, for each $\mathbf{n} \in \mathbb{N}^d$ and $k \geq 0$,

$$P_{(k+1)t_0} H_{\mathbf{n}} = e^{-|\mathbf{n}|(k+1)t_0/2} H_{\mathbf{n}}. \tag{51}$$

Thus, applying $P_{(k+1)t_0}$, and combining equations (51) and (50) yields,

$$P_{(k+1)t_0} f_{\mathrm{data}} = \sum_{\mathbf{n}\in\mathbb{N}^d} e^{-|\mathbf{n}|(k+1)t_0/2} \langle f_{\mathrm{data}}, H_{\mathbf{n}}\rangle_{L^2(\gamma)} H_{\mathbf{n}}. \tag{52}$$

Substituting (52) into (49),

$$f^\star_\infty = \alpha \sum_{k=0}^{\infty} (1-\alpha)^k \sum_{\mathbf{n}\in\mathbb{N}^d} e^{-|\mathbf{n}|(k+1)t_0/2} \langle f_{\mathrm{data}}, H_{\mathbf{n}}\rangle H_{\mathbf{n}}. \tag{53}$$

Since P_t is a contraction on $L^2(\gamma)$ and the coefficients are square-summable, the series converges absolutely in $L^2(\gamma)$, which justifies exchanging the sums. Hence,

$$f^\star_\infty = \sum_{\mathbf{n}\in\mathbb{N}^d} \left[\alpha \sum_{k=0}^{\infty} (1-\alpha)^k e^{-|\mathbf{n}|(k+1)t_0/2}\right] \langle f_{\mathrm{data}}, H_{\mathbf{n}}\rangle H_{\mathbf{n}}.$$

Now looking at each term inside the series on $\mathbb{N}^d$, we observe that, for each $\mathbf{n} \in \mathbb{N}^d$,

$$\alpha \sum_{k=0}^{\infty} (1-\alpha)^k e^{-|\mathbf{n}|(k+1)t_0/2} = \alpha e^{-|\mathbf{n}|t_0/2} \sum_{k=0}^{\infty} \left[(1-\alpha)e^{-|\mathbf{n}|t_0/2}\right]^k$$
$$= \alpha e^{-|\mathbf{n}|t_0/2} \frac{1}{1-(1-\alpha)e^{-|\mathbf{n}|t_0/2}} \qquad ((1-\alpha)e^{-|\mathbf{n}|t_0/2} < 1).$$

Denoting $m_{\mathbf{n}}(\alpha, t_0) = \frac{\alpha e^{-|\mathbf{n}|t_0/2}}{1-(1-\alpha)e^{-|\mathbf{n}|t_0/2}}$, yields (12):

$$\langle f_\infty^\star, H_{\mathbf{n}} \rangle_{L^2(\gamma)} = m_{\mathbf{n}}(\alpha, t_0) \langle f_{\text{data}}, H_{\mathbf{n}} \rangle_{L^2(\gamma)},$$

Therefore, each Hermite mode of the data distribution is attenuated by the explicit factor $m_{\mathbf{n}}(\alpha, t_0)$. This shows that the collapse distribution is a spectrally filtered version of p_{data}, with preferential suppression of high-frequency modes. □

D.5 Proof of the finite-moment attenuation identity (Remark 3.5)

Proof. Fix $\mathbf{n} \in \mathbb{N}^d$. Assume $\int_{\mathbb{R}^d} \|\mathbf{x}\|^{|\mathbf{n}|} \, p_{\text{data}}(\mathrm{d}\mathbf{x}) < \infty$, let $\mathbf{X} \sim p_{\text{data}}$ and $\mathbf{Z} \sim \mathcal{N}(0, \mathbf{I}_d)$ be independent, and write $\sigma_t := \sqrt{1-e^{-t}}$. $H_{\mathbf{n}}$ is a polynomial of total degree $|\mathbf{n}|$, i.e. a finite sum $\sum_{|\boldsymbol{\beta}| \le |\mathbf{n}|} c_{\boldsymbol{\beta}} \, \mathbf{x}^{\boldsymbol{\beta}}$. Since $|\mathbf{x}^{\boldsymbol{\beta}}| \le \|\mathbf{x}\|^{|\boldsymbol{\beta}|} \le 1 + \|\mathbf{x}\|^{|\mathbf{n}|}$, we have $|H_{\mathbf{n}}(\mathbf{x})| \le C_{\mathbf{n}}(1 + \|\mathbf{x}\|^{|\mathbf{n}|})$ with $C_{\mathbf{n}} := \sum_\beta |c_\beta|$. By Proposition B.1(iv),

$$\sup_{t \ge 0} \left\{ \mathbb{E}_{\mathcal{U}_t p_{\text{data}}}[|H_{\mathbf{n}}|] \right\} \le C_{\mathbf{n}} \Big(1 + \big(M_{|\mathbf{n}|}(p_{\text{data}})^{1/|\mathbf{n}|} + (\mathbb{E}\|\mathbf{Z}\|^{|\mathbf{n}|})^{1/|\mathbf{n}|} \big)^{|\mathbf{n}|} \Big) =: B_{\mathbf{n}} < \infty. \quad (54)$$

Fubini's theorem therefore applies, and conditioning on $\mathbf{X}$ gives

$$\mathbb{E}_{\mathcal{U}_t p_{\text{data}}}[H_{\mathbf{n}}] = \mathbb{E}\big[H_{\mathbf{n}}(e^{-t/2}\mathbf{X} + \sigma_t \mathbf{Z})\big] = \mathbb{E}\big[(P_t H_{\mathbf{n}})(\mathbf{X})\big] \overset{\text{Prop. } B.3}{=} e^{-|\mathbf{n}|t/2} \mathbb{E}_{p_{\text{data}}}\big[H_{\mathbf{n}}\big]. \quad (55)$$

Recall that $p_\infty^\star$ is defined in (9) as,

$$p_\infty^\star = \sum_{j \ge 0} \alpha(1-\alpha)^j \, \mathcal{U}_{(j+1)t_0} p_{\text{data}}.$$

so that, by monotone convergence, for every measurable $g \ge 0$,

$$\int g \, \mathrm{d}p_\infty^\star = \sum_j \alpha(1-\alpha)^j \int g \, \mathrm{d}(\mathcal{U}_{(j+1)t_0} p_{\text{data}}).$$

Taking $g = |H_{\mathbf{n}}|$, (54) gives $\mathbb{E}_{p_\infty^\star}|H_{\mathbf{n}}| \le B_{\mathbf{n}} < \infty$ and using (55),

$$\mathbb{E}_{p_\infty^\star}[H_{\mathbf{n}}] = \sum_{j \ge 0} \alpha(1-\alpha)^j \, \mathbb{E}_{\mathcal{U}_{(j+1)t_0} p_{\text{data}}}[H_{\mathbf{n}}]$$
$$= \alpha \, \mathbb{E}_{p_{\text{data}}}[H_{\mathbf{n}}] \sum_{j \ge 0} (1-\alpha)^j \, e^{-|\mathbf{n}|(j+1)t_0/2} = m_{\mathbf{n}}(\alpha, t_0) \, \mathbb{E}_{p_{\text{data}}}[H_{\mathbf{n}}].$$

□

D.6 Proof of Theorem 4.1

Proof. We consider the error-free recursion

$$\hat{p}^{i+1} = \mathcal{U}_{t_0^{(i)}}\big(\alpha \, p_{\text{data}} + (1-\alpha)\hat{p}^i\big), \qquad i \ge 0. \quad (56)$$

We first prove (15) by induction on N. For $N = 1$, by linearity of $\mathcal{U}_t$ (Proposition B.1 *(i)*),

$$\hat{p}^1 = \mathcal{U}_{t_0^{(0)}}\big(\alpha \, p_{\text{data}} + (1-\alpha)\hat{p}^0\big) = \alpha \, \mathcal{U}_{t_0^{(0)}} p_{\text{data}} + (1-\alpha) \, \mathcal{U}_{t_0^{(0)}} p_{\text{data}},$$

which agrees with (15), since

$$\sigma_{0,1} = t_0^{(0)}, \qquad s_{0,1} = t_0^{(0)}.$$

Assume now that (15) holds for some $N \geq 1$, that is,

$$\hat{p}^N = \alpha \sum_{m=0}^{N-1} (1-\alpha)^m \mathcal{U}_{\sigma_{m,N}} p_{\text{data}} + (1-\alpha)^N \mathcal{U}_{s_{0,N}} p_{\text{data}},$$

where $\sigma_{m,N} = \sum_{\ell=N-1-m}^{N-1} t_0^{(\ell)}$ and $s_{0,N} = \sum_{\ell=0}^{N-1} t_0^{(\ell)}$. Once again, by linearity of $\mathcal{U}_t$ and leveraging the induction assumption,

$$\begin{aligned}
\hat{p}^{N+1} &= \mathcal{U}_{t_0^{(N)}} \left(\alpha\, p_{\text{data}} + (1-\alpha) \hat{p}^N \right) \\
&= \alpha \mathcal{U}_{t_0^{(N)}} p_{\text{data}} + (1-\alpha)\, \mathcal{U}_{t_0^{(N)}} \hat{p}^N \\
&= \alpha \mathcal{U}_{t_0^{(N)}} p_{\text{data}} + (1-\alpha)\, \mathcal{U}_{t_0^{(N)}} \left[\alpha \sum_{m=0}^{N-1} (1-\alpha)^m \mathcal{U}_{\sigma_{m,N}} p_{\text{data}} + (1-\alpha)^N \mathcal{U}_{s_{0,N}} \hat{p}^0 \right] \\
&= \alpha \mathcal{U}_{t_0^{(N)}} p_{\text{data}} + \left[\alpha \sum_{m=0}^{N-1} (1-\alpha)^{m+1} \mathcal{U}_{t_0^{(N)}} \mathcal{U}_{\sigma_{m,N}} p_{\text{data}} + (1-\alpha)^{N+1} \mathcal{U}_{t_0^{(N)}} \mathcal{U}_{s_{0,N}} \hat{p}^0 \right]
\end{aligned}$$

But, by semigroup property of $\mathcal{U}_t$ (Proposition B.1, *(ii)*), , $\mathcal{U}_{t_0^{(N)}} \mathcal{U}_{\sigma_{m,N}} = \mathcal{U}_{t_0^{(N)} + \sigma_{m,N}}$ and similarly $\mathcal{U}_{t_0^{(N)}} \mathcal{U}_{s_{0,N}} = \mathcal{U}_{t_0^{(N)} + s_{0,N}}$. Observing that $t_0^{(N)} + \sigma_{m,N} = \sum_{\ell=N-m-1}^{N} t_0^{(\ell)} = \sigma_{m+1,N+1}$ and $t_0^{(N)} + s_{0,N} = s_{0,N+1}$ finally yields,

$$\hat{p}^{N+1} = \alpha \sum_{m=0}^{N} (1-\alpha)^m \mathcal{U}_{\sigma_{m,N+1}} p_{\text{data}} + (1-\alpha)^{N+1} \mathcal{U}_{s_{0,N+1}} p_{\text{data}}, \tag{57}$$

which is exactly (15) with N replaced by $N+1$. This proves the representation formula for all $N \geq 1$.

Asymptotic behavior when $t_0^{(i)} \to t_\infty$. Assume that $t_0^{(i)} \to t_\infty \in [0, \infty)$. We define the geometric weights $\pi_m(\alpha) = \alpha(1-\alpha)^m$ and the *annealed truncation limiting distribution* $p_\infty^{(t_\infty)}$ as,

$$p_\infty^{(t_\infty)} := \sum_{m=0}^{\infty} \pi_m(\alpha)\, \mathcal{U}_{(m+1)t_\infty} p_{\text{data}}.$$

The objective of this part is to show that $\mathcal{W}_2(\hat{p}^N, p_\infty^{(t_\infty)}) \underset{N\to\infty}{\longrightarrow} 0$. Observe that, for any fixed $m \geq 0$,

$$\sigma_{m,N} := \sum_{\ell=N-1-m}^{N-1} t_0^{(\ell)} \longrightarrow (m+1) t_\infty \quad \text{as } N \to \infty. \tag{58}$$

By continuity of $t \mapsto \mathcal{U}_t(p_{\text{data}})$ in $\mathcal{W}_2$ (Proposition B.1, *(vii)*),

$$\mathcal{U}_{\sigma_{m,N}}(p_{\text{data}}) \underset{N\to\infty}{\xrightarrow{\mathcal{W}_2}} \mathcal{U}_{(m+1)t_\infty}(p_{\text{data}}) \qquad \text{for each fixed } m. \tag{59}$$

We now compare $\hat{p}^N$ and $p_\infty^{(t_\infty)}$ as mixtures with the same total weights. To that end, we decompose $p_\infty^{(t_\infty)}$ as follows,

$$p_\infty^{(t_\infty)} = \sum_{m=0}^{N-1} \pi_m(\alpha)\, \mathcal{U}_{(m+1)t_\infty}(p_{\text{data}}) + (1-\alpha)^N r_N,$$

where

$$r_N := \frac{1}{(1-\alpha)^N} \sum_{m=N}^{\infty} \pi_m(\alpha) \mathcal{U}_{(m+1)t_\infty} p_{\text{data}} = \alpha \sum_{j=0}^{\infty} (1-\alpha)^j \mathcal{U}_{(N+j+1)t_\infty} p_{\text{data}}.$$

The measure r_N is a probability measure. By convexity of $\mathcal{W}_2^2$ under mixtures (Proposition B.6), we have

$$\mathcal{W}_2^2 \left(\hat{p}^N, p_\infty^{(t_\infty)} \right) \leq \sum_{m=0}^{N-1} \pi_m(\alpha) \mathcal{W}_2^2 \left(\mathcal{U}_{\sigma_{m,N}} p_{\text{data}}, \mathcal{U}_{(m+1)t_\infty} p_{\text{data}} \right) + (1-\alpha)^N \mathcal{W}_2^2 \left(\mathcal{U}_{s_{0,N}} p_{\text{data}}, r_N \right). \tag{60}$$

We treat the two terms separately.

First term. Fix $M \geq 0$, then

$$\sum_{m=0}^{N-1} \pi_m(\alpha) \mathcal{W}_2^2\big(\mathcal{U}_{\sigma_{m,N}} p_{\text{data}}, \mathcal{U}_{(m+1)t_\infty} p_{\text{data}}\big) \leq \sum_{m=0}^{M} \pi_m(\alpha) \mathcal{W}_2^2\big(\mathcal{U}_{\sigma_{m,N}} p_{\text{data}}, \mathcal{U}_{(m+1)t_\infty} p_{\text{data}}\big) + \sum_{m=M+1}^{N-1} \pi_m(\alpha) B,$$

where $B < \infty$ is a uniform second-moment bound. Proposition B.1 *(iv)* ensures that, for any t, $\mathcal{U}_t p_{\text{data}}$ has uniformly bounded second moment,

$$\sup_{t \geq 0} \int \|\mathbf{x}\|_2^2 \, \mathcal{U}_t p_{\text{data}}(\mathrm{d}\mathbf{x}) \leq \int \|\mathbf{x}\|_2^2 \, p_{\text{data}}(\mathrm{d}\mathbf{x}) + d, \tag{61}$$

Therefore, all the above pairwise $\mathcal{W}_2$-distances are bounded uniformly.

For fixed M, the finite sum converges to zero by the continuity invoked in (59). The tail satisfies

$$\sum_{m=M+1}^{N-1} \pi_m(\alpha) B = B \sum_{m=M+1}^{N-1} \alpha(1-\alpha)^m \leq B \sum_{m=M+1}^{\infty} \alpha(1-\alpha)^m = B(1-\alpha)^{M+1}.$$

which can be made arbitrarily small by choosing M large. Hence the first term in (60) converges to zero.

Second term. For the second term in (60), as stated in (61), $\mathcal{U}_{s_{0,N}} p_{\text{data}}$ has uniformly bounded second moments by Proposition B.1 *(iv)*. As r_N is a convex combination of measures of the form $\mathcal{U}_{t_j} p_{\text{data}}$ for $j \geq 0$, its second moment is also bounded. Therefore, there exists a constant $B_0 < \infty$ such that,

$$\mathcal{W}_2^2\big(\mathcal{U}_{s_{0,N}}(\hat{p}^0), r_N\big) \leq B_0$$

for all N. Since $(1-\alpha)^N \to 0$, the second term in (60) also converges to zero.

Combining the two estimates yields the desired convergence

$$\mathcal{W}_2\Big(\hat{p}^N, p_\infty^{(t_\infty)}\Big) \to 0.$$

Case $t_\infty = 0$. If $t_\infty = 0$, then $\mathcal{U}_{(m+1)t_\infty} = \mathcal{U}_0 = \mathrm{Id}$ for every m, and thus

$$p_\infty^{(0)} = \alpha \sum_{m=0}^{\infty} (1-\alpha)^m \, p_{\text{data}} = p_{\text{data}},$$

since $\sum_{m=0}^{\infty} \alpha(1-\alpha)^m = 1$. This proves that the recursive compounding effect disappears asymptotically when $t_0^{(i)} \to 0$. □

D.7 Proof of the finite-N schedule bound

Proof. Let $h(t) := \min\big\{d\,t + \frac{t^2}{4}\,\mathbb{E}_{p_{\text{data}}}\|\mathbf{X}\|^2,\ 4\mathcal{W}_2^2(p_{\text{data}}, \gamma)\big\}$, where the two expressions inside the minimum come from the following bounds, valid for every $t \geq 0$:

1. By the proof of Proposition B.1(vii) together with $1 - e^{-x} \leq x$

$$\mathcal{W}_2^2(\mathcal{U}_t p_{\text{data}}, p_{\text{data}}) \leq (1 - e^{-t/2})^2 \, \mathbb{E}_{p_{\text{data}}}\|\mathbf{X}\|^2 + d\,(1 - e^{-t});$$

2. By the proof of Corollary 3.2,

$$\mathcal{W}_2(\mathcal{U}_t p_{\text{data}}, p_{\text{data}}) \leq \mathcal{W}_2(\mathcal{U}_t p_{\text{data}}, \gamma) + \mathcal{W}_2(\gamma, p_{\text{data}}) \leq (e^{-t/2} + 1)\, \mathcal{W}_2(p_{\text{data}}, \gamma).$$

Hence $\mathcal{W}_2^2(\mathcal{U}_t p_{\text{data}}, p_{\text{data}}) \le h(t)$. Decomposing $p_{\text{data}} = \alpha \sum_{m=0}^{N-1} (1-\alpha)^m \, p_{\text{data}} + (1-\alpha)^N p_{\text{data}}$ and applying the convexity of $\mathcal{W}_2^2$ under mixtures (Proposition B.6) term by term against the representation (15) of Theorem 4.1 yields

$$\mathcal{W}_2^2(\hat{p}^N, p_{\text{data}}) \;\le\; \alpha \sum_{m=0}^{N-1} (1-\alpha)^m \, h(\sigma_{m,N}) \;+\; (1-\alpha)^N \, h(s_{0,N}).$$

□

D.8 Proof of Theorem 5.2

Proof. The proof proceeds in three steps: a one-step recursion, its iteration by induction, and verification of the two regimes.

(a) One-step Recursion.

By the triangle inequality,

$$D_{N+1} = \mathcal{W}_2(\hat{p}^{N+1}, p_\infty^\star) \le \mathcal{W}_2(\hat{p}^{N+1},\, \mathcal{T}(\hat{p}^N)) + \mathcal{W}_2(\mathcal{T}(\hat{p}^N),\, p_\infty^\star)$$

For the first term: since $\hat{p}^{N+1} = \hat{\mathcal{S}}_N(q_N)$ (imperfect sampler) and $\mathcal{T}(\hat{p}^N) = \mathcal{U}_{t_0}(q_N)$ (ideal sampler applied to the same training mixture q_N), the definition of δ_N in (16) gives

$$\mathcal{W}_2(\hat{p}^{N+1},\, \mathcal{T}(\hat{p}^N)) = \mathcal{W}_2(\hat{\mathcal{S}}_N(q_N),\, \mathcal{U}_{t_0}(q_N)) \le \delta_N.$$

For the second term: since $p_\infty^\star = \mathcal{T}(p_\infty^\star)$, we need to bound $\mathcal{W}_2(\mathcal{T}(\hat{p}^N),\, \mathcal{T}(p_\infty^\star))$. But $\mathcal{T}$ is a contraction (Proposition C.1) thus,

$$\mathcal{W}_2(\mathcal{T}(\hat{p}^N),\, p_\infty^\star) = \mathcal{W}_2(\mathcal{T}(\hat{p}^N),\, \mathcal{T}(p_\infty^\star)) \le \sqrt{1-\alpha}\, e^{-t_0/2}\, D_N = \kappa\, D_N,$$

where $\kappa := \sqrt{1-\alpha}\, e^{-t_0/2} < 1$. Combining the two bounds:

$$D_{N+1} \le \kappa\, D_N + \delta_N, \qquad N \ge 0. \tag{62}$$

(b) Full Induction.

We prove (17) by induction on N. The base case $N = 1$ follows directly from (62): $D_1 \le \kappa\, D_0 + \delta_0$. Assume the bound holds at generation N:

$$D_N \le \kappa^N\, D_0 + \sum_{i=0}^{N-1} \kappa^{N-1-i}\, \delta_i.$$

Then, substituting into (62),

$$D_{N+1} \le \kappa\, D_N + \delta_N \le \kappa\Big(\kappa^N\, D_0 + \sum_{i=0}^{N-1} \kappa^{N-1-i}\, \delta_i\Big) + \delta_N = \kappa^{N+1}\, D_0 + \sum_{i=0}^{N} \kappa^{N-i}\, \delta_i,$$

which completes the induction and establishes (17).

We now prove the two asymptotic regimes of Theorem 5.2.

(i) Summable perturbations. Assume $\sum_{i\ge 0} \delta_i < \infty$, which implies $\delta_N \to 0$ as $N \to \infty$. The first term $\kappa^N\, D_0 \to 0$. For the convolution term, fix $\varepsilon > 0$ and choose M such that $\sum_{i=M}^{\infty} \delta_i < \varepsilon$. For $N > M$, split:

$$\sum_{i=0}^{N-1} \kappa^{N-1-i}\, \delta_i = \underbrace{\sum_{i=0}^{M-1} \kappa^{N-1-i}\, \delta_i}_{\to\, 0 \text{ as } N\to\infty \,(M \text{ fixed},\, \kappa<1)} + \underbrace{\sum_{i=M}^{N-1} \kappa^{N-1-i}\, \delta_i}_{\le \sum_{i=M}^{N-1} \delta_i \le \sum_{i=M}^{\infty} \delta_i < \varepsilon}.$$

Since ε is arbitrary, $D_N \to 0$.

(ii) Constant perturbation floor. If $\delta_i < \delta$, the geometric sum evaluates to $\sum_{i=0}^{N-1} \kappa^{N-1-i} = \sum_{j=0}^{N-1} \kappa^j = (1-\kappa^N)/(1-\kappa)$, so

$$D_N \le \kappa^N\, D_0 + \frac{\delta\,(1-\kappa^N)}{1-\kappa} \xrightarrow{N\to\infty} \frac{\delta}{1-\kappa}.$$

□

D.9 Explicit characterization of perturbation

Proposition D.3 (Linear perturbation recursion)**.** *Define the perturbation $\xi_N := \hat{p}^N - p^\star_\infty$ and, for each generation i, let $\eta_i := \hat{p}^{i+1} - \mathcal{U}_{t_0}(q_i)$ be the* one-step perturbation *at generation i. Then, the imperfect distribution $\hat{p}^N$ satisfies:*

$$\hat{p}^N \;=\; p^\star_\infty \;+\; \underbrace{(1-\alpha)^N \mathcal{U}_{Nt_0}(\xi_0)}_{\textit{initialization transient}} \;+\; \underbrace{\sum_{i=0}^{N-1} (1-\alpha)^{N-1-i} \mathcal{U}_{(N-1-i)t_0}(\eta_i)}_{\textit{accumulated perturbations}}. \tag{63}$$

Proof. Define the perturbation $\xi_N := \hat{p}^N - p^\star_\infty$. We first prove it satisfies the following linear recursion:

$$\xi_{N+1} \;=\; (1-\alpha)\,\mathcal{U}_{t_0}(\xi_N) \;+\; \eta_N. \tag{64}$$

Indeed, by definition of η_N, we have $\hat{p}^{N+1} = \mathcal{U}_{t_0}(q_N) + \eta_N$. Since $p^\star_\infty = \mathcal{U}_{t_0}(\alpha\, p_{\text{data}} + (1-\alpha)p^\star_\infty)$, subtracting and using the linearity of $\mathcal{U}_{t_0}$ on signed measures, we obtain the following:

$$\begin{aligned}
\xi_{N+1} &= \hat{p}^{N+1} - p^\star_\infty \\
&= \mathcal{U}_{t_0}(q_N) - \mathcal{U}_{t_0}(\alpha\, p_{\text{data}} + (1-\alpha)p^\star_\infty) + \eta_N \\
&= \mathcal{U}_{t_0}(\alpha p_{\text{data}} + (1-\alpha)\hat{p}^N) - \mathcal{U}_{t_0}(\alpha\, p_{\text{data}} + (1-\alpha)p^\star_\infty) + \eta_N \\
&= \alpha\mathcal{U}_{t_0}(p_{\text{data}}) + (1-\alpha)\mathcal{U}_{t_0}(\hat{p}^N) - \alpha\mathcal{U}_{t_0}(p_{\text{data}}) - (1-\alpha)\mathcal{U}_{t_0}(p^\star_\infty) + \eta_N \qquad \text{by linearity of } \mathcal{U}_{t_0} \\
&= \mathcal{U}_{t_0}\big((1-\alpha)(\hat{p}^N - p^\star_\infty)\big) + \eta_N \qquad \text{by linearity of } \mathcal{U}_{t_0} \\
&= (1-\alpha)\,\mathcal{U}_{t_0}(\xi_N) + \eta_N.
\end{aligned}$$

Unrolling the recursion and applying the semigroup property $\mathcal{U}^k_{t_0} = \mathcal{U}_{kt_0}$ yields the end result:

$$\hat{p}^N \;=\; p^\star_\infty \;+\; (1-\alpha)^N \mathcal{U}_{Nt_0}(\xi_0) \;+\; \sum_{i=0}^{N-1} (1-\alpha)^{N-1-i} \mathcal{U}_{(N-1-i)t_0}(\eta_i).$$

□

Equation (63) reveals that each iterate $\hat{p}^N$ decomposes as $p^\star_\infty$ plus two correction terms. The initialization transient decays geometrically *and* is progressively smoothed by Nt_0 units of OU evolution. Each perturbation η_i introduced at generation i is subsequently smoothed by $(N-1-i)t_0$ units of OU evolution before being observed at generation N, with amplitude damped by $(1-\alpha)^{N-1-i}$.

D.10 Proof of Proposition 5.3

Proof. In Proposition D.3, we explicitly wrote $\hat{p}^N$ in presence of errors as,

$$\hat{p}^N \;=\; p^\star_\infty \;+\; (1-\alpha)^N \mathcal{U}_{Nt_0}(\xi_0) \;+\; \sum_{i=0}^{N-1} (1-\alpha)^{N-1-i} \mathcal{U}_{(N-1-i)t_0}(\eta_i),$$

where,

$$\xi_N = \hat{p}^N - p^\star_\infty, \quad \text{and} \quad \eta_N = \hat{p}^{N+1} - \mathcal{U}_{t_0}(q_N).$$

Now, projecting perturbation ξ_N onto the Hermite polynomials basis $H_{\mathbf{n}}$, since $\mathcal{U}_{t_0}$ acts on mode $\mathbf{n}$ by multiplication by $e^{-|\mathbf{n}|t_0/2}$ (Proposition Prop. B.4):

$$\begin{aligned}
\langle \xi_{N+1}, H_{\mathbf{n}} \rangle &= \langle (1-\alpha)\mathcal{U}_{t_0}(\xi_N) + \eta_N, H_{\mathbf{n}} \rangle \\
&= (1-\alpha)\, e^{-|\mathbf{n}|t_0/2} \langle \xi_N, H_{\mathbf{n}} \rangle + \langle \eta_N, H_{\mathbf{n}} \rangle \\
&= \kappa_{\mathbf{n}} \langle \xi_N, H_{\mathbf{n}} \rangle + \langle \eta_N, H_{\mathbf{n}} \rangle.
\end{aligned}$$

Unrolling gives (18). □

E Self-Regularization and Annealed Truncation

In this section, we investigate the well-posedness of annealed truncation schedules, which consists in considering adaptive truncation times $t_0^{(i)}$ that converge to 0 as the generation $i \to \infty$.

E.1 Fisher Information Instability and Properties

Instability of the reverse diffusion around $t \downarrow 0$ We first formally state the instability problem that arises near the reverse diffusion endpoint $t = 0$. Recall that, in order to sample from the reverse diffusion, one needs to estimate the score of the reverse time-t marginal at all times t. For a target distribution q_i at generation i, we denote the score of the reverse time-t marginal by $\nabla \log q_{i,t}$. The key quantity that captures how hard these are to estimate is the Fisher information of $q_{i,t}$ [61], defined as the expected squared magnitude of the score:

$$J(q_{i,t}) = \int \|\nabla \log q_{i,t}(\mathbf{x})\|_2^2 \, q_{i,t}(\mathbf{x}) \mathrm{d}\mathbf{x}. \tag{65}$$

In diffusion training, the score matching objective [38, 39, 68] at generation i contains terms of the form,

$$\mathbb{E}_{\mathbf{x}_t \sim q_{i,t}} \big[\|\mathbf{s}_\theta(\mathbf{X}_t, t) - \nabla \log q_{i,t}(\mathbf{X}_t)\|_2^2 \big], \tag{66}$$

suggesting that if $J(q_{i,t})$ defined in (65) is large, the target score itself (66) has large L^2-norm.

Concretely, for OU diffusion, one has

$$\mathcal{U}_t q_i = \mathrm{Law}(\mathbf{X}_t^i) = \mathrm{Law}(e^{-t/2}\mathbf{X}_0^i + \sqrt{1 - e^{-t}}\mathbf{Z})$$

where $\mathbf{Z} \sim \mathcal{N}(0, \mathbf{I}_d)$ and $\mathbf{X}_0^i \sim q_i$. Therefore $\mathcal{U}_t q_i$ approaches the raw data distribution q_i as $t \to 0$. If the training distribution is empirical, singular, sharp, multimodal or concentrated near a low-dimensional set, then its density develops steep gradients at small t. The generic Fisher bound captures exactly this ((ii) in Proposition E.1) :

$$J(\mathcal{U}_t q_i) \le \frac{d}{1 - e^{-t}} \underset{t \to 0}{\sim} \frac{d}{t}, \qquad \text{since } 1 - e^{-t} \underset{t \to 0}{\sim} t.$$

Proposition E.1 (Fisher control for OU smoothing)**.** *Let $\mu \in \mathcal{P}(\mathbb{R}^d)$ be absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$. Denote by $J(\mu)$ its Fisher information,*

$$J(\mu) = \int_{\mathbb{R}^d} \|\nabla \log \mu(\mathbf{x})\|^2 \mu(\mathbf{x}) \, \mathrm{d}\mathbf{x} = \int_{\mathbb{R}^d} \frac{\|\nabla \mu(\mathbf{x})\|^2}{\mu(\mathbf{x})} \, \mathrm{d}\mathbf{x}. \tag{67}$$

We equivalently write $J(\mathbf{X})$ to denote $J(\mu)$ when $\mathbf{X} \sim \mu$. Let $(\mathcal{U}_t)_{t \ge 0}$ denote the OU sampling operator (defined in (24)) Then the following properties hold.

(i) ***Quadratic scaling under dilation.*** *For any $a \neq 0$,*

$$J(a\mathbf{X}) = a^{-2} J(\mathbf{X}).$$

(ii) ***Universal OU smoothing bound.*** *For every $t > 0$,*

$$J(\mathcal{U}_t \mu) \le \frac{d}{1 - e^{-t}}.$$

(iii) ***Fisher control under OU evolution.*** *If $J(\mu) < \infty$, then for every $t \ge 0$,*

$$J(\mathcal{U}_t \mu) \le e^t J(\mu).$$

In particular, for t in a bounded interval $[0, T]$,

$$\sup_{0 \le t \le T} J(\mathcal{U}_t \mu) \le e^T J(\mu).$$

(iv) ***Convexity under mixtures.*** *Let $\mu_1, \dots, \mu_N$ have smooth positive densities, and let $w_1, \dots, w_N \ge 0$ satisfy $\sum_{k=1}^N w_k = 1$. Then*

$$J\left(\sum_{k=1}^N w_k \mu_k\right) \le \sum_{k=1}^N w_k J(\mu_k).$$

These results are standard, and detailed proofs can be found in [65, 42]. We provide proofs for completeness.

Proof. (i) Quadratic scaling under dilation. Let $\mathbf{Y} = a\mathbf{X}$ with $a \neq 0$. If $\mathbf{X}$ has density μ, then $\mathbf{Y}$ has density

$$\mu_a(\mathbf{y}) = |a|^{-d}\mu(\mathbf{y}/a).$$

Therefore,

$$\nabla_{\mathbf{y}} \log \mu_a(\mathbf{y}) = \nabla_{\mathbf{y}} \log \mu(\mathbf{y}/a) = a^{-1}\nabla \log \mu(\mathbf{y}/a).$$

Hence,

$$J(\mathbf{Y}) = \int_{\mathbb{R}^d} \|\nabla \log \mu_a(\mathbf{y})\|^2 \mu_a(\mathbf{y})\, \mathrm{d}\mathbf{y} = \int_{\mathbb{R}^d} \left\|a^{-1}\nabla \log \mu(\mathbf{y}/a)\right\|^2 |a|^{-d}\mu(\mathbf{y}/a)\, \mathrm{d}\mathbf{y}. \tag{68}$$

By change of variables $\mathbf{y} = a\mathbf{x}$, (68) yields the desired results:

$$J(\mathbf{Y}) = a^{-2}\int_{\mathbb{R}^d} \|\nabla \log \mu(\mathbf{x})\|^2 \mu(\mathbf{x})\, \mathrm{d}\mathbf{x} = a^{-2}J(\mathbf{X}).$$

(ii) Universal OU smoothing bound. Fix $\mu \in \mathcal{P}(\mathbb{R}^d)$ and let $\mathbf{Y}_t = e^{-t/2}\mathbf{X} + \sqrt{1-e^{-t}}\mathbf{Z}$ where $\mathbf{X} \sim \mu \perp\!\!\!\perp \mathbf{Z} \sim \mathcal{N}(0, \mathbf{I}_d)$. Also, let $\sigma_t^2 = 1 - e^{-t}$. Using the standard Fisher-information inequality [21, 42] $J(\mathbf{X}_1 + \mathbf{X}_2) \leq J(\mathbf{X}_2)$ for any two independent random variables $\mathbf{X}_1$ and $\mathbf{X}_2$, to $\mathbf{X}_1 = e^{-t/2}\mathbf{X}$ and $\mathbf{X}_2 = \sigma_t\mathbf{Z}$ gives,

$$J(\mathbf{Y}_t) \leq J(\sigma_t\mathbf{Z}).$$

Since $\sigma_t\mathbf{Z} \sim \mathcal{N}(0, \sigma_t^2\mathbf{I}_d)$, its Fisher information is

$$J(\sigma_t\mathbf{Z}) = \frac{d}{\sigma_t^2} = \frac{d}{1-e^{-t}}.$$

Therefore,

$$J(\mathcal{U}_t\mu) = J(\mathbf{Y}_t) \leq \frac{d}{1-e^{-t}}.$$

(iii) Fisher control under OU evolution. Once again, let $\mathbf{Y}_t$ be defined as $\mathbf{Y}_t = e^{-t/2}\mathbf{X} + \sqrt{1-e^{-t}}\mathbf{Z}$. By the same monotonicity property Fisher information used in *(ii)*,

$$J(\mathbf{Y}_t) \leq J(e^{-t/2}\mathbf{X}).$$

Using the scaling property proved in *(i)* with $a = e^{-t/2}$,

$$J(e^{-t/2}\mathbf{X}) = e^t J(\mathbf{X}).$$

Therefore,

$$J(\mathcal{U}_t\mu) = J(\mathbf{Y}_t) \leq e^t J(\mu).$$

If $t \in [0, T]$, then $e^t \leq e^T$, and hence

$$\sup_{0\leq t\leq T} J(\mathcal{U}_t\mu) \leq e^T J(\mu).$$

(iv) Convexity under mixtures. Let $\mu_1, \ldots, \mu_N$ denote both the probability measures and their smooth positive densities and define the convex mixture distribution $\mu = \sum_{k=1}^N w_k\mu_k$. Then, by taking the gradient of μ,

$$\nabla\mu = \sum_{k=1}^N w_k \nabla\mu_k.$$

Noticing that the map $(u, \mathbf{v}) \mapsto \frac{\|\mathbf{v}\|^2}{u}$ for $u > 0$ and $\mathbf{v} \in \mathbb{R}^d$ is convex, one has pointwise in $\mathbf{x}$ that

$$\frac{\|\nabla\mu(\mathbf{x})\|^2}{\mu(\mathbf{x})} = \frac{\left\|\sum_{k=1}^N w_k\nabla\mu_k(\mathbf{x})\right\|^2}{\sum_{k=1}^N w_k\mu_k(\mathbf{x})} \leq \sum_{k=1}^N w_k \frac{\|\nabla\mu_k(\mathbf{x})\|^2}{\mu_k(\mathbf{x})}.$$

Integrating over $\mathbb{R}^d$ gives the final result,

$$J(\mu) = \int_{\mathbb{R}^d} \frac{\|\nabla\mu(\mathbf{x})\|^2}{\mu(\mathbf{x})}\, \mathrm{d}\mathbf{x} \leq \sum_{k=1}^N w_k \int_{\mathbb{R}^d} \frac{\|\nabla\mu_k(\mathbf{x})\|^2}{\mu_k(\mathbf{x})}\, \mathrm{d}\mathbf{x} = \sum_{k=1}^N w_k J(\mu_k).$$

□

E.2 Annealed Truncation

The instability of score estimation near $t = 0$ highlights the limits of considering truncation times $t_0^{(i)}$ that tend to zero. The recursive structure partially mitigates this problem, because synthetic samples have already passed through previous OU smoothing steps. However, it does not imply that the whole synthetic distribution has accumulated the full cumulative smoothing $\sum_{j<i} t_0^{(j)}$ since, in our setting, fresh data are reintroduced at each generation, and recently injected components have only been smoothed for the most recent truncation times. The following decomposition makes this precise.

Proposition E.2 (Fisher decomposition under annealed truncation)**.** *Consider the error-free annealed recursion with schedules* $(t_0^{(i)})_{i\geq 0}$*,*

$$\hat{p}^{i+1} = \mathcal{U}_{t_0^{(i)}}\big(\alpha p_{\mathrm{data}} + (1-\alpha)\hat{p}^i\big), \qquad \hat{p}^0 = p_{\mathrm{data}}, \tag{69}$$

where $t_0^{(i)} > 0$*. Then, for each generation* $N \geq 1$*,*

$$\begin{aligned} J(q_{N,t_0^{(N)}}) \leq \alpha J(\mathcal{U}_{t_0^{(N)}} p_{\mathrm{data}}) + (1-\alpha)\alpha \sum_{m=0}^{N-1} (1-\alpha)^m J(\mathcal{U}_{t_0^{(N)}+\sigma_{m,N}} p_{\mathrm{data}}) \\ + (1-\alpha)^{N+1} J(\mathcal{U}_{t_0^{(N)}+s_{0,N}} p_{\mathrm{data}}), \end{aligned} \tag{70}$$

where, as in Theorem 4.1, $\sigma_{m,N} := \sum_{\ell=N-1-m}^{N-1} t_0^{(\ell)}$ *is the smoothing accumulated by the fresh data injected at generation* $N-1-m$*, and* $s_{0,N} := \sum_{\ell=0}^{N-1} t_0^{(\ell)} = \sigma_{N-1,N}$*.*

Proof. By Theorem 4.1, for any finite horizon $N \geq 1$, $\hat{p}^N$ can be written as

$$\hat{p}^N = \alpha \sum_{m=0}^{N-1} (1-\alpha)^m \mathcal{U}_{\sigma_{m,N}} p_{\mathrm{data}} + (1-\alpha)^N \mathcal{U}_{s_{0,N}} p_{\mathrm{data}}, \tag{71}$$

For any $t \geq 0$, $q_{N,t} := \mathcal{U}_t q_N$ can be written as

$$q_{N,t} = \mathcal{U}_t(\alpha p_{\mathrm{data}} + (1-\alpha)\hat{p}^N) = \alpha \mathcal{U}_t p_{\mathrm{data}} + (1-\alpha)\mathcal{U}_t \hat{p}^N,$$

where we have used the linearity of $\mathcal{U}_t$ (Proposition B.1). Thus, plugging the expression of $\hat{p}^N$ given in (71) and using linearity again, we obtain

$$\begin{aligned} q_{N,t} &= \alpha \mathcal{U}_t p_{\mathrm{data}} + (1-\alpha)\mathcal{U}_t \hat{p}^N \\ &= \alpha \mathcal{U}_t p_{\mathrm{data}} + \alpha \sum_{m=0}^{N-1} (1-\alpha)^{m+1} \mathcal{U}_{t+\sigma_{m,N}} p_{\mathrm{data}} + (1-\alpha)^{N+1} \mathcal{U}_{t+s_{0,N}} p_{\mathrm{data}}. \end{aligned} \tag{72}$$

The Fisher bound (70) follows from (72) using the convexity of Fisher information under mixtures (Proposition E.1*(iv)*) and taking $t = t_0^{(N)}$. □

Proposition E.2 clearly differentiates between two terms. On the one hand, components injected many generations in the past have accumulated several steps of smoothing and are regularized by the recursion. On the other hand, components injected more recently have less smoothing, and the fresh component $\alpha \mathcal{U}_t p_{\mathrm{data}}$ has only been smoothed by the current truncation time $t_0^{(N)}$. Therefore, recursive self-regularization does not by itself give a uniform Fisher bound as $t \downarrow 0$ for arbitrary raw data. We distinguish the settings in which these annealed truncations are well-defined.

E.2.1 Case 1: Under A3 – finite Fisher information

Under the standing assumption A3, annealed truncation schedules are well-defined since the unknown target distribution p_{data} has finite Fisher information.

Corollary E.3 (Annealing under finite-Fisher regularity)**.** *If* p_{data} *has finite Fisher information, i.e. if* $J(p_{\mathrm{data}}) < \infty$*, then, for a bounded positive truncation schedule* $(t_0^{(N)})_{N\geq 0}$*,*

$$\sup_N J\big(q_{N,t_0^{(N)}}\big) < \infty.$$

Proof. Let $T_0 := \sup_N t_0^{(N)} < \infty$. By (72), evaluated at $t = t_0^{(N)}$, we have

$$q_{N,t_0^{(N)}} = \alpha \mathcal{U}_{t_0^{(N)}} p_{\mathrm{data}} + \alpha \sum_{m=0}^{N-1} (1-\alpha)^{m+1} \mathcal{U}_{t_0^{(N)}+\sigma_{m,N}} p_{\mathrm{data}} + (1-\alpha)^{N+1} \mathcal{U}_{t_0^{(N)}+s_{0,N}} p_{\mathrm{data}}.$$

The coefficients in this decomposition are non-negative and sum to one. Hence, by convexity of Fisher information under mixtures,

$$\begin{aligned} J(q_{N,t_0^{(N)}}) \le \alpha J(\mathcal{U}_{t_0^{(N)}} p_{\mathrm{data}}) + \alpha \sum_{m=0}^{N-1} (1-\alpha)^{m+1} J(\mathcal{U}_{t_0^{(N)}+\sigma_{m,N}} p_{\mathrm{data}}) \\ + (1-\alpha)^{N+1} J(\mathcal{U}_{t_0^{(N)}+s_{0,N}} p_{\mathrm{data}}). \end{aligned} \tag{73}$$

We now show that every term appearing on the right-hand side is uniformly bounded. Since $J(p_{\mathrm{data}}) < \infty$, Proposition E.1(iii) implies that, for $0 \le s \le T_0$,

$$J(\mathcal{U}_s p_{\mathrm{data}}) \le e^{T_0} J(p_{\mathrm{data}}).$$

On the other hand, for $s \ge T_0$, Proposition E.1(ii) gives the universal smoothing bound

$$J(\mathcal{U}_s p_{\mathrm{data}}) \le \frac{d}{1-e^{-s}} \le \frac{d}{1-e^{-T_0}}.$$

Consequently, for all $s \ge 0$,

$$J(\mathcal{U}_s p_{\mathrm{data}}) \le C_{T_0,p_{\mathrm{data}}} := \max\left\{ e^{T_0} J(p_{\mathrm{data}}), \frac{d}{1-e^{-T_0}} \right\}.$$

Applying this bound to each component in the convex decomposition (73) yields

$$J(q_{N,t_0^{(N)}}) \le C_{T_0,p_{\mathrm{data}}}$$

for every N, which yields the result. □

Remark E.4 (Interpretation of the finite Fisher information assumption ($J(p_{\mathrm{data}}) < \infty$))**.** The condition $J(p_{\mathrm{data}}) < \infty$ is the regularity assumption A3 of the main text. It fails on raw empirical measures which have no Lebesgue densities and hence no classical scores, and may also not be satisfied for idealized distributions supported on low-dimensional sets. Hence, in such cases, annealing the truncation all the way to zero is not justified by Fisher information control alone.

E.2.2 Case 2: Beyond A3 – target distribution without finite Fisher information

We distinguish between two alternatives in the case where $J(p_{\mathrm{data}}) < \infty$ fails.

1. *Strictly positive truncation floor.* A first option is to consider truncation schedules $(t_0^{(N)})_{N\ge 0}$ that are decreasing down to a small positive floor $t_\infty > 0$ for which Fisher scores remain bounded independently of the regularity of p_{data}, by ((ii)-Proposition E.1),

$$J(\mathcal{U}_{t_\infty} p_{\mathrm{data}}) \le \frac{d}{1-e^{-t_\infty}}$$

2. *Coupled annealing of fresh data and truncation.* A way to remove the finite-Fisher-information requirement on p_{data} while still letting the truncation time vanish is to anneal the fresh-data proportion together with the truncation schedule. Consider the generation-dependent recursion

$$\hat{p}^{N+1} = \mathcal{U}_{t_0^{(N)}}\Big(\alpha_N p_{\mathrm{data}} + (1-\alpha_N)\hat{p}^N\Big), \qquad t_0^{(N)} \downarrow 0, \qquad \alpha_N \downarrow 0.$$

For any generation $N \ge 1$, linearity of $\mathcal{U}_{t_0^{(N)}}$ (Proposition B.1) and convexity of Fisher information (Proposition E.1),

$$\begin{aligned} J(\hat{p}^{N+1}) &= J\Big(\alpha_N \mathcal{U}_{t_0^{(N)}} p_{\mathrm{data}} + (1-\alpha_N)\mathcal{U}_{t_0^{(N)}} \hat{p}^N\Big) \\ &\le \alpha_N J(\mathcal{U}_{t_0^{(N)}} p_{\mathrm{data}}) + (1-\alpha_N) J(\mathcal{U}_{t_0^{(N)}} \hat{p}^N). \end{aligned}$$

The first term is the only term involving freshly injected, potentially irregular data. By the universal OU smoothing bound ((ii) in Proposition E.1),

$$\alpha_N J(\mathcal{U}_{t_N} p_{\mathrm{data}}) \le \alpha_N \frac{d}{1 - e^{-t_N}}.$$

Consequently, the fresh-data contribution remains uniformly bounded provided

$$\sup_{N \ge 0} \frac{\alpha_N}{1 - e^{-t_N}} < \infty,$$

or equivalently, since $1 - e^{-t_N} \sim t_N$ as $t_N \to 0$,

$$\alpha_N = \mathcal{O}(t_N) \qquad \text{as } t_N \downarrow 0.$$

This condition controls the newly injected raw-data component. To obtain a uniform Fisher bound for the whole training distribution, one must additionally control the synthetic term

$$(1 - \alpha_N) J(\mathcal{U}_{t_N} \hat{p}^N),$$

either by an induction argument, by an explicit unrolling of the recursion, or by imposing a stronger global summability condition on the whole sequence of weights and accumulated smoothing times. Thus, coupled annealing of α_N and t_N provides an alternative to $J(p_{\mathrm{data}}) < \infty$ for the fresh component, but full uniform Fisher control requires the same convexity argument to be applied to all mixture components.

F Experiments

We validate our theoretical results on (i) synthetic 2D Gaussian mixtures, where the limiting distribution $p^\star_\infty$ (Theorem 3.1) can be computed in closed form and serves as a controlled environment for Corollary 3.2; (ii) CIFAR-10 [47] with fixed truncation, where we test the geometric rate of convergence to $p^\star_\infty$ (Theorems 3.1 and 5.2) and the spectral structure of collapse (Proposition 3.4); and (iii) CIFAR-10 with annealed truncation schedules, where we test the recovery of p_{data} predicted by Theorem 4.1. All synthetic two-dimensional experiments were run on CPU workers, whereas image experiments (on CIFAR-10) were run on a single NVIDIA RTX 6000 Blackwell GPU with 96GB memory; a full 8-generation CIFAR-10 recursive training run required approximately 12 hours of wall-clock time.

F.1 Synthetic 2D Gaussian mixtures

We set p_{data} to a three-component isotropic mixture in $\mathbb{R}^2$ and run the error-free recursion (8) for 50 generations at $\alpha \in \{0.1, 0.5, 0.9\}$, using a Variance Preserving SDE [64] with $\beta \equiv 1$ so that the OU semigroup matches our analysis exactly and the contraction constant $\kappa = \sqrt{1-\alpha}\, e^{-t_0/2}$ holds without approximation. For p_{data} a Gaussian mixture, $p^\star_\infty$ in (9) is itself a countably-infinite mixture of Gaussians that can be sampled from KL divergences and admit accurate Monte-Carlo estimators.

Limiting behaviors (Corollary 3.2). Figure 4 sweeps $\alpha \in [0.02, 0.98]$ at fixed t_0 (left and center panels) and $t_0 \in [0, 5]$ at fixed α (right panel), computing $p^\star_\infty$ analytically at each grid point. The three predicted limits are recovered: $\mathrm{KL}(p^\star_\infty \,\|\, \mathcal{U}_{t_0} p_{\mathrm{data}}) \to 0$ as $\alpha \to 1$, $\mathrm{KL}(p^\star_\infty \,\|\, \mathcal{N}(0, \mathbf{I})) \to 0$ as $\alpha \to 0$, and $\mathrm{KL}(p^\star_\infty \,\|\, p_{\mathrm{data}}) \to 0$ as $t_0 \to 0$, with strict positivity at every interior (α, t_0).

Truncation schedules (Theorem 4.1). Figure 2 compares four truncation schedules over 50 generations at $\alpha \in \{0.1, 0.5, 0.9\}$: a fixed schedule $t_0^{(i)} \equiv t_0$, a shifted schedule with $t_0^{(i)} \to t_\infty = 0.2 > 0$,and two annealed schedules $t_0^{(i)} = t_0/(1+i)^\beta$ with $\beta \in \{1, 2\}$, both converging to 0 (only the latter is summable; Theorem 4.1 requires only $t_0^{(i)} \to 0$). The fixed schedule plateaus at the predicted irreducible floor $\mathrm{KL}(p^\star_\infty, p_{\mathrm{data}})$, and the shifted schedule plateaus at the lower predicted limit $\mathrm{KL}(p^{(t_\infty)}_\infty, p_{\mathrm{data}})$, both consistent with the asymptotic statement of Theorem 4.1. The two annealed schedules drive the recursion back toward p_{data}, with the decay rate increasing in β, confirming that recursive compounding is asymptotically eliminated whenever $t_0^{(i)} \to 0$. The qualitative behavior is consistent across α, with smaller α producing slower convergence, as expected from the contraction constant $\kappa = \sqrt{1-\alpha}\, e^{-t_0/2}$.

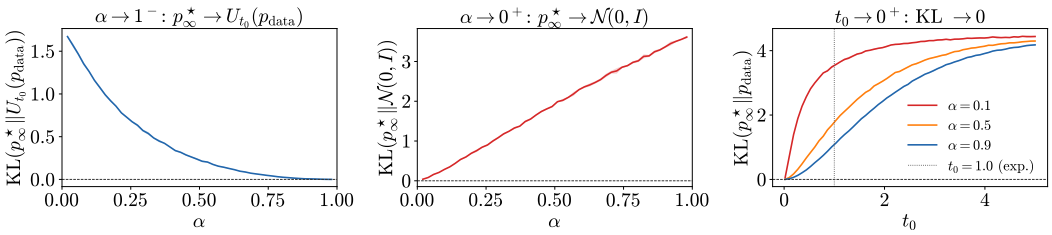


Figure 4: **Limiting behaviours of $p^\star_\infty$ on a 2D Gaussian mixture.** The closed-form Neumann-series representation of $p^\star_\infty$ (Theorem 3.1) recovers the three predicted limits of Corollary 3.2: convergence to a single OU-smoothed copy of p_{data} as $\alpha \to 1$ (i), convergence to the standard Gaussian as $\alpha \to 0$ (ii), and recovery of p_{data} as $t_0 \to 0$ (iii); strict positivity at all interior (α, t_0) is visible in panel (iii).

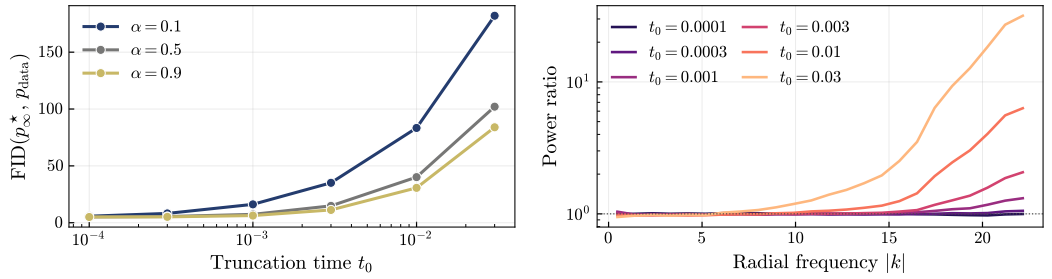


Figure 5: **Truncation time controls the severity of collapse (CIFAR-10 [47]).** *Left*: FID between the limiting collapse distribution $p^\star_\infty$ and the data distribution p_{data} as a function of the reverse-time truncation t_0, for several fresh-data proportions α. Increasing t_0 amplifies the bias induced by each recursive generation, and the effect is strongest when α is small, i.e. when synthetic samples dominate the next training distribution. *Right*: absolute radial Fourier power of $p^\star_\infty$, divided by that of p_{data}, for increasing values of t_0. Since the ratio is not normalized to unit total spectral mass, values above one indicate excess broadband power induced by the truncation noise rather than a redistribution of fixed spectral energy. As $t_0 \to 0$, the ratio approaches one, consistently with the recovery of p_{data}.

F.2 CIFAR-10: limiting collapse and spectral signature

Setup. We use a U-Net [55] score network with self-attention [66] at the two coarsest spatial scales, trained with the min-SNR-γ objective [34] ($\gamma = 5$) and EMA decay 0.9999. Each generation is trained for 300 epochs at $\mathrm{lr} = 2 \times 10^{-4}$ with cosine decay, batch size 128, and AdamW [50] (weight decay 10^{-2}). The truncation time is fixed at $t_0 = 10^{-3}$. At generation $i \geq 1$, the training loader returns each example with probability α from CIFAR-10 and probability $1 - \alpha$ from a pool of 50,000 samples drawn from generation $i - 1$ via DPM-Solver++ [51] with 50 steps. We run 8 generations for $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. FID [36] is computed against the full CIFAR-10 training split using 30,000 generated samples per checkpoint.

Empirical proxy for $p^\star_\infty$. The collapse distribution $p^\star_\infty$ is not directly accessible, since p_{data} is unknown. We consider an empirical proxy of $p^\star_\infty$ by approximating $p_{\text{data}} \approx \hat{p}_{\text{trained}}$, where $\hat{p}_{\text{trained}}$ is a well-trained diffusion model on CIFAR-10, solely on real data, and from which one can thus sample. This allows to sample from $p^\star_\infty$ by leveraging its closed-form formula (infinite mixture of Gaussian-smoothed version of p_{data}), which is required to compute the FID.

High-frequency energy. For a probability measure p on $\mathbb{R}^{3\times32\times32}$ approximated by $N_{\text{spec}} = 2000$ samples $\{x^{(j)}\}_{j=1}^{N_{\text{spec}}}$, we define the *high-frequency energy* as

$$\mathrm{HF}(p) \;:=\; \frac{1}{N_{\text{spec}}} \sum_{j=1}^{N_{\text{spec}}} \frac{1}{|\mathcal{A}|} \sum_{\mathbf{k}\in\mathcal{A}} \left|\mathcal{F}[\bar{x}^{(j)}](\mathbf{k})\right|^2, \tag{74}$$

where $\bar{x}^{(j)} = 0.2989\, x_R^{(j)} + 0.5870\, x_G^{(j)} + 0.1140\, x_B^{(j)}$ is the luminance projection of the j-th sample, $\mathcal{F}$ is the zero-centered 2D discrete Fourier transform, and $\mathcal{A} := \{\mathbf{k} \in \mathbb{Z}^2 : \|\mathbf{k}\|_2 \geq 15\}$ is the high-frequency annulus in the 32×32 Fourier plane (cutoff radius $k^\star = 15$, slightly below the Nyquist limit of 16). We define the *high-frequency deficit* of generation N as

$$\mathrm{D}_{\mathrm{HF}}(\hat{p}^N) \;:=\; 1 - \frac{\mathrm{HF}(\hat{p}^N)}{\mathrm{HF}(p_{\mathrm{data}})}, \tag{75}$$

where $\mathrm{HF}(p_{\mathrm{data}})$ is computed on N_{spec} real CIFAR-10 images. Larger D_{HF} indicates stronger loss of fine-scale Fourier content relative to the data. Note that HF is quadratic in the samples, so it depends on p only through its second-moment matrix, i.e. through Hermite coefficients of degree at most 2; it is therefore not a proxy for the high-degree Hermite attenuation of Proposition 3.4. In particular, Proposition D.1 gives

$$\mathrm{HF}(p_\infty^\star) = c_2\, \mathrm{HF}(p_{\mathrm{data}}) + (1 - c_2)\, \mathrm{HF}(\gamma),$$

so truncation alone *increases* high-frequency power whenever $\mathrm{HF}(\gamma) > \mathrm{HF}(p_{\mathrm{data}})$, as is the case for CIFAR-10 (Figure 5, right).

Convergence to $p_\infty^\star$ (Figure 3a, Theorem 5.2). We report $\mathrm{FID}(\hat{p}^N, \hat{p}_\infty^{(\alpha)})$, where $\hat{p}_\infty^{(\alpha)}$ denotes the limiting distribution of the recursion with proportion α, across the first 8 recursive generations for each α. The contraction is fastest at large α and slowest at $\alpha = 0.1$, where the trajectory is clearly transient: the iterate drops from FID ≈ 33 at $N = 1$ to FID ≈ 5 by $N = 8$. By the eighth generation, all trajectories settle near a common empirical floor of FID ≈ 5, which we interpret as the residual $\mathrm{FID}(\hat{p}_\infty^{(\alpha)}, p_\infty^\star)$ plus measurement noise from finite-sample FID. This pattern is consistent with the contraction-rate ordering predicted by $\kappa(\alpha) = \sqrt{1-\alpha}\, e^{-t_0/2}$ in Theorem 5.2 and with the stability statement (ii) of the same theorem.

Spectral signature (Figure 3b, Propositions 3.4 and 5.3). The high-frequency deficit $\mathrm{D}_{\mathrm{HF}}(\hat{p}^N)$ across the first 8 generations is broadly increasing in both N and $1 - \alpha$. At $\alpha = 0.1$, the deficit reaches ≈ 0.5 by $N = 8$, indicating that roughly half of the high-frequency Fourier energy of CIFAR-10 has been suppressed; at $\alpha = 0.9$, the deficit remains below 0.15 over the same horizon. The mode-dependent contraction rate $\kappa_{\mathbf{n}}(\alpha) = (1-\alpha)\, e^{-|\mathbf{n}| t_0/2}$ in Proposition 5.3 predicts exactly this joint monotonicity in N (faster compounding for larger $|\mathbf{n}|$) and in $1 - \alpha$ (smaller fresh-data injection magnifies attenuation).

Truncation severity (Figure 5, Corollary 3.2). The FID gap $\mathrm{FID}(p_\infty^\star, p_{\mathrm{data}})$ grows with t_0 and decays toward zero as $t_0 \to 0$, while the radial Fourier spectrum of $p_\infty^\star$ approaches that of p_{data} at all frequencies as t_0 shrinks.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's main theoretical contributions. These claims are matched to formal results in Sections 3, 4, and 5, and to empirical validation described in the main text and in Appendix F.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Section 6 (Discussion and Future Work). These are mainly theoretical since the contribution is also theoretical.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All assumptions are explicitly stated and referenced (e.g. Assumptions (A1, A2, ...), and appear in order depending on which result they are needed for. Theorems are stated in the main text, and their proofs are systematically included in the appendix and appropriately referenced in the main text.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We include an experiments Section in Appendix F in which we describe the both the synthetic and real-data settings (dataset, architecture, optimization details etc.). These details are sufficient to understand and reproduce the reported empirical trends supporting the theoretical claims.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Justification: Code for all experiments, with instructions to reproduce every figure, is available at GitHub. CIFAR-10 is a publicly available dataset.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix F specifies the main experimental settings, including the synthetic-data and the CIFAR-10 diffusion setup, U-Net architecture, optimizer, learning rate, batch size, training epochs, sampling method, number of generations, number of seeds, and FID evaluation protocol.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The experiments are primarily used as qualitative and diagnostic validation of the theoretical predictions. While experiments on synthetic datasets report error bars or confidence intervals, our setting (recursive training) involves training diffusion models over several generations on real-image datasets, so obtaining accurate error bars is computationally prohibitive.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: These details can be found in Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research is mainly theoretical on diffusion-model collapse, and the empirical validation using synthetic data and standard public image datasets. It does not involve human-subject experiments, private data collection, deceptive practices, or the release of high-risk models.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: The paper is primarily foundational. We do not believe that there are potential societal consequences that need to be highlighted, so the paper does not currently include a societal impact discussion.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not introduce or release a high-risk pretrained generative model, scraped dataset, or deployment-ready system. The experiments use synthetic distributions and standard public benchmark data for scientific validation.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: The paper cites the existing datasets, models, and methods used in the experiments, including CIFAR-10 and standard diffusion/sampling components, but the current version does not explicitly list licenses or terms of use for all external assets.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not introduce a new dataset, benchmark, pretrained model, or other reusable asset. The main contributions are theoretical results and empirical validation.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: This theoretical work did not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This theoretical work did not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core methodology does not use LLMs as an important, original, or non-standard component. Any use of LLMs, if any, was limited to writing, editing, or formatting and did not affect the scientific content or methodology.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.