Low-Dimensional Adaptation of Rectified Flow: A Diffusion and Stochastic Localization Perspective

Saptarshi Roy
CMS Division
Mohamed bin Zayed University of Artificial Intelligence
Abu Dhabi, UAE

Alessandro Rinaldo
Department of Statistics and Data Science
University of Texas at Austin, USA

Purnamrita Sarkar
Department of Statistics and Data Science
University of Texas at Austin, USA

Abstract

In recent years, Rectified flow (RF) has gained considerable popularity largely due to its generation efficiency and state-of-the-art performance. In this paper, we investigate how well RF automatically adapts to the intrinsic low dimensionality of the support of the target distribution to accelerate sampling. We show that, using a carefully designed end-point adapted time-discretization scheme and with sufficiently accurate drift estimates, the RF sampler enjoys an iteration complexity of order $O(k/\varepsilon)$ (up to log factors), where ε is the precision in total variation distance and k is the intrinsic dimension of the target distribution. In addition, we show that the denoising diffusion probabilistic model (DDPM) procedure is equivalent to a stochastic version of RF by establishing a novel connection between these processes and stochastic localization. Building on this connection, we further design a stochastic RF sampler that also adapts to the low-dimensionality of the target distribution under mild requirements on the accuracy of the drift estimates, and also with a specific time schedule. We illustrate the efficacy of newly designed time-discretization schedules with simulations on the synthetic data and text-to-image (T2I) data experiments.

1 Introduction

Recently, Rectified flow or Flow matching [43, 39] has gained much attention due to its state-of-the-art generative performance across various modalities including image [47, 68], audio [66, 41], and video [10, 42]. Its effectiveness arises from a deterministic ordinary differential equation (ODE) based sampler and the ability of reflow to learn nearly straight transport paths, resulting in faster sampling. This has naturally sparked theoretical interests [7, 21, 69, 30] to understand the generation quality of RF from a convergence viewpoint under a suitable metric. However, these recent results suffer from *curse of dimensionality*. Therefore, they fail to explain the impressive performance of RF for *high-dimensional* data, e.g., image data, even though they reside in a low-dimensional space [5, 54]. In contrast, recent works [55, 6, 26, 37, 38, 31] have established low-dimensional adaptation of DDPM and denoising diffusion implicit models (DDIM) during sampling. This explains the accelerated and high-quality generative capabilities of DDPM and DDIM in practice, even for high-dimensional data. To the best of our knowledge such results are not currently present for RF models. To this end, we make three key contributions in this paper:

1. **Low-dimensional adaptation of RF:** We establish that a class of endpoint-adapted time-discretization schedules provably enables RF to exploit the intrinsic low-dimensional structure (*cf.* Theorem 4.6 and Theorem 4.7) of the target distribution during the *first reflow*. Our theoretical findings complement the empirical observations of [34], who demonstrated that U-shaped time discretization improves the generative quality of the first-reflow model. In particular, the RF sampler enjoys $O(k/\varepsilon)$ (upto log factors) iteration complexity, where ε is the precision of the generated samples in total variation distance and k is the intrinsic dimen-

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

sion of the target distribution. This is particularly useful because it can eliminate repeated reflow steps, which are computationally costly and often cause model collapse [70, 27].

2. **Unification of samplers through stochastic localization:** Using the stochastic localization [16, 17] perspective, we demonstrate that DDPM is equivalent to the stochastic version of RF (which we refer to as STOC-RF) under a proper time change. Stochastic localization has been recently shown to provide a unifying and effective framework for representing and analyzing diffusion-based sampling processes [8, 52, 15]. This equivalence is convenient as it allows us to view STOC-RF through the lens of well-understood DDPM processes, which are known to adapt to the low-dimensionality of the target distribution. In fact, Section 3 shows that stochastic localization can be leveraged to provide a unified and convenient representation of a variety of sampling methods.
3. **Designing efficient stochastic RF sampler:** Using the connection between DDPM and STOC-RF, we also show that the stochastic RF sampler provably adapts to the low-dimensional structure (*cf.* Theorem 4.9) of the target distributions under less stringent assumptions on the learned model. We illustrate its competitive performance through various simulation experiments.

Notation. Throughout the paper, with some abuse of notations, we will identify a probability distribution with its Lebesgue density, which we will assume exists unless otherwise noted. For two distributions P and Q on $\mathbb{R}^d$ with Lebesgue densities p and q respectively, their total variation distance is denoted by $\mathsf{TV}(p,q) := \sup_A |P(A) - Q(A)| = 1/2 \int_{\mathbb{R}^d} |p(x) - q(x)|dx$, where the supremum is over all Borel subsets of $\mathbb{R}^d$.

2 Background and Preliminaries

2.1 Rectified Flow

In this section, we briefly introduce the basics of Rectified flow, a generative model that transitions between two distributions p_0 and p_1 by solving ODEs. We refer the reader to [44, 43] for an exhaustive treatment. We let $p_1 := \text{Law}(X_1)$ be the target data distribution with its support $\mathcal{X} \subseteq \mathbb{R}^d$, the data generating distribution from which we would like to draw samples. The linear-interpolation process is defined by

$$X_t = tX_1 + (1-t)X_0, \quad 0 \le t \le 1 \tag{1}$$

where p_0 is the *standard Gaussian* distribution. We will assume an independent initial coupling, i.e., $(X_0, X_1) \sim p_0 \otimes p_1$ and let $p_t = \text{Law}(X_t)$, for any $t \in [0,1]$. The RF procedure is a deterministic sampler arising from an ODE with *drift (or velocity) function* $v : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ defined as the solution to the optimization problem

$$v = \arg\min_f \int \mathbb{E}\|X_1 - X_0 - f(X_t, t)\|_2^2 dt, \tag{2}$$

where the minimization is over all functions $f : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$. For each t, The above objective in (2) is minimized at $(x,t) \mapsto v_t(x) := v(x,t) = \mathbb{E}[X_1 - X_0 \mid X_t = x]$. Under appropriate regularity conditions (investigated in [7, 50]), the ODE

$$dZ_t = v_t(Z_t)\, dt, \quad Z_0 \sim \mathcal{N}(0, I_d) \tag{3}$$

satisfies the marginal preserving property, i.e., $\text{Law}(Z_t) = \text{Law}(X_t) = p_t$, owing to the Fokker-Planck equation

$$\frac{\partial p_t}{\partial t} + \nabla.(v_t p_t) = 0. \tag{4}$$

Hence, ODE (3) can be used for sampling.

To implement this procedure, one faces two key challenges. First, given a sample from the target distribution p_1, one has to estimate the drift function, which is naturally done by empirically minimizing the objective in (2) over a large and expressive function class (e.g., U-Net). Secondly, for sampling, one must rely on a time-discretization scheme of the ODE (3). One such scheme is the Eulerian update rule, which, for a given discretization of the time course in N intervals, with $0 = t_0 < t_1 < \cdots < t_N = 1$, evaluates

$$Y_{t_{i+1}} = Y_{t_i} + \eta_i \widehat{v}_{t_i}(Y_{t_i});\ \eta_i := t_{i+1} - t_i, 0 \le i \le N-1, \tag{5}$$

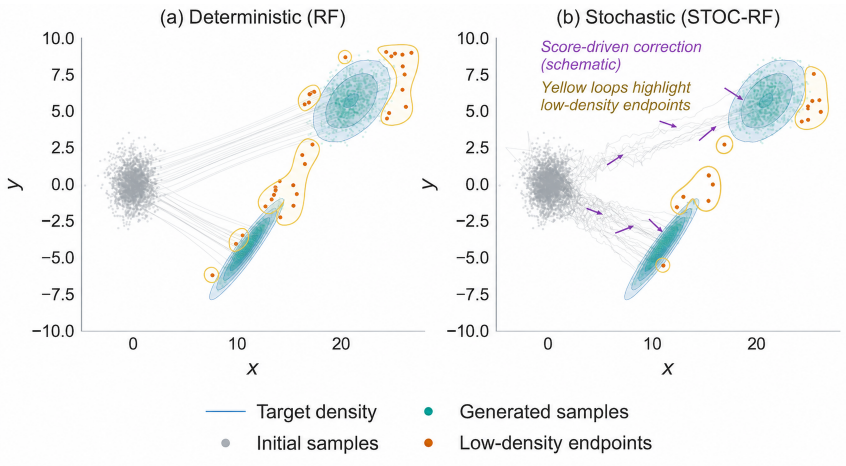


Figure 1: Trajectories and generated samples from deterministic (RF) and stochastic (STOC-RF) samplers targeting a two-component Gaussian mixture. Blue contours indicate target density, and highlighted endpoints lie in low-density regions. Purple arrows illustrate the score-driven component of the Langevin correction, which locally directs samples toward higher density and can counteract accumulated sampling errors—acting as a soft "guardrail." The stochastic sampler exhibits fewer low-density endpoints in this example.

where $Y_0 \sim \mathcal{N}(0, I_d)$ and $\widehat{v}_t(\cdot)$ is the estimated velocity at time t. Let $\widehat{p}_1 := \text{Law}(Y_1)$ denote the distribution of the final step. Other types of time discretized schemes of the ODE (3) can be chosen for the sampling step, and the resulting discretization errors impact the sampling fidelity differently. This has been the focus of a recent line of works [8, 11, 35, 36, 26, 38] aimed at identifying time-discretized versions of DDPM and DDIM samplers to achieve tight convergence rates.

The velocity function. Before proceeding, we illustrate the connection between the velocity $v_t(x)$ and the score function $s_t(x) := \nabla \log p_t(x)$. Recall that $X_0 \sim \mathcal{N}(0, I_d)$ and independent of $X_1 \sim p_1$. Therefore, Tweedie's formula [14, 51] immediately yields $v_t(x) = \frac{x}{t} + \frac{(1-t)}{t} s_t(x)$. Thus, estimating v_t is essentially equivalent to estimating the score s_t for $t \in (0, 1)$.

2.2 Stochastic rectified flow (STOC-RF)

Since the work of DDIM [62] and score-based diffusion models [63], it is well known that an SDE admits a corresponding probability-flow ODE that maintains the same marginal. Conversely, an ODE can also be augmented with appropriately matched drift and diffusion terms to construct a stochastic sampler that preserves these marginals. However, the joint distributions could be different; resulting in mismatch between the stochastic processes. In fact, prior works [1, 67, 24] suggest that stochastic samplers typically enjoy better generation quality. One reason behind the inferior performance of deterministic samplers is the possibility of error accumulation during the sampling step along the discrete trajectories stemming from inaccurate velocity estimates. In other words, if Y_{t_i} differs

significantly from Z_{t_i} for some t_i, then the Euler step (5) accumulates error due to inaccurate evaluation of $\widehat{v}_{t_i}(Y_{t_i})$. Therefore, this can further reinforce the deviation in evaluating $Y_{t_{i+1}}$. A way to mitigate the compounding error from successive Euler steps is to introduce a stochastic Langevin correction at each step. Empirically, [25] showed that the resulting stochastic RF samplers enjoy better text rendering quality over deterministic samplers. We further illustrate this point with Figure 1, showing that the ODE-based RF trajectories appear to generate more outliers compared to the SDE-based STOC-RF sampler.

Following e.g. [44, Chapter 5], the stochastic version of RF is specified by the SDE

$$d\tilde{Z}_t = v_t(\tilde{Z}_t)dt + \underbrace{\gamma_t s_t(\tilde{Z}_t)dt + \sqrt{2\gamma_t}d\tilde{B}_t}_{\text{Langevin correction}}; \tag{6}$$

for $t \in [t_0, 1)$, where $t_0 > 0$ and $\tilde{Z}_{t_0} \sim p_{t_0}$. This is obtained from adding a Langevin correction to the original RF ODE (3). Above, $\gamma_t \geq 0$ is a possibly time-varying diffusion coefficient, and $\{\tilde{B}_t\}_{t\geq 0}$ is the standard Brownian motion in $\mathbb{R}^d$. One can check (see, e.g., [44, Section 5.2] or [63]) that the marginal $\tilde{p}_t := \text{Law}(\tilde{Z}_t)$ satisfies the continuity equation

$$\frac{\partial \tilde{p}_t}{\partial t} + \nabla.(\tilde{v}_t \tilde{p}_t) = 0, \tag{7}$$

where $\tilde{v}_t(x) = v_t(x) + \gamma_t s_t(x) - \gamma_t \nabla \log \tilde{p}_t(x)$. With this matching initial law, the standard uniqueness argument for the marginal evolution gives $\tilde{p}_t = p_t$, i.e., $\text{Law}(\tilde{Z}_t) = \text{Law}(Z_t)$, for $t \in [t_0, 1)$. Thus the SDE follows the RF probability path towards p_1. In practice, this is accomplished via an appropriate time discretization scheme

$$\tilde{Z}_{t_{i+1}} = \tilde{Z}_{t_i} + \tilde{\eta}_i \left\{ v_{t_i}(\tilde{Z}_{t_i}) + \gamma_{t_i} s_{t_i}(\tilde{Z}_{t_i}) \right\} + \sqrt{2\tilde{\eta}_i \gamma_{t_i}} \xi_{t_i}, \tag{8}$$

where $\{t_i\}_{i=0}^N$ is an appropriate schedule of times, the ξ_{t_i}'s are i.i.d. draws from $\mathcal{N}(0, I_d)$, and the $\tilde{\eta}_i$'s are (typically small) step-sizes. The Langevin term essentially acts as a course-corrector to the trajectory of (8) to adjust the trajectory distribution closer to p_t. Finally, one has freedom to choose the diffusion coefficients γ_t. In this paper we will restrict ourselves to the choice $\gamma_t = (1-t)/t$, which, as we will see below, is the *appropriate noise scaling that yields equivalence between the STOC-RF SDE* (6) *and DDPM through the lens of Stochastic localization*. Endpoint behavior is discussed in Appendix D.2. This also addresses an question raised during the ICML tutorial talk [45] regarding the motivation of such choice.

3 Stochastic localization, RF and DDPM: A Unified View

In this section, we present a unified framework for studying and analyzing the stochastic DDPM sampler and the deterministic RF sampler. We show that stochastic localization acts as a unifying bridge between STOC-RF and DDPM, and it allows us to seamlessly transit the convergence property of one to another. This framework also extends to deterministic samplers, including RF and DDIM; see Appendix A. Figure 2 provides a schematic of the equivalence results.

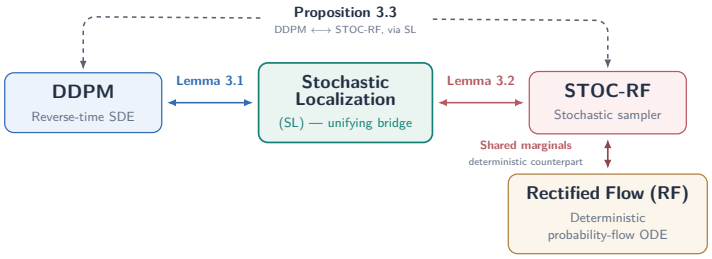


Figure 2: Stochastic localization (SL) provides a unifying bridge between DDPM and STOC-RF. Lemmas 3.1 and 3.2 establish the respective equivalences in law under suitable time reparameterizations and scalings, leading to Proposition 3.3, which connects the DDPM and STOC-RF samplers.

Stochastic localization (SL). SL is a probabilistic technique to study the properties of high-dimensional distributions due to [16, 17] that provides a powerful unifying framework to understand, analyze and improve generative models and sampling algorithms based on scores. See [15], [52] and the survey by [60]. To sample from the target data distribution p_1, SL focuses on an appropriate measure-valued stochastic process $\{\nu_s\}_{s\geq 0}$ such that ν_s "localizes" towards p_1 as $s \to \infty$. A straightforward way to formulate this process is to let, for any $s \geq 0$, $\nu_s := \text{Law}(X_1 \mid U_s)$, where $X_1 \sim p_1$ and

$$U_s = sX_1 + B_s, \tag{9}$$

with $\{B_s\}_{s\geq 0}$ a standard Brownian motion in $\mathbb{R}^d$ [52], independent of X_1. Thus, ν_s is a random measure depending on U_s. Note that $U_s/s \to X_1$ almost surely as $s \to \infty$, i.e., ν_s does indeed localize around p_1. More details on SL is deferred to Appendix A.

SL and DDPM. DDPMs are a class of diffusion-based sampling mechanism for generative AI introduced by [23]. DDPM first starts with a *forward process* $(Y'_\tau)_{\tau\geq 0}$ starting from $Y'_0 = X_1 \sim p_1$ and evolving according to the SDE

$$dY'_\tau = -\beta(\tau)Y'_\tau d\tau + \sqrt{2\beta(\tau)}dB'_\tau,\ Y'_0 \sim p_1, \tag{10}$$

where $\beta(\tau) > 0$ is a continuous diffusion coefficient with $\int_0^\infty \beta(u)\,du = \infty$ and $\{B'_\tau\}_{\tau\geq 0}$ is standard Brownian motion. If $\beta(\tau) = 1$, then the SDE (10) reduces to the classical Ornstein-Uhlenbeck (OU) process. Typically, The forwards process is continued till a large time $\tau = N$ such that Y'_N is approximately Gaussian. Let $q_\tau(y'_\tau)$ be the marginals of y'_τ following (10). Then in order to sample from the target distribution p_1, a *reverse process* $\{\overleftarrow{Y}_\tau\}_{\tau\in[0,N]}$ is formulated, starting at $\overleftarrow{Y}_0 = Y'_N$ and evolving according to the reverse-time SDE

$$d\overleftarrow{Y}_\tau = \{\overleftarrow{Y}_\tau + 2\nabla \log q_{N-\tau}(\overleftarrow{y}_\tau)\}\beta(N-\tau)d\tau + \sqrt{2\beta(N-\tau)}d\overleftarrow{B}_\tau, \tag{11}$$

where $\{\overleftarrow{B}_\tau\}_{\tau\geq 0}$ is a standard Brownian motion. Classical results in SDE [4, 22] show that $\overleftarrow{Y}_\tau \overset{d}{=} Y'_{N-\tau}$ which allows us to generate samples from p_1 via simulating the backward SDE (11).

It is known that DDPMs and stochastic localization are equivalent [52]. In particular, [8] showed that the forward process (10) is *equivalent in law* to the SL process (9) under an appropriate time change. In this section, we restate this result in a slightly different fashion that will allow us to easily see the connection between DDPM and RF. To this end, we define $\omega_\tau := \exp(-2\int_0^\tau \beta(u)du)$. Then the following lemma shows that the the *forward process* (10) *and the SL process* (9) *are equivalent:*

Lemma 3.1. *Let* $\{\tau(s)\}_{s>0}$ *satisfy* $\frac{1-\omega_{\tau(s)}}{\omega_{\tau(s)}} = \frac{1}{s}$. *Then,* $\left\{\frac{Y'_{\tau(s)}}{\sqrt{\omega_{\tau(s)}}}\right\}_{s>0} \overset{d}{=} \left\{\frac{U_s}{s}\right\}_{s>0}$.

If $\beta(\tau) = 1$ for all τ (OU process), then the time transformation in Lemma 3.1 simplifies to be $\tau(s) = \frac{1}{2}\log(1+s^{-1})$, and we get $\left\{\frac{Y'_{\tau(s)}}{e^{-\tau(s)}}\right\}_{s>0} \overset{d}{=} \left\{\frac{U_s}{s}\right\}_{s>0}$. Note, this matches the findings in [8, Section 1.2]. The proof is deferred to Appendix D.1.

SL, RF and STOC-RF. Next, we will also show that the probability path of RF is also equivalent to the SL process (9). Recall that that RF starts by constructing the linear process $X_t = tX_1 + (1-t)X_0$, where $X_1 \sim p_1, X_0 \sim \mathcal{N}(0, I_d)$ and $t \in [0,1]$. To this end, we consider the process

$$\tilde{X}_t = tX_1 + tW_{(1-t)^2/t^2}, \quad t \in (0,1] \tag{12}$$

where $\{W_\tau\}_{\tau\geq 0}$ is a Brownian motion independent of X_1. Note that $\text{Law}(\tilde{X}_t) = \text{Law}(X_t)$ for all $t \in (0,1]$. In fact, the entire process (12) has the same law of the SL process (9), up to time-change.

Lemma 3.2. *Let* $\{t(s)\}_{s>0}$ *satisfy* $\left(\frac{1-t(s)}{t(s)}\right)^2 = \frac{1}{s}$. *Then,* $\left\{\frac{\tilde{X}_{t(s)}}{t(s)}\right\}_{s>0} \overset{d}{=} \left\{\frac{U_s}{s}\right\}_{s>0}$.

The proof is in Appendix D.1. Additionally, Lemma 3.1 and Lemma 3.2 leads to the following:

$$\left\{Y'_\tau/\sqrt{\omega_\tau}\right\}_{\tau\geq 0} \overset{d}{=} \left\{\tilde{X}_{t(\tau)}/t(\tau)\right\}_{\tau\geq 0}, \tag{13}$$

where $t(\tau) = \frac{\sqrt{\omega_\tau}}{\sqrt{\omega_\tau}+\sqrt{1-\omega_\tau}}$. Now, on any positive-time interval with initial law p_{t_0}, (12) has the same law as the solution of (6) with $\gamma_t = (1-t)/t$, and (13) suggests that under the same time-change the STOC-RF process (6) and the backward process (11) are equivalent.

Proposition 3.3. *Let* $t(\tau) = \frac{\sqrt{\omega_\tau}}{\sqrt{\omega_\tau}+\sqrt{1-\omega_\tau}}$, *and let* $\{\tilde{Z}_t\}_{t\in[t(N),1)}$ *solve* (6) *with* $\gamma_t = (1-t)/t$ *and* $\tilde{Z}_{t(N)} \sim p_{t(N)}$. *Then,* $\left\{\frac{\sqrt{\omega_{N-\tau}}\tilde{Z}_{t(N-\tau)}}{t(N-\tau)}\right\}_{0\leq\tau<N}$ *is a solution of* (11) *with initial law* q_N. *Conversely, if* $\{\overleftarrow{Y}_\tau\}_{0\leq\tau<N}$ *solves* (11) *with* $\overleftarrow{Y}_0 \sim q_N$, *then* $\left\{\frac{\overleftarrow{Y}_{\tau(t)}}{\sqrt{\omega_{N-\tau(t)}}+\sqrt{1-\omega_{N-\tau(t)}}}\right\}_{t\in[t(N),1)}$ *solves* (6) *with initial law* $p_{t(N)}$. *Here* $\tau(t) \in [0,N)$ *is the unique solution of* $\int_0^{N-\tau(t)} \beta(u)\,du = \log\left(\sqrt{1+\frac{(1-t)^2}{t^2}}\right)$.

The above proposition is very useful in analyzing the convergence rate of STOC-RF. Essentially, Proposition 3.3 allows us to transit between DDPM and STOC-RF under simple time change and scaling. Therefore, convergence properties of one process can be carried over to the other one with the same transformation. This is the key insight that we will use to prove our convergence result of STOC-RF in later sections. The detailed proof is deferred to Appendix D.2.

4 Main Results

In this section, we present our main convergence results for both deterministic and stochastic samplers. We begin by introducing and commenting on our key assumptions. Throughout, $\{t_i\}_{i=0}^N$ refers to the time schedule for both the RF and STOC-RF procedures, as in (5) and (8), respectively.

4.1 Low-dimensionality of target distribution

To formalize the notion of low-dimensionality of p_1, we turn to the geometric complexity of its support $\mathcal{X}$ as measured by its metric entropy.

Assumption 4.1 (Low-dimensionality)**.** Let $\epsilon_0 = N^{-c_{\epsilon_0}}$ for some sufficiently large $c_{\epsilon_0} > 0$. The ϵ_0 covering number of $\mathcal{X}$ with respect to $\|\cdot\|_2$ norm, $\mathcal{N}(\mathcal{X}, \|\cdot\|_2, \epsilon_0)$, satisfies $\log \mathcal{N}(\mathcal{X}, \|\cdot\|_2, \epsilon_0) \leq C_{\text{cover}} k \log N$, for some constant $C_{\text{cover}} > 0$ and integer k. The quantity k is referred as the intrinsic-dimension of $\mathcal{X}$.

This assumption was proposed in recent studies by [26, 37, 38] on adapting diffusion methods to the intrinsic low-dimensional structure of the target data. The covering resolution and the entropy bound in this assumption both depend on the number of steps N. Informally, Assumption 4.1 requires the support of p_1 to be concentrated *on or near* a k-dimensional set. The above definition of intrinsic dimension is fairly general, as it covers a variety of important low-dimensional structures. Examples satisfying Assumption 4.1 include k-dimensional linear subspace in $\mathbb{R}^d$ and k-dimensional non-linear manifolds (provided $\mathcal{X}$ is polynomially bounded as in Assumption 4.2). Structures with doubling-dimension k [12, 28] also satisfies Assumption 4.1. More detailed discussion on this can be found in Section 4.1.1 of [26]. Moreover, in practice, target distribution (e.g., image data) is supported on a lower dimensional manifold and the intrinsic dimension $k \ll d$ [18, 5, 54].

Assumption 4.2 (Bounded support)**.** There exists a universal constant $c_R > 0$ such that $\sup_{x\in\mathcal{X}} \|x\|_2 \leq R$, where $R = N^{c_R}$.

The bounded support assumption is not as stringent as it may appear at first, as the diameter of $\mathcal{X}$ is allowed to scale polynomially (with arbitrarily large degree) in the number of iterations of the sampler, which accommodates a wide range of practical applications, such as image generation.

4.2 Assumptions on velocity approximation error

Our next set of assumptions are concerned with the approximation error of the estimator $\hat{v}_t$.

Assumption 4.3. Define $(\varepsilon_i^{\text{V}})^2 := \mathbb{E}\|v_{t_i}(X_{t_i}) - \widehat{v}_{t_i}(X_{t_i})\|_2^2$. Then, $\frac{1}{N}\sum_{i=1}^{N-2}(\varepsilon_i^{\text{V}})^2 \leq \varepsilon_{\text{v}}^2$.

The preceding assumption effectively calls for strong control of the mean-squared error in the estimated velocity field. Recent studies [9, 29] have demonstrated that modern neural networks can attain low approximation error through low-dimensional adaptation. Similar assumptions have been considered in [7, 21] to prove Wasserstein convergence of RF. In the context of score-matching algorithms, similar assumptions related to the score-approximation error have also appeared in prior works such as [8, 11, 36, 38].

Assumption 4.4 (Higher-order approximations)**.**

(a) Let $(\varepsilon_i^{\text{J},1})^2 := \mathbb{E}\|\nabla v_{t_i}(X_{t_i}) - \nabla\widehat{v}_{t_i}(X_{t_i})\|_F^2$. Then, $N^{-1}\sum_{i=1}^{N-2}(\varepsilon_i^{\text{J},1})^2 \leq \varepsilon_{\text{J},1}^2$. Also, we assume $\eta_i u^\top \nabla\widehat{v}_{t_i}(x)u \geq -\|u\|_2^2/2$ for all $u, x \in \mathbb{R}^d$ and $i \in \{0, \ldots N-2\}$.

(b) Let $(\varepsilon_i^{\text{J},2})^2 := \mathbb{E}[\text{Tr}(\nabla v_{t_i}(X_{t_i}) - \nabla\widehat{v}_{t_i}(X_{t_i}))^2]$. Then, $N^{-1}\sum_{i=1}^{N-2}(\varepsilon_i^{\text{J},2})^2 \leq \varepsilon_{\text{J},2}^2$.

(c) Let $(\varepsilon_i^{\text{H}})^2 := \mathbb{E}\|\nabla \text{Tr}(\nabla v_{t_i}(X_{t_i}) - \nabla\widehat{v}_{t_i}(X_{t_i}))\|_2^2$. Then, $N^{-1}\sum_{i=1}^{N-2}(\varepsilon_i^{\text{H}})^2 \leq \varepsilon_{\text{H}}^2$.

These assumptions are standard and similar conditions have been used in [38] in the context of DDIM. In practice, one may adopt higher-order velocity/score matching framework in [48] for controlling the error terms above. However, it is well known that estimating higher-order derivatives is a difficult problem in nonparametric statistics [19, 65].

4.3 Convergence rate of rectified flow

Our convergence guarantees for the RF sampler (5) depend crucially on a carefully constructed end-point adapted time schedule. We first make the following assumptions on the time schedule.

Assumption 4.5 (Endpoint concentration)**.** There exists small enough $\delta, h \in (0,1)$ such that $t_{N-1} \leq 1-\delta$, and

$$\max_{1\leq i\leq N-2}\left\{\frac{\eta_i}{t_i}\mathbb{1}\{t_i \leq 1/2\} + \frac{\eta_i}{1-t_{i+1}}\mathbb{1}\{t_i > 1/2\}\right\} \leq h.$$

The first condition simply guarantees that the second-to-last time step stays sufficiently far from 1. The second condition guarantees that the time points are tightly concentrated near the endpoints $t=0$ and $t=1$. This brings us to our first main result of this section.

Theorem 4.6. *Let Assumptions 4.1-4.4 hold, with covering resolution satisfying* (37)*, and write* $\mu_1 = \mathbb{E}_{X\sim p_1}[X]$*. Let* $\{Y_{t_i}\}_{i\geq 0}$ *denote the RF updates of* (5) *with time schedule satisfying Assumption 4.5. For sufficiently small* h *(depending on* k *and* N*), we have*

$$\begin{aligned}\mathsf{TV}(p_{X_{t_{N-1}}}, p_{Y_{t_{N-1}}}) \lesssim{}& (h+Nh^2)k\log N + Nh^2\varepsilon_{\mathrm{J},1}^2 + Nh^2\varepsilon_{\mathrm{v}}^2 + Nh(\varepsilon_{\mathrm{v}} + \varepsilon_{\mathrm{J},1} + \varepsilon_{\mathrm{J},2} + \varepsilon_{\mathrm{H}}) \\ &+ \eta_0^2\mathbb{E}\|X_1-\mu_1\|_2^2 + \eta_0^2(\varepsilon_0^{\mathrm{v}})^2 + \eta_0^2(\varepsilon_0^{\mathrm{J},1})^2 + (\varepsilon_0^{\mathrm{v}} + \varepsilon_0^{\mathrm{J},1} + \varepsilon_0^{\mathrm{J},2} + \varepsilon_0^{\mathrm{H}}) \\ &+ \frac{h+Nh^2}{\delta^2 N^{10}} + \frac{h^3}{\delta^6 N^9}.\end{aligned} \tag{14}$$

Adaptation to low-dimensionality and accelerated convergence. The weighted covariance-trace sum contributes an additional initial boundary term of order $hk\log N$. Assuming $(h+Nh^2)\log N = o(1)$, the upper bound (14) scales linearly only in k, indicating that the RF sampler adapts to the low-dimensional structure of the target distribution. For the geometric schedule below, $h \lesssim Nh^2$ and $h \asymp \log(1/\delta)/N$. Thus, assuming perfect knowledge of the velocity drift (i.e. $\widehat{v}_t = v_t$ and the higher approximation terms of Assumption 4.4 are all zero), the RF sampler (5) needs $\tilde{O}(k/\epsilon)$ iterations to achieve ϵ-approximate sampling. Empirical validation of the above scaling can be found in Appendix B.1. If the intrinsic dimension $k \ll d$, then the RF sampler automatically accelerates without any prior knowledge of the low-dimensionality of p_1. In the worst case, the moment term $\mathbb{E}\|X_1-\mu_1\|_2^2$ can scale as $O(d)$. However, if $\eta_0 = O(1/d)$, the moment term in the upper bound can be atmost $O(1/d)$.

Example of time schedule. The sufficient leading-term condition for low-dimensional (low-d) adaptation in the general bound is $(h+Nh^2)\log N = o(1)$. If we choose the linear uniform schedule, i.e., $t_i = i/N$, then $\eta_1/t_1 = 1$, so the small-step condition in Theorem 4.6 is not satisfied. We discuss a detailed recipe for constructing an admissible time schedule in Appendix E.1 and point out several examples in Table 5 that includes well known log-uniform, and EDM style time schedules. The general bound requires control of the endpoint step ratios through h, together with the displayed approximation-error and remainder terms. In this section, we adopt one such transparent example for clarity. For given, user-specified values $h \in (0,1)$ and $\delta \geq 0$, we consider the non-uniform time partition

$$\begin{aligned}&1.\ t_0 = 0, t_1 = \delta, t_j = (1+h)t_{j-1} \text{ for } 2 \leq j \leq N/2, \\ &2.\ t_N = 1, t_{N-1} = 1-\delta, t_j = 1-(1+h)(1-t_{j+1}) \quad \text{for } N/2 < j \leq N-2.\end{aligned} \tag{15}$$

We set h such that $t_{N/2} = \delta(1+h)^{(N-2)/2} = 1/2$. Using $N-2 > N/2$ (for $N>4$), we arrive at

$$\frac{h}{2} \leq \log(1+h) = \frac{2\log(\frac{1}{2\delta})}{N-2} \Rightarrow h \leq \frac{8\log(\frac{1}{2\delta})}{N}. \tag{16}$$

We remark that the proposed time schedule is practicable, as it does not depend on any properties of p_1. Typically, δ is chosen to be small, as this give rise to a U-shaped time discretization as shown

in Figure 7(a). Furthermore, the schedule enjoys certain properties (*cf.* Lemma E.2) that ensure $h + Nh^2 = O(\log^2(1/\delta)/N)$. This type of choice is not unusual [34], as it yields more accurate estimates $\widehat{v}_t$ during training when $t \approx 0$ and $t \approx 1$. Prior empirical work by non-uniform time discretizations have also been considered by [26, 37, 38] in the context DDPM and DDIM to establish low-dimensional adaptation of the models. It is important to emphasize that the specific form of the time schedule (15), while conforming to best empirical practices, is theoretically grounded, and it was derived directly as a result of our technical analysis of the RF dynamics (see Step 7 of the proof of Theorem 4.7, Appendix C.2). We now present the convergence result that establishes low-d adaptation under (15).

Theorem 4.7. *Let Assumptions 4.1-4.4 hold, with covering resolution satisfying* (37)*, and write* $\mu_1 = \mathbb{E}_{X\sim p_1}[X]$*. Let* $\{Y_{t_i}\}_{i\geq 0}$ *denote the RF updates of* (5) *with time schedule* (15)*. For* $\delta = 1/(N \vee d)$*, suppose* N *is sufficiently large so that* $h \leq 1/4$ *and* $16C_3kh\log N < 1$*, where* C_3 *is the constant in* (38)*. Then, we have*

$$\begin{aligned}\mathsf{TV}(p_{X_{t_{N-1}}}, p_{Y_{t_{N-1}}}) \lesssim\ & \frac{k\log^3(\frac{1}{\delta})}{N} + \frac{\log^2(\frac{1}{\delta})}{N}\varepsilon_{\mathrm{J},1}^2 + \frac{\log^2(\frac{1}{\delta})}{N}\varepsilon_{\mathrm{v}}^2 + \log(1/\delta)\left(\varepsilon_{\mathrm{v}} + \varepsilon_{\mathrm{J},1} + \varepsilon_{\mathrm{J},2} + \varepsilon_{\mathrm{H}}\right) \\ & + \delta^2\mathbb{E}\|X_1 - \mu_1\|_2^2 + \delta^2(\varepsilon_0^{\mathrm{v}})^2 + \delta^2(\varepsilon_0^{\mathrm{J},1})^2 + (\varepsilon_0^{\mathrm{v}} + \varepsilon_0^{\mathrm{J},1} + \varepsilon_0^{\mathrm{J},2} + \varepsilon_0^{\mathrm{H}}) \\ & + \frac{\log(1/\delta) + \log^2(1/\delta)}{\delta^2 N^{11}} + \frac{\log^3(1/\delta)}{\delta^6 N^{12}}.\end{aligned} \tag{17}$$

Guarantee on perturbed data. The TV guarantee in Theorem 4.7 is on the penultimate update $Y_{t_{N-1}}$ of (5), and not on $Y_{t_N} = Y_1$. It is worthwhile to mention that $\mathsf{TV}(p_{X_1}, p_{Y_{t_N}})$ might not be a useful or even meaningful quantity, as X_1 and Y_1 may have supports of different dimensions. For example, X_1 could be supported on a low-dimensional space, whereas Y_{t_N} would typically have full dimensional support, thereby rendering $\mathsf{TV}(X_1, Y_{t_N}) = 1$. However, note that for any $\delta > 0$, $X_\delta = (1-\delta)X_1 + \delta X_0$ has full-dimensional support due to the slight Gaussian perturbation. Therefore, the TV distance in Theorem 4.7 is well defined.

Higher-order approximation error terms. The upper bound in Theorem 4.7 is heavily dependent on the higher-order approximation errors of $\widehat{v}_t$. These terms are required to be small as deterministic samplers are unable to self-correct their trajectories. In addition, at $t = 0$, the true drift $v_0(x) = \mu_1 - x$ can be estimated by $\widehat{v}_0(x) = n^{-1}\sum_{i\in[n]} X_1^{(i)} - x$, where $\{X_1^{(i)}\}_{i\in[n]} \overset{i.i.d.}{\sim} p_1$. With this choice of $\widehat{v}_0$, we have $\varepsilon_0^{\mathrm{J},1} = \varepsilon_0^{\mathrm{J},2} = \varepsilon_0^{\mathrm{H}} = 0$, and $\varepsilon_0^{\mathrm{v}} = \sqrt{\mathbb{E}\|X_1 - \mu_1\|_2^2/n}$.

4.4 Convergence rate of STOC-RF

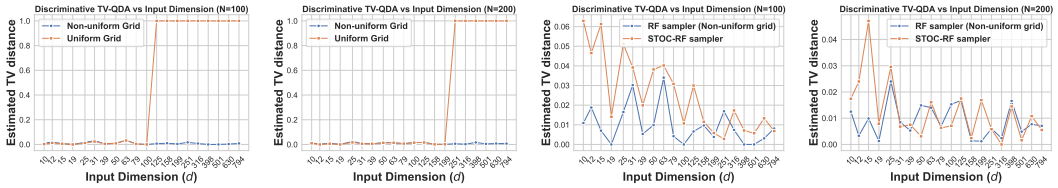
In this section, we design an SDE sampler that also adapts to the low-dimensionality of p_1 without the restrictive Assumption 4.4. The key idea here is to leverage the equivalence between DDPM and STOC-RF (*cf.* Figure 2) and exploit the low-dimensional adaptivity of DDPM samplers shown in [38]. In this regard, Proposition 3.3 and (13) provides the crucial time-change map to facilitate a simple reduction argument — the map $\tau \in [0,\infty) \mapsto t(\tau) := \frac{\sqrt{\omega_\tau}}{\sqrt{\omega_\tau}+\sqrt{1-\omega_\tau}}$ ensures that the forward process $\{Y'_\tau\}_{\tau\geq 0}$ in (10) and the STOC-RF process $\{\tilde{Z}_t\}_{t\in[0,1]}$ enjoy $\frac{Y'_\tau}{\sqrt{\omega_\tau}} \overset{d}{=} \frac{\tilde{Z}_{t(\tau)}}{t(\tau)}$.

In practice, we simulate the forward process by discretizing τ to *positive integers* in $\{1,\ldots,N\}$ and setting specific values for ω_τ. Following [38], we let $\omega_\tau = \prod_{j=1}^{\tau}\alpha_j$ with α_j's defined implicitly as

$$\beta_1 := 1 - \alpha_1 = 1/N^{c_0}, \beta_{\tau+1} := 1 - \alpha_{\tau+1} = \frac{c_1\log N}{N}\min\left\{\beta_1\left(1 + \frac{c_1\log N}{N}\right)^{\tau}, 1\right\}, \tag{18}$$

for a user-specified choice of the parameters $c_0, c_1 > 0$. Similar choices also appear in [55, 36, 37, 26] to analyze the convergence properties of DDPM. The discretized DDPM sampler then takes the form

$$\begin{aligned}&\hat{Y}'_{\tau-1} = \frac{1}{\sqrt{\alpha_\tau}}\left\{\hat{Y}'_\tau + \delta_\tau\widehat{s}_{Y'_\tau}(\hat{Y}'_\tau) + \nu_\tau\xi_\tau\right\}, \hat{Y}'_N \sim \mathcal{N}(0, I_d) \\ &\delta_\tau = 1 - \alpha_\tau, \nu_\tau = \sqrt{\frac{(\alpha_\tau - \omega_\tau)(1-\alpha_\tau)}{1-\omega_\tau}}, \tau = N, N-1, \ldots, 2, \quad \{\xi_\tau\}_{\tau\geq 1} \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d),\end{aligned} \tag{19}$$



(a) (b) (c) (d)

Figure 3: (a)-(b) *Estimated* TV distance between *blurred* low-rank Gaussian distribution and the generated samples by the RF sampler (5) under uniform and non-uniform time-grid under varying d. (c)-(d) *Estimated* TV distance between *blurred* low-rank Gaussian distribution and the generated samples by the RF and STOC-RF samplers under uniform and non-uniform time-grid for varying d.

where $\widehat{s}_{Y'_\tau}$ is an approximate score of Y'_τ.

Next, it remains to convert (19) to a STOC-RF sampler using the equivalence between Y'_τ and $\tilde{Z}_{t(\tau)}$. To this end, we let $\sigma_t^2 = (1-t)^2 + t^2$ for any $t \in [0,1]$, so that

$$\sigma^2_{t(\tau)} = \frac{1}{(\sqrt{\omega(\tau)} + \sqrt{1-\omega(\tau)})^2} = \frac{t^2(\tau)}{\omega_\tau}.$$

Therefore, we have $Y'_\tau \overset{d}{=} \tilde{Z}_{t(\tau)}/\sigma_{t(\tau)}$, and simple change of variable yields $s_{Y'_\tau}(y) = \sigma_{t(\tau)} s_{t(\tau)}(\sigma_{t(\tau)} y)$. Additionally, if we define $t_i := t(N-i)$ for $i = 0, 1, \ldots, N-1$, and $R_i = t_i/\sigma_{t_i}$, we arrive at an equivalent reformulation of (19) for RF sampling (*cf.* Appendix C.3) as described below: For $i = 0, \ldots, N-2$ update as

$$\frac{Y_{t_{i+1}}}{\sigma_{t_{i+1}}} = \frac{R_{i+1}}{R_i}\left\{\frac{Y_{t_i}}{\sigma_{t_i}} + \eta_i \sigma_{t_i} \widehat{s}_{t_i}(Y_{t_i}) + \sqrt{\psi_i} W_{t_i}\right\}, \ \frac{Y_{t_0}}{\sigma_{t_0}} \sim N(0, I_d), \ \{W_{t_i}\}_{i\geq 0} \overset{i.i.d.}{\sim} N(0, I_d), \tag{20}$$

where $\eta_i = 1 - \frac{R_i^2}{R_{i+1}^2}$, $\psi_i = \frac{R_i^2}{R_{i+1}^2}\frac{1-R_{i+1}^2}{1-R_i^2}\left(1 - \frac{R_i^2}{R_{i+1}^2}\right)$, and $\widehat{s}_{t_i}(x) = \frac{t_i \widehat{v}_{t_i}(x) - x}{1-t_i}$. Here η_i denotes the stochastic sampler coefficient; the time increment remains $t_{i+1} - t_i$. See Appendix C.3 for details. Lastly, we impose an assumption on the approximation error of $\widehat{s}_{t_i}$.

Assumption 4.8. Define $(\varepsilon_i^{\mathrm{s}})^2 := \mathbb{E}\|s_{t_i}(X_{t_i}) - \widehat{s}_{t_i}(X_{t_i})\|_2^2$. Then, $\frac{1}{N}\sum_{i=0}^{N-2}(\varepsilon_i^{\mathrm{s}})^2 \leq \varepsilon_{\mathrm{s}}^2$.

The above error terms are typically hard to control due to singularities near $t \approx 1$. However, higher weights can be assigned to training loss near end-points (*cf.* [34]) to achieve training-time stability. Next, the equivalence between (19) and (20) leads to the main convergence guarantee for STOC-RF.

Theorem 4.9. *Under Assumption 4.1, 4.2 and 4.8, the stoc-RF sampler* (20) *satisfies*

$$\mathsf{TV}(Y_{t_{N-1}}, X_{t_{N-1}}) \lesssim \frac{k \log^3 N}{N} + \varepsilon_{\mathrm{s}}\sqrt{\log N}.$$

Note that $t_{N-1} = \frac{\sqrt{\omega_1}}{\sqrt{\omega_1}+\sqrt{1-\omega_1}} \approx 1$. Therefore, $\mathrm{Law}(X_{t_{N-1}}) \approx p_1$. In comparison with Theorem 4.7, STOC-RF sampler does not require any conditions on higher-order derivatives (Assumption 4.4) of the score or velocity field. That is, STOC-RF sampler can achieve similar performance under less stringent requirements. This is primarily due to the Gaussian perturbation term in (20) which acts as a corrector step. The proof is in Appendix C.3.

Remark 1. Using the equivalence relation (13), it is also possible to directly relate the deterministic RF sampler to the DDIM sampler studied by [38, Theorem 1] (Appendix C.4) via the same time change transformation considered in this section. The resulting time schedule also follows a U-shaped distribution and the sampler will also adapts to the low-dimensionality of p_1.

5 Experiments

Synthetic data. We set the intrinsic dimension $k = 8$, and set our target distribution to be $p_1 = \mathcal{N}(\mu, \Sigma)$, where $\mu = (8, 8, \ldots, 8)^\top \in \mathbb{R}^d$ and Σ is a block-diagonal matrix with blocks I_k and

Code for the experiments is available at `https://github.com/roysaptaumich/Low-dim-RF`.

(a) T2I experiments.

Method	Mean CLIP (↑)	Mean Pick (↑)	Mean CER (↓)
ODE + Lin	34.1130	0.2295	1.0904
ODE + Geom	**34.1650**	0.2301	1.1339
SDE (vanilla)	29.3820	0.2085	1.6738
STOC-RF	33.8270	**0.2307**	**0.9718**

(b) MNIST experiments with different time grids.

Time Grid	FID (↓)	MMD (↓)
Linear	6.373 ± 0.733	0.002026 ± 0.000003
Geometric	$\mathbf{6.135 \pm 0.727}$	$\mathbf{0.002023 \pm 0.000004}$
Logit	6.358 ± 1.143	0.002024 ± 0.000004

Table 1: Comparison of sampling methods and time discretization schedules. (a) CLIP, Pick, and CER metrics for text-to-image (T2I) generation. (b) FID and MMD metrics on MNIST under different time grids. Best results are highlighted in bold.

$\mathbf{0}_{(d-k)\times(d-k)}$. Tweedie's formula (Section 2.1) yields

$$v_{t_i}(x) = \begin{bmatrix} \frac{2t_i-1}{\sigma_{t_i}^2} I_k & \mathbf{0}_{k\times(d-k)} \\ \mathbf{0}_{(d-k)\times k} & -\frac{1}{1-t_i} I_{d-k} \end{bmatrix} x + \begin{bmatrix} \frac{1-t_i}{\sigma_{t_i}^2} I_k & \mathbf{0}_{k\times(d-k)} \\ \mathbf{0}_{(d-k)\times k} & \frac{1}{1-t_i} I_{d-k} \end{bmatrix} \mu. \tag{21}$$

We use (21) for generating new samples from p_1. We vary d between 10 and 800, and set $N \in \{100, 200\}$. We use both a non-uniform time-grid (15) (with $\delta = \min\{1/N, 1/d\}$) and a uniform time-grid for the generation task. We generate 2000 samples using (15) and measure the TV distance with respect to another batch of 2000 samples independently generated from the slightly blurred target distribution Law$((1-\delta)X_1 + \delta Z)$, where $Z \sim \mathcal{N}(0, I_d)$. We use the discriminative approach proposed in [64] to estimate the TV distance between the distributions via 10 independent Monte Carlo rounds. For brevity, we refer to the non-uniform grid (15) as the geometric (Geom) grid in this section. Figure 3(a)-(b) show that Geom grid has better adaptation to the low-dimensional structure of p_1 for large d. We observe a sharp phase transition in TV distance across time schedules. This may reflect the sensitivity of TV to support mismatch, which becomes more pronounced as discretization errors accumulate in higher dimensions. The geometric schedule appears to mitigate this effect by better controlling errors near the endpoints. For the SDE sampler (20), we also compare its TV convergence with that of the deterministic sampler (5) with the non-uniform time grid (15). Figure 3(c)-(d) show that both samplers have similar performance across varying dimensions.

MNIST experiments. We train an RF model on MNIST and compare sampling time schedules using FID and MMD averaged over 10 independent runs. Table 1(b) shows that the geometric schedule achieves the lowest FID and MMD by modest margins, suggesting competitive performance on real-world data exhibiting low-dimensional structure.

T2I experiments. We perform experiments using Flux [32, 33] to demonstrate the usefulness of the U-shaped time schedule (15). We collect 6 different text prompts from [46, 25] and apply different samplers for 5 independent rounds and report the mean score in Table 1(a). We see that generation quality of RF is competitive under the Geom grid (15). In terms of mean CLIP and Pick score, Geom grid is slightly better than the linear/ uniform grid. On the contrary, linear grid shows better performance in mean CER score. Figure 6 compares the generation quality of STOC-RF sampler (20) and the vanilla DDPM (uniform grid) and ODE sampler. Table 1(a) shows that the generation quality is slightly better (on average) for STOC-RF in terms of mean PickScore and CER. The limitations of deterministic samplers have also been observed in prior works [13, 56, 61]. More experimental details and studies are deferred to Appendix B.2–B.4.

6 Conclusions

We show that U-shaped time schedules allow pretrained RF samplers to accelerate convergence by adapting to low-dimensional target structure under the stated assumptions. Endpoint step ratios, controlled by h, govern the leading discretization error and provide a principled criterion for comparing schedules beyond empirical performance. The benefits arise without additional reflow iterations. Through stochastic localization, we unify DDPM and stochastic RF, and establish low-dimensional adaptation for STOC-RF under weaker approximation assumptions.

Limitations. Our guarantees require regularity and estimation assumptions; practical performance depends on the time grid and velocity/score accuracy. Dimension-dependent constants, schedule optimality, broader empirical evaluation, and tighter finite-step guarantees remain subjects for future theoretical and empirical investigation.

Acknowledgment and Disclosures

We gratefully acknowledge NSF grants CCF-2019844, CCF-2505865, and 2217069.

AI use statement. The main technical ideas in the paper emerged from the authors. GPT-5.6 Sol assisted with writing and editing this paper in conjunction with the authors. The authors take full responsibility for the paper's content.

References

[1] Michael Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

[2] Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *The Eleventh International Conference on Learning Representations*, 2023.

[3] Tom Alberts, Yiming Xu, and Qiang Ye. Joint stochastic localization and applications. *arXiv preprint arXiv:2505.13410*, 2025.

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

[5] Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[6] Iskander Azangulov, George Deligiannidis, and Judith Rousseau. Convergence of diffusion models under the manifold hypothesis in high-dimensions. *arXiv preprint arXiv:2409.18804*, 2024.

[7] Vansh Bansal, Saptarshi Roy, Alessandro Rinaldo, and Purnamrita Sarkar. On the convergence and straightness of rectified flow. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026.

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

[9] Nicholas Boffi, Arthur Jacot, Stephen Tu, and Ingvar Ziemann. Shallow diffusion networks provably learn hidden low-dimensional structure. In *International Conference on Learning Representations*, volume 2025, pages 52889–52923, 2025.

[10] Shoufa Chen, Chongjian Ge, Yuqi Zhang, Yida Zhang, Fengda Zhu, Hao Yang, Hongxiang Hao, Hui Wu, Zhichao Lai, Yifei Hu, et al. Goku: Flow based video generative foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23516–23527, 2025.

[11] 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 *The Eleventh International Conference on Learning Representations*, 2023.

[12] Sanjoy Dasgupta and Yoav Freund. Random projection trees and low dimensional manifolds. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 537–546, 2008.

[13] Valentin De Bortoli, Iryna Korshunova, Andriy Mnih, and Arnaud Doucet. Schrodinger bridge flow for unpaired data translation. *Advances in Neural Information Processing Systems*, 37:103384–103441, 2024.

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

[15] Ahmed El Alaoui and Andrea Montanari. An information-theoretic view of stochastic localization. *IEEE Transactions on Information Theory*, 68(11):7423–7426, 2022.

[16] Ronen Eldan. Thin shell implies spectral gap up to polylog via a stochastic localization scheme. *Geometric and Functional Analysis*, 23(2):532–569, 2013.

[17] Ronen Eldan. Taming correlations through entropy-efficient measure decompositions with applications to mean-field approximation. *Probability Theory and Related Fields*, 176(3):737–755, 2020.

[18] Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):12140, 2017.

[19] Jianqing Fan. *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*. Routledge, 2018.

[20] Ruiqi Gao, Emiel Hoogeboom, Jonathan Heek, Valentin De Bortoli, Kevin Patrick Murphy, and Tim Salimans. Diffusion models and gaussian flow matching: Two sides of the same coin. In *The Fourth Blogpost Track at ICLR 2025*, 2025.

[21] Yunrui Guan, Krishnakumar Balasubramanian, and Shiqian Ma. Mirror flow matching with heavy-tailed priors for generative modeling on convex domains. *arXiv preprint arXiv:2510.08929*, 2025.

[22] Ulrich G Haussmann and Etienne Pardoux. Time reversal of diffusions. *The Annals of Probability*, pages 1188–1205, 1986.

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[24] Xixi Hu, Runlong Liao, Keyang Xu, Bo Liu, Yeqing Li, Eugene Ie, Hongliang Fei, and Qiang Liu. Improving rectified flow with boundary conditions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2025.

[25] Xixi Hu, Keyang Xu, Bo Liu, Qiang Liu, and Hongliang Fei. Amo sampler: Enhancing text rendering with overshooting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13157–13166, 2025.

[26] Zhihan Huang, Yuting Wei, and Yuxin Chen. Denoising diffusion probabilistic models are optimally adaptive to unknown low dimensionality. *Mathematics of Operations Research*, 2026.

[27] Shin Seong Kim, Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Balanced conic rectified flow. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[28] Samory Kpotufe and Sanjoy Dasgupta. A tree-based regressor that adapts to intrinsic dimension. *Journal of Computer and System Sciences*, 78(5):1496–1515, 2012.

[29] Shivam Kumar, Yixin Wang, and Lizhen Lin. Flow matching is adaptive to manifold structures. *arXiv preprint arXiv:2602.22486*, 2026.

[30] Syamantak Kumar, Dheeraj Nagaraj, Saptarshi Roy, and Purnamrita Sarkar. Conditional flow matching for transport between markov processes. *arXiv preprint arXiv:2610.07229*, 2026.

[31] Syamantak Kumar, Dheeraj Nagaraj, and Purnamrita Sarkar. Dimension-free score matching and time bootstrapping for diffusion models. *Advances in Neural Information Processing Systems*, 38:32224–32277, 2026.

[32] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

[33] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025.

[34] Sangyun Lee, Zinan Lin, and Giulia Fanti. Improving the training of rectified flows. *Advances in neural information processing systems*, 37:63082–63109, 2024.

[35] Gen Li, Zhihan Huang, and Yuting Wei. Towards a mathematical theory for consistency training in diffusion models. *arXiv preprint arXiv:2402.07802*, 2024.

[36] Gen Li, Yuting Wei, Yuejie Chi, and Yuxin Chen. A sharp convergence theory for the probability flow odes of diffusion models. *arXiv preprint arXiv:2408.02320*, 2024.

[37] Gen Li and Yuling Yan. Adapting to unknown low-dimensional structures in score-based diffusion models. *Advances in Neural Information Processing Systems*, 37:126297–126331, 2024.

[38] Jiadong Liang, Zhihan Huang, and Yuxin Chen. Low-dimensional adaptation of diffusion models: Convergence in total variation. In *The Thirty Eighth Annual Conference on Learning Theory*, pages 3723–3729. PMLR, 2025.

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

[40] R.S. Liptser, A.B. Aries, and A.N. Shiryaev. *Statistics of Random Processes I: General Theory*. Stochastic Modelling and Applied Probability. Springer New York, 2013.

[41] Huadai Liu, Jialei Wang, Rongjie Huang, Yang Liu, Heng Lu, Zhou Zhao, and Wei Xue. Flashaudio: Rectified flow for fast and high-fidelity text-to-audio generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13694–13710, 2025.

[42] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di ZHANG, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving video generation with human feedback. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

[43] Qiang Liu. Rectified flow: A marginal preserving approach to optimal transport. *arXiv preprint arXiv:2209.14577*, 2022.

[44] Qiang Liu. Let us flow together: Notes on rectified flow. `https://www.cs.utexas.edu/~lqiang/PDF/flow_book.pdf`, December 2024. Working Draft; version as of December 24, 2024.

[45] Qiang Liu. Flowing through continuous-time generative models: A clear and systematic tour. Tutorial presented at the 42nd International Conference on Machine Learning (ICML 2025), Vancouver, Canada, 2025.

[46] Rosanne Liu, Dan Garrette, Chitwan Saharia, William Chan, Adam Roberts, Sharan Narang, Irina Blok, RJ Mical, Mohammad Norouzi, and Noah Constant. Character-aware models improve visual text rendering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16270–16297, 2023.

[47] Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, et al. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *The Twelfth International Conference on Learning Representations*, 2024.

[48] Cheng Lu, Kaiwen Zheng, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Maximum likelihood training for score-based diffusion odes by high order denoising score matching. In *International conference on machine learning*, pages 14429–14460. PMLR, 2022.

[49] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pages 23–40. Springer, 2024.

[50] Gonzalo Mena, Arun Kumar Kuchibhotla, and Larry Wasserman. Statistical properties of rectified flow. *arXiv preprint arXiv:2511.03193*, 2025.

[51] Chenlin Meng, Yang Song, Wenzhe Li, and Stefano Ermon. Estimating high order gradients of the data distribution by denoising. *Advances in Neural Information Processing Systems*, 34:25359–25369, 2021.

[52] Andrea Montanari. Sampling, diffusions, and stochastic localization. *arXiv preprint arXiv:2305.10690*, 2023.

[53] Bernt Øksendal. Stochastic differential equations. In *Stochastic differential equations: an introduction with applications*, pages 38–50. Springer, 2003.

[54] Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021.

[55] Peter Potaptchik, Iskander Azangulov, and George Deligiannidis. Linear convergence of diffusion models under the manifold hypothesis. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291, pages 4668–4685. PMLR, 2025.

[56] Teodora Reu, Sixtine Dromigny, Michael Bronstein, and Francisco Vargas. Gradient variance reveals failure modes in flow-based generative models. *Advances in Neural Information Processing Systems*, 38:128536–128570, 2026.

[57] Daniel Revuz and Marc Yor. *Continuous martingales and Brownian motion*, volume 293. Springer Science & Business Media, 2013.

[58] Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic image inversion and editing using rectified stochastic differential equations. In *The Thirteenth International Conference on Learning Representations*, 2025.

[59] Michael Ruzhansky and Mitsuru Sugimoto. On global inversion of homogeneous maps. *Bulletin of Mathematical Sciences*, 5(1):13–18, 2015.

[60] Bobby Shi, Kevin Tian, and Matthew S Zhang. Perspectives on stochastic localization. *arXiv preprint arXiv:2510.04460*, 2025.

[61] Yuyang Shi, Valentin De Bortoli, Andrew Campbell, and Arnaud Doucet. Diffusion schrödinger bridge matching. *Advances in neural information processing systems*, 36:62183–62223, 2023.

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

[63] 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*, 2021.

[64] Lan Tao, Shirong Xu, Chi-Hua Wang, Namjoon Suh, and Guang Cheng. Discriminative estimation of total variation distance: A fidelity auditor for generative data. *arXiv preprint arXiv:2405.15337*, 2024.

[65] Alexandre B Tsybakov and Vladimir Zaiats. *Introduction to nonparametric estimation*, volume 11. Springer, 2009.

[66] Yongqi Wang, Wenxiang Guo, Rongjie Huang, Jiawei Huang, Zehan Wang, Fuming You, Ruiqi Li, and Zhou Zhao. Frieren: Efficient video-to-audio generation network with rectified flow matching. *Advances in Neural Information Processing Systems*, 37:128118–128138, 2024.

[67] Kaiwen Xue, Yuhao Zhou, Shen Nie, Xu Min, Xiaolu Zhang, Jun Zhou, and Chongxuan Li. Unifying bayesian flow networks and diffusion models through stochastic differential equations. In *International Conference on Machine Learning*, pages 55656–55681. PMLR, 2024.

[68] Hanshu Yan, Xingchao Liu, Jiachun Pan, Jun Hao Liew, Qiang Liu, and Jiashi Feng. Perflow: Piecewise rectified flow as universal plug-and-play accelerator. *Advances in Neural Information Processing Systems*, 37:78630–78652, 2024.

[69] Zhengyu Zhou and Weiwei Liu. An error analysis of flow matching for deep generative modeling. In *Forty-second International Conference on Machine Learning*, 2025.

[70] Huminhao Zhu, Fangyikang Wang, Tianyu Ding, Qing Qu, and Zhihui Zhu. Analyzing and mitigating model collapse in rectified flow models. *arXiv preprint arXiv:2412.08175*, 2024.

Appendix

The appendix is organized as follows.

1. Section A discusses more deatils on the SL process.
2. Section B contains additional experimental results.
3. Section C contains detailed proofs from Sections 4 and the description of the RF sampler.
4. Section D contains detailed proofs of results about Section 3.
5. Finally, Section E contains all other helpful lemmas and their proofs.

A More details on Stochastic localization

In this section, we present a unified framework for studying and analyzing the stochastic DDPM sampler and the deterministic RF sampler.

Stochastic localization (SL). SL is a probabilistic technique to study the properties of high-dimensional distributions due to [16, 17] that provides a powerful, unifying framework for understanding, analyzing, and improving score-based generative models and sampling algorithms. See [15], [52] and the survey by [60]. To sample from the target data distribution p_1, SL focuses on an appropriate measure-valued stochastic process $\{\nu_s\}_{s\geq 0}$ such that ν_s "localizes" towards p_1 as $s \to \infty$. A straightforward way to formulate this process is by letting, for any $s \geq 0$, $\nu_s := \text{Law}(X_1 \mid U_s)$, where $X_1 \sim p_1$ and

$$U_s = sX_1 + B_s,$$

with $\{B_s\}_{s\geq 0}$ a standard Brownian motion in $\mathbb{R}^d$ [52], independent of X_1. Thus, ν_s is a random measure depending on U_s. Note that $U_s/s \to X_1$ almost surely as $s \to \infty$, i.e., ν_s does indeed localize around p_1.

While enlightening, the SL process, as stated, does not yield a sampler from p_1, as it uses $X_1 \sim p_1$ in the definition of U_s. Therefore, it is instead convenient to consider the alternative process $\{\tilde{U}_s\}_{s\geq 0}$ defined as the solution of the SDE

$$d\tilde{U}_s = a_s(\tilde{U}_s)\, ds + d\tilde{B}_s, \tag{22}$$

where $\tilde{B}_s$ is an independent Brownian motion and, for any $s \geq 0$, $a_s(u) = \mathbb{E}[X_1 \mid U_s = u]$. By classical SDE results – see [53, Theorem 8.4.3] or [40, Theorem 7.12] – $\{U_s\}_{s\geq 0}$ and $\{\tilde{U}_s\}_{s\geq 0}$ are equivalent in law. Therefore, (22) can be used for sampling from p_1 as long as we have access to the regression functions $\{a_s(\cdot), s \geq 0\}$, which can be estimated from the data. We also define the posterior covariance as $\mathbf{A}_s = \text{Cov}(X_1 \mid U_s)$. The next lemma from [17] allows us to control the growth of the expected posterior covariance.

Lemma A.1. *For all $s \geq 0$, $\frac{d}{ds}\mathbb{E}[\mathbf{A}_s] = -\mathbb{E}\left[\mathbf{A}_s^2\right]$.*

This is a crucial result for controlling the discretization error of ODE and SDE samplers of RF. Crucially, Lemma 3.2 yields the following Corollary to Lemma A.1 which is used to control the discretization error of the RF sampler.

Corollary A.2. *Let $\mathbf{\Sigma}_t := Cov(X_1 \mid X_t)$ for $t \in [0, 1)$. Then $\frac{d}{dt}\mathbb{E}(\mathbf{\Sigma}_t) = -\frac{2t}{(1-t)^3}\mathbb{E}(\mathbf{\Sigma}_t^2)$.*

See Appendix E.2 for a complete proof. The corollary above is extremely useful for controlling the quadratic term involving the Frobenius norm of $\mathbf{\Sigma}_t$ (*cf.* Proposition C.2) in the discretization error of the RF sampler. One may use vanilla Cauchy-Schwarz inequality to bound $\mathbb{E}\left(\|\mathbf{\Sigma}_t\|_F^2\right)$. However, this will lead to a quadratic dependence on k. Corollary A.2 provides much sharper control over the Frobenius norm by connecting it with the rate of change of the trace of $\mathbf{\Sigma}_t$ that only scales linearly in k. We point the readers to *Step 7* of the main proof of Theorem 4.6 in Section C.2.1.

SL and Stochastic Interpolants. Lemma 3.2 and the equivalence result in (13) can be extended to cover general stochastic interpolation [2, 1] schemes beyond RF. In detail, instead of a linear interpolation, one can consider a general deterministic affine interpolation process $I_\theta = a_\theta X_1 + b_\theta X_0$, where $a_{(\cdot)}$ and $b_{(\cdot)}$ are smooth functions of $\theta \in [0,1]$, with $a_0 = b_1 = 0$, and $a_1 = b_0 = 1$. When $X_0 \sim N(0, I_d)$, the associated stochastic interpolant is the process $(\tilde{I}_\theta)$ at times $\theta > 0$ for which $a_\theta \neq 0$, where $\tilde{I}_\theta = a_\theta X_1 + a_\theta W_{r_\theta^2}$, with $r_\theta = b_\theta / a_\theta$ and $\{W_t\}_{t\geq 0}$ a Brownian motion. It is easy to see that $\tilde{I}_\theta \overset{d}{=} I_\theta$, for each θ. Moreover, each stochastic interpolant can be rescaled and time-transformed to return the SL process, in a sense made precise by the next result.

Proposition A.3. $\left\{ \frac{\tilde{I}_{\theta(s)}}{a_{\theta(s)}} \right\}_{s>0} \overset{d}{=} \left\{ \frac{U_s}{s} \right\}_{s>0}$, *for positive s for which the time change $r^2_{\theta(s)} = 1/s$ is defined.*

In particular, in virtue of Lemma 3.2, RF is equal in law to any interpolant process I_θ under the time change $t(\theta) = \frac{a_\theta}{a_\theta + b_\theta}$. This equivalence among stochastic interpolants is not new, as it has also been shown in [44] by direct arguments. Connection between diffusion and flow matching has also been studied in [1, 2, 20, 49, 58] through different perspectives. By leveraging the representation properties of the SL process, we establish this result in a more unified manner, which has the added benefit of directly applying to both stochastic interpolants and SDEs. The proof of Proposition A.3 is deferred to Appendix D.3.

B Additional Experiments

Synthetic data experiments are conducted in standard Macbook with 8GB RAM and M2 processor. The T2I experiments have been done in the Lightning AI cloud with NVIDIA L40S GPU.

B.1 Checking tightness of the scaling of TV upperbounds

Experimental Setup and Findings

Setup: We evaluated the ODE-geometric sampler across a grid of:

- **Matrix dimensions:** $d \in \{100, 122, \ldots, 300\}$ (10 values evenly spaced)
- **Ranks:** $k \in \{3, 7, 11, 15, 20\}$ (5 values evenly spaced)
- **Tolerance levels:** $TV_{tol} \in \{0.01, 0.03, 0.05, 0.07, 0.10\}$ (5 levels)

For each triple (d, k, TV_{tol}), we recorded $N(d,k)$ — the **minimum number of discretization steps** required to achieve TV distance below TV_{tol}. We then compared the two candidate rates $k/N(d,k)$ and $d/N(d,k)$ as empirical proxies for TV_{tol} across all (d,k) pairs. To reduce noise, we binned each ratio across all (d,k) pairs and computed the **mean** TV_{tol} **per bin**, then fitted an OLS line to the binned means. We show the plots in Figure 4.

Linearity of k/N: The binned mean TV_{tol} exhibits a **clear and monotone linear trend** with k/N, with OLS achieving $R^2 = 0.878$ and slope $= 0.0094$ ($p = 1.83 \times 10^{-3}$). The through-origin fit

$$TV_{tol} \approx C \cdot \frac{k}{N}$$

also has smaller OLS residuals per bin (std $= 0.0082$) compared to

$$TV_{tol} \approx C \cdot \frac{d}{N}$$

fit. See the Figures for OLS statistics.

Superiority of k/N over d/N

By contrast, the same analysis applied to d/N yields substantially weaker results. The key comparisons are summarised below:

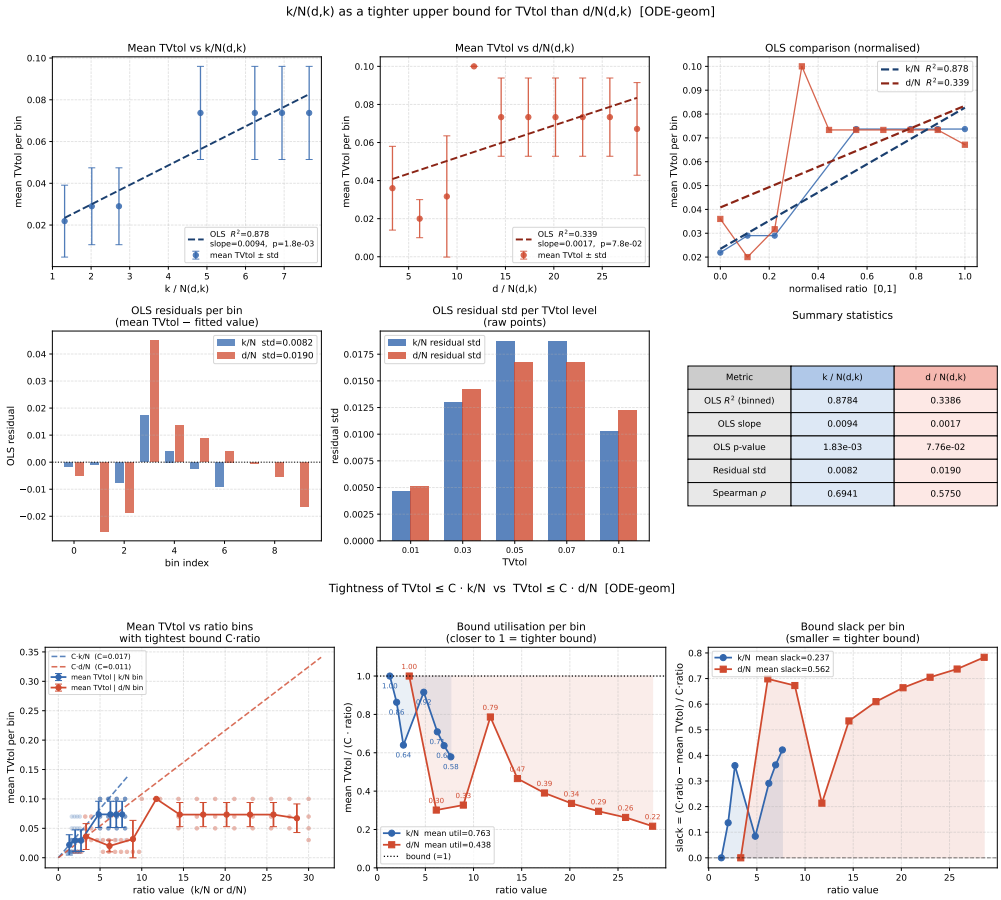


Figure 4: Empirical Validation of TV Distance Scaling Laws and Bound Utilisation.

OLS slope for d/N is **not statistically significant** ($p = 0.078 > 0.05$) - **no evidence** of a linear relationship between d/N and TV_{tol}. In contrast, k/N yields a highly significant slope with residual std more than **twice as small** (0.0082 vs 0.0190).

Tightness of the bound.
For each ratio $r \in \{\frac{k}{N}, \frac{d}{N}\}$, we estimate the tightest valid constant

$$C = \max_{d,k} \frac{TV_{tol}}{r},$$

and report two diagnostics per bin:

- **Bound utilisation:**

$$\frac{TV_{tol}}{Cr} \in (0, 1],$$

  where 1 indicates a perfectly tight bound.

- **Slack:**

$$\frac{Cr - TV_{tol}}{Cr} \in [0, 1),$$

The normalized gap between bound and observation: the k/N bound achieves mean utilization **0.763**, while d/N drops to **0.438**, making it roughly **1.74× looser**. Mean slack shows the same pattern: **0.237** for k/N vs. **0.562** for d/N, meaning the d/N bound leaves over twice the unused margin.

B.2 Comparison of RF sampler under uniform and non-uniform grid in T2I setting

We have performed various experiments to demonstrate the usefulness of the U-shaped time schedule (15) in the context of T2I. We use the Flux model [32, 33] which is a pretrained RF model. For our experiments, we use the prompts designed by [46, 25] to test and assess the generation quality. The list is given below:

1. `Photo of a robot lecturer writing the words 'Representation Learning' in cursive on a black- board, with math formulas and diagrams.`
2. `A close-up of a t-shirt with the phrase "I love AI".`
3. `Grape vines in the shape of text 'open your mind' sprouting out of a head with flowers and butterflies. DSLR photo.`
4. `A sign that says "Please refrain from arguing with the chimpanzees"`
5. `A photo of a cat holding a sign Rectified flow.`
6. ` car made of cheese on a wooden table.`
7. `the hubble telescope and the milky way, with the text ''the universe is a mystery, but we are here to solve it"`

In these examples, we always set $\delta = 1/N$ for a specified choices of N and report the mean CLIP, PickScores and CER scores in Table 2 across 5 different rounds for each of the first 6 prompts above. In summary, from Table 1 we observe the following:

- SDE (vanilla) is, on average, comparatively worse than other methods.
- STOC-RF is very competitive with other methods, both in terms of text rendering quality and text-image alignment.
- Among the ODE methods, non-uniform scheduling clearly improves or remains competitive in terms of CLIP score and PickScore. In terms of text rendering, ODE + uniform is slightly better than ODE + non-uniform.

In Figure 5 we see that generation quality of RF is superior under a non-uniform time-grid (15) compared to uniform time-grid. In particular, the experiments indicate that a non-uniform grid helps to reduce hallucinations, enjoys better text rendering and achieves better semantic completeness.

We also compare the generation quality of STOC-RF sampler (20) and the vanilla DDPM sampler (uniform grid) in Figure 6. The generated plots show that the quality is generally better for STOC-RF sampler (20) compared to the RF sampler and the vanilla DDPM sampler (uniform grid).

B.3 Ablation on Time Discretization

We investigate the effect of different time discretization schedules on text-to-image generation using five representative prompts from Appendix B.2. Specifically, we compare linear, geometric, cosine, and logit schedules while keeping the sampling budget fixed. Table 3 reports the corresponding CLIP, PickScore, and CER metrics. The geometric schedule achieves competitive performance across the considered prompts, although no single schedule uniformly outperforms the others. These results suggest that the proposed geometric schedule is a robust choice for time discretization in practical sampling.

B.4 Quality–Speed Trade-off

We further investigate the quality–speed trade-off between deterministic RF (ODE) and STOC-RF (SDE) by varying the number of function evaluations (NFE). We consider two prompts from Appendix B.2 and measure generation quality using CLIP, PickScore, and CER, together with wall-clock sampling time on an NVIDIA L40S GPU. For deterministic RF, we use the geometric time schedule.

As shown in Table 4, the sampling time increases approximately linearly with NFE for both methods, with comparable computational costs. Deterministic RF generally achieves better generation quality, particularly at smaller NFE. At larger NFE, the performance gap becomes smaller for some metrics, although deterministic RF remains competitive or superior across most reported configurations.

Table 2: Comparison of generative sampling methods across six T2I prompts from Appendix B.2 using CLIP, PickScore, and CER. Best results for each prompt are highlighted in bold.

Method	CLIP (↑)	PickScore (↑)	CER (↓)
Prompt: Robot Lecturer			
ODE + Lin	31.420	0.225	2.574
ODE + Geom	30.520	0.228	2.957
SDE (vanilla)	**33.710**	0.211	4.939
STOC-RF	31.310	**0.237**	**2.044**
Prompt: T-shirt ("I love AI")			
ODE + Lin	39.590	0.239	0.844
ODE + Geom	40.910	**0.241**	0.711
SDE (vanilla)	27.510	0.183	1.400
STOC-RF	**41.760**	0.238	**0.689**
Prompt: Grape Vines			
ODE + Lin	25.900	**0.230**	0.917
ODE + Geom	25.460	0.229	0.901
SDE (vanilla)	25.830	0.221	0.944
STOC-RF	**26.310**	**0.230**	**0.889**
Prompt: Sign ("Chimpanzees")			
ODE + Lin	38.710	**0.207**	0.522
ODE + Geom	**38.830**	**0.207**	0.528
SDE (vanilla)	24.430	0.172	0.976
STOC-RF	38.470	0.206	**0.508**
Prompt: Cat with sign			
ODE + Lin	**35.540**	**0.240**	**0.685**
ODE + Geom	35.450	**0.240**	0.706
SDE (vanilla)	32.460	0.227	0.783
STOC-RF	35.440	**0.240**	0.702
Prompt: Cheese Car			
ODE + Lin	33.520	0.237	1.000
ODE + Geom	**33.820**	0.235	1.000
SDE (vanilla)	32.350	**0.238**	1.000
STOC-RF	29.670	0.234	1.000

C Proof of main results

C.1 Velocity function and its first order identities

Before proceeding, we illustrate the connection between the velocity $v_t(x)$ and the score function $s_t(x) := \nabla \log p_t(x)$. To this end, we define two key quantities:

$$\mu_{1|t}(x) = \mathbb{E}[X_1 \mid X_t = x],\ \mathrm{Cov}_{1|t}(x) = \mathrm{Cov}(X_1 \mid X_t = x).$$

For $0 < t < 1$, the true drift can be re-written as

$$v_t(x) = \mathbb{E}[X_1 - X_0 \mid X_t = x] = \frac{1}{1-t}\left(\mu_{1|t}(x) - x\right). \tag{23}$$

By the first and second order Tweedie's formula [14, 51], it follows that, for any choice of $t \in (0,1)$ and $x \in \mathbb{R}^d$,

$$\mu_{1|t}(x) = \frac{x}{t} + \frac{(1-t)^2}{t} s_t(x),$$
$$\mathrm{Cov}_{1|t}(x) = \frac{(1-t)^2}{t^2} I_d + \frac{(1-t)^4}{t^2} \nabla s_t(x).$$

Substituting this in (23), we obtain that

$$v_t(x) = \frac{x}{t} + \frac{(1-t)}{t} s_t(x), \quad \text{and} \quad \nabla v_t(x) = -\frac{I_d}{1-t} + \frac{t}{(1-t)^3}\mathrm{Cov}_{1|t}(x).$$



(a) **Prompt:** A photo of a cat holding a sign Rectified flow. (Less hallucination, $N = 50$)



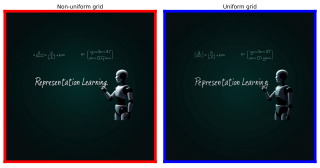
(b) **Prompt:** Grape vines in the shape of text 'open your mind' sprouting out of a head with flowers and butterflies. DSLR photo. (Better text rendering, $N = 50$)



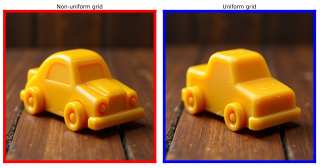
(c) **Prompt:** the hubble telescope and the milky way, with the text "the universe is a mystery, but we are here to solve it" (Better fidelity/ prompt adherence, $N = 50$)



(d) **Prompt:** A sign that says "Please refrain from arguing with the chimpanzees". (Minimal mistakes, $N = 50$)



(e) **Prompt:** Photo of a robot lecturer writing the words 'Representation Learning' in cursive on a blackboard, with math formulas and diagrams. (Correct spelling, $N = 50$)



(f) **Prompt:** A car made of cheese on a wooden table. (Semantic completeness, $N = 80$)

Figure 5: Generated images using Flux for specific prompts via Euler's scheme (5) under non-uniform time-grid (15) (Red frame) and uniform time-grid (Blue frame).

C.2 Proof of convergence of RF sampler (Theorem 4.6 and Theorem 4.7)

C.2.1 Proof of Theorem 4.6

Notational conventions. We begin with some notational conventions for the ease of presentation.

1. We define transition map for the Euler's scheme as
$$\Phi_{t_i}(x) := x + \eta_i \widehat{v}_{t_i}(x) \quad \text{with } \eta_i = t_{i+1} - t_i,$$
where $\widehat{v}_t(x)$ is the velocity estimate.
2. For notational brevity, we will use index i to denote the time t_i. In this index notation we have $X_N \equiv X_1 \sim p_1$
3. With the above notations in mind, we have
$$X_i = (1 - t_i)X_0 + t_i X_N, \quad \text{and} \quad Y_{i+1} = \Phi_i(Y_i). \tag{24}$$

Step 1: TV recursion. We start by obtaining a recursion relation between the TVdistances between the updates Y_i and the ground truth X_i. For any $i \leq N - 2$, we have

$$\begin{aligned}
\mathsf{TV}\left(p_{X_{i+1}}, p_{Y_{i+1}}\right) &= \sup_{\mathcal{A}\subseteq\mathbb{R}^d} \left\{\mathbb{P}_{X_{i+1}}(\mathcal{A}) - \mathbb{P}_{Y_{i+1}}(\mathcal{A})\right\} = \sup_{\mathcal{A}\subseteq\mathbb{R}^d} \left\{\mathbb{P}_{X_{i+1}}(\mathcal{A}) - \mathbb{P}_{Y_i}\left(\Phi_i^{-1}(\mathcal{A})\right)\right\} \\
&\leq \sup_{\mathcal{A}\subseteq\mathbb{R}^d} \left\{\mathbb{P}_{X_{i+1}}(\mathcal{A}) - \mathbb{P}_{X_i}\left(\Phi_i^{-1}(\mathcal{A})\right)\right\} + \sup_{\mathcal{A}\subseteq\mathbb{R}^d} \left\{\mathbb{P}_{X_i}\left(\Phi_i^{-1}(\mathcal{A})\right) - \mathbb{P}_{Y_i}\left(\Phi_i^{-1}(\mathcal{A})\right)\right\} \\
&\leq \sup_{\mathcal{A}\subseteq\mathbb{R}^d} \left\{\mathbb{P}_{X_{i+1}}(\mathcal{A}) - \mathbb{P}_{\Phi_i(X_i)}(\mathcal{A})\right\} + \mathsf{TV}\left(p_{X_i}, p_{Y_i}\right) \\
&= \mathsf{TV}\left(p_{X_{i+1}}, p_{\Phi_i(X_i)}\right) + \mathsf{TV}\left(p_{X_i}, p_{Y_i}\right)
\end{aligned} \tag{25}$$

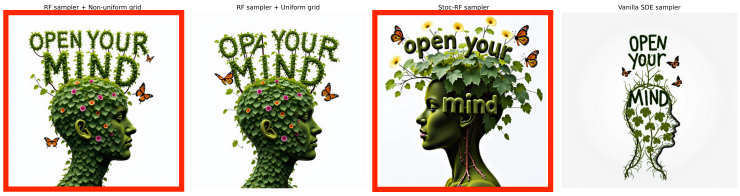
where the first identity arises from the basic property of the TV distance.

Step 2: Simplifying $\mathsf{TV}(p_{X_{i+1}}, p_{\Phi_i(X_i)})$**.** Next, we recall that $v_i(x) = \frac{x}{t_i} + \frac{1-t_i}{t_i} s_i(x)$. Therefore, by [44, Lemma 13] we have

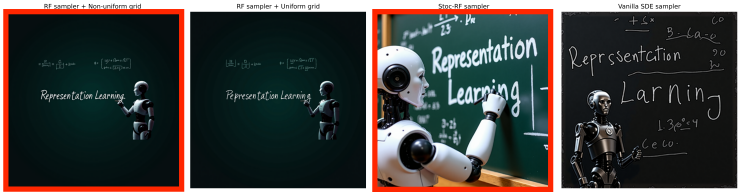
$$\frac{\partial v_i(x)}{\partial x} = -\frac{1}{1-t_i} I + \frac{t_i}{(1-t_i)^3} \mathrm{Cov}_{1|t_i}(x_i) \tag{26}$$



(a) **Prompt:** A photo of a cat holding a sign Rectified flow. ($N = 50$)



(b) **Prompt:** Grape vines in the shape of text 'open your mind' sprouting out of a head with flowers and butterflies. DSLR photo. ($N = 50$)



(c) **Prompt:** Photo of a robot lecturer writing the words 'Representation Learning' in cursive on a blackboard, with math formulas and diagrams. ($N = 50$)



(d) **Prompt:** A car made of cheese on a wooden table. ($N = 80$)

Figure 6: Generated images using Flux for specific prompts via Euler's scheme (5) under non-uniform time-grid (15) (Columns 1), uniform time-grid (Column 2), STOC-RF sampler (Column 3), and vanilla SDE sampler (Column 4). Red color indicates the proposed methods in this paper. The generation quality of Stoc-Rf is generally better out of the four samplers.

Further, for any vector $u \in \mathbb{R}^d$ with $\|u\|_2 = 1$ Assumption 4.4(a) tells us that for $i \leq N-2$

$$u^\top \frac{\partial \Phi_i(x)}{\partial x} u = u^\top \left(I + \eta_i \frac{\partial \widehat{v}_i(x)}{\partial x} \right) u \geq \left(1 - \frac{1}{2} \right) \|u\|_2^2 = \frac{1}{2}.$$

Here, the penultimate relation follows from the positive semi-definiteness of the covariance matrix $\mathrm{Cov}_{1|t}(x)$ and the inequality $\frac{\eta_i}{1-t_i} \leq \frac{\eta_i}{1-t_{i+1}} \leq h$ (*cf.* Assumption 4.5).

Invertibility of $\Phi_i(\cdot)$: The above inequality implies that $\frac{\partial \Phi_i(x)}{\partial x}$ is positive definite uniformly over all x. This entails that $\Phi_i(\cdot)$ is *proper*, i.e, $\lim_{\|x\|_2 \to \infty} \|\Phi(x)\|_2 = \infty$. To see this, we define $\phi(t) := u^\top \Phi_i(tu)$ for a fixed *unit vector* u.Then, we have $\phi'(t) = u^\top \nabla \Phi_i(tu) u > 1/2$. This shows that

$$\phi(t) - \phi(0) = \int_0^t \phi'(u)\, du \geq t/2.$$

An application Cauchy-Schwarz inequality yields

$$\|\Phi_i(tu)\|_2 \geq u^\top \Phi_i(tu) \geq t/2 + u^\top \Phi_i(0) \geq t/2 - \|\Phi_i(0)\|_2.$$

Table 3: Ablation study of time discretization schedules on five T2I prompts from Appendix B.2. Best results for each prompt are highlighted in bold.

Time Grid	CLIP (↑)	PickScore (↑)	CER (↓)
Prompt: Cat with sign			
Linear	$\mathbf{35.54 \pm 1.93}$	0.2402 ± 0.0068	$\mathbf{0.6851 \pm 0.0454}$
Geometric	35.44 ± 0.98	0.2401 ± 0.0025	0.7064 ± 0.0159
Cosine	35.27 ± 2.33	$\mathbf{0.2414 \pm 0.0033}$	0.7489 ± 0.0834
Logit	35.24 ± 0.94	0.2403 ± 0.0023	0.7064 ± 0.0085
Prompt: T-shirt ("I love AI")			
Linear	39.59 ± 2.67	0.2389 ± 0.0072	0.8444 ± 0.1333
Geometric	40.87 ± 1.22	0.2414 ± 0.0025	$\mathbf{0.7111 \pm 0.1333}$
Cosine	$\mathbf{41.64 \pm 0.82}$	0.2374 ± 0.0042	0.7778 ± 0.1988
Logit	40.84 ± 1.11	$\mathbf{0.2417 \pm 0.0023}$	$\mathbf{0.7111 \pm 0.1663}$
Prompt: Robot Lecturer			
Linear	31.42 ± 1.49	0.2247 ± 0.0112	$\mathbf{2.5739 \pm 0.9245}$
Geometric	30.61 ± 1.57	0.2267 ± 0.0079	2.9304 ± 1.3950
Cosine	31.04 ± 1.41	$\mathbf{0.2318 \pm 0.0048}$	2.6696 ± 0.8947
Logit	$\mathbf{31.45 \pm 1.54}$	0.2254 ± 0.0099	2.9304 ± 1.5200
Prompt: Cheese car			
Linear	33.52 ± 1.85	0.2366 ± 0.0043	$\mathbf{1.0000 \pm 0.0000}$
Geometric	$\mathbf{33.77 \pm 0.99}$	0.2356 ± 0.0043	$\mathbf{1.0000 \pm 0.0000}$
Cosine	32.01 ± 1.39	$\mathbf{0.2390 \pm 0.0054}$	$\mathbf{1.0000 \pm 0.0000}$
Logit	31.61 ± 1.13	0.2333 ± 0.0038	$\mathbf{1.0000 \pm 0.0000}$
Prompt: Sign ("Chimpanzees")			
Linear	38.71 ± 0.65	0.2066 ± 0.0044	0.5224 ± 0.0778
Geometric	38.82 ± 0.79	0.2069 ± 0.0038	0.5254 ± 0.1385
Cosine	38.42 ± 0.35	$\mathbf{0.2101 \pm 0.0063}$	0.5433 ± 0.0889
Logit	$\mathbf{39.06 \pm 0.99}$	0.2075 ± 0.0040	$\mathbf{0.4627 \pm 0.0433}$

Table 4: Quality–speed trade-off between deterministic RF (ODE) and STOC-RF (SDE) on two T2I prompts. Sampling time is measured on an NVIDIA L40S GPU. NFE denotes the number of function evaluations.

	Deterministic RF (ODE + Geometric)				**STOC-RF (SDE)**			
NFE	Time (s)	CLIP (↑)	PickScore (↑)	CER (↓)	Time (s)	CLIP (↑)	PickScore (↑)	CER (↓)
Prompt: Cat with sign								
5	2.8	31.01	0.2124	0.8553	2.1	23.15	0.1910	1.0000
10	5.5	34.15	0.2326	0.7745	4.8	23.15	0.1910	1.0000
20	11.4	34.35	0.2399	0.7362	10.8	18.46	0.1671	1.0000
30	17.0	34.31	0.2403	0.7021	16.3	34.97	0.2294	0.8468
50	27.9	35.44	0.2401	0.7064	27.4	32.39	0.2333	0.8511
80	45.0	36.28	0.2398	0.6766	44.7	33.63	0.2379	0.8340
Prompt: Sign ("Chimpanzees")								
5	2.7	23.02	0.1729	0.8597	2.1	23.30	0.1949	1.0000
10	5.5	34.11	0.1975	0.6388	4.8	23.30	0.1949	1.0000
20	11.1	37.59	0.2071	0.4657	10.5	18.99	0.1527	1.0000
30	16.5	38.57	0.2075	0.4478	16.1	38.31	0.2002	0.5522
50	28.3	38.82	0.2069	0.5254	27.4	37.06	0.2013	0.6358
80	45.0	39.36	0.2074	0.5343	44.6	37.58	0.2052	0.5552

Setting $t = \|x\|_2$ and $u = x/\|x\|_2$, we have $\|\Phi_i(x)\|_2 \geq \|x\|_2/2 - \|\Phi_i(0)\|_2$. This implies that $\Phi_i(\cdot)$ is proper. Then, by Theorem 2.2 of [59] shows that $\Phi_i(\cdot)$ is a diffeomorphism. Therefore, we can find a unique x_i such that $\Phi_i(x_i) = x_{i+1}$, which in turn allows us to derive

$$p_{\Phi_i(X_i)}(x_{i+1}) = p_{X_i}\left(\Phi_i^{-1}(x_{i+1})\right) \cdot \det\left(\frac{\partial \Phi_i^{-1}(x_{i+1})}{\partial x_{i+1}}\right) = p_{X_i}(x_i) \cdot \det\left(\frac{\partial x_i}{\partial x_{i+1}}\right)$$

Consequently, we see that: for any $i \geq 1$,

$$\begin{aligned}
p_{\Phi_i(X_i)}\left(x_{i+1}\right) - p_{X_{i+1}}\left(x_{i+1}\right) &= p_{X_i}\left(x_i\right) \det\left(\frac{\partial x_i}{\partial x_{i+1}}\right) - p_{X_{i+1}}\left(x_{i+1}\right) \\
&= \int \left\{ p_{X_i|X_N}\left(x_i \mid x_N\right) \det\left(\frac{\partial x_i}{\partial x_{i+1}}\right) - p_{X_{i+1}|X_N}\left(x_{i+1} \mid x_N\right) \right\} p_{X_N}\left(x_N\right) \mathrm{d}x_N \\
&= \int \Big\{ 1 - \underbrace{\frac{p_{X_{i+1}|X_N}\left(x_{i+1} \mid x_N\right)}{p_{X_i|X_N}\left(x_i \mid x_N\right)} \det\left(\frac{\partial x_{i+1}}{\partial x_i}\right)}_{=:\mathcal{T}\left(x_i, x_N\right)} \Big\} p_{X_i, X_N}\left(x_i, x_N\right) \det\left(\frac{\partial x_i}{\partial x_{i+1}}\right) \mathrm{d}x_N \\
&= \int \left\{1 - \mathcal{T}\left(x_i, x_N\right)\right\} p_{X_i, X_N}\left(x_i, x_N\right) \det\left(\frac{\partial x_i}{\partial x_{i+1}}\right) \mathrm{d}x_N
\end{aligned} \tag{27}$$

Step 3: Analysis of $\mathcal{T}(x_i, x_N)$. The next step is then to analyze $\mathcal{T}\left(x_i, x_N\right)$. Controlling $\mathcal{T}\left(x_i, x_N\right)$: Given how X_{i+1} and X_i are generated, we observe that

$$\begin{aligned}
\mathcal{T}\left(x_{i+1}, x_N\right) &= \frac{p_{X_{i+1}|X_N}\left(x_{i+1} \mid x_N\right)}{p_{X_i|X_N}\left(x_i \mid x_N\right)} \det\left(\frac{\partial x_{i+1}}{\partial x_i}\right) = \frac{\left(\frac{1}{1-t_{i+1}}\right)^d \exp\left\{\frac{-\left\|x_{i+1} - t_{i+1} x_N\right\|_2^2}{2\left(1-t_{i+1}\right)^2}\right\}}{\left(\frac{1}{1-t_i}\right)^d \exp\left\{\frac{-\left\|x_i - t_i x_N\right\|_2^2}{2\left(1-t_i\right)^2}\right\}} \det\left(\frac{\partial x_{i+1}}{\partial x_i}\right) \\
&= \underbrace{\left(\frac{1-t_i}{1-t_{i+1}}\right)^d \det\left(\frac{\partial x_{i+1}}{\partial x_i}\right)}_{=:\mathcal{T}_1\left(x_i, x_N\right)} \underbrace{\exp\left\{\frac{\left\|x_i - t_i x_N\right\|_2^2}{2\left(1-t_i\right)^2} - \frac{\left\|x_{i+1} - t_{i+1} x_N\right\|_2^2}{2\left(1-t_{i+1}\right)^2}\right\}}_{=:\mathcal{T}_2\left(x_t, x_0\right)}
\end{aligned}$$

leaving us with two terms to control. Let us first study the term $\mathcal{T}_1\left(x_t, x_0\right)$, towards which we see that

$$\begin{aligned}
\mathcal{T}_1\left(x_t, x_0\right) &= \left(\frac{1-t_i}{1-t_{i+1}}\right)^d \det\left(\frac{\partial(x_i + \eta_i \widehat{v}_i(x_i))}{\partial x_i}\right) \\
&= \left(\frac{1-t_i}{1-t_{i+1}}\right)^d \det\left(I + \eta_t \frac{\partial}{\partial x_t} v_i\left(x_i\right) + \eta_i \left(\frac{\partial}{\partial x_i} \widehat{v}_i\left(x_i\right) - \frac{\partial}{\partial x_i} v_i\left(x_i\right)\right)\right) \\
&\overset{(a)}{=} \left(\frac{1-t_i}{1-t_{i+1}}\right)^d \det\left(\left(1 - \frac{\eta_i}{1-t_i}\right) I + \frac{\eta_i t_i}{\left(1-t_i\right)^3} \operatorname{Cov}_{1|t_i}(x) + \eta_t \varepsilon_i^J\left(x_i\right)\right) \\
&= \det\left(\frac{1-t_i-\eta_i}{1-t_{i+1}} I + \frac{\eta_i t_i}{(1-t_{i+1})(1-t_i)^2} \operatorname{Cov}_{1|t_i}\left(x_i\right) + \frac{\eta_i(1-t_i)}{(1-t_{i+1})} \varepsilon_i^J\left(x_t\right)\right) \\
&\overset{(b)}{=} \det\left(I + \frac{\eta_i t_i}{(1-t_{i+1})(1-t_i)^2} \operatorname{Cov}_{1|t_i}\left(x_i\right) + \frac{\eta_i(1-t_i)}{(1-t_{i+1})} \varepsilon_i^J\left(x_i\right)\right).
\end{aligned}$$

Here, (a) arises from Tweedie's formula, whereas (b) follows since

$$\frac{1-t_i-\eta_i}{1-t_{i+1}} = \frac{1-t_{i+1}}{1-t_{i+1}} = 1.$$

Next, we turn attention to the term $\mathcal{T}_2\left(x_t, x_0\right)$, which satisfies

$$\begin{aligned}
\log \mathcal{T}_2\left(x_i, x_N\right) &= \frac{\left\|x_i - t_i x_N\right\|_2^2}{2\left(1-t_i\right)^2} - \frac{\left\|x_{i+1} - t_{i+1} x_N\right\|_2^2}{2\left(1-t_{i+1}\right)^2} \\
&= \frac{\left\|x_i - t_i x_N\right\|_2^2}{2\left(1-t_i\right)^2} - \frac{\left\|x_i + \eta_i \widehat{v}_i(x_i) - t_{i+1} x_N\right\|_2^2}{2\left(1-t_{i+1}\right)^2}
\end{aligned} \tag{28}$$

We focus on the term $x_i + \eta_i \widehat{v}_i(x_i) - t_{i+1} x_N$:

$$\begin{aligned} x_i + \eta_i \widehat{v}_i(x_i) - t_{i+1} x_N &= \left(1 - \frac{\eta_i}{1-t_i}\right) x_i + \frac{\eta_i}{1-t_i} \mu_{1|t_i}(x_i) - (t_i + \eta_i) x_N + \eta_i \left(\widehat{v}_i(x_i) - v_i(x_i)\right) \\ &= \left(1 - \frac{\eta_i}{1-t_i}\right)(x_i - t_i x_N) + \frac{\eta_i}{1-t_i} \mu_{1|t_i}(x_i) - \frac{\eta_i}{1-t_i} t_i x_N - \eta_i x_N + \eta_t \left(\widehat{v}_i(x_i) - v_i(x_i)\right) \\ &= \left(\frac{1-t_{i+1}}{1-t_i}\right)(x_i - t_i x_N) + \frac{\eta_i}{1-t_i}\left(\mu_{1|t_i}(x_i) - x_N\right) + \eta_i \varepsilon_i^{\mathrm{v}}(x_i) \end{aligned}$$

Substitution into (28) yields

$$\begin{aligned} &\log \mathcal{T}_2(x_i, x_N) \\ &= -\frac{\eta_i}{(1-t_i)^2(1-t_{i+1})}(x_i - t_i x_N)^\top (\mu_{1|t_i}(x_i) - x_N) - \frac{\eta_i}{(1-t_i)(1-t_{i+1})}(x_i - t_i \mu_{1|t_i}(x_i))^\top \varepsilon_i^{\mathrm{v}}(x_i) \\ &- \frac{\eta_i t_{i+1}}{(1-t_{i+1})^2}\left(\mu_{1|t_i}(x_i) - x_N\right) \varepsilon_i^{\mathrm{v}}(x_i) - \frac{\eta_i^2 \|\mu_{1|t_i}(x_i) - x_N\|_2^2}{2(1-t_i)^2(1-t_{i+1})^2} - \frac{\eta_i^2 \|\varepsilon_i^{\mathrm{v}}(x_i)\|_2^2}{2(1-t_{i+1})^2}. \end{aligned} \tag{29}$$

Now Define:

$$\begin{aligned} \xi(x_i, x_N) := &-\frac{\eta_i (x_i - t_i x_N)^\top (\mu_{1|t_i}(x_i) - x_N)}{(1-t_i)^2(1-t_{i+1})} - \frac{\eta_i t_{i+1}\left(\mu_{1|t_i}(x_i) - x_N\right) \varepsilon_i^{\mathrm{v}}(x_i)}{(1-t_{i+1})^2} \\ &- \frac{\eta_i^2 \|\mu_{1|t_i}(x_i) - x_N\|_2^2}{2(1-t_i)^2(1-t_{i+1})^2} + \frac{\eta_i}{(1-t_i)^2(1-t_{i+1})}\left(t_i + \frac{\eta_i}{2(1-t_{i+1})}\right) \mathrm{Tr}(\mathrm{Cov}_{1|t_i}) \end{aligned}$$

It is easy to see that

$$\int \xi(x_i, x_N) p_{X_N|X_i=x_i}(x_N)\, dx_N = 0.$$

Thus, based on (29), we can further simplify $\log \mathcal{T}_2(x_i, x_N)$ as follows:

$$\begin{aligned} \log \mathcal{T}_2(x_i, x_N) = &\xi(x_i, x_N) - \frac{\eta_i (x_i - t_i \mu_{1|t_i}(x_i))^\top \varepsilon_i^{\mathrm{v}}(x_i)}{(1-t_i)(1-t_{i+1})} - \frac{\eta_i^2 \|\varepsilon_i^{\mathrm{v}}(x_i)\|_2^2}{2(1-t_{i+1})^2} \\ &- \frac{\eta_i}{(1-t_i)^2(1-t_{i+1})}\left(t_i + \frac{\eta_i}{2(1-t_{i+1})}\right) \mathrm{Tr}(\mathrm{Cov}_{1|t_i}) \end{aligned}$$

Now, for ease of presentation, define

$$\begin{aligned} W(x_i) := &\log\det\left(I + \frac{\eta_i t_i}{(1-t_{i+1})(1-t_i)^2}\,\mathrm{Cov}_{1|t_i}(x_i) + \frac{\eta_i(1-t_i)}{(1-t_{i+1})}\varepsilon_i^J(x_i)\right) - \frac{\eta_i^2 \|\varepsilon_i^{\mathrm{v}}(x_i)\|_2^2}{2(1-t_{i+1})^2} \\ &- \frac{\eta_i}{(1-t_i)^2(1-t_{i+1})}\left(t_i + \frac{\eta_i}{2(1-t_{i+1})}\right) \mathrm{Tr}(\mathrm{Cov}_{1|t_i}) - \frac{\eta_i (x_i - t_i \mu_{1|t_i}(x_i))^\top \varepsilon_i^{\mathrm{v}}(x_i)}{(1-t_i)(1-t_{i+1})} \end{aligned}$$

Then for any $x_i \in \mathcal{X}_{\text{data}}$, it holds that

$$\int_{x_N} \left(1 - e^{\xi(x_t, x_N)}\right) e^{W(x_i)} p_{X_N|X_i}(x_N \mid x_i)\, \mathrm{d}x_N \le -e^{W(x_i)} \int_{x_N} \xi(x_i, x_N)\, p_{X_N|X_i}(x_N \mid x_i)\, \mathrm{d}x_N = 0 \tag{30}$$

where the inequality results from the elementary inequality $1 - e^x \le -x$ for all $x \in \mathbb{R}$. Using the above quantities and (27) we get: for any $\mathcal{A} \subseteq \mathcal{X}_{\text{data}} \cap \mathcal{E}_i$,

$$
\begin{aligned}
&\mathbb{P}_{\Phi_i(X_i)}(\mathcal{A}) - \mathbb{P}_{X_{i+1}}(\mathcal{A}) \\
&= \int_{\mathcal{A}} \{1 - \mathcal{T}(x_i, x_N)\}\, p_{X_i, X_N}(x_i, x_N) \det\left(\frac{\partial x_i}{\partial x_{i+1}}\right) \mathrm{d}x_N\, \mathrm{d}x_{i+1} \\
&= \int_{\mathcal{A}\times\mathcal{X}_{\text{data}}} \{1 - \mathcal{T}(x_i, x_N)\}\, p_{X_N|X_i=x_i}(x_N)\, p_{X_i}(x_i) \det\left(\frac{\partial x_i}{\partial x_{i+1}}\right) \mathrm{d}x_N\, \mathrm{d}x_{i+1} \\
&= \int_{\Phi_i^{-1}(\mathcal{A})\times\mathcal{X}_{\text{data}}} \{1 - \mathcal{T}(x_i, x_N)\}\, p_{X_N|X_i=x_i}(x_N)\, p_{X_i}(x_i) \mathrm{d}x_N\, \mathrm{d}x_i \\
&= \int_{\Phi_i^{-1}(\mathcal{A})\times\mathcal{X}_{\text{data}}} \left\{1 - e^{\xi(x_i,x_N)} \cdot e^{W(x_i)}\right\} p_{X_N|X_i=x_i}(x_N)\, p_{X_i}(x_i) \mathrm{d}x_N\, \mathrm{d}x_i \\
&= \int_{\Phi_i^{-1}(\mathcal{A})\times\mathcal{X}_{\text{data}}} \left\{\left(1 - e^{\xi(x_i,x_N)}\right) e^{W(x_i)} + \left(1 - e^{W(x_i)}\right)\right\} p_{X_N|X_i=x_i}(x_N)\, p_{X_i}(x_i) \mathrm{d}x_N\, \mathrm{d}x_i \\
&\overset{(a)}{\leq} \int_{x_i, x_N \in \Phi_i^{-1}(\mathcal{A})\times\mathcal{X}_{\text{data}}} \left\{1 - e^{W(x_i)}\right\} p_{X_i, X_N}(x_i, x_N)\, \mathrm{d}x_i\, \mathrm{d}x_N \\
&\leq \int_{x_i \in \Phi_i^{-1}(\mathcal{A})} -W(x_i)\, p_{X_i}(x_i)\, \mathrm{d}x_i,
\end{aligned}
$$

where, (a) invokes (30), and the last inequality follows from the elementary inequality $1 - e^x \leq -x$.

Step 4: Analyzing $W(x_i)$. First, we introduce the following lemma.

Lemma C.1. *Let $M \in \mathbb{R}^{d\times d}$ satisfy $(M + M^\top)/2 \succeq I/2$. Then $\det M > 0$, and*

$$\log\det M \geq \mathrm{Tr}(M - I) - 2\|M - I\|_F^2.$$

Proof. Write $B = M - I$ and $M(s) = I + sB$ for $s \in [0, 1]$. The symmetric part of $M(s)$ is bounded below by $(1 - s/2)I$. Thus, for every $u \in \mathbb{R}^d$,

$$\|u\|_2 \|M(s)u\|_2 \geq u^\top M(s) u \geq (1 - s/2)\|u\|_2^2.$$

Hence $M(s)$ is invertible and $\|M(s)^{-1}\|_{\text{op}} \leq 2$. Since $\det M(0) = 1$, continuity gives $\det M(s) > 0$ along the entire path. Setting $f(s) = \log\det M(s)$, we obtain

$$f'(s) = \mathrm{Tr}(M(s)^{-1}B), \qquad f''(s) = -\,\mathrm{Tr}\left((M(s)^{-1}B)^2\right) \geq -4\|B\|_F^2,$$

where we used $\mathrm{Tr}(C^2) \leq \|C\|_F^2$ for any real square matrix C, and $\|M(s)^{-1}B\|_F \leq 2\|B\|_F$. Therefore, Taylor's formula yields

$$\log\det M = f(0) + f'(0) + \int_0^1 (1 - s) f''(s)\, ds \geq \mathrm{Tr}(B) - 2\|B\|_F^2.$$

□

To apply the lemma, recall that $\varepsilon_i^J(x_i) = \nabla\widehat{v}_{t_i}(x_i) - \nabla v_{t_i}(x_i)$. Using (26) and $\eta_i = t_{i+1} - t_i$, the matrix inside the log determinant can be written as

$$
\begin{aligned}
M_i &:= I + \frac{\eta_i t_i}{(1 - t_{i+1})(1 - t_i)^2}\mathrm{Cov}_{1|t_i}(x_i) + \frac{\eta_i(1 - t_i)}{1 - t_{i+1}}\varepsilon_i^J(x_i) \\
&= \frac{1 - t_i}{1 - t_{i+1}}\left(I + \eta_i \nabla\widehat{v}_{t_i}(x_i)\right).
\end{aligned}
$$

Assumption 4.4(a) gives, for every $u \in \mathbb{R}^d$,

$$u^\top M_i u \geq \frac{1 - t_i}{2(1 - t_{i+1})}\|u\|_2^2 \geq \frac{1}{2}\|u\|_2^2.$$

Thus $(M_i + M_i^\top)/2 \succeq I/2$, and Lemma C.1 applies. Using also $\|A + D\|_F^2 \le 2\|A\|_F^2 + 2\|D\|_F^2$ for the two terms in $M_i - I$, we have

$$
\begin{aligned}
&- \log\det\left(I + \frac{\eta_i t_i}{(1-t_{i+1})(1-t_i)^2}\,\mathrm{Cov}_{1|t_i}\,(x_i) + \frac{\eta_i(1-t_i)}{(1-t_{i+1})}\varepsilon_i^J\,(x_i)\right)\\
&\le \frac{4\eta_i^2 t_i^2}{(1-t_{i+1})^2(1-t_i)^4}\|\mathrm{Cov}_{1|t_i}\|_F^2 + \frac{4\eta_i^2(1-t_i)^2}{(1-t_{i+1})^2}\|\varepsilon_i^J(x_i)\|_F^2\\
&\quad - \frac{\eta_i t_i}{(1-t_{i+1})(1-t_i)^2}\,\mathrm{Tr}(\mathrm{Cov}_{1|t_i}) - \frac{\eta_i(1-t_i)}{(1-t_{i+1})}\,\mathrm{Tr}(\varepsilon_i^J(x_i)),
\end{aligned}
$$

This yields

$$
\begin{aligned}
-W\,(x_i) &= -\log\det\left(I + \frac{\eta_i t_i}{(1-t_{i+1})(1-t_i)^2}\,\mathrm{Cov}_{1|t_i}\,(x_i) + \frac{\eta_i(1-t_i)}{(1-t_{i+1})}\varepsilon_i^J\,(x_i)\right) + \frac{\eta_i^2\|\varepsilon_i^{\mathrm{v}}(x_i)\|_2^2}{2(1-t_{i+1})^2}\\
&\quad + \frac{\eta_i}{(1-t_i)^2(1-t_{i+1})}\left(t_i + \frac{\eta_i}{2(1-t_{i+1})}\right)\mathrm{Tr}(\mathrm{Cov}_{1|t_i}) + \frac{\eta_i(x_i - t_i\mu_{1|t_i}(x_i))^\top\varepsilon_i^{\mathrm{v}}(x_i)}{(1-t_i)(1-t_{i+1})}\\
&\le \frac{4\eta_i^2 t_i^2}{(1-t_{i+1})^2(1-t_i)^4}\|\mathrm{Cov}_{1|t_i}(x_i)\|_F^2 + \frac{4\eta_i^2(1-t_i)^2}{(1-t_{i+1})^2}\|\varepsilon_i^J(x_i)\|_F^2 + \frac{\eta_i^2\|\varepsilon_i^{\mathrm{v}}(x_i)\|_2^2}{2(1-t_{i+1})^2}\\
&\quad + \frac{\eta_i(x_i - t_i\mu_{1|t_i}(x_i))^\top\varepsilon_i^{\mathrm{v}}(x_i)}{(1-t_i)(1-t_{i+1})} + \frac{\eta_i}{(1-t_i)^2(1-t_{i+1})}\left(t_i + \frac{\eta_i}{2(1-t_{i+1})} - t_i\right)\mathrm{Tr}(\mathrm{Cov}_{1|t_i}(x_i))\\
&\quad - \frac{\eta_i(1-t_i)}{(1-t_{i+1})}\,\mathrm{Tr}(\varepsilon_i^J(x_i))\\
&\le \frac{4\eta_i^2 t_i^2}{(1-t_{i+1})^2(1-t_i)^4}\|\mathrm{Cov}_{1|t_i}(x_i)\|_F^2 + \frac{4\eta_i^2(1-t_i)^2}{(1-t_{i+1})^2}\|\varepsilon_i^J(x_i)\|_F^2 + \frac{\eta_i^2\|\varepsilon_i^{\mathrm{v}}(x_i)\|_2^2}{2(1-t_{i+1})^2}\\
&\quad + \frac{\eta_i^2}{2(1-t_i)^2(1-t_{i+1})^2}\,\mathrm{Tr}(\mathrm{Cov}_{1|t_i}(x_i)) + \underbrace{\frac{\eta_i(x_i - t_i\mu_{1|t_i}(x_i))^\top\varepsilon_i^{\mathrm{v}}(x_i)}{(1-t_i)(1-t_{i+1})}}_{\Delta(\varepsilon_i^{\mathrm{v}}(x_i))}\\
&\quad - \frac{\eta_i(1-t_i)}{(1-t_{i+1})}\,\mathrm{Tr}(\varepsilon_i^J(x_i))\\
&\le \frac{4\eta_i^2 t_i^2}{(1-t_{i+1})^2(1-t_i)^4}\|\mathrm{Cov}_{1|t_i}(x_i)\|_F^2 + \frac{4\eta_i^2(1-t_i)^2}{(1-t_{i+1})^2}\|\varepsilon_i^J(x_i)\|_F^2 + \frac{\eta_i^2\|\varepsilon_i^{\mathrm{v}}(x_i)\|_2^2}{2(1-t_{i+1})^2}\\
&\quad + \frac{\eta_i^2}{2(1-t_i)^2(1-t_{i+1})^2}\,\mathrm{Tr}(\mathrm{Cov}_{1|t_i}(x_i)) + \underbrace{\frac{\eta_i(1-t_i)}{(1-t_{i+1})}\frac{(x_i - t_i\mu_{1|t_i}(x_i))^\top\varepsilon_i^{\mathrm{v}}(x_i)}{(1-t_i)^2}}_{\Delta(\varepsilon_i^{\mathrm{v}}(x_i))}\\
&\quad - \frac{\eta_i(1-t_i)}{(1-t_{i+1})}\,\mathrm{Tr}(\varepsilon_i^J(x_i))
\end{aligned}
\tag{31}
$$

Step 5: Controlling the error arising from $\Delta(\varepsilon_i^{\mathrm{v}}(x_i))$. We start with the following integral quantity for any set $\mathcal{A}$:

$$\begin{aligned}
&\int_{x_i\in\mathcal{A}} \frac{1}{(1-t_i)^2}(x_i - t_i\mu_{1|t_i}(x_i))^\top \varepsilon_i^{\mathrm{v}}(x_i) p_{X_i}(x_i)\,\mathrm{d}x_i \\
&= \int_{x_i,x_N\in\mathcal{A}\times\mathcal{X}_{\mathrm{data}}} \frac{1}{(1-t_i)^2}(x_i - t_i\mu_{1|t_i}(x_i))^\top \varepsilon_i^{\mathrm{v}}(x_i) p_{X_i,X_N}(x_i,x_N)\,\mathrm{d}x_i\,\mathrm{d}x_N \\
&= \int_{x_i,x_N\in\mathcal{A}\times\mathcal{X}_{\mathrm{data}}} \frac{1}{(1-t_i)^2}(x_i - t_i\mu_{1|t_i}(x_i))^\top \varepsilon_i^{\mathrm{v}}(x_i) p_{X_i|X_N=x_N}(x_i) p_{X_N}(x_N)\,\mathrm{d}x_i\,\mathrm{d}x_N \\
&= \int_{x_i,x_N\in\mathcal{A}\times\mathcal{X}_{\mathrm{data}}} \frac{1}{(1-t_i)^2}\left\{(x_i - t_i x_N)^\top \varepsilon_i^{\mathrm{v}}(x_i) p_{X_i|X_N=x_N}(x_i)\,\mathrm{d}x_i\right\} p_{X_N}(x_N)\,\mathrm{d}x_N \\
&\le \int_{x_i,x_N\in\mathbb{R}^d\times\mathcal{X}_{\mathrm{data}}} \frac{1}{(1-t_i)^2}\left\{\left|(x_i - t_i x_N)^\top \varepsilon_i^{\mathrm{v}}(x_i)\right| p_{X_i|X_N=x_N}(x_i)\,\mathrm{d}x_i\right\} p_{X_N}(x_N)\,\mathrm{d}x_N \\
&\le \left(\int_{x_i,x_N\in\mathbb{R}^d\times\mathcal{X}_{\mathrm{data}}} \left\{\frac{1}{(1-t_i)^4}\left|(x_i - t_i x_N)^\top \varepsilon_i^{\mathrm{v}}(x_i)\right|^2 p_{X_i|X_N=x_N}(x_i)\,\mathrm{d}x_i\right\} p_{X_N}(x_N)\,\mathrm{d}x_N\right)^{1/2} \\
&= \left(\int_{x_i,x_N\in\mathbb{R}^d\times\mathcal{X}_{\mathrm{data}}} \left\{\frac{1}{(1-t_i)^4}\left\langle (x_i - t_i x_N)(x_i - t_i x_N)^\top, \varepsilon_i^{\mathrm{v}}(x_i)\varepsilon_i^{\mathrm{v}}(x_i)^\top\right\rangle p_{X_i|X_N=x_N}(x_i)\,\mathrm{d}x_i\right\} p_{X_N}(x_N)\,\mathrm{d}x_N\right)^{1/2}
\end{aligned} \tag{32}$$

Now, recall that $X_i \mid X_N = x_N \sim N(t_i x_N, (1-t_i)^2 I)$. Using standard techniques from [38, Section B.5], we can obtain bound on (32):

$$\int_{x_i\in\mathcal{A}} \frac{1}{(1-t_i)^2}(x_i - t_i\mu_{1|t_i}(x_i))^\top \varepsilon_i^{\mathrm{v}}(x_i) p_{X_i}(x_i)\,\mathrm{d}x_i \le \frac{2\varepsilon_i^{\mathrm{v}}}{(1-t_i)} + \varepsilon_i^{\mathrm{J},1} + \varepsilon_i^{\mathrm{J},2} + \varepsilon_i^{\mathrm{H}}.$$

Substituting this in (31), we have

$$\begin{aligned}
&\int_{x_i\in\mathcal{A}} \left\{\Delta(\varepsilon_i^{\mathrm{v}}(x_i)) - \frac{\eta_i(1-t_i)}{(1-t_{i+1})}\operatorname{Tr}(\varepsilon_i^{J}(x_i))\right\} p_{X_i}(x_i) \\
&\le \frac{2\varepsilon_i^{\mathrm{v}}}{(1-t_i)} + \varepsilon_i^{\mathrm{J},1} + \varepsilon_i^{\mathrm{J},2} + \varepsilon_i^{\mathrm{H}} + \frac{\eta_i(1-t_i)}{(1-t_{i+1})}\left(\int \operatorname{Tr}(\varepsilon_i^{J}(x_i))^2 p_{X_i}(x_i)\mathrm{d}x_i\right)^{1/2} \\
&\le \frac{2\eta_i}{(1-t_{i+1})}(\varepsilon_i^{\mathrm{v}} + \varepsilon_i^{\mathrm{J},1} + \varepsilon_i^{\mathrm{J},2} + \varepsilon_i^{\mathrm{H}}).
\end{aligned}$$

Here, we used the fact that $\frac{\eta_i(1-t_i)}{(1-t_{i+1})} \le \frac{\eta_i}{(1-t_{i+1})}$.

Step 6: Bounding the terms involving $\mathbf{Cov}_{1|t_i}(x_i)$. Here we always assume that $k \ge \log d$.[1] We begin defining some quantities first.

- Let $\{x_i^*\}_{1\le i\le N_{\epsilon_0}}$ be an ϵ_0-net of $\mathcal{X}_{\mathrm{data}}$, with N_{ϵ_0} denoting its cardinality. Let $\{B_i\}_{1\le i\le N_{\epsilon_0}}$ be a disjoint ϵ_0-cover for $\mathcal{X}_{\mathrm{data}}$ such that $x_i^* \in B_i$ for each i.
- Define the following two sets:

$$\mathcal{I} := \{1 \le i \le N_{\epsilon_0} : \mathbb{P}(X_N \in B_i) \ge \exp(-C_1 k \log N)\} \tag{33}$$

and

$$\mathcal{G} := \left\{w \in \mathbb{R}^d : \|w\|_2 \le 2\sqrt{d} + \sqrt{C_1 k \log N},\ \left|(x_i^* - x_j^*)^\top w\right| \le \sqrt{C_1 k \log N}\,\|x_i^* - x_j^*\|_2,\ \forall\, 1 \le i,j \le N_{\epsilon_0}\right\}. \tag{34}$$

for some sufficiently large universal constant $C_1 > 0$. As it turns out, $\bigcup_{i\in\mathcal{I}} B_i$ and $\mathcal{G}$ form certain high-probability sets related to the random vector $X_N \sim p_1$ and the standard Gaussian random vector in $\mathbb{R}^d$, respectively.

- Recall (24), we can express

$$X_{t_i} \equiv X_i = t_i X_N + (1-t_i)X_0, \tag{35}$$

for some random vector $X_0 \sim \mathcal{N}(0, I_d)$.

[1] If $k < \log d$, then we can redefine k as $\log d$. This does not change the results in Theorem 4.7 and Theorem 4.9.

Next, we define the set $\mathcal{T}_t := \{tv_1 + (1-t)\omega : v_1 \in \cup_{i\in\mathcal{I}} B_i, \omega \in \mathcal{G}\}$.

By the argument of [38, Lemma 1], we have

$$\mathbb{P}\left(X_i \notin \mathcal{T}_{t_i}\right) \le \exp\left(-C_2 k \log N\right), \quad \text{for } C_2 \gg C_{\text{cover}}, \text{ and } 0 \le t_i < 1. \tag{36}$$

For the trace bound, we additionally take the covering resolution sufficiently fine that

$$\epsilon_0 \ll \frac{\delta}{1-\delta} \min\left\{1, \sqrt{\frac{k \log N}{d}}\right\}. \tag{37}$$

Under the normalization $\alpha = t^2/(t^2 + (1-t)^2)$, we have $(1-\alpha)/\alpha = (1-t)^2/t^2$. Thus, the covering-resolution condition used in [38, Appendix G.2] holds uniformly for $t \in [t_1, t_{N-1}]$. The same proof and the posterior-moment argument of [38, Corollary 1] then yield

$$\mathrm{Tr}(\mathrm{Cov}_{1|t_i}(x_i)) \le \frac{C_3(1-t_i)^2}{t_i^2} k \log N, \quad \text{for } t_i \in [t_1, t_{N-1}] \text{ and } x_i \in \mathcal{T}_{t_i}. \tag{38}$$

Step 7: Bound on the term involving squared Frobenius norm using linearization: One can use (38) directly to bound the term $\mathbb{E}_{X_i}\left(\|\mathrm{Cov}_{1|t_i}(X_i)\|_F^2\right)$ in (31). However, this will lead to a quadratic dependence on k. Therefore, we present the following lemma that will yield a linear dependence on k. The proof is deferred to Appendix E.3.

Proposition C.2. *For all $1 \le i \le N-2$, provided $h \le 1/4$ and $16C_3 kh \log N < 1$, the following holds:*

$$\frac{\eta_i t_i}{(1-t_{i+1})(1-t_i)^2}\mathbb{E}(\|\mathit{Cov}_{1|t_i}(X_i)\|_F^2) \le \frac{3}{2}\left(\mathbb{E}[\mathrm{Tr}(\mathit{Cov}_{1|t_i}(X_i))] - \mathbb{E}[\mathrm{Tr}(\mathit{Cov}_{1|t_{i+1}}(X_{i+1}))]\right) + \frac{24h\eta_i t_i}{(1-t_{i+1})^3(1-t_i)^2 N^{10}}.$$

Using Proposition C.2 and retaining the expected covariance traces until summation, we have

$$\begin{aligned}
&\mathbb{P}_{\Phi_i(X_i)}(\mathcal{A}) - \mathbb{P}_{X_{i+1}}(\mathcal{A}) \\
&\le \frac{6\eta_i t_i}{(1-t_i)^2(1-t_{i+1})}\left(\mathbb{E}[\mathrm{Tr}(\mathrm{Cov}_{1|t_i}(X_i))] - \mathbb{E}[\mathrm{Tr}(\mathrm{Cov}_{1|t_{i+1}}(X_{i+1}))]\right) \\
&\quad + \frac{\eta_i^2}{2(1-t_i)^2(1-t_{i+1})^2}\mathbb{E}[\mathrm{Tr}(\mathrm{Cov}_{1|t_i}(X_i))] \\
&\quad + \frac{4\eta_i^2(1-t_i)^2}{(1-t_{i+1})^2}(\varepsilon_i^{\mathrm{J},1})^2 + \frac{\eta_i^2}{2(1-t_{i+1})^2}(\varepsilon_i^{\mathrm{v}})^2 \\
&\quad + \frac{2\eta_i}{(1-t_{i+1})}(\varepsilon_i^{\mathrm{v}} + \varepsilon_i^{\mathrm{J},1} + \varepsilon_i^{\mathrm{J},2} + \varepsilon_i^{\mathrm{H}}) + \frac{96h\eta_i^2 t_i^2}{(1-t_{i+1})^4(1-t_i)^4 N^{10}}.
\end{aligned} \tag{39}$$

We next sum (39) using only Assumption 4.5. Throughout this argument, take $h \le 1/4$ and $16C_3 kh \log N < 1$, as required by Proposition C.2. First, bounded support and (36) extend (38) to

$$\mathbb{E}[\mathrm{Tr}(\mathrm{Cov}_{1|t_i}(X_i))] \le C_3 \frac{(1-t_i)^2}{t_i^2} k \log N + N^{-10}.$$

Indeed, the exceptional-set contribution is at most $R^2 \exp(-C_2 k \log N) \le N^{-10}$ after choosing C_2 sufficiently large relative to c_R.

The coefficients of the trace differences in (39) need not be increasing for a general admissible schedule. However, the schedule assumption gives the increasing upper bound

$$\frac{6\eta_i t_i}{(1-t_i)^2(1-t_{i+1})} \le 24h \frac{t_i^2}{(1-t_i)^2}.$$

For $t_i \le 1/2$, this follows from $\eta_i \le h t_i$ and $1 - t_{i+1} \ge (1-h)/2$; for $t_i > 1/2$, it follows from $\eta_i \le h(1-t_{i+1})$ and $t_i \le 2t_i^2$. Lemma E.1 shows that each expected trace difference is nonnegative. We may therefore apply this upper bound and then collect the equal-index trace terms:

$$\begin{aligned}
&\sum_{i=1}^{N-2} \frac{6\eta_i t_i}{(1-t_i)^2(1-t_{i+1})} \left(\mathbb{E}[\mathrm{Tr}(\mathrm{Cov}_{1|t_i}(X_i))] - \mathbb{E}[\mathrm{Tr}(\mathrm{Cov}_{1|t_{i+1}}(X_{i+1}))]\right) \\
&\le 24h \Bigg\{ \frac{t_1^2}{(1-t_1)^2} \mathbb{E}[\mathrm{Tr}(\mathrm{Cov}_{1|t_1}(X_1))] \\
&\quad + \sum_{i=2}^{N-2} \left(\frac{t_i^2}{(1-t_i)^2} - \frac{t_{i-1}^2}{(1-t_{i-1})^2} \right) \mathbb{E}[\mathrm{Tr}(\mathrm{Cov}_{1|t_i}(X_i))] \\
&\quad - \frac{t_{N-2}^2}{(1-t_{N-2})^2} \mathbb{E}[\mathrm{Tr}(\mathrm{Cov}_{1|t_{N-1}}(X_{N-1}))] \Bigg\} \\
&\le 24 C_3 hk \log N \left\{ 1 + \sum_{i=2}^{N-2} \left[1 - \left(\frac{t_{i-1}(1-t_i)}{t_i(1-t_{i-1})} \right)^2 \right] \right\} + \frac{24h}{\delta^2 N^{10}}.
\end{aligned}$$

In the last step we dropped the nonpositive terminal term, applied the unconditional trace bound only to nonnegative coefficients, and telescoped the exceptional-set coefficients.

For every step covered by the schedule assumption,

$$\frac{\eta_i}{t_i} \le h, \qquad \frac{\eta_i}{1-t_{i+1}} \le 2h.$$

Consequently, using $1 - e^{-x} \le x$ for $x \ge 0$,

$$\begin{aligned}
1 - \left(\frac{t_{i-1}(1-t_i)}{t_i(1-t_{i-1})} \right)^2 &\le 2 \log \left(\frac{t_i(1-t_{i-1})}{t_{i-1}(1-t_i)} \right) \\
&= 2 \log \left(1 + \frac{\eta_{i-1}}{t_{i-1}} \right) + 2 \log \left(1 + \frac{\eta_{i-1}}{1-t_i} \right) \le 6h.
\end{aligned}$$

Thus the sum of the trace-difference terms is at most

$$24 C_3 h(1 + 6Nh) k \log N + \frac{24h}{\delta^2 N^{10}}.$$

For convenience, write

$$s_N := \sum_{i=1}^{N-2} \frac{\eta_i^2}{t_i^2 (1-t_{i+1})^2}. \tag{40}$$

This quantity is used only to collect the linear trace terms and the remainders before applying $s_N \lesssim Nh^2$. The remaining linear trace terms satisfy

$$\sum_{i=1}^{N-2} \frac{\eta_i^2}{2(1-t_i)^2(1-t_{i+1})^2} \mathbb{E}[\mathrm{Tr}(\mathrm{Cov}_{1|t_i}(X_i))] \le \frac{C_3}{2} s_N k \log N + \frac{s_N}{2\delta^2 N^{10}}.$$

Also, the same schedule assumption gives $s_N \lesssim Nh^2$. The corrected remainders in (39) obey

$$\begin{aligned}
\sum_{i=1}^{N-2} \frac{96 h \eta_i^2 t_i^2}{(1-t_{i+1})^4 (1-t_i)^4 N^{10}} &\le \frac{96h}{\delta^6 N^{10}} \sum_{i=1}^{N-2} \frac{\eta_i^2}{t_i^2 (1-t_{i+1})^2} \\
&= \frac{96 h s_N}{\delta^6 N^{10}} \lesssim \frac{h^3}{\delta^6 N^9}.
\end{aligned}$$

Combining these estimates, using the averaged approximation-error assumptions, and iterating (25), we obtain

$$\begin{aligned}
\mathsf{TV}(p_{X_{N-1}}, p_{Y_{N-1}}) \le\ & C_1'(h + Nh^2) k \log N + C_2' N h^2 \varepsilon_{\mathrm{J},1}^2 + C_3' N h^2 \varepsilon_{\mathrm{v}}^2 \\
&+ C_4' N h (\varepsilon_{\mathrm{v}} + \varepsilon_{\mathrm{J},1} + \varepsilon_{\mathrm{J},2} + \varepsilon_{\mathrm{H}}) \\
&+ C_5' \left(\frac{h + Nh^2}{\delta^2 N^{10}} + \frac{h^3}{\delta^6 N^9} \right) + \mathsf{TV}(p_{X_1}, p_{Y_1}).
\end{aligned}$$

The $hk \log N$ term comes from the initial boundary term in the weighted sum. It is retained for a general schedule; no relation between h, N, and δ specific to the geometric schedule has been imposed.

Handling $\mathsf{TV}(p_{X_1}, p_{Y_1})$ separately: Recall that (31) is also valid for $t_0 = 0, t_1 = t_0 + \eta_0 = \eta_0$. Therefore, we have

$$\begin{aligned}
&- W(x_0) \\
&\le \frac{4\eta_0^2}{(1-\eta_0)^2}\|\varepsilon_0^J(x_0)\|_F^2 + \frac{\eta_0^2\|\varepsilon_0^{\mathrm{v}}(x_0)\|_2^2}{2(1-\eta_0)^2} + \frac{\eta_0^2}{2(1-\eta_0)^2}\mathsf{Trace}(\mathrm{Cov}_1) + \underbrace{\frac{\eta_0}{1-\eta_0}x_0^\top \varepsilon_0^{\mathrm{v}}(x_0)}_{\Delta(\varepsilon_0^{\mathrm{v}}(x_0))} \\
&\quad - \frac{\eta_0^2}{(1-\eta_0)}\mathsf{Trace}(\varepsilon_0^J(x_0)).
\end{aligned}$$

Therefore, applying similar argument, we get

$$\mathsf{TV}(p_{X_1}, p_{Y_1}) \le A_1\eta_0^2\mathsf{Trace}(\mathrm{Cov}_1) + A_2\eta_0^2(\varepsilon_0^{\mathrm{v}})^2 + A_3\eta_0^2(\varepsilon_0^{\mathrm{J},1})^2 + A_4(\varepsilon_0^{\mathrm{v}} + \varepsilon_0^{\mathrm{J},1} + \varepsilon_0^{\mathrm{J},2} + \varepsilon_0^{\mathrm{H}}) + \underbrace{\mathsf{TV}(p_{X_0}, p_{Y_0})}_{=0}.$$

Therefore, combining the preceding bounds gives

$$\begin{aligned}
&\mathsf{TV}(p_{X_{N-1}}, p_{Y_{N-1}}) \\
&\le C_1''(h + Nh^2)k\log N + C_2''Nh^2\varepsilon_{\mathrm{J},1}^2 + C_3''Nh^2\varepsilon_{\mathrm{v}}^2 \\
&\quad + C_4''Nh(\varepsilon_{\mathrm{v}} + \varepsilon_{\mathrm{J},1} + \varepsilon_{\mathrm{J},2} + \varepsilon_{\mathrm{H}}) + C_5''\left(\frac{h + Nh^2}{\delta^2 N^{10}} + \frac{h^3}{\delta^6 N^9}\right) \\
&\quad + A_1\eta_0^2\mathrm{Tr}(\mathrm{Cov}_1) + A_2\eta_0^2(\varepsilon_0^{\mathrm{v}})^2 + A_3\eta_0^2(\varepsilon_0^{\mathrm{J},1})^2 + A_4(\varepsilon_0^{\mathrm{v}} + \varepsilon_0^{\mathrm{J},1} + \varepsilon_0^{\mathrm{J},2} + \varepsilon_0^{\mathrm{H}}),
\end{aligned} \tag{41}$$

where $\mathrm{Cov}_1 := \mathrm{Cov}_{X_1 \sim p_1}(X_1)$. Now, note that one can estimate $v_0(x) = \mu_{p_{X_1}} - x$ by $\widehat{v}_0(x) := \widehat{\mu}_1 - x$, where $\widehat{\mu}_1 = \sum_{i=1}^n x_{i,1}/n$ and n is the (training) sample size. In this case, we have

- $\varepsilon_0^{\mathrm{J},1} = \varepsilon_0^{\mathrm{J},2} = \varepsilon^{\mathrm{H}} = 0$.
- $(\varepsilon_0^{\mathrm{v}})^2 = \mathbb{E}\|\widehat{\mu}_1 - \mu_{p_{X_1}}\|_2^2 = O\left(\frac{\mathsf{Trace}(\mathrm{Cov}_1)}{n}\right)$.

This finishes the proof of Theorem 4.6.

C.2.2 Proof of Theorem 4.7

The proof follows by applying the result of the previous section to the time schedule (15).

Recalling properties of time scheduling (15) **(see Lemma E.2):**

1. For all $i \in [N-2]$, we have $\frac{\eta_i}{1-t_i} \le h$ and $\frac{\eta_i}{1-t_{i+1}} \le h$.
2. For all $i \in [N-2]$, we have $\frac{\eta_i^2(1-t_i)^2}{(1-t_{i+1})^2} \le h^2(1-t_i)^2 \le h^2$, as $\frac{\eta_i}{1-t_{i+1}} \le h$ and $(1-t_i) \le 1$.
3. $\sum_{i=1}^{N-2}\frac{\eta_i^2}{(1-t_{i+1})^2 t_i^2} = \sum_{\substack{1\le i\le N-2\\ t_i<1/2}}\frac{\eta_i^2}{(1-t_{i+1})^2 t_i^2} + \sum_{\substack{1\le i\le N-2\\ t_i\ge 1/2}}\frac{\eta_i^2}{(1-t_{i+1})^2 t_i^2} \le 4h^2 N$.

This shows that Assumption 4.5 is satisfied. We take N sufficiently large so that $h \le 1/4$ and $16C_3kh\log N < 1$, as required in the general theorem. By (16), these conditions hold when N is sufficiently large relative to $k\log N\log(1/\delta)$.

Now, recall that $\delta = 1/(N \vee d) \le 1/N$, and that the geometric schedule satisfies

$$\log(1+h) = \frac{2\log(1/(2\delta))}{N-2}.$$

Using $h/2 \le \log(1+h) \le h$, we obtain $h \asymp \log(1/\delta)/N$ and $Nh \gtrsim 1$. Therefore,

$$s_N \lesssim Nh^2 \lesssim \frac{\log^2(1/\delta)}{N}, \qquad h + Nh^2 \lesssim Nh^2.$$

In particular, the initial boundary term in the weighted trace sum is absorbed into the leading term. Since $\log N \le \log(1/\delta)$, we have

$$(h + Nh^2)k \log N \lesssim \frac{k \log^3(1/\delta)}{N}.$$

The remainder terms in (41) satisfy

$$\frac{h + Nh^2}{\delta^2 N^{10}} + \frac{h^3}{\delta^6 N^9} \lesssim \frac{\log(1/\delta) + \log^2(1/\delta)}{\delta^2 N^{11}} + \frac{\log^3(1/\delta)}{\delta^6 N^{12}}.$$

Also, $\eta_0 = t_1 = \delta$, so the bound for the first step is already included in (41). Putting all of these facts together and substituting in (41), we get

$$\begin{aligned}
&\mathsf{TV}(p_{X_{N-1}}, p_{Y_{N-1}}) \\
&\le C_1'' \frac{k \log^3(1/\delta)}{N} + C_2'' \frac{\log^2(1/\delta)}{N} \varepsilon_{\mathrm{J},1}^2 + C_3'' \frac{\log^2(1/\delta)}{N} \varepsilon_{\mathrm{v}}^2 \\
&\quad + C_4'' \log(1/\delta)(\varepsilon_{\mathrm{v}} + \varepsilon_{\mathrm{J},1} + \varepsilon_{\mathrm{J},2} + \varepsilon_{\mathrm{H}}) \\
&\quad + C_5'' \left(\frac{\log(1/\delta) + \log^2(1/\delta)}{\delta^2 N^{11}} + \frac{\log^3(1/\delta)}{\delta^6 N^{12}} \right) \\
&\quad + A_1 \delta^2 \operatorname{Tr}(\mathrm{Cov}_1) + A_2 \delta^2 (\varepsilon_0^{\mathrm{v}})^2 + A_3 \delta^2 (\varepsilon_0^{\mathrm{J},1})^2 + A_4 (\varepsilon_0^{\mathrm{v}} + \varepsilon_0^{\mathrm{J},1} + \varepsilon_0^{\mathrm{J},2} + \varepsilon_0^{\mathrm{H}}),
\end{aligned}$$

where $\mathrm{Cov}_1 := \mathrm{Cov}_{X_1 \sim p_1}(X_1)$. Now, note that one can estimate $v_0(x) = \mu_{p_{X_1}} - x$ by $\widehat{v}_0(x) := \widehat{\mu}_1 - x$, where $\widehat{\mu}_1 = \sum_{i=1}^n x_{i,1}/n$ and n is the (training) sample size. In this case, we have

- $\varepsilon_0^{\mathrm{J},1} = \varepsilon_0^{\mathrm{J},2} = \varepsilon_0^{\mathrm{H}} = 0$.
- $(\varepsilon_0^{\mathrm{v}})^2 = \mathbb{E}\|\widehat{\mu}_1 - \mu_{p_{X_1}}\|_2^2 = O\left(\frac{\mathsf{Trace}(\mathrm{Cov}_1)}{n}\right)$.

This finishes the proof of Theorem 4.7.

C.3 Proof of convergence of STOC-RF (Theorem 4.9)

The proof of this result essentially follows by reduction and leverages key results of [38]. We start by recalling some key facts: $t(\tau) = \frac{\sqrt{\omega_\tau}}{\sqrt{1-\omega_\tau}+\sqrt{\omega_\tau}}$, and $\frac{Y'_\tau}{\sqrt{\omega_\tau}} \overset{d}{=} \frac{\tilde{Z}_{t(\tau)}}{t(\tau)}$. With this time transformation, we have

$$\alpha_\tau = \frac{\omega_\tau}{\omega_{\tau-1}} = \frac{(t_i/\sigma_{t_i})^2}{(t_{i+1}/\sigma_{t_{i+1}})^2} = \frac{R_i^2}{R_{i+1}^2}, \quad \text{where } t_i = t(\tau). \tag{42}$$

We also have $s_{Y'_\tau}(y) = \sigma_t s_{\tilde{Z}_t}(\tilde{z})$, where $\tilde{z} = \sigma_t y$. Next, we recall the DDPM sampler considered in [38]:

$$\begin{aligned}
&\hat{Y}'_{\tau-1} = \frac{1}{\sqrt{\alpha_\tau}} \left\{ \hat{Y}'_\tau + \delta_\tau \widehat{s}_{Y'_\tau}(\hat{Y}'_\tau) + \nu_\tau \xi_\tau \right\}, \hat{Y}'_N \sim \mathcal{N}(0, I_d) \\
&\delta_\tau = 1 - \alpha_\tau, \nu_\tau = \sqrt{\frac{(\alpha_\tau - \omega_\tau)(1-\alpha_\tau)}{1-\omega_\tau}}, \\
&\tau = N, N-1, \ldots, 2, \quad \{\xi_\tau\}_{\tau \ge 1} \overset{i.i.d.}{\sim} N(0, I_d),
\end{aligned} \tag{43}$$

Now, if we substitute all the quantities of the above sampler by their corresponding RF versions, i.e., we get $\hat{Y}_\tau \to Y_t/\sigma_t, \widehat{s}_{Y'_\tau}(y) \to \sigma_t \widehat{s}_{t_i}(Y_{t_i})$, and $\alpha_\tau \to \frac{R_i^2}{R_{i+1}^2}$, then it leads to the following sampler:

$$\begin{aligned}
&\frac{Y_{t_{i+1}}}{\sigma_{t_{i+1}}} = \frac{R_{i+1}}{R_i} \left\{ \frac{Y_{t_i}}{\sigma_{t_i}} + \eta_i \sigma_{t_i} \widehat{s}_{t_i}(Y_{t_i}) + \sqrt{\psi_i} W_{t_i} \right\}, \\
&\frac{Y_{t_0}}{\sigma_{t_0}} \sim N(0, I_d), \quad \{W_{t_i}\}_{i \ge 0} \overset{i.i.d.}{\sim} N(0, I_d), \\
&i = 0, 1, \ldots, N-2,
\end{aligned} \tag{44}$$

where $\eta_i = 1 - \frac{R_i^2}{R_{i+1}^2}, \psi_i = \frac{R_i^2}{R_{i+1}^2} \cdot \frac{1-R_{i+1}^2}{1-R_i^2} \cdot \left(1 - \frac{R_i^2}{R_{i+1}^2}\right)$, and $\widehat{s}_{t_i}(x) = \frac{t_i \widehat{v}_{t_i}(x) - x}{1-t_i}$. Here η_i has the same meaning as in (20). The above is exactly the sampler described in (20). In other words, we have equivalence between two above samplers. Therefore, we can use the convergence result of (43).

Defining few auxiliary processes: All updates below are for $\tau = N, \ldots, 2$, so that the final index is $\tau = 1$, corresponding to $i = N - 1$. We define the auxiliary processes in exact way defined in [38, Section C.1].

- First, we define $\{Y_\tau^\star\}_{\tau=N}^1$ by

$$Y_N^\star = Y_N' \sim \mathcal{N}(0, I_d), \quad Y_{\tau-1}^\star = \frac{1}{\sqrt{\alpha_\tau}} \left(Y_\tau^\star + \delta_\tau s_{Y_\tau'}(Y_\tau^\star) + \nu_\tau \xi_\tau\right).$$

  Note that, the above update rule is just the analog of $\{\hat{Y}_\tau\}_\tau$ process with the true score.
- We construct an auxiliary process $\bar{Y}_\tau$ that follows the same transition dynamics as $Y_\tau^\star$:

$$\bar{Y}_{\tau-1}^- \left| \bar{Y}_\tau \sim Y_{\tau-1}^\star \right| Y_\tau^\star, \quad \bar{Y}_{\tau-1} \left| \left\{ \bar{Y}_{\tau-1}^- = y_{\tau-1}^- \right\} \right. = \begin{cases} y_{\tau-1}^-, & \text{with prob. } \frac{p_{Y_{\tau-1}'}(y_{\tau-1}^-)}{p_{\bar{Y}_{\tau-1}^-}(y_{\tau-1}^-)} \wedge 1 \\ \infty, & \text{otherwise} \end{cases} \tag{45}$$

for any $y_{\tau-1}^- \neq \infty$, where we recall that $a \wedge b := \min\{a, b\}$. It is straightforward to show that

$$p_{\bar{Y}_\tau}(y_\tau) = \int_{\mathbb{R}^d} \left(p_{Y_\tau'}(y_\tau^-) \wedge p_{\bar{Y}_\tau^-}(y_\tau^-)\right) \delta(y_\tau - y_\tau^-) \, \mathrm{d}y_\tau^- = p_{Y_\tau'}(y_\tau) \wedge p_{\bar{Y}_\tau^-}(y_\tau) \tag{46}$$

for any $y_\tau \neq \infty$, where $\delta(\cdot)$ denotes the Dirac measure.

- To account for the score estimation error, we introduce another auxiliary process $\widetilde{Y}_\tau$ based on the dynamics of Y_τ :

$$\widetilde{Y}_{\tau-1}^- \left| \widetilde{Y}_\tau \sim \hat{Y}_{\tau-1}' \right| \hat{Y}_\tau', \quad \widetilde{Y}_{\tau-1} \left| \left\{ \widetilde{Y}_{\tau-1}^- = y_{\tau-1}^- \right\} \right. = \begin{cases} y_{\tau-1}^-, & \text{with prob. } \frac{p_{Y_{\tau-1}'}(y_{\tau-1}^-)}{p_{\bar{Y}_{\tau-1}^-}(y_{\tau-1}^-)} \wedge 1 \\ \infty, & \text{otherwise.} \end{cases} \tag{47}$$

Next, we observe that for $t_i = t(\tau)$

$$(\varepsilon_\tau^s)^2 := \mathbb{E}\|\widehat{s}_{Y_\tau'}(Y_\tau') - s_{Y_\tau'}(Y_\tau')\|_2^2 = \sigma_{t_i}^2 \mathbb{E}\|\widehat{s}_{t_i}(X_{t_i}) - s_{t_i}(X_{t_i})\|_2^2 \leq (\varepsilon_i^{\mathrm{s}})^2. \tag{48}$$

To see the weight of the score error, the Gaussian transitions with the true and approximate scores have the same covariance $\nu_\tau^2 I_d / \alpha_\tau$. Their conditional KL divergence at y is therefore

$$\frac{(1-\alpha_\tau)^2}{2\nu_\tau^2} \|s_{Y_\tau'}(y) - \widehat{s}_{Y_\tau'}(y)\|_2^2 = \frac{(1-\alpha_\tau)(1-\omega_\tau)}{2(\alpha_\tau - \omega_\tau)} \|s_{Y_\tau'}(y) - \widehat{s}_{Y_\tau'}(y)\|_2^2.$$

For sufficiently large N, the schedule in (18) gives $\alpha_\tau \geq 1/2$ and $\beta_\tau \leq 2\beta_{\tau-1}$ for $\tau \geq 2$. Since $1 - \omega_{\tau-1} \geq \beta_{\tau-1}$, we have

$$\frac{1-\omega_\tau}{\alpha_\tau - \omega_\tau} = 1 + \frac{\beta_\tau}{\alpha_\tau(1-\omega_{\tau-1})} \leq 5.$$

Thus, the coefficient is bounded by a constant times $1 - \alpha_\tau \lesssim \log N / N$. The common acceptance step cannot increase KL divergence, and (46) bounds the finite part of $p_{\bar{Y}_\tau}$ by $p_{Y_\tau'}$. Hence, the expected score error is bounded by $(\varepsilon_\tau^s)^2$. Finally, $i = N - \tau$ maps $\tau = 2, \ldots, N$ to $i = 0, \ldots, N-2$. Now, we can exactly replicate the analysis in [38, Section C.2] till Equation (106) in Step 6, which yields the following for a universal constant $L > 0$:

$$\mathsf{KL}\left(p_{\bar{Y}_1} \| p_{\widetilde{Y}_1}\right) \leq \sum_{\tau=2}^{N} L(1-\alpha_\tau)(\varepsilon_\tau^s)^2 \lesssim \frac{\log N}{N} \sum_{i=0}^{N-2} (\varepsilon_i^{\mathrm{s}})^2 \lesssim \varepsilon_{\mathrm{s}}^2 \log N.$$

After this, Step 7 of the proof follows verbatim to yield the final bound in Theorem 4.9.

C.4 Description of the new RF sampler in Remark 1

The DDIM sampler considered in [38] takes the form

$$\begin{aligned}
&\hat{Y}'_{\tau-1} = \frac{1}{\sqrt{\alpha_\tau}}\left\{\hat{Y}'_\tau + \delta_\tau \widehat{s}_{Y'_\tau}(\hat{Y}'_\tau)\right\},\ \hat{Y}'_N \sim \mathcal{N}(0, I_d)\\
&\delta_\tau = \frac{1-\alpha_\tau}{1+\sqrt{\frac{\alpha_\tau-\omega_\tau}{1-\omega_\tau}}},\\
&\tau = N, N-1, \ldots, 1.
\end{aligned} \tag{49}$$

Using the same transformations detailed in Section C.3 and plugging them in (49), we arrive at the RF sampler

$$\begin{aligned}
&\frac{Y_{t_{i+1}}}{\sigma_{t_{i+1}}} = \frac{R_{i+1}}{R_i}\left\{\frac{Y_{t_i}}{\sigma_{t_i}} + \eta_i \sigma_{t_i} \widehat{s}_{t_i}(Y_{t_i})\right\},\\
&\frac{Y_{t_0}}{\sigma_{t_0}} \sim N(0, I_d),\\
&i = 0, 1, \ldots, N-1,
\end{aligned} \tag{50}$$

where $\eta_i = \frac{1-\frac{R_i^2}{R_{i+1}^2}}{1+\sqrt{\frac{R_i^2}{R_{i+1}^2}\cdot\frac{1-R_{i+1}^2}{1-R_i^2}}}$, $R_i = t_i/\sigma_{t_i}$.

Algebraic simplification. Using $R_i = t_i/\sigma_{t_i}$, we have

$$\frac{R_{i+1}}{R_i} = \frac{t_{i+1}}{t_i}\frac{\sigma_{t_i}}{\sigma_{t_{i+1}}}, \qquad \frac{R_i^2}{R_{i+1}^2}\cdot\frac{1-R_{i+1}^2}{1-R_i^2} = \frac{t_i^2(1-t_{i+1})^2}{t_{i+1}^2(1-t_i)^2}, \qquad \eta_i = \frac{1-\frac{t_i^2\sigma_{t_{i+1}}^2}{t_{i+1}^2\sigma_{t_i}^2}}{1+\frac{t_i(1-t_{i+1})}{t_{i+1}(1-t_i)}}.$$

Multiplying (50) by $\sigma_{t_{i+1}}$ yields

$$Y_{t_{i+1}} = \frac{t_{i+1}}{t_i}\Big(Y_{t_i} + \eta_i\, \sigma_{t_i}^2\, \widehat{s}_{t_i}(Y_{t_i})\Big).$$

A direct calculation using $\sigma_t^2 = t^2 + (1-t)^2$ shows that

$$\eta_i \sigma_{t_i}^2 = \frac{(t_{i+1}-t_i)\big(1-t_i\big)}{t_{i+1}}.$$

Therefore, the deterministic sampler can be written equivalently as

$$Y_{t_{i+1}} = \frac{t_{i+1}}{t_i}\, Y_{t_i} + \frac{(t_{i+1}-t_i)(1-t_i)}{t_i}\widehat{s}_{t_i}(Y_{t_i}) = Y_{t_i} + \Delta_i \underbrace{\left\{\frac{Y_{t_i}}{t_i} + \frac{1-t_i}{t_i}\widehat{s}_{t_i}(Y_{t_i})\right\}}_{=\widehat{v}_{t_i}(Y_{t_i})}, \tag{51}$$

where $\Delta_i = t_{i+1} - t_i$. This expression matches exactly the form of the RF sampler in (5). However, the time steps Δ_i are different from those we propose in (15), which we used to obtain the RF convergence guarantees given in Theorem 4.7. Although the time-schedule is still U-shaped as shown in Figure 7 (b), however, it is not symmetric and, as a result, might take more discretization steps at either the beginning or the end of the sampling stage. Whether a symmetric or asymmetric schedule is better remains an open problem. Furthermore, we observe that empirically it appears to be relatively sensitive to the choice of the parameters c_0 and c_1 in (18).

D Proof of Lemmas related to Stochastic Localization (Section 3)

D.1 Proof of Lemma 3.1 and Lemma 3.2

We first show that the solution to the SDE (10) takes the form

$$\frac{Y'_\tau}{\sqrt{\omega_\tau}} = X_1 + \widetilde{B}'_{\frac{1-\omega_\tau}{\omega_\tau}}, \tag{52}$$

where $\omega_\tau = \exp(-2\int_0^\tau \beta(u)du)$ and $\left\{\widetilde{B}'_\tau\right\}_{\tau\geq 0}$ is a standard Brownian motion. First, it is easy to see that the solution of SDE (10) is

$$\frac{Y'_\tau}{\sqrt{\omega_\tau}} = X_1 + \underbrace{\int_0^\tau \sqrt{2\beta(t)}\exp\left(\int_0^t \beta(u)du\right)dB'_t}_{:=M_\tau}. \tag{53}$$

We note that M_τ is continuous local Martingale with $M_0 = 0$. Also, the quadratic variation of M_τ is

$$\langle M\rangle_\tau = \int_0^\tau 2\beta(t)\exp\left(2\int_0^t \beta(u)du\right)dt = \exp\left(2\int_0^\tau \beta(u)du\right) - 1 = \frac{1-\omega_\tau}{\omega_\tau}.$$

Then, by Dambis–Dubins–Schwarz theorem [57, Theorem 1.6] there exists a Brownian motion $\left\{\widetilde{B}'_u\right\}_{u\geq 0}$ such that $\{M_\tau\}_{\tau\geq 0} = \left\{\widetilde{B}'_{\frac{1-\omega_\tau}{\omega_\tau}}\right\}_{\tau\geq 0}$. Therefore, we have (52) by substituting M_τ in (53).

Now, for $s > 0$, we define two processes: $\hat{U}_s := sY'_{\tau(s)}/\sqrt{\omega_{\tau(s)}} = sX_1 + s\widetilde{B}'_{\frac{1-\omega_{\tau(s)}}{\omega_{\tau(s)}}}$, and $\tilde{U}_s := s\tilde{X}_{t(s)}/t(s) = sX_1 + s\tilde{B}_{\frac{(1-t(s))^2}{t^2(s)}}$. Now, under the prescribed time change in Lemma 3.1 and Lemma 3.2, we have

$$s\widetilde{B}'_{\frac{1-\omega_{\tau(s)}}{\omega_{\tau(s)}}} = s\widetilde{B}'_{1/s}, \quad \text{and} \quad s\tilde{B}_{\frac{(1-t(s))^2}{t^2(s)}} = s\tilde{B}_{1/s}.$$

By Brownian time inversion, both $\left\{s\widetilde{B}'_{1/s}\right\}_{s>0}$ and $\left\{s\tilde{B}_{1/s}\right\}_{s>0}$ extend continuously to Brownian motions by setting their values at $s = 0$ to zero. Hence $\hat{U}_s$ and $\tilde{U}_s$, with the same zero extension, are equivalent in law to the SL process. The identities involving U_s/s are asserted only for $s > 0$.

D.2 Proof of Lemma 3.3: Equivalence of SDE (6) and SDE (11)

Behavior at the initial endpoint. The process (12) is defined for $t > 0$. As $t \downarrow 0$, its marginals converge weakly to $p_0 = \mathcal{N}(0, I_d)$, but the process has no continuous extension to $t = 0$. Indeed, writing $G_t = tW_{(1-t)^2/t^2}$, each coordinate satisfies $\mathrm{Var}(G_t - G_{2t}) = (1-t)^2 \to 1$ for $0 < t < 1/2$. Thus G_t is not Cauchy in probability at zero; neither is $\tilde{X}_t$, since $tX_1 \to 0$ almost surely. Also, for $\gamma_t = (1-t)/t$, we have $\int_0^\epsilon 2\gamma_t\, dt = \infty$, so the Brownian integral in (6) cannot start at zero in the usual Itô formulation. We therefore formulate the SDE equivalence on positive-time intervals with matching initial laws. A Gaussian initialization at positive time approximates p_{t_0} and is distinct from this exact initialization.

SDE (6) to SDE (11):

We work on $t \in [t(N), 1)$ and $0 \leq \tau < N$, with the matching initial laws in Proposition 3.3. The differential calculations hold on compact subintervals of these domains. We begin with the solution $\tilde{Z}_t$ of STOC-RF SDE (6):

$$d\tilde{Z}_t = \left\{\frac{\tilde{Z}_t}{t} + 2\left(\frac{1-t}{t}\right)s_t(\tilde{Z}_t)\right\}dt + \sqrt{2\left(\frac{1-t}{t}\right)}dB_t, \tag{54}$$

where $s_t(\cdot)$ is the score of $\tilde{Z}_t$, and $\{B_t\}_{t\geq 0}$ is Brownian motion.

Recall the transformation $t(\tau) = \frac{\sqrt{\omega_\tau}}{\sqrt{\omega_\tau}+\sqrt{1-\omega_\tau}}$, where $\omega(\tau) = \exp\left(-2\int_0^\tau \beta(u)du\right)$.

Now, we define $\overleftarrow{Y}_\tau := \frac{\sqrt{\omega_{N-\tau}}\tilde{Z}_{t(N-\tau)}}{t(N-\tau)}$. We first begin by noting that

$$\frac{\sqrt{\omega_{N-\tau}}}{t(N-\tau)} = \frac{1}{\sigma_{t(N-\tau)}} \tag{55}$$

We also have $\log \omega_{N-\tau} = -2\int_0^{N-\tau} \beta(u)du$. Therefore, we have

$$\begin{aligned}
&\frac{1}{\omega_{N-\tau}}.\frac{d\omega_{N-\tau}}{dt} = 2\beta(N-\tau)\frac{d\tau}{dt} \\
&\Rightarrow \frac{d\tau}{dt} \overset{(i)}{=} \frac{1}{2\beta(N-\tau)\omega(N-\tau)}.\frac{d\omega_{N-\tau}}{dt} = \frac{1}{2\beta(N-\tau)\omega(N-\tau)}.\frac{d(t^2/\sigma_t^2)}{dt} \\
&\Rightarrow \frac{d\tau}{dt} = \frac{1}{\beta(N-\tau)}.\frac{1-t}{t\sigma_t^2}
\end{aligned} \tag{56}$$

Step (i) follows from (55). Next, we define the backward time $t_\tau := t(N-\tau)$, and then we can write $\tilde{Z}_{t_\tau} = \sigma_{t_\tau}\overleftarrow{Y}_\tau$. This also yields that the score of $\overleftarrow{Y}_\tau$ satisfies $s_\tau^{\leftarrow}(\overleftarrow{Y}_\tau) = \sigma_{t_\tau} s_t(\tilde{Z}_{t_\tau})$. Moreover, (13) yields that $\sigma_{t_\tau} s_t(\tilde{Z}_{t_\tau}) = \nabla \log q_{N-\tau}(\overleftarrow{Y}_\tau)$. Next, with (56) and (54) in mind, we can write

$$\begin{aligned}
d\overleftarrow{Y}_\tau &= \frac{1}{\sigma_{t_\tau}} d\tilde{Z}_{t_\tau} - \frac{2t_\tau - 1}{\sigma_{t_\tau}^3}\tilde{Z}_{t_\tau}dt_\tau \\
&= \left\{\frac{\tilde{Z}_{t_\tau}}{t_\tau\sigma_{t_\tau}} + 2\left(\frac{1-t_\tau}{t_\tau\sigma_{t_\tau}}\right)s_{t_\tau}(\tilde{Z}_{t_\tau}) - \frac{2t_\tau-1}{\sigma_{t_\tau}^3}\tilde{Z}_{t_\tau}\right\}dt_\tau + \sqrt{2\left(\frac{1-t_\tau}{\sigma_{t_\tau}^2 t_\tau}\right)}dB_{t_\tau} \\
&= \left\{\frac{1-t_\tau}{t_\tau\sigma_{t_\tau}^3}\tilde{Z}_{t_\tau} + 2\left(\frac{1-t_\tau}{t_\tau\sigma_{t_\tau}}\right)s_{t_\tau}(\tilde{Z}_{t_\tau})\right\}dt_\tau + \sqrt{2\left(\frac{1-t_\tau}{\sigma_{t_\tau}^2 t_\tau}\right)}dB_{t_\tau} \\
&= \left\{\frac{1-t_\tau}{t_\tau\sigma_{t_\tau}^2}\overleftarrow{Y}_\tau + 2\left(\frac{1-t_\tau}{t_\tau\sigma_{t_\tau}^2}\right)s_\tau^{\leftarrow}(\overleftarrow{Y}_\tau)\right\} \times \beta(N-\tau) \times \frac{t_\tau\sigma_{t_\tau}^2}{1-t_\tau}d\tau + \sqrt{2\left(\frac{1-t_\tau}{\sigma_{t_\tau}^2 t_\tau}\right)} \times \sqrt{\beta(N-\tau)\times\frac{t_\tau\sigma_{t_\tau}^2}{1-t_\tau}}d\overleftarrow{B}_\tau \\
&= \beta(N-\tau)\left\{\overleftarrow{Y}_\tau + 2s_\tau^{\leftarrow}(\overleftarrow{Y}_\tau)\right\}d\tau + \sqrt{2\beta(N-\tau)}d\overleftarrow{B}_\tau \\
&= \beta(N-\tau)\left\{\overleftarrow{Y}_\tau + 2\nabla\log q_{N-\tau}(\overleftarrow{Y}_\tau)\right\}d\tau + \sqrt{2\beta(N-\tau)}d\overleftarrow{B}_\tau,
\end{aligned}$$

where $\left\{\overleftarrow{B}_\tau\right\}_{\tau\geq 0}$ is another Brownian motion.

Initial condition check: Recall that $\overleftarrow{Y}_0 = \frac{\sqrt{\omega_N}\tilde{Z}_{t(N)}}{t(N)} = (\sqrt{\omega_N} + \sqrt{1-\omega_N})\tilde{Z}_{t(N)}$. Therefore, we have

$$\text{Law}(\overleftarrow{Y}_0) = \text{Law}\left[(\sqrt{\omega_N} + \sqrt{1-\omega_N}).\{t(N)X_1 + (1-t(N))X_0\}\right] = \text{Law}(\sqrt{\omega_N}X_1 + \sqrt{1-\omega_N}X_0) = \text{Law}(Y_N').$$

This shows that $\overleftarrow{Y}_\tau$ satisfies the backward DDPM SDE (11).

SDE (11) **to SDE** (6)**:**

For the other direction, we let $\left\{\overleftarrow{Y}_\tau\right\}_{0\leq\tau<N}$ be a solution of SDE (11), with $\overleftarrow{Y}_0 \sim q_N$. Now, to generate a solution of SDE (6), we need to define

$$\tilde{Z}_t = \frac{t(N-\tau(t))\overleftarrow{Y}_{\tau(t)}}{\sqrt{\omega_{N-\tau(t)}}} = \frac{\overleftarrow{Y}_{\tau(t)}}{\sqrt{\omega_{N-\tau(t)}} + \sqrt{1-\omega_{N-\tau(t)}}},$$

where $\tau(t)$ should satisfy the following:

$$\begin{aligned} t(N-\tau(t)) = \quad & \frac{\sqrt{\omega_{N-\tau(t)}}}{\sqrt{\omega_{N-\tau(t)}}+\sqrt{1-\omega_{N-\tau(t)}}} = t \\ & \Rightarrow \omega_{N-\tau(t)} = \frac{t^2}{t^2+(1-t)^2} \\ & \Rightarrow \exp\left(-2\int_0^{N-\tau(t)} \beta(u)du\right) = \frac{t^2}{t^2+(1-t)^2} \\ & \Rightarrow \int_0^{N-\tau(t)} \beta(u)du = \log\left(\sqrt{1+\frac{(1-t)^2}{t^2}}\right). \end{aligned} \tag{57}$$

This shows that $\tilde{Z}_t = \sigma_t \overleftarrow{Y}_{\tau(t)}$. Additionally, (13) yields that the score of $\tilde{Z}_t$ is $s_t(\cdot)$. Now, for convenience, we now write τ to denote $\tau(t)$. Taking derivative w.r.t. t on both sides of (57) yields

$$\frac{d\tau}{dt} = \frac{1}{\beta(N-\tau)}.\frac{(1-t)}{t\sigma_t^2}.$$

Using these facts, we finally get

$$\begin{aligned} d\tilde{Z}_t &= \frac{(2t-1)}{\sigma_t}\overleftarrow{Y}_\tau\, dt + \sigma_t d\overleftarrow{Y}_\tau \\ &= \frac{(2t-1)}{\sigma_t^2}\tilde{Z}_t\, dt + \sigma_t\left\{\overleftarrow{Y}_\tau + 2s_\tau^{\leftarrow}(\overleftarrow{Y}_\tau)\right\}\beta(N-\tau)d\tau + \sigma_t\sqrt{2\beta(N-\tau)}d\overleftarrow{B}_\tau \\ &= \frac{(2t-1)}{\sigma_t^2}\tilde{Z}_t\, dt + \sigma_t\left\{\frac{\tilde{Z}_t}{\sigma_t} + 2\sigma_t s_t(\tilde{Z}_t)\right\}\frac{(1-t)}{t\sigma_t^2}dt + \sqrt{2\left(\frac{1-t}{t}\right)}d\tilde{B}_t \\ &= \left(\frac{2t-1}{\sigma_t^2}+\frac{1-t}{t\sigma_t^2}\right)\tilde{Z}_t\, dt + 2\left(\frac{1-t}{t}\right)s_t(\tilde{Z}_t)\, dt + \sqrt{2\left(\frac{1-t}{t}\right)}d\tilde{B}_t \\ &= \frac{\tilde{Z}_t}{t}\, dt + 2\left(\frac{1-t}{t}\right)s_t(\tilde{Z}_t)\, dt + \sqrt{2\left(\frac{1-t}{t}\right)}d\tilde{B}_t. \end{aligned}$$

Therefore, we $\tilde{Z}_t$ satisfies SDE (6).

D.3 Proof of Proposition A.3

Recall that $\frac{\tilde{I}_{\theta(s)}}{a_{\theta(s)}} = X_1 + W_{r^2_{\theta(s)}}$, where $r^2_{\theta(s)} = 1/s$. Now, we define $\hat{U}_s := \frac{\tilde{I}_{\theta(s)}}{a_{\theta(s)}}$. Then, we have $s\hat{U}_s = sX_1 + sW_{1/s}$. Similar to the proofs of Lemma 3.1 and Lemma 3.2, the process $\hat{W}_s = sW_{1/s}$ for $s > 0$, extended by $\hat{W}_0 = 0$, is also a Brownian motion. This finishes the proof by the description of the SL process in (9).

E Auxiliary results and discussions

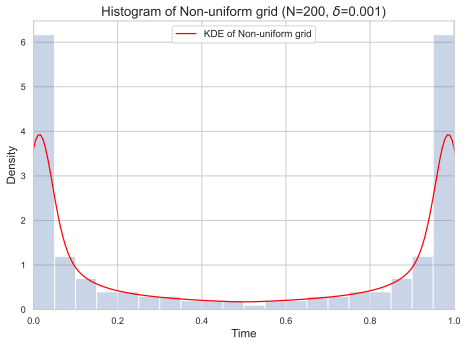
E.1 Time discretization and admissible schedules

The leading discretization term in Theorem 4.6 is $(h + Nh^2)k\log N$, where h controls the endpoint step ratios in Assumption 4.5. To compare different time schedules, write

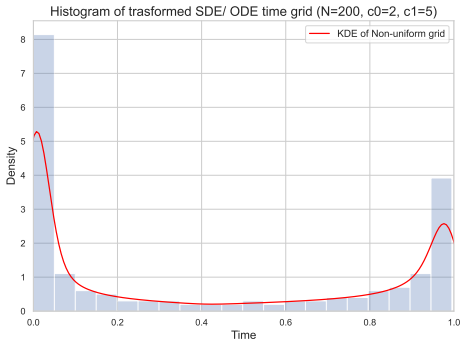
$$R_N := \max_{1\le i\le N-2}\left\{\frac{\eta_i}{t_i}\mathbb{1}\{t_i \le 1/2\} + \frac{\eta_i}{1-t_{i+1}}\mathbb{1}\{t_i > 1/2\}\right\}, \qquad \eta_i = t_{i+1} - t_i.$$

We may take $h = R_N$, provided $R_N \le 1/4$ and $16C_3kR_N\log N < 1$. Therefore, a sufficient condition for the coefficient of k in the leading term to vanish is

$$(R_N + NR_N^2)\log N = o(1).$$



(a) Histogram and kernel density estimation (KDE) plot of time-grid (15) showing U-shaped distribution.



(b) Histogram and KDE plot of the transformed time-grid (18) showing U-shaped distribution.

Figure 7: Histogram of non-uniform time grids.

Table 5: Examples of time discretization schedules and their leading discretization coefficients. The small-step condition $16C_3kh\log N < 1$ and the remainder bounds must be checked separately. For the EDM-type row, $\rho > \alpha + \beta$ is assumed.

Schedule	Construction	Endpoint bound	$(h + Nh^2)\log N$	Tends to zero?
Linear	$t_i = i/N$	$R_N = 1$	$(1+N)\log N$ at $h = R_N$	No; small-step condition fails
Two-sided geometric	$\eta_i \lesssim h_N t_i$ for $t_i \le 1/2$; $\eta_i \lesssim h_N(1-t_{i+1})$ for $t_i > 1/2$	$R_N = O(h_N)$	$O((h_N + Nh_N^2)\log N)$	If $h_N = o((N\log N)^{-1/2})$
EDM-type	$r_i^{1/\rho}$ equally spaced; $r_i = t_i/(1-t_i)$	$R_N \asymp N^{-1+(\alpha+\beta)/\rho}$	$\Theta(N^{-1+2(\alpha+\beta)/\rho}\log N)$	If $\rho > 2(\alpha+\beta)$
Logit-uniform	$\mathrm{logit}(t_i)$ equally spaced	$R_N = O(\log N/N)$	$O(\log^3 N/N)$	Yes, for fixed $\alpha, \beta > 0$
Logit-warped	$\mathrm{logit}(t_i) = a_N + (b_N - a_N)$ $\times g((i-1)/(N-2))$	$R_N = O(\log N/N)$	$O(\log^3 N/N)$	Yes, if $\|g'\|_\infty$ is bounded

In particular, $R_N = o((N\log N)^{-1/2})$ is sufficient. If k also grows with N, we additionally require $k(R_N + NR_N^2)\log N = o(1)$ for the leading term itself to vanish. The approximation-error, first-step, and remainder terms in the theorem must also be controlled.

The definition of R_N explains why the time steps should become finer near both endpoints. Near $t = 0$, we require η_i to be small relative to t_i, whereas near $t = 1$, we require it to be small relative to $1 - t_{i+1}$. This naturally leads to "U-shaped" schedules, with more discretization points near the endpoints and larger steps in the interior.

A convenient way to construct such schedules is to introduce the odds coordinate $r(t) := t/(1-t)$ and the normalized increment

$$a_i := \frac{\eta_i}{t_i(1-t_{i+1})} = \frac{r(t_{i+1})}{r(t_i)} - 1.$$

Since both endpoint step ratios are bounded by a_i, we have $R_N \le \max_i a_i$. Thus, an upper bound $\max_i a_i \le \gamma_N$ allows us to take $h = \gamma_N$. This gives a simple sufficient condition for admissibility, together with the small-step conditions of the theorem. The auxiliary sum s_N used in the proof is not needed for this comparison.

Table 5 summarizes several representative schedules. For the EDM-type and logit schedules, we take

$$t_1 = N^{-\alpha}, \qquad t_{N-1} = 1 - \delta, \qquad \delta = N^{-\beta},$$

where $\alpha, \beta > 0$ are fixed. These choices make the dependence on both endpoints explicit. The choice $\delta = 1/(N \vee d)$ has this form when $N \vee d = N^\beta$. The last column concerns only the leading coefficient of k; the remaining conditions are discussed below.

E.1.1 Verification of admissibility

We briefly verify the endpoint step bounds and the leading coefficients for the schedules in Table 5.

Linear schedule. For $t_i = i/N$, the two relevant ratios are $1/i$ and $1/(N-i-1)$, respectively. Hence $R_N = 1$. In particular, no $h \le 1/4$ can satisfy the endpoint step bound, so this schedule does not meet the hypotheses of the general theorem.

Two-sided geometric schedule. Suppose $\eta_i \lesssim h_N t_i$ for $t_i \le 1/2$ and $\eta_i \lesssim h_N(1-t_{i+1})$ for $t_i > 1/2$. Then $R_N = O(h_N)$ directly from its definition. Taking h to be a sufficiently large constant multiple of h_N, the leading coefficient is $O((h_N + Nh_N^2)\log N)$. Therefore, $h_N = o((N\log N)^{-1/2})$ is sufficient for this coefficient to vanish, and the small-step condition holds when $kh_N \log N$ is sufficiently small.

For the particular schedule (15), we have $t_1 = \delta$, $h \asymp \log(1/\delta)/N$, and $h \lesssim Nh^2$. Thus, for $\delta = N^{-\beta}$ with fixed $\beta > 0$, the leading coefficient is $O(\log^3 N/N)$. The sufficient condition $k\log^2 N/N = o(1)$ ensures $16C_3 kh\log N < 1$ for sufficiently large N.

EDM-type schedule. Consider the power-law interpolation of the odds coordinate in the table. Let $r_i = z_i^\rho$, where

$$z_i = z_{\min} + (i-1)\Delta, \qquad z_{\min} = r(t_1)^{1/\rho}, \qquad z_{\max} = r(t_{N-1})^{1/\rho}, \qquad \Delta = \frac{z_{\max} - z_{\min}}{N-2}.$$

The prescribed endpoints give

$$z_{\min} \asymp N^{-\alpha/\rho}, \qquad z_{\max} \asymp N^{\beta/\rho}, \qquad \frac{\Delta}{z_{\min}} \asymp N^{-1+(\alpha+\beta)/\rho}.$$

In particular, Δ is not of order N^{-1} when the terminal odds grow with N. For $\rho > \alpha+\beta$, uniformly in i,

$$a_i = \left(1 + \frac{\Delta}{z_i}\right)^\rho - 1 = (1+o(1))\frac{\rho\Delta}{z_i}.$$

Thus $R_N \le \max_i a_i = a_1$. Also, $\eta_1/t_1 = a_1(1-t_2) \sim a_1$, since $t_2 \to 0$. Consequently,

$$R_N \asymp N^{-1+(\alpha+\beta)/\rho}, \qquad (R_N + NR_N^2)\log N \asymp N^{-1+2(\alpha+\beta)/\rho}\log N.$$

Therefore, $\rho > 2(\alpha+\beta)$ ensures decay of the leading coefficient. A sufficient small-step condition is $kN^{-1+(\alpha+\beta)/\rho}\log N = o(1)$. These calculations concern the specific odds-coordinate construction displayed in the table.

Logit-uniform schedule. If $r(t_{i+1}) = q_N r(t_i)$, then

$$a_i = q_N - 1, \qquad \log q_N = \frac{\log r(t_{N-1}) - \log r(t_1)}{N-2} \sim \frac{(\alpha+\beta)\log N}{N}.$$

It follows that $R_N \le q_N - 1 = O(\log N/N)$. Taking $h = q_N - 1$, we obtain $(h + Nh^2)\log N = O(\log^3 N/N)$. Again, $k\log^2 N/N = o(1)$ suffices for the small-step condition.

Logit-warped schedule. Let $a_N = \mathrm{logit}(t_1)$, $b_N = \mathrm{logit}(t_{N-1})$, and

$$u_i := \mathrm{logit}(t_i) = a_N + (b_N - a_N)g\left(\frac{i-1}{N-2}\right),$$

where g is strictly increasing, $g(0) = 0$, $g(1) = 1$, and $\|g'\|_\infty$ is bounded independently of N. Since $b_N - a_N = O(\log N)$,

$$0 < u_{i+1} - u_i \le \frac{\|g'\|_\infty (b_N - a_N)}{N-2} = O\left(\frac{\log N}{N}\right).$$

Using $a_i = e^{u_{i+1}-u_i} - 1$, we obtain $R_N = O(\log N/N)$. Taking h of this order gives the same leading coefficient and sufficient small-step condition as the logit-uniform schedule.

Remainder and approximation-error terms. The schedule conditions above control the leading coefficient and the applicability of the covariance proposition. To obtain a vanishing total error, we must also control

$$\frac{h + Nh^2}{\delta^2 N^{10}} + \frac{h^3}{\delta^6 N^9},$$

as well as the approximation-error and first-step terms in Theorem 4.6. For the geometric and logit schedules with $h = O(\log N/N)$ and $\delta = N^{-\beta}$, the displayed remainder is bounded by

$$O\left(N^{2\beta-11}\log^2 N + N^{6\beta-12}\log^3 N\right).$$

Thus $0 < \beta < 2$ is sufficient for this remainder bound to vanish. For $\delta = 1/(N \vee d)$, the corresponding expressions are

$$O\left(\frac{(N \vee d)^2 \log^2(N \vee d)}{N^{11}} + \frac{(N \vee d)^6 \log^3(N \vee d)}{N^{12}}\right).$$

These terms need not vanish for arbitrary growth of d.

For the EDM-type construction with $\rho > 2(\alpha + \beta)$, substitution of $h \asymp N^{-1+(\alpha+\beta)/\rho}$ gives the remainder bound

$$O\left(N^{2\beta-11+2(\alpha+\beta)/\rho} + N^{6\beta-12+3(\alpha+\beta)/\rho}\right).$$

Both exponents must be negative for this bound to vanish. These additional requirements illustrate why decay of the leading coefficient alone is not sufficient for convergence of the full bound.

E.2 Controlling posterior covariance of RF

We recall the SL process (9) to study the dynamics of the conditional covariance matrix $\mathbf{\Sigma}_t := \mathrm{Cov}_{1|t}(X_t)$ where $X_t = tX_1 + (1-t)X_0$ and $X_0 \sim \mathcal{N}(0, I_d)$. Note that $\mathbf{\Sigma}_t$ is a random matrix. Recall the defintion of $\tilde{X}_t$ in (12). It is evident that $\tilde{\mathbf{\Sigma}}_t := \mathrm{Cov}_{1|t}(\tilde{X}_t) \overset{d}{=} \mathbf{\Sigma}_t$. Due to Lemma 3.2, the SL process (9) is equivalent to the process (12) under time change $t(s) := \frac{\sqrt{s}}{1+\sqrt{s}}$. Then, one can substitute this in Lemma A.1 to get the following result:

Lemma E.1 (Corollary A.2)**.** *for $t \in [0, 1)$, we have $\frac{d}{dt}\mathbb{E}(\mathbf{\Sigma}_t) = -\frac{2t}{(1-t)^3}\mathbb{E}(\mathbf{\Sigma}_t^2)$.*

Proof. Now, we note that $\mathbf{A}_s$ and $\tilde{\mathbf{\Sigma}}_{t(s)}$ (so is $\mathbf{\Sigma}_{t(s)}$) are marginally equal in law. Therefore, by chain rule we have the following:

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}(\mathbf{\Sigma}_t) &= \frac{d}{dt}\mathbb{E}(\tilde{\mathbf{\Sigma}}_t) \\
&= \frac{d}{dt}\mathbb{E}(\mathbf{A}_s) \\
&= \frac{2t}{(1-t)^3}\frac{d}{ds}\mathbb{E}(\mathbf{A}_s) \\
&= -\frac{2t}{(1-t)^3}\mathbb{E}(\mathbf{A}_s^2) \\
&= -\frac{2t}{(1-t)^3}\mathbb{E}(\mathbf{\Sigma}_t^2).
\end{aligned}$$

□

As mentioned in previous section, the above result allows to control the discretization error of the ODE and SDE sampler of RF. More concretely, Lemma E.1 allows us to bound the spectra of $\mathbf{\Sigma}_t^2$ in terms of the spectra of $\mathbf{\Sigma}_t$, thereby producing tighter bounds for convergence rate.

E.3 Proof of Proposition C.2

We recall the definitions of $(\mathbf{A}_s)_{s>0}$ and $(\tilde{\mathbf{\Sigma}}_{t(s)})_{s>0}$ in Section 3 and Section E.2. Note that $(\mathbf{A}_s)_{s>0}$ and $(\tilde{\mathbf{\Sigma}}_{t(s)})_{s>0}$ (see Section 3 and E.2) are also equivalent in law. This is simply due to the fact that $\tilde{\mathbf{\Sigma}}_{t(s)} = \mathrm{Cov}_{1|t(s)}(\tilde{X}_{t(s)}) = \mathrm{Cov}_{1|t(s)}(s\tilde{X}_{t(s)}/t(s))$, and the distributional equivalence of the two processes $(U_s)_{s>0}$ and $(s\tilde{X}_{t(s)}/t(s))_{s>0}$. It is known that the random matrix valued process $(\mathbf{A}_s)_{s>0}$ obeys the following SDE:

$$d\mathbf{A}_s = -\mathbf{A}_s^2 ds + \mathcal{M}_s^{(3)} dB_s, \tag{58}$$

where $\mathcal{M}_s^{(\ell)} = \mathbb{E}[(X_1 - \mathbb{E}[X_1 \mid U_s])^{\otimes \ell} \mid U_s]$. More details on this can be found in [17, Section 4.2.1] or [3, Section 3]. Since the two processes $(\mathbf{A}_s)_{s>0}$ and $(\tilde{\mathbf{\Sigma}}_{t(s)})_{s>0}$ are equivalent in law, SDE (58) immediately yields

$$d\tilde{\mathbf{\Sigma}}_t = -\frac{2t}{(1-t)^3}\tilde{\mathbf{\Sigma}}_t^2 dt + \tilde{\mathcal{M}}_t^{(3)} dB_{t^2/(1-t)^2}, \tag{59}$$

where $\tilde{\mathcal{M}}_t^{(\ell)} = \mathbb{E}[(X_1 - \mathbb{E}[X_1 \mid \tilde{X}_t])^{\otimes \ell} \mid \tilde{X}_t]$. Now, an application of Ito's formula on (58) yields

$$d(\mathsf{Trace}(\mathbf{A}_s^2)) = -2\left\langle \mathbf{A}_s, \mathbf{A}_s^2 \right\rangle ds + \left\langle \mathcal{M}_s^{(3)}, \mathcal{M}_s^{(3)} \right\rangle ds + 2\left\langle \mathbf{A}_s, \mathcal{M}_s^{(3)} \right\rangle dB_s$$

Taking expectation on both side we arrive at the following inequality:

$$d(\mathsf{Trace}[\mathbb{E}(\mathbf{A}_s^2)]) = -2\mathbb{E}\left[\left\langle \mathbf{A}_s, \mathbf{A}_s^2 \right\rangle\right] ds + \mathbb{E}\left[\left\langle \mathcal{M}_s^{(3)}, \mathcal{M}_s^{(3)} \right\rangle\right] ds.$$

Finally, using the fact that $\mathbf{A}_s$ and $\mathbf{\Sigma}_{t(s)}$ have the same law, chain rule yields the following:

$$d(\mathsf{Trace}[\mathbb{E}(\mathbf{\Sigma}_t^2)]) = -\frac{4t}{(1-t)^3}\mathbb{E}\left[\left\langle \mathbf{\Sigma}_t, \mathbf{\Sigma}_t^2 \right\rangle\right] dt + \frac{2t}{(1-t)^3}\mathbb{E}\left[\left\langle \tilde{\mathcal{M}}_t^{(3)}, \tilde{\mathcal{M}}_t^{(3)} \right\rangle\right] dt. \tag{60}$$

We first focus on the first term on the right-hand side of (60). Using the symmetry of $\mathbf{\Sigma}_t$, we observe that

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \mathbf{\Sigma}_t, \mathbf{\Sigma}_t^2 \right\rangle\right] = \mathbb{E}\left[\mathsf{Trace}(\mathbf{\Sigma}_t^3)\right] \\
&\le \mathbb{E}\left[\|\mathbf{\Sigma}_t\|_{\mathrm{op}} \|\mathbf{\Sigma}_t\|_F^2\right] \\
&\le \mathbb{E}\left[\mathsf{Trace}\left(\mathbf{\Sigma}_t\right) \|\mathbf{\Sigma}_t\|_F^2\right] \\
&= \mathbb{E}\left[\mathsf{Trace}\left(\mathbf{\Sigma}_t\right) \mathbb{1}\{X_t \in \mathcal{T}_t\} \|\mathbf{\Sigma}_t\|_F^2\right] + \mathbb{E}\left[\mathsf{Trace}\left(\mathbf{\Sigma}_t\right) \mathbb{1}\{X_t \notin \mathcal{T}_t\} \|\mathbf{\Sigma}_t\|_F^2\right] \\
&\le C_3 \frac{(1-t)^2}{t^2}(k \log N)\mathbb{E}\left[\|\mathbf{\Sigma}_t\|_F^2\right] + \mathbb{E}\left[\mathsf{Trace}\left(\mathbf{\Sigma}_t\right) \mathbb{1}\{X_t \notin \mathcal{T}_t\} \|\mathbf{\Sigma}_t\|_F^2\right],
\end{aligned} \tag{61}$$

where the last inequality follows due to (38). Regarding the last term in above display, Assumption 4.2 and (36) yields

$$\mathbb{E}\left[\mathsf{Trace}\left(\mathbf{\Sigma}_t\right) \mathbb{1}\{X_t \notin \mathcal{T}_t\} \|\mathbf{\Sigma}_t\|_F^2\right] \le R^6 \mathbb{P}(X_t \notin \mathcal{T}_t) = N^{6c_R}\mathbb{P}(X_t \notin \mathcal{T}_t) \le \frac{1}{N^{10}}.$$

Here we used $\mathsf{Trace}(\mathbf{\Sigma}_t) \le R^2$ and $\|\mathbf{\Sigma}_t\|_F^2 \le R^4$, and chose C_2 sufficiently large that $C_2 k \ge 6c_R + 10$. Therefore, we arrive at

$$\mathbb{E}\left[\left\langle \mathbf{\Sigma}_t, \mathbf{\Sigma}_t^2 \right\rangle\right] \le C_3 \frac{(1-t)^2}{t^2}(k \log N)\mathbb{E}\left[\|\mathbf{\Sigma}_t\|_F^2\right] + \frac{1}{N^{10}}. \tag{62}$$

Substituting this in (61) yields,

$$\begin{aligned}
d\mathbb{E}(\|\mathbf{\Sigma}_t\|_F^2) &= -\frac{4t}{(1-t)^3}\mathbb{E}\left[\left\langle \mathbf{\Sigma}_t, \mathbf{\Sigma}_t^2 \right\rangle\right] dt + \frac{2t}{(1-t)^3}\mathbb{E}\left[\left\langle \tilde{\mathcal{M}}_t^{(3)}, \tilde{\mathcal{M}}_t^{(3)} \right\rangle\right] dt \\
&\ge -\frac{4t}{(1-t)^3}\mathbb{E}\left[\left\langle \mathbf{\Sigma}_t, \mathbf{\Sigma}_t^2 \right\rangle\right] dt \\
&\ge -\frac{4C_3 k \log N}{(1-t)t}\mathbb{E}\left[\|\mathbf{\Sigma}_t\|_F^2\right] dt - \frac{4t}{(1-t)^3} \cdot \frac{1}{N^{10}} dt \\
&\ge -\frac{4C_3 k \log N}{(1-t_i)t_{i-1}}\mathbb{E}\left[\|\mathbf{\Sigma}_t\|_F^2\right] dt - \frac{4t_i}{(1-t_i)^3} \cdot \frac{1}{N^{10}} dt,
\end{aligned}$$

where the last line follows holds provided $t \in [t_{i-1}, t_i]$ with $i > 1$. Therefore, we can conclude that

$$\begin{aligned}
&\exp\left\{\frac{4C_3kt\log N}{(1-t_i)t_{i-1}}\right\}\mathbb{E}(\|\mathbf{\Sigma}_t\|_F^2) - \exp\left\{\frac{4C_3kt_{i-1}\log N}{(1-t_i)t_{i-1}}\right\}\mathbb{E}(\|\mathbf{\Sigma}_{t_{i-1}}\|_F^2)\\
&\geq -\frac{4t_i}{(1-t_i)^3}.\frac{1}{N^{10}}\int_{t_{i-1}}^{t}\exp\left\{\frac{4C_3ku\log N}{(1-t_i)t_{i-1}}\right\}du\\
&\geq -\frac{t_it_{i-1}}{C_3k(1-t_i)^2\log N}.\frac{1}{N^{10}}\left(\exp\left\{\frac{4C_3kt\log N}{(1-t_i)t_{i-1}}\right\} - \exp\left\{\frac{4C_3kt_{i-1}\log N}{(1-t_i)t_{i-1}}\right\}\right)
\end{aligned}$$

Dividing both sides by $\exp\left\{\frac{4C_3kt_{i-1}\log N}{(1-t_i)t_{i-1}}\right\}$ and rearranging, we obtain

$$\begin{aligned}
\mathbb{E}(\|\mathbf{\Sigma}_{t_{i-1}}\|_F^2) \leq \exp\left\{\frac{4C_3k(t-t_{i-1})\log N}{(1-t_i)t_{i-1}}\right\}\mathbb{E}(\|\mathbf{\Sigma}_t\|_F^2)\\
+ \frac{t_it_{i-1}}{C_3k(1-t_i)^2\log N}\frac{1}{N^{10}}\left(\exp\left\{\frac{4C_3k(t-t_{i-1})\log N}{(1-t_i)t_{i-1}}\right\} - 1\right).
\end{aligned}$$

Now, Assumption 4.5 and $h \leq 1/4$ give

$$\frac{\eta_i}{t_i(1-t_{i+1})} \leq \begin{cases} \dfrac{2h}{1-h} \leq 4h, & t_i \leq 1/2,\\ 2h \leq 4h, & t_i > 1/2. \end{cases}$$

Indeed, in the first case $\eta_i \leq ht_i$ and $1 - t_{i+1} \geq (1-h)/2$, including when the step crosses $1/2$. In the second case, $\eta_i \leq h(1-t_{i+1})$ and $t_i > 1/2$. Hence, for $t \in [t_{i-1}, t_i]$,

$$\frac{4C_3k(t-t_{i-1})\log N}{(1-t_i)t_{i-1}} \leq \frac{4C_3k\eta_{i-1}\log N}{(1-t_i)t_{i-1}} \leq 16C_3kh\log N < 1.$$

Using $e^x \leq 3$ and $e^x - 1 \leq 3x$ for $x \in [0,1]$, while retaining the factor t_it_{i-1} in the remainder, gives

$$\begin{aligned}
\mathbb{E}(\|\mathbf{\Sigma}_{t_{i-1}}\|_F^2) &\leq 3\mathbb{E}(\|\mathbf{\Sigma}_t\|_F^2) + \frac{12t_i(t-t_{i-1})}{(1-t_i)^3N^{10}}\\
&\leq 3\mathbb{E}(\|\mathbf{\Sigma}_t\|_F^2) + \frac{24h}{(1-t_i)^2N^{10}}.
\end{aligned}$$

Here we used $t_i \leq 1$, $t - t_{i-1} \leq \eta_{i-1}$, and $\eta_{i-1}/(1-t_i) \leq 2h$. The last bound follows from the same schedule assumption: for $t_{i-1} \leq 1/2$ it is at most $h/(1-h) \leq 2h$, and for $t_{i-1} > 1/2$ it is at most h. Replacing i by $i+1$, we obtain

$$\mathbb{E}(\|\mathbf{\Sigma}_{t_i}\|_F^2) \leq 3\mathbb{E}(\|\mathbf{\Sigma}_t\|_F^2) + \frac{24h}{(1-t_{i+1})^2N^{10}}, \quad t \in [t_i, t_{i+1}], \quad 1 \leq i \leq N-2.$$

Now we use Lemma E.1 to get

$$\begin{aligned}
\mathbb{E}[\mathsf{Trace}(\mathbf{\Sigma}_{t_{i+1}})] - \mathbb{E}[\mathsf{Trace}(\mathbf{\Sigma}_{t_i})] &= -\int_{t_i}^{t_{i+1}}\frac{2t}{(1-t)^3}\mathbb{E}(\|\mathbf{\Sigma}_t\|_F^2)dt\\
&\leq -\int_{t_i}^{t_{i+1}}\frac{2t}{3(1-t)^3}\mathbb{E}(\|\mathbf{\Sigma}_{t_i}\|_F^2)dt + \frac{8h}{(1-t_{i+1})^2N^{10}}\int_{t_i}^{t_{i+1}}\frac{2t}{(1-t)^3}dt\\
&= -\frac{\eta_i(t_i + t_{i+1} - 2t_it_{i+1})}{3(1-t_{i+1})^2(1-t_i)^2}\mathbb{E}(\|\mathbf{\Sigma}_{t_i}\|_F^2)\\
&\quad + \frac{8h\eta_i(t_i + t_{i+1} - 2t_it_{i+1})}{(1-t_{i+1})^4(1-t_i)^2N^{10}}.
\end{aligned}$$

Rearranging the above inequality, dividing by $t_i + t_{i+1} - 2t_it_{i+1} > 0$, and using $t_i + t_{i+1} - 2t_it_{i+1} \geq 2t_i(1-t_{i+1})$ together with the nonnegativity of $\mathbb{E}[\mathsf{Trace}(\mathbf{\Sigma}_{t_i})] - \mathbb{E}[\mathsf{Trace}(\mathbf{\Sigma}_{t_{i+1}})]$, we obtain

$$\begin{aligned}
\frac{\eta_it_i}{(1-t_{i+1})(1-t_i)^2}\mathbb{E}(\|\mathbf{\Sigma}_{t_i}\|_F^2) \leq \frac{3}{2}\left(\mathbb{E}[\mathsf{Trace}(\mathbf{\Sigma}_{t_i})] - \mathbb{E}[\mathsf{Trace}(\mathbf{\Sigma}_{t_{i+1}})]\right)\\
+ \frac{24h\eta_it_i}{(1-t_{i+1})^3(1-t_i)^2N^{10}}.
\end{aligned} \tag{63}$$

E.4 Properties of time scheduling (15)

Lemma E.2. *The time-steps* (15) *satisfies the following for* $i \in \{1, \dots, N-2\}$*:*

(a) We have $\eta_i = ht_i \mathbb{1}\{t_i < 1/2\} + h(1-t_{i+1})\mathbb{1}\{t_i \geq 1/2\}$.

(b) We have $\frac{\eta_i}{1-t_i} \leq h$ *and* $\frac{\eta_i}{1-t_{i+1}} \leq h$.

(c) We have $\frac{\eta_i^2(1-t_i)^2}{(1-t_{i+1})^2} \leq h^2$*, as* $\frac{\eta_i}{(1-t_{i+1})} \leq h$ *and* $(1-t_i) \leq 1$.

(d) We have $\sum_{i:t_i>0} \frac{\eta_i^2}{(1-t_{i+1})^2 t_i^2} = \sum_{i:0<t_i<1/2} \frac{\eta_i^2}{(1-t_{i+1})^2 t_i^2} + \sum_{i:t_i \geq 1/2} \frac{\eta_i^2}{(1-t_{i+1})^2 t_i^2} \leq 4h^2 N$.

Proof.

- **Part (a):** If $t_i < 1/2$, we have $\eta_i = t_{i+1} - t_i = (1+h)t_i - t_i = ht_i$. For the case $t_i \geq 1/2$, we have
$$\begin{aligned}\eta_i &= t_{i+1} - t_i \\ &= (1-t_i) - (1-t_{i+1}) \\ &= h(1-t_{i+1}).\end{aligned}$$
- **Part (b):** For $t_i < 1/2$, we have $\frac{\eta_i}{1-t_i} = ht_i/(1-t_i) \leq h$. We use the fact that $\frac{t}{1-t} < 1$ for $t \leq 1/2$. Similarly, $\frac{\eta_i}{1-t_{i+1}} = \frac{ht_i}{1-t_{i+1}}$. Now, note that $t_i < 1/2$ implies that $t_{i+1} \leq 1/2$. Therefore, $\frac{t_i}{1-t_{i+1}} \leq 1$, and it yields $\frac{\eta_i}{1-t_{i+1}} = \frac{ht_i}{1-t_{i+1}} < h$.

  For $t_i \geq 1/2$, we have $\frac{\eta_i}{1-t_i} = \frac{h(1-t_{i+1})}{1-t_i} < h$. Similarly, $\frac{\eta_i}{1-t_{i+1}} = \frac{h(1-t_{i+1})}{1-t_{i+1}} = h$. This yields the coveted bound.
- **Part (c):** The proof immediately follows from Part (b).
- **Part (d):** Fist, from Part (a), we can note that $\frac{\eta_i}{t_i} \leq h$. Also, Part (b) yields that $\frac{\eta_i}{1-t_{i+1}} \leq h$. Therefore,
$$\begin{aligned}&\sum_{i:t_i>0} \frac{\eta_i^2}{(1-t_{i+1})^2 t_i^2} \\ &= \sum_{i:0<t_i<1/2} \frac{\eta_i^2}{t_i^2} \cdot \frac{1}{(1-t_{i+1})^2} + \sum_{i:t_i \geq 1/2} \frac{\eta_i^2}{(1-t_{i+1})^2} \cdot \frac{1}{t_i^2} \\ &\leq \sum_{i:0<t_i<1/2} 4h^2 + \sum_{i:t_i \geq 1/2} 4h^2 \\ &\leq 4h^2 N.\end{aligned}$$

□

Lemma E.3. *Let* $t_{i+1} = t_i + \eta_i$ *with* $0 < t_i \leq t_{i+1} \leq 1$*. Then*
$$\frac{(1-t_i)^2}{t_i^2} - \frac{(1-t_{i+1})^2}{t_{i+1}^2} \leq \frac{2\eta_i}{t_i^3}.$$

Proof. Define $f(t) := \frac{(1-t)^2}{t^2}$ for $t > 0$. A direct calculation gives
$$f'(t) = \frac{2(t-1)}{t^3}.$$
By the mean value theorem, there exists $\xi \in (t_i, t_{i+1})$ such that
$$f(t_i) - f(t_{i+1}) = -f'(\xi)\,(t_{i+1} - t_i) = \frac{2(1-\xi)}{\xi^3}\,\eta_i.$$
Since $\xi \geq t_i$ and $\xi \leq 1$, we have $(1-\xi)/\xi^3 \leq 1/t_i^3$, which implies the desired bound. □

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: All the main claims made in abstract has been proven appendix and discussed in the paper.

   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: We point out some limitations of our theoretical results in the paper.

   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 the assumptions have been discussed and detailed proofs are provided in appendix.

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: [No]

Justification: The paper describes the sampling algorithms and experimental comparisons, with additional results in Appendix B. However, the current description does not fully specify the MNIST training setup or the T2I checkpoint and evaluation settings needed to reproduce all results.

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: [No]

   Justification: The paper provides a code link at `https://github.com/roysaptaumich/Low-dim-RF`. Complete instructions covering the environments, data preparation, and commands for reproducing all reported experiments are not documented in the supplemental material.

   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: [No]

   Justification: The paper and Appendix B describe the experimental comparisons and sampling schedules. Some training and evaluation details, including MNIST hyperparameters and the exact T2I checkpoint and settings, are not specified.

   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 MNIST results are averaged over ten independent runs and include reported variation, but the meaning and calculation of the error bars are not specified. Uncertainty estimates are not provided for all experimental comparisons.

   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: [No]

   Justification: Appendix B reports the hardware used for the synthetic and T2I experiments, and sampling times for selected T2I settings. MNIST training resources and a complete account of execution time and total compute are not provided.

   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: No conflict with Ethics code has been found.

   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 work has no negative societal impact.

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 poses no such risks.

    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 experiments use MNIST, a pretrained FLUX model, prompts from prior work, and existing evaluation metrics. The paper cites the FLUX model and prompt sources, but does not provide a complete account of asset versions, licenses, and terms of use.

    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: [No]

    Justification: The paper links experimental code, but complete documentation of its environment, usage, and license is not provided in the supplemental material.

    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: The paper does 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: The paper poses no such risks.

    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 method development in this research does not involve LLMs.

    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.