NOISE SCHEDULE DESIGN FOR DIFFUSION MODELS: AN OPTIMAL CONTROL PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

We develop a principled framework for analyzing and designing noise schedules in diffusion models. We show that one can recast this design problem as an optimal control problem, whose state is the Fisher information of the diffusion process, which evolves according to an ODE, and whose control input is the noise schedule. The objective of the optimal control problem is a functional involving the Fisher information, which is shown to be an upper bound on the Kullback-Leibler sampling error. By solving this optimal control problem, we obtain sufficient conditions on noise schedules under which $\tilde{\mathcal{O}}(d/n)$ sampling error is achievable, where d is the data dimension and n is the number of discretization steps. While existing theoretical works also prove that $\tilde{\mathcal{O}}(d/n)$ sampling error bounds are achievable, these results hold for specific noise schedules that differ from those used in practice. Under additional assumptions on the data distribution, we derive a family of noise schedules that recovers the functional forms of schedules widely used in practice. We show that these schedules achieve the $\tilde{\mathcal{O}}(d/n)$ error bound when their parameters are chosen according to our analysis. Our theory offers a principled approach to noise schedule design, yielding explicit functional forms and guidance for selecting their parameters.

1 INTRODUCTION

Score-based generative models (Ho et al., 2020; Song et al., 2021c;a; Liu et al., 2023) have emerged as a leading framework for approximate sampling. These methods formulate the generative process as the simulation of a time-reversed stochastic differential equation (SDE) driven by a learned score function. While the training objective is well-understood (Song et al., 2021b; Chen et al., 2023c;b; Zhang et al., 2025), the design of noise schedules governing the forward and reverse processes remains largely heuristic. Consequently, standard noise schedules (Ho et al., 2020; Nichol & Dhariwal, 2021; Jabri et al., 2023; Xu et al., 2022; von Platen et al., 2022) are often selected through empirical sweeps rather than theoretical guidance.

Noise schedule design plays a critical role in generating high-quality data using diffusion models (Song et al., 2021c; Nichol & Dhariwal, 2021; Chen, 2023; Reeves & Pfister, 2025). This importance is partially reflected in existing theory. When utilizing constant noise schedules and uniform discretization steps, the best-known bounds on the sampling error measured in the Kullback-Leibler (KL) divergence remain sub-optimal in the problem constants (Block et al., 2022; Chen et al., 2023c;a; Conforti et al., 2025). Specifically, these upper bounds typically scale as $\tilde{\mathcal{O}}(d^2/n)$ or $\tilde{\mathcal{O}}(\max\{\mathcal{J}_\star, d\}/n)$, where $\mathcal{J}_\star$ is the Fisher information of the data distribution, d is the data dimension, and n is the number of sampling steps. Because $\mathcal{J}_\star$ can be much larger than d, these guarantees suffer from severe problem-dependent penalties.

More recent works (Benton et al., 2024; Conforti et al., 2025) obtain improved $\widetilde{\mathcal{O}}(d/n)$ bounds on the KL divergence by carefully constructing nonuniform discretization grids. The discretization grids used to establish these guarantees are rarely deployed in practice (von Platen et al., 2022). Conversely, the noise schedules favored by practitioners lack comparable theoretical guarantees with their conventional parameter choices, leaving a gap between diffusion model theory and practice. We refer the reader to Section A for a more detailed discussion of related work.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In this work, we establish a new theoretical framework for designing noise schedules, which enlarges the class of schedules that are known to achieve $\tilde{\mathcal{O}}(d/n)$ KL error bounds. Directly minimizing the KL divergence is analytically intractable because the densities are governed by the Fokker-Planck partial differential equation (PDE). Using the Fisher information dynamics, we derive a variational upper bound on the KL sampling error whose optimization yields noise schedules achieving the $\tilde{\mathcal{O}}(d/n)$ rate. The bound is established under mild regularity conditions, and we consider additional assumptions on the data distribution to obtain explicit expressions for the resulting schedules. This family of Affinely Coupled Schedules (ACS) recovers the functional forms of noise schedules widely used in practice and prescribes how their parameters should scale with problem constants and the sampling budget to achieve $\tilde{\mathcal{O}}(d/n)$ KL error. Recent work (de Bortoli et al., 2025) also considers variational schedule optimization for an objective different from the KL divergence. Our formulation establishes how noise schedules used in practice can attain the $\tilde{\mathcal{O}}(d/n)$ error bound for the more commonly-used KL divergence.

Our contributions are summarized below:

1. **Optimal Control:** We formulate noise schedule design as an optimal control problem. The cost functional involves a variational bound on the discretization error with a non-trivial Lagrangian that admits path-dependent variations, and it replaces the intractable PDE representation of probability density evolution with a one-dimensional Fisher information ODE. In addition to Girsanov's theorem which is now widely used to theoretically study diffusion-based generative AI models (Song et al., 2021b; Chen et al., 2023c; Liu et al., 2022), we also use a number of other tools from probability and information theory such as Bochner's formula, concavity of entropy power, and Blachman-Stam inequality to derive the optimal control formulation.
2. **Error Bounds for a Class of Noise Schedules:** Under mild regularity conditions on the data distribution, we establish that the optimal control construction achieves $\tilde{\mathcal{O}}(d/n)$ KL sampling error for a class of noise schedules, provided we have knowledge of the Fisher information of the target distribution.
3. **Implementable Budget-Dependent Schedule Design:** Under additional assumptions on the target distribution, we show that the parameters of the noise schedules can be chosen as a function of n (the computational budget) without knowledge of the Fisher information to achieve $\tilde{\mathcal{O}}(d/n)$ sampling error.

General notation. The notation $\nabla\phi$ is used to denote the gradient when ϕ is a scalar-valued function, or the Jacobian when ϕ is a vector-valued function. Throughout, $\|\cdot\|$ denotes the standard element-wise ℓ_2 norm: the Euclidean norm for vectors, the Frobenius norm for matrices, and the Hilbert-Schmidt norm for higher-order tensors. For any probability density p and a vector-valued function ϕ, let $\|\phi\|_{L_2(p)} = \sqrt{\mathbb{E}_{X\sim p}\|\phi(X)\|^2}$. We use $\lesssim$ to denote an inequality ignoring absolute (problem-independent) constants in an inequality. The set of continuous functions defined on the interval $[0, T]$ is denoted by $C([0, T])$.

2 Preliminaries and Problem Setup

Diffusion models are based on two processes: the *forward process*, which transforms data into noise, and its time reversal, often referred to as the *reverse process*, which transforms noise back into data. The *noise schedule* controls the rates at which these transformations occur.

Forward process. Let $p_\star$ denote the density function of the target data distribution on $\mathbb{R}^d$. The forward process is a stochastic differential equation (SDE) of the following form:

$$dX_t = -f(t)X_t dt + \sqrt{g(t)}dB_t, \quad 0 \le t \le T, \quad X_0 \sim p_\star, \tag{1}$$

where $(B_t : t \in [0, T])$ is a standard Brownian motion on $\mathbb{R}^d$ independent of X_0, and $f(\cdot)$ and $g(\cdot)$ are real-valued functions of time. Let p_t denote the density function of X_t. We call the function pair (f, g) the *noise schedule*. Many theoretical studies of diffusion models focus on the simplified setting $f(t) = 1$ and $g(t) = 2$ for all t, in which case the forward process is the standard Ornstein–Uhlenbeck (OU) process.

To see how the noise schedule controls the distribution of the forward process, note that the marginal distribution of X_t admits the representation

$$X_t \overset{d}{=} \alpha_t X_0 + \sigma_t Z, \quad Z \sim \mathcal{N}(0, I), \tag{2}$$

where

$$\alpha_t = \exp\left(-\int_0^t f(\tilde{t})d\tilde{t}\right), \quad \sigma_t^2 = \alpha_t^2 \int_0^t \frac{g(\tilde{t})}{\alpha_{\tilde{t}}^2} d\tilde{t}.$$

In the OU setting where $f(t) = 1, g(t) = 2$, these quantities reduce to $\alpha_t = e^{-t}, \sigma_t^2 = 1 - e^{-2t}$. More generally, a larger value of $f(t)$ leads to a faster decay of the contribution of X_0 in the distribution of X_t, while a larger value of $g(t)$ corresponds to a faster injection of Gaussian noise. In this paper, we consider the class of noise schedules $\Theta = \mathcal{F} \times \mathcal{G}$ defined as

$$\begin{aligned} \mathcal{F} &= \{f \in C([0,T]) : f(t) \geq 0 \text{ for all } t \in [0,T]\}, \\ \mathcal{G} &= \{g \in C([0,T]) : g(t) > 0 \text{ for all } t \in [0,T]\}. \end{aligned} \tag{3}$$

The noise schedules used in most diffusion model implementations (von Platen et al., 2022) fall under the class Θ.

Reverse process. The reverse process is the time reversal of the forward process $(X_t : t \in [0,T])$ in (1). That is, the reverse process $(X_\tau^{\leftarrow} : \tau \in [0,T])$ is defined by $X_\tau^{\leftarrow} \overset{d}{=} X_{T-\tau}, \tau \in [0,T]$. Then the reverse process admits the following SDE representation (Anderson, 1982; Haussmann & Pardoux, 1986):

$$\begin{gathered} dX_\tau^{\leftarrow} = [f(T-\tau)X_\tau^{\leftarrow} + g(T-\tau)s_{T-\tau}(X_\tau^{\leftarrow})]\, d\tau + \sqrt{g(T-\tau)}dB_\tau, \\ 0 \leq \tau \leq T, \quad X_0^{\leftarrow} \sim p_T, \end{gathered} \tag{4}$$

where $(B_\tau : 0 \leq \tau \leq T)$ is a standard Brownian motion on $\mathbb{R}^d$ independent of $X_0^{\leftarrow}$, and the function $s_t(\cdot)$ defined as $s_t(x) \triangleq \nabla \log p_t(x)$ is called the score function.

2.1 Noise Schedule Optimization

To turn the reverse process (4) into an implementable algorithm for generating data, several approximations need to be made: (1) The score function s_t is unknown since it depends on the unknown data distribution, and is therefore replaced by an approximation $\hat{s}_t$, typically learned through a training process called score matching; (2) The reverse SDE cannot be solved exactly, but rather must be solved numerically with time discretization; (3) The reverse SDE cannot be initialized exactly from p_T, and is instead initialized from a Gaussian distribution so the initial state can be easily sampled.

A central question in the analysis of diffusion models is how these approximations affect the error between the distribution of the generated samples and the target distribution (Block et al., 2022; Chen et al., 2023c; Benton et al., 2024; Conforti et al., 2025). Existing analysis shows that the overall error can be separated into three additive terms: (1) score-matching error, (2) discretization error, and (3) initialization error, corresponding to the three approximations above. The score-matching error arises from the training stage, while the discretization error and the initialization error arise from the sampling stage when one solves the reverse SDE to generate data samples.

In this work, we establish a noise schedule design framework for sampling. When using diffusion models for generative tasks, a score function is first learned during a training phase. The trained network can then be reused with different noise schedules without changing its parameters, provided that the training phase covers the required SNR range as in (Karras et al., 2022). This allows us to study noise schedule design separately from the training procedure. In the main body of the paper, we focus on how noise schedules affect initialization and discretization errors assuming that the score function is known perfectly. Since the score matching error enters the sampling error bound as an additive term, it is easy to extend the results to the case where there is an error in the score function estimation; see Section B.2.

Specifically, we consider the following method for solving the reverse SDE. We select discretization time points $0 = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_n = T$, where $\tau_k = kh, k = 0, 1, 2, \ldots, n$ with a step

size $h > 0$, and solve the following SDE:

$$d\hat{X}_\tau^{\leftarrow} = \left[f(T-\tau)\hat{X}_\tau^{\leftarrow} + g(T-\tau)s_{T-\tau_k}(\hat{X}_{\tau_k}^{\leftarrow}) \right] d\tau + \sqrt{g(T-\tau)}dB_\tau, \\ \tau \in [\tau_k, \tau_{k+1}), \quad k = 0, 1, 2, \ldots, n-1, \quad \hat{X}_0^{\leftarrow} \sim \mathcal{N}(0, \sigma_T^2 I). \tag{5}$$

Note that here we take the approach where we use a constant step h and let the noise schedule (f, g) vary over time. This is without loss of generality, as discussed in Section B.3. We allow the schedule $(f, g) \in \Theta$ to depend on the number of discretization steps n.

Let $\hat{p}_\star$ denote the density of $\hat{X}_T^{\leftarrow}$, i.e., the output of the approximated reverse SDE in (5). Recall that $p_\star$ is the density of the data distribution. Then our goal is to minimize the Kullback-Leibler (KL) divergence between $p_\star$ and $\hat{p}_\star$, i.e., to solve the following problem

$$\begin{aligned} \underset{(f,g) \in \Theta}{\text{minimize}} \quad & \mathrm{KL}(p_\star \| \hat{p}_\star) \\ \text{subject to} \quad & p_\star \text{ is the density of } X_T^{\leftarrow} \text{ in (4)}, \\ & \hat{p}_\star \text{ is the density of } \hat{X}_T^{\leftarrow} \text{ in (5)}, \end{aligned} \tag{6}$$

where the constraints are specified by partial differential equations (PDEs) for the density functions of $X_t^{\leftarrow}$ and $\hat{X}_t^{\leftarrow}$ given by the Fokker–Planck equations. Solving optimization problems subject to PDE constraints is notoriously difficult (Hinze et al., 2008; Risken, 1989), especially in high dimensions. In this work, we present a principled framework to find the optimal noise schedule while circumventing the difficulty involved with solving this PDE-constrained optimization problem.

Regularity assumption on data distribution. For convenience, we use $X_\star \sim p_\star$ to denote a sample from the data distribution with density $p_\star$. We also generally use the subscript $\star$ for quantities associated with the data distribution. Let $s_\star \triangleq \nabla \log p_\star$ be the score function of $p_\star$. In our analysis, we frequently use the *Fisher information* of a distribution. For a density p with score function s, its Fisher information is defined as

$$\mathcal{J}(p) \triangleq \|\nabla \log p\|_{L_2(p)}^2 = \|s\|_{L_2(p)}^2. \tag{7}$$

The Fisher information of the data distribution is denoted by $\mathcal{J}_\star \coloneqq \|s_\star\|_{L_2}^2$.

Assumption 1. *We assume that* $\|X_\star\|_{L_2}^2$, $\|s_\star\|_{L_2}^2$, $\|\nabla s_\star\|_{L_2}$, *and* $\|\nabla^2 s_\star\|_{L_2}$ *are all finite.*

The L_2 norms in Assumption 1 are all with respect to $p_\star$. When Assumption 1 is not satisfied, early stopping can be applied where the target distribution $p_\star$ is replaced by p_δ for some $\delta > 0$ as in (Chen et al., 2023c; Benton et al., 2024; Bortoli, 2023; Lee et al., 2023).

3 Main Results

Our main results consist of two parts. In Section 3.1, we present an optimal control formulation for the noise schedule design problem, which leads to a class of schedules that achieve $\tilde{\mathcal{O}}(d/n)$ sampling error. In Section 3.2, we derive noise schedules with closed-form expressions under additional assumptions on the target data distribution. We will provide proof sketches of these results in Section 4, and defer the full proofs to Sections D and E.

3.1 An Optimal Control Formulation for Noise Schedule Design

Our objective is to solve the optimization problem (6), which minimizes $\mathrm{KL}(p_\star \| \hat{p}_\star)$, the KL-divergence between the data distribution $p_\star$ and the output distribution $\hat{p}_\star$. Here, the densities $p_\star$ and $\hat{p}_\star$ are functions on the high-dimensional space $\mathbb{R}^d$, and they depend on the noise schedule (f, g) through PDEs given by the Fokker–Planck equations. This PDE dependence makes directly solving the KL-minimization problem analytically intractable.

To circumvent this difficulty, a key idea in our approach is to work with the Fisher information $\mathcal{J}_t = \mathcal{J}(p_t)$. Unlike the density p_t, the Fisher information $\mathcal{J}_t$ is a scalar quantity. Thus, rather than

having its dynamics described by a PDE as p_t does, $\mathcal{J}_t$ evolves according to an ODE, derived using the heat flow and the Bochner formula (Lemma 3):

$$\dot{\mathcal{J}}_t = 2f(t)\mathcal{J}_t - g(t)\|\nabla s_t\|_{L_2}^2. \tag{8}$$

Theorem 1 below shows that the KL objective can be upper bounded in terms of $\mathcal{J}_t$.

Theorem 1. *Suppose Assumption 1 holds. For any noise schedule $(f, g) \in \Theta$ and any sufficiently small step size h, it holds that*

$$\mathrm{KL}(p_\star \| \hat{p}_\star) \lesssim \frac{\alpha_T^2}{\sigma_T^2}\|X_\star\|_{L_2(p_\star)}^2 + hd\int_0^T \left(2f(t) - \frac{\dot{\mathcal{J}}_t}{\mathcal{J}_t}\right)^2 dt. \tag{9}$$

Theorem 1 allows us to formulate the noise schedule design problem as the following optimal control problem:

$$\begin{aligned} \underset{(f,g)\in\Theta}{\text{minimize}} \quad & \frac{\alpha_T^2}{\sigma_T^2}\|X_\star\|_{L_2(p_\star)}^2 + hd\int_0^T \left(2f(t) - \frac{\dot{\mathcal{J}}_t}{\mathcal{J}_t}\right)^2 dt \\ \text{subject to} \quad & \dot{\mathcal{J}}_t = 2f(t)\mathcal{J}_t - g(t)\|\nabla s_t\|_{L_2}^2, \\ & \mathcal{J}_0 = \mathcal{J}_\star, \quad \mathcal{J}_T = J_T. \end{aligned} \tag{10}$$

This optimal control problem (10) is significantly more tractable thanks to the ODE constraint. Moreover, we show in Theorem 2 that its solution, although obtained by minimizing an upper bound on the KL objective, indeed achieves $\tilde{\mathcal{O}}(d/n)$ sampling error in the KL divergence.

Theorem 2. *Suppose Assumption 1 holds. For any sufficiently large n, let $f \in \mathcal{F}$ satisfy $f_{\max} \le (2T)^{-1}\log(n)$. Then, there exists $g^* \in \mathcal{G}$ such that*

$$\mathrm{KL}(p_\star \| \hat{p}_\star) \lesssim \frac{d}{n}\log^2\left(\frac{\|X_\star\|_{L_2}^2}{d}\cdot\frac{\mathcal{J}_\star}{d}\cdot n\right). \tag{11}$$

Compared with the constant-schedule bounds discussed in Section 1, this bound improves the dependence on d and $\mathcal{J}_\star$. Consistent with the best-known bounds for the constant schedule, we show in Section F that under our framework, an optimized bound for VP-constant schedules retains a linear dependence on $\mathcal{J}_\star$, even when constants depend on problem constants or n.

The proof of Theorem 2 reveals that g^* takes the form

$$g^*(t) = \lambda\frac{\mathcal{J}_t^*}{\|\nabla s_t\|_{L_2}^2}, \tag{12}$$

where λ is a constant independent of t and $(\mathcal{J}_t^* : t \in [0, T])$ is the optimal Fisher information trajectory. The objective in (10) is invariant to f when g is chosen as $g = g^*$. Therefore, one can choose f freely (e.g., as a function of n), subject to the constraints of Theorem 2, and then choose the corresponding g^* that solves (10). The choice of f is important in how it affects the resulting form of g^*. Some choices of f yield schedules that are easier to implement, as we show in Corollary 1.

3.2 Noise Schedule Design

In this section, we consider target data distributions that satisfy the following additional assumption.

Assumption 2. *The target data distribution is either strongly log-concave (SLC), or belongs to a class of isotropic Gaussian mixture models (GMMs). Specifically, the density $p_\star$ of the data distribution satisfies one of the following conditions:*

(i) (SLC) For some $m_\star, M_\star > 0$,

$$m_\star I \preceq -\nabla^2 \log p_\star \preceq M_\star I. \tag{13}$$

(ii) (GMM) For some mixture weights $\{\pi_i\}_{i=1}^L$, means $\{\mu_i\}_{i=1}^L$ with $\mu_i \in \mathbb{R}^d$, and $\nu > 0$, the density $p_\star(x) = \sum_{i=1}^L \pi_i \varphi(x; \mu_i, \nu^2 I)$, where $\varphi(x; \mu_i, \nu^2 I)$ is the density of $\mathcal{N}(\mu_i, \nu^2 I)$.

Although real-world data distributions are highly complex and may not satisfy Assumption 2, we use this assumption to derive an implementable class of noise schedules. Under Assumption 2, we show in Lemma 7 that the ratio $\mathcal{J}_t^2/\|\nabla s_t\|_{L_2}^2$ is bounded above and below by positive multiples of d, uniformly in $t \in [0,T]$, with constants depending only on the target distribution. Using these bounds, we construct an explicit family of noise schedules whose parameters can be chosen as functions of problem constants and n to attain the $\tilde{\mathcal{O}}(d/n)$ sampling error rate.

Theorem 3. *Suppose Assumptions 1 and 2 hold. For every sufficiently large n and any continuous $f \in \mathcal{F}$, let q be a continuous function such that $q(t) \geq 2f(t)$ for all $t \in [0,T]$. Then, for any such q and constants $c, C > 0$ independent of n and d,*

$$\sup_{t\in[0,T]} q(t) \leq C\log(1+n), \quad \int_0^T \exp\left(\int_0^t q(\tilde{t})\,d\tilde{t}\right) dt \geq \frac{c}{g_0}\frac{\|X_\star\|_{L_2}^2}{d}\frac{n}{\log(1+n)},$$

where g_0 satisfies $0 < g_0 \leq C\frac{d}{\mathcal{J}_\star}\log(1+n)$. Then, the schedule (f, g_q°) defined as

$$g_q^\circ(t) = g_0 \exp\left(-\int_0^t \big(2f(\tilde{t}) - q(\tilde{t})\big)\,d\tilde{t}\right) \tag{14}$$

achieves $\tilde{\mathcal{O}}(d/n)$ sampling error in the KL divergence, ignoring problem constants in Assumption 2.

Theorem 2 establishes a $\tilde{\mathcal{O}}(d/n)$ sampling error bound under the mild regularity conditions in Assumption 1. Theorem 3 imposes additional distributional assumptions to obtain an explicit family of schedules attaining this rate. As we show below, this family encompasses the functional forms of several schedules used in practice, including VE-exponential, VP-sigmoid, and VP-linear. Beyond recovering these forms, the theorem provides insights on how their parameters should scale with problem constants and the sampling budget n to obtain the $\tilde{\mathcal{O}}(d/n)$ KL error guarantee. These conditions guide both the choice of parameters for familiar schedules and the construction of new schedules with the same guarantee.

Theorem 3 gives a family of schedules through choices of f and q. We now impose an affine relationship between f and g to show how Theorem 3 can be used for noise schedule design.

$$\Theta_{\text{ACS}} = \{(f,g) \in \Theta : f(t) = \theta g(t) + \omega \text{ for some } \theta, \omega \geq 0\}. \tag{15}$$

We call this class affinely coupled schedules (ACSs). This ACS family provides a theoretical framework for several empirically successful noise schedules, and identifies conditions under which they attain the $\tilde{\mathcal{O}}(d/n)$ sampling error rate.

Corollary 1. *For a schedule (f, g_q°) of the form* (14) *in the ACS class* (15)*,*

$$g_q^\circ(t) = \frac{g_0 \exp\left(\int_0^t \big(q(\tilde{t}) - 2\omega\big)\,d\tilde{t}\right)}{1 + 2\theta g_0 \int_0^t \exp\left(\int_0^s \big(q(\tilde{t}) - 2\omega\big)\,d\tilde{t}\right) ds}, \qquad f(t) = \theta g_q^\circ(t) + \omega. \tag{16}$$

The ACS family recovers the VE-exponential schedule (Song & Ermon, 2019) and VP-linear (Ho et al., 2020). It also yields sigmoidal VP schedules similar to those used in (Xu et al., 2022; Jabri et al., 2023; Chen, 2023). The corresponding choices of q, θ, ω, and g_0 are:

- **VE-Exponential:** Setting $q(t) = \bar{q} = (2/T)\log(1+n)$ and $(\theta, \omega) = (0,0)$ gives $g_q^\circ(t) = g_0 e^{\bar{q}t}$. A choice of g_0 independent of n and the above q satisfy the conditions of Theorem 3 for all sufficiently large n.
- **VP-Sigmoid:** Setting $q(t) = \bar{q} = \frac{2}{T}\log(1+n)$, $g_0 > 0$ independent of n, and $(\theta, \omega) = (1/2, 0)$ gives
$$g_q^\circ(t) = \frac{g_0 \bar{q}}{g_0 + (\bar{q} - g_0)e^{-\bar{q}t}}.$$
The conditions of Theorem 3 are satisfied for sufficiently large n.
- **VP-Linear:** For $(\theta, \omega) = (1/2, 0)$, $g_{\min} = g_0 > 0$ independent of n, and $g_{\max} = g_{\min} + \frac{2}{T}\log(1+n)$,
$$q(t) = g_{\min} + \frac{g_{\max} - g_{\min}}{T}t + \frac{g_{\max} - g_{\min}}{T g_{\min} + (g_{\max} - g_{\min})t}$$
satisfies the conditions of Theorem 3 for sufficiently large n. The corresponding $g_q^\circ(t)$ is the linear schedule $g_q^\circ(t) = g_{\min} + (g_{\max} - g_{\min})t/T$.

We show in Section E.6 that the choice of q in VP-Sigmoid and VP-Linear above satisfies the conditions of Theorem 3 for sufficiently large n, and therefore achieves the $\tilde{\mathcal{O}}(d/n)$ error.

Corollary 1 goes beyond recovering the functional forms of VP-Sigmoid and VP-Linear. Standard implementations typically choose schedule parameters independently of n, whereas the examples above provide explicit n-dependent parameterizations. Once the n-independent constants are chosen, these formulas adapt each schedule to the inference budget without separate retuning.

4 Proof Sketches for Theorems 1 to 3

Here we present proof sketches for Theorems 1 to 3. The full proofs and intermediate lemmas are presented in Sections D and E.

Proof Sketch for Theorem 1. The goal is to upper bound $\mathrm{KL}(p_\star \| \hat{p}_\star)$, the KL-divergence between the data distribution $p_\star$ and the output distribution $\hat{p}_\star$.

Step 1: Error Decomposition. Following standard techniques that use Girsanov theorem and data processing inequality to analyze SDE samplers (see, e.g., (Song et al., 2021b; Chen et al., 2023c; Liu et al., 2022)), we can decompose the total sampling error $\mathrm{KL}(p_\star \| \hat{p}_\star)$ into the initialization error and the accumulated discretization error over the reverse SDE steps:

$$\mathrm{KL}(p_\star \| \hat{p}_\star) \le \underbrace{\frac{\alpha_T^2}{2\sigma_T^2} \|X_\star\|^2_{L_2(p_\star)}}_{\text{initialization error}} + \underbrace{\frac{1}{2} \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} g(t) \left\| s_t(X_t) - s_{t_{k+1}}(X_{t_{k+1}}) \right\|^2_{L_2} dt}_{\text{discretization error}}. \tag{17}$$

Here the discretization error is written in terms of forward process quantities using the distributional equivalence $X_\tau^{\leftarrow} \overset{d}{=} X_{T-\tau}, \tau \in [0,T]$, and $t_k = T - \tau_{n-k}$. The remainder of the proof focuses on bounding the discretization error.

Step 2: Local Discretization Error. We now bound $\left\| s_t(X_t) - s_{t_{k+1}}(X_{t_{k+1}}) \right\|^2_{L_2}$ for $t \in [t_k, t_{k+1}]$. For this, it is more convenient to work with $s_{T-\tau}(X_\tau^{\leftarrow})$ and show that it satisfies an SDE (Lemma 2):

$$ds_{T-\tau}(X_\tau^{\leftarrow}) = -f(T-\tau) s_{T-\tau}(X_\tau^{\leftarrow}) d\tau + \sqrt{g(T-\tau)} \nabla s_{T-\tau}(X_\tau^{\leftarrow}) dB_\tau. \tag{18}$$

When analyzing the discretization error of an SDE within a small time interval of length h, the key fact to leverage is the Itô isometry, which says for a stochastic integral with respect to Brownian motion, $\mathbb{E}[(\int Y_t dB_t)^2] = \int \mathbb{E}[Y_t^2] dt$. For sufficiently small h, the discretization error due to the drift term is smaller than the error arising from the diffusion term. Integrating and using $g \in \mathcal{G}$ gives

$$\int_{\tau_k}^{\tau_{k+1}} g(T-\tau) \left\| s_{T-\tau}(X_\tau^{\leftarrow}) - s_{T-\tau_k}(X_{\tau_k}^{\leftarrow}) \right\|^2_{L_2} d\tau \lesssim h \int_{\tau_k}^{\tau_{k+1}} g(T-\tau)^2 \|\nabla s_{T-\tau}(X_\tau^{\leftarrow})\|^2_{L_2} d\tau. \tag{19}$$

Step 3: Fisher Information ODE. Now a key term to tackle is $\|\nabla s_t(X_t)\|^2_{L_2}$ in the bound (19). Using Bochner's formula, we show that it is related to the Fisher information via the following ODE (Lemma 3):

$$\dot{\mathcal{J}}_t = 2f(t)\mathcal{J}_t - g(t)\|\nabla s_t(X_t)\|^2_{L_2} \quad \Leftrightarrow \quad g(t)\|\nabla s_t(X_t)\|^2_{L_2} = 2f(t)\mathcal{J}_t - \dot{\mathcal{J}}_t. \tag{20}$$

Step 4: Final Error Bound. Substituting (20) into the discretization error gives

$$\text{discretization error} \lesssim h \int_0^T g(t)(2f(t)\mathcal{J}_t - \dot{\mathcal{J}}_t) dt. \tag{21}$$

However, this is still not easy to optimize since it involves $f(t), g(t), J_t$, which are related. Therefore, we eliminate $g(t)$ by writing the bound as

$$h \int_0^T g(t)(2f(t)\mathcal{J}_t - \dot{\mathcal{J}}_t) dt = h \int_0^T \frac{(2f(t)\mathcal{J}_t - \dot{\mathcal{J}}_t)^2}{\|\nabla s_t(X_t)\|^2_{L_2}} dt, \tag{22}$$

and using the concavity of entropy power $\mathcal{J}_t^2 \le d\|\nabla s_t(X_t)\|^2_{L_2}$ (Lemma 4), which gives the final error bound

$$\text{discretization error} \lesssim hd \int_0^T \left(2f(t) - \frac{\dot{\mathcal{J}}_t}{\mathcal{J}_t} \right)^2 dt. \tag{23}$$

Proof Sketch for Theorem 2. For any f, we aim to find the optimal g^*, which must produce a trajectory $(\mathcal{J}_t^* : t \in [0,T])$ that minimizes the upper bound on the discretization error in (23). This problem can be solved via the Euler-Lagrange equation, which states that to minimize a functional $\int_0^T \mathcal{L}(t, u_t, \dot{u}_t)dt$, the optimal condition is described by

$$\frac{\partial \mathcal{L}}{\partial u_t} - \frac{d}{dt}\frac{\partial \mathcal{L}}{\partial \dot{u}_t} = 0. \tag{24}$$

Let $u_t = \log \mathcal{J}_t$, so $\dot{u}_t = \dot{\mathcal{J}}_t/\mathcal{J}_t$ and $\mathcal{L}(t, u_t, \dot{u}_t) = (2f(t) - \dot{u}_t)^2$. Then the optimal condition for the optimal trajectory $(\mathcal{J}_t^* : t \in [0,T])$ is

$$2f(t) - \frac{\dot{\mathcal{J}}_t^*}{\mathcal{J}_t^*} = \lambda, \tag{25}$$

where $\lambda > 0$ is a time-independent constant. This condition also gives the characterization of the optimal g^* in (12). Then the discretization error bound becomes

$$\text{discretization error} \lesssim hd\lambda^2 T.$$

Balancing the rates of the initialization and discretization errors yields the $\mathrm{KL}(p_\star \| \hat{p}_\star) = \tilde{\mathcal{O}}(d/n)$ bound.

Proof Sketch for Theorem 3. We show that the choice of $g_q^\circ(t)$ in (14) implies the $\tilde{\mathcal{O}}(d/n)$ rate under Assumption 2. The key implication of Assumption 2 is that lower and upper bounds on $\mathcal{J}_t^2/\|\nabla s_t\|_{L_2}^2$ can be established: for some problem constant κ,

$$\frac{d}{\kappa^2} \le \frac{\mathcal{J}_t^2}{\|\nabla s_t\|_{L_2}^2} \le d.$$

The schedule $g_q^\circ(t)$ was constructed to satisfy $\frac{\dot{g^\circ}_q(t)}{g_q^\circ(t)} = q(t) - 2f(t)$ so that the ODE (20) yields

$$\frac{d}{dt}\big(g_q^\circ(t)\mathcal{J}_t\big) \le q(t) g_q^\circ(t)\mathcal{J}_t - \frac{\big(g_q^\circ(t)\mathcal{J}_t\big)^2}{d}.$$

By the choice of g_0 in Theorem 3, $g_0\mathcal{J}_\star \le Cd\log(1+n)$. Together with the condition $\sup_{t\in[0,T]} q(t) \le C\log(1+n)$, this gives

$$g_q^\circ(t)\mathcal{J}_t \le Cd\log(1+n), \qquad t \in [0,T].$$

Moreover, $q(t) \ge 2f(t)$ implies that g_q° is nondecreasing. Following the proof of Theorem 1, we use these properties to obtain that the discretization error is bounded by a multiple of

$$\frac{d\log^2(1+n)}{n} + \frac{d\log^3(1+n)}{n^2}.$$

The second condition on q ensures that the initialization error is $\tilde{\mathcal{O}}(d/n)$, and we conclude $\mathrm{KL}(p_\star \| \hat{p}_\star) = \tilde{\mathcal{O}}(d/n)$.

5 Experiments

Our theoretical analyses show that ACS encompasses several existing noise schedules and identify conditions under which these schedules achieve the $\mathrm{KL}(p_\star \| \hat{p}_\star) = \tilde{\mathcal{O}}(d/n)$ sampling error rate. This does not imply equivalent finite-budget performance: different schedules induce different Fisher-information trajectories and may differ in the constants and tradeoffs suppressed by the asymptotic bound, leaving open how different members of the ACS family compare on real-world data.

We describe our choice of parametrization for the noise schedule (f_n, g_n) in Theorem 3. For some time-varying function ψ, let $q_n(t) = \lambda_n \psi(t)$. Following the proof of Theorem 3 in Section E.4, the bound on the sampling error can be written approximately as

$$C_1 e^{-\lambda_n} + C_2 \frac{\lambda_n^2}{n},$$

where C_1, C_2 are problem constants. For any n, the minimizer of this approximate upper bound is given by $\lambda_n = \mathcal{W}(Kn)$ for some constant K independent of n, where $\mathcal{W}$ is the Lambert-W function. This choice of λ_n balances the initialization error proportional to $e^{-\lambda_n}$ and the discretization error $\sim \lambda_n^2/n$. Therefore, we choose (f_n, g_n) to vary with n only through their dependence on λ_n. The exact parameterization is given in Section G.

We evaluate all schedules using identical pretrained checkpoints from (Karras et al., 2022; 2024), without fine-tuning, corresponding to each dataset and the same ancestral sampler. "Power" uses the published (Karras et al., 2022) EDM power law noise grid ($\sigma_{\min} = 0.002, \sigma_{\max} = 80, \rho = 7$). The ACS and all other schedule hyperparameters were tuned once to minimize the mean log-FID scores at NFE 10 and 50 with 10,000 samples on CIFAR-10 and ImageNet 64×64, then held fixed for all NFEs and datasets. Performance on these held-out datasets is reported in Figure 1, and results on the tuning datasets are reported in Figure 2.

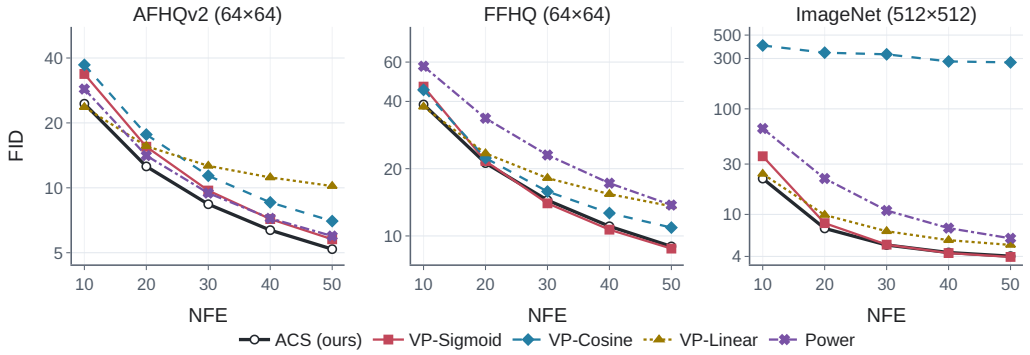

Figure 1: FID on 50,000 generated samples across inference budgets on datasets not used for hyperparameter selection. VP-Linear yields substantially lower FID than VP-Sigmoid at NFE 10, whereas VP-Sigmoid generally yields lower FID at higher NFE budgets. ACS balances these two regimes and remains competitive across the full range of inference budgets. The cosine schedule performed well on its tuning datasets and the held-out datasets AFHQv2 and FFHQ, but transferred poorly to ImageNet 512×512.

Figure 1 shows that the relative performance of existing schedules depends on the inference budget: VP-Linear and Power are competitive at low NFEs, whereas VP-Sigmoid becomes more competitive as NFE increases. ACS remains competitive across the full NFE range. These results show that the n-dependent ACS construction balances the inference budget tradeoffs exhibited by standard schedules.

6 Conclusion

We introduced an optimal control framework for designing noise schedules, utilizing a Fisher information ODE to bypass intractable PDE constraints. Our approach guarantees $\tilde{\mathcal{O}}(d/n)$ bounds on the KL divergence and yields a novel class of Affinely Coupled Schedules (ACS) whose n-dependent parameterization achieves low FID across image-generation tasks and inference step counts. Our sampling error guarantee holds under mild regularity conditions, and the additional SLC or GMM assumptions are used to obtain explicit schedule constructions. Future work will seek similarly explicit constructions under weaker distributional assumptions.

AI use statement

Generative AI tools were primarily used to proofread this manuscript and review code used to run the experiments. They were also used to facilitate experiments, for example, writing shell scripts for efficient job allocations on a high-performance cluster or assistance with editing scripts to generate plots. We have reviewed all AI-assisted work.

We have not used generative AI tools to help develop theoretical models or conceptual frameworks, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, or assist with translation. The remaining required disclosures of using generative AI tools to generate synthetic datasets, clean and reformat dataset, support qualitative and thematic data analysis, or interpret results are not applicable to this work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

The assumptions underlying our theoretical results are stated in Sections 2–3, and complete proofs are provided in Appendices D–F. Appendix G specifies the noise-schedule parameterization, hyperparameter search spaces and selected values, pretrained models, datasets, sampling protocol, and evaluation procedure used in our experiments.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2623–2631, 2019.

Brian D.O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3), 1982.

Muhammad H. Ashiq, Samanyu Arora, Abhinav N. Harish, Ishaan Kharbanda, Hung Yun Tseng, and Grigorios G. Chrysos. Why ddim hallucinates more than ddpm: A theoretical analysis of reverse dynamics, 2026. URL https://arxiv.org/abs/2605.06831.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*, volume 348. Springer Science & Business Media, 2013.

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

N. Blachman. The convolution inequality for entropy powers. *IEEE Transactions on Information Theory*, 11(2):267–271, April 1965.

Adam Block, Youssef Mroueh, and Alexander Rakhlin. Generative Modeling with Denoising Auto-Encoders and Langevin Sampling. October 2022.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, (arXiv:2208.05314), May 2023.

Herm Jan Brascamp and Elliott H Lieb. On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *Journal of functional analysis*, 22(4):366–389, 1976.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pp. 4735–4763. PMLR, 2023a.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *International Conference on Machine Learning*, pp. 4672–4712. PMLR, 2023b.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru 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*, 2023c.

Sitan Chen, Vasilis Kontonis, and Kulin Shah. Learning general gaussian mixtures with efficient score matching. *arXiv preprint arXiv:2404.18893*, 2024.

Ting Chen. On the importance of noise scheduling for diffusion models, 2023.

Yifan Chen, Eric Vanden-Eijnden, and Jiawei Xu. Lipschitz-guided design of interpolation schedules in generative models. *arXiv preprint arXiv:2509.01629*, 2025.

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

Valentin de Bortoli, Romuald Elie, Anna Kazeykina, Zhenjie Ren, and Jiacheng Zhang. Dimension-free error estimate for diffusion model and optimal scheduling, 2025. URL https://arxiv.org/abs/2512.01820.

A. Dembo. Simple proof of the concavity of the entropy power with respect to added Gaussian noise. *IEEE Transactions on Information Theory*, 35(4):887–888, July 1989. ISSN 1557-9654.

R. Esposito. On a relation between detection and estimation in decision theory. *Inf. Control*, 12(2): 116–120, February 1968.

Xuefeng Gao and Lingjiong Zhu. Convergence analysis for general probability flow odes of diffusion models in wasserstein distances. *arXiv preprint arXiv:2401.17958*, 2024.

Khashayar Gatmiry, Jonathan Kelner, and Holden Lee. Learning mixtures of gaussians using diffusion models. *arXiv preprint arXiv:2404.18869*, 2024.

Leonard Gross. Logarithmic Sobolev Inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975. ISSN 0002-9327. doi: 10.2307/2373688.

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

Michael Hinze, René Pinnau, Michael Ulbrich, and Stefan Ulbrich. *Optimization with PDE constraints*. Springer Science & Business Media, 2008.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4 (2):251–257, 1991.

Allan Jabri, David Fleet, and Ting Chen. Scalable adaptive computation for iterative generation, 2023.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24174–24184. IEEE, 2024.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pp. 946–985. PMLR, 2023.

Xingchao Liu, Lemeng Wu, Mao Ye, and Qiang Liu. Let us build bridges: Understanding and extending diffusion generative models. *arXiv preprint arXiv:2208.14699*, 2022.

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

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*. PMLR, 2021.

Galen Reeves and Henry D. Pfister. Information-theoretic proofs for diffusion sampling, 2025. URL https://arxiv.org/abs/2502.02305.

Hannes Risken. Fokker-planck equation. In *The Fokker-Planck equation: methods of solution and applications*. Springer, 1989.

H. Robbins. An empirical Bayes approach to statistics. In *Third Berkeley Symp. Math Statist. Probab*, 1956.

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

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34, 2021b.

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*, 2021c.

A.J. Stam. Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Information and Control*, 2(2), June 1959.

C. Villani. A short proof of the "concavity of entropy power". *IEEE Transactions on Information Theory*, 46(4), July 2000.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations*, 2022.

Yiran Zhang, Weihang Xu, Mo Zhou, Maryam Fazel, and Simon Shaolei Du. Convergence dynamics of over-parameterized score matching for a single gaussian. *arXiv preprint arXiv:2511.22069*, 2025.

A Additional Related Works

Early theoretical analyses of diffusion models (Block et al., 2022; Bortoli, 2023; Chen et al., 2023c) focused predominantly on SDE samplers utilizing first-order discretization schemes under constant noise schedules. A primary objective in this line of work has been to obtain a bound on the sampling error, typically measured in the total variation (TV) distance. Because their analysis is restricted to constant noise schedules, their bounds scale sub-optimally with respect to the data dimension d and problem constants. Chen et al. (2023a); Conforti et al. (2025) strengthened these results by establishing bounds on the KL divergence, which implies a similar scaling in the TV distance through Pinsker's inequality.

The effect of time-varying noise schedules has been explored more recently. Benton et al. (2024); Conforti et al. (2025) showed that by utilizing time-varying noise schedules, the KL error bound's dependence on the dimension d and problem constants can be improved significantly such that it grows linearly with d. Their results imply that the iteration complexity $n \asymp d/\epsilon^2$ is sufficient to achieve $\mathrm{KL}(p_\star \| \hat{p}_\star) \lesssim \epsilon^2$ error bounds. These improved error bounds are achieved by considering constant (f, g) and time-varying discretization step size $\{h_k\}$. As mentioned in (Chen et al., 2023a, Remark 3), such schedules can be viewed as fixing a constant-in-time f and using g that increases exponentially with time. While these works establish $\tilde{\mathcal{O}}(d/n)$ rates, the specific schedule constructions required by their analyses depart from the standard affinely coupled schedules commonly favored in empirical implementations.

Beyond tuning the standard time-reversal schedule, a conceptually related problem explored by Chen et al. (2025) seeks to design a reverse process that differs from the exact time-reversal of the forward process. To design this mismatched schedule, Chen et al. (2025) propose to minimize the Lipschitz constant of the resulting reverse drift. However, such generation processes also depart from standard empirical pipelines. In contrast, our framework aligns with practice by maintaining the exact time-reversal of the true forward SDE and directly minimizes a rigorous upper bound on the KL sampling error, rather than relying on a regularity proxy.

Lastly, noise schedule design was considered in the context of Probability Flow ODEs. Gao & Zhu (2024) analyze the Wasserstein-2 error, requiring the data distribution be strongly log-concave. Under this regime, they conclude that VE schedules exhibit worse iteration complexity than VP schedules. Although ODE samplers can be effective at low NFE, stochastic samplers can produce higher-quality samples (Ashiq et al., 2026). Our work analyzes stochastic SDE samplers using the KL divergence and propose an optimal control framework based on Fisher information dynamics. We require only mild regularity conditions to prove $\tilde{\mathcal{O}}(d/n)$ bounds, where a strongly log-concave or a Gaussian mixture model is assumed only to derive closed-form parametrizations. Under this setting, we prove that both several existing schedules for SDE samplers can achieve the $\tilde{\mathcal{O}}(d/n)$ error rate if their parameters are chosen to scale with the number of discretization steps n appropriately.

B Supplement to Section 2

B.1 Statistical Equivalence of Score Functions

We describe how to map a score function trained for one noise schedule to the score function corresponding to another noise schedule using their statistical equivalence as discussed in (Chen et al., 2025). In this section, we use τ_t to be a map from one time-scale t to another time-scale τ_t.

Let $X_t^{(1)} = \alpha_t^{(1)} X_\star + \sigma_t^{(1)} Z$ and $X_t^{(2)} = \alpha_t^{(2)} X_\star + \sigma_t^{(2)} Z$ be two random variables parameterized by $(\alpha_t^{(i)}, \sigma_t^{(i)})$. We first identify the equivalent time τ_t in the first schedule that matches the SNR of the second schedule at time t:

$$\frac{\alpha_{\tau_t}^{(1)}}{\sigma_{\tau_t}^{(1)}} = \frac{\alpha_t^{(2)}}{\sigma_t^{(2)}}$$

Because the SNRs are identical, $X_t^{(2)}$ and $X_{\tau_t}^{(1)}$ differ strictly by a scaling factor. Specifically,

$$X_t^{(2)} = \sigma_t^{(2)} \left(\frac{\alpha_t^{(2)}}{\sigma_t^{(2)}} X_\star + Z \right) = \sigma_t^{(2)} \left(\frac{\alpha_{\tau_t}^{(1)}}{\sigma_{\tau_t}^{(1)}} X_\star + Z \right) = \frac{\sigma_t^{(2)}}{\sigma_{\tau_t}^{(1)}} X_{\tau_t}^{(1)}.$$

Applying the change of variables formula for probability densities, $p_t^{(2)}(x) = \left(\frac{\sigma_{\tau_t}^{(1)}}{\sigma_t^{(2)}}\right)^d p_{\tau_t}^{(1)}\left(\frac{\sigma_{\tau_t}^{(1)}}{\sigma_t^{(2)}}x\right)$. Taking the gradient of the logarithm of this density with respect to x yields the translation:

$$s_t^{(2)}(x) = \frac{\sigma_{\tau_t}^{(1)}}{\sigma_t^{(2)}} s_{\tau_t}^{(1)}\left(\frac{\sigma_{\tau_t}^{(1)}}{\sigma_t^{(2)}}x\right) \tag{26}$$

Score Function from Noise Prediction Model: It is common to train a noise prediction model

$$S_t(x) = \mathbb{E}[Z \mid X_t = x]$$

in place of the score function s_t. The score of the conditional distribution of $(X_t \mid X_\star) \sim \mathcal{N}(\alpha_t X_\star, \sigma_t^2 I)$ is given by

$$\nabla_{x_t} \log p(x_t \mid x_\star) = -\frac{x_t - \alpha_t x_\star}{\sigma_t^2} = -\frac{Z}{\sigma_t}.$$

The marginal score is the expectation of the conditional score, which gives the general relationship

$$s_t(x) = -\frac{1}{\sigma_t} S_t(x)$$

B.2 Non-Zero Score Matching Error

For any noise schedule (f, g) and reverse-time grid $\{\tau_k\}_{k=0}^n$, define the score matching error ϵ_{sc} as

$$\epsilon_{sc}^2 \coloneqq \sum_{k=0}^{n-1}\left(\int_{\tau_k}^{\tau_{k+1}} g(T-\tau)\, d\tau\right) \|s_{T-\tau_k} - \hat{s}_{T-\tau_k}\|_{L_2(p_{T-\tau_k})}^2 .$$

Using that

$$s_{T-\tau}(X_\tau^{\leftarrow}) - \hat{s}_{T-\tau_k}(X_{\tau_k}^{\leftarrow}) = s_{T-\tau}(X_\tau^{\leftarrow}) - s_{T-\tau_k}(X_{\tau_k}^{\leftarrow}) + s_{T-\tau_k}(X_{\tau_k}^{\leftarrow}) - \hat{s}_{T-\tau_k}(X_{\tau_k}^{\leftarrow}),$$

and the Girsanov theorem and the data processing inequality as in (Chen et al., 2023c; Benton et al., 2024; Conforti et al., 2025), the bound (17) is replaced by

$$\mathrm{KL}(p_\star \| \hat{p}_\star) \le \epsilon_{sc}^2 + \frac{\alpha_T^2}{2\sigma_T^2}\|X_\star\|_{L_2}^2 + \sum_{k=0}^{n-1}\int_{\tau_k}^{\tau_{k+1}} g(T-\tau)\left\|s_{T-\tau}(X_\tau^{\leftarrow}) - s_{T-\tau_k}(X_{\tau_k}^{\leftarrow})\right\|_{L_2}^2 d\tau. \tag{27}$$

To address the impact of the score matching error ϵ_{sc}^2, we evaluate two distinct paradigms: utilizing an existing pre-trained score network, and training a new model from scratch.

Pre-Trained Statistical Equivalence: Let (f_1, g_1) and (f_2, g_2) denote two distinct noise schedules with corresponding marginal distributions characterized by parameters $(\alpha_t^{(1)}, \sigma_t^{(1)})$ and $(\alpha_t^{(2)}, \sigma_t^{(2)})$. Suppose there exists a monotonically increasing time-map τ_t such that the Signal-to-Noise Ratios (SNRs) of the two schedules are strictly equivalent, i.e.,

$$\frac{\alpha_{\tau_t}^{(1)}}{\sigma_{\tau_t}^{(1)}} = \frac{\alpha_t^{(2)}}{\sigma_t^{(2)}}, \quad \forall t \in [0, T_2]. \tag{28}$$

As established in Section B.1, under this SNR equivalence, the random variables $X_t^{(2)}$ and $X_{\tau_t}^{(1)}$ differ only by a deterministic spatial scaling. Consequently, the score network $\hat{s}^{(2)}$ corresponding to schedule (f_2, g_2) can be derived from a pre-trained score function $\hat{s}^{(1)}$ as in (26). Applying this change of variables to both the exact and approximate scores gives

$$\left\|s_t^{(2)} - \hat{s}_t^{(2)}\right\|_{L_2(p_t^{(2)})}^2 = \left(\frac{\sigma_{\tau_t}^{(1)}}{\sigma_t^{(2)}}\right)^2 \left\|s_{\tau_t}^{(1)} - \hat{s}_{\tau_t}^{(1)}\right\|_{L_2(p_{\tau_t}^{(1)})}^2 .$$

Thus, ϵ_{sc}^2 for (f_2, g_2) is controlled by the pre-trained model's errors at the matched grid times, with the spatial scaling factors above and the weights in its definition.

Universal Approximation: Alternatively, when training a score network $\hat{s}_t(x)$ from scratch under noise schedule $(f, g) \in \Theta$, the score matching error can be driven arbitrarily close to zero. To formalize this, we rewrite the score matching objective as an L_2 distance with respect to a joint space-time measure. Define the measure on $\mathbb{R}^d \times [0, T]$ by

$$d\mu(x, t) = \sum_{k=0}^{n-1} \left(\int_{\tau_k}^{\tau_{k+1}} g(T - \tau)\, d\tau \right) p_{T-\tau_k}(x)\, dx\, \delta_{T-\tau_k}(dt),$$

where $\delta_{T-\tau_k}$ is the Dirac measure at $T - \tau_k$. The score matching error is then equivalently given by:

$$\epsilon_{sc}^2 = \int_0^T \int_{\mathbb{R}^d} \|\hat{s}_t(x) - s_t(x)\|^2 d\mu(x, t). \tag{29}$$

Because $p_t(x)$ integrates to 1 over $\mathbb{R}^d$, the total measure is $\mu(\mathbb{R}^d \times [0, T]) = \int_0^T g(t)dt$. For any schedule $(f, g) \in \Theta$, this integral is finite, making μ a finite measure. Furthermore, the target score function $s_t(x)$ belongs to $L_2(\mu)$, since

$$\|s_t\|_{L_2(\mu)}^2 = \sum_{k=0}^{n-1} \left(\int_{\tau_k}^{\tau_{k+1}} g(T - \tau)\, d\tau \right) \mathcal{J}_{T-\tau_k} < \infty. \tag{30}$$

The finiteness of this sum follows from the integrability of $g(t)$ and Assumption 1. Under these conditions, the universal approximation theorem (e.g., (Hornik, 1991)) guarantees that the class of neural networks is dense in $L_2(\mu)$. Therefore, for each fixed schedule and grid, there exists a network parameterization with arbitrarily small ϵ_{sc}^2 given sufficient model capacity.

B.3 Constant Step Size Suffices for Time-Varying Noise Schedules

We show that choosing (f, g) with constant step size is without loss of generality. Consider a choice of noise schedule (f_1, g_1) and time-varying step size with time grid $\{\tau_k^{(1)}\}$. Denote by $\{\hat{X}^{\leftarrow,(1)}_{\tau_k^{(1)}}\}$ the EI discretized process according to this noise schedule and time grid. We show that there exists a noise schedule (f_2, g_2) and time grid $\{\tau_k^{(2)}\}$ generated by constant step size h, i.e., $\tau_k^{(2)} = kh$, such that the EI discretized process $\{\hat{X}^{\leftarrow,(2)}_{\tau_k^{(2)}}\}$ admits the same distribution as $\{\hat{X}^{\leftarrow,(1)}_{\tau_k^{(1)}}\}$.

Fix (f_1, g_1), $\{\tau_k^{(1)}\}$, and $\{\tau_k^{(2)}\}$ where $\tau_k^{(2)} = kh$. Let $\phi : [0, T] \to [0, T]$ be a continuously differentiable map with $\phi'(\tau) > 0$ for all $\tau \in [0, T]$ and $\phi(\tau_k^{(2)}) = \tau_k^{(1)}$ for all k. We consider the construction

$$f_2(T - \tau^{(2)}) = f_1(T - \phi(\tau^{(2)})) \cdot \phi'(\tau^{(2)}), \quad g_2(T - \tau^{(2)}) = g_1(T - \phi(\tau^{(2)})) \cdot \phi'(\tau^{(2)}), \tag{31}$$

for all $\tau^{(2)} \in [0, T]$. The SDE governing $\{\hat{X}^{\leftarrow,(1)}_{\tau_k^{(1)}}\}$ is

$$d\hat{X}^{\leftarrow,(1)}_{\tau^{(1)}} = \left[f_1(T - \tau^{(1)}) X^{(1)}_{\tau^{(1)}} + g_1(T - \tau^{(1)}) s_{T-\tau_k^{(1)}}(X^{(1)}_{\tau_k^{(1)}}) \right] d\tau^{(1)} + \sqrt{g_1(T - \tau^{(1)})} dB_{\tau^{(1)}} \tag{32}$$

Substituting the differentials

$$d\tau^{(1)} = \phi'(\tau^{(2)}) d\tau^{(2)}, \quad dB_{\phi(\tau^{(2)})} = \sqrt{\phi'(\tau^{(2)})} dB_{\tau^{(2)}}$$

for (32), we obtain

$$\begin{aligned} d\hat{X}^{\leftarrow,(2)}_{\tau^{(2)}} = \Big[& f_1(T - \phi(\tau^{(2)})) \hat{X}^{\leftarrow,(2)}_{\tau^{(2)}} \\ & + g_1(T - \phi(\tau^{(2)})) s_{T-\tau_k^{(1)}}(\hat{X}^{(2)}_{\tau_k^{(2)}}) \Big] \phi'(\tau^{(2)}) d\tau^{(2)} + \sqrt{g_1(T - \phi(\tau^{(2)})) \phi'(\tau^{(2)})} dB_{\tau^{(2)}}. \end{aligned}$$

Substituting (31) for the above equation, we obtain

$$\begin{aligned} d\hat{X}^{\leftarrow,(2)}_{\tau^{(2)}} = & \left[f_2(T - \tau^{(2)}) \hat{X}^{\leftarrow,(2)}_{\tau^{(2)}} + g_2(T - \tau^{(2)}) s_{T-\phi(\tau_k^{(2)})}(X^{\leftarrow,(2)}_{\tau_k^{(2)}}) \right] d\tau^{(2)} \\ & + \sqrt{g_2(T - \tau^{(2)})} dB_{\tau^{(2)}}. \end{aligned}$$

Since $\hat{X}^{\leftarrow,(2)}_{\tau^{(2)}} \overset{d}{=} \hat{X}^{\leftarrow,(1)}_{\tau^{(1)}}$, the sequence of variables at the respective grid points $\{\tau_k^{(1)}\}$ and $\{\tau_k^{(2)}\}$ have the same distribution.

C PRELIMINARIES

Throughout the paper, we use the notation (α_t, σ_t) to denote the coefficients of $X_t = \alpha_t X_\star + \sigma_t Z$, which is the solution to the forward SDE (1). The coefficients can be evaluated as $\alpha_t = \exp(-\int_0^t f(\tilde{t})d\tilde{t})$ and σ_t^2 as a solution to the ODE

$$\frac{d}{dt}\sigma_t^2 = -2f(t)\sigma_t^2 + g(t), \quad \sigma_0 = 0. \tag{33}$$

The name "variance preserving" is used to refer to schedules satisfying $f(t) > 0$ and $g(t) = 2f(t)$ because they preserve the variance when $\mathrm{Var}(X_\star) \coloneqq \|X_\star\|_{L_2}^2 = \|Z\|_{L_2}^2$.

Notations: For the proofs in the appendix, we use t and $\tau = T - t$ to denote the corresponding forward and reverse time indices. The notation $\otimes$ is used to denote the tensor product, and $\|\cdot\|$ is used to denote the Hilbert-Schmidt norm for order r tensors in $(\mathbb{R}^d)^{\otimes r}$. The divergence of a vector-valued function ϕ is denoted by $\nabla \cdot \phi$.

C.1 NOISE SCHEDULES

In this section, we describe the coefficients (f, g) that correspond to various noise schedules used in practice.

Exponential: Variance exploding models were used in (Song & Ermon, 2019), where the corresponding SDE form (Song et al., 2021c) is given by

$$dX_t = \sqrt{g(t)}dB_t,$$

where $g(t)$ is chosen to grow exponentially with t.

Sigmoid: We consider the sigmoid schedule of Chen (2023). Let $\sigma(x) = 1/(1 + e^{-x})$ and $z(t) = (u_- + (u_+ - u_-)t/T)/\eta_{\text{sig}}$, where $u_- < u_+$ and $\eta_{\text{sig}} > 0$. Excluding the clipping, the schedule uses

$$\alpha_t^2 = \frac{\sigma(u_+/\eta_{\text{sig}}) - \sigma(z(t))}{\sigma(u_+/\eta_{\text{sig}}) - \sigma(u_-/\eta_{\text{sig}})}.$$

The corresponding VP coefficients satisfy $f(t) = g_{\text{sigmoid}}(t)/2$, where

$$g_{\text{sigmoid}}(t) = -\frac{d}{dt}\log\alpha_t^2 = \frac{u_+ - u_-}{T\eta_{\text{sig}}}\,\frac{\sigma(z(t))(1 - \sigma(z(t)))}{\sigma(u_+/\eta_{\text{sig}}) - \sigma(z(t))}. \tag{34}$$

For values of t such that $1 - \sigma(u_+/\eta_{\text{sig}})$ is small relative to $1 - \sigma(z(t))$, this rate is approximately $\frac{u_+ - u_-}{T\eta_{\text{sig}}}\sigma(z(t))$.

Linear: The linear schedule, proposed in (Ho et al., 2020), parametrizes the sample path solution

$$X_t = \sqrt{\bar{\alpha}_t}X_\star + \sqrt{1 - \bar{\alpha}_t}Z, \quad \bar{\alpha}_{t_k} = \prod_{j=0}^{k-1}(1 - \beta_{t_j}). \tag{35}$$

The sequence $\{\beta_i\}$ is a linear function $\beta_t = \beta_{\min} + (\beta_{\max} - \beta_{\min}) \cdot (t/T)$. As shown in (Song et al., 2021c), the expression for f in the limit $n \to \infty$ is

$$f_{lin}(t) \coloneqq \frac{1}{2}\left(\beta_{\min} + (\beta_{\max} - \beta_{\min})\frac{t}{T}\right).$$

The parameters are set with $\beta_{\min} = 0.1/T, \beta_{\max} = 20/T$. While (f_{lin}, g_{lin}) are obtained in the limit $n \to \infty$, we refer to the pair as linear schedules.

Cosine: The cosine schedule (Nichol & Dhariwal, 2021) refers to the choice

$$\sqrt{\bar{\alpha}_t} = \frac{\cos\left(\frac{t/T+s}{1+s} \cdot \frac{\pi}{2}\right)}{\cos\left(\frac{s}{1+s} \cdot \frac{\pi}{2}\right)}, \qquad s = 0.008.$$

Let $\theta(t) = (t/T + s)\pi/(2(1+s))$ so that $\sqrt{\bar{\alpha}_t} = \cos\theta(t)/\cos\theta(0)$. In the SDE (1), this is equivalently written as the VP constraint $f = g/2$ with $\bar{\alpha}_t = \exp\left(-2\int_0^t f(\tilde{t})d\tilde{t}\right)$, which yields

$$\begin{aligned} f(t) = -\frac{d}{dt}\log\sqrt{\bar{\alpha}_t} &= -\frac{d}{dt}\left(\log\cos\theta(t) - \log\cos\theta(0)\right) \\ &= \frac{\sin\theta(t)}{\cos\theta(t)}\theta'(t) \\ &= \frac{\pi}{2T(1+s)}\tan\left(\frac{t/T+s}{1+s}\cdot\frac{\pi}{2}\right). \end{aligned}$$

As a result, the cosine schedule corresponds to $g(t) = (\pi/T(1+s))\tan((t/T+s)/(1+s)\cdot(\pi/2))$.

C.2 Generalized Blachman-Stam Inequality

Recall the Blachman-Stam inequality (Stam, 1959; Blachman, 1965) which states that for $X_t = \alpha_t X_\star + \sigma_t Z$,

$$\frac{1}{\mathcal{J}(\alpha_t X_\star + \sigma_t Z)} \geq \frac{1}{\mathcal{J}(\alpha_t X_\star)} + \frac{1}{\mathcal{J}(\sigma_t Z)} = \frac{\alpha_t^2}{\mathcal{J}_\star} + \frac{\sigma_t^2}{d}.$$

Denote by $\mathrm{PS}(a,b)$ for two positive scalars a, b the parallel sum $(a^{-1}+b^{-1})^{-1}$. Rearranging and using the notation $\mathcal{J}_t = \mathcal{J}(\alpha_t X_\star + \sigma_t Z)$, we obtain the Blachman-Stam inequality

$$\mathcal{J}_t \leq \mathrm{PS}\left(\frac{\mathcal{J}_\star}{\alpha_t^2}, \frac{d}{\sigma_t^2}\right).$$

We prove a generalization of Stam's inequality that holds for higher order derivatives.

Lemma 1. *Let p_t and $p_{t|\star}$ be the density and (Gaussian) transition kernels of $X_t = \alpha_t X_\star + \sigma_t Z$ with $X_\star \perp Z$. Suppose $m \geq 1$, $t \in (0,T]$, and $\nabla^m p_\star \in L_1(\mathrm{Leb}^d)$. Then,*

$$\left\|\frac{\nabla^m p_t}{p_t}\right\|_{L_2}^2 \leq \mathrm{PS}\left(\alpha_t^{-2m}\left\|\frac{\nabla^m p_\star}{p_\star}\right\|_{L_2}^2, \frac{\|H_m\|_{L_2(\gamma^d)}^2}{\sigma_t^{2m}}\right). \tag{36}$$

Importantly, this bound interpolates the squared L_2 norm of $\nabla^m p_\star/p_\star$ and that of the limit $\alpha_t/\sigma_t \to 0$.

Proof. First, we show that Assumption 1 implies that the boundary condition vanishes sufficiently fast for the integration by parts formula

$$\int \nabla_{x_\star}^m p_{t|\star}(x|x_\star)p_\star(x_\star)dx_\star = (-1)^j \int \nabla^{m-j} p_{t|\star}(x|x_\star)\otimes\nabla^j p_\star(x_\star)dx_\star, \quad \forall j \in \{1,\cdots,m\} \tag{37}$$

to hold. Denote by ϕ the Gaussian transition kernel. The case for $j=1$ simplifies to

$$\int(\nabla^2\phi)p_\star dx = -\int \nabla\phi\otimes\nabla p_\star dx \Leftrightarrow \int\nabla(p_\star\nabla\phi)dx_\star = 0.$$

The second identity above holds if $p_\star\nabla\phi$ and $\nabla(p_\star\nabla\phi)$ are in L_1. Since $\nabla\phi$ is bounded,

$$\int\|p_\star\nabla\phi\|dx_\star \leq \|\nabla\phi\|_\infty\int p_\star dx_\star < \infty.$$

Therefore, $p_\star\nabla\phi \in L_1(\mathrm{Leb}^d)$. Observe $\nabla\cdot(p_\star\nabla\phi) = p_\star\nabla^2\phi + \langle\nabla\phi,\nabla p_\star\rangle$. Since $\nabla^2\phi$ is bounded, the first term is in L_1. The second term is bounded in L_1 since

$$\int\|\nabla p_\star\|dx_\star = \int\frac{\|\nabla p_\star\|}{p_\star}p_\star dx_\star \leq \left(\int\|\nabla\log p_\star\|^2 p_\star dx_\star\right)^{1/2} < \infty.$$

The case for $j=2$ follows by observing that

$$-\int\nabla\phi\otimes\nabla p_\star dx_\star = \int\phi\nabla^2 p_\star dx_\star \Leftrightarrow \int\nabla(\phi\nabla p_\star) = 0,$$

and repeating the same steps above, observing that $\nabla \cdot (\phi \nabla p_\star)$ involves $\nabla \log p_\star$ and $\nabla^2 \log p_\star$. The same argument applies to any m and $j \in \{1, \cdots, m\}$.

Let $H_m(x) = (-1)^m e^{x^2/2} \nabla^m e^{-x^2/2}$ be the m-th order Hermite polynomial. It holds by the marginalization $p_t(x) = \int p_{t|x}(x|x_\star) p_\star(x_\star) dx_\star$ and the identity

$$\nabla^m p_{t|0}(x|x_\star) = \frac{(-1)^m}{\sigma_t^m} H_m\left(\frac{x - \alpha_t x_\star}{\sigma_t}\right) p_{t|0}(x|x_\star) \tag{38}$$

that

$$\begin{aligned}\frac{\nabla^m p_t(x)}{p_t(x)} &= \int \frac{(-1)^m}{\sigma_t^m} H_m(z) \frac{p_{t|0}(x|x_\star)}{p_t(x)} p_\star(x_\star) dx_\star \\ &= \frac{(-1)^m}{\sigma_t^m} \mathbb{E}\left[H_m(Z) | X_t = x\right], \end{aligned} \tag{39}$$

where $m = 1$ recovers Tweedie's formula (Robbins, 1956; Esposito, 1968)

$$s_t(x) = \frac{\alpha_t \mathbb{E}[X_\star | X_t = x] - x}{\sigma_t^2}. \tag{40}$$

Observe the identity

$$\nabla_x p_{t|\star}(x|x_\star) = -\frac{1}{\alpha_t} \nabla_{x_\star} p_{t|\star}(x|x_\star),$$

which holds by $p_{t|\star}(x|x_\star)$ being a Gaussian transition probability $\mathcal{N}(\alpha_t x_\star; \sigma_t I)$. Applying the same derivatives m times, we obtain

$$\nabla_x^m p_{t|\star}(x|x_\star) = \frac{(-1)^m}{\alpha_t^m} \nabla_{x_\star}^m p_{t|\star}(x|x_\star).$$

Marginalizing, we obtain

$$\nabla_x^m p_t(x) = \nabla_x^m \int p_{t|\star}(x|x_\star) p_\star(x_\star) dx_\star = \int \frac{(-1)^m}{\alpha_t^m} \nabla_{x_\star}^m p_{t|\star}(x|x_\star) p_\star(x_\star) dx_\star$$

Using the integration by parts property (37) repeatedly, we obtain

$$\begin{aligned}\frac{(-1)^m}{\alpha_t^m} \int \nabla_{x_\star}^m p_{t|\star}(x|x_\star) p_\star(x_\star) dx_\star &= \frac{(-1)^{m-1}}{\alpha_t^m} \int \nabla^{m-1} p_{t|\star}(x|x_\star) \otimes \nabla p_\star(x_\star) dx_\star \\ &= \frac{1}{\alpha_t^m} \int p_{t|\star}(x|x_\star) \nabla^m p_\star(x_\star) dx_\star.\end{aligned}$$

Dividing by $p_t(x)$, we obtain the identity

$$\begin{aligned}\frac{\nabla^m p_t(x)}{p_t(x)} &= \frac{1}{\alpha_t^m} \int \frac{\nabla^m p_\star(x_\star)}{p_\star(x_\star)} \frac{p_{t|\star}(x|x_\star) p_\star(x_\star)}{p_t(x)} dx_\star \\ &= \alpha_t^{-m} \mathbb{E}\left[\frac{\nabla^m p_\star(X_\star)}{p_\star(X_\star)} \mid X_t = x\right].\end{aligned}$$

To summarize, we obtained above and in (39) the two identities

$$\begin{aligned}\frac{\nabla^m p_t(x)}{p_t(x)} &= \alpha_t^{-m} \mathbb{E}\left[\frac{\nabla^m p_\star(X_\star)}{p_\star(X_\star)} \mid X_t = x\right], \\ \frac{\nabla^m p_t(x)}{p_t(x)} &= \sigma_t^{-m} (-1)^m \mathbb{E}\left[H_m(Z) \mid X_t = x\right].\end{aligned}$$

For any θ, it holds that their linear combination is equal to $\nabla^m p_t(x)/p_t(x)$. By Jensen's inequality and $X_\star \perp Z$, we obtain

$$\begin{aligned}\left\| \frac{\nabla^m p_t}{p_t} \right\|_{L_2}^2 &\le \mathbb{E} \left\| (1-\theta) \alpha_t^{-m} \frac{\nabla^m p_\star(X_\star)}{p_\star(X_\star)} + \theta \sigma_t^{-m} (-1)^m H_m(Z) \right\|^2 \\ &= (1-\theta)^2 \alpha_t^{-2m} \left\| \frac{\nabla^m p_\star}{p_\star} \right\|_{L_2}^2 + \theta^2 \sigma_t^{-2m} \left\| H_m \right\|_{L_2(\gamma^d)}^2,\end{aligned}$$

where the cross term vanishes because $Z \perp X_\star$ and $\mathbb{E} H_m(Z) = 0$. Minimizing over θ, we obtain (36). □

A consequence of Stam's inequality is that $\alpha_T^2\|X_\star\|_{L_2}^2 \ll \sigma_T^2 d$ implies $\mathcal{J}_T \approx \sigma_T^{-2} d$. Recall Tweedie's formula (40), which implies

$$\mathcal{J}_T = \frac{1}{\sigma_T^2}\mathbb{E}\|\mathbb{E}[Z|X_T]\|^2 = \frac{1}{\sigma_T^2}\left(d - \mathbb{E}\mathrm{TrCov}(Z|X_T)\right).$$

Using $Z = \sigma_T^{-1}(X_T - \alpha_T X_\star)$, we have

$$\mathbb{E}\mathrm{TrCov}(Z|X_T) = \frac{\alpha_T^2}{\sigma_T^2}\mathbb{E}\mathrm{TrCov}(X_\star|X_T) \le \frac{\alpha_T^2}{\sigma_T^2}\mathrm{TrCov}(X_\star).$$

Substituting for Tweedie's formula above, we have

$$\frac{1}{\sigma_T^2}\left(d - \frac{\alpha_T^2}{\sigma_T^2}\|X_\star\|_{L_2}^2\right) \le \mathcal{J}_T \le \frac{d}{\sigma_T^2}, \tag{41}$$

which implies $\mathcal{J}_T \approx \sigma_T^{-2} d$ when $\alpha_T^2\|X_\star\|_{L_2}^2 \ll \sigma_T^2 d$.

C.3 Score Function Dynamics Along the Reverse SDE

The SDE describing the score function $s_{T-\tau}(X_\tau^{\leftarrow}) = \nabla \log p_{T-\tau}(X_\tau^{\leftarrow})$ along the reverse process $(X_\tau^{\leftarrow})$ with constant schedules was established in (Conforti et al., 2025, Proposition 3.2). Assuming the regularity conditions in Assumption 1, we derive the corresponding SDE for time-varying schedules.

Lemma 2. *Under Assumption 1, it holds that*

$$ds_{T-\tau}(X_\tau^{\leftarrow}) = -f(T-\tau)s_{T-\tau}(X_\tau^{\leftarrow})d\tau + \sqrt{g(T-\tau)}\nabla s_{T-\tau}(X_\tau^{\leftarrow})dB_\tau$$

is well-defined, and $\|s_t\|_{L_2}, \|\nabla s_t\|_{L_2} < \infty$ *for all* $t \in [0,T]$.

Proof. We use $t = T - \tau$. The Fokker-Planck equation governing the evolution of the density p_t along the SDE (1) is

$$\frac{\partial}{\partial t}p_t = \nabla \cdot (f(t)xp_t) + \frac{1}{2}g(t)\nabla \cdot \nabla p_t. \tag{42}$$

By the chain rule, we obtain the point-wise evaluation

$$\frac{\partial}{\partial t}\log p_t = f(t)d + f(t)\langle x, s_t\rangle + \frac{1}{2}g(t)(\nabla \cdot s_t + \|s_t\|^2).$$

Using the identities $\nabla(x \cdot s_t) = s_t + (\nabla s_t)x$, $\nabla\|s_t\|^2 = 2(\nabla s_t)s_t$ and taking the gradient on both sides of the above equation, we obtain

$$\partial_t s_t = f(t)s_t + f(t)(\nabla s_t)x + \frac{1}{2}g(t)\nabla(\nabla \cdot s_t) + g(t)(\nabla s_t)s_t, \tag{43}$$

where $(\nabla s_t)x$ is understood as the matrix-vector product of the Jacobian ∇s_t and the vector x. Using the Itô formula, we obtain that the drift driving $ds_{T-\tau}(X_\tau^{\leftarrow})$ along the reverse process is

$$\begin{aligned}&\left(-\partial_t s_t\right)(X_\tau^{\leftarrow}) + \nabla s_t(X_\tau^{\leftarrow})\left[f(t)X_\tau^{\leftarrow} + g(t)s_t(X_\tau^{\leftarrow})\right] + \frac{g(t)}{2}\nabla(\nabla \cdot s_t)(X_\tau^{\leftarrow})\\ &= -f(t)s_t(X_\tau^{\leftarrow}).\end{aligned}$$

Substituting for the Itô formula, we obtain

$$ds_{T-\tau}(X_\tau^{\leftarrow}) = -f(t)s_t(X_\tau^{\leftarrow})d\tau + \sqrt{g(t)}(\nabla s_t)(X_\tau^{\leftarrow})dB_\tau.$$

Since $\nabla s_t = \frac{\nabla^2 p_t}{p_t} - \frac{\nabla p_t \nabla p_t^\dagger}{p_t^2}$, the finiteness of $\|\nabla s_t\|_{L_2}$ follows from Lemma 1. □

C.4 Regularity Properties of the Score Function

Gaussian mixture models form an important class of distributions that is often used to model real-world data. We show that Assumption 1 is satisfied by a Gaussian mixture model. Consider $p_0(x) = \sum_{i=1}^{\ell} w_i \mathcal{N}(x; \mu_i, \Sigma_i)$, where each Σ_i is positive definite. The score function $s_0(x)$ can be evaluated as in (Chen et al., 2024), where

$$s_0(x) = -\sum_{i=1}^{\ell} \pi_i(x)\Sigma_i^{-1}(x-\mu_i), \quad \pi_i(x) = \frac{w_i \mathcal{N}(x;\mu_i,\Sigma_i)}{p_0(x)}.$$

Taking the gradient on both sides, we obtain

$$\nabla s_0(x) = -\sum_{i=1}^{\ell} \pi_i(x)\Sigma_i^{-1} + \sum_{i=1}^{\ell} \pi_i(x)(-\Sigma_i^{-1}(x-\mu_i) - s_0(x))(-\Sigma_i^{-1}(x-\mu_i) - s_0(x))^{\dagger}.$$

Since the score function of each component is affine and $0 \le \pi_i(x) \le 1$, the score and its first two derivatives grow at most polynomially in $\|x\|$. A finite Gaussian mixture has finite moments of every order, so the required L_2 norms are finite. For $t > 0$, p_t remains a finite Gaussian mixture, and the same argument applies.

C.5 Fisher Information Dynamics

The following states the Fisher information dynamics in (8).

Lemma 3. *Suppose Assumption 1 holds. For any* $(f, g) \in \Theta$,

$$\dot{\mathcal{J}}_t = 2f(t)\mathcal{J}_t - g(t)\|\nabla s_t\|_{L_2}^2.$$

Proof. In this proof, we use the notation $p_t^X = p_t$ to distinguish between the density p_t^X of the process (X_t) from the process (Y_t) defined as

$$X_t = \alpha_t Y_t. \tag{44}$$

By the Itô formula, the two processes are related through the SDE $dX_t = \dot{\alpha}_t Y_t dt + \alpha_t dY_t$. Substituting $\dot{\alpha}_t = -f(t)\alpha_t$ and using (44), we obtain

$$dY_t = -\frac{\dot{\alpha}_t}{\alpha_t} Y_t dt + dX_t = f(t)Y_t dt - f(t)\frac{X_t}{\alpha_t}dt + \frac{\sqrt{g(t)}}{\alpha_t}dB_t = \frac{\sqrt{g(t)}}{\alpha_t}dB_t. \tag{45}$$

The density p_t^Y of Y_t is a convolution between p_0^Y and a Gaussian kernel, and differentiating under the integral yields the well-known identity (e.g., (Stam, 1959, Section 5))

$$\partial_t p_t^Y = \frac{g(t)}{2\alpha_t^2}\Delta p_t^Y.$$

The Bochner formula (Villani, 2000) yields the identity

$$\frac{d}{dt}\mathcal{J}(p_t^Y) = -\frac{g(t)}{\alpha_t^2}\|\nabla^2 \log p_t^Y\|_{L_2}^2. \tag{46}$$

Moreover, using the change of variables $p_t^X(x) = \alpha_t^{-d} p_t^Y(x/\alpha_t) \Rightarrow \nabla \log p_t^X(x) = \alpha_t^{-1}\nabla \log p_t^Y(x/\alpha_t)$, we obtain the relation between the Fisher information of the two laws:

$$\mathcal{J}(p_t^X) = \alpha_t^{-2}\mathcal{J}(p_t^Y). \tag{47}$$

Differentiating $\mathcal{J}(p_t^X)$ in (47), we obtain

$$\begin{aligned}\frac{d}{dt}\mathcal{J}(p_t^X) &= \left[-2\alpha^{-3}\dot{\alpha}_t\right]\mathcal{J}(p_t^Y) + \alpha_t^{-2}\frac{d}{dt}\mathcal{J}(p_t^Y) \\ &= 2f(t)\alpha_t^{-2}\mathcal{J}(p_t^Y) + \alpha_t^{-2}\frac{d}{dt}\mathcal{J}(p_t^Y).\end{aligned}$$

Substituting (47) and $\nabla_x^2 \log p_t^X(x) = \alpha_t^{-2}\nabla^2 \log p_t^Y(x/\alpha_t)$, we obtain

$$\frac{d}{dt}\mathcal{J}(p_t^X) = 2f(t)\mathcal{J}(p_t^X) - \frac{g(t)}{\alpha_t^4(t)}\|\nabla^2 \log p_t^Y\|_{L_2}^2 = 2f(t)\mathcal{J}(p_t^X) - g(t)\|\nabla^2 \log p_t^X\|_{L_2}^2.$$

□

We use the following "concavity of entropy power" theorem (Dembo, 1989; Villani, 2000).

Lemma 4. *For any $X_\star$ with density $p_\star$ satisfying Assumption 1, it holds that*

$$\mathcal{J}_t^2 \le d\|\nabla s_t\|_{L_2}^2.$$

Proof. Recall that the solution (X_t) to (1) can be written as (2). Because the density p_t is obtained as a convolution of $p_\star$ with a Gaussian density, p_t is infinitely differentiable and the integration by parts formula

$$\mathbb{E}\nabla \log p_t \nabla \log p_t^\dagger = -\mathbb{E}\nabla^2 \log p_t, \quad \forall t \in (0, T]$$

holds. The relation $\|\nabla s_t\|_{L_2,F}^2 \ge \mathcal{J}_t^2/d$ is shown next:

$$\begin{aligned}\mathcal{J}_t = \mathbb{E}\|\nabla \log p_t\|^2 = \mathrm{Tr}\mathbb{E}\nabla \log p_t \nabla \log p_t^\dagger &\underset{(*)}{=} -\mathrm{Tr}\mathbb{E}\nabla^2 \log p_t \\ &= -\mathrm{Tr}\mathbb{E}\nabla s_t,\end{aligned}$$

where $(*)$ follows from integration by parts. By Jensen's inequality and the Cauchy-Schwartz inequality,

$$\mathcal{J}_t^2 = (\mathbb{E}\mathrm{Tr}\nabla s_t)^2 \le \mathbb{E}(\mathrm{Tr}\nabla s_t)^2 \le d\mathrm{Tr}\mathbb{E}(\nabla s_t)^2 \coloneqq d\|\nabla s_t\|_{L_2}^2.$$

□

D Proofs of Theorems 1 and 2 (Section 3.1)

D.1 Proof of Theorem 1

We first prove the bound on the initialization error. When (f, g) are constants, the log-Sobolev inequality (Gross, 1975) and the contraction of the KL divergence along the Ornstein-Uhlenbeck process (Bakry et al., 2013) can be used to obtain a bound as in (Conforti et al., 2025). A different bound can be obtained as in (Benton et al., 2024, Proposition 4) and (Chen et al., 2023a, Lemma 9) based on the convexity of the KL divergence. For time-varying schedules, we find it to be more useful to utilize a small modification to the bound emerging from the latter argument.

By (2), we have $(X_T|X_\star) \sim \mathcal{N}(\alpha_T X_\star, \sigma_T^2 I)$. Let $\mathcal{I}$ denote the mutual information so that $\mathcal{I}(X_\star; X_T) = \mathbb{E}_{p_\star}[\mathrm{KL}(\mathcal{N}(\alpha_T X_\star, \sigma_T^2 I)\|p_T)]$. Using the chain rule, we obtain the bound

$$\begin{aligned}\mathrm{KL}(\mathrm{Law}(X_T)\|\mathcal{N}(0, \sigma_T^2 I)) &= \mathbb{E}_{p_\star}\left[\mathrm{KL}(\mathrm{Law}(X_T|X_0)\|\mathcal{N}(0, \sigma_T^2 I))\right] - \mathcal{I}(X_0; X_T) \\ &\le \mathbb{E}_{p_\star}\left[\mathrm{KL}(\mathcal{N}(\alpha_T X_\star, \sigma_T^2 I)\|\mathcal{N}(0, \sigma_T^2 I))\right] \\ &= \frac{\alpha_T^2}{2\sigma_T^2}\|X_\star\|_{L_2(p_\star)}^2.\end{aligned}$$

where the inequality follows from non-negativity of $\mathcal{I}$.

Recall the Girsanov bound (17). We use Lemma 2 to obtain the identity

$$s_t(X_\tau^{\leftarrow}) - s_{T-\tau_k}(X_{\tau_k}^{\leftarrow}) = -\int_{\tau_k}^{\tau}\left[f(\tilde{t})s_{\tilde{t}}(X_{\tilde{\tau}}^{\leftarrow})\right]d\tilde{\tau} + \int_{\tau_k}^{\tau}\sqrt{g(\tilde{t})}\nabla s_{\tilde{t}}(X_{\tilde{\tau}}^{\leftarrow})dB_{\tilde{\tau}}, \tag{48}$$

where $t = T - \tau$ and $\tilde{t} = T - \tilde{\tau}$. Evaluating the squared L_2 norm, we have the bound

$$\begin{aligned}&\left\|s_t(X_\tau^{\leftarrow}) - s_{T-\tau_k}(X_{\tau_k}^{\leftarrow})\right\|_{L_2}^2 \\ &\le 2(\tau - \tau_k)\int_{\tau_k}^{\tau} f(\tilde{t})^2\left\|s_{\tilde{t}}(X_{\tilde{\tau}}^{\leftarrow})\right\|_{L_2}^2 d\tilde{\tau} + 2\int_{\tau_k}^{\tau} g(\tilde{t})\left\|\nabla s_{\tilde{t}}(X_{\tilde{\tau}}^{\leftarrow})\right\|_{L_2}^2 d\tilde{\tau} \\ &\le 2h\int_{\tau_k}^{\tau} f(\tilde{t})^2\mathcal{J}_{\tilde{t}}d\tilde{\tau} + 2\int_{\tau_k}^{\tau} g(\tilde{t})\left\|\nabla s_{\tilde{t}}(X_{\tilde{\tau}}^{\leftarrow})\right\|_{L_2}^2 d\tilde{\tau} \\ &\lesssim \int_{\tau_k}^{\tau} g(\tilde{t})\|\nabla s_{\tilde{t}}(X_{\tilde{\tau}}^{\leftarrow})\|_{L_2}^2 d\tilde{\tau},\end{aligned} \tag{49}$$

where the last holds when h is sufficiently small such that $h < 1$ and satisfies

$$h \leq \min_{k \leq n-1} \min_{\tau \in [\tau_k, \tau_{k+1}]} \frac{\int_{\tau_k}^{\tau} g(\tilde{t}) \|\nabla s_{\tilde{t}}(X_{\tilde{\tau}}^{\leftarrow})\|_{L_2}^2 d\tilde{\tau}}{\int_{\tau_k}^{\tau} f(\tilde{t})^2 \mathcal{J}_{\tilde{t}} d\tilde{\tau}}. \tag{50}$$

We show that $h \lesssim 1/n$ satisfies (50) in Proposition 1.

Since g is positive and continuous on $[0, T]$, for sufficiently small h, $g(T-\tau) \leq 2g(T-\tilde{\tau})$ whenever $|\tau - \tilde{\tau}| \leq h$. Integrating the preceding inequality and interchanging the order of integration, we obtain

$$\begin{aligned}
&\int_{\tau_k}^{\tau_{k+1}} g(t) \left\| s_t(X_\tau^{\leftarrow}) - s_{T-\tau_k}(X_{\tau_k}^{\leftarrow}) \right\|_{L_2}^2 d\tau \\
&\lesssim \int_{\tau_k}^{\tau_{k+1}} g(T-\tau) \int_{\tau_k}^{\tau} g(T-\tilde{\tau}) \|\nabla s_{T-\tilde{\tau}}(X_{\tilde{\tau}}^{\leftarrow})\|_{L_2}^2 \, d\tilde{\tau} \, d\tau \\
&\leq 2 \int_{\tau_k}^{\tau_{k+1}} (\tau_{k+1} - \tilde{\tau}) g(T-\tilde{\tau})^2 \|\nabla s_{T-\tilde{\tau}}(X_{\tilde{\tau}}^{\leftarrow})\|_{L_2}^2 \, d\tilde{\tau} \\
&\leq 2h \int_{\tau_k}^{\tau_{k+1}} g(t)^2 \|\nabla s_t(X_\tau^{\leftarrow})\|_{L_2}^2 \, d\tau.
\end{aligned} \tag{51}$$

By $X_{\tilde{t}} \overset{d.}{=} X_{\tilde{\tau}}^{\leftarrow}$, it holds that $\|\nabla s_t(X_\tau^{\leftarrow})\|_{L_2} = \|\nabla s_t(X_t)\|_{L_2}$ and we denote this quantity by $\|\nabla s_t\|_{L_2}$. Recall (8), where

$$\dot{\mathcal{J}}_t = 2f(t)\mathcal{J}_t - g(t)\|\nabla s_t\|_{L_2}^2.$$

Substituting $g(t)\|\nabla s_t\|_{L_2}^2 = 2f(t)\mathcal{J}_t - \dot{\mathcal{J}}_t$ into the integrated bound (51), we obtain

$$\begin{aligned}
\int_{\tau_k}^{\tau_{k+1}} g(t) \left\| s_t(X_\tau^{\leftarrow}) - s_{T-\tau_k}(X_{\tau_k}^{\leftarrow}) \right\|_{L_2}^2 d\tau &\lesssim h \int_{\tau_k}^{\tau_{k+1}} g(t) \left(2f(t)\mathcal{J}_t - \dot{\mathcal{J}}_t \right) d\tau \\
&= h \int_{\tau_k}^{\tau_{k+1}} \frac{\left(2f(t)\mathcal{J}_t - \dot{\mathcal{J}}_t \right)^2}{\|\nabla s_t\|_{L_2}^2} d\tau \\
&\leq hd \int_{\tau_k}^{\tau_{k+1}} \left(2f(t) - \frac{\dot{\mathcal{J}}_t}{\mathcal{J}_t} \right)^2 d\tau.
\end{aligned}$$

Summing over all k, we obtain

$$\sum_{k=0}^{n-1} \int_{\tau_k}^{\tau_{k+1}} g(t) \left\| s_t(X_\tau^{\leftarrow}) - s_{T-\tau_k}(X_{\tau_k}^{\leftarrow}) \right\|_{L_2}^2 d\tau \lesssim hd \int_0^T \left(2f(t) - \frac{\dot{\mathcal{J}}_t}{\mathcal{J}_t} \right)^2 dt$$

D.2 Discretization Step Size

Proposition 1. *Under Assumption 1, it holds for any $(f, g) \in \Theta$ that any h satisfying*

$$h \leq \frac{g_{\min}}{f_{\max}^2} \min_{t \in [0,T]} \frac{1}{\alpha_t^2 \|X_\star\|_{L_2}^2 / d + \sigma_t^2} \tag{52}$$

necessarily satisfies for any $\{\tau_k\}_{k=0}^n$ with $h = \tau_{k+1} - \tau_k$ and $\tau_n = T$ the inequality (50)*.*

Proof. Recall the inequality (50):

$$h \leq \min_k \min_{\tau \in [\tau_k, \tau_{k+1}]} \frac{\int_{\tau_k}^{\tau} g(\tilde{t}) \|\nabla s_{\tilde{t}}(X_{\tilde{\tau}}^{\leftarrow})\|_{L_2}^2 d\tilde{\tau}}{\int_{\tau_k}^{\tau} f(\tilde{t})^2 \mathcal{J}_{\tilde{t}} d\tilde{\tau}}$$

We obtain a lower bound on

$$\frac{\int_{\tau_k}^{\tau} g(\tilde{t}) \|\nabla s_{\tilde{t}}(X_{\tilde{\tau}}^{\leftarrow})\|_{L_2}^2 d\tilde{\tau}}{\int_{\tau_k}^{\tau} f(\tilde{t})^2 \mathcal{J}_{\tilde{t}} d\tilde{\tau}},$$

such that any h chosen to be smaller than the lower bound necessarily satisfies (50).

By Lemma 4, it holds that $\|\nabla s_t\|_{L_2} \geq \frac{\mathcal{J}_t}{\sqrt{d}}$, and we have

$$\begin{aligned}\frac{\int_{\tau_k}^{\tau} g(\tilde{t})\|\nabla s_{\tilde{t}}\|_{L_2}^2 d\tilde{\tau}}{\int_{\tau_k}^{\tau} f(\tilde{t})^2 \mathcal{J}_{\tilde{t}} d\tilde{\tau}} &\geq \frac{1}{d}\frac{\int_{\tau_k}^{\tau} g(\tilde{t})\|\nabla s_{\tilde{t}}\|_{L_2}\mathcal{J}_{\tilde{t}} d\tilde{\tau}}{\int_{\tau_k}^{\tau} f(\tilde{t})^2\|\nabla s_{\tilde{t}}\|_{L_2} d\tilde{\tau}} \\ &\geq \left(\frac{\min_{\tilde{t}\in[\tau_k,\tau]}\mathcal{J}_{\tilde{t}}}{d}\right)\frac{\int_{\tau_k}^{\tau} g(\tilde{t})\|\nabla s_{\tilde{t}}\|_{L_2} d\tilde{\tau}}{\int_{\tau_k}^{\tau} f(\tilde{t})^2\|\nabla s_{\tilde{t}}\|_{L_2} d\tilde{\tau}}. \end{aligned} \tag{53}$$

The Cramér-Rao lower bound (CRLB) states that a random variable X whose density p is defined against the Lebesgue measure satisfies

$$\mathrm{Cov}(X) \succeq \left(\mathbb{E}\nabla \log p \nabla \log p^{\dagger}\right)^{-1}.$$

Let $A \in \mathbb{R}^{d\times d}$ be a positive definite matrix with eigenvalues $\{\lambda_i\}$. By the arithmetic-harmonic mean (AM-HM) inequality,

$$\frac{1}{d}\sum_{i=1}^{d}\lambda_i \geq \frac{d}{\sum_{i=1}^{d}\frac{1}{\lambda_i}}.$$

Rearranging, we obtain that $(\mathrm{Tr}A)(\mathrm{Tr}A^{-1}) \geq d^2$. Using the above property, we have the inequality

$$\|X\|_{L_2}^2 \mathcal{J}(p) \geq \mathrm{TrCov}(X)\mathrm{Tr}\mathbb{E}\nabla \log p \nabla \log p^{\dagger} \geq d^2.$$

Rearranging, we obtain the following form of the CRLB:

$$\mathcal{J}(p) \geq \frac{d^2}{\|X\|_{L_2}^2}. \tag{54}$$

Substituting for the denominator in (53), we obtain

$$\begin{aligned}\frac{\int_{\tau_k}^{\tau} g(\tilde{t})\|\nabla s_{\tilde{t}}\|_{L_2}^2 d\tilde{\tau}}{\int_{\tau_k}^{\tau} f(\tilde{t})^2 \mathcal{J}_{\tilde{t}} d\tilde{\tau}} &\geq \left(\min_{\tilde{t}\in[\tau_k,\tau]}\frac{d}{\|X_t\|_{L_2}^2}\right)\frac{\int_{\tau_k}^{\tau} g(\tilde{t})\|\nabla s_{\tilde{t}}\|_{L_2} d\tilde{\tau}}{\int_{\tau_k}^{\tau} f(\tilde{t})^2\|\nabla s_{\tilde{t}}\|_{L_2} d\tilde{\tau}} \\ &\geq \frac{g_{\min}}{f_{\max}^2}\min_{\tilde{t}\in[\tau_k,\tau]}\frac{d}{\|X_t\|_{L_2}^2}.\end{aligned}$$

Minimizing both sides over k and τ, we obtain that a sufficient condition for h to satisfy (50) is

$$h \leq \frac{g_{\min}}{f_{\max}^2}\min_{t\in[0,T]}\frac{d}{\|X_t\|_{L_2}^2}.$$

□

D.3 Solving the Euler-Lagrange Equation

The discretization error bound in Theorem 1 is minimized over g. Under a change of variables $u(t) = \log \mathcal{J}_t$, the second term's integral in (9) is given by

$$\int_0^T \left[2f(t) - \dot{u}(t)\right]^2 dt.$$

To find the optimal g^*, we first solve for the optimal trajectory:

$$\min_{u(t)} \int_0^T \left(2f(t) - \dot{u}(t)\right)^2 dt, \quad u(0) = \log \mathcal{J}_\star, \quad u(T) = \log \mathcal{J}_T. \tag{55}$$

The integrand $\mathcal{L}(t, u, \dot{u}) = (2f(t) - \dot{u})^2$ is jointly convex in $(u, \dot{u})$, and the optimality condition is given by the Euler–Lagrange equation

$$\frac{\partial \mathcal{L}}{\partial u} - \frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \dot{u}}\right) = 0.$$

Evaluating $\partial_u \mathcal{L} = 0, \partial_{\dot{u}} \mathcal{L} = -2(2f(t) - \dot{u}) = 2\dot{u} - 4f(t)$, we obtain the optimality condition

$$\frac{d}{dt}(\dot{u}_t - 2f(t)) = 0.$$

In other words, the optimality condition is given by (25), which we re-state:

$$2f(t) - \frac{\dot{\mathcal{J}}_t}{\mathcal{J}_t} = \lambda,$$

where the value of λ is determined by the boundary condition

$$\lambda T = 2\int_0^T f(t)dt + \log \mathcal{J}_\star - \log \mathcal{J}_T. \tag{56}$$

The above proof delineates an optimal trajectory $(\mathcal{J}_t^*)_{t\in[0,T]}$. We substitute $\int_0^T f(t)dt = -\log \alpha_T$ for (56), which yields

$$\lambda = \frac{1}{T}\log \frac{\mathcal{J}_\star}{\alpha_T^2 \mathcal{J}_T}. \tag{57}$$

Using $\mathcal{J}_T \le d/\sigma_T^2$ from (41) and the choice $\alpha_T^2/\sigma_T^2 = d/(\|X_\star\|_{L_2}^2 n)$, we obtain

$$\lambda \ge \frac{1}{T}\log \frac{\mathcal{J}_\star \sigma_T^2}{d\alpha_T^2} = \frac{1}{T}\log \frac{\mathcal{J}_\star \|X_\star\|_{L_2}^2 n}{d^2}, \tag{58}$$

which implies $\lambda \ge T^{-1}\log n$ by the CRLB (54). Hence, $f_{\max} \le (2T)^{-1}\log(n) \le \lambda/2$.

From (47), the optimal trajectory $\mathcal{J}_t^*$ defines a corresponding trajectory $\mathcal{J}(p_t^Y) = \alpha_t^2 \mathcal{J}_t^*$. Since α_t is a deterministic function of a fixed schedule $f(t)$, the optimal target $\mathcal{J}_t^*$ uniquely maps (as shown later) to a continuous variance trajectory V_t^*. The noise schedule is then uniquely defined by differentiating this variance:

$$g^*(t) = \alpha_t^2 \frac{d}{dt} V_t^*.$$

From the Fisher ODE (8) and the optimality condition (25), we obtain the explicit expression for g^*:

$$g^*(t) = \frac{2f(t)\mathcal{J}_t^* - \dot{\mathcal{J}}_t^*}{\|\nabla s_t\|_{L_2}^2} = \frac{\lambda \mathcal{J}_t^*}{\|\nabla s_t\|_{L_2}^2}. \tag{59}$$

Using $2f_{\max} \le \lambda$, we obtain $g^*(t)\|\nabla s_t\|_{L_2}^2 = \lambda \mathcal{J}_t^*$ is nonincreasing.

We next verify that $h = T/n$ is admissible. By the assumption on $f_{\max}$ and the lower bound (58) on λ, we obtain

$$h f_{\max}^2 \le \frac{\log^2 n}{4Tn} \le \frac{\log n}{T} \le \lambda.$$

Combining with (59), we obtain

$$h f(t)^2 \mathcal{J}_t^* \le h f_{\max}^2 \mathcal{J}_t^* \le g^*(t)\|\nabla s_t\|_{L_2}^2 \Rightarrow h \le \frac{g^*(t)\|\nabla s_t\|_{L_2}^2}{f(t)^2 \mathcal{J}_t^*}.$$

Therefore, $h = T/n$ satisfies (50). When $f(t) = 0$, the inequality for $h f(t)^2 \mathcal{J}_t^*$ implies that $h = T/n$ satisfies (50).

Because (51) was established for h sufficiently small given g, we next show that the bound holds for g^* in (59) which may depend on n. Using $f_{\max}^2 \le \lambda/h$ and that $\lambda \mathcal{J}_t^*$ is nonincreasing in t, we obtain

$$\begin{aligned}\left\| s_{T-\tau}(X_\tau^{\leftarrow}) - s_{T-\tau_k}(X_{\tau_k}^{\leftarrow}) \right\|_{L_2}^2 &\le 4\int_{\tau_k}^{\tau} g^*(T-\tilde{\tau}) \left\|\nabla s_{T-\tilde{\tau}}\right\|_{L_2}^2 d\tilde{\tau} \\ &\le 4h\, g^*(T-\tau) \left\|\nabla s_{T-\tau}\right\|_{L_2}^2.\end{aligned}$$

Multiplying by $g^*(T-\tau)$ and using $(\mathcal{J}_t^*)^2 \le d\|\nabla s_t\|_{L_2}^2$, we obtain

$$\sum_k \int_{\tau_k}^{\tau_{k+1}} g^*(T-\tau) \left\| s_{T-\tau}(X_\tau^{\leftarrow}) - s_{T-\tau_k}(X_{\tau_k}^{\leftarrow}) \right\|_{L_2}^2 d\tau \le 4h\lambda^2 \int_0^T \frac{(\mathcal{J}_t^*)^2}{\|\nabla s_t\|_{L_2}^2} dt \le 4hd\lambda^2 T.$$

Therefore, we have the bound

$$\mathrm{KL}(p_\star \| \hat{p}_\star) \le \frac{\alpha_T^2}{2\sigma_T^2} \|X_\star\|_{L_2(p_\star)}^2 + 2hd\lambda^2 T. \tag{60}$$

Substituting $h = T/n$, the lower bound (41) on $\mathcal{J}_T$, and (57) for λ, we obtain that the discretization error bound in (60) is

$$hd\lambda^2 T \lesssim \frac{d}{n} \left(\log \frac{\mathcal{J}_\star \sigma_T^2}{d\alpha_T^2} \right)^2 = \frac{d}{n} \log^2 \left(\frac{\mathcal{J}_\star}{d} \cdot \frac{\|X_\star\|_{L_2}^2}{d} n \right),$$

where the last step uses the choice of $\alpha_T^2/\sigma_T^2 = d/(\|X_\star\|_{L_2}^2 n)$. Adding the initialization error back, we obtain

$$\mathrm{KL}(p_\star \| \hat{p}_\star) \lesssim \frac{d}{2n} + \frac{d}{n} \log^2 \left(\frac{\mathcal{J}_\star}{d} \cdot \frac{\|X_\star\|_{L_2}^2}{d} n \right). \tag{61}$$

Existence: While the Euler-Lagrange optimization was solved with respect to $u_t = \log \mathcal{J}_t$, we must verify that the optimal trajectory $\mathcal{J}_t^*$ corresponds to a valid noise schedule $g^* \in \mathcal{G}$. We establish this mapping using the process $Y_t = X_t/\alpha_t$ introduced in Section C.5.

We show that there exists a bijection $\mathcal{J}_t^* \mapsto g^*(t)$, which implies that the trajectory $(\mathcal{J}_t^*)$ can be controlled by a choice of g^*. Using the relation $V_t := \sigma_t^2/\alpha_t^2 = \int_0^t g(\tilde{t})/\alpha_{\tilde{t}}^2 d\tilde{t}$ which implies $\dot{V}_t = g(t)/\alpha_t^2$, the identity (46) can be written as

$$\frac{d}{dV_t} \mathcal{J}(p_t^Y) = -\|\nabla^2 \log p_t^Y\|_{L_2}^2.$$

Because the RHS is strictly negative, $\mathcal{J}(p_t^Y)$ is strictly decreasing as a function of the variance V_t. Consequently, the map $V_t \mapsto \mathcal{J}(p_t^Y)$ is a bijection.

Finally, we verify that $g^*(t) > 0$ so that it satisfies the constraints in (3). By Lemma 1, $\|\nabla s_t\|_{L_2}^2 < \infty$ for all $t \in [0, T]$. Since $\mathcal{J}_t > 0$ (e.g., via the CRLB (54)), we conclude from (59) that $g^*(t) > 0$.

This concludes the proof of Theorem 2.

E Proofs of Theorem 3 and Corollary 1 (Section 3.2)

E.1 Higher Order Score and Fisher Information

Under the strong log-concavity (SLC) condition in Assumption 2(i), we have the reverse inequality of Lemma 4,

Lemma 5. *Under Assumption 2(i), it holds that*

$$\|\nabla s_\star\|_{L_2}^2 \le \frac{1}{d} \left(\frac{M_\star}{m_\star} \right)^2 \mathcal{J}_\star^2.$$

Recall the integration by parts formula $\mathbb{E}\|s_\star\|^2 = -\mathbb{E}\mathrm{Tr}\nabla s_\star$ By the log-concavity assumption, it holds that $-\mathrm{Tr}\mathbb{E}\nabla s_\star \ge m_\star d$. Moreover, the upper bound implies $\|\nabla s_\star\|_{L_2}^2 = \mathbb{E}\mathrm{Tr}(\nabla s_\star)^2 \le M_\star^2 d$. Combining, we obtain

$$\frac{\|\nabla s_\star\|_{L_2}^2}{\mathcal{J}_\star^2} = \frac{\mathbb{E}\mathrm{Tr}(\nabla s_\star)^2}{(\mathrm{Tr}\mathbb{E}\nabla s_\star)^2} \le \frac{M_\star^2 d}{(m_\star d)^2} = \frac{1}{d} \left(\frac{M_\star}{m_\star} \right)^2.$$

E.2 PRÉKOPA–LEINDLER INEQUALITY

The constants (m_t, M_t) of the SLC condition in Assumption 2(i) are evaluated along the SDE (1). For the following, denote by p_Z the standard Gaussian density (of the random variable Z).

Lemma 6. *Suppose the density $p_\star$ satisfies* (13)*. For any positive constants (α_t, σ_t), it holds that the density p_t of $X_t = \alpha_t X_\star + \sigma_t Z$ satisfies*

$$\mathrm{PS}\left(\frac{m(p_\star)}{\alpha_t^2}, \frac{m(p_Z)}{\sigma_t^2}\right) I \preceq -\nabla^2 \log p_t \preceq \mathrm{PS}\left(\frac{M(p_\star)}{\alpha_t^2}, \frac{M(p_Z)}{\sigma_t^2}\right) I. \tag{62}$$

Proof. Let $p_{\star|t}(x_\star \mid x)$ be the conditional density of $X_\star$ given $X_t = x$, where

$$p_{\star|t}(x_\star \mid x) = p_\star(x_\star) \exp\left(-\frac{\|x - \alpha_t x_\star\|^2}{2\sigma_t^2}\right).$$

By Assumption 2(i),

$$(m_\star + \alpha_t^2/\sigma_t^2) I \preceq -\nabla^2 \log p_{\star|t}(x_\star \mid x) \preceq (M_\star + \alpha_t^2/\sigma_t^2) I.$$

The Brascamp–Lieb (Brascamp & Lieb, 1976) and Cramér–Rao inequalities therefore give

$$\left(M_\star + \frac{\alpha_t^2}{\sigma_t^2}\right)^{-1} I \preceq \mathrm{Cov}(X_\star \mid X_t = x) \preceq \left(m_\star + \frac{\alpha_t^2}{\sigma_t^2}\right)^{-1} I.$$

Using Tweedie's formula

$$-\nabla^2 \log p_t(x) = \frac{1}{\sigma_t^2} I - \frac{\alpha_t^2}{\sigma_t^4} \mathrm{Cov}(X_\star \mid X_t = x),$$

we obtain

$$\left(\frac{\alpha_t^2}{m_\star} + \sigma_t^2\right)^{-1} I \preceq -\nabla^2 \log p_t(x) \preceq \left(\frac{\alpha_t^2}{M_\star} + \sigma_t^2\right)^{-1} I.$$

□

E.3 EQUIVALENCE BETWEEN SCORE FUNCTION'S JACOBIAN AND FISHER INFORMATION

Lemma 7. *Suppose Assumptions 1 and 2 hold. Then*

$$\frac{d}{\kappa^2} \le \frac{\mathcal{J}_t^2}{\|\nabla s_t\|_{L_2}^2} \le d, \quad \forall t \in [0, T], \tag{63}$$

where the constant κ is defined as follows:

- *When $p_\star$ is SLC under Assumption 2(i): $\kappa = \frac{M_\star}{m_\star}$;*
- *When $p_\star$ is a GMM under Assumption 2(ii):*

$$\kappa = \left(1 + \frac{\Delta_\mu^2}{4\nu^2}\right)^{1/2} \left(1 + \frac{1}{\nu^2} \cdot \frac{1}{d} \sum_{i=1}^{L} \pi_i \|\mu_i - \bar{\mu}\|^2\right), \tag{64}$$

where $\bar{\mu} = \sum_{i=1}^{L} \pi_i \mu_i$ and $\Delta_\mu = \max_{1 \le i, \ell \le L} \|\mu_i - \mu_\ell\|$.

In particular, if $\max_{1 \le i, \ell \le L} \|\mu_i - \bar{\mu}\| \le R_\mu$ for some $R_\mu > 0$, then

$$\kappa \le \left(1 + \frac{R_\mu^2}{\nu^2}\right) \left(1 + \frac{1}{d} \frac{R_\mu^2}{\nu^2}\right). \tag{65}$$

Proof. The upper bound $\frac{\mathcal{J}_t^2}{\|\nabla s_t\|_{L_2}^2} \le d$ directly follows from Lemma 4, and it does not require Assumption 2.

In the remainder of the proof, we focus on proving the lower bound $\frac{\mathcal{J}_t^2}{\|\nabla s_t\|_{L_2}^2} \geq \frac{d}{\kappa^2}$, or equivalently, $\frac{\|\nabla s_t\|_{L_2}^2}{\mathcal{J}_t^2} \leq \frac{\kappa^2}{d}$.

Log-Concave Distributions. Because p_t remains log-concave for all t, its Hessian can be bounded in the positive-definite (Löwner) order $\prec$: $m_t I \preceq -\nabla^2 \log p_t \preceq M_t I$. We define $\kappa_t = M_t/m_t$ as the condition number of this log-concave density. Adding Gaussian noise decreases the condition number and hence, the time-varying condition number is bounded by that associated with the target density: $\kappa = \sup_t M_t/m_t = M_\star/m_\star$.

The above remark is described formally. By Lemma 5, there exists some $(m_\star, M_\star)$ such that

$$\|\nabla s_\star\|_{L_2}^2 \leq \frac{1}{d}\left(\frac{M_\star}{m_\star}\right)^2 \mathcal{J}_\star^2.$$

By Lemma 6, for any $t \in [0,T]$, the condition number of the smoothed density p_t can be bounded pointwise. Using $m(\gamma^d) = M(\gamma^d) = 1$ and m_t, M_t defined the lower and upper bounds of (62), the condition number $\kappa_t = M_t/m_t$ evaluates to

$$\kappa_t \coloneqq \frac{M_t}{m_t} = \left(\frac{M_\star}{m_\star}\right)\frac{1 + m_\star \sigma_t^2/\alpha_t^2}{1 + M_\star \sigma_t^2/\alpha_t^2}.$$

The condition number κ_t is bounded uniformly over time (over all values of SNR α_t^2/σ_t^2), and consequently there exists a global constant $\kappa = M_\star/m_\star$ such that:

$$\|\nabla s_t\|_{L_2}^2 \leq \kappa^2 \frac{\mathcal{J}_t^2}{d}, \quad \forall t \in [0,T].$$

Gaussian Mixture Models. Recall that we assume that the data distribution is a mixture of L Gaussian components. Specifically, $p_\star(x) = \sum_{i=1}^L \pi_i \varphi(x; \mu_i, \nu^2 I)$, where $\varphi(x; \mu_i, \nu^2 I)$ is the density of $\mathcal{N}(\mu_i, \nu^2 I)$.

From the solution of the forward process $X_t \overset{d}{=} \alpha_t X_\star + \sigma_t Z$, it holds that the law of X_t is a GMM with density

$$p_t(x) = \sum_{i=1}^L \pi_i \varphi(x; \alpha_t \mu_i, v_t^2 I), \quad v_t^2 = \alpha_t^2 \nu^2 + \sigma_t^2. \tag{66}$$

Define the random variable M which takes the value μ_i with probability π_i. Let $\tilde{Z}$ be a standard Gaussian vector independent of Z and M, so that

$$X_t \overset{d}{=} \alpha_t M + v_t \tilde{Z}. \tag{67}$$

Observe that $(\alpha_t M \mid X_t = x)$ takes values in $\{\alpha_t \mu_1, \cdots, \alpha_t \mu_L\}$, where $(\alpha_t M \mid X_t = x)$ is $\alpha_t \mu_i$ with probability

$$\frac{\pi_i \varphi(x; \alpha_t \mu_i, v_t^2 I)}{\sum_{j=1}^L \pi_j \varphi(x; \alpha_t \mu_j, v_t^2 I)}. \tag{68}$$

The score function for GMMs can be evaluated using Tweedie's formula as in (Gatmiry et al., 2024, Section 3.2), where

$$\nabla \log p_{X_t}(x) = \frac{1}{v_t^2}\left(\mathbb{E}[\alpha_t M \mid X_t = x] - x\right).$$

The Jacobian of the conditional expectation with respect to x is then

$$\nabla \mathbb{E}[\alpha_t M \mid X_t = x] = \frac{1}{v_t^2}\mathrm{Cov}(\alpha_t M \mid X_t = x) = \frac{\alpha_t^2}{v_t^2}\mathrm{Cov}(M \mid X_t = x).$$

Therefore, the second order Tweedie's formula evaluates to

$$-\nabla^2 \log p_t(x) = \frac{1}{v_t^2} I - \frac{\alpha_t^2}{v_t^4}\mathrm{Cov}(M \mid X_t = x)$$

and we use integration by parts as in the proof of Lemma 4 to obtain

$$\mathcal{J}_t = -\mathbb{E}\mathrm{Tr}\nabla^2 \log p_t(X_t) = \frac{d}{v_t^2} - \frac{\alpha_t^2}{v_t^4}\mathbb{E}\mathrm{TrCov}(M \mid X_t).$$

From $\mathcal{J}_t \geq 0$, we have

$$\frac{\alpha_t^2}{v_t^2}\mathbb{E}\mathrm{TrCov}(M \mid X_t) \leq d. \tag{69}$$

Using the integration by parts $\mathbb{E}\|s_t\|^2 = -\mathbb{E}\mathrm{Tr}\nabla s_t$, we obtain

$$\begin{aligned}
\|\nabla s_t\|_{L_2(p_t)}^2 &= \mathbb{E}\mathrm{Tr}\left(-\nabla^2 \log p_t\right)^2 \\
&= \frac{d}{v_t^4} - \frac{2\alpha_t^2}{v_t^6}\mathbb{E}\mathrm{TrCov}(M \mid X_t) + \frac{\alpha_t^4}{v_t^8}\mathbb{E}\mathrm{Tr}\left(\mathrm{Cov}(M \mid X_t)\right)^2 \\
&\leq \frac{d}{v_t^4} + \frac{\alpha_t^4}{v_t^8}\mathbb{E}\mathrm{Tr}\left(\mathrm{Cov}(M \mid X_t)\right)^2.
\end{aligned} \tag{70}$$

By the semi-positive definiteness of $\mathrm{Cov}(M \mid X_t = x)$, there exists some $\lambda_{\max}$ such that $\mathrm{Cov}(M \mid X_t = x) \preceq \lambda_{\max} I$. Multiplying both sides by $\mathrm{Cov}(M \mid X_t = x)$ and taking the trace, we obtain

$$\mathrm{Tr}\left(\mathrm{Cov}(M \mid X_t = x)\right)^2 \leq \lambda_{\max}\mathrm{TrCov}(M \mid X_t = x).$$

which implies

$$\mathbb{E}\mathrm{Tr}\left(\mathrm{Cov}(M \mid X_t = x)\right)^2 \leq \lambda_{\max} d\frac{v_t^2}{\alpha_t^2}. \tag{71}$$

We next obtain an upper bound on $\lambda_{\max}$. Define $\bar{\mu}_x = \mathbb{E}[M \mid X_t = x]$. For any unit vector $u \in \mathbb{R}^d$,

$$\begin{aligned}
u^\dagger \mathrm{Cov}(M \mid X_t = x)u &= \sum_{i=1}^{L} \mathbb{P}\left(M = \mu_i \mid X_t = x\right) u^\dagger(\mu_i - \bar{\mu}_x)(\mu_i - \bar{\mu}_x)^\dagger u \\
&= \mathrm{Var}(u^\dagger M \mid X_t = x).
\end{aligned}$$

The random variable $u^\dagger M$ takes values in the set $\{u^\dagger\mu_1, u^\dagger\mu_2, \cdots, u^\dagger\mu_L\}$, and

$$|u^\dagger(\mu_i - \mu_\ell)| \leq \|u\|\|\mu_i - \mu_\ell\| \leq \Delta_\mu.$$

Using the definition $\lambda_{\max} = \sup_{\|u\|_2=1} u^\dagger \mathrm{Cov}(M \mid X_t = x)u$, we have

$$\lambda_{\max} = \max_{u:\|u\|_2=1} u^\dagger \mathrm{Cov}(M \mid X_t = x)u \leq \frac{1}{4}\Delta_\mu^2.$$

Substituting for (71), we obtain

$$\mathbb{E}\mathrm{Tr}\left(\mathrm{Cov}(M \mid X_t = x)\right)^2 \leq \frac{d}{4}\frac{v_t^2}{\alpha_t^2}\Delta_\mu^2.$$

Substituting for (70), we obtain

$$\|\nabla s_t\|_{L_2}^2 \leq \frac{d}{v_t^4} + \frac{\alpha_t^4}{v_t^8}\frac{d}{4}\frac{v_t^2}{\alpha_t^2}\Delta_\mu^2 = \frac{d}{v_t^4}\left(1 + \frac{1}{4}\frac{\alpha_t^2}{v_t^2}\Delta_\mu^2\right). \tag{72}$$

A lower bound on $\mathcal{J}_t$ is obtained using the CRLB (54):

$$\mathcal{J}_t \geq \frac{d^2}{\mathrm{TrCov}(X_t)} = \frac{d^2}{\mathrm{Tr}\left\{\alpha_t^2\mathrm{Cov}(M) + v_t^2 I\right\}}.$$

Substituting

$$\mathrm{Cov}(M) = \sum_{i=1}^{L} \pi_i(\mu_i - \bar{\mu})(\mu_i - \bar{\mu})^\dagger,$$

we obtain

$$\mathcal{J}_t \geq \frac{d^2}{\alpha_t^2 \sum_{i=1}^L \pi_i \|\mu_i - \bar{\mu}\|^2 + v_t^2 d}. \tag{73}$$

Combining (72) and (73), we obtain

$$\begin{aligned}\frac{d\|\nabla s_t\|_{L_2}^2}{\mathcal{J}_t^2} &\leq \frac{d^2}{v_t^4}\left(1 + \frac{1}{4}\frac{\alpha_t^2}{v_t^2}\Delta_\mu^2\right) \cdot \left(\frac{\alpha_t^2 \sum_{i=1}^L \pi_i \|\mu_i - \bar{\mu}\|^2 + v_t^2 d}{d^2}\right)^2 \\ &= \left(1 + \frac{1}{4}\frac{\alpha_t^2}{v_t^2}\Delta_\mu^2\right)\left(1 + \frac{\alpha_t^2}{v_t^2}\frac{1}{d}\sum_{i=1}^L \pi_i \|\mu_i - \bar{\mu}\|^2\right)^2.\end{aligned}$$

The bound is further simplified by using $v_t^2 \geq \alpha_t^2 \nu^2$, which yields

$$\frac{d\|\nabla s_t\|_{L_2}^2}{\mathcal{J}_t^2} \leq \left(1 + \frac{1}{4\nu^2}\Delta_\mu^2\right)\left(1 + \frac{1}{\nu^2}\frac{1}{d}\sum_{i=1}^L \pi_i \|\mu_i - \bar{\mu}\|^2\right)^2.$$

The upper bound in (65) is obtained by observing that $\max_{1\leq i,\ell\leq L}\|\mu_i - \bar{\mu}\| \leq R_\mu$ implies $\Delta_\mu^2 \leq 4R_\mu^2$.

□

E.4 Proof of Theorem 3

Let $g = g_q^\circ$. By Lemmas 3 and 7,

$$h\int_0^T g(t)^2 \|\nabla s_t\|_{L_2}^2 \, dt \leq \frac{h\kappa^2}{d}\int_0^T [g(t)\mathcal{J}_t]^2 dt. \tag{74}$$

The upper bound on $\mathcal{J}_t^2/\|\nabla s_t\|_{L_2}^2$ in Lemma 7 also gives

$$\dot{\mathcal{J}}_t \leq 2f(t)\mathcal{J}_t - \frac{g(t)}{d}\mathcal{J}_t^2. \tag{75}$$

Since $g'(t)/g(t) = q(t) - 2f(t)$, (75) implies

$$\frac{d}{dt}\big(g(t)\mathcal{J}_t\big) \leq q(t)g(t)\mathcal{J}_t - \frac{\big(g(t)\mathcal{J}_t\big)^2}{d}.$$

Using $\sup_{t\in[0,T]} q(t) \leq C\log(1+n)$, we obtain

$$\frac{d}{dt}\big(g(t)\mathcal{J}_t\big) \leq C\log(1+n)g(t)\mathcal{J}_t - \frac{\big(g(t)\mathcal{J}_t\big)^2}{d}.$$

Whenever $g(t)\mathcal{J}_t \geq Cd\log(1+n)$, the upper bound is negative. Since $g_0\mathcal{J}_\star \leq Cd\log(1+n)$, we have

$$g(t)\mathcal{J}_t \leq Cd\log(1+n), \qquad t \in [0,T].$$

Since $q(t) \geq 2f(t)$, (14) implies that g is nondecreasing. For each reverse-time step, (49) and $g(T-\tau) \leq g(T-\tilde{\tau})$ for $\tilde{\tau} \leq \tau$ give

$$\begin{aligned}&\int_{\tau_k}^{\tau_{k+1}} g(T-\tau)\left\|s_{T-\tau}(X_\tau^{\leftarrow}) - s_{T-\tau_k}(X_{\tau_k}^{\leftarrow})\right\|_{L_2}^2 d\tau \\ &\lesssim h\int_{\tau_k}^{\tau_{k+1}}\int_{\tau_k}^{\tau} g(T-\tilde{\tau})f(T-\tilde{\tau})^2\mathcal{J}_{T-\tilde{\tau}}\, d\tilde{\tau}\, d\tau + \int_{\tau_k}^{\tau_{k+1}}\int_{\tau_k}^{\tau} g(T-\tilde{\tau})^2 \left\|\nabla s_{T-\tilde{\tau}}\right\|_{L_2}^2 d\tilde{\tau}\, d\tau \\ &\leq h^2\int_{\tau_k}^{\tau_{k+1}} g(T-\tilde{\tau})f(T-\tilde{\tau})^2\mathcal{J}_{T-\tilde{\tau}}\, d\tilde{\tau} + h\int_{\tau_k}^{\tau_{k+1}} g(T-\tilde{\tau})^2\left\|\nabla s_{T-\tilde{\tau}}\right\|_{L_2}^2 d\tilde{\tau}.\end{aligned}$$

Summing over k and changing variables to forward time bounds the discretization error in (17) by

$$h^2\int_0^T g(t)f(t)^2\mathcal{J}_t\, dt + h\int_0^T g(t)^2\|\nabla s_t\|_{L_2}^2\, dt.$$

The first term is bounded using $f(t) \le q(t)/2 \le C\log(1+n)/2$ and $g(t)\mathcal{J}_t \le Cd\log(1+n)$. For the second term, Lemma 7 gives

$$h\int_0^T g(t)^2 \|\nabla s_t\|_{L_2}^2\, dt \le \frac{h\kappa^2}{d}\int_0^T [g(t)\mathcal{J}_t]^2 dt.$$

Thus, with $h = T/n$, the discretization error is bounded by a constant multiple of

$$\frac{\kappa^2 C^2 dT^2 \log^2(1+n)}{n} + \frac{C^3 dT^3 \log^3(1+n)}{n^2}.$$

For the initialization error, (14) and the second condition $\int_0^T \exp\left(\int_0^t q(\tilde{t})\, d\tilde{t}\right) dt \ge \frac{c}{g_0}\frac{\|X_\star\|_{L_2}^2}{d}\frac{n}{\log(1+n)}$ assumed in Theorem 3 give

$$\frac{\sigma_T^2}{\alpha_T^2} = g_0\int_0^T \exp\left(\int_0^t q(\tilde{t})\, d\tilde{t}\right) dt \ge c\frac{\|X_\star\|_{L_2}^2 n}{d\log(1+n)}.$$

Consequently, $\alpha_T^2\|X_\star\|_{L_2}^2/\sigma_T^2 \le d\log(1+n)/(cn)$. Combining these bounds with (17), we obtain

$$\mathrm{KL}(p_\star\|\hat{p}_\star) \lesssim \frac{d}{n}\left[\frac{\log(1+n)}{c} + \kappa^2 C^2 T^2 \log^2(1+n) + \frac{C^3 T^3 \log^3(1+n)}{n}\right].$$

E.5 Proof of Corollary 1

Differentiating (14) and substituting $f(t) = \theta g_q^\circ(t) + \omega$ gives

$$\frac{d}{dt} g_q^\circ(t) = (q(t) - 2\omega) g_q^\circ(t) - 2\theta g_q^\circ(t)^2.$$

Consequently,

$$\frac{d}{dt}\left[\frac{\exp\left(\int_0^t (q(\tilde{t}) - 2\omega)\, d\tilde{t}\right)}{g_q^\circ(t)}\right] = 2\theta \exp\left(\int_0^t (q(\tilde{t}) - 2\omega)\, d\tilde{t}\right).$$

Integrating from 0 to t and using $g_q^\circ(0) = g_0$ yields (16).

E.6 Verifying Theorem 3 for VP Schedules

VP-Sigmoid. For $q(t) = \bar{q} = 2\log(1+n)/T$ and $g_0 > 0$ independent of n, we have $g_0 < \bar{q}$ for sufficiently large n. The schedule in (16) then satisfies $2f(t) = g_q^\circ(t) < \bar{q} = q(t)$ for all $t \in [0, T]$, and

$$\sup_{t\in[0,T]} q(t) = \frac{2}{T}\log(1+n).$$

Moreover,

$$\int_0^T \exp\left(\int_0^t q(\tilde{t})\, d\tilde{t}\right) dt = \frac{e^{\bar{q}T} - 1}{\bar{q}} = \frac{Tn(n+2)}{2\log(1+n)}.$$

For sufficiently large n, this is larger than the lower bound required in Theorem 3, so the conditions on q in that theorem hold.

VP-linear. For $g_0 > 0$ independent of n, let

$$g_{\min} = g_0, \qquad g_{\max} = g_{\min} + \frac{2}{T}\log(1+n),$$

and define

$$g(t) = g_{\min} + \frac{g_{\max} - g_{\min}}{T}t, \qquad f(t) = \frac{g(t)}{2}, \qquad q(t) = g(t) + \frac{g_{\max} - g_{\min}}{Tg(t)}. \tag{76}$$

Then $g'(t)/g(t) = q(t) - 2f(t)$, and (14) recovers the VP-linear schedule. Since $g'(t) > 0$, we have $q(t) \geq g(t) = 2f(t)$, and

$$\sup_{t\in[0,T]} q(t) \leq g_{\max} + \frac{g_{\max} - g_{\min}}{T g_{\min}} = \mathcal{O}(\log(1+n)).$$

Moreover, $q(t) = g(t) + g'(t)/g(t)$ implies

$$\int_0^T \exp\left(\int_0^t q(\tilde{t})\, d\tilde{t}\right) dt = \frac{e^{\int_0^T g(t)\, dt} - 1}{g_{\min}} = \frac{e^{T g_{\min}}(1+n) - 1}{g_{\min}}.$$

For sufficiently large n, this is larger than the lower bound required in Theorem 3, so the conditions on q required to achieve the $\tilde{\mathcal{O}}(d/n)$ rate hold.

F Error Bound for the Constant Schedule

Proposition 2. *Suppose Assumptions 1 and 2 hold. For sufficiently small h, the VP-constant schedule (f_{const}, g_{const}) with $f_{const} = g_{const}/2$ achieves the error bound*

$$\mathrm{KL}(p_\star \| \hat{p}_\star) \lesssim \frac{\alpha_T^2}{2\sigma_T^2}\|X_\star\|_{L_2}^2 + hd\kappa^2 g_{const}\left[\frac{\mathcal{J}_\star}{d} + \log\left(1 + \frac{\mathcal{J}_\star}{d}(e^{g_{const}T} - 1)\right)\right]. \tag{77}$$

Proof. See Section F.1. □

As discussed earlier, the best bounds (Chen et al., 2023c; Conforti et al., 2025) available for VP-constant schedules grow linearly with $\mathcal{J}_\star/d$. We show that, under our analysis, a constant schedule suffers from the larger rate even when the constants are chosen to scale with problem constants as in Theorem 3. Let $T = 1$ and define $\mathcal{E} = \int_0^1 f(t)dt$. For the constant schedule, $g_{const} = 2\mathcal{E}$, and Proposition 2 gives

$$\mathrm{KL}(p_\star \| \hat{p}_\star) \lesssim \frac{\|X_\star\|_{L_2}^2}{2(e^{g_{const}} - 1)} + \frac{d\kappa^2}{n} g_{const}\left[\frac{\mathcal{J}_\star}{d} + \log\left(1 + \frac{\mathcal{J}_\star}{d}(e^{g_{const}} - 1)\right)\right] \tag{78}$$

The choice of g_{const}^* that minimizes the bound yields

$$\mathrm{KL}(p_\star \| \hat{p}_\star) \lesssim \frac{\kappa^2}{n}\left(d\log^2 n + \mathcal{J}_\star \log n\right). \tag{79}$$

Therefore, under our framework, the constant schedule cannot eliminate the linear dependence on $\mathcal{J}_\star/d$ even with a problem-dependent choice of g_{const}.

F.1 Proof of Proposition 2

Under the VP constraint, (75) gives

$$\dot{\mathcal{J}}_t \leq g_{const}\mathcal{J}_t\left(1 - \frac{\mathcal{J}_t}{d}\right).$$

Solving this differential inequality yields

$$\mathcal{J}_t \leq \frac{d e^{g_{const}t}}{e^{g_{const}t} - (1 - d/\mathcal{J}_\star)}.$$

Therefore,

$$\int_0^T \mathcal{J}_t dt \leq \frac{d}{g_{const}} \log\left(1 + \frac{\mathcal{J}_\star}{d}(e^{g_{const}T} - 1)\right). \tag{80}$$

Applying the integration-by-parts argument for the constant schedule gives

$$\int_0^T (g_{const}\mathcal{J}_t)^2 dt \leq d g_{const}\mathcal{J}_\star + d g_{const}^2 \int_0^T \mathcal{J}_t dt. \tag{81}$$

Substituting (80), we obtain Proposition 2.

G SUPPLEMENT TO EXPERIMENTS (SECTION 5)

The time-varying function q is parametrized by a rational cubic Bernstein basis. We set $T = 1$ and use the schedule in Corollary 1. Let

$$r_{\ell,\epsilon}(t) = \frac{\sum_{j=0}^{3} e^{\ell_j} \binom{3}{j} t^j (1-t)^{3-j}}{(t+\epsilon)(1-t+\epsilon_{\text{term}})}, \qquad \psi(t) = \gamma \frac{r_{\ell,\epsilon}(t)}{\int_0^1 r_{\ell,\epsilon}(s)ds}, \tag{82}$$

where $\ell_0 = 0$ and $\epsilon_{\text{term}} = 10^{-6}$. Thus, ψ is positive and satisfies $\int_0^1 \psi(t)dt = \gamma$. For an NFE budget n, define

$$\lambda_n = \mathcal{W}(Kn), \qquad \mathcal{E}_n = \frac{\lambda_n}{2}, \qquad q_n(t) = \lambda_n \psi(t), \qquad \omega_n = \rho \mathcal{E}_n. \tag{83}$$

In the implementation, we set $K = \frac{2\mathcal{E}_{10}\exp(2\mathcal{E}_{10})}{10}$ which ensures that $\lambda_{10} = 2\mathcal{E}_{10}$. The initial diffusion coefficient is determined by the energy constraint $\int_0^1 f_n(t)dt = \mathcal{E}_n$:

$$g_{0,n} = \frac{\exp\big(2(\mathcal{E}_n - \omega_n)\big) - 1}{2\theta \int_0^1 \exp\left(\int_0^s (q_n(u) - 2\omega_n)du\right) ds}. \tag{84}$$

Substituting q_n, ω_n, and $g_{0,n}$ into (16) determines $(f_n, g^{\circ}_{q_n})$ for every n.

Hyperparameter selection. We selected a single parameter vector for each schedule by minimizing the mean log-FID over CIFAR-10 and ImageNet 64×64 at NFE 10 and 50. Each objective evaluation used 10,000 generated samples. The universal selection study used 32 trials per schedule in Optuna (Akiba et al., 2019), with scrambled Sobol points for initialization and CMA-ES for multidimensional search spaces. After selection, we fixed every schedule's parameters across all NFE budgets and evaluation datasets.

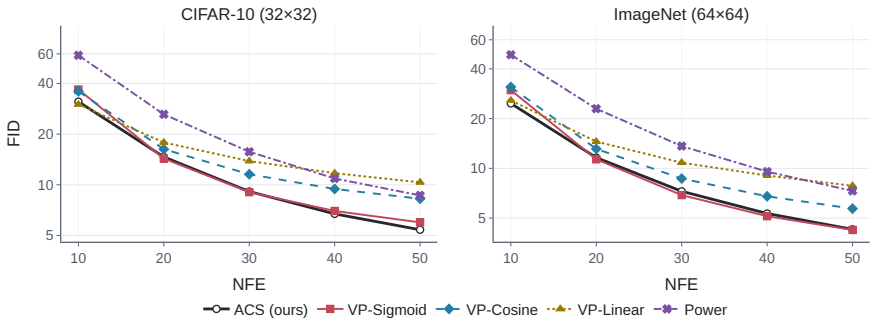

Figure 2: FID on 50,000 generated samples across inference budgets on CIFAR-10 and ImageNet 64×64, the datasets used for hyperparameter selection. Hyperparameters were selected using 10,000-sample FID at NFE 10 and 50 and then fixed across all inference budgets.

Generation and evaluation. All reverse processes were initialized with $X_0^{\leftarrow} \sim \mathcal{N}(0, \sigma_T^2 I)$. We used ancestral sampling (Ho et al., 2020) with pretrained EDM models for CIFAR-10, ImageNet 64×64, FFHQ 64×64, and AFHQv2 64×64 (Karras et al., 2022), and a pretrained EDM2-XS model (Karras et al., 2024) for ImageNet 512×512. We applied each schedule to the corresponding pretrained model through the statistical equivalence in Section B.1, without fine-tuning the model.

Figure 3 provides a qualitative comparison of samples generated for ImageNet 512×512. The visual quality of the baseline schedules varies across examples, whereas ACS consistently produces samples comparable in quality to the strongest baseline in each example.

Figure 4 plot $(f(t), g(t))$ and (α_t, σ_t) for ACS, VP-Linear, VP-Cosine, and VP-Sigmoid using the parameters selected for the generation experiments. We show the schedules at NFEs 10 and 50 because the ACS parameters depend on the NFE budget.

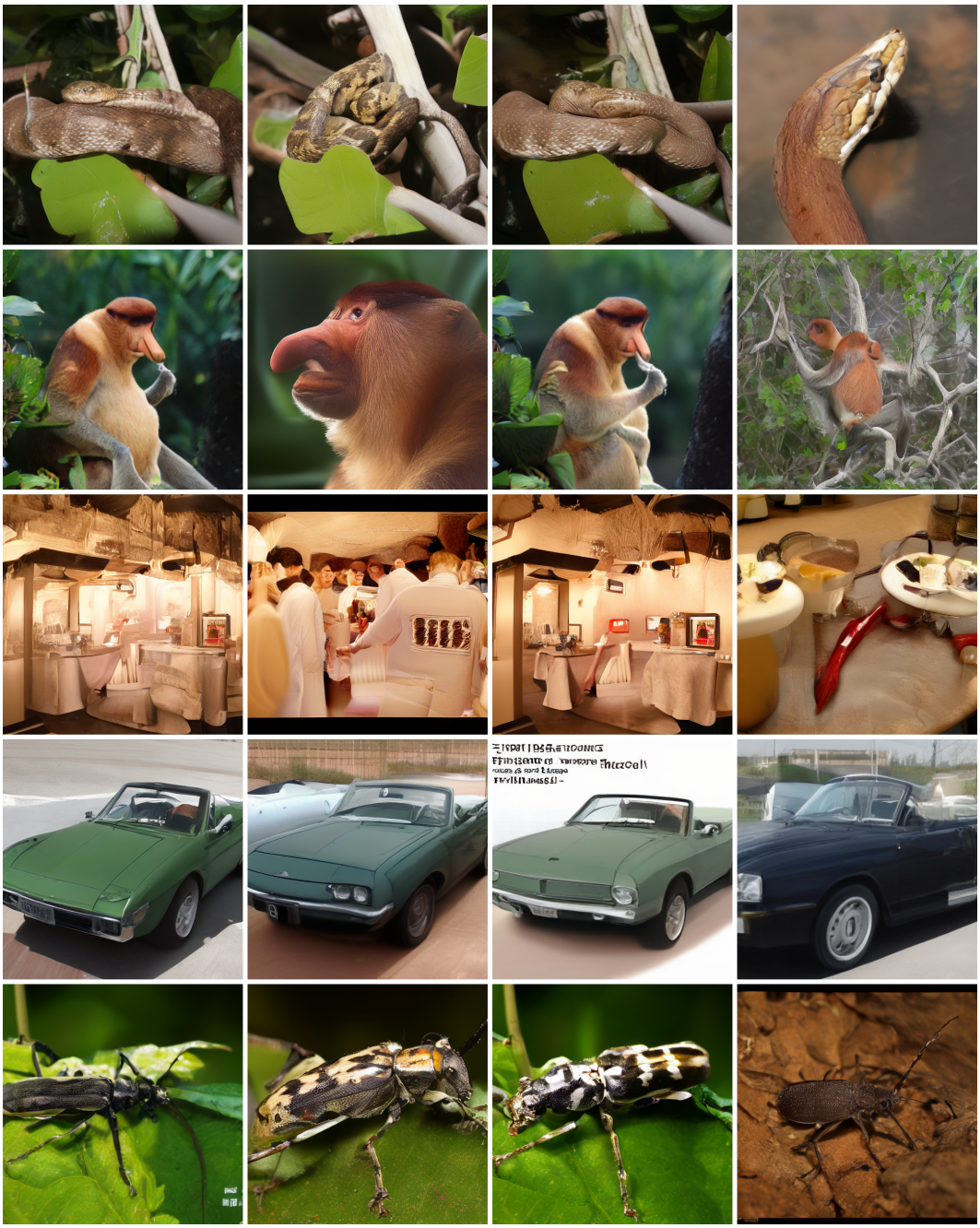

Figure 3: ImageNet 512×512 samples generated at NFE 50. From left to right, the schedules are ACS, VP-Linear, VP-Sigmoid, and Power. The rows are in order of: "Water snake", "Proboscis monkey", "Restaurant", "Convertible", and "Long-horned beetle". The samples generated using VP-Cosine were excluded because they were illegible for the ImageNet 512×512 dataset.

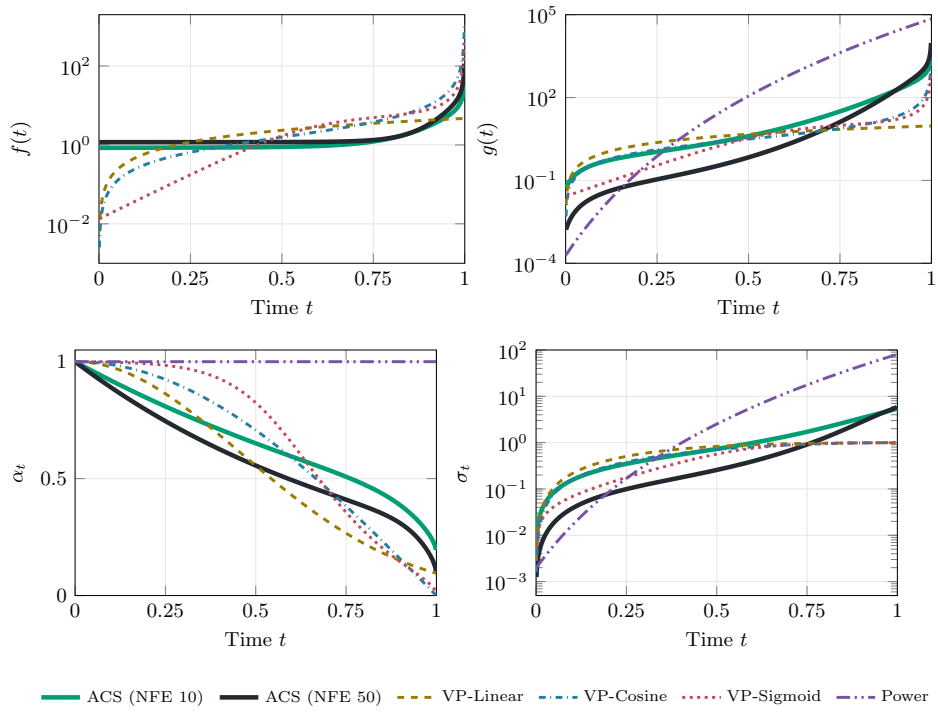

Figure 4: Noise schedules used in the generation experiments. The upper plots show logarithmic scales of $f(t)$ and $g(t)$, with Power's $f(t) = 0$ excluded. The lower plots show α_t and σ_t. All curves are evaluated on $t \in [10^{-3}, 1 - 10^{-3}]$.

Table 1: Search spaces and selected parameters used for the generation experiments. Logarithmic search spaces are indicated by "log."

Schedule	Parameter	Search space	Selected value
VP-Linear	$\beta_{\min}$	$[10^{-4}, 0.5]$ (log)	0.003318
	$\beta_{\max}$	$[1, 50]$ (log)	9.400954
VP-Cosine	offset	$[10^{-5}, 0.1]$ (log)	1.778279×10^{-5}
VP-Sigmoid	start	$[-6, -1]$	-4.576358
	end	$[1, 6]$	3.408840
	temperature	$[0.25, 2.5]$ (log)	0.765914
ACS	$\mathcal{E}_{10}$	$[0.0456, 2.625]$	1.634306
	γ	$[1, 8]$ (log)	4.443826
	θ	$[10^{-3}, 1]$ (log)	0.012058
	ρ	$[0, 0.99]$	0.515484
	ϵ	$[10^{-5}, 0.2]$ (log)	0.016110
	ℓ_1	$[-10, 5]$	-4.330955
	ℓ_2	$[-10, 5]$	1.366422
	ℓ_3	$[-10, 5]$	-0.586175