No-Sampling Predictors for Diffusion Quality via Spectral Kernels and Training-Time Residuals

Liz Lemma   Future Detective

January 19, 2026

Abstract

Sampling error in score-based generative models is driven by an interplay between score estimation noise (from finite data and non-convergent SGD), discretization, and truncation/early stopping. Recent Gaussian-linear analyses show that the end-to-end Wasserstein sampling error can be written as a kernel-type norm of the data power spectrum, with kernels determined by training and sampling parameters. We operationalize this perspective into a practical diagnostic that predicts the sampling error curve—versus stopping time and step schedule—without running the sampler. Our method estimates (i) spectral summaries of the data distribution in a convenient basis (e.g., Fourier bandpower) and (ii) score-error covariances from training-time denoising residuals on a held-out set. In the Gaussian-linear setting, we prove that the resulting estimator is consistent and give finite-sample uniform concentration bounds, yielding provable guarantees for selecting near-optimal stopping times. Beyond Gaussians, we provide kernel-based upper bounds with explicit remainder terms controlled by score nonlinearity certificates. Empirically (to be validated), the predictor forecasts optimal stopping and relative quality across runs while reducing evaluation cost by orders of magnitude.

Table of Contents

4. 4. Kernel Predictor Construction: derive the explicit kernel functional for (A) Langevin with fixed σ and (B) diffusion with schedule σ_t; define the estimator $\widehat{E}(r)$ from spectral and residual statistics; discuss variants for ODE vs SDE samplers.

5. 5. Finite-Sample Guarantees in the Gaussian-Linear Regime: (i) unbiasedness/consistency; (ii) uniform concentration over a grid of stopping times; (iii) regret bounds for selecting $\hat{r}^*$; (iv) improved rates under bandpower structure/spectral decay.

6. 6. Beyond Gaussian: Robust Upper Bounds with Nonlinearity Certificates: state sufficient conditions (local linearization, Lipschitz drift perturbation bounds) and derive $E(r) \leq \widehat{E}(r) + \mathrm{Nonlin}(r)$; show how to estimate/upper bound $\mathrm{Nonlin}(r)$ from validation probes.

7. 7. Complexity and Lower Bounds: sample complexity needed to estimate spectra/residual covariances; reductions from covariance estimation/trace estimation; show near-tightness of upper bounds (up to logs) under mild assumptions.

8. 8. Implementation Details (to strengthen contribution): scalable spectrum estimation (Hutch++/randomized bandpower); residual-statistic collection during training; numerical stability; calibration mapping from predicted W2 to FID/sliced-W2; failure modes.

9. 9. Experimental Plan (recommended): synthetic Gaussians and mixtures for exact ground truth; real-image diffusion runs (medium scale) comparing predicted vs measured curves; ablations on basis choice, correlation modeling, and solver choice.

10. 10. Discussion and Future Work: using correlated-time kernels (shared-network covariance), extending to distillation and guidance, integrating into automated schedulers.

1 Introduction

Empirical evaluation of diffusion and Langevin samplers is commonly performed by actually running the reverse-time procedure, generating a large collection of samples, and measuring a downstream discrepancy (e.g. a perceptual metric, a classifier-based score, or a proxy for distributional mismatch). This approach is computationally expensive for two intertwined reasons. First, each evaluation of a candidate stopping rule (or terminal time) requires producing full trajectories whose length scales with the number of discretization steps between the stopping time and the terminal noise level. Second, the evaluation must be repeated across multiple random seeds to control Monte Carlo variance, and across multiple candidate stopping times to obtain an error curve rather than a single number. In the regimes of interest—high-dimensional data, moderately large batch sizes, and tens to hundreds of solver steps—the cost of sampler-based evaluation is often comparable to (or larger than) the cost of training-time ablations, which makes systematic tuning of stopping times and step schedules impractical.

We study a different objective: given a trained model and access to held-out data, we aim to *predict* the end-to-end sampling error curve as a function of the reverse-time terminal gap. Concretely, for a grid of stopping gaps $\mathcal{R}$, we seek an estimator $\widehat{E}(r)$ of

$$E(r) \;:=\; \mathbb{E}\big[W_2^2(p_{\text{data}}, q_r)\big],$$

where q_r denotes the distribution of the sampler output when the reverse-time procedure is stopped at gap r (equivalently, the sampler is run only from time T down to time $t_k = T - r$). The operational goal is then to select a recommended stopping rule $\hat{r}^* \in \arg\min_{r\in\mathcal{R}} \widehat{E}(r)$ without generating long reverse trajectories. The constraint is essential: we allow forward noising of held-out data, and we allow querying the denoiser/score network on corrupted inputs, but we disallow iterative sampling for evaluation.

Our approach is based on a kernel-norm view of sampling error that becomes exact in a Gaussian-linear regime. When p_{data} is Gaussian and the reverse-time drift (induced by the learned score) is affine in the state, the reverse dynamics preserve Gaussianity: for each r, the law q_r is itself Gaussian with mean and covariance obtained by linear propagation. In that case, the squared Wasserstein distance admits a closed form in terms of the means and covariances; for example, for Gaussians $\mathcal{N}(m_1, \Sigma_1)$ and $\mathcal{N}(m_2, \Sigma_2)$,

$$W_2^2\big(\mathcal{N}(m_1,\Sigma_1), \mathcal{N}(m_2,\Sigma_2)\big) = \|m_1 - m_2\|^2 + \operatorname{tr}\Big(\Sigma_1 + \Sigma_2 - 2(\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2})^{1/2}\Big).$$

In the same regime, discretization error and score error can be propagated mode-by-mode along the eigendirections of the data covariance C, yielding an expression for $E(r)$ as a sum (or integral) of contributions weighted by sampler-dependent kernels. Informally, each eigenvalue λ_i contributes to the total error through a scalar weight determined by the noise schedule $\{\sigma_t\}$ and the discretization rule Γ, and the contribution is modulated by second moments of the score error along that mode.

The crucial consequence is that, under linearization of the score error,

$$e_t(x) \;=\; s_\theta(x,t) - \nabla \log p_t(x) \;\approx\; -\Delta_t x + \delta_t,$$

the dependence of $E(r)$ on the learned model enters only through (i) the spectrum of the data covariance and (ii) the second moments of (Δ_t, δ_t) across times. Both of these are estimable from held-out data *without* running the reverse-time sampler. The spectrum of C can be estimated either exactly in low dimensions or approximately via bandpower summaries in a suitable basis (e.g. Fourier bands). The score-error moments can be accessed through denoising residuals evaluated on forward-corrupted data: for $x \sim p_{\text{data}}$ and $w \sim \mathcal{N}(0, I)$, the quantity

$$r_t(x,w) \;:=\; \frac{x - D_{\sigma_t}(x + \sigma_t w; \theta)}{\sigma_t^2}$$

is a standard proxy for the score, and its fluctuations across held-out samples and noise draws provide unbiased estimates of the covariance terms that appear in the kernel expression. Thus, the expensive operation—reverse-time iterative sampling—is replaced by two inexpensive primitives: forward noising and network evaluation.

We therefore formulate the following practical task. Given a trained diffusion (or Langevin) model with a fixed schedule and solver, we compute a predicted curve $r \mapsto \widehat{E}(r)$ on a user-chosen grid $\mathcal{R}$, together with a recommended stopping gap $\hat{r}^*$. The predicted curve decomposes into interpretable components: a truncation/noise-floor term, a discretization term (depending on the solver and step schedule), and a score-error term which is an explicit kernel functional of estimated spectral summaries and residual covariances. In the Gaussian-linear setting, this plug-in construction is unbiased for each r, and with sufficiently many held-out samples and residual draws it concentrates uniformly over $\mathcal{R}$, which implies near-optimal stopping-time selection by standard argmin stability arguments.

Our contributions are accordingly: (a) a no-sampling evaluation method that produces an entire predicted error curve and a recommended stopping rule; (b) a finite-sample analysis showing uniform deviation bounds for $\widehat{E}(r)$ and regret guarantees for $\hat{r}^*$ under natural concentration assumptions; (c) high-dimensional implementations via bandpower and randomized spectral estimation that reduce dependence on d to dependence on the number of spectral bands; and (d) an extension beyond Gaussianity in which the same estimator yields a certified upper bound with an additional remainder term controlled by measurable nonlinearity/curvature proxies of the learned score on the validation distribution. The remainder of the paper formalizes the diffusion/Langevin setup, the residual estimators, and the spectral approximations, after which we state the kernel formulas and the associated guarantees.

2 Preliminaries

We collect the diffusion/Langevin notation used throughout, recall the closed-form expression of W_2 for Gaussians, and formalize the validation-time residual quantities from which we estimate score error moments. We also summarize the spectral and bandpower statistics required to evaluate the kernel functionals that appear later.

Forward noising and reverse-time dynamics. Let $x_0 \sim p_{\text{data}}$ on $\mathbb{R}^d$. We consider a continuous-time diffusion with noise schedule $\{\sigma_t\}_{t\in[0,T]}$ such that the forward marginal p_t is obtained by Gaussian corruption of the form

$$x_t \;=\; x_0 + \sigma_t w, \qquad w \sim \mathcal{N}(0, I), \tag{1}$$

which is the marginal relation in variance-exploding (VE) diffusions and is also the primitive used in denoising score matching (DSM). In this setting, the (Stein) score of the forward marginal is $\nabla \log p_t(x)$, and the learned model provides either a score estimate $s_\theta(x,t)$ or a denoiser $D_{\sigma_t}(x;\theta)$ from which one can obtain a score proxy.

The reverse-time sampling procedure is determined by the choice of reverse dynamics (reverse SDE, or the probability-flow ODE) and a discretization rule with step schedule Γ. Concretely, we consider a time grid

$$T = t_0 > t_1 > \cdots > t_K \geq 0, \qquad \gamma_k := t_k - t_{k+1} > 0,$$

and an update map of the generic form

$$X_{t_{k+1}} \;=\; \Psi_k(X_{t_k};\, s_\theta(\cdot, t_k), \gamma_k)\,,$$

where Ψ_k is Euler–Maruyama for the reverse SDE, or an ODE integrator for the probability-flow ODE. We will index early stopping by the *terminal gap*

$$r \;:=\; T - t_k,$$

meaning we run the reverse procedure from T down to $t_k = T - r$ and stop. We denote by q_r the output distribution at this stopping gap.

W_2 and Gaussian closed forms. For probability measures π, ν on $\mathbb{R}^d$ with finite second moments, the squared 2-Wasserstein distance is

$$W_2^2(\pi,\nu) \;:=\; \inf_{\gamma\in\Pi(\pi,\nu)} \int \|x-y\|^2\, d\gamma(x,y),$$

where $\Pi(\pi, \nu)$ is the set of couplings with marginals π and ν. When both measures are Gaussian, $\pi = \mathcal{N}(m_1, \Sigma_1)$ and $\nu = \mathcal{N}(m_2, \Sigma_2)$, we have the classical identity

$$W_2^2(\pi, \nu) = \|m_1 - m_2\|^2 + \operatorname{tr}\Big(\Sigma_1 + \Sigma_2 - 2(\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2})^{1/2}\Big). \tag{2}$$

In particular, if Σ_1 and Σ_2 commute (e.g. they are simultaneously diagonalizable in a common eigenbasis), then (2) reduces to a mode-wise sum. Writing $\Sigma_j = \sum_{i=1}^{d} \lambda_i^{(j)} u_i u_i^\top$ with the same $\{u_i\}$, we obtain

$$W_2^2(\pi, \nu) \; = \; \|m_1 - m_2\|^2 + \sum_{i=1}^{d} \left(\sqrt{\lambda_i^{(1)}} - \sqrt{\lambda_i^{(2)}}\right)^2. \tag{3}$$

This mode-wise form is the reason that, in the Gaussian-linear regime considered later, the sampling error decomposes into a sum of scalar contributions indexed by the eigendirections of the data covariance.

Denoising residuals and score error. We assume access to a denoiser $D_{\sigma_t}(\cdot; \theta)$ (possibly implemented implicitly via a score network). Under Gaussian corruption (1), the Bayes-optimal denoiser is $D^\star_{\sigma_t}(y) = \mathbb{E}[x_0 \mid x_t = y]$. The following identity (Tweedie's formula) relates this conditional mean to the score of the noisy marginal:

$$\nabla \log p_t(y) \; = \; \frac{D^\star_{\sigma_t}(y) - y}{\sigma_t^2}. \tag{4}$$

Accordingly, given a learned denoiser, we form the standard score proxy

$$r_t(x, w) \; := \; \frac{x - D_{\sigma_t}(x + \sigma_t w; \theta)}{\sigma_t^2}, \qquad x \sim p_{\text{data}}, \; w \sim \mathcal{N}(0, I), \tag{5}$$

which coincides with $-\nabla \log p_t(x + \sigma_t w)$ when $D_{\sigma_t} = D^\star_{\sigma_t}$ and $\mu = 0$ (more generally one accounts for the mean shift in the usual way). We denote the learned score by $s_\theta(\cdot, t)$ and define the pointwise score error at time t by

$$e_t(y) \; := \; s_\theta(y, t) - \nabla \log p_t(y).$$

In the regime where the reverse-time drift is well-approximated by an affine map in y (e.g. near a Gaussian reference), we model this error by a linearization

$$e_t(y) \; \approx \; -\Delta_t y + \delta_t, \tag{6}$$

where $\Delta_t \in \mathbb{R}^{d \times d}$ and $\delta_t \in \mathbb{R}^d$ are random (capturing randomness induced by the learned network and the data/noise), with $\mathbb{E}[\delta_t] = 0$ and finite second moments. The prediction formulas we use later depend on $\operatorname{Cov}(\delta_t)$ and on second moments of Δ_t (typically through $\operatorname{Cov}(\operatorname{vec}(\Delta_t))$). Our validation-time procedure estimates these moments from fluctuations of (5) across held-out samples and independent noise draws.

Spectrum and bandpower summaries. In the Gaussian setting, we write $p_{\text{data}} = \mathcal{N}(\mu, C)$ with $C \succ 0$ and eigendecomposition $C = \sum_{i=1}^{d} \lambda_i u_i u_i^\top$. The kernel expressions for sampling error depend on C through scalar functionals of its spectrum (e.g. sums of the form $\sum_i f(\lambda_i)$ for sampler-dependent f). When d is small, we may estimate C directly from held-out data and compute its eigenvalues. In high dimensions, we instead compute *compressed spectral statistics* that are sufficient for a given approximation class.

One convenient option is a fixed orthogonal basis (e.g. Fourier) with a partition into B spectral bands. Let Π_b denote the orthogonal projector onto band b. We define the bandpowers

$$P_b \;:=\; \mathbb{E}\big[\|\Pi_b(x-\mu)\|^2\big] \;=\; \mathrm{tr}(\Pi_b C), \tag{7}$$

and estimate them by the empirical averages $(1/n)\sum_{i=1}^{n} \|\Pi_b(x_i - \bar{x})\|^2$ on held-out data. These summaries replace explicit eigenvalues by B aggregate variances, which reduces both computation and statistical complexity from scaling with d to scaling with B in the subsequent kernel evaluation.

More generally, when the kernels require quantities of the form $\mathrm{tr}(f(C))$ for scalar functions f, we may use randomized trace estimation (e.g. Hutchinson-type estimators) applied to the sample covariance operator, yielding estimates without forming a full $d \times d$ matrix. The later sections specify exactly which spectral summaries are needed for a given sampler and kernel approximation, and how these summaries interface with the residual covariance estimates to produce a predicted curve $r \mapsto \widehat{E}(r)$.

3 Problem formulation

We formalize the *no-sampling prediction* task addressed in this work. Fix a trained diffusion or Langevin model θ, together with its sampling specification: a noise family $\{\sigma_t\}_{t\in[0,T]}$ (or a fixed σ in the Langevin-DSM setting), a reverse-time dynamic (reverse SDE or probability-flow ODE), and a discretization rule with step schedule Γ on a descending grid $T = t_0 > t_1 > \cdots > t_K \geq 0$. For a terminal gap $r := T - t_k$, we denote by q_r the output distribution produced by running the reverse procedure from T down to $t_k = T - r$ and stopping. Our target is the *end-to-end sampling error curve*

$$E(r) \;:=\; \mathbb{E}\big[W_2^2(p_{\text{data}}, q_r)\big]\,, \qquad r \in [0, T],$$

where the expectation is over the sampler randomness (and, when relevant, over any randomness in the learned score through its stochastic error model).

No-sampling prediction task. We assume access to a held-out dataset $\{x_i\}_{i=1}^{n} \sim p_{\text{data}}$ and oracle access to the trained denoiser/score in the following restricted sense: for user-chosen time points t and for sampled $w \sim$

$\mathcal{N}(0, I)$, we may query

$$y \;=\; x_i + \sigma_t w, \qquad \text{and evaluate} \qquad D_{\sigma_t}(y;\theta) \;\text{ and/or }\; s_\theta(y,t).$$

Equivalently, we may form the standard DSM residual/score proxy

$$r_t(x_i, w) \;:=\; \frac{x_i - D_{\sigma_t}(x_i + \sigma_t w;\theta)}{\sigma_t^2},$$

and compute empirical moments of $r_t(x_i, w)$ across held-out samples and independent noise draws. Crucially, we disallow *reverse-time rollout*: we do not generate trajectories $(X_{t_k})_k$ from the sampler, even for a small number of steps, since the aim is to predict long-horizon behavior without incurring the computational cost (or design entanglement) of running the sampler itself. The only permitted operations are (i) forward corruptions of held-out data, (ii) network evaluations at those corrupted points, and (iii) lightweight spectral estimation routines on the held-out data (e.g. bandpowers in a fixed orthogonal basis, or randomized trace estimation).

Allowed statistics and interface to the predictor. We constrain the predictor to depend on the held-out data and model only through a finite collection of summary statistics, computed on a modest set of probe times $\mathcal{T} \subset [0, T]$ and a finite stopping grid $\mathcal{R} \subset [0, T]$. Concretely, we permit:

1. *Spectral summaries of the data covariance.* In the Gaussian regime these are functions of C (or of C projected to a bandpower basis), such as $\mathrm{tr}(\Pi_b C)$ for bands $\{\Pi_b\}_{b=1}^B$, or more general functionals $\mathrm{tr}(f(C))$ estimated by randomized linear algebra. The predictor may use any such summaries as long as they are computable from $\{x_i\}$ without forming long Markov chains.

2. *Per-time residual/error covariances.* For each $t \in \mathcal{T}$, from m independent noise draws per x_i (or an equivalent batching strategy), we may estimate second moments of the score error model in (6), summarized by empirical estimators $\widehat{V_t}$ and $\widehat{W_t}$ targeting $\mathrm{Cov}(\delta_t)$ and $\mathrm{Cov}(\mathrm{vec}(\Delta_t))$, respectively. We emphasize that we do not require estimating full $d \times d$ covariances when d is large: diagonal, bandpower-projected, or low-rank sketches are admissible provided the subsequent kernel evaluation only depends on these projections.

Given these statistics, the predictor outputs a curve $\widehat{E} : \mathcal{R} \to \mathbb{R}_+$ and a recommended stopping gap

$$\hat{r}^* \in \arg\min_{r \in \mathcal{R}} \widehat{E}(r).$$

Optionally, the predictor may also output a recommendation for modifying the discretization (e.g. a refined step schedule Γ), but in the present formulation we treat Γ as fixed and known.

Evaluation metrics. We evaluate the quality of the predicted curve and the induced stopping rule via two complementary criteria. First, we measure the accuracy of the curve estimate uniformly over the grid:

$$\mathcal{E}_{\mathrm{unif}}(\widehat{E}; E) \;:=\; \sup_{r\in\mathcal{R}} \big|\widehat{E}(r) - E(r)\big|. \tag{8}$$

In applications one may also consider an average version $\frac{1}{|\mathcal{R}|}\sum_{r\in\mathcal{R}} |\widehat{E}(r) - E(r)|$, but the uniform metric is the natural quantity for guaranteeing stable argmin selection on a discrete grid.

Second, we measure the quality of the predicted early-stopping choice by its *regret* relative to the best grid point:

$$\mathrm{Regret}(\hat{r}^*) \;:=\; E(\hat{r}^*) - \min_{r\in\mathcal{R}} E(r). \tag{9}$$

This notion isolates the operational consequence of prediction error: even if $\widehat{E}$ has nontrivial pointwise deviations, it may still yield a near-optimal stopping time if the minimizer is robust. In later sections we relate (9) to (8) through standard argmin stability conditions.

Statistical goal under restricted access. Under the Gaussian-linear assumptions stated in the enclosing scope, we aim to design a plug-in estimator $\widehat{E}(r)$ built solely from the allowed statistics above, such that for given accuracy/failure parameters (ε, δ),

$$\Pr\Big[\mathcal{E}_{\mathrm{unif}}(\widehat{E}; E) \le \varepsilon\Big] \ge 1 - \delta,$$

with sample sizes (n, m) and computational cost that scale tractably in dimension (or in the compressed dimension B under bandpower structure). The subsequent section constructs $\widehat{E}$ by identifying an explicit kernel functional $\mathcal{K}_{\mathrm{sampler}}(r; \cdot)$ for the chosen sampler (SDE or ODE, fixed-noise Langevin or scheduled diffusion), and by specifying how the spectral and residual statistics are assembled into that kernel to produce the predicted curve and stopping recommendation.

4 Kernel predictor construction

We construct an explicit plug-in predictor by reducing the sampler, under the Gaussian–linear assumptions, to a family of decoupled one-dimensional linear systems in the eigenbasis of the data covariance. Throughout we write $C = \sum_{i=1}^{d} \lambda_i u_i u_i^\top$ and analyze each coordinate $\xi_{t,i} := \langle u_i, X_t - \mu\rangle$ separately; all kernels below are scalar functions of λ_i and the sampler specification.

Ideal dynamics and error injection. Let p_t denote the forward noised distribution at time t. In the VE (additive-noise) setting, $p_t = \mathcal{N}(\mu, C + \sigma_t^2 I)$ and hence

$$\nabla \log p_t(x) = -(C + \sigma_t^2 I)^{-1}(x - \mu).$$

We decompose the learned score as $s_\theta(x, t) = \nabla \log p_t(x) + e_t(x)$ and impose the linearized error model $e_t(x) \approx -\Delta_t x + \delta_t$ with $\mathbb{E}[\delta_t] = 0$ and second moments accessible through validation residual statistics. Substituting into a reverse-time scheme yields a linear(ized) recursion whose mean and covariance admit closed-form propagation; the contribution of $\mathrm{Cov}(\delta_t)$ and $\mathrm{Cov}(\Delta_t)$ appears linearly through kernels determined by the solver.

4.1 (A) Fixed-noise Langevin (DSM) kernel

We first treat the fixed-noise DSM/Langevin setting at noise level σ. The target of the ideal Langevin chain is the σ-smoothed density $p_\sigma = \mathcal{N}(\mu, C + \sigma^2 I)$, with score $s^*(x) = -(C + \sigma^2 I)^{-1}(x - \mu)$. Consider the Euler–Maruyama discretization with constant step $\gamma > 0$:

$$X_{k+1} = X_k + \gamma\, s_\theta(X_k) + \sqrt{2\gamma}\, Z_k, \qquad Z_k \sim \mathcal{N}(0, I). \tag{10}$$

In the eigen-direction u_i, the ideal drift coefficient is $a_i := 1/(\lambda_i + \sigma^2)$. Writing $\xi_{k,i} := \langle u_i, X_k - \mu \rangle$ and linearizing $e(x) \approx -\Delta x + \delta$ yields

$$\xi_{k+1,i} = \underbrace{(1 - \gamma a_i)}_{=:\rho_i}\, \xi_{k,i} \;-\; \gamma \langle u_i, \Delta X_k \rangle \;+\; \gamma \langle u_i, \delta \rangle \;+\; \sqrt{2\gamma}\, z_{k,i}, \tag{11}$$

where $z_{k,i} \sim \mathcal{N}(0, 1)$. Ignoring higher-order products between Δ and the state (which vanish in expectation under the Gaussian–linear closure assumed in the enclosing scope), we obtain an explicit kernel representation for the second moment of $\xi_{K,i}$ as a weighted sum of per-step error covariances. Concretely, let $w_{i,j}^{(\delta)} := \gamma\, \rho_i^{K-1-j}$ denote the sensitivity of $\xi_{K,i}$ to an additive perturbation at step j. Then the additive-score-error contribution takes the form

$$\mathbb{E}\big[\xi_{K,i}^2\big] \;\supset\; \sum_{j=0}^{K-1} \big(w_{i,j}^{(\delta)}\big)^2\, \mathrm{Var}\big(\langle u_i, \delta_j \rangle\big), \tag{12}$$

and the multiplicative error Δ_j contributes analogously with weights proportional to ρ_i^{K-1-j} and factors depending on $\mathbb{E}[\xi_{j,i}^2]$ (hence, ultimately, on λ_i). Collecting these terms defines a sampler-dependent kernel $\mathcal{K}_{\mathrm{Lan}}$ which is polynomial in ρ_i and linear in the required projected covariances of δ and Δ.

For the *truncation/noise-floor* term, we note that stopping Langevin at stationarity would match p_σ rather than p_{data}, so the irreducible gap is

$$E_{\mathrm{Lan}}^{(0)}(\sigma) = W_2^2\big(\mathcal{N}(\mu, C), \mathcal{N}(\mu, C + \sigma^2 I)\big) = \sum_{i=1}^{d} \big(\sqrt{\lambda_i + \sigma^2} - \sqrt{\lambda_i}\big)^2, \tag{13}$$

which is computable from spectral summaries of C. Discretization error $E^{\mathrm{disc}}_{\mathrm{Lan}}$ is obtained by comparing the discrete covariance propagation induced by (10) under the *ideal* score to the corresponding continuous-time Ornstein–Uhlenbeck flow; in the eigenbasis this comparison is again scalar and depends only on (λ_i, γ, K).

4.2 (B) Scheduled diffusion kernel: reverse SDE versus probability-flow ODE

We next treat a scheduled diffusion with noise levels $\{\sigma_t\}_{t\in[0,T]}$ discretized on $T = t_0 > \cdots > t_K \geq 0$. For a broad class of VE parameterizations, an Euler step of the reverse SDE has the schematic form

$$X_{t_{k+1}} = X_{t_k} + b_k\, s_\theta(X_{t_k}, t_k)\, \Delta t_k + \sqrt{b_k\, \Delta t_k}\, Z_k, \tag{14}$$

with known scalar coefficient b_k (determined by the schedule) and $\Delta t_k := t_{k+1} - t_k < 0$. In the eigen-direction u_i, the ideal linear drift coefficient is $b_k/(\lambda_i + \sigma^2_{t_k})$. Defining the per-step propagator

$$\Pi_i(k \to \ell) := \prod_{j=k}^{\ell-1} \Big(1 - \frac{b_j}{\lambda_i + \sigma^2_{t_j}}\, \Delta t_j\Big), \qquad k < \ell,$$

we obtain weights describing how an error injected at time t_k influences the stopped output at t_ℓ. In particular, for additive error δ_{t_k} the contribution is governed by

$$w^{(\delta)}_{i,k}(r) := \Pi_i(k \to k_r)\, b_k\, \Delta t_k, \qquad k_r := \min\{j :\ T - t_j \geq r\}, \tag{15}$$

and the associated kernel term is a Riemann-sum approximation of $\sum_k (w^{(\delta)}_{i,k}(r))^2\, \mathrm{Var}(\langle u_i, \delta_{t_k}\rangle)$ (with an analogous multiplicative term for Δ_{t_k}). The truncation term $E^{(0)}(r)$ corresponds to stopping at nonzero t_{k_r} even with a perfect solver; in the VE Gaussian case it reduces to

$$E^{(0)}_{\mathrm{diff}}(r) = W_2^2\big(\mathcal{N}(\mu, C), \mathcal{N}(\mu, C + \sigma^2_{t_{k_r}} I)\big) = \sum_{i=1}^{d} \big(\sqrt{\lambda_i + \sigma^2_{t_{k_r}}} - \sqrt{\lambda_i}\big)^2. \tag{16}$$

For the probability-flow ODE, the update is identical to (14) but without the stochastic term and with the standard drift scaling change (replacing b_k by a known multiple, typically $b_k/2$ under the usual SDE–ODE correspondence). Consequently, the kernel weights (15) and the discretization term E^{disc} are modified deterministically, while the plug-in dependence on $\mathrm{Cov}(\delta_t)$ and $\mathrm{Cov}(\Delta_t)$ remains linear with the same projected statistics.

4.3 Plug-in estimator from spectral and residual statistics

Given (i) spectral summaries of C (either $\{\hat{\lambda}_i\}$ in low dimension or bandpowers $\{\widehat{P}_b\}_{b=1}^B$ in a fixed basis) and (ii) per-time residual-based estimates $\widehat{V}_t$ and $\widehat{W}_t$, we define

$$\widehat{E}(r) := \widehat{E}^{(0)}(r) + \widehat{E}^{\mathrm{disc}}(r) + \widehat{E}^{\mathrm{score}}(r).$$

Here $\widehat{E}^{(0)}$ and $\widehat{E}^{\mathrm{disc}}$ are computed by running the *ideal* mean/covariance recursions in the eigen- or bandpower-reduced representation, using the known schedule and solver. The term $\widehat{E}^{\mathrm{score}}(r)$ is obtained by evaluating the sampler-specific kernel (Langevin or diffusion; SDE or ODE variant) on the estimated projections of $\widehat{V}_t$ and $\widehat{W}_t$:

$$\widehat{E}^{\mathrm{score}}(r) = \sum_{t\in\mathcal{T}:\ t\geq T-r} \ \sum_{b=1}^{B} \Big(\kappa_{b,t}^{(\delta)}(r)\, \widehat{v}_{b,t} \ + \ \kappa_{b,t}^{(\Delta)}(r)\, \widehat{w}_{b,t}\Big), \tag{17}$$

where $\widehat{v}_{b,t}$ and $\widehat{w}_{b,t}$ denote the bandpower (or eigen-coordinate) projections of the error covariances, and the coefficients $\kappa_{b,t}^{(\delta)}(r), \kappa_{b,t}^{(\Delta)}(r)$ are determined by the propagators of the chosen sampler. The finite-sample properties of this plug-in construction are established in the next section by combining unbiasedness of the residual statistics with concentration of the spectral and covariance estimates.

5 Finite-sample guarantees in the Gaussian–linear regime

We now state finite-sample properties of the plug-in curve $\widehat{E}(\cdot)$ when the enclosing Gaussian–linear assumptions hold, so that $\mathrm{Rem}(r) = 0$ and $E(r)$ is an explicit multilinear functional of (i) spectral summaries of C and (ii) second moments of the linearized score-error parameters. Throughout, we fix a finite stopping grid $\mathcal{R}$ and a finite probe-time set $\mathcal{T} \subset [0, T]$ used to form the residual statistics, and we regard the sampler specification (schedule, solver, and discretization rule) as fixed and known.

5.1 Unbiasedness and consistency of the plug-in curve

The first guarantee is that, in the idealized regime where the spectral quantities entering the kernel are known exactly, the residual-based portion of $\widehat{E}(r)$ is unbiased for the corresponding contribution to $E(r)$. Concretely, the score-error term admits a representation of the form

$$E^{\mathrm{score}}(r) = \sum_{t\in\mathcal{T}:\ t\geq T-r} \Big\langle A_t(r),\, \mathrm{Cov}(\delta_t)\Big\rangle + \sum_{t\in\mathcal{T}:\ t\geq T-r} \Big\langle B_t(r),\, \mathrm{Cov}(\mathrm{vec}(\Delta_t))\Big\rangle,$$

for deterministic coefficient operators $A_t(r), B_t(r)$ determined by the sampler propagators and by C only through its spectral summaries. By construction of the residual proxies (DSM/Tweedie identities under the Gaussian forward corruption model), our estimators satisfy $\mathbb{E}[\widehat{V}_t] = \mathrm{Cov}(\delta_t)$ and $\mathbb{E}[\widehat{W}_t] = \mathrm{Cov}(\mathrm{vec}(\Delta_t))$, whence linearity yields

$$\mathbb{E}\Big[\widehat{E}^{\mathrm{score}}(r) \,\Big|\, \text{spectral inputs}\Big] = E^{\mathrm{score}}(r) \qquad \text{for each } r \in \mathcal{R}.$$

The remaining terms $\widehat{E}^{(0)}(r)$ and $\widehat{E}^{\mathrm{disc}}(r)$ depend only on the sampler and C (through scalar functions of its eigenvalues). If we are in a low-dimensional regime where C is estimated by the sample covariance and eigenvalues are plugged in, these terms are generally only *consistent* (since eigenvalues are nonlinear in the sample covariance). However, in the bandpower/trace-functional setting the dependence on C can be arranged to be linear in moments such as $\mathrm{tr}(f(C))$ for explicit scalar functions f (rational functions in the VE Gaussian case), and then standard randomized trace estimators yield unbiased estimates of these spectral inputs. In either case, as $n, m \to \infty$ with $|\mathcal{T}|, |\mathcal{R}|$ fixed, we have $\widehat{E}(r) \to E(r)$ in probability pointwise for each r, and under the concentration bounds below the convergence is uniform over $\mathcal{R}$.

5.2 Uniform concentration over a stopping grid

We next bound $\sup_{r\in\mathcal{R}} |\widehat{E}(r) - E(r)|$ with high probability. The argument is a stability inequality for the kernel functional plus concentration of the estimated inputs. We write schematically

$$\widehat{E}(r) - E(r) = \underbrace{\mathrm{Err}_{\mathrm{spec}}(r)}_{\text{spectral summary error}} + \underbrace{\mathrm{Err}_{\mathrm{res}}(r)}_{\text{residual-covariance error}},$$

where $\mathrm{Err}_{\mathrm{spec}}(r)$ captures the effect of using $\widehat{\lambda}$ (or bandpowers $\widehat{P}_b$) in place of the population spectral summaries, and $\mathrm{Err}_{\mathrm{res}}(r)$ captures the effect of using $\widehat{V}_t, \widehat{W}_t$ in place of their expectations.

Since p_{data} is Gaussian and the residual proxies are (conditionally) sub-Gaussian under the corruption model, standard matrix concentration (e.g. matrix Bernstein for sample covariances and Hanson–Wright for quadratic forms) yields, for each fixed t and each fixed band/eigendirection summary, deviations of order $O\big(\sqrt{\log(1/\delta)/(nm)}\big)$ (for residual statistics) and $O\big(\sqrt{\log(1/\delta)/n}\big)$ (for spectral summaries). The kernel assembly step is Lipschitz in these inputs: there exists a sampler-dependent constant L such that, for all $r \in \mathcal{R}$,

$$|\widehat{E}(r) - E(r)| \le L\Big(\|\widehat{\mathrm{spec}} - \mathrm{spec}\| + \max_{t\in\mathcal{T}} \|\widehat{V}_t - \mathrm{Cov}(\delta_t)\| + \max_{t\in\mathcal{T}} \|\widehat{W}_t - \mathrm{Cov}(\mathrm{vec}(\Delta_t))\|\Big),$$

with norms chosen compatibly with the representation (coordinatewise for eigen-directions, bandpowerwise for Fourier bands, or operator/Frobenius norms for low-rank models). Applying a union bound over $|\mathcal{T}|$ and the finite grid $\mathcal{R}$ (or, more precisely, over the finite set of kernel coefficients that are re-used for all r) yields the stated uniform guarantee: there exists an explicit $N(\cdot)$ such that if

$$n, m \gtrsim \frac{B + \log\big(|\mathcal{R}||\mathcal{T}|/\delta\big)}{\varepsilon^2} \quad \text{(bandpower model)}, \qquad n, m \gtrsim \frac{d + \log\big(|\mathcal{R}||\mathcal{T}|/\delta\big)}{\varepsilon^2} \quad \text{(full-spectrum m}$$

then

$$\Pr\Big[\sup_{r\in\mathcal{R}} |\widehat{E}(r) - E(r)| \le \varepsilon\Big] \ge 1 - \delta,$$

up to absolute constants and mild dependence on the schedule through L. The key point is that no reverse-time sampling is needed: all randomness comes from held-out data and forward corruptions.

5.3 Stopping-time selection and regret

Let $\hat{r}^* \in \arg\min_{r\in\mathcal{R}} \widehat{E}(r)$ and $r^* \in \arg\min_{r\in\mathcal{R}} E(r)$. On the event $\sup_{r\in\mathcal{R}} |\widehat{E}(r) - E(r)| \le \varepsilon$, we have the deterministic inequality

$$E(\hat{r}^*) \le \widehat{E}(\hat{r}^*) + \varepsilon \le \widehat{E}(r^*) + \varepsilon \le E(r^*) + 2\varepsilon,$$

hence the regret is bounded by $E(\hat{r}^*) - \min_{r\in\mathcal{R}} E(r) \le 2\varepsilon$. If, additionally, $E(\cdot)$ satisfies a discrete margin condition on $\mathcal{R}$ (e.g. strong convexity around its minimizers), then the same uniform deviation bound converts directly into a bound on the distance from $\hat{r}^*$ to the set of minimizers, with scaling proportional to ε divided by the margin parameter.

5.4 Improved rates under bandpower structure and spectral decay

The preceding bounds scale with d only through the complexity of estimating the spectral inputs required by the kernel. When we replace full eigenvalue dependence by a bandpower approximation in a fixed basis (e.g. Fourier bands for image-like data), the effective dimension becomes $B \ll d$, yielding the sample complexity $n, m = \tilde{O}(B/\varepsilon^2)$ plus an approximation bias term Bias_B in the final curve. This bias is deterministic and can be bounded by smoothness of the scalar kernel functions in λ together with the bandwidth of the spectral partition: if the kernel depends on λ through a Lipschitz function f, then Bias_B is controlled by the within-band variation of f times the mass of the spectrum in that band.

More generally, when the spectrum of C decays, one can quantify an *effective rank* dependence. Many kernel components reduce to trace functionals of the form $\mathrm{tr}(f(C))$ for monotone, bounded rational f (arising from resolvents $(C + \alpha I)^{-1}$ along the schedule). For such f, standard effective-dimension quantities (e.g. $d_{\text{eff}}(\alpha) := \operatorname{tr}\big(C(C+\alpha I)^{-1}\big)$ or $r_{\text{eff}} := \operatorname{tr}(C)/\|C\|_{\text{op}}$) control the variance of randomized trace estimators and, correspondingly, the required n for a given accuracy. In this regime the curve predictor can be implemented with rates depending on $d_{\text{eff}}(\alpha)$ (uniformly over the relevant α induced by σ_t), rather than on the ambient dimension d, without altering the no-sampling nature of the procedure.

5.5 Beyond Gaussian: robust upper bounds with nonlinearity certificates

When p_{data} is not Gaussian (and, concomitantly, when the learned score is not globally affine on the region explored by the reverse dynamics), the remainder term $\operatorname{Rem}(r)$ in the kernel decomposition need not vanish. In this regime we do not interpret $\widehat{E}(r)$ as an unbiased predictor of $E(r)$; instead, we use it as a *baseline* and derive a certified upper bound of the form

$$E(r) \;\le\; \widehat{E}(r) \;+\; \operatorname{Nonlin}(r),$$

where $\operatorname{Nonlin}(r)$ is a computable quantity estimated from the same validation probes used to form $\widehat{V}_t, \widehat{W}_t$.

A surrogate linearized sampler and a drift-gap inequality. Fix a probe set $\mathcal{T} \subset [0, T]$ and construct, for each $t \in \mathcal{T}$, an affine approximation to the learned score (or, equivalently, to the learned reverse drift on that time slice). Concretely, we fit parameters $(\widehat{\Delta}_t, \widehat{\delta}_t)$ (by least squares over validation probes, bandwise regression, or any restricted parametric fit compatible with the kernel representation) and define the *surrogate* score

$$\tilde{s}_t(x) \;:=\; -\widehat{\Delta}_t x + \widehat{\delta}_t.$$

Let q_r denote the distribution produced by the actual sampler (using s_θ) when stopped at gap r, and let $\tilde{q}_r$ denote the distribution produced by the same sampler specification (same σ_t and discretization rule) but with s_θ replaced by $\tilde{s}_t$ at the probed times (or interpolated between them). By construction, the quantity $\widehat{E}(r)$ is precisely the kernel evaluation associated with this affine surrogate (together with the estimated second moments); hence $\widehat{E}(r)$ is the natural proxy for $W_2^2(p_{\text{data}}, \tilde{q}_r)$ in the regime where the kernel derivation is accurate.

We then compare q_r and $\tilde{q}_r$ by a drift perturbation bound. Writing the reverse-time dynamics abstractly as an SDE

$$dX_t = b_t(X_t)\,dt + \sqrt{2}\,dB_t, \qquad t \in [T - r, T],$$

where b_t is the (schedule-dependent) drift induced by the sampler and score, we decompose

$$b_t(x) \;=\; \tilde{b}_t(x) \;+\; \Delta b_t(x), \qquad \Delta b_t(x) \;:=\; b_t(x) - \tilde{b}_t(x),$$

with $\tilde{b}_t$ the drift obtained by replacing s_θ with $\tilde{s}_t$. For standard VE/VP samplers, Δb_t is a known scalar multiple of $s_\theta(\cdot,t) - \tilde{s}_t(\cdot)$; we denote this scalar factor by α_t so that $\|\Delta b_t(x)\| \le \alpha_t \|s_\theta(x,t) - \tilde{s}_t(x)\|$.

Assume a one-sided Lipschitz (contractivity) condition for the surrogate drift: there exists a measurable $\kappa_t \in \mathbb{R}$ such that

$$\langle x-y, \tilde{b}_t(x) - \tilde{b}_t(y)\rangle \;\le\; -\kappa_t \|x-y\|^2 \qquad \text{for all } x, y \text{ in the relevant region.} \tag{18}$$

Under (18), the standard synchronous coupling argument yields a stability inequality of the schematic form

$$W_2^2(q_r, \tilde{q}_r) \;\le\; \int_{T-r}^{T} \exp\Big(-2\int_t^T \kappa_u \, du\Big)\, \mathbb{E}\big[\|\Delta b_t(\tilde{X}_t)\|^2\big] \, dt, \tag{19}$$

where $(\tilde{X}_t)_{t\in[T-r,T]}$ denotes the surrogate process. (If κ_t is negative, (19) still holds with an exponential growth factor; in practice this merely worsens the certificate.)

Finally, by the triangle inequality and $(a+b)^2 \le a^2 + 2ab + b^2$, we obtain

$$E(r) = W_2^2(p_{\text{data}}, q_r) \;\le\; W_2^2(p_{\text{data}}, \tilde{q}_r) + 2\, W_2(p_{\text{data}}, \tilde{q}_r)\, W_2(\tilde{q}_r, q_r) + W_2^2(\tilde{q}_r, q_r). \tag{20}$$

We therefore define the nonlinearity remainder by upper bounding the last two terms in (20) using (19) and the baseline proxy $\widehat{E}(r) \approx W_2^2(p_{\text{data}}, \tilde{q}_r)$:

$$\text{Nonlin}(r) \;:=\; 2\sqrt{\widehat{E}(r)}\cdot \mathsf{D}(r) \;+\; \mathsf{D}(r)^2, \qquad \mathsf{D}(r)^2 \;\ge\; W_2^2(q_r, \tilde{q}_r),$$

where $\mathsf{D}(r)$ is any computable upper bound on $W_2(q_r, \tilde{q}_r)$.

Estimating the certificate from validation probes. It remains to upper bound the integrand in (19) without running the reverse sampler. We introduce the pointwise *nonlinearity residual*

$$\eta_t(x) \;:=\; \|s_\theta(x,t) - \tilde{s}_t(x)\|, \qquad \text{so that} \qquad \|\Delta b_t(x)\| \;\le\; \alpha_t\, \eta_t(x).$$

We estimate moments of η_t on the forward noised validation distribution. Specifically, for each $t \in \mathcal{T}$ we draw $x_i \sim p_{\text{data}}$ from held-out data and $w_{ij} \sim \mathcal{N}(0, I)$, form $y_{ij,t} := x_i + \sigma_t w_{ij}$, and compute

$$\widehat{\eta}_{ij,t} \;:=\; \big\|s_\theta(y_{ij,t}, t) - \tilde{s}_t(y_{ij,t})\big\|.$$

Then

$$\widehat{M}_t^{(2)} \;:=\; \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} \widehat{\eta}_{ij,t}^2$$

is an unbiased (and, under sub-Gaussian tails, concentrated) estimator of $\mathbb{E}_{Y\sim p_t}[\eta_t(Y)^2]$, where p_t is the forward-corrupted distribution at noise level σ_t.

To connect this forward expectation to the surrogate-path expectation in (19), we impose a localization condition stating that the surrogate reverse process does not leave the region where the validation probes are representative. A convenient sufficient form is an absolute-continuity domination inequality: there exists $\Lambda_t \geq 1$ such that

$$\mathbb{E}\big[\eta_t(\tilde{X}_t)^2\big] \;\leq\; \Lambda_t \; \mathbb{E}_{Y\sim p_t}\big[\eta_t(Y)^2\big] \qquad \text{for } t \in [T-r, T],$$

which can be justified under log-concavity/contractivity or verified empirically by monitoring the score norm and denoiser residual norms on the validation distribution.

Combining these ingredients, we obtain the computable bound

$$\mathsf{D}(r)^2 \;:=\; \int_{T-r}^{T} \exp\Big(-2\int_{t}^{T} \kappa_u \, du\Big)\, \alpha_t^2 \, \Lambda_t \, \widehat{M}_t^{(2)} \, dt,$$

(with a Riemann-sum discretization over $t \in \mathcal{T}$), and hence the certified inequality $E(r) \leq \widehat{E}(r) + \text{Nonlin}(r)$. In summary, the same forward-corruption probes used to estimate $\widehat{V}_t, \widehat{W}_t$ also yield a quantitative *nonlinearity certificate* via the fitted-score residual η_t; the only additional sampler-dependent inputs are the scalar factors α_t and a stability profile κ_t (which may be chosen conservatively, e.g. via an upper bound on the Jacobian norm of the drift estimated by randomized directional finite differences on the same validation probes).

5.6 Complexity and lower bounds

We separate the cost of the predictor into (i) estimating spectral information about the data covariance (or its compressed surrogates) and (ii) estimating the second moments of the score-error surrogates from validation residuals. Both are unavoidable, in the sense that they match natural statistical lower bounds inherited from classical covariance/trace estimation problems.

Computational complexity of the no-sampling predictor. Fix a probe set of times $\mathcal{T} \subset [0, T]$, a grid of stopping gaps $\mathcal{R}$, and m noise draws per held-out point and per probed time. The dominant cost is the collection of residual statistics. For each $t \in \mathcal{T}$ and each held-out x_i we form m corrupted inputs $y_{ij,t} = x_i + \sigma_t w_{ij}$ and evaluate the denoiser/score to compute residual proxies. Thus the number of network evaluations is $n\, m\, |\mathcal{T}|$ (up to constant factors depending on whether we evaluate both D_{σ_t} and $s_\theta(\cdot, t)$). Kernel assembly is negligible by comparison: once spectral summaries and per-time error summaries are available, the curve evaluation for all $r \in \mathcal{R}$ reduces to deterministic algebra whose complexity depends on the representation. In a full-eigendecomposition regime (small d), this is $O(|\mathcal{R}|\, d^2)$; in a bandpower regime with B bands, it is $O(|\mathcal{R}|\, B^2)$ (or $O(|\mathcal{R}|\, B)$ if the kernel depends only on bandwise diagonal statistics).

Spectrum estimation is the other nontrivial cost. If we work in an orthogonal basis with fast transforms (e.g. Fourier), and define Π_b as the projector onto band b, then bandpowers $\widehat{P}_b = (1/n) \sum_i \|\Pi_b x_i\|^2$ cost $O(n\, d \log d)$ with FFT-type primitives. More generally, randomized trace estimators (Hutchinson/Hutch++) estimate functionals $\operatorname{tr}(f(C))$ at cost $O(s)$ matrix-vector products with the empirical covariance; implemented in streaming form, this becomes $O(s\, n\, d)$ arithmetic with small s, avoiding $d \times d$ storage. In all cases, memory is dominated by the activation footprint of the denoiser evaluations; the predictor-side state is $O(|\mathcal{T}|\, B)$ (or $O(|\mathcal{T}|\, d)$ in diagonal form).

Statistical complexity: how n and m enter. The predictor depends on two classes of random inputs: the held-out data $\{x_i\}_{i=1}^n$ (governing the spectrum estimate) and the injected noises $\{w_{ij}\}$ (governing residual estimates). Under sub-Gaussian assumptions (satisfied in the Gaussian-linear regime), the relevant spectral/bandpower estimates concentrate at rate $O(\sqrt{B/n})$ (or $O(\sqrt{d/n})$ without compression), while residual second-moment estimates concentrate at rate $O(\sqrt{P/(nm)})$, where P is the number of retained parameters per time slice (e.g. $P = d$ for a diagonal $\widehat{V}_t$, $P = B$ for bandwise $\widehat{V}_t$, and potentially larger if cross-band or low-rank structure is modeled). A union bound over $|\mathcal{T}|$ probed times and $|\mathcal{R}|$ stopping points introduces only logarithmic factors, yielding sample requirements of the schematic form

$$n \gtrsim \frac{\mathrm{SpecDim} + \log(|\mathcal{R}||\mathcal{T}|/\delta)}{\varepsilon^2}, \qquad nm \gtrsim \frac{\mathrm{ErrDim} + \log(|\mathcal{R}||\mathcal{T}|/\delta)}{\varepsilon^2},$$

where $\mathrm{SpecDim} \in \{d, B\}$ and ErrDim is the effective parameter count needed to represent the score-error moments at the fidelity demanded by the kernel. This separation is operationally useful: when denoiser calls are expensive, one may increase n and keep m small (even $m = 1$) while still driving down the dominant variance term, whereas if the held-out set is small one may partially compensate by increasing m to reduce residual-noise variance.

Lower bounds via covariance and trace estimation reductions. The preceding rates are not artifacts of our analysis: in the Gaussian-linear regime they are near-tight up to logarithmic factors, because predicting $E(r)$ uniformly over r contains covariance estimation as a special case.

We sketch a standard two-point reduction. Consider two data distributions $p_0 = \mathcal{N}(0, C_0)$ and $p_1 = \mathcal{N}(0, C_1)$ with

$$C_0 = I_d, \qquad C_1 = I_d + \tau\, uu^\top,$$

for a fixed unit vector u. For many sampler/kernel specifications, the kernelized component of $E(r)$ depends on C through spectral functionals of the form $\sum_{i=1}^{d} f_r(\lambda_i)$ or, in bandpower form, $\sum_{b=1}^{B} f_{r,b}(P_b)$. In the rank-one spiked construction above, the functional gap is typically linear in τ for small τ, i.e. there exists r (or a finite subset of r values) such that

$$|E_1(r) - E_0(r)| \;\geq\; c\,\tau$$

for a constant c depending on the schedule and kernel family. Choosing $\tau = \Theta(\varepsilon)$ makes the $E(r)$-gap of order ε. On the other hand, the Kullback–Leibler divergence between n i.i.d. samples satisfies

$$\mathrm{KL}\big(p_0^{\otimes n} \,\|\, p_1^{\otimes n}\big) = \frac{n}{2}\Big(\mathrm{tr}(C_1^{-1}C_0) - d - \log\det(C_1^{-1}C_0)\Big) = \Theta(n\,\tau^2),$$

so that taking $\tau = \Theta(\varepsilon)$ keeps the divergence $O(n\varepsilon^2)$. By Le Cam's method, any procedure that estimates $E(r)$ to additive error $o(\varepsilon)$ with constant success probability would distinguish p_0 from p_1, which requires $\mathrm{KL} = \Omega(1)$ and hence $n = \Omega(1/\varepsilon^2)$. To recover the dimension dependence, we randomize u over an orthonormal set (or use a packing of rank-one perturbations) so that distinguishing among d alternatives requires $n = \Omega(d/\varepsilon^2)$; this is the classical covariance-estimation lower bound transported through the kernel functional. Under a B-band model where the predictor only accesses bandpowers, the same argument with perturbations supported within a band yields $n = \Omega(B/\varepsilon^2)$.

A parallel lower bound applies to residual-statistic estimation: if the kernel includes a term of the form $\int \langle K_t, \mathrm{Cov}(\delta_t)\rangle dt$ (or its discrete analogue), then estimating $E(r)$ implies estimating at least ErrDim mean-square parameters of δ_t (or linear functionals thereof). By reduction from mean estimation in $\mathbb{R}^{\mathrm{ErrDim}}$ (or from covariance estimation when cross-terms are retained), one obtains $nm = \Omega(\mathrm{ErrDim}/\varepsilon^2)$ for fixed $\mathcal{T}$, again matching the concentration-based upper bounds up to logs.

These lower bounds clarify what can and cannot be improved: substantial gains are only possible by exploiting structure (small B, spectral decay, low-rank residual covariances, or smoothness across t) and by careful implementation so that the constants in the denoiser-evaluation budget are controlled.

5.7 Implementation details

Scalable spectrum estimation. The kernel assembly step requires access to spectral functionals of the data covariance C (or of its noised variants implicit in the sampler). Since C is not formed explicitly at image scale, we implement two interchangeable estimators whose outputs match the sufficient statistics assumed by the chosen kernel approximation.

First, in a *bandpower* regime, we fix an orthogonal transform F admitting a fast multiply (typically FFT/DCT or a learned orthobasis with fast application) and define bands $\{\Pi_b\}_{b=1}^{B}$ as disjoint coordinate blocks in the transform domain. Writing $z_i = Fx_i$, we estimate $P_b = \mathbb{E}\|\Pi_b z\|_2^2$ via the streaming average

$$\widehat{P}_b \;=\; \frac{1}{n}\sum_{i=1}^{n} \|\Pi_b F(x_i - \widehat{\mu})\|_2^2,$$

where $\widehat{\mu}$ is the held-out mean (or the training-set mean if fixed preprocessing is used). This estimator is $O(n\,d\log d)$ with an FFT and does not require storing z_i. If the kernel uses additional spectral moments (e.g. $\sum_i f(\lambda_i)$ for a small family of f), we can also estimate bandwise higher moments $\widehat{M}_{b,k} = (1/n)\sum_i \|\Pi_b z_i\|_2^{2k}$ when the approximation calls for kurtosis-like corrections, though in our baseline predictor we restrict to second moments for robustness.

Second, in a *randomized trace* regime, we estimate quantities of the form $\mathrm{tr}(f(C))$ without diagonalizing C. We use Hutch++: for i.i.d. $g_\ell \sim \mathcal{N}(0, I)$ we estimate $\mathrm{tr}(A)$ by $\frac{1}{s}\sum_\ell g_\ell^\top A g_\ell$ and reduce variance by a low-rank range finder on A; here $A = f(\widehat{C})$ for $\widehat{C}$ the empirical covariance operator. Crucially, we never materialize $\widehat{C}$; rather, we implement $v \mapsto \widehat{C}v$ as a single pass over the held-out set:

$$\widehat{C}v \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{\mu})\,\langle x_i - \widehat{\mu}, v\rangle.$$

When f is a polynomial or a rational approximation of the kernel-required map (as is typical when the kernel depends on $(C + \alpha I)^{-1}$ or similar), we apply $f(\widehat{C})$ through repeated calls to this linear operator, using Chebyshev polynomials or conjugate gradients with preconditioning. This approach is basis-agnostic and reduces to $O(s\,n\,d)$ arithmetic with small s.

Residual-statistic collection as a training-time callback. To avoid a separate post-hoc pass, we implement residual-statistic estimation as an optional validation callback that runs every K training steps. For each probed time $t \in \mathcal{T}$ and each minibatch $\{x_i\}$ from the held-out loader, we draw noises $w_{ij} \sim \mathcal{N}(0, I)$ and evaluate either the score $s_\theta(\cdot, t)$ or the denoiser $D_{\sigma_t}(\cdot; \theta)$ to form the DSM proxy

$$r_t(x_i, w_{ij}) \;=\; \frac{x_i - D_{\sigma_t}(x_i + \sigma_t w_{ij}; \theta)}{\sigma_t^2},$$

which equals $\nabla \log p_t(\cdot)$ in the idealized Gaussian-linear calibration. We then aggregate second moments needed for $\widehat{V}_t$ and $\widehat{W}_t$ in the same representation as the spectrum estimator: either diagonal-in-pixel, diagonal-in-transform-band, or a small set of random projections. Concretely, for bandpower residuals we store

$$\widehat{v}_{t,b} \approx \mathbb{E}\big\|\Pi_b F(\delta_t)\big\|_2^2, \qquad \widehat{w}_{t,b} \approx \mathbb{E}\big\|\Pi_b F(\Delta_t)\big\|_F^2,$$

where the second quantity is implemented via Jacobian-vector products or, more simply, via linear regression of the residual on x in the chosen basis when we restrict to a diagonal (per-band) Δ_t. We update these statistics with numerically stable streaming formulas (Welford updates for means and second moments), and we store the effective sample counts per (t, b) to expose variance diagnostics.

Numerical stability and variance reduction. Several implementation choices materially affect stability. (i) We explicitly center data by $\widehat{\mu}$ when estimating $\widehat{P}_b$, and we treat preprocessing (e.g. scaling to $[-1, 1]$) as part of the definition of C. (ii) For small σ_t, the proxy r_t can have large magnitude; we therefore compute in `float64` for the accumulation even if the network runs in `float16/float32`, and we optionally clip r_t only for diagnostic plotting, not for the estimator (since clipping introduces bias). (iii) We use antithetic noise pairs w and $-w$ to reduce odd-moment fluctuations in residual estimates without changing expectations. (iv) When the kernel assembly requires integrals over t, we precompute quadrature weights consistent with the solver discretization and use the same $\mathcal{T}$ grid to avoid interpolation artifacts.

Calibration from $\widehat{E}(r)$ to practical metrics. Although W_2^2 is mathematically convenient, practitioners often report FID or sliced-W_2. We treat calibration as a *monotone* post-processing map g applied to the predicted curve. Specifically, we run a small number of short sampling experiments at a few gaps $\{r_\ell\}$ (far fewer than would be required to map the entire curve) to obtain empirical pairs $(\widehat{E}(r_\ell), \mathrm{FID}(r_\ell))$ or $(\widehat{E}(r_\ell), \mathrm{sW}_2(r_\ell))$, and we fit g via isotonic regression or a low-degree spline constrained to be increasing. The output $g(\widehat{E}(r))$ then provides a calibrated predictor for the chosen metric while preserving the predicted minimizer ordering in typical regimes; we emphasize that this calibration is optional and separate from the no-sampling guarantee for $\widehat{E}(r)$ itself.

Common failure modes and diagnostics. The predictor can fail in structured ways. The most important is model mismatch: if the score error is strongly nonlinear so that $e_t(x) \approx -\Delta_t x + \delta_t$ is inaccurate on the validation distribution, the plug-in curve may be over-optimistic; this typically manifests as a systematic underprediction at intermediate r and can be detected by monitoring a nonlinearity certificate such as the residual of the best affine fit of r_t versus x in the chosen basis. A second failure mode is an inadequate spectral basis: if C is far from diagonal in the selected transform, bandpower compression merges incompatible directions and introduces Bias_B, often visible as sensitivity of $\widehat{E}(r)$ to the number of bands B. Third, solver mismatch matters: the kernel $\mathcal{K}_{\mathrm{sampler}}$ must match the actual discretization (SDE vs ODE, predictor–corrector vs Euler), otherwise the discretization term is misspecified. Finally, distribution shift between the held-out set and the sampler's effective training distribution (e.g. due to augmentations or classifier-free guidance) can invalidate both spectrum and residual estimates; we therefore recommend reporting predictor-side uncertainty bars derived from empirical variability across held-out shards and across independent noise seeds.

5.8 Experimental plan

We organize experiments to (i) validate the plug-in kernel predictor in regimes where the ground truth curve $E(r)$ is known or can be estimated to negligible error, (ii) test the predictor end-to-end on standard image diffusion models under realistic resource constraints, and (iii) isolate the contributions of spectral compression, score-error modeling, and solver specification through targeted ablations.

Synthetic Gaussians with exact ground truth. We first work in the setting of Thm. 1 where $p_{\text{data}} = \mathcal{N}(\mu, C)$ and where we can either (a) construct a score model whose error is exactly affine with prescribed (Δ_t, δ_t), or (b) train a small denoiser on Gaussian data and empirically verify that its residuals are well-approximated by an affine map on the validation distribution. We generate C with controlled spectra: (i) isotropic ($C = \alpha I$), (ii) power-law decay $\lambda_i \propto i^{-p}$, and (iii) spiked models with a small number of large eigenvalues. For each case we fix a noise schedule σ_t and a discretization rule (SDE Euler–Maruyama and ODE probability flow) and compute the kernel-defined $E(r)$ either in closed form (when the induced q_r is Gaussian and the dynamics are linear) or to numerical precision by simulating the resulting linear Gaussian state-space model (which yields exact mean/covariance recursions). Since W_2^2 between Gaussians is explicit,

$$W_2^2\big(\mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}(\mu_2, \Sigma_2)\big) = \|\mu_1 - \mu_2\|_2^2 + \operatorname{tr}\Big(\Sigma_1 + \Sigma_2 - 2(\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2})^{1/2}\Big),$$

this yields a ground-truth curve $E(r)$ without Monte Carlo error. We then run KernelPredict using only held-out samples and denoiser queries, producing $\widehat{E}(r)$ on a grid $\mathcal{R}$ (log-spaced in r to emphasize small terminal gaps). The primary evaluation is the uniform deviation $\sup_{r\in\mathcal{R}} |\widehat{E}(r) - E(r)|$ as a function of (n, m), compared to the $\tilde{O}(1/\sqrt{n})$ and $\tilde{O}(1/\sqrt{m})$ scaling suggested by Thm. 2. We also report stopping-time regret $E(\hat{r}^*) - \min_{r\in\mathcal{R}} E(r)$ to validate the argmin stability in Thm. 3.

Mixtures and controlled non-Gaussianity. To probe robustness beyond the Gaussian-linear regime while retaining tractable evaluation, we consider mixtures of Gaussians with known parameters, including symmetric two-component mixtures and higher-K mixtures with separated means. In these settings $W_2^2(p_{\text{data}}, q_r)$ is no longer closed-form even if q_r were known, so we estimate $E(r)$ by a high-accuracy offline procedure used *only* for evaluation: for each r we generate a large batch from the sampler (with many independent chains), and we approximate W_2^2 via sliced Wasserstein (many random projections) or via entropic OT on a moderate subsample. We treat this as ground truth for experimental purposes and compare it to $\widehat{E}(r)$ and to the certified upper bound interpretation $E(r) \leq \widehat{E}(r) + \text{Nonlin}(r)$ by instantiating a nonlinearity certificate based on the residual of the best affine fit of r_t versus x (measured in the same basis used by the predictor). We expect systematic underprediction when mixture separation induces strong curvature in $\nabla \log p_t$ at intermediate t, and we quantify whether $\text{Nonlin}(r)$ correctly tracks this deviation.

Real-image diffusion: predicted versus measured curves. We next evaluate on medium-scale image diffusion models where sampling is expensive but still feasible for limited sweeps. We select a standard pretrained model (e.g. on CIFAR-10 or ImageNet-64) and fix a reference sampler (SDE Euler, probability-flow ODE, and optionally a predictor–corrector variant). We compute $\widehat{E}(r)$ using a held-out set (e.g. $n = 10^4$ images), a modest time grid $|\mathcal{T}| \in [32, 128]$, and a small number of noises per time $m \in [1, 8]$, reporting runtime and memory overhead. For evaluation we perform a *sparse* sampling sweep: for a small subset of gaps $\{r_\ell\} \subset \mathcal{R}$ we run the actual sampler to obtain empirical metrics, including (i) a proxy for W_2^2 via sliced-W_2 in a fixed feature space and (ii) FID. We then assess (a) rank consistency of gaps (Kendall τ between $\widehat{E}(r)$ and measured error), (b) accuracy of the predicted minimizer $\hat{r}^*$ relative to the measured best r, and (c) whether the curve shape (monotonicity at extremes, presence of an interior optimum) is correctly recovered.

Ablations: basis choice, correlation modeling, solver mismatch. We perform three ablation families designed to stress each modeling decision.

Basis choice. We compare (i) pixel-diagonal statistics, (ii) Fourier/DCT bandpowers with varying B, and (iii) randomized projections (Hutch++-style traces) holding compute fixed. We report sensitivity of $\widehat{E}(r)$ and $\hat{r}^*$ to B and to the transform, thereby empirically characterizing Bias_B.

Correlation modeling. Our baseline treats per-time residual statistics independently. Since a single network induces correlated errors across t, we compare this baseline to a correlated-time model in which we estimate cross-time covariances on a coarse subset of $\mathcal{T}$ and interpolate, and we evaluate whether incorporating these correlations improves prediction of the *shape* of $\widehat{E}(r)$ (particularly around the optimum).

Solver choice. We deliberately assemble kernels using an incorrect solver (e.g. ODE kernel for an SDE sampler) and quantify the resulting bias. This isolates discretization-model mismatch as a distinct failure mode and motivates solver-aware kernel libraries.

Across all experiments we fix seeds and report uncertainty bars from held-out sharding and noise resampling, emphasizing that the predictor is itself a statistical estimator whose variability should be visible to the user.

6 Discussion and Future Work

Our formulation treats the predicted curve $r \mapsto \widehat{E}(r)$ as a plug-in evaluation of an explicit sampler-dependent kernel functional, driven by two classes of inputs: spectral summaries of the held-out data and second-moment statistics of the model error as revealed by denoising residuals. The experiments above are designed to validate the estimator under controlled regimes; here we discuss extensions which, in our view, are structurally natural within the same “no-reverse-sampling” access model.

Correlated-time kernels induced by a shared network. The baseline predictor estimates per-time covariances $\mathrm{Cov}(\delta_t)$ and $\mathrm{Cov}(\mathrm{vec}(\Delta_t))$ independently across $t \in \mathcal{T}$. This is statistically convenient, but it is also a modeling choice: a single parameter vector θ induces strongly coupled errors across time, and the resulting dynamics accumulate these errors through time-integrated propagators. In the Gaussian-linear regime one can make this dependence explicit. Writing schematically the linearized drift error along the reverse dynamics as

$$e_t(x) \approx -\Delta_t x + \delta_t,$$

the contribution of score error to the terminal mean/covariance (hence to W_2^2) depends not only on $\mathbb{E}[\delta_t \delta_t^\top]$ and $\mathbb{E}[\Delta_t \otimes \Delta_t]$ but also on cross-time second moments such as $\mathbb{E}[\delta_t \delta_{t'}^\top]$ and $\mathbb{E}[\Delta_t \otimes \Delta_{t'}]$ for $t \neq t'$. A correlated-time refinement therefore replaces the diagonal-in-time summary

$$\{\widehat{V}_t, \widehat{W}_t\}_{t \in \mathcal{T}} \quad \text{by} \quad \{\widehat{V}_{t,t'}, \widehat{W}_{t,t'}\}_{t,t' \in \mathcal{T}},$$

with $\widehat{V}_{t,t'} \approx \mathrm{Cov}(\delta_t, \delta_{t'})$ and an analogous definition for $\widehat{W}_{t,t'}$. The resulting kernel assembly becomes a quadratic form in these matrices, reflecting the fact that the terminal error is obtained by integrating (or summing, in discrete time) propagated perturbations. Practically, full estimation of $\widehat{V}_{t,t'}$ at fine $|\mathcal{T}|$ is expensive, but the structure of shared networks suggests compressions: (i) low-rank models in the time index (e.g. a small number of temporal factors), (ii) coarse-to-fine schemes estimating correlations on a sparse skeleton of times and interpolating, or (iii) parametric covariance models (e.g.

Matérn/Gaussian kernels in t) fit to empirical cross-covariances. Any such approach yields a principled bias–variance trade-off: the uncorrelated baseline corresponds to enforcing $\widehat{V}_{t,t'} = 0$ for $t \neq t'$, whereas correlated-time kernels aim to reduce systematic misprediction of the *shape* of $\widehat{E}(r)$ near its minimizer when temporal error correlations dominate.

Predicting the effect of distillation and step reduction. Diffusion distillation methods (progressive distillation, consistency models, and related student–teacher procedures) primarily alter the effective solver: they reduce the number of steps, modify the time grid, and sometimes replace the reverse-time dynamics by an alternative parameterization. Our predictor is, by design, solver-aware through $\mathcal{K}_{\text{sampler}}$ and Γ. This suggests a direct use case: given a candidate distilled sampler specified by (σ_t, Γ) and a fixed trained network (teacher or student), we can compute $\widehat{E}(r)$ for that sampler without generating samples. When distillation changes the network itself, we can still use the same pipeline by recomputing residual statistics on held-out data for the student. This yields a mechanism for selecting, among a family of distilled variants, a recommended terminal gap $\hat{r}^*$ and step schedule before any expensive visual evaluation. Conceptually, the predictor separates two effects that are often conflated empirically: the discretization effect E^{disc} driven by step size and solver choice, and the score-error effect driven by model mismatch, which distillation may either improve (by training objective alignment) or worsen (by reducing capacity).

Guidance as a controllable perturbation of the score. Classifier guidance and classifier-free guidance replace the base score $s_\theta(x,t)$ by a guided score $s_\theta^{(g)}(x,t)$, typically of the form

$$s_\theta^{(g)}(x,t) = s_\theta(x,t) + g \cdot a(x,t),$$

where a is a guidance term (e.g. $\nabla_x \log p(y \mid x, t)$ or the conditional–unconditional difference in classifier-free guidance) and $g \geq 0$ is a guidance scale. Even if s_θ is well-calibrated, guidance alters both the drift magnitude and the effective error statistics, often causing stiffness near small σ_t. Our estimator can be adapted by measuring residual proxies for the *guided* score directly on corrupted held-out inputs, thereby obtaining $\widehat{V}_t^{(g)}, \widehat{W}_t^{(g)}$ as functions of g. This would enable predicting an error curve $\widehat{E}_g(r)$ that can be optimized jointly over (g,r), and potentially traded against perceptual metrics or conditional accuracy when a differentiable surrogate is available. From a robustness perspective, guidance is also a stress test for the non-Gaussian remainder: large g may push trajectories into regions where the linearization certificate η_t (cf. Thm. 4) is large, and the upper-bound interpretation $E(r) \leq \widehat{E}(r) + \text{Nonlin}(r)$ becomes operational.

Integration into automated schedulers and model-selection loops. Finally, we view $\widehat{E}(r)$ as a primitive for automated configuration. Given a fixed network, one may wish to choose: (i) a stopping rule r, (ii) a step schedule Γ (including adaptive step sizes), and (iii) a solver family (SDE vs. ODE vs. predictor–corrector). Since our estimator is fast relative to sampling, we can place it inside an outer-loop optimizer that searches over discrete design choices. A minimal version performs a grid search over candidate schedules and selects the minimizer of $\widehat{E}$; more ambitious variants exploit the smooth dependence of the kernel on Γ to perform continuous optimization under constraints (e.g. a fixed budget of function evaluations). Importantly, because the predictor is itself statistical, a scheduler should propagate uncertainty: if $\widehat{E}$ is flat within confidence bands over a range of r, one should prefer conservative choices or incorporate secondary objectives (runtime, stability) rather than overfit to estimator noise. We expect this "predict–then-optimize" perspective to be most effective when combined with compressed spectral representations (bandpowers or randomized traces), yielding a practical tool that can be run routinely during training checkpoints to monitor sampler performance without repeated sampling sweeps.