Correlated-Time Kernel Expansions for End-to-End Wasserstein Error in Diffusion Models

Liz Lemma Future Detective

January 20, 2026

Abstract

Score-based diffusion models are trained with a single shared network across noise levels, inducing strong correlations in score error across time—an effect not captured by analyses that assume independent per-time errors. Building on recent fine-grained Gaussian-linear error decompositions for (i) constant-step SGD denoising score matching and (ii) diffusion/Langevin sampling in Wasserstein distance, we derive a second-order, sampler-specific expansion of the end-to-end sampling error for discretized reverse diffusion. In the Gaussian setting, we show that the expected W_2^2 error decomposes into (a) a truncation/terminal-time term, (b) a discretization term, and (c) a new *correlated-time quadratic form* depending on the full time–time covariance of the score error process induced by shared-parameter training. This quadratic term takes the form of a double-integral (or double-sum on the discretization grid) against an explicit kernel operator determined by the diffusion schedule and the chosen solver, reducing to earlier single-integral formulas when errors are independent across time. We provide tightness statements (matching lower bounds) for Gaussian error processes, and propose diagnostics to estimate the relevant covariance structure from training-time residuals. Experiments on synthetic anisotropic Gaussians with controlled temporal correlations validate the kernel prediction and show how correlation structure explains which noise bands dominate quality.

Table of Contents

3. 3. Problem formulation: shared-network score error as a stochastic process in time; define the target quantity $\mathbb{E}W_2^2(p_{\text{data}}, q_k)$ and the object to characterize (time–time covariance functional).

4. 4. Linear-Gaussian reverse dynamics with correlated affine score perturbations: explicit propagation of mean/covariance under correlated perturbations; stability and well-posedness conditions.

5. 5. Main theorem: correlated-time kernel expansion of $\mathbb{E}W_2^2$; explicit kernels for Euler–Maruyama reverse SDE and for probability-flow ODE; reduction to the independent-time case.

6. 6. Tightness and lower bounds: show the quadratic functional is unavoidable by constructing Gaussian score-error processes matching any prescribed time–time covariance; discuss identifiability/sample complexity of estimating correlations.

7. 7. Practical estimators and diagnostics: how to estimate (approximations of) $\mathrm{Cov}(e_s, e_{s'})$ from denoising residuals and how to compute the kernel numerically; optional implications for schedule selection.

8. 8. Experiments: synthetic anisotropic Gaussians and shared random-feature scores; controlled correlation ablations; kernel-predicted error vs measured W_2; solver comparison; discuss where remainders become visible.

9. 9. Discussion and extensions: toward non-Gaussian/local linearization, modern solvers, and linking to optimizer dynamics (SGD/Adam) for predicting covariance structure.

10. 10. Limitations and open problems: beyond Gaussianity, non-asymptotic control of remainders, and bridging to large-scale image diffusion.

1 Introduction and motivation

Diffusion generative models are commonly justified through an idealized statement: if we sample the reverse-time dynamics using the *exact* score $\nabla \log p_t$, then the terminal distribution matches the data distribution (up to terminal-time truncation and numerical discretization). In practice, however, sampling uses a single learned network queried at many times, and the induced score error is neither small nor independent across time. The present work isolates this latter point. We regard the score error as a stochastic process $\{e_t\}_{t\in[0,T]}$ and argue that its *time–time correlation structure* is a missing piece in end-to-end error propagation theory.

The usual diagnostics for score accuracy are time-marginal: one reports a denoising score-matching loss, a per-time mean-squared error, or bounds involving $\sup_t \|e_t\|$ and its integral. These quantities cannot distinguish between (i) small errors that are rapidly decorrelated in time and therefore partially cancel, and (ii) errors of the same marginal magnitude that are coherently aligned across time and therefore accumulate. The sampling map from the entire trajectory $\{e_t\}$ to the terminal sample is intrinsically path-dependent. When the same network is evaluated at different t, architectural biases and shared features naturally induce correlations between e_s and $e_{s'}$, even when the latent randomness driving the sampler is independent. Consequently, controlling only $\mathbb{E}\|e_t\|^2$ at each t is insufficient to predict sampling quality in metrics sensitive to mean and covariance, such as W_2.

Our starting point is that, for Gaussian targets and the linear reverse-time dynamics induced by common diffusion schedules, the influence of a perturbation of the score on the terminal distribution admits a linear sensitivity representation. Concretely, when p_{data} is Gaussian, the exact reverse dynamics preserves Gaussianity, and the sampler output is characterized by its mean and covariance. If we write the learned score as a perturbation of the true score, then to leading order the induced perturbations of terminal mean and covariance are linear functionals of the entire error trajectory. This observation converts the analysis of sampling error into an operator calculus: we propagate the error process through linear "state transition" operators determined by the schedule and by the chosen discretization scheme.

The metric we target is the squared Wasserstein-2 distance. For Gaussians, W_2^2 decomposes into a mean term and a covariance term, the latter being the Bures squared distance. Hence, once we express the terminal mean and covariance as perturbations around their unperturbed values, a second-order Taylor expansion of the Gaussian W_2^2 yields a quadratic form in these perturbations. Combining the linear sensitivity representation with this second-order expansion implies a structural consequence: the leading score-error contribution to $\mathbb{E}W_2^2$ depends only on *second moments* of the score error process, i.e. on the full time–time covariance $\mathrm{Cov}(e_s, e_{s'})$. In particular, the relevant object is not a single-time variance but a two-time

kernel, and the expected W_2^2 error can be written as a bilinear functional of $\mathrm{Cov}(e_s, e_{s'})$.

To make this dependence explicit, we model the score error in the affine form

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

where Δ_t and δ_t are jointly zero-mean random coefficients that may be correlated across time. This form is not merely a convenient abstraction: for Gaussian marginals, the true score is affine in x, and the leading approximation error of a learned affine map is naturally expressed through a random matrix and vector. Under this parametrization, the correlated-time effect separates into a matrix-error part (driving covariance perturbations and, through coupling, possibly mean perturbations) and a vector-error part (directly driving mean perturbations). The resulting quadratic term admits a decomposition into two bilinear forms, one involving $\mathrm{Cov}(\Delta_s, \Delta_{s'})$ and one involving $\mathrm{Cov}(\delta_s, \delta_{s'})$, together with cross terms when present. The associated kernels are explicit in the Gaussian-linear setting and are computable from schedule-dependent propagation operators and the local Hessian of the Bures metric.

This perspective generalizes and clarifies prior analyses that yield "kernel norms" or time-weighted integrals of per-time score errors. Such results implicitly impose a diagonal-in-time approximation, effectively replacing $\mathrm{Cov}(e_s, e_{s'})$ by its time-diagonal. When errors are independent across time (or sufficiently mixing so that off-diagonal correlations are negligible), our bilinear form reduces to a single integral with a one-time weight, recovering the familiar structure. The point of our formulation is that this reduction is *not* valid in general, and the discrepancy is not a higher-order effect: two error processes can share identical time-marginal variances and yet produce different expected sampling errors because the kernel sees their cross-time covariance.

Our contributions are therefore organized around an explicit decomposition of the expected sampling error into three terms: (i) a terminal-time bias due to truncating the reverse dynamics at time $r > 0$; (ii) a solver-dependent discretization bias controlled by the step size γ and the numerical order; and (iii) a correlated-time quadratic term of order $\Theta(\varepsilon^2)$, where ε controls the magnitude of score errors in second moment. The third term is the main novelty: it is invariant to solver order in the sense that improving discretization does not remove the effect of correlated score errors, but only modifies the sensitivity operators appearing in the kernel. Moreover, within the Gaussian framework, the quadratic term is tight: without additional structure on the error process, the dependence on the full time–time covariance cannot be reduced to time-marginal quantities.

The kernel representation also suggests a practical diagnostic. If we can estimate $\mathrm{Cov}(\Delta_{t_a}, \Delta_{t_b})$ and $\mathrm{Cov}(\delta_{t_a}, \delta_{t_b})$ on a discrete time grid from residual

samples of the learned network, then we can contract these empirical covariances with the precomputed kernels to predict the expected W_2^2 degradation of a given sampler. This yields a tool for comparing samplers, schedules, and noise injection parameters α under a common learned model, and for deciding whether improvements in sampling error should be sought through finer discretization (reducing the discretization term) or through training interventions that reduce cross-time coherence (reducing the quadratic term). At the same time, the necessity of time–time covariance brings an intrinsic cost: estimating an unrestricted covariance matrix on an m-point grid is quadratically expensive in m unless one exploits structure such as low-rank or banded correlations. This observation motivates the structural approximations we later discuss.

Finally, we emphasize scope. We work in a regime where the forward marginals and the reverse dynamics are Gaussian-linear, so that the propagation operators and the Bures expansion are tractable and the resulting kernel can be written explicitly. While this setting is restrictive, it isolates the role of time correlation without conflating it with nonlinear score geometry. The qualitative message extends beyond Gaussians: any end-to-end sampling analysis that compresses score error into independent-time statistics will, in general, fail to capture coherent accumulation across time. The remainder of the paper develops the required preliminaries, establishes the kernelized expansion, and relates it to estimation and computation on a time grid.

2 Background and preliminaries

2.1 Forward diffusion and Gaussian marginals

We work with a linear forward noising process on $\mathbb{R}^d$ specified by a schedule (β_t, ξ_t) on $t \in [0, T]$. Concretely, we consider the (time-inhomogeneous) SDE

$$dX_t \;=\; -\frac{1}{2}\beta_t X_t\, dt \;+\; \xi_t\, dW_t, \tag{1}$$

where W_t is a standard d-dimensional Brownian motion and $\xi_t \geq 0$ is interpreted as an isotropic diffusion magnitude. This unified form contains common parameterizations as special cases (e.g. variance-preserving choices correspond to $\xi_t = \sqrt{\beta_t}$ up to conventional rescalings). Throughout, we assume $X_0 \sim p_{\text{data}} = \mathcal{N}(\mu, C)$ with $C \succ 0$.

Since (1) is linear with additive Gaussian noise, each marginal p_t is Gaussian, $p_t = \mathcal{N}(\mu_t, \Sigma_t)$, with (μ_t, Σ_t) solving deterministic ODEs. In the isotropic-noise setting above one has

$$\dot{\mu}_t \;=\; -\frac{1}{2}\beta_t \mu_t, \qquad \dot{\Sigma}_t \;=\; -\beta_t \Sigma_t \;+\; \xi_t^2 I, \tag{2}$$

with initial conditions $(\mu_0, \Sigma_0) = (\mu, C)$. In particular, the true score is affine:

$$\nabla \log p_t(x) \;=\; -\Sigma_t^{-1}(x - \mu_t). \tag{3}$$

This affine structure motivates the error model used later.

2.2 Reverse-time dynamics and probability-flow limit

Sampling requires integrating a reverse-time dynamics from a tractable terminal law at time T to an approximation of p_{data} at time 0. With the exact score (3), the (formal) reverse-time SDE has drift corrected by the score. For our purposes, it suffices to record a parameterized family interpolating between the stochastic reverse SDE and the deterministic probability-flow ODE. Writing the reverse-time variable as $r = T - t$ (so r increases as we sample), we view the sampler as evolving a state Z_r forward in r with coefficients inherited from (β_t, ξ_t) and score evaluations at the corresponding forward time $t = T - r$. Abstractly, we write the reverse dynamics in the schematic form

$$dZ_r \;=\; a_r(Z_r)\, dr + \alpha\, b_r\, d\bar{W}_r, \qquad a_r(z) \;\propto\; (\text{linear term in } z) - \xi_{T-r}^2\, \widehat{s}_{T-r}(z), \tag{4}$$

where $\bar{W}_r$ is a Brownian motion independent of W_t and $\alpha \geq 0$ controls the injected noise. The case $\alpha > 0$ corresponds to a reverse SDE with stochasticity, while the probability-flow ODE is obtained in the limit $\alpha \to 0$ (removing the stochastic term). Under the exact score $\widehat{s}_t = \nabla \log p_t$, both variants transport the terminal Gaussian marginal back to p_{data} in the idealized continuous-time setting (up to any terminal-time truncation if sampling begins at $r > 0$ rather than $r = 0$).

In the Gaussian-linear regime and for affine scores, the reverse dynamics preserves Gaussianity: for each r the law of Z_r is characterized by a mean and covariance solving linear (matrix) ODEs/SDE moment equations. This reduction is the starting point for an end-to-end analysis in terms of perturbations of these moments.

2.3 Discretization on a time grid

In practice, sampling is performed on a grid $0 = t_0 < t_1 < \cdots < t_K = T$ with step sizes $h_k = t_{k+1} - t_k$ and maximum step size $\gamma := \max_k h_k$. We allow nonuniform grids. A discretization scheme (Euler–Maruyama for SDEs, explicit/implicit Runge–Kutta for ODEs, etc.) defines a one-step map producing iterates whose laws we denote by q_k.

For later use, we emphasize that in the present setting the update can be written as an affine transformation of the current state plus injected Gaussian noise. In a generic notation, a single step can be expressed as

$$x_k \;=\; \mathsf{A}_k x_{k+1} \;+\; \mathsf{b}_k \;+\; \mathsf{G}_k\, \eta_k, \qquad \eta_k \sim \mathcal{N}(0, I), \tag{5}$$

where $\mathsf{A}_k, \mathsf{b}_k, \mathsf{G}_k$ are determined by the schedule, the solver, α, and the score evaluations. When the score is affine, the dependence of b_k on the score error can be made explicit and linearized. Mean-square stability assumptions used later amount to requiring that the products of the linear maps A_k remain uniformly controlled in a manner sufficient to bound the propagation of second moments.

2.4 Denoising score matching and the induced score error

A learned score network is typically trained by denoising score matching (DSM): one samples $t \sim \pi$ on $[0, T]$, draws $x_0 \sim p_{\mathrm{data}}$, generates a noisy x_t from the forward process, and minimizes a weighted regression loss towards the true score $\nabla \log p_t(x_t)$. In its idealized form, DSM targets the population objective

$$\min_{\widehat{s}} \ \mathbb{E}_{t\sim\pi}\, \mathbb{E}_{x_t\sim p_t}\Big[w(t)\, \|\widehat{s}_t(x_t) - \nabla \log p_t(x_t)\|^2\Big], \tag{6}$$

for some weight $w(t)$. The key modeling point in our analysis is that the *same* network is queried across all times, so the residuals $\widehat{s}_t(x) - \nabla \log p_t(x)$ at different times are generally statistically dependent, reflecting shared features and inductive biases rather than independent noise.

Because (3) is affine in x, it is natural (and exact for affine predictors) to parameterize the score error by an affine perturbation

$$e_t(x) \ := \ \widehat{s}_t(x) - \nabla \log p_t(x) \ = \ -(\Delta_t)x + \delta_t, \tag{7}$$

where $\Delta_t \in \mathbb{R}^{d\times d}$ and $\delta_t \in \mathbb{R}^d$ are random coefficients, jointly zero-mean, and possibly correlated across time. The decomposition (7) separates a matrix component that directly perturbs covariance evolution from a vector component that directly perturbs the mean evolution. Later, the dependence of sampling error on *time–time* covariances such as $\mathrm{Cov}(\mathrm{vec}(\Delta_s), \mathrm{vec}(\Delta_{s'}))$ and $\mathrm{Cov}(\delta_s, \delta_{s'})$ will be explicit.

2.5 Wasserstein-2 distance for Gaussians and the Bures metric

Our performance metric is W_2. For Gaussian measures $P = \mathcal{N}(m, \Sigma)$ and $Q = \mathcal{N}(m', \Sigma')$ on $\mathbb{R}^d$, the squared Wasserstein-2 distance has the closed form

$$W_2^2(P, Q) \ = \ \|m - m'\|^2 \ + \ B^2(\Sigma, \Sigma'), \tag{8}$$

where B^2 is the Bures squared distance

$$B^2(\Sigma, \Sigma') \ := \ \mathrm{tr}(\Sigma) + \mathrm{tr}(\Sigma') - 2\,\mathrm{tr}\Big((\Sigma^{1/2}\Sigma'\Sigma^{1/2})^{1/2}\Big). \tag{9}$$

We use the Frobenius pairing $\langle A, B\rangle := \mathrm{tr}(A^\top B)$ and the Euclidean pairing for vectors. The relevance of (8) is that once we represent the sampler output

q_k as a Gaussian (exactly, in our regime) and express its mean/covariance as perturbations around the exact-score baseline, a second-order expansion of (8) reduces expected sampling error to quadratic expressions in those perturbations.

2.6 Operator notation and kernel contractions

We will repeatedly map perturbations injected at intermediate times to terminal perturbations of mean and covariance. On a grid, we denote by $\Phi_{k\leftarrow a}$ a (solver-dependent) linear sensitivity operator that propagates an infinitesimal score perturbation at time index a to its contribution at terminal index k. In the Gaussian-linear setting, $\Phi_{k\leftarrow a}$ can be written in closed form by unrolling the linear recurrence implied by (20). In continuous time, $\Phi_{r\leftarrow s}$ is the corresponding state-transition operator obtained from variation-of-constants.

The second-order expansion of (8) yields a quadratic form at terminal time whose expectation depends only on second moments of the score error process. Concretely, after linearizing terminal mean/covariance perturbations in $\{\Delta_t, \delta_t\}$ and taking expectations, we obtain bilinear contractions of the form

$$\sum_{a,b\geq k} \big\langle K_k^{\Delta}[a,b],\, \mathrm{Cov}(\Delta_a, \Delta_b)\big\rangle \;+\; \sum_{a,b\geq k} \big\langle K_k^{\delta}[a,b],\, \mathrm{Cov}(\delta_a, \delta_b)\big\rangle, \tag{10}$$

and similarly in continuous time as double integrals. Here the kernel blocks K^{Δ}, K^{δ} are determined by composing (i) propagation/sensitivity operators Φ with (ii) the local quadratic structure induced by (8)–(9). This operator viewpoint is the mechanism by which correlated errors across time enter the expected W_2^2 error, and it motivates the problem formulation in the next section, where we treat $\{e_t\}_{t\in[0,T]}$ as a stochastic process whose full time–time covariance is the primary object of interest.

3 Problem formulation: correlated score error as a stochastic process

Our objective is to characterize, for a fixed discretized sampler, how a *shared-network* score approximation propagates to terminal sampling error when its residuals are statistically dependent across time. We therefore elevate the score residual to a time-indexed stochastic process and seek an explicit functional relationship between its *time–time covariance* and the expected terminal Wasserstein error.

3.1 Sampler, terminal index, and the performance target

Fix a discretization scheme (for the reverse SDE with parameter $\alpha > 0$ or for the probability-flow ODE as $\alpha \to 0$) on a grid $0 = t_0 < t_1 < \cdots < t_K = T$ with maximum step size $\gamma := \max_k(t_{k+1} - t_k)$. Denote by q_k the law of the sampler iterate at index k (corresponding to time t_k). We measure sampling quality at step k by

$$\mathcal{E}_k \;:=\; \mathbb{E}\, W_2^2\big(p_{\mathrm{data}},\, q_k\big), \tag{11}$$

where the expectation is taken over all randomness present in the construction of q_k, including the reverse-time noise (if $\alpha > 0$), any randomized components of the numerical scheme, and the randomness induced by the learned score approximation (made precise below). We also use the reverse-time variable $r = T - t$ and write $r_k := T - t_k$ to emphasize that terminal-time effects appear naturally as functions of r rather than t.

When the sampler is run with the *exact* score $\nabla \log p_t$, we denote the corresponding law by $q_k^\star$. Even in this idealized setting, two unavoidable error sources remain: (i) *terminal-time truncation* if the sampler is initialized from a proxy law at time T and/or if we evaluate at $r_k > 0$, and (ii) *discretization bias* induced by the finite grid. Our aim is to isolate these components and then quantify the additional degradation caused by *correlated* score errors.

3.2 Shared-network score error as a time-indexed random element

We model the learned score as

$$\widehat{s}_t(x) \;=\; \nabla \log p_t(x) \;+\; e_t(x), \qquad t \in [0, T], \tag{12}$$

where the residual e_t is viewed as a stochastic process in t. The central modeling point is that a single network, trained once, is queried at all times; thus $\{e_t\}_{t\in[0,T]}$ need not be independent in time and, in general, exhibits nontrivial cross-time dependence.

In the Gaussian setting, the true score is affine, and we restrict to affine residuals,

$$e_t(x) \;=\; -(\Delta_t)x + \delta_t, \tag{13}$$

where $\Delta_t \in \mathbb{R}^{d\times d}$ and $\delta_t \in \mathbb{R}^d$ are random coefficients. We impose the basic centering and small-error conditions

$$\mathbb{E}\Delta_t = 0, \qquad \mathbb{E}\delta_t = 0, \qquad \sup_{t\in[0,T]} \Big(\mathbb{E}\|\Delta_t\|_F^2 + \mathbb{E}\|\delta_t\|^2\Big) \le \varepsilon^2, \tag{14}$$

for a parameter $\varepsilon \to 0$. We allow arbitrary cross-time correlations subject to bounded second moments. In particular, for $s, s' \in [0, T]$ we define the

time–time covariance objects

$$\Gamma^{\Delta}(s,s') := \mathrm{Cov}\big(\mathrm{vec}(\Delta_s),\, \mathrm{vec}(\Delta_{s'})\big), \tag{15}$$

$$\Gamma^{\delta}(s,s') := \mathrm{Cov}\big(\delta_s,\, \delta_{s'}\big), \tag{16}$$

and, when needed, the cross-covariances

$$\Gamma^{\Delta\delta}(s,s') := \mathrm{Cov}\big(\mathrm{vec}(\Delta_s),\, \delta_{s'}\big), \qquad \Gamma^{\delta\Delta}(s,s') := \mathrm{Cov}\big(\delta_s,\, \mathrm{vec}(\Delta_{s'})\big). \tag{17}$$

On a grid $\{t_a\}_{a=0}^{K}$ we use the corresponding block matrices Γ^{Δ}_{ab}, Γ^{δ}_{ab}, etc. The setting (13)–(17) captures both structured approximation error (e.g. systematic under/over-estimation of the score's linear coefficient) and stochastic residuals that may be temporally correlated due to shared features, architectural biases, or correlated training noise.

3.3 The object to characterize: a covariance functional for expected W_2^2

We seek an explicit expansion of (11) in the joint regime of small step size and small score error. Concretely, for each terminal index k we aim to express

$$\mathcal{E}_k \;=\; E^{(0)}_{r_k} \;+\; E^{\mathrm{disc}}_{r_k}(\gamma) \;+\; \mathcal{Q}_{r_k}\big[\Gamma^{\Delta}, \Gamma^{\delta}, \Gamma^{\Delta\delta}\big] \;+\; o(\gamma) \;+\; o(\varepsilon^2), \tag{18}$$

where $E^{(0)}_{r_k}$ depends only on the terminal-time initialization/truncation at reverse time r_k, $E^{\mathrm{disc}}_{r_k}(\gamma)$ depends only on the numerical scheme and schedule (and vanishes as $\gamma \to 0$ at the scheme's order), and $\mathcal{Q}_{r_k}$ is a *quadratic* contribution due to score error. The key requirement is that $\mathcal{Q}_{r_k}$ depend on the score error process only through its *second-order structure*, namely the time–time covariances in (15)–(17), and that it admit an explicit kernel representation.

On a grid, the desired structure is a double sum of bilinear contractions,

$$\mathcal{Q}_{r_k} \;=\; \sum_{a,b\geq k} \big\langle K^{\Delta}_k[a,b],\, \Gamma^{\Delta}_{ab}\big\rangle + \sum_{a,b\geq k} \big\langle K^{\delta}_k[a,b],\, \Gamma^{\delta}_{ab}\big\rangle + \sum_{a,b\geq k} \big\langle K^{\Delta\delta}_k[a,b],\, \Gamma^{\Delta\delta}_{ab}\big\rangle, \tag{19}$$

with an analogous double-integral form in continuous time. The kernel blocks $K^{\Delta}_k[a,b]$, $K^{\delta}_k[a,b]$, and $K^{\Delta\delta}_k[a,b]$ are deterministic objects determined by the schedule (β_t, ξ_t), the reverse-time parameter α, and the chosen discretization through sensitivity/propagation operators. Importantly, (19) reduces to a single sum only under an *independent-time* assumption (diagonal Γ in the time indices), which we do *not* impose.

Finally, we require that the expansion (18) be meaningful under stability: score perturbations injected at intermediate times must not be amplified without control by the discretized dynamics. We therefore assume

mean-square stability conditions sufficient to bound products of the discretized linear maps and to justify interchanging expectation with perturbative expansions. Under these conditions, the next section makes (19) explicit by unrolling the linear-Gaussian reverse dynamics, expressing terminal mean/covariance perturbations as linear functionals of $\{(\Delta_t, \delta_t)\}$, and combining this with the second-order structure of W_2^2 for Gaussians.

4 4. Linear-Gaussian reverse dynamics with correlated affine score perturbations: explicit propagation of mean/covariance under correlated perturbations; stability and well-posedness conditions.

We now make explicit the linear–Gaussian structure of the discretized reverse dynamics under the affine score perturbation (13). Throughout, we write $X_k \in \mathbb{R}^d$ for the sampler state at index k (time t_k), so that the numerical scheme updates $X_{k+1} \mapsto X_k$.

In the Gaussian setting the exact reverse drift is affine in x at each time, hence any one-step discretization of the reverse SDE/ODE can be written (possibly after collecting terms) in the generic form

$$X_k = A_k^\star X_{k+1} + a_k^\star + \mathbf{1}_{\{\alpha>0\}}\, G_k^\star Z_k \qquad (Z_k \sim \mathcal{N}(0, I_d)), \tag{20}$$

when the exact score is used, where $A_k^\star \in \mathbb{R}^{d\times d}$, $a_k^\star \in \mathbb{R}^d$, and $G_k^\star \in \mathbb{R}^{d\times d}$ are deterministic and depend on (β_t, ξ_t), α, and the chosen solver. (For ODE solvers, the noise term is absent.) The key point for our purposes is that (20) is affine-Gaussian, so $q_k^\star$ is Gaussian with mean/covariance obeying the standard recurrences

$$m_k^\star = A_k^\star m_{k+1}^\star + a_k^\star, \qquad \Sigma_k^\star = A_k^\star \Sigma_{k+1}^\star (A_k^\star)^\top + \mathbf{1}_{\{\alpha>0\}}\, G_k^\star (G_k^\star)^\top. \tag{21}$$

Under the learned score $\widehat{s}_t = \nabla \log p_t + e_t$ with $e_t(x) = -(\Delta_t)x + \delta_t$, the same discretization produces an update which remains affine in X_{k+1} conditional on the error coefficients. Namely, for each step k there exist deterministic solver-dependent matrices B_k^Δ and B_k^δ (depending on the schedule and on whether we run the reverse SDE or ODE) such that the error enters the update as

$$X_k = \big(A_k^\star - B_k^\Delta \Delta_{t_{k+1}}\big) X_{k+1} + \big(a_k^\star + B_k^\delta \delta_{t_{k+1}}\big) + \mathbf{1}_{\{\alpha>0\}}\, G_k^\star Z_k \;+\; R_k, \tag{22}$$

where R_k collects higher-order (in γ) terms specific to the discretization if the score is evaluated at intermediate stages; for Euler-type schemes one may take $R_k \equiv 0$ in the linear-Gaussian model. Since (22) is affine-Gaussian conditional on $\{(\Delta_{t_a}, \delta_{t_a})\}$, we have that $q_k \,|\, \{(\Delta_{t_a}, \delta_{t_a})\}_{a\geq k}$ is Gaussian.

Denoting conditional moments by $(\widetilde{m}_k, \widetilde{\Sigma}_k)$, we obtain the exact conditional recurrences

$$\widetilde{m}_k = \big(A_k^\star - B_k^\Delta \Delta_{t_{k+1}}\big)\widetilde{m}_{k+1} + a_k^\star + B_k^\delta \delta_{t_{k+1}} + \mathbb{E}[R_k \mid \Delta, \delta], \tag{23}$$

$$\widetilde{\Sigma}_k = \big(A_k^\star - B_k^\Delta \Delta_{t_{k+1}}\big)\widetilde{\Sigma}_{k+1}\big(A_k^\star - B_k^\Delta \Delta_{t_{k+1}}\big)^\top + \mathbf{1}_{\{\alpha>0\}}\, G_k^\star (G_k^\star)^\top + \mathrm{Cov}(R_k \mid \Delta, \delta). \tag{24}$$

The unconditional law q_k is therefore a mixture of Gaussians in general; however, our expansion of $\mathbb{E}W_2^2$ will only require the first two moments of $(\widetilde{m}_k, \widetilde{\Sigma}_k)$ up to second order in ε.

To expose the dependence on the full error trajectory, we unroll (23)–(24) around the unperturbed trajectory (21). Define the unperturbed propagation operator

$$\Phi^\star_{k\leftarrow a} := A_k^\star A_{k+1}^\star \cdots A_{a-1}^\star, \qquad k < a, \qquad \Phi^\star_{k\leftarrow k} := I. \tag{25}$$

Ignoring R_k for simplicity of exposition (it can be absorbed into the discretization term later), a first-order variation-of-constants expansion yields

$$\widetilde{m}_k = m_k^\star + \sum_{a\geq k} \Phi^\star_{k\leftarrow a}\, B_a^\delta\, \delta_{t_{a+1}} - \sum_{a\geq k} \Phi^\star_{k\leftarrow a}\, B_a^\Delta\, \Delta_{t_{a+1}}\, m_{a+1}^\star \;+\; O(\varepsilon^2), \tag{26}$$

where B_a^Δ, B_a^δ are the step-a coefficients appearing in (22) (with a shift in indices depending on whether the solver evaluates the score at t_a or t_{a+1}; this is immaterial for the present structural statement). Importantly, (26) shows that, to first order, the terminal mean perturbation is a *linear functional* of the entire time-indexed perturbation process $\{\Delta_{t_a}, \delta_{t_a}\}$ with deterministic weights given by $\Phi^\star_{k\leftarrow a}$.

For the covariance, we linearize (24) around $\Sigma_k^\star$. Let $\delta\Sigma_k := \widetilde{\Sigma}_k - \Sigma_k^\star$. Expanding to first order in Δ and using that $\Sigma_{k+1}^\star$ is deterministic, we obtain the affine recursion

$$\delta\Sigma_k = A_k^\star\, \delta\Sigma_{k+1}\, (A_k^\star)^\top - A_k^\star \Sigma_{k+1}^\star (B_k^\Delta \Delta_{t_{k+1}})^\top - (B_k^\Delta \Delta_{t_{k+1}})\Sigma_{k+1}^\star (A_k^\star)^\top + O(\varepsilon^2). \tag{27}$$

Thus, defining the linear operator $\mathcal{A}_k(M) := A_k^\star M (A_k^\star)^\top$ and the step forcing

$$\mathcal{F}_k(\Delta) := -A_k^\star \Sigma_{k+1}^\star (B_k^\Delta \Delta)^\top - (B_k^\Delta \Delta)\Sigma_{k+1}^\star (A_k^\star)^\top, \tag{28}$$

we can unroll (27) as

$$\delta\Sigma_k = \sum_{a\geq k} \Big(\mathcal{A}_k \circ \mathcal{A}_{k+1} \circ \cdots \circ \mathcal{A}_{a-1}\Big)\big(\mathcal{F}_a(\Delta_{t_{a+1}})\big) \;+\; O(\varepsilon^2). \tag{29}$$

Equations (26)–(29) are the structural input for the kernel expansion: they exhibit terminal mean and covariance perturbations as linear functionals of (Δ, δ), hence any second-order metric expansion (such as the Gaussian W_2^2

expansion) becomes a quadratic form in the error process, and therefore depends only on time–time covariances.

The remaining requirement is well-posedness and stability. We impose a mean-square stability condition on the unperturbed propagators: there exists $M < \infty$ such that for all $k \le a \le K$,

$$\|\Phi^\star_{k\leftarrow a}\| \le M, \qquad \Big\|\mathcal{A}_k \circ \cdots \circ \mathcal{A}_{a-1}\Big\|_{\mathrm{op}} \le M, \tag{30}$$

together with boundedness of the solver coefficients $\sup_k \|B^\Delta_k\| + \|B^\delta_k\| + \|G^\star_k\| < \infty$. Under (30) and the small-error regime (14), a perturbation argument shows that the perturbed one-step maps in (22) remain uniformly stable for ε sufficiently small (and γ sufficiently small when stability is only guaranteed asymptotically). Concretely, one obtains uniform moment bounds of the form $\sup_k \mathbb{E}\|X_k\|^2 < \infty$ and $\sup_k \mathbb{E}\|\widetilde{\Sigma}_k\|_F < \infty$, ensuring that the $O(\varepsilon^2)$ remainders in (26)–(29) are controlled uniformly over k.

With these propagation identities and stability bounds in place, the next section can treat W_2^2 between Gaussians as a smooth function of (m_k, Σ_k) around $(m^\star_k, \Sigma^\star_k)$ and identify the explicit correlated-time kernel obtained by contracting the time–time covariances of (Δ_t, δ_t) against the deterministic sensitivity operators induced by (26)–(29).

5 5. Main theorem: correlated-time kernel expansion of $\mathbb{E}W_2^2$; explicit kernels for Euler–Maruyama reverse SDE and for probability-flow ODE; reduction to the independent-time case.

$$W_2^2\big(\mathcal{N}(\mu, C), \mathcal{N}(m, \Sigma)\big) = \|\mu - m\|^2 + B^2(C, \Sigma), \tag{31}$$

and we regard the right-hand side as a smooth functional of $(m, \Sigma) \in \mathbb{R}^d \times \mathbb{S}^d_{++}$. Although q_k is generally a mixture (since (Δ, δ) are random), the conditional law $q_k \,|\, \{(\Delta_{t_a}, \delta_{t_a})\}$ is Gaussian with moments $(\widetilde{m}_k, \widetilde{\Sigma}_k)$. We therefore expand the random quantity

$$F_k := \|\mu - \widetilde{m}_k\|^2 + B^2\big(C, \widetilde{\Sigma}_k\big)$$

around the deterministic reference $(m^\star_k, \Sigma^\star_k)$ and then take expectation. Writing $\delta m_k := \widetilde{m}_k - m^\star_k$ and $\delta\Sigma_k := \widetilde{\Sigma}_k - \Sigma^\star_k$, the first-order term in δm_k vanishes after expectation because $\mathbb{E}\,\delta m_k = 0$ under $\mathbb{E}\Delta = \mathbb{E}\delta = 0$ and the stability bounds; similarly, $\mathbb{E}\,\delta\Sigma_k = 0$ at first order. Consequently, the leading dependence on the score error is quadratic and is completely determined by the time–time second moments.

To make this dependence explicit, we introduce the deterministic sensitivity weights appearing in the first-order unrollings. For the mean, define

for $a \geq k$ the matrices

$$S_{k,a} := \Phi^{\star}_{k \leftarrow a}\, B^{\delta}_a \in \mathbb{R}^{d\times d}, \qquad T_{k,a} := \Phi^{\star}_{k \leftarrow a}\, B^{\Delta}_a \in \mathbb{R}^{d\times d},$$

so that the first-order perturbation can be written as

$$\delta m_k = \sum_{a\geq k} S_{k,a}\, \delta_{t_{a+1}} - \sum_{a\geq k} T_{k,a}\, \Delta_{t_{a+1}}\, m^{\star}_{a+1} + O(\varepsilon^2).$$

For the covariance, define the linear propagation operator on matrices

$$\Psi_{k,a} := \mathcal{A}_k \circ \cdots \circ \mathcal{A}_{a-1}, \qquad a \geq k,$$

and the induced linear map $\mathcal{P}_{k,a} : \mathbb{R}^{d\times d} \to \mathbb{S}^d$,

$$\mathcal{P}_{k,a}(\Delta) := \Psi_{k,a}\big(\mathcal{F}_a(\Delta)\big),$$

so that

$$\delta\Sigma_k = \sum_{a\geq k} \mathcal{P}_{k,a}\big(\Delta_{t_{a+1}}\big) + O(\varepsilon^2).$$

Finally, let $\mathcal{H}_k : \mathbb{S}^d \to \mathbb{S}^d$ denote the Hessian (second Fréchet derivative) of $\Sigma \mapsto B^2(C,\Sigma)$ evaluated at $\Sigma^{\star}_k$. Concretely, for $U, V \in \mathbb{S}^d$ we define the bilinear form

$$\langle U,\, \mathcal{H}_k(V)\rangle := \frac{d^2}{d\tau\, d\sigma}\bigg|_{\tau=\sigma=0} B^2\big(C,\, \Sigma^{\star}_k + \tau U + \sigma V\big), \tag{32}$$

which is well-defined and bounded on compact subsets of $\mathbb{S}^d_{++}$.

We can now state the correlated-time expansion in a form that directly identifies the kernel blocks contracting the time–time covariances.

Theorem 5.1 (Correlated-time kernel expansion on a grid)**.** *Under the standing assumptions (Gaussian data, mean-square stability, bounded moments, and* $\sup_t \mathbb{E}\|\Delta_t\|_F^2 + \mathbb{E}\|\delta_t\|^2 \leq \varepsilon^2$*), we have for each* k

$$\mathbb{E}\, W_2^2\big(p_{\mathrm{data}}, q_k\big) = E^{(0)}_{r_k} + E^{\mathrm{disc}}_{r_k}(\gamma) + \mathcal{Q}^{\Delta}_k + \mathcal{Q}^{\delta}_k + \mathcal{Q}^{\Delta\delta}_k + o(\gamma) + o(\varepsilon^2),$$

where the correlated-time contributions are the quadratic forms

$$\mathcal{Q}^{\delta}_k = \sum_{a,b\geq k} \mathrm{tr}\Big(S_{k,a}\, \mathrm{Cov}(\delta_{t_{a+1}}, \delta_{t_{b+1}})\, S_{k,b}^{\top}\Big), \tag{33}$$

$$\mathcal{Q}^{\Delta}_k = \sum_{a,b\geq k} \Big\langle K^{\Delta,\mathrm{mean}}_k[a,b] + K^{\Delta,\mathrm{cov}}_k[a,b],\, \mathrm{Cov}\big(\mathrm{vec}(\Delta_{t_{a+1}}), \mathrm{vec}(\Delta_{t_{b+1}})\big)\Big\rangle, \tag{34}$$

$$\mathcal{Q}^{\Delta\delta}_k = -2 \sum_{a,b\geq k} \mathbb{E}\Big[\delta_{t_{a+1}}^{\top} S_{k,a}^{\top} T_{k,b}\, \Delta_{t_{b+1}} m^{\star}_{b+1}\Big], \tag{35}$$

with deterministic kernel blocks given by

$$K_k^{\Delta,\mathrm{mean}}[a,b] := \big(m_{a+1}^\star \otimes T_{k,a}\big)^\top \big(m_{b+1}^\star \otimes T_{k,b}\big), \tag{36}$$

$$K_k^{\Delta,\mathrm{cov}}[a,b] := \mathcal{P}_{k,a}^\dagger \circ \mathcal{H}_k \circ \mathcal{P}_{k,b}, \tag{37}$$

where $\mathcal{P}_{k,a}^\dagger$ denotes the adjoint of $\mathcal{P}_{k,a}$ with respect to the Frobenius inner product on matrices and the Euclidean inner product on $\mathbb{R}^{d^2}$ after vectorization.

In the Gaussian-linear case, the kernels in Theorem 5.1 are explicit once the solver coefficients are fixed. We record the resulting one-step coefficients for Euler-type schemes in the commonly used unified parametrization

$$dX_t = -\tfrac{1}{2}\beta_t X_t\,dt + \sqrt{\beta_t}\,\xi_t\,dW_t, \qquad g_t^2 := \beta_t \xi_t^2.$$

For Euler–Maruyama applied to the reverse SDE with noise parameter α (so that the injected diffusion coefficient is αg_t), the score enters the drift with prefactor g_t^2, hence over a step $\gamma_k := t_{k+1} - t_k$ the score error contributes as $-\gamma_k g_{t_{k+1}}^2 e_{t_{k+1}}(X_{k+1})$. In the affine model $e_t(x) = -(\Delta_t)x + \delta_t$, this yields the concrete choice

$$B_k^\Delta = -\eta_k I, \qquad B_k^\delta = -\eta_k I, \qquad \eta_k := \gamma_k g_{t_{k+1}}^2, \tag{38}$$

with the unperturbed affine map determined by the exact Gaussian score $\nabla \log p_t(x) = -\Sigma_t^{-1}(x - \mu_t)$:

$$A_k^\star = I + \gamma_k\Big(\tfrac{1}{2}\beta_{t_{k+1}} I - \eta_k\,\Sigma_{t_{k+1}}^{-1}\Big), \qquad a_k^\star = \gamma_k\,\eta_k\,\Sigma_{t_{k+1}}^{-1}\mu_{t_{k+1}}, \qquad G_k^\star = \alpha\sqrt{\gamma_k}\,g_{t_{k+1}}\,I,$$

so that all propagation operators in (33)–(37) are computable by forward recursion. For the probability-flow ODE (the limit $\alpha \to 0$ with the standard drift modification), the score prefactor is halved; correspondingly (38) holds with η_k replaced by $\eta_k/2$, and $G_k^\star \equiv 0$.

Finally, the reduction to the independent-time case is immediate at the level of the quadratic forms. If, for instance, $\mathrm{Cov}(\delta_{t_a}, \delta_{t_b}) = 0$ and $\mathrm{Cov}(\mathrm{vec}(\Delta_{t_a}), \mathrm{vec}(\Delta_{t_b})) = 0$ for $a \neq b$ (and similarly for the cross-covariance), then (33)–(35) collapse to single sums:

$$\mathcal{Q}_k^\delta = \sum_{a\geq k} \mathrm{tr}\Big(S_{k,a}\,\mathrm{Var}(\delta_{t_{a+1}})\,S_{k,a}^\top\Big), \qquad \mathcal{Q}_k^\Delta = \sum_{a\geq k} \Big\langle K_k^{\Delta,\mathrm{mean}}[a,a] + K_k^{\Delta,\mathrm{cov}}[a,a],\ \mathrm{Var}\big(\mathrm{vec}(\Delta_{t_{a+1}})\big)\Big\rangle,$$

which recovers the familiar "diagonal-in-time" structure. Theorem 5.1 shows that in general no such reduction is valid: correlated score errors contribute through the full time–time covariance via the deterministic kernel blocks induced by the sampler dynamics and by the Bures geometry.

6 6. Tightness and lower bounds: show the quadratic functional is unavoidable by constructing Gaussian score-error processes matching any prescribed time–time covariance; discuss identifiability/sample complexity of estimating correlations.

We justify that the quadratic functional of the score-error process identified above is not merely an artifact of the proof technique: in the small-error regime it is the correct leading-order object, and its dependence on the *full* time–time covariance (rather than only time-marginal variances) is in general unavoidable.

Tightness in the Gaussian score-error class. Fix a sampling grid $\{t_k\}_{k=0}^K$ and consider the affine error model $e_{t_a}(x) = -(\Delta_{t_a})x + \delta_{t_a}$ with $\mathbb{E}\Delta_{t_a} = 0$ and $\mathbb{E}\delta_{t_a} = 0$. Let us collect all random coefficients into a single finite-dimensional vector

$$Z := \Big(\mathrm{vec}(\Delta_{t_{k+1}}), \dots, \mathrm{vec}(\Delta_{t_K}),\ \delta_{t_{k+1}}, \dots, \delta_{t_K}\Big) \in \mathbb{R}^{(K-k)d^2+(K-k)d}.$$

Any choice of positive semidefinite covariance matrix $\Sigma_Z \succeq 0$ defines a centered Gaussian law $Z \sim \mathcal{N}(0, \Sigma_Z)$, hence a jointly Gaussian score-error process on the grid realizing prescribed time–time covariances

$$\mathrm{Cov}\big(\mathrm{vec}(\Delta_{t_a}), \mathrm{vec}(\Delta_{t_b})\big), \qquad \mathrm{Cov}(\delta_{t_a}, \delta_{t_b}), \qquad \mathrm{Cov}\big(\mathrm{vec}(\Delta_{t_a}), \delta_{t_b}\big), \quad a, b \in \{k{+}1, \dots, K\}.$$

For such Gaussian inputs, the unrolled perturbations $(\delta m_k, \delta\Sigma_k)$ are (to first order in ε) *linear* functionals of Z, hence themselves jointly Gaussian at leading order. Therefore, when we expand $F_k = \|\mu - \widetilde{m}_k\|^2 + B^2(C, \widetilde{\Sigma}_k)$ around $(m_k^\star, \Sigma_k^\star)$, all odd-order terms in $(\delta m_k, \delta\Sigma_k)$ vanish under expectation by symmetry, while the second-order term is exact up to the controlled remainder coming from (i) the $O(\varepsilon^2)$ truncation in the linearization of the dynamics and (ii) the third-order Taylor remainder of the smooth functional $(m, \Sigma) \mapsto \|\mu - m\|^2 + B^2(C, \Sigma)$. Concretely, for a centered Gaussian Z with $\mathbb{E}\|Z\|^2 = O(\varepsilon^2)$ and mean-square stability ensuring uniform bounds on the sensitivity operators, we obtain

$$\mathbb{E}F_k = F_k^\star + \frac{1}{2}\,\mathbb{E}\Big[\big\langle (\delta m_k, \delta\Sigma_k),\ \nabla^2 F_k^\star\,(\delta m_k, \delta\Sigma_k)\big\rangle\Big] + O(\varepsilon^3),$$

where $F_k^\star := \|\mu - m_k^\star\|^2 + B^2(C, \Sigma_k^\star)$ and the Hessian is evaluated at the unperturbed point. Since $(\delta m_k, \delta\Sigma_k)$ are linear in Z to first order, the second-order term is a quadratic form in Z, and taking expectation yields precisely a bilinear contraction against Σ_Z, i.e., a sum/double-sum of the form appearing in (33)–(35). This establishes that, within the jointly Gaussian affine-error

class, the quadratic functional is *attainable* and the residual is of strictly higher order (under the stated moment and stability controls). In particular, without additional structure on Z beyond second moments, one cannot improve the dependence on $\mathrm{Cov}(e_s, e_{s'})$.

Why off-diagonal time correlations cannot be ignored. The kernel expansion shows that the leading error contribution is a quadratic form in the entire error trajectory, hence depends on the covariance *operator* in time. To see that off-diagonal terms are essential, it suffices to exhibit two error processes with identical time-marginal variances but different cross-time covariances that yield different values of the quadratic form.

Consider, for simplicity, only the δ-part and two times t_a, t_b (with $a \neq b$), and assume $\Delta \equiv 0$ and $m^\star \equiv 0$ to eliminate the mean–Δ coupling. Let $\delta_{t_a}, \delta_{t_b} \in \mathbb{R}^d$ be centered jointly Gaussian with

$$\mathrm{Var}(\delta_{t_a}) = \mathrm{Var}(\delta_{t_b}) = \sigma^2 I, \qquad \mathrm{Cov}(\delta_{t_a}, \delta_{t_b}) = \rho\, \sigma^2 I,$$

where $\rho \in [-1, 1]$. The diagonal-in-time statistics (the per-time variances) are independent of ρ, yet the quadratic contribution becomes

$$\mathcal{Q}_k^\delta(\rho) = \mathrm{tr}\big(S_{k,a}\sigma^2 I S_{k,a}^\top\big) + \mathrm{tr}\big(S_{k,b}\sigma^2 I S_{k,b}^\top\big) + 2\rho\, \mathrm{tr}\big(S_{k,a}\sigma^2 I S_{k,b}^\top\big).$$

Unless the cross-sensitivity trace $\mathrm{tr}(S_{k,a} S_{k,b}^\top)$ vanishes (a nongeneric condition tied to special choices of schedule/grid/solver), the value of $\mathcal{Q}_k^\delta(\rho)$ varies with ρ. Thus any diagnostic that only measures $\mathrm{Var}(\delta_{t_a})$ at each time (e.g., per-time denoising MSE) cannot, in general, predict the sampling error: two trained networks may have identical time-marginal validation losses while producing different sampling quality due to different cross-time correlations induced by shared weights and correlated features. The same phenomenon holds, *a fortiori*, for Δ and for mixed Δ–δ cross-covariances, where the kernel blocks couple different times through the propagation operators and the Bures Hessian.

Identifiability and the cost of estimating correlations. The preceding argument is not only qualitative: it implies an information-theoretic obstruction. The object that controls $\mathcal{Q}$ is the full time–time covariance, which on a grid of size $m := K - k$ comprises m^2 blocks (scalar-valued when $d = 1$, matrix-valued for $d > 1$). Even in the scalar case, estimating an $m \times m$ covariance matrix from i.i.d. samples incurs a minimax sample complexity scaling proportional to m^2 for constant Frobenius accuracy. More precisely, for R i.i.d. samples $z^{(1)}, \ldots, z^{(R)} \in \mathbb{R}^m$ from $\mathcal{N}(0, \Sigma)$ with bounded spectrum, the empirical covariance concentrates as

$$\mathbb{E}\|\widehat{\Sigma} - \Sigma\|_F^2 \asymp \frac{m^2}{R},$$

so achieving $\mathbb{E}\|\widehat{\Sigma}-\Sigma\|_F \le \eta$ requires $R = \Omega(m^2/\eta^2)$. In contrast, estimating only the diagonal variances needs $R = \Omega(m/\eta^2)$. Therefore, any attempt to compute the correlated-time term $\mathcal{Q}$ from residual samples must either (i) accept quadratic scaling in the grid size, or (ii) exploit additional structure in time such as low-rank covariances, bandedness (short-range correlations), parametric kernels (e.g. exponential decay), or stationarity assumptions enabling spectral estimation.

These tightness and identifiability considerations motivate the next step: given a trained score model, we must decide what portion of $\mathrm{Cov}(e_s, e_{s'})$ is practically estimable from denoising residuals, what structural approximations are reasonable, and how to contract estimated covariances with the deterministic kernels to obtain actionable predictors and diagnostics.

7 7. Practical estimators and diagnostics: how to estimate (approximations of) $\mathrm{Cov}(e_s, e_{s'})$ from denoising residuals and how to compute the kernel numerically; optional implications for schedule selection.

7. Practical estimators and diagnostics

Our expansion reduces the leading score-error contribution to a deterministic kernel contraction against the (generally unknown) time–time covariances of the affine coefficients (Δ_t, δ_t). We now describe how we estimate approximations of these covariances from denoising residuals, and how we compute the kernel numerically on a time grid.

Observable residuals on a common probability space. For each time t we construct paired samples (X_0, ε) with $X_0 \sim p_{\mathrm{data}}$ and $\varepsilon \sim \mathcal{N}(0, I)$, and then generate the corresponding noisy state X_t via the known forward map of the chosen diffusion schedule (e.g. $X_t = a(t)X_0 + b(t)\varepsilon$ in the Gaussian-forward setting). This coupling is essential: it induces a joint law for (X_{t_a}, X_{t_b}) and allows us to estimate *cross*-time statistics from i.i.d. draws of (X_0, ε). We then evaluate the trained network to obtain $\widehat{s}_t(X_t)$. In the Gaussian target setting, $\nabla \log p_t(x)$ is available in closed form; hence we can form the pointwise score residual

$$R_t \;:=\; \widehat{s}_t(X_t) - \nabla \log p_t(X_t).$$

In standard noise-prediction parameterizations, R_t can equivalently be obtained (up to a known scalar factor) from the observable residual $\widehat{\varepsilon}_t(X_t) - \varepsilon$; thus the procedure does not rely on direct access to $\nabla \log p_t$ beyond the synthetic/controlled setting.

From residual fields to affine coefficients. The kernel formulas are stated in terms of the affine decomposition $e_t(x) = -(\Delta_t)x + \delta_t$. When p_t is Gaussian, the true score is linear in x and, empirically, $\widehat{s}_t(x)$ is often well-approximated by an affine map on the typical set of p_t. We therefore estimate (Δ_t, δ_t) by projecting the residual field $x \mapsto R_t(x)$ onto affine functions under p_t. Concretely, given a batch $\{X_t^{(j)}, R_t^{(j)}\}_{j=1}^B$, we compute the (ridge) least-squares fit

$$(\widehat{\Delta}_t, \widehat{\delta}_t) \;\in\; \arg\min_{\Delta,\delta} \; \frac{1}{B}\sum_{j=1}^{B} \big\| R_t^{(j)} + \Delta X_t^{(j)} - \delta \big\|^2 \;+\; \lambda \|\Delta\|_F^2,$$

with a small $\lambda \geq 0$ for numerical stability. This estimator targets the $L^2(p_t)$-best affine approximation of $R_t(\cdot)$. In the diagonalizable Gaussian regime (working in an eigenbasis of Σ_t), we often further restrict Δ to be diagonal (or block-diagonal), which reduces variance and aligns with the per-mode kernel evaluation described below.

The preceding fit yields one pair $(\widehat{\Delta}_t, \widehat{\delta}_t)$ per time. To estimate time–time covariances, we need *random* affine coefficients. We obtain this randomness by repeating the entire coupled draw (X_0, ε), thereby producing i.i.d. realizations of the fitted coefficients

$$(\widehat{\Delta}_t^{(r)}, \widehat{\delta}_t^{(r)})_{t \in \{t_{k+1}, \ldots, t_K\}}, \qquad r = 1, \ldots, R,$$

where each repetition r uses an independent batch of coupled forward samples and refits the affine map. This "refit-per-repetition" construction yields empirical covariance estimates that capture the induced cross-time dependence created by a shared network evaluated on correlated inputs. When d is large, we replace refitting by a single fit of the mean affine map and estimate fluctuations by linearizing (e.g. via Jacobians), but we keep the refit-based estimator as a conceptually simple baseline in controlled experiments.

Estimating $\mathrm{Cov}(\Delta, \Delta)$, $\mathrm{Cov}(\delta, \delta)$, and cross terms. On a grid of size $m := K - k$, we form empirical covariances

$$\widehat{\Sigma}^{\delta}_{ab} \;:=\; \frac{1}{R-1}\sum_{r=1}^{R} \big(\widehat{\delta}^{(r)}_{t_a} - \overline{\delta}_{t_a}\big)\big(\widehat{\delta}^{(r)}_{t_b} - \overline{\delta}_{t_b}\big)^{\top}, \qquad \overline{\delta}_{t_a} := \frac{1}{R}\sum_{r=1}^{R} \widehat{\delta}^{(r)}_{t_a},$$

and analogously for $\widehat{\Sigma}^{\Delta}_{ab}$ using vectorization $\mathrm{vec}(\widehat{\Delta}^{(r)}_{t_a})$. Cross-covariances $\widehat{\Sigma}^{\Delta\delta}_{ab} := \mathrm{Cov}(\mathrm{vec}(\Delta_{t_a}), \delta_{t_b})$ are estimated similarly when we include the mixed kernel blocks. In regimes where m is moderately large, we regularize these block matrices by enforcing symmetry and positive semidefiniteness (e.g. eigenvalue clipping on the full stacked covariance), which improves stability of subsequent contractions.

Structured approximations in time. Because unrestricted estimation scales poorly in m, we implement and compare several structured models for the time–time covariance. (i) *Banded time covariance:* we set $\widehat{\Sigma}_{ab} = 0$ for $|a - b| > w$, with bandwidth w chosen by validation on held-out repetitions. (ii) *Low-rank in time:* we compute a truncated SVD of the empirical $m \times m$ covariance in each scalar mode (or after trace aggregation) and retain the leading r components. (iii) *Separable (Kronecker) model:* we fit $\mathrm{Cov}(\delta_s, \delta_{s'}) \approx k(s, s')\,\Sigma_\delta$ with k parametric (e.g. exponential decay), which reduces estimation to a small number of time-kernel parameters. These approximations interpolate between feasibility and fidelity, and they are directly compatible with the kernel contraction, since $\mathcal{Q}$ is linear in each covariance argument.

Numerical computation of the kernel. On the same time grid, we compute sensitivity operators that map perturbations at time index a to first-order perturbations of terminal statistics at step k. In the Gaussian-linear setting, these operators are obtained by unrolling the linearized mean/covariance recurrences, yielding matrices $S^{\delta}_{k\leftarrow a}$ and $S^{\Delta}_{k\leftarrow a}$ such that

$$\delta m_k \;\approx\; \sum_{a\geq k} S^{\delta}_{k\leftarrow a}\,\delta_{t_a} \;+\; \sum_{a\geq k} S^{\Delta,m}_{k\leftarrow a}\,\mathrm{vec}(\Delta_{t_a}), \qquad \delta\Sigma_k \;\approx\; \sum_{a\geq k} S^{\Delta,\Sigma}_{k\leftarrow a}\,\mathrm{vec}(\Delta_{t_a}).$$

We then assemble the kernel blocks by composing these sensitivities with the Hessian of $(m, \Sigma) \mapsto \|m - \mu\|^2 + B^2(C, \Sigma)$ at the unperturbed terminal pair $(m^\star_k, \Sigma^\star_k)$. Computationally, the only nontrivial step is evaluating the Bures Hessian action, which we implement via spectral factorization in the eigenbasis where C and $\Sigma^\star_k$ are diagonal (or approximately so). In that basis, the contraction decouples across coordinates and the kernel reduces to per-mode scalar weights, leading to an $O(m^2 d)$ assembly cost.

Diagnostics and schedule implications. Given $(K^{\Delta}_k[a, b], K^{\delta}_k[a, b])$ and estimated covariances, we report: (i) the predicted terminal error curve $k \mapsto E^{(0)}_{r_k} + E^{\mathrm{disc}}_{r_k}(\gamma) + \widehat{\mathcal{Q}}_{r_k}$; (ii) a *time interaction matrix* with entries

$$I^{\delta}_{ab} := \big\langle K^{\delta}_k[a, b], \widehat{\Sigma}^{\delta}_{ab} \big\rangle, \qquad I^{\Delta}_{ab} := \big\langle K^{\Delta}_k[a, b], \widehat{\Sigma}^{\Delta}_{ab} \big\rangle,$$

which localizes whether error arises primarily from marginal variances ($a = b$) or from correlations ($a \neq b$); and (iii) bootstrap confidence intervals over repetitions. Finally, since the kernel depends on the schedule through the sensitivities, a practical implication is that schedule or grid selection can be informed by minimizing $\widehat{\mathcal{Q}}$ subject to computational constraints, e.g. allocating smaller step sizes in time regions where the product "kernel magnitude $\times$ estimated covariance" is largest.

8 8. Experiments: synthetic anisotropic Gaussians and shared random-feature scores; controlled correlation ablations; kernel-predicted error vs measured W_2; solver comparison; discuss where remainders become visible.

8. Experiments: synthetic anisotropic Gaussians and controlled time–time correlations

We validate the correlated-time expansion by constructing settings in which (i) the target is anisotropic and hence the Bures contribution is nontrivial, (ii) the sampler output remains exactly Gaussian (so W_2^2 is computable without Monte Carlo bias), and (iii) we can tune the time–time covariance of the score error while keeping time-marginal error magnitudes essentially fixed.

Synthetic anisotropic targets. We take $p_{\mathrm{data}} = \mathcal{N}(\mu, C)$ with $\mu = 0$ and

$$C = \mathrm{diag}(\lambda_1, \dots, \lambda_d), \qquad \lambda_i \propto i^{-p}, \quad p \in \{0, 1, 2\},$$

normalized so that $\mathrm{tr}(C) = d$. This yields progressively stronger anisotropy, ranging from isotropic ($p = 0$) to heavy-tailed spectra ($p = 2$). We fix a diffusion schedule for which the forward marginals are Gaussian with known (μ_t, Σ_t); in all experiments we work in the eigenbasis of C so that the kernel assembly decouples per coordinate as described in the preceding section. We report $W_2^2(p_{\mathrm{data}}, q_k)$ at several terminal indices k (equivalently reverse times r_k), computed in closed form from the mean and covariance of q_k.

Shared random-feature score models with affine errors. To obtain an exactly affine learned score with controllable cross-time dependence, we define a family of *random-feature* score perturbations

$$e_t(x) = -(\Delta_t)x + \delta_t$$

by generating a shared random matrix $U \in \mathbb{R}^{m \times d}$ with i.i.d. $\mathcal{N}(0, 1/m)$ entries and letting

$$\Delta_t := U^\top A_t U, \qquad \delta_t := U^\top b_t,$$

where $A_t \in \mathbb{R}^{m \times m}$ and $b_t \in \mathbb{R}^m$ are low-dimensional time-indexed random coefficients. This construction induces nontrivial correlations in Δ_t and δ_t across t through the shared embedding U, while allowing us to modulate time correlation by controlling (A_t, b_t). We then define $\widehat{s}_t(x) = \nabla \log p_t(x) + e_t(x)$ and run reverse-time sampling with this inexact score. Since the drift remains affine in x (and the diffusion is state-independent), the sampler output is

Gaussian for any discretization, so $q_k = \mathcal{N}(m_k, \Sigma_k)$ can be tracked exactly by propagating (m_k, Σ_k).

Correlation ablations at fixed marginal error. We compare three regimes designed to separate the effect of time–time covariance from per-time error magnitudes.

1. *Independent-in-time:* draw (A_{t_a}, b_{t_a}) independently over grid points. This makes $\mathrm{Cov}(\Delta_{t_a}, \Delta_{t_b}) \approx 0$ and $\mathrm{Cov}(\delta_{t_a}, \delta_{t_b}) \approx 0$ for $a \neq b$ up to the shared-feature effect, which we suppress by resampling U per time.

2. *Fully shared in time:* draw a single pair (A, b) and set $(A_{t_a}, b_{t_a}) \equiv (A, b)$ for all times. This keeps $\mathbb{E}\|\Delta_{t_a}\|_F^2$ and $\mathbb{E}\|\delta_{t_a}\|^2$ comparable to the independent regime but makes cross-time covariances maximal.

3. *Band-correlated:* generate an AR(1) process in time for (A_{t_a}, b_{t_a}) with correlation $\rho^{|a-b|}$, varying $\rho \in [0, 0.99]$ to interpolate between the two extremes.

In each regime we scale the coefficient variances to enforce $\max_a \mathbb{E}\|\Delta_{t_a}\|_F^2 + \mathbb{E}\|\delta_{t_a}\|^2 \approx \varepsilon^2$ for a prescribed ε, so that changes in sampling error can be attributed primarily to changes in cross-time structure rather than marginal magnitude.

Kernel prediction versus measured W_2^2. For each configuration we compute (a) the measured quantity $\mathbb{E}W_2^2(p_{\mathrm{data}}, q_k)$ by averaging the closed-form Gaussian W_2^2 over independent draws of the entire error trajectory, and (b) the predicted quantity given by the expansion

$$E_{r_k}^{(0)} + E_{r_k}^{\mathrm{disc}}(\gamma) + \sum_{a,b \geq k} \langle K_k^{\Delta}[a,b], \mathrm{Cov}(\Delta_{t_a}, \Delta_{t_b}) \rangle + \sum_{a,b \geq k} \langle K_k^{\delta}[a,b], \mathrm{Cov}(\delta_{t_a}, \delta_{t_b}) \rangle,$$

including cross terms when present in the synthetic generator. We emphasize that in this synthetic setting the covariances are known analytically (as functions of (U, ρ) and the coefficient variances), so we can separate *estimation* error from *model* error and directly test the correctness of the kernel contraction.

Across all spectra $p \in \{0, 1, 2\}$ we observe that, for small ε and stable step sizes, the predicted curve matches the measured curve uniformly over k up to a relative discrepancy consistent with $o(\varepsilon^2) + o(\gamma)$. The largest improvements over diagonal-in-time predictors occur in the fully shared and band-correlated regimes: holding the per-time variances fixed, the measured W_2^2 increases substantially with ρ, and this increase is captured by the off-diagonal kernel contraction terms. Conversely, a predictor that uses only the diagonal blocks $a = b$ remains nearly constant as ρ varies, demonstrating in practice the necessity highlighted by the tightness statement for correlated Gaussian processes.

Solver comparison and discretization scaling. We compare several discretizations on the same grid family: Euler–Maruyama for $\alpha > 0$, and first- and higher-order ODE solvers in the $\alpha \to 0$ regime. In all cases we keep the score-error process fixed and decrease γ to isolate solver effects. Empirically, after subtracting the estimated quadratic score-error term (kernel contraction) from the total error, the residual scales with the expected solver order, consistent with $E_r^{\mathrm{disc}}(\gamma) = O(\gamma^q)$. Meanwhile, the quadratic term itself remains essentially invariant to solver order except through the sensitivity operators inside the kernel, which produces only mild quantitative changes when the solver is changed but γ is kept small.

Where the remainders become visible. We probe the boundary of the small-error/stable regime in two ways. First, we increase ε at fixed γ and observe a clear deviation from quadratic scaling: plotting $\mathbb{E}W_2^2$ versus ε shows an ε^2 regime followed by noticeable curvature, which we attribute to the $O(\varepsilon^3)$ remainder and to the breakdown of the second-order Taylor approximation of the Bures term when $\delta\Sigma_k$ is no longer small relative to $\Sigma_k^\star$. Second, we increase γ toward the stability limit of the linear recurrence; in this case the discretization term and the interaction between discretization and score error (a higher-order effect not retained in our leading expansion) becomes visible, and the kernel predictor begins to systematically under-estimate the measured error. These observations delineate the practical range in which the correlated-time kernel contraction is quantitatively predictive and, more importantly, confirm that the dominant failure modes match the remainder terms suggested by the theory.

9 9. Discussion and extensions: toward non-Gaussian/local linearization, modern solvers, and linking to optimizer dynamics (SGD/Adam) for predicting covariance structure.

9. Discussion and extensions: beyond Gaussianity, solver families, and predicting time–time covariance from training dynamics

Our analysis isolates a single object that governs the leading score-induced degradation of sampling quality: the time–time covariance of the learned score error, pushed through an explicit sensitivity–metric kernel. In the Gaussian-linear setting the kernel can be written in closed form and the remainder is controllable, but the structure of the expansion suggests several extensions in which the same *operator viewpoint* remains meaningful.

From Gaussian targets to local linearization around trajectories. For general p_{data}, the reverse-time dynamics is nonlinear and the sampler output need not remain Gaussian. Nevertheless, the derivation of a quadratic functional in the score error is not inherently Gaussian; what is Gaussian-specific is the availability of an exact W_2^2 formula in terms of mean and covariance and the fact that affine perturbations of the score preserve Gaussianity. A natural extension is therefore to replace the global Gaussian closure by a *local linearization* along the unperturbed sampling flow (defined by the exact score and the chosen solver). Concretely, writing a generic reverse-time update as

$$x_{k-1} = \Psi_k(x_k) + \Gamma_k\, \widehat{s}_{t_k}(x_k) + \text{noise},$$

we may linearize Ψ_k and the score error around a reference path $x_k^\star$ (or around the evolving mean of $q_k^\star$) to obtain an approximate affine recursion for perturbations $\delta x_k := x_k - x_k^\star$:

$$\delta x_{k-1} \approx J_k\, \delta x_k + \Gamma_k\, e_{t_k}(x_k^\star) + \Gamma_k\, (\nabla_x e_{t_k})(x_k^\star)\, \delta x_k.$$

In this form, the role played by (Δ_t, δ_t) is assumed by the Jacobian and the value of e_t along the reference trajectory. If we further approximate the law of δx_k as Gaussian (a standard closure in weak-error analysis), then the same kernel contraction logic applies with *effective* matrix and vector errors derived from $(\nabla_x e_t)(x_k^\star)$ and $e_t(x_k^\star)$. The main technical change is that the kernels become *path-dependent* random operators, and the relevant covariances are conditional on the reference flow. In practice one may estimate these conditional covariances by running the sampler with the learned model and recording residuals along trajectories, thereby obtaining an empirical analogue of $\mathrm{Cov}(e_s, e_{s'})$ in the region of state space actually visited by the sampler.

Non-Gaussian metrics and the role of the Bures Hessian. Even if we retain Gaussian closures, one may wish to replace W_2^2 by alternative discrepancies used in diffusion evaluation (e.g. χ^2, KL, MMD). Our derivation separates (i) a linear sensitivity map from score perturbations to terminal distributional parameters, from (ii) a second-order metric expansion. For any discrepancy admitting a second-order expansion around a reference distribution, the quadratic term remains a bilinear form in the error process. In the Gaussian case, the metric component is the Hessian of the Bures term at $\Sigma_k^\star$; for alternative metrics one obtains a different curvature operator, but the *necessity of time–time covariance* persists whenever the perturbation is an integral/sum over time and the discrepancy is quadratic to leading order.

Modern solver families as kernel modifiers. Theorem 2 emphasizes that solver order affects $E_r^{\mathrm{disc}}(\gamma)$, while the score-error term remains $\Theta(\varepsilon^2)$

and depends on the solver only through sensitivity operators. This is consistent with the practical behavior of higher-order samplers (Heun, Runge–Kutta, predictor–corrector, exponential integrators, and multistep methods): their benefit is primarily to reduce discretization bias for a fixed number of function evaluations, not to suppress model error due to imperfect scores. In our framework, adopting a different solver replaces Φ (or its discrete analogue $\Phi_{k\leftarrow a}$) by a new propagation operator, hence modifying K_r^{Δ} and K_r^{δ} but not changing the fundamental quadratic dependence on $\mathrm{Cov}(e_s, e_{s'})$. This suggests a principled way to compare solvers in the presence of *correlated* score errors: two solvers with similar discretization error may still differ because they weight early/late-time score perturbations differently, thereby amplifying or attenuating correlations across specific time windows.

Guidance, conditioning, and the emergence of structured cross-time correlations. Classifier-free guidance and related conditioning mechanisms effectively replace the score by a linear combination of scores (e.g. conditional and unconditional). If each component has its own error process and the same network parameters generate correlated residuals across time, then the guided error inherits both *within-time* covariance (across score components) and *across-time* covariance. In our notation this appears through additional cross terms between multiple δ_t and Δ_t processes. The kernel formalism extends directly: one augments the covariance object to include cross-covariances between the constituent errors and contracts with corresponding block kernels. The main message is that guidance can change sampling error not only by scaling marginal error magnitudes, but also by altering the correlation structure across t through shared computations.

Linking $\mathrm{Cov}(e_s, e_{s'})$ to training dynamics (SGD/Adam). A central open direction is to predict (or at least parametrize) the time–time covariance of score errors from properties of the trained network and the optimizer. Let θ denote network parameters and define $e_t(x;\theta)$ as the score residual at time t. Linearizing around the terminal iterate $\hat{\theta}$ yields

$$e_t(x;\theta) \approx e_t(x;\hat{\theta}) + J_t(x)\,(\theta - \hat{\theta}), \qquad J_t(x) := \nabla_\theta e_t(x;\hat{\theta}).$$

If the randomness in e_t is dominated by parameter uncertainty induced by stochastic optimization (or by ensembling/checkpoint averaging), then for fixed evaluation distribution of x we obtain an approximate covariance factorization

$$\mathrm{Cov}\big(e_s(x;\theta), e_{s'}(x';\theta)\big) \approx \mathbb{E}\big[J_s(x)\,\mathrm{Cov}(\theta)\,J_{s'}(x')^\top\big],$$

with expectations over (x, x') drawn from appropriate forward marginals. This identifies a concrete source of cross-time correlations: the same random parameter perturbation $(\theta-\hat{\theta})$ simultaneously affects all times. In particular,

if $\mathrm{Cov}(\theta)$ is approximately low-rank (as suggested by implicit regularization and sharpness structure), then $\mathrm{Cov}(e_s, e_{s'})$ is also low-rank in time after projection through J_t. Such a structure would reconcile the information-theoretic barrier in Proposition 4 with feasible estimation, since one may estimate a small number of dominant temporal modes rather than the full $m \times m$ covariance.

Parametric time-covariance models and identifiable summaries. Motivated by optimizer-induced structure, we can posit models of the form

$$\mathrm{Cov}(\delta_s, \delta_{s'}) \approx \sum_{\ell=1}^{L} u_\ell(s)\, u_\ell(s')\, S_\ell, \qquad \mathrm{Cov}(\mathrm{vec}(\Delta_s), \mathrm{vec}(\Delta_{s'})) \approx \sum_{\ell=1}^{L} v_\ell(s)\, v_\ell(s')\, M_\ell,$$

with small L and unknown PSD matrices S_ℓ, M_ℓ. Given such a model, the kernel contraction reduces to a sum of L separable contributions, and one may estimate $\{u_\ell, v_\ell\}$ from training logs (e.g. per-time losses, gradient norms, or checkpoint differences) while fitting $\{S_\ell, M_\ell\}$ from a limited set of cross-time residual measurements. This shifts the emphasis from estimating an unrestricted covariance matrix to identifying the *kernel-relevant* subspace of correlations, namely the directions that produce large $\langle K_r, \mathrm{Cov}\rangle$.

Implications for practice. The expansion suggests two complementary levers for improving sampling: reducing ε^2 (better score accuracy) and reshaping $\mathrm{Cov}(e_s, e_{s'})$ (decorrelating errors across time in kernel-sensitive directions). The latter is not addressed by standard per-time training losses, and may require explicit regularizers or architectural interventions (e.g. time-embedding designs or per-time adapters) that reduce shared-mode coupling. Our kernel viewpoint provides a target: correlations matter only insofar as they align with the kernel; hence one may aim to penalize empirical cross-time covariance projected onto dominant kernel eigenmodes rather than attempting to decorrelate all residuals uniformly.

10 10. Limitations and open problems: beyond Gaussianity, non-asymptotic control of remainders, and bridging to large-scale image diffusion.

10. Limitations and open problems: beyond Gaussianity, non-asymptotic control of remainders, and bridging to large-scale image diffusion

We conclude by delineating what the present kernelized covariance viewpoint does *not* yet resolve. Our results provide a second-order, small-error

description of $\mathbb{E}W_2^2(p_{\text{data}}, q_k)$ in a Gaussian-linear regime, and they isolate the time–time covariance of score errors as the relevant object. However, several aspects of modern diffusion practice lie outside the current technical envelope, and closing these gaps appears to require new ideas rather than incremental refinements.

Beyond Gaussianity: identifying the correct state variables. The Gaussian setting is restrictive not merely because real data are non-Gaussian, but because the entire reduction to mean/covariance and the explicit Bures curvature are special. In non-Gaussian settings the map $e_\cdot \mapsto q_k$ is still a time-accumulated perturbation, but the observable one would like to control (e.g. W_2^2 or an image metric surrogate) depends on the full law and not on finitely many moments. A basic open problem is to identify a tractable set of *sufficient coordinates* for perturbation analysis: for instance, (i) a finite-dimensional projection of the law (moments, score-matching residuals, or feature statistics), (ii) a local Gaussian closure along the sampling flow, or (iii) a functional analytic formulation in which the perturbation is measured in a Sobolev-type norm of the density. Each choice leads to a different "metric Hessian" and hence a different kernel operator. At present we do not know which choice yields a theory that is both mathematically controlled and empirically predictive for high-dimensional images.

Affine error modeling and its failure modes. We model $e_t(x)$ as affine in x so that the perturbed reverse dynamics remains Gaussian and sensitivity operators are linear. In practice, e_t can be highly nonlinear and state-dependent, especially under guidance, clipping, or latent-space parameterizations. A natural extension is to view

$$e_t(x) \approx e_t(x_t^\star) + (\nabla_x e_t)(x_t^\star)(x - x_t^\star)$$

along a reference trajectory $x_t^\star$, but this introduces two complications: (i) the effective coefficients $(\nabla_x e_t)(x_t^\star)$ and $e_t(x_t^\star)$ are random and coupled to the sampler state, and (ii) the relevant covariances become conditional and path-dependent. Establishing a kernel representation in this setting appears to require a stability theory for random linearizations and an understanding of how trajectory dispersion feeds back into score-error statistics. We do not currently have such a theory at a level that would justify replacing global covariances $\mathrm{Cov}(e_s, e_{s'})$ by trajectory-conditioned covariances without additional assumptions.

Non-asymptotic remainder control and uniformity in d. Our expansion is asymptotic in two small parameters: γ (mesh size) and ε (score-error magnitude). While this is the appropriate regime for isolating leading mechanisms, it leaves open whether the remainder terms are small at the parameter values used in large-scale sampling. In particular, for high-dimensional

problems one must ask for *dimension-uniform* bounds: a statement of the form

$$\left|\mathbb{E}W_2^2 - (E^{(0)} + E^{\mathrm{disc}} + \mathcal{Q})\right| \le c_1(d)\,\gamma^{q+1} + c_2(d)\,\varepsilon^3$$

is only meaningful if $c_1(d), c_2(d)$ do not grow prohibitively with d. Controlling these constants seems difficult because even in Gaussian settings, W_2^2 scales with trace-like quantities, and the Bures curvature can amplify perturbations in poorly conditioned directions of C. A sharp open problem is to develop non-asymptotic inequalities that (i) track spectral conditioning explicitly, (ii) separate the dependence on low-variance and high-variance modes, and (iii) remain informative when d is large but the data lie near a lower-dimensional manifold.

Estimating time–time covariances at scale and identifying kernel-relevant structure. The kernel term depends on $\mathrm{Cov}(e_s, e_{s'})$ across pairs of times, and Proposition 4 formalizes the generic quadratic-in-grid-size cost of estimating an unstructured covariance. For large image models, we face an additional constraint: even computing $e_t(x)$ (or its proxies) at many times on many samples is expensive. Thus, a practical and theoretical challenge is to identify *compressed* summaries of the covariance that are sufficient for predicting $\langle K, \mathrm{Cov}\rangle$. One direction is to treat the kernel itself as defining a seminorm on covariance objects and to seek low-dimensional parameterizations that approximate $\mathrm{Cov}(e, e)$ only in that seminorm. Concretely, if K is approximately low-rank in time or concentrated on a few time windows, then only a small number of temporal modes are estimable and relevant. Formalizing such statements requires spectral analysis of the operator $(s, s') \mapsto K_r(s, s')$ and a corresponding minimax theory for estimating $\langle K, \mathrm{Cov}\rangle$ directly, without estimating Cov pointwise.

Solver interaction beyond weak order: adaptivity and stiffness. While higher-order solvers reduce discretization bias, in practice their performance depends on stability under stiff schedules and on how they query the score network. Our current framework treats the solver through a sensitivity operator, but does not explain when adaptive step sizes, multistep memory, or stochastic correctors improve (or worsen) the *interaction* with correlated score errors. An open problem is to characterize optimal time grids $\{t_k\}$ when $\mathrm{Cov}(e_s, e_{s'})$ is nontrivial: the best grid is plausibly not the one minimizing discretization error alone, but one that trades off discretization bias against kernel-weighted amplification of correlated errors. This suggests a design problem of the form $\min_{\{t_k\}} E^{\mathrm{disc}}(\gamma) + \mathcal{Q}$ under a budget constraint on the number of score evaluations, which we do not solve here.

Bridging to image diffusion: from W_2^2 to perceptual metrics and latent representations. Finally, even if one could extend the kernel ex-

pansion beyond Gaussians, it remains unclear whether W_2^2 is the right end metric for image generators evaluated by FID, precision/recall, or human preference. The practical implication is that a theory predicting W_2^2 may fail to predict perceived quality unless one can relate W_2^2 along the sampling distribution to changes in feature-space statistics. A promising direction is to push our analysis through a feature map φ (e.g. Inception features) and study W_2^2 or Bures distances in that feature space, where Gaussian approximations are empirically more plausible. This raises its own questions: the induced dynamics in feature space is not Markov, the effective score is not available, and the feature map can introduce strong anisotropy. Establishing a principled connection between kernel-weighted score-error covariance and downstream perceptual metrics therefore remains open, and we view it as a necessary step for translating operator-level insights into actionable diagnostics for large-scale diffusion models.