EFFICIENCY OF DIFFUSION MODELS FOR INFINITE-DIMENSIONAL DATA GENERATION: DIMENSION INDEPENDENT APPROXIMATION AND ESTIMATION ERRORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite the empirical success of diffusion models in generation of extremely high dimensional data such as images and videos, existing theoretical guarantees are largely confined to finite-dimensional spaces and suffer from the curse of dimensionality. This work provides the first rigorous statistical analysis of diffusion models in infinite-dimensional Hilbert spaces. To handle the absence of a Lebesgue measure in $\mathbb{R}^\infty$, we reformulate the score function using Radon-Nikodym derivatives relative to a Gaussian base measure, thereby eliminating the need for restrictive density lower-bound assumptions. By adopting the γ-*smooth function space* with different degrees of smoothness depending on directions and a novel Gaussian CDF transform, we derive polynomial-order, "dimension-independent" estimation error bounds for neural network score estimators. Our results establish a formal theoretical foundation for the scalability of diffusion models to infinite-dimensional data.

1 INTRODUCTION

Diffusion models have recently emerged as one of the most powerful generative modeling frameworks (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020; Song et al., 2021; Vahdat et al., 2021), achieving remarkable performance in diverse applications such as image synthesis (Ho et al., 2020; Song et al., 2021; Dhariwal & Nichol, 2021), text generation (Nie et al., 2025), and molecular design (Hoogeboom et al., 2022). One of the remarkable properties of diffusion models is that it works quite well for ultra high-dimensional data such as images and videos. Indeed, the capable resolution for image generation is getting higher and higher so that it can be considered as an almost infinite dimensional object. This situation raises the following questions for us:

Why do diffusion models work for extremely high dimensional and even infinite dimensional data?

Alongside standard diffusion models intended to be applied to finite dimensional data, there are also formulations of diffusion models designed for infinite dimensional objects that utilize Gaussian processes in infinite dimensional Hilbert spaces, which are specifically called *infinite dimensional diffusion models* (Lim et al., 2025; Kerrigan et al., 2023; Pidstrigach et al., 2024; Franzese et al., 2023). Actually, several studies have applied this approach to data of functional spaces based on infinite dimensional stochastic differential equations (SDEs) (Baker et al., 2024; Lim et al., 2023; Franzese & Michiardi, 2025).

However, the theoretical analyses of diffusion models have been conducted mainly for finite dimensional spaces and they do not support success of diffusion models on ultra high dimensional and infinite dimensional spaces.

(Approximation error analysis) Song et al. (2021) first proposed to bound the Kullback–Leibler (KL) divergence for the continuous-time dynamics via Girsanov's theorem. Lee et al. (2022) derived a predictive error bound which is polynomial with respect to the error in score estimate assuming the smoothness and log-Sobolev inequality (LSI) for the true density. Chen et al. (2023c) and Lee et al. (2023) eliminated the LSI with sufficient smoothness conditions. More improved analyses have been given in Chen et al. (2023a); Benton et al. (2024); Li & Yan (2025).

(Estimation error analysis) In addition to those studies that analyzed the effect of score approximation error, there are also statistical analyses that provide the estimation error of the score functions

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

with a finite number of training data points (Oko et al., 2023; Fu et al., 2024; Wibisono et al., 2024). Remarkably, they successfully showed (minimax) optimality of score based diffusion probabilistic models. One of the biggest issues in this line of research is to investigate how a low dimensional structure such as manifold assumption mitigates the curse of dimensionalities (Oko et al., 2023; Chen et al., 2023b; Tang & Yang, 2024; Azangulov et al., 2024; Yakovlev & Puchkin, 2025). In these analyses, the dimensionality of the manifold on which data exist plays crucial role, and moreover the hidden constants depend strongly on the ambient dimensionality which prevents us from naively extending their results to infinite dimensional spaces.

(Infinite dimensional analysis) Although there is a lot of remarkable theoretical research for finite dimensional models as described above, theories on infinite dimensional spaces are still at its early stage of exploration. To deal with ultra-high and even infinite dimensional data, it is essential to determine the convergence rate that can be applied to an arbitrary dimensionality in order to theoretically account for the real-world success of diffusion models. For that purpose, Pidstrigach et al. (2024); Wang & Wang (2024) provided theoretical analyses of numerical properties such as the time-discretization errors and the effects of score approximation errors. Although these analyses are useful to understand numerical errors, there is no statistical analysis (sample complexity) that is applicable to infinite dimensional diffusion models.

In this work, we analyze the approximation and estimation abilities of diffusion models with infinite-dimensional inputs. To achieve this, we adopt the γ-*smooth function space* (Okumoto & Suzuki, 2022), a function class that constrains smoothness and is well suited for analyzing models in infinite-dimensional regimes. We present solid theoretical results that extend the finite-dimensional guarantees to this broader context to discuss how diffusion models work in infinite dimensional space. The primary contributions of this paper are summarized as follows:

- We provide the first rigorous estimation error bounds for infinite-dimensional diffusion models, deriving polynomial-order dimension-independent rates for a neural network-based score matching estimator under coordinate-dependent smoothness assumptions.
- In the infinite-dimensional space $\mathbb{R}^\infty$, the absence of a non-trivial Lebesgue measure prevents the score function from being defined via the standard log-density function. To overcome this fundamental challenge, we consider the Radon-Nikodym derivative (density ratio) with respect to an infinite-dimensional Gaussian base measure. By focusing on the density ratio, we eliminate the need for lower-bound assumptions on the density, which are typically required in existing literature (Oko et al., 2023).
- We employ the γ-smooth function class to model the density ratio as an analogue to Sobolev spaces. This framework enables us to deal with coordinate-dependent smoothness, providing dimension independent approximation errors on infinite-dimensional Hilbert spaces, yielding fast convergence rates with respect to the Wasserstein distances. To bridge the gap between Gaussian measures and existing theory, we utilize a novel Gaussian cumulative distribution function (CDF) transform. Applied coordinatewise, this transform pushes the Gaussian reference measure forward to the uniform product measure on the infinite-dimensional unit cube, allowing us to fully leverage the theories developed for infinite-dimensional unit cubes (Okumoto & Suzuki, 2022; Nishimura & Suzuki, 2024).

2 Preliminaries

In this section, we introduce the mathematical basis of infinite dimensional diffusion models and the problem setting.

2.1 Finite Dimensional Diffusion Models

Before we introduce the infinite dimensional diffusion models, we start from a finite-dimensional counter part. We follow the SDE characterization of diffusion models (Song et al., 2021). Let $(B_t)_{[0,\overline{T}]}$ denote the d-dimensional Brownian motion and $\beta_t \colon [0, \overline{T}] \to \mathbb{R}_+$ be a weighting function. We denote by P_t and p_t the distribution and its density of X_t, where P_0 is the true data distribution that we want to estimate. The diffusion models consist of two processes: the *forward process* and the *reverse process*. As a forward process, $(X_t)_{[0,\overline{T}]}$ in $\mathbb{R}^d$, we consider the following Ornstein–Uhlenbeck (OU) process:

$$X_0 \sim P_0, \quad \mathrm{d}X_t = -\beta_t X_t \mathrm{d}t + \sqrt{2\beta_t}\mathrm{d}B_t.$$

Then, we can easily see that $X_t|X_0 \sim \mathcal{N}(m_t X_0, \sigma_t^2 I_d)$, where $m_t = \exp(-\int_0^t \beta_s \mathrm{d}s), \sigma_t^2 = 1 - \exp(-2\int_0^t \beta_s \mathrm{d}s)$. Following Assumption 2.5 of Oko et al. (2023), we assume that $\beta \in C^\infty([0,\infty))$ and, for fixed constants $0 < \underline{\beta} \le \overline{\beta} < \infty$, $\underline{\beta} \le \beta_t \le \overline{\beta} \quad (t \ge 0)$, $\sup_{k\in\mathbb{N}_0} \sup_{t\ge 0} |\beta^{(k)}(t)| \le 1$.

These conditions hold uniformly over time, with constants independent of sample size and truncation dimension. Writing $\tau(t) := \int_0^t \beta_s \, \mathrm{d}s$, we have $\underline{\beta} t \le \tau(t) \le \overline{\beta} t$, $e^{-\overline{\beta}t} \le m_t \le e^{-\underline{\beta}t}$, and $1 - e^{-2\underline{\beta}t} \le \sigma_t^2 \le 1 - e^{-2\overline{\beta}t}$. In particular, $\sigma_t^2 \asymp \min\{t, 1\}$, with constants depending only on $\underline{\beta}$ and $\overline{\beta}$. The *reverse process* $(Y_t)_{[0,\overline{T}]}$ traces back the forward process in a sense that the law of Y_t is equivalent to that of $X_{\overline{T}-t}$ (Haussmann & Pardoux, 1986):

$$Y_0 \sim P_{\overline{T}}, \quad \mathrm{d}Y_t = \beta_{\overline{T}-t}(Y_t + 2\nabla \log p_{\overline{T}-t}(Y_t))\mathrm{d}t + \sqrt{2\beta_{\overline{T}-t}}\mathrm{d}B_t,$$

where B_t is again another independent Brownian motion with slight abuse of notations. $\nabla \log p_t(x)$ is called the *score* that is what we estimate from training data to generate samples nearly from the true distribution because $Y_{\overline{T}} \sim P_0$. In addition to that, since it is known that P_t approaches $\mathcal{N}(0, I_d)$ exponentially fast in terms of KL-divergence (Chen et al., 2023c), we may take $\overline{T} = \tilde{\mathcal{O}}(1)$[1] and replace the initial noise distribution of Y_0 by $\mathcal{N}(0, I_d)$. Then the modified reverse process $(\hat{Y}_t)_{[0,\overline{T}]}$ is defined as

$$\hat{Y}_0 \sim \mathcal{N}(0, I_d), \quad \mathrm{d}\hat{Y}_t = \beta_{\overline{T}-t}(\hat{Y}_t + 2\hat{s}(\hat{Y}_t, \overline{T} - t))\mathrm{d}t + \sqrt{2\beta_{\overline{T}-t}}\mathrm{d}B_t.$$

Score matching The score network $\hat{s}$ is ideally estimated by minimizing the *denoising score matching loss*: $\hat{s} \in \mathrm{argmin}_{s\in\mathcal{S}} \mathbb{E}_t \left[\mathbb{E}_{x_0}\left[\mathbb{E}_{x_t|x_0}[\|s(x_t,t) - \nabla \log p_t(x_t|x_0)\|^2]\right]\right]$, where $t \sim \mathrm{Unif}[\underline{T}, \overline{T}]$, $x_0 \sim P_0$ and $x_t|x_0 \sim P_t(x_t|x_0)$. Here, $\underline{T} > 0$ is the time for *early stopping* that is known to be effective to achieve the optimal rate as the score function $\nabla \log p_t$ is less smooth than $\log p_t$ itself around $t = 0$ causing numerical instability (Oko et al., 2023; Azangulov et al., 2024). Girsanov's theorem bounds the KL-divergence from $P_{\underline{T}}$ to the law of $\hat{Y}_{\overline{T}-\underline{T}}$ by the initialization error $\mathrm{KL}(P_{\overline{T}}\|\mathcal{N}(0, I_d))$ plus the time-integrated, β_t-weighted squared score error (Chen et al., 2023c). The latter is bounded by $\overline{\beta}(\overline{T} - \underline{T})$ times the denoising score matching loss above. The early-stopping error between P_0 and $P_{\underline{T}}$ is controlled separately in Wasserstein distance. Given finite training data $\{x_{0,i}\}_{i=1}^n$ ($x_{0,i} \sim P_0$, i.i.d.), we select $\hat{s}$ to minimize the following empirical loss, which replaces $\mathbb{E}_{x_0}$ by the sample mean: $\frac{1}{n}\sum_{i=1}^n \mathop{\mathbb{E}}_{\substack{t\sim\mathrm{Unif}[\underline{T},\overline{T}] \\ x_t\sim P_t(x_t|x_{0,i})}} [\|s(x_t,t) - \nabla \log p_t(x_t|x_{0,i})\|^2]$. Here $P_t(x_t|x_{0,i})$ corresponds to $\mathcal{N}(m_t X_{0,i}, \sigma_t^2 I_d)$.

2.2 Extension to Infinite-Dimensional Setting

Here, we extend the diffusion models defined above to an infinite-dimensional one. We employ the formulation induced by Pidstrigach et al. (2024) (see also Wang & Wang (2024) for more advanced analyses) in this paper.

2.2.1 Gaussian measure in infinite-dimensional space

The biggest problem when we extend to an infinite dimensional one is that the norm of the standard Gaussian random variable $\mathcal{N}(0, I_d)$ diverges to infinity almost surely as $d \to \infty$. As in the standard way in the literature of SPDE (Da Prato & Zabczyk, 2014), Pidstrigach et al. (2024) and Lim et al. (2025) addressed this issue by introducing the Cameron–Martin space as the foundation for defining the noise process. Let $U = C^{1/2}H$ be the Cameron–Martin space associated with a Gaussian measure $\mathcal{N}(0, C)$, where C is a covariance operator on a separable Hilbert space H. Formally, for any $g, h \in H$, it is defined by $\langle g, Ch\rangle_H = \mathbb{E}\left[\langle X, g\rangle_H \langle X, h\rangle_H\right]$, where $X \sim \mathcal{N}(0, C)$. The operator C is symmetric, positive-definite, and trace-class, i.e., $\mathrm{Tr}(C) = \sum_{i\ge 1} c_i < \infty$ for its eigenvalues $\{c_i\}$, ensuring that the Gaussian measure $\mathcal{N}(0, C)$ is supported on H. It characterizes the correlation and smoothness structure of the Gaussian measure, and its inverse C^{-1} is typically unbounded, implying that $U = C^{1/2}H$ is a smaller and smoother subspace of H. Let W_t be a cylindrical Wiener process on H. Since $C^{1/2}$ is Hilbert–Schmidt, the stochastic integral $W_t^C := \int_0^t C^{1/2}\, \mathrm{d}W_s$ defines an H-valued C-Wiener process, whose increments satisfy $W_{t+\Delta t}^C - W_t^C \sim \mathcal{N}(0, \Delta t C)$. This formulation provides a rigorous way to define stochastic noise in infinite dimensions, even in the absence of standard Brownian motion. In this paper, we assume that C is diagonal, i.e., $C = \mathrm{diag}\{c_1, c_2, \ldots\}$.

[1]The symbol $\tilde{\mathcal{O}}(\cdot)$ represents the order notation hiding polylog-order, i.e., $\tilde{O}(a_n) = O(a_n \mathrm{polylog}(n))$.

2.2.2 RADON-NIKODYM FORMULATION OF THE SCORE

In infinite-dimensional Hilbert spaces, the score function $\nabla \log p_t(x)$ cannot be defined directly because there is no Lebesgue measure. To address this issue, Pidstrigach et al. (2024) formulate the reverse drift using conditional expectations, which remain well-defined in infinite dimensions. In this paper, we instead employ an equivalent formulation based on the Radon–Nikodym derivative with respect to a Gaussian reference measure, which is more convenient for approximation-theoretic analysis. To formulate the score in infinite dimensions, we introduce a Gaussian reference measure and define the corresponding Radon–Nikodym derivative.

Reference measure. For that purpose, we introduce the reference measure to define the density. By extending the finite dimensional OU-process, we consider an "anisotropic" version of the forward process (Ornstein–Uhlenbeck process) on H, governed by the following SDE:

$$X_0 \sim P_0, \quad \mathrm{d}X_t = -\beta_t X_t \mathrm{d}t + \sqrt{2\beta_t}\sqrt{C}\mathrm{d}W_t,$$

where β_t satisfies the noise-schedule assumptions above, and W_t is a cylindrical Wiener process on H. Suppose that C_0 is a diagonal, positive-definite, and trace-class operator $C_0 : H \to H$, and $C_0 = \mathrm{diag}\{c_{0,1}, c_{0,2}, \dots\}$. Let μ_{C_0} denote the Gaussian reference measure with mean zero and covariance operator C_0.

Assumption 2.1. *We assume that the initial distribution P_0 is absolutely continuous with respect to μ_{C_0}, and we denote its Radon-Nikodym derivative by $\rho_0(x) := \frac{\mathrm{d}P_0}{\mathrm{d}\mu_{C_0}}(x)$.*

We define the time-dependent covariance operator C_t as a linear combination of the initial and diffusion covariances: $C_t := m_t^2 C_0 + \sigma_t^2 C = \mathrm{diag}\{c_{t,1}, c_{t,2}, \dots\}$, where m_t and σ_t are scalar coefficients associated with the solution of the OU process. We posit that the Radon-Nikodym derivative $\rho_t(x) := \frac{\mathrm{d}P_t}{\mathrm{d}\mu_{C_t}}(x)$ is well-defined for all $t \geq 0$. This assumption is theoretically justified by the preservation of absolute continuity under the dynamics of the SDE. Specifically, the solution X_t at time t can be represented as $X_t = m_t X_0 + \xi_t$, where ξ_t is a Gaussian random variable independent of X_0 with covariance $\sigma_t^2 C$. Similarly, a random variable Z_t distributed according to the reference measure μ_{C_t} satisfies $Z_t = m_t Z_0 + \xi_t$, where $Z_0 \sim \mu_{C_0}$ and ξ_t is the same noise term. Since P_0 is absolutely continuous with respect to μ_{C_0} ($P_0 \ll \mu_{C_0}$) by Assumption 2.1, the joint distribution of (X_0, ξ_t) is absolutely continuous with respect to that of (Z_0, ξ_t). Since X_t and Z_t are obtained by the same measurable map $(x, \xi) \mapsto m_t x + \xi$ applied to these pairs, the absolute continuity is preserved for the marginals, ensuring $P_t \ll \mu_{C_t}$.

Assumption 2.2. *We assume that C_0 and C are positive definite and satisfy $K := \max\{\sup_{i\in\mathbb{N}} c_{0,i}/c_i, \sup_{i\in\mathbb{N}} c_i/c_{0,i}\} < \infty$. Moreover, there exists $0 < r_0 < 1$ such that $c_{0,i}, c_i \simeq i^{-2/r_0}$ and $\max\{c_{0,i}, c_i\} \leq 1$.*

The bounds on the ratios between C_0 and C are also assumed in (Pidstrigach et al., 2024; Wang & Wang, 2024). The second assumption is usually assumed as a source condition in the literature of kernel method (and Gaussian process) analyses (Caponnetto & De Vito, 2007). The bound $\max\{c_{0,i}, c_i\} \leq 1$ is not essential. As long as they are bounded, we may rescale the variable so that this condition is satisfied.

2.2.3 REVERSE PROCESS VIA RADON–NIKODYM DERIVATIVE

Although we cannot deal with the density function p_t w.r.t. the Lebesgue measure, we can formulate the reverse process by a score function with respect to the Radon–Nikodym derivative ρ_t:

$$Y_0 \sim P_{\overline{T}}, \quad \mathrm{d}Y_t = \beta_{\overline{T}-t}(Y_t + 2C\nabla \log \rho_{\overline{T}-t}(Y_t) - 2CC_{\overline{T}-t}^{-1} Y_t)\mathrm{d}t + \sqrt{2\beta_{\overline{T}-t}}\sqrt{C}\mathrm{d}W_t.$$

Roughly speaking, this formula can be derived from the following calculation on the finite dimensional space and taking $d \to \infty$ (see Pidstrigach et al. (2024); Wang & Wang (2024) for more details):

$$\nabla \log \frac{\mathrm{d}P_{\overline{T}-t}}{\mathrm{d}\lambda^d}(Y_t) = \nabla \log \frac{\mathrm{d}P_{\overline{T}-t}}{\mathrm{d}\mu_{C_{\overline{T}-t}}}(Y_t) + \nabla \log \frac{\mathrm{d}\mu_{C_{\overline{T}-t}}}{\mathrm{d}\lambda^d}(Y_t) = \nabla \log \rho_{\overline{T}-t}(Y_t) - C_{\overline{T}-t}^{-1} Y_t,$$

where λ^d is the Lebesgue measure on $\mathbb{R}^d$. Accordingly, we define *score function* $s(x,t)$ as

$$s(x,t) := 2C\nabla \log \rho_t(x) - 2CC_t^{-1}x, \tag{1}$$

and we consider approximating $s(x,t)$ using a deep neural network $\hat{s}(x,t)$. Then, our generative model is described by the following infinite dimensional SDE:

$$\hat{Y}_0 \sim \mathcal{N}(0, C), \quad \mathrm{d}\hat{Y}_t = \beta_{\overline{T}-t}(\hat{Y}_t + \hat{s}(\hat{Y}_t, \overline{T} - t))\mathrm{d}t + \sqrt{2\beta_{\overline{T}-t}}\sqrt{C}\mathrm{d}W_t.$$

Although we defined $(\hat{Y}_t)_t$ as an infinite dimensional process, we consider a model $\hat{s}(\cdot, t)$ that only depends on finite dimensional components, i.e., the top d_{out} components $\hat{Y}_{t,[1:d_{\mathrm{out}}]}$. This means that we only simulate the top d_{out} components in practice.

Score matching estimator. Next, we introduce our empirical risk to estimate the score function. Since the KL-divergence between distributions on infinite dimensional space can be easily infinity (Da Prato, 2006), we employ the Wasserstein metric as the performance measure instead:

$$\mathcal{W}_q(P, Q) := \left(\inf_\pi \int \|x - y\|_H^q \mathrm{d}\pi(x, y)\right)^{1/q},$$

where $1 \le q < \infty$, and π runs over all measures which has marginals P and Q. The Wasserstein error decomposition in Lemma 5.2 bounds the error of the terminally projected generator in terms of the weighted score estimation error, together with initialization, early-stopping, coordinate-truncation, and projection errors. This motivates estimating the score by the empirical loss introduced below.

Score matching for infinite dimensional diffusion. We construct a score estimator $\hat{s}$ so that it takes d_{out}-dimensional input and provide d_{out}-dimensional vector as an approximate of the true score. Since we cannot implement an infinite dimensional SDE in practice, the score estimator $\hat{s}$ is intended to simulate the dynamics of the top d_{out}-dimensional elements of X_t. Suppose that $\tilde{X}_t$ is the sub-vector of the top d_{out} components of X_t where X_t follows the true dynamics and let $\tilde{P}_t(\cdot|x_0)$ be the d_{out}-dimensional marginal distribution of $\tilde{X}_t$ given $X_0 = x_0$, with density $\tilde{p}_t(\cdot|x_0)$. Noticing that each component of X_t is independent to each other conditioned by X_0, we have an explicit and exact formula of the marginal score $2C\nabla \log \tilde{p}_t(\cdot|x_0)$. Throughout this finite-dimensional formulation, we write C for $C_{[1:d_{\mathrm{out}}]}$ by abuse of notation; in particular, C^{-1} denotes the inverse of this finite-dimensional covariance matrix.

Then, our estimator can be defined as the minimizer of the following empirical score matching loss in our neural network class:

$$\hat{R}_A(s) := \frac{1}{n}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\substack{t\sim \mathrm{Unif}[\underline{T},\overline{T}] \\ \tilde{x}_t \sim \tilde{P}_t(\tilde{x}_t|x_{0,i})}} \left[\|s(\tilde{x}_t, t) - 2C\nabla \log \tilde{p}_t(\tilde{x}_t|x_{0,i})\|_A^2\right],$$

where $\|v\|_A^2 = v^\top A v$. We take $A = I$ by default. On the other hand, to estimate a finite-dimensional approximation, it is still natural to minimize the KL-divergence, for which we take $A = C^{-1}$. The corresponding population score error in the C^{-1}-weighted norm controls the finite-dimensional KL-divergence through Girsanov's theorem, with the time-integration factor and the initialization error made explicit in Appendix H. For $A = C^{-1}$, the coordinate transform $\hat{Y}_t \leftarrow C^{-1/2}\hat{Y}_t$ reduces the reverse process and the empirical risk to the standard ones in finite-dimensional spaces. Let $\mathcal{S}$ be the set of our neural network model as defined below. Then, we define $\hat{s} = \arg\min_{s\in\mathcal{S}} \hat{R}_A(s)$ with **(Option 1)** $A = I$ or **(Option 2)** $A = C^{-1}$.

Terminal projection. For the finite-dimensional generators, let $\hat{Y}_t^{(d_{\mathrm{out}})}$ denote the simulated reverse process, initialized from $\mathcal{N}(0, C_{[1:d_{\mathrm{out}}]})$ and driven by the learned score $\hat{s}$. In these finite-dimensional statements, we write $\hat{Q}_{\overline{T}-\underline{T}}$ for the law of its unprocessed terminal output $\hat{Y}_{\overline{T}-\underline{T}}^{(d_{\mathrm{out}})}$. For $R > 0$, define the radial projection onto the Euclidean ball of radius R by $\Pi_R(y) = \frac{R}{R\vee\|y\|}\, y$. For $n \ge 2$, set $R_n = K_{\mathrm{proj}}\sqrt{\log n}$, where $K_{\mathrm{proj}} > 0$ is a sufficiently large fixed constant independent of n and d_{out}, and define $\hat{Q}^{\mathrm{proj}}_{\overline{T}-\underline{T}} := (\Pi_{R_n})_{\#}\hat{Q}_{\overline{T}-\underline{T}}$. Here, $(\Pi_{R_n})_{\#}Q$ denotes the law of $\Pi_{R_n}(Y)$ for $Y \sim Q$. When compared with P_0 on H, this distribution is embedded into H by padding the coordinates beyond d_{out} with zeros. The projection is applied only to the final output and does not change the training losses or the reverse SDE.

2.3 Class of neural networks and score model

As usual in approximation theories of neural networks (Yarotsky, 2017; Schmidt-Hieber, 2020; Suzuki, 2019), we introduce a class of deep neural networks with the ReLU activation function $\mathrm{ReLU}(x) = \max\{0, x\}$ (operated element-wise for a vector) (Nair & Hinton, 2010; Glorot et al., 2011) with a sparsity constraint (on the number of non-zero parameters) as the hypothesis class of scores.

Definition 2.3 (Deep neural networks). *A class of neural networks $\Phi(L, W, S, B; D_{\mathrm{in}}, D_{\mathrm{out}})$ with height L, width W, sparsity constraint S, norm constraint B, and input-output dimensionalities*

$D_{\text{in}}, D_{\text{out}}$ *is defined as* $\Phi(L, W, S, B; D_{\text{in}}, D_{\text{out}}) := \{(A^{(L)}\text{ReLU}(\cdot) + b^{(L)}) \circ \cdots \circ (A^{(1)}x + b^{(1)}) | \, A^{(i)} \in \mathbb{R}^{W_{i+1} \times W_i}, b^{(i)} \in \mathbb{R}^{W_{i+1}}, \sum_{i=1}^{L}(\|A^{(i)}\|_0 + \|b^{(i)}\|_0) \leq S, \, \max_i \|A^{(i)}\|_\infty \vee \|b^{(i)}\|_\infty \leq B, \max_{2 \leq i \leq L} W_i \leq W, W_{L+1} = D_{\text{out}}, W_1 = D_{\text{in}}\}$. *When the dimensionalities* D_{in} *and* D_{out} *are obvious from the context, we may simply write as* $\Phi(L, W, S, B)$.

The score network maps $(x, t) \in \mathbb{R}^{d_{\text{out}}} \times [\underline{T}, \overline{T}]$ to $\mathbb{R}^{d_{\text{out}}}$. To impose output bounds on the entire hypothesis class, we append fixed coordinatewise clipping layers to both subnetworks.

For $b = (b_i)_{i=1}^k \in (0, \infty)^k$, define $[\text{clip}_b(z)]_i := -b_i + \text{ReLU}(z_i + b_i) - \text{ReLU}(z_i - b_i)$, $i = 1, \ldots, k$. Thus $[\text{clip}_b(z)]_i = \min\{b_i, \max\{-b_i, z_i\}\}$. This map is realized exactly by a fixed ReLU network.

Let $n \geq 2$, and set $u_n := \sqrt{\log(en)}$, $b_{\text{grad},i} := R_{\text{grad}}\sqrt{c_i}\, i^{-1/2-\delta_0}$, $b_{\text{linear},i} := R_{\text{linear}}\sqrt{c_i}\, u_n$, where $R_{\text{grad}}, R_{\text{linear}} > 0$ are sufficiently large fixed constants independent of n and d_{out}. Write $b_{\text{grad}} = (b_{\text{grad},i})_{i=1}^{d_{\max}}$ and $b_{\text{linear}} = (b_{\text{linear},i})_{i=1}^{d_{\text{out}}}$. Let $E_{d_{\max}, d_{\text{out}}} : \mathbb{R}^{d_{\max}} \to \mathbb{R}^{d_{\text{out}}}$ denote zero-padding.

We define the hypothesis class by

$$\mathcal{S} := \big\{(x,t) \mapsto \phi_{\text{score}}(x,t) = \phi_{\text{score,grad}}(x,t) + \phi_{\text{score,linear}}(x,t) :$$
$$\phi_{\text{score,grad}}(x,t) = E_{d_{\max}, d_{\text{out}}} \, \text{clip}_{b_{\text{grad}}}\big(g(x_{[1:d_{\max}]}, t)\big), \; \phi_{\text{score,linear}}(x,t) = \text{clip}_{b_{\text{linear}}}\big(\ell(x,t)\big),$$
$$g \in \Phi(L_{\text{grad}}, W_{\text{grad}}, S_{\text{grad}}, B_{\text{grad}}; d_{\max} + 1, d_{\max}),$$
$$\ell \in \Phi(L_{\text{linear}}, W_{\text{linear}}, S_{\text{linear}}, B_{\text{linear}}; d_{\text{out}} + 1, d_{\text{out}})\big\}.$$

The networks $\phi_{\text{score,grad}}$ and $\phi_{\text{score,linear}}$ approximate the relative-score component $2C\nabla \log \rho_t(x)$ and the linear component $-2CC_t^{-1}x$, respectively. Both components are learned by neural networks. The clipping thresholds are fixed in advance and are not optimized during training.

3 γ-SMOOTH FUNCTION CLASS FOR THE TRUE DENSITY RATIO

Here, we introduce a function class where we assume that the true density ratio ρ_0 is included. The concept is based on the idea of Okumoto & Suzuki (2022), but the domain is extended from $[0,1]^\infty$ to $\mathbb{R}^\infty$. To connect the theory with respect to the uniform distribution on $[0,1]^\infty$ to that of the Gaussian measure on $\mathbb{R}^\infty$, we utilize a coordinate transformation induced by the *Gaussian CDF*.

For a set $\mathbb{D}$, let $\mathbb{D}^\infty := \{(s_1, \ldots, s_i, \ldots) : s_i \in \mathbb{D}\}$. For $s \in \mathbb{D}^\infty$, let $\text{supp}(s) := \{i \in \mathbb{N} : s_i \neq 0\}$. Let $\mathbb{N}_0^\infty := \{l \in (\mathbb{N} \cup \{0\})^\infty : |\text{supp}(l)| < \infty\}$ and let $\mathbb{Z}_0^\infty := \{l \in \mathbb{Z}^\infty : |\text{supp}(l)| < \infty\}$. First, we introduce trigonometric bases functions that compose the γ-smooth function class. For a given $l \in \mathbb{Z}_0^\infty$, define $\psi_l : \mathbb{R}^\infty \to \mathbb{R}$ as

$$\psi_l(x) := \textstyle\prod_{i=1}^\infty \psi_{l_i}(x_i),$$

$$\psi_{l_i}(x_i) := \cos\left(2\pi l_i \Phi\left(x_i/\sqrt{c_{0,i}}\right)\right) - \sin\left(2\pi l_i \Phi\left(x_i/\sqrt{c_{0,i}}\right)\right) = \sqrt{2}\cos\left(2\pi l_i \Phi\left(x_i/\sqrt{c_{0,i}}\right) + \pi/4\right).$$

Here, $\Phi : \mathbb{R} \to (0,1)$ means the cumulative distribution function of standard normal distribution, which is defined as $\Phi(x) := \int_{-\infty}^x (2\pi)^{-1/2} e^{-t^2/2} \mathrm{d}t$. Using this distribution function Φ, we can transfer the argument on the L^2-space on $[0,1]$ equipped with the uniform distribution to that on the L^2-space on $\mathbb{R}$ equipped with a Gaussian distribution. Let $L^2(\mathbb{R}^\infty, \mu_{C_0}) := \left\{f : \mathbb{R}^\infty \to \mathbb{R} \middle| \int_{\mathbb{R}^\infty} f^2(x) \mathrm{d}\mu_{C_0}(x) < \infty\right\}$ equipped with the inner product $\langle f, g \rangle := \int_{\mathbb{R}^\infty} f(x) g(x) \mathrm{d}\mu_{C_0}(x)$ for $f, g \in L^2(\mathbb{R}^\infty, \mu_{C_0})$. Then, $(\psi_l)_{l \in \mathbb{Z}_0^\infty}$ forms a complete orthonormal system of $L^2(\mathbb{R}^\infty, \mu_{C_0})$, that is, $f \in L^2(\mathbb{R}^\infty, \mu_{C_0})$ can be expanded as $f(x) = \sum_{l \in \mathbb{Z}_0^\infty} \langle f, \psi_l \rangle \psi_l(x)$. For $s \in \mathbb{N}_0^\infty$, let $J(s)$ be $J(s) := \{l \in \mathbb{Z}_0^\infty : \lfloor 2^{s_i - 1} \rfloor \leq |l_i| < 2^{s_i}\}$ and let $\delta_s(f) : \mathbb{R}^\infty \to \mathbb{R}$ be $\delta_s(f)(\cdot) := \sum_{l \in J(s)} \langle f, \psi_l \rangle \psi_l(\cdot)$, which can be seen as the frequency component of f of frequency $|l_i| \simeq 2^{s_i}$toward each coordinate. We also define $\|f\|_p := \left(\int_{\mathbb{R}^\infty} |f|^p \mathrm{d}\mu_{C_0}(x)\right)^{1/p}$ for $p \geq 1$. Then, we define a function space with a general smoothness configuration as follows.

Definition 3.1 (γ-smooth function space)**.** *For a given* $\gamma : \mathbb{N}_0^\infty \to \mathbb{R}_+$ *which is monotonically non-decreasing with respect to each coordinate and for* $p, q \geq 1$*, we define the* γ*-smooth space as*

$$\mathcal{F}_{p,q}^\gamma(\mathbb{R}^\infty) := \left\{f \in L^p(\mathbb{R}^\infty, \mu_{C_0}) \;\middle|\; \|f\|_{\mathcal{F}_{p,q}^\gamma} < \infty\right\}, \quad \|f\|_{\mathcal{F}_{p,q}^\gamma} := \left(\sum_{s \in \mathbb{N}_0^\infty} 2^{q\gamma(s)} \|\delta_s(f)\|_p^q\right)^{1/q}.$$

Recall that $\delta_s(f)$ represents the frequency component associated with the frequency $(2^{s_i})_{i=1}^\infty$, and then the norm of the γ-smooth space imposes weight $2^{q\gamma(s)}$ on each frequency component associated with s. In that sense, $\gamma(s)$ controls the weight of each frequency component.

Okumoto & Suzuki (2022) considered the γ-smooth space whose domain is limited in $[0,1]^\infty$ and based on Lebesgue measure, demonstrating that dilated CNNs can achieve approximation and estimation errors that depend only on smoothness, not dimensionality, for infinite-dimensional inputs. In addition, Nishimura & Suzuki (2024) showed the minimax optimality of dilated CNNs for estimating a map between infinite-dimensional input and output spaces. Ponnoprat & Imaizumi (2025) employed the γ-smooth space to discuss the minimax optimality of estimating the optimal transport map in an infinite dimensional space where the optimal transport map is included in the γ-smooth space.

4 Approximation of the True Score

In this section, we consider approximating the true score $s(x,t)$ using a deep neural network and derive the approximation error bound.

Assumption 4.1 (Function class specification)**.** *The true density ratio belongs to the ball of radius $R_\rho\,(>1)$ centered at zero in the γ-smooth function class: $\rho_0 \in \mathcal{F}^\gamma_{p,q}(\mathbb{R}^\infty)$ and $\|\rho_0\|_{\mathcal{F}^\gamma_{p,q}} \le R_\rho$.*

The γ-smooth structure provides a theoretically meaningful abstraction of the effective low-dimensional structures observed in practical diffusion models. In latent diffusion models (Vahdat et al., 2021; Rombach et al., 2022), data are first encoded into a latent representation by a VAE, diffusion is performed in the latent space, and the samples are decoded back to the original space. Such latent representations are typically organized so that low-index coordinates capture dominant semantic information, while higher-index coordinates have smaller influence. Moreover, U-Net architectures, which are widely used in modern diffusion models (Ho et al., 2020; Dhariwal & Nichol, 2021), have recently been theoretically related to multi-resolution decompositions closely connected to wavelet analysis (Williams et al., 2023), where coarse-scale components capture dominant global structures while fine-scale high-frequency components contribute progressively smaller corrections. Similarly, diffusion models for functional data are often applied to spectral coefficients induced by Fourier representations (Lim et al., 2025) or operator eigenfunction expansions (Gallon et al., 2026), where higher-order coefficients naturally exhibit decay.

Assumption 4.2 (Regularity condition)**.** *We assume that $\|\rho_0 - 1\|_{L^\infty} \le 1/2$ and $C\nabla\rho_0$ is L-Lipschitz continuous. Moreover, we assume that $\|\partial_{x_i}\rho_0\|_{L^\infty} \lesssim i^{-1/2-\delta_0}/\sqrt{c_{0,i}}$ $(i \in \mathbb{N})$ for some $\delta_0 > 0$.*

The assumption on $\partial_{x_i}\rho_0$ is a natural condition because the density ratio ρ_0 is assumed to become progressively smoother along higher-index coordinates. Moreover, the Gaussian conditional representation and Lemma A.7 give $\|\partial_{x_i}\rho_t\|_{L^\infty} \le \frac{m_t c_{0,i}}{c_{t,i}}\|\partial_{x_i}\rho_0\|_{L^\infty}$. By Assumption 2.2, $c_{t,i} \asymp c_{0,i} \asymp c_i$ uniformly in t and i. Together with $\rho_t \ge 1/2$, this yields $|2c_i\partial_{x_i}\log\rho_t(x)| \lesssim m_t\sqrt{c_i}\,i^{-1/2-\delta_0} \le \sqrt{c_i}\,i^{-1/2-\delta_0}$, which justifies the clipping on $\mathcal{S}$.

To obtain a concrete rate of convergence, we put the following condition on $\gamma(s)$.

Assumption 4.3 (Mixed and anisotropic smoothness)**.** *We assume that there exists $a \in \mathbb{R}^\infty_{>0}$ such that either of the following holds:*

- **Mixed smoothness :** $\gamma(s) = \langle a, s\rangle = \sum_{i=1}^\infty a_i s_i$.
- **Anisotropic smoothness :** $\gamma(s) = \max_{i\in\mathbb{N}}\{a_i s_i\}$.

Moreover, we assume that there exists $0 < \eta < \infty$ such that the smoothness parameter $a = (a_i)_{i=1}^\infty$ satisfies $a_i \simeq i^\eta$ for $0 < \eta < \infty$. We additionally assume that the smoothness sequence is nondecreasing: $0 < a_1 \le a_2 \le \cdots$.

The intuition of this assumption is that a_i represents the smoothness toward the i-th coordinate and the smoothness increases as the coordinate index increases. Accordingly, we define $a^\dagger := a_1$ for mixed smoothness and $a^\dagger := (\sum_{i=1}^\infty a_i^{-1})^{-1}$ for anisotropic smoothness. Then, $a^\dagger$ represents the effective smoothness which summarizes how strongly high-index coordinates are penalized. Larger $a^\dagger$ means easier approximation and estimation.

Theorem 4.4 (Approximation error of the true score by FNNs)**.** *Let $v := \max\{1/p - 1/2, 0\}$ and $0 < r_0 < 1$. Assume that Assumptions 2.2, 4.1, 4.2, and 4.3 hold, with $v/a^\dagger < 1$. For sufficiently large $\Gamma \ge 2$, we may choose the network budgets as*

$$L_{\mathrm{score}} \lesssim 2^{2\Gamma/a^\dagger}\Gamma^{2\max\{1,1/\eta\}}\log^2\Gamma,\ W_{\mathrm{score}} \lesssim 2^{(4+\log_2 3)\Gamma/a^\dagger}\Gamma^{1/\eta+4\max\{1,1/\eta\}}\log^4\Gamma + d_{\mathrm{out}}\Gamma^3,$$

$$S_{\mathrm{score}} \lesssim 2^{(5+\log_2 3)\Gamma/a^\dagger}\Gamma^{1/\eta+5\max\{1,1/\eta\}}\log^5\Gamma + d_{\mathrm{out}}\Gamma^4 + d_{\mathrm{out}}\,2^{2\Gamma/a^\dagger}\Gamma^{2\max\{1,1/\eta\}}\log^2\Gamma,$$

$$\log B_{\mathrm{score}} \lesssim 2^{\Gamma/a^\dagger}\Gamma^{\max\{1,1/\eta\}}\log\Gamma + \log d_{\mathrm{out}}$$

Then there exists a score network $\phi_{\text{score}} \in \Phi(L_{\text{score}}, W_{\text{score}}, S_{\text{score}}, B_{\text{score}}; d_{\text{out}}+1, d_{\text{out}})$ *that satisfies, for every* $t \in [\underline{T}, \overline{T}]$,

$$\int_x \left\|\phi_{\text{score}}(x_{[1:d_{\text{out}}]}, t) - s(x,t)\right\|_H^2 \, \mathrm{d}P_t(x) \lesssim \frac{\Gamma}{\sigma_t^2} 2^{-2(1-v/a^\dagger)\Gamma} + \frac{r_0}{2-r_0} d_{\text{out}}^{1-2/r_0}.$$

The proof can be found in Appendix D. This theorem shows the trade-off between the model size and the approximation error through the choice of Γ and d_{out}. We see that as Γ and d_{out} become large, the model size becomes large accordingly. In Theorem 5.1, the parameters are chosen adaptively depending on the sample size n. The precise selection of these parameters in each case is provided in Appendix H.

Proof Sketch. We approximate the first d_{out} coordinates of the score function $s^i(x,t) = 2c_i\partial_{x_i}\log\rho_t(x) - 2c_i c_{t,i}^{-1}x_i$ by neural networks and truncate the remaining coordinates. The truncation error is controlled by the covariance decay assumption $c_i \simeq i^{-2/r_0}$, yielding the term d_{out}^{1-2/r_0}.

For the first gradient term, we use $2c_i\partial_{x_i}\log\rho_t(x) = 2c_i\partial_{x_i}\rho_t(x)/\rho_t(x)$, and separately approximate the numerator and denominator by neural networks. Since ρ_0 belongs to the γ-smooth class, its basis expansion is dominated by low-frequency components. Accordingly, we truncate the expansion at frequency level Γ and define $R_\Gamma(\rho_0) := \sum_{s:\gamma(s)<\Gamma}\sum_{l\in J(s)}\langle\rho_0,\psi_l\rangle\psi_l$. The approximation error caused by discarding the remaining high-frequency components is controlled by the decay condition imposed through the γ-smooth norm, yielding the error term $2^{-2(1-v/a^\dagger)\Gamma}$. Applying the OU semigroup preserves this structure, so $\rho_t = \mathcal{P}_t\rho_0$ and $\partial_{x_i}\rho_t$ can be approximated by neural networks constructed from low-frequency basis convolutions. The approximation error decays as $2^{-2(1-v/a^\dagger)\Gamma}$. Finally, the second linear term $-2c_i c_{t,i}^{-1}x_i$ is approximated by combining neural network approximations of m_t^2, reciprocal functions, and multiplication.

5 Estimation Error Bound

In this section, we finally evaluate how close the distribution of our estimated generative model is to the true distribution. For that purpose, we establish an estimation error bound based on the approximation error bound given in the previous section.

We now bound the Wasserstein distance between the estimated and true distributions. The proof is given in Appendix H.

Theorem 5.1. *Assume that* $\gamma : \mathbb{N}_0^\infty \to \mathbb{R}_{>0}$ *satisfies* $v/a^\dagger < 1$ *and Assumption 2.1, 2.2, 4.1, 4.2 and 4.3 holds where* $0 < r_0 < 1$. *We set* $\kappa := 9$ *and* $\kappa := \log_2 3 + 8a^\dagger/a_1$ *for mixed and anisotropic smoothness, respectively. Then we have the following estimation error bounds. Here,* $\mathbb{E}_n$ *denotes expectation over the training data.*

$$\mathbb{E}_n\left[\mathcal{W}_1^2\left(P_0, \hat{Q}^{\text{proj}}_{\overline{T}-\underline{T}}\right)\right] = \begin{cases} \tilde{\mathcal{O}}\left(n^{-\min\left\{\frac{2-r_0}{4-r_0}\frac{2(a^\dagger-v)}{2(a^\dagger-v)+\kappa}, \frac{2-r_0}{4}\right\}}\right) & \textit{Case(i): the estimator of Option 1} \\ \tilde{\mathcal{O}}\left(n^{-\min\left\{\frac{(2-r_0)(a^\dagger-v)}{2(a^\dagger-v)+\kappa}, \frac{2-r_0}{2+r_0}\right\}}\right) & \textit{Case(ii): the estimator of Option 2.} \end{cases}$$

This theorem shows that diffusion models achieve polynomial-order estimation rates even in infinite-dimensional Hilbert spaces under coordinate-dependent smoothness assumptions on the true density ratio ρ_0. Importantly, the convergence rates depend on the effective smoothness $a^\dagger$ and the covariance decay parameter r_0, rather than on an ambient dimensionality itself. The minimization structure in the rates reflects the competition between the regularity of the true density ratio and the covariance structure of the infinite-dimensional diffusion process. When $a^\dagger - v$ is sufficiently large relative to κ, the exponents of the squared $\mathcal{W}_1$ bounds saturate at $(2-r_0)/4$ in Case (i) and $(2-r_0)/(2+r_0)$ in Case (ii). These ceilings arise from balancing the covariance tail d_{out}^{1-2/r_0} against the dimension-dependent estimation terms. Thus faster covariance decay (smaller r_0) improves these ceilings, whereas additional smoothness alone does not improve the resulting polynomial exponents. This is a feature of the present upper bounds.

For comparison with finite-dimensional rates, formally take $(a_1, \ldots, a_d, a_{d+1}, \ldots) = (s, \ldots, s, \infty, \ldots)$ in the anisotropic setting. Then $a^\dagger = s/d$, $a_1 = s$, and $v = 0$ when $p \geq 2$. As $r_0 \downarrow 0$, the polynomial exponent in Case (ii) approaches $2s/(2s + d\log_2 3 + 8)$. Taking a square root by Jensen's inequality gives an expected $\mathcal{W}_1$ exponent approaching $s/(2s + d\log_2 3 + 8)$. This is smaller even than the TV-based exponent $s/(2s+d)$ in Oko et al. (2023). Their Theorem 5.4 obtains

the sharper expected $\mathcal{W}_1$ exponent $(s+1-\delta)/(2s+d)$ for arbitrarily small fixed $\delta > 0$ under their assumptions. Thus both network complexity and the conversion from score error to Wasserstein error matter in this comparison; the distribution classes and risk criteria must also be distinguished.

Appendix H.2 also provides $\mathcal{W}_2$ guarantees for a modified estimator under additional regularity assumptions, as well as projection-based guarantees for the original estimators without those additional assumptions.

Proof Sketch. For Cases (i) and (ii), we combine terminal projection, a tail bound for the true distribution, Pinsker's inequality, and Girsanov's theorem to reduce distribution estimation to score estimation. This argument does not require a tail bound for the unprojected learned distribution.

Write $C_{d_{\mathrm{out}}} := C_{[1:d_{\mathrm{out}}]}$ and $s_{d_{\mathrm{out}}}(x,t) := 2C_{d_{\mathrm{out}}} \nabla \log \tilde{p}_t(x)$, where $\tilde{P}_t$ is the d_{out}-dimensional marginal of the true distribution, with density $\tilde{p}_t$. For a symmetric positive definite matrix A, define

$$\bar{R}_{T_0,T_1;A}(F,F') := \mathbb{E}_{t\sim \mathrm{Unif}[T_0,T_1],\tilde{X}_t\sim\tilde{P}_t}\left[\|F(\tilde{X}_t,t) - F'(\tilde{X}_t,t)\|_A^2\right].$$

Then we obtain the following decomposition.

Lemma 5.2 (Wasserstein error decomposition)**.** *Under the standing assumptions, let* $0 < \underline{T} \le 1$ *and* $\overline{T} > \underline{T}$*. For* $n \ge 2$ *and* $R_n = K_{\mathrm{proj}}\sqrt{\log n}$ *with a sufficiently large fixed* K_{proj}*, it holds that*

$$\mathbb{E}_n\left[\mathcal{W}_1^2\left(P_0, \hat{Q}^{\mathrm{proj}}_{\overline{T}-\underline{T}}\right)\right] \lesssim \underline{T} + d_{\mathrm{out}}^{1-2/r_0} + n^{-2} + \log n\left[(\overline{T}-\underline{T})\,\mathbb{E}_n\left[\bar{R}_{\underline{T},\overline{T};C^{-1}_{d_{\mathrm{out}}}}(\hat{s}, s_{d_{\mathrm{out}}})\right] + K_{\mathrm{init}}\right],$$

where $K_{\mathrm{init}} := \mathrm{KL}\left(\tilde{P}_{\overline{T}} \,\middle\|\, \mathcal{N}(0, C_{d_{\mathrm{out}}})\right)$*. The projected distribution is identified with its zero-padding embedding into* H*.*

The proof is given in Appendix H. The terms correspond to early stopping, coordinate truncation, projection bias, score estimation, and initialization. The factor $\overline{T} - \underline{T}$ converts the time-averaged score risk into a time integral.

The score error is controlled by the oracle inequality in Proposition H.3, combining the approximation bound with a finite covering argument, localization, and Bernstein's inequality. Option 2 directly controls the required weighted error, whereas Option 1 requires an additional comparison with the unweighted error. Balancing these bounds with the remaining terms yields the estimation rates.

6 Conclusion

We analyzed approximation and statistical estimation errors of diffusion models for distributions on infinite-dimensional Hilbert spaces. Using density ratios relative to Gaussian measures and coordinate-dependent γ-smoothness, we obtained polynomial Wasserstein error bounds governed by effective smoothness and covariance decay. The comparison between Options 1 and 2 shows the benefit of matching the training loss to the covariance-weighted score error used in the distributional bound.

The rates in Theorem 5.1 also exhibit saturation: once the effective smoothness is sufficiently large relative to κ, their exponents are limited by r_0. This reflects the balance between the covariance tail discarded by coordinate truncation and the statistical cost of increasing the output dimension. Faster covariance decay permits faster rates in this regime. This saturation describes our upper bounds, rather than a minimax lower bound for the distribution class.

Limitations and future work The gap from the finite-dimensional results of Oko et al. (2023) has several sources in our analysis. The frequency-dependent network construction enlarges the covering-number bound, producing the complexity factor κ. Moreover, the KL-to-TV-to-$\mathcal{W}_1$ argument does not exploit the sharper Wasserstein analysis used in that work. In the infinite-dimensional setting, covariance weighting and coordinate truncation introduce further costs. Improving the approximation architecture and its entropy bound, and deriving a sharper Wasserstein comparison, are therefore distinct directions for improving the rates. Further directions include establishing matching lower bounds to identify intrinsic covariance-dependent losses and quantifying numerical discretization error for the continuous-time reverse dynamics analyzed here.

AI USE STATEMENT

In this work, we used generative AI tools for formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting with translation, and assisting in the writing of proofs. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, interpreting results, or proposing or refining hypotheses. Designing or providing feedback on research methodology or experiments, generating synthetic data sets, implementing methods, cleaning and reformatting datasets, and supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for summarizing or analysing existing literature and editing the paper to improve readability. We have reviewed all AI-assisted work. We checked the AI-assisted mathematical claims and proofs for correctness and logical consistency. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work develops theoretical approximation and estimation guarantees for diffusion models in infinite-dimensional Hilbert spaces. It does not involve human-subject studies, personal data, or empirical datasets. The results concern statistical accuracy under mathematical assumptions and do not establish privacy, fairness, or safety guarantees for deployed generative models. Applications of these models should therefore assess data provenance, privacy, bias, and potential misuse in their specific contexts.

REPRODUCIBILITY STATEMENT

The results in this paper are theoretical and do not depend on numerical experiments or empirical datasets. The probabilistic setup and score-matching objectives are described in Section 2, and the assumptions for the approximation and estimation results are stated in Theorems 4.4 and 5.1 and the assumptions they reference. Appendix C presents the neural network building blocks, while Appendices D, E, and G detail the score approximation construction and its analysis. The estimation arguments and choices of truncation level, output dimension, and time endpoints are given in Appendix H.

REFERENCES

Iskander Azangulov, George Deligiannidis, and Judith Rousseau. Convergence of diffusion models under the manifold hypothesis in high-dimensions. *arXiv preprint arXiv.org/abs/:2409.18804*, 2024. URL https://arxiv.org/abs/2409.18804.

Elizabeth Louise Baker, Gefan Yang, Michael L. Severinsen, Christy Anna Hipsley, and Stefan Sommer. Conditioning non-linear and infinite-dimensional diffusion processes. In *Advances in Neural Information Processing Systems*, volume 37, pp. 10801–10826. Curran Associates, Inc., 2024.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=r5njV3BsuD.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4735–4763. PMLR, 23–29 Jul 2023a.

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

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023c.

Giuseppe Da Prato. *An Introduction to Infinite-Dimensional Analysis*. Springer, Berlin, Heidelberg, 1 edition, 2006. ISBN 978-3-540-29021-6. doi: 10.1007/3-540-29021-4.

Giuseppe Da Prato and Jerzy Zabczyk. *Stochastic Equations in Infinite Dimensions*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2 edition, 2014.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 8780–8794, 2021.

Lutz Dümbgen. Bounding Standard Gaussian Tail Probabilities, 2010. URL https://arxiv.org/abs/1012.2063.

Giulio Franzese and Pietro Michiardi. Generative diffusion models in infinite dimensions: a survey. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 383(2299):20240322, 06 2025. doi: 10.1098/rsta.2024.0322. URL https://doi.org/10.1098/rsta.2024.0322.

Giulio Franzese, Giulio Corallo, Simone Rossi, Markus Heinonen, Maurizio Filippone, and Pietro Michiardi. Continuous-time functional diffusion processes. In *Advances in Neural Information Processing Systems*, volume 36, pp. 37370–37400. Curran Associates, Inc., 2023.

Hengyu Fu, Zhuoran Yang, Mengdi Wang, and Minshuo Chen. Unveil conditional diffusion models with classifier-free guidance: A sharp statistical theory. *ArXiv*, abs/2403.11968, 2024. URL https://api.semanticscholar.org/CorpusID:268532529.

Davide Gallon, Philippe von Wurstemberger, Patrick Cheridito, and Arnulf Jentzen. Physics-informed diffusion models in spectral space, 2026. URL https://arxiv.org/abs/2602.09708.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 315–323, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.

R. D. Gordon. Values of Mills' Ratio of Area to Bounding Ordinate and of the Normal Probability Integral for Large Values of the Argument. *The Annals of Mathematical Statistics*, 12(3):364–366, 1941. doi: 10.1214/aoms/1177731721.

U. G. Haussmann and É. Pardoux. Time reversal of diffusions. *The Annals of Probability*, 14(4): 1188–1205, October 1986. doi: 10.1214/aop/1176992362.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In *International conference on machine learning*, pp. 8867–8887. PMLR, 2022.

Ioannis Karatzas, Ioannis Karatzas, Steven Shreve, and Steven E Shreve. *Brownian motion and stochastic calculus*, volume 113. Springer Science & Business Media, 1991.

Gavin Kerrigan, Justin Ley, and Padhraic Smyth. Diffusion generative models in infinite dimensions. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 9538–9563. PMLR, 2023.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 22870–22882. Curran Associates, Inc., 2022.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory, February 20-23, 2023, Singapore*, volume 201 of *Proceedings of Machine Learning Research*, pp. 946–985. PMLR, 2023.

Gen Li and Yuling Yan. O(d/T) Convergence Theory for Diffusion Probabilistic Models under Minimal Assumptions. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4EjdYiNRzE.

Jae Hyun Lim, Nikola B. Kovachki, Ricardo Baptista, Christopher Beckham, Kamyar Azizzadenesheli, Jean Kossaifi, Vikram Voleti, Jiaming Song, Karsten Kreis, Jan Kautz, Christopher Pal, Arash Vahdat, and Anima Anandkumar. Score-based diffusion models in function space. *Journal of Machine Learning Research*, 26(158):1–62, 2025. URL http://jmlr.org/papers/v26/23-1472.html.

Sungbin Lim, EUN BI YOON, Taehyun Byun, Taewon Kang, Seungwoo Kim, Kyungjae Lee, and Sungjoon Choi. Score-based Generative Modeling through Stochastic Evolution Equations in Hilbert Spaces. In *Advances in Neural Information Processing Systems*, volume 36, pp. 37799–37812. Curran Associates, Inc., 2023.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pp. 807–814, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21(174):1–38, 2020.

R. J. Nessel and G. Wilmes. Nikolskii-type inequalities for trigonometric polynomials and entire functions of exponential type. *Journal of the Australian Mathematical Society*, 25(1):7–18, 1978. doi: 10.1017/S1446788700038878.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models, 2025. URL https://arxiv.org/abs/2502.09992.

Yuto Nishimura and Taiji Suzuki. Minimax optimality of convolutional neural networks for infinite dimensional input-output problems and separation from kernel methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=EW8ZExRZkJ.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning*, pp. 26517–26582. PMLR, 2023.

Sho Okumoto and Taiji Suzuki. Learnability of convolutional neural networks for infinite dimensional input via mixed and anisotropic smoothness. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=dgxFTxuJ50e.

Jakiw Pidstrigach, Youssef Marzouk, Sebastian Reich, and Sven Wang. Infinite-dimensional diffusion models. *Journal of Machine Learning Research*, 25(414):1–52, 2024.

Donlapark Ponnoprat and Masaaki Imaizumi. Minimax rates of estimation for optimal transport map between infinite-dimensional spaces. *arXiv preprint arXiv:2505.13570*, 2025.

Herbert Robbins. A Remark on Stirling's Formula. *The American Mathematical Monthly*, 62(1): 26–29, 1955. doi: 10.2307/2308012.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875, 2020.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 11895–11907, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Taiji Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Rong Tang and Yun Yang. Adaptivity of diffusion models to manifold structures. In *International Conference on Artificial Intelligence and Statistics*, pp. 1648–1656. PMLR, 2024.

Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 11287–11302. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/5dca4c6b9e244d24a30b4c45601d9720-Paper.pdf.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Xixian Wang and Zhongjian Wang. Wasserstein Bounds for generative diffusion models with Gaussian tail targets. *arXiv preprint arXiv:2412.11251*, 2024.

Andre Wibisono, Yihong Wu, and Kaylee Yingxi Yang. Optimal score estimation via empirical Bayes smoothing. In Shipra Agrawal and Aaron Roth (eds.), *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 4958–4991. PMLR, 30 Jun–03 Jul 2024.

Christopher Williams, Fabian Falck, George Deligiannidis, Christopher C. Holmes, Arnaud Doucet, and Saifuddin Syed. A unified framework for u-net design and analysis. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=43ruO2fMjq.

Konstantin Yakovlev and Nikita Puchkin. Generalization error bound for denoising score matching under relaxed manifold assumption. In Nika Haghtalab and Ankur Moitra (eds.), *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 5824–5891. PMLR, 30 Jun–04 Jul 2025. URL https://proceedings.mlr.press/v291/yakovlev25a.html.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks.*, 94: 103–114, 2017. doi: 10.1016/j.neunet.2017.07.002.

CONTENTS

1 Introduction **1**

2 Preliminaries **2**

2.1 Finite Dimensional Diffusion Models 2

2.2 Extension to Infinite-Dimensional Setting 3

2.2.1 Gaussian measure in infinite-dimensional space 3

2.2.2 Radon-Nikodym formulation of the score 4

2.2.3 Reverse process via Radon–Nikodym derivative 4

2.3 Class of neural networks and score model 5

3 γ-smooth Function Class for the True Density Ratio **6**

4 Approximation of the true score **7**

5 Estimation error bound **8**

6 Conclusion **9**

A Mathematical basic lemmas **17**

A.1 the Evolution of Relative Density ρ_t 18

A.2 Contraction Property between $L^2(\mu_{C_t})$ and $L^2(\mu_{C_0})$ 20

B Some featured quantities of γ-smooth space **21**

B.1 Effective dimension and frequency counts 21

B.2 Validity of Assumption 4.2 22

C Construction of Neural Networks **22**

C.1 Composition and Combination Lemmas 22

C.2 Approximations of rational functions 23

C.3 Approximation of m_t and σ_t 23

C.4 Complexity of the clipped score class 23

D Proof of Theorem 4.4 : Approximation of score function $s(x,t)$ **24**

D.1 Proof Sketch 24

D.2 Detailed proof 26

E Approximation of ρ_t **33**

E.1 The error bound of (I) 34

E.2 The error bound of (II) 35

F ReLU approximation of the normalized one-dimensional factor **37**

F.1 Outer-domain approximation 38
F.2 Polynomial approximation on an enlarged inner domain 38
F.2.1 Step 1: Truncation of the Gaussian integral 39
F.2.2 Step 2: Polynomialization of the integrand 39
F.2.3 Network construction for polynomial approximation 42
F.3 Continuous joining and global approximation 46

G Evaluation of $\int_x \left|\phi^i_{2c\nabla\rho_t} - 2c_i\partial_{x_i}\rho_t\right|^2 \rho_t(x)\mathrm{d}\mu_{C_t}(x)$ 48
G.1 Evaluation of the derivative of the truncation remainder 49
G.2 Evaluation of $\int_x \left|\phi^i_{2c\nabla\rho_t} - 2c_i\partial_{x_i}\mathcal{P}_t R_\Gamma(\rho_0)\right|^2 \mathrm{d}P_t(x)$ 50

H Proof of Theorem 5.1 : The estimation error bound 54
H.1 Wasserstein-1 bounds 54
H.2 Wasserstein-2 bounds 68
H.2.1 Case (iii): estimating the residual after removing the stationary linear score 69
H.2.2 Case (iv): a projection-based bound without corrected Lipschitz regularity . 72

NOTATION LIST

Table 1: Notation List

notation	definition
n	sample size
$\tilde{\mathcal{O}}(\cdot)$	the order notation hiding polylog-order, i.e., $\tilde{O}(a_n) = O(a_n \mathrm{polylog}(n))$
$\mathcal{W}_q(P, Q)$	Wasserstein distance of order q between probability measures P, Q
$\|\cdot\|_{L^p(\mu)}$	L^p norm with respect to measure μ (often μ_{C_0} or μ_{C_t})
$\|\cdot\|_A$	weighted Euclidean norm $\|v\|_A^2 := v^\top A v$ (e.g. $A = C^{-1}$)
φ, Φ	standard normal Gaussian PDF and CDF
λ^d	Lebesgue measure on $\mathbb{R}^d$ (used for finite-dimensional heuristics)
μ_Σ	infinite-dimensional gaussian measure with means 0 and covariance Σ
$\mathcal{N}(m, \Sigma)$	Gaussian distribution with mean m and covariance Σ
X_t	forward OU process on H
Y_t	reverse-time process (ideal reverse SDE) driven by the true score
$\hat{Y}_t$	reverse-time process generated using an estimated score $\hat{s}$
β_t	noise schedule, with $0 < \underline{\beta} \le \beta_t \le \overline{\beta}$
$\underline{T}, \overline{T}$	lower and upper bounds of the time interval
$m_t,\ \sigma_t$	$m_t = \exp(-\int_0^t \beta_s \mathrm{d}s),\ \sigma_t = \sqrt{1 - m_t^2}$
$s(x, t)$	(infinite-dimensional) score function defined by $s(x,t) := 2C\nabla \log \rho_t(x) - 2CC_t^{-1}x$
$\hat{s}(x, t)$	learned/estimated score model (neural network-based)
W_t	cylindrical Wiener process
$C = \mathrm{diag}\{c_1, c_2, \ldots\}$	(diagonal) diffusion covariance operator in the OU dynamics
$C_0 = \mathrm{diag}\{c_{0,1}, c_{0,2}, \ldots\}$	reference covariance operator for the initial Gaussian measure μ_{C_0}
$C_t = \mathrm{diag}\{c_{t,1}, c_{t,2}, \ldots\}$	time-dependent covariance operator $C_t := m_t^2 C_0 + \sigma_t^2 C$
r_0	covariance decay exponent: $c_i, c_{0,i} \simeq i^{-2/r_0}$
P_t	law of X_t under the physical measure (with $X_0 \sim P_0$)
ρ_t	Radon–Nikodym derivative $\rho_t := \frac{dP_t}{d\mu_{C_t}}$ (relative density)
$\mathcal{P}_t$	the evolution operator ($\mathcal{P}_t \rho_0 = \rho_t$)
d_{out}	output dimension of estimated score function
$\Phi(L, W, S, B)$	set of fully connected networks with depth L, width W, sparsity level S, norm bound B
$a = (a_i)_{i=1}^\infty$	smoothness with respect to each coordinate
$s = (s_i)_{i=1}^\infty$	frequency with respect to each coordinate
$\psi_l(x) = \prod_{i=1}^\infty \psi_{l_i}(x_i)$	trigonometric orthonormal basis functions
$\delta_s(f)$	basis expansion of f over $l \in \mathbb{Z}_0^\infty$ such that $\lfloor 2^{s_i - 1} \rfloor \le \|l_i\| < 2^{s_i}$
$J(s)$	frequency block index set $J(s) := \{\ell \in \mathbb{Z}_0^\infty :\ \lfloor 2^{s_i-1} \rfloor \le \|\ell_i\| < 2^{s_i}\}$
$\mathcal{F}_{p,q}^\gamma(\mathbb{R}^\infty)$	γ-smooth function class
$\gamma(s)$	penalty on each frequency component
Γ	truncation level for the smoothness expansion
$R_\Gamma(\rho_0)$	the truncation of the basis expansion of ρ_0 obtained by retaining only the basis elements satisfying $\gamma(s) < \Gamma$
$I(\Gamma, \gamma)$	set of features contributing a frequency component s with $\gamma(s) < \Gamma$
$d_{\max}(\Gamma, \gamma)$	$\|I(\Gamma, \gamma)\|$ (axial complexity)
$f_{\max}(\Gamma, \gamma)$	$\max_{s \in \mathbb{N}_0^\infty :\, \gamma(s) < \Gamma} \max_{i \in \mathbb{N}} s_i$ (frequency direction complexity)
v	$\left(\frac{1}{p} - \frac{1}{2}\right)_+$
$\alpha(\gamma)$	$\sup_{s \in \mathbb{N}_0^\infty} \frac{\sum_{i=1}^\infty s_i}{\gamma(s)}$
$G(\Gamma, \gamma)$	$\sum_{s \in \mathbb{N}_0^\infty :\, \gamma(s) < \Gamma} 2^s \quad (\text{where } 2^s := 2^{\sum_{i=1}^\infty s_i})$
η	the rate of a_i ($a_i \simeq i^\eta$)
$a^\dagger$	effective smoothness: mixed: $a^\dagger = a_1$, anisotropic: $a^\dagger = \left(\sum_i a_i^{-1}\right)^{-1}$ (so $a^\dagger = 1/\alpha(\gamma)$)
$\mathbb{E}_n[\cdot]$	expectation with respect to training data

A MATHEMATICAL BASIC LEMMAS

Here, we introduce some basic mathematical lemmas that are necessary for the proof of approximation theory.

Lemma A.1 (A Stirling-type bound for $1/n!$ (Robbins, 1955)). *For every integer $n \geq 1$,*

$$\frac{1}{n!} < \frac{1}{\sqrt{2\pi n}} \left(\frac{e}{n}\right)^n.$$

Lemma A.2 (Mills ratio inequality for Gaussian tails (Gordon, 1941; Dümbgen, 2010)). *Let*

$$\varphi(x) := \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad \Phi(x) := \int_{-\infty}^{x} \varphi(t)\, dt, \qquad \bar{\Phi}(x) := 1 - \Phi(x) = \int_{x}^{\infty} \varphi(t)\, dt.$$

Then for every $x > 0$,

$$\frac{x}{1+x^2}\varphi(x) < \bar{\Phi}(x) < \frac{1}{x}\varphi(x).$$

In particular, for $Z \sim \mathcal{N}(0,1)$,

$$\mathbb{P}(Z \geq x) = \bar{\Phi}(x) \leq \frac{1}{\sqrt{2\pi}} \frac{e^{-\frac{x^2}{2}}}{x}, \qquad \mathbb{P}(|Z| \geq x) = 2\bar{\Phi}(x) \leq \frac{2}{\sqrt{2\pi}} \frac{e^{-\frac{x^2}{2}}}{x}.$$

Remark A.3 (Upper bound of $\int_x^\infty t^2\varphi(t)\mathrm{d}t$). *By using Lemma A.2, we can get the upper bound of $\int_x^\infty t^2\varphi(t)\mathrm{d}t$:*

$$\int_x^\infty t^2\varphi(t)\mathrm{d}t \leq \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \left(x + \frac{1}{x}\right).$$

Proposition A.4 (Girsanov change of drift and entropy bound (Karatzas et al., 1991)). *Consider two finite-dimensional SDEs on $[0,T]$,*

$$\mathrm{d}Z_t = b(Z_t,t)\,\mathrm{d}t + \Sigma_t\,\mathrm{d}B_t, \qquad Z_0 \sim \mu,$$
$$\mathrm{d}Z'_t = b'(Z'_t,t)\,\mathrm{d}t + \Sigma_t\,\mathrm{d}B'_t, \qquad Z'_0 \sim \nu,$$

where Σ_t is deterministic, square-integrable in time, and invertible for almost every t. Let $\mathbb{P}$ and $\mathbb{Q}$ denote the corresponding path laws, and let p_t and q_t denote their time-t marginals.

Assume that the SDEs admit nonexplosive solutions and that the local uniqueness conditions needed for the stopped changes of drift hold. For example, it is sufficient to have spatially locally Lipschitz drifts with appropriate time-integrable Lipschitz bounds and a uniform linear-growth bound, together with bounded Σ_t.

Define

$$\theta_t(x) := \Sigma_t^{-1}\big(b(x,t) - b'(x,t)\big).$$

Suppose that

$$\mathrm{KL}(\mu\|\nu) < \infty, \qquad \mathbb{E}_{\mathbb{P}} \int_0^T \|\theta_t(Z_t)\|_2^2\,\mathrm{d}t < \infty.$$

Then

$$\begin{aligned}\mathrm{KL}(p_T\|q_T) &\leq \mathrm{KL}(\mathbb{P}\|\mathbb{Q}) \\ &\leq \mathrm{KL}(\mu\|\nu) + \frac{1}{2}\mathbb{E}_{\mathbb{P}} \int_0^T \|\theta_t(Z_t)\|_2^2\,\mathrm{d}t.\end{aligned}$$

This inequality can be established by localization without imposing a global Novikov condition.

For comparison, when $\mu = \nu$ and

$$\mathbb{E}_{\mathbb{P}} \exp\left(\frac{1}{2}\int_0^T \|\theta_t(Z_t)\|_2^2\,\mathrm{d}t\right) < \infty,$$

the usual Novikov argument gives

$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}(Z) = \exp\left(-\int_0^T \langle\theta_t(Z_t), \mathrm{d}B_t\rangle - \frac{1}{2}\int_0^T \|\theta_t(Z_t)\|_2^2\,\mathrm{d}t\right),$$

where B is the Brownian motion under $\mathbb{P}$. Under the finite-energy assumption above, this yields the corresponding path-space entropy equality.

Proof. We prove the localized entropy inequality. Let $\mathbb{Q}^\mu$ be the path law of the SDE with drift b' and initial distribution μ. The initial change of measure satisfies

$$\frac{\mathrm{d}\mathbb{Q}^\mu}{\mathrm{d}\mathbb{Q}} = \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(Z_0).$$

On canonical path space, define the stopping time

$$\tau_k := \inf\left\{t \in [0,T] : \int_0^t \|\theta_s(Z_s)\|_2^2 \,\mathrm{d}s \geq k\right\} \wedge T.$$

Under $\mathbb{Q}^\mu$, the stopped stochastic exponential

$$\exp\left(\int_0^{\tau_k} \langle \theta_t(Z_t), \mathrm{d}B_t^{\mathbb{Q}^\mu}\rangle - \frac{1}{2}\int_0^{\tau_k} \|\theta_t(Z_t)\|_2^2 \,\mathrm{d}t\right)$$

is a true martingale, because its quadratic variation is bounded by k. Let $\mathbb{P}^{(k)}$ be the resulting changed measure. Its drift is b before τ_k and b' afterwards, and its initial distribution remains μ. By uniqueness of the stopped dynamics, $\mathbb{P}^{(k)}$ and $\mathbb{P}$ have the same stopped-path law up to τ_k. The stopped Girsanov formula therefore gives

$$\begin{aligned}\mathrm{KL}(\mathbb{P}^{(k)}\|\mathbb{Q}) &= \mathrm{KL}(\mu\|\nu) + \frac{1}{2}\mathbb{E}_{\mathbb{P}^{(k)}}\int_0^{\tau_k} \|\theta_t(Z_t)\|_2^2 \,\mathrm{d}t \\ &= \mathrm{KL}(\mu\|\nu) + \frac{1}{2}\mathbb{E}_{\mathbb{P}}\int_0^{\tau_k} \|\theta_t(Z_t)\|_2^2 \,\mathrm{d}t.\end{aligned}$$

The finite-energy assumption implies

$$\mathbb{P}(\tau_k < T) \leq \frac{1}{k}\mathbb{E}_{\mathbb{P}}\int_0^T \|\theta_t(Z_t)\|_2^2 \,\mathrm{d}t \longrightarrow 0.$$

Since the stopped laws agree and the dynamics are unchanged on paths that do not stop before T, $\mathbb{P}^{(k)}$ converges to $\mathbb{P}$ in total variation. Lower semicontinuity of relative entropy gives

$$\begin{aligned}\mathrm{KL}(\mathbb{P}\|\mathbb{Q}) &\leq \liminf_{k\to\infty} \mathrm{KL}(\mathbb{P}^{(k)}\|\mathbb{Q}) \\ &\leq \mathrm{KL}(\mu\|\nu) + \frac{1}{2}\mathbb{E}_{\mathbb{P}}\int_0^T \|\theta_t(Z_t)\|_2^2 \,\mathrm{d}t.\end{aligned}$$

Finally, applying the terminal evaluation map and data processing for relative entropy yields the terminal-marginal inequality. The Novikov statement is the standard unstopped Girsanov formula. □

Lemma A.5 (Pinsker's inequality (see, e.g., (3.59) of Wainwright (2019))). *Let P and Q be probability measures on a measurable space such that $P \ll Q$. Then,*

$$\|P - Q\|_{\mathrm{TV}} \leq \sqrt{\frac{1}{2}\mathrm{KL}(P\|Q)},$$

where $\|P - Q\|_{\mathrm{TV}} := \sup_{A\in\mathcal{F}} |P(A) - Q(A)|$ is the total variation distance, and

$$\mathrm{KL}(P\|Q) := \int \log\left(\frac{dP}{dQ}\right) dP$$

denotes the Kullback–Leibler divergence.

A.1 The Evolution of Relative Density ρ_t

Let $\mathbb{P}$ denote the physical probability measure where the initial distribution is P_0, and let $\mathbb{Q}$ denote the reference probability measure where the initial distribution is the Gaussian measure μ_{C_0}. The relative density at time $t = 0$ is given by the Radon-Nikodym derivative:

$$\rho_0(x) = \frac{\mathrm{d}P_0}{\mathrm{d}\mu_{C_0}}(x).$$

Our goal is to find the expression for the relative density at time t, denoted by $\rho_t(x) = \frac{\mathrm{d}P_t}{\mathrm{d}\mu_{C_t}}(x)$, where P_t and μ_{C_t} are the push-forward measures of P_0 and μ_{C_0} under the Ornstein-Uhlenbeck dynamics, respectively.

Lemma A.6 (The Evolution of Relative Density ρ_t)**.** *$\rho_t(x)$ can be written with $\widetilde{m}_t = m_t C_0 C_t^{-1}$ and $\widetilde{\sigma}_t = \sigma_t \sqrt{C C_t^{-1}}$ as follows:*

$$\rho_t(x) = \int_{\mathbb{R}^\infty} \rho_0(\widetilde{m}_t x + \widetilde{\sigma}_t y)\, \mathrm{d}\mu_{C_0}(y).$$

Proof. By the property of Radon-Nikodym derivatives under filtration (change of measure formula), the density ρ_t satisfies:

$$\rho_t(x) = \mathbb{E}^{\mathbb{Q}}\left[\frac{\mathrm{d}P_0}{\mathrm{d}\mu_{C_0}}(X_0) \,\middle|\, X_t = x\right] = \mathbb{E}^{\mathbb{Q}}\left[\rho_0(X_0) \mid X_t = x\right]. \tag{2}$$

Here, the expectation $\mathbb{E}^{\mathbb{Q}}$ is taken with respect to the *reference process* where the process starts from μ_{C_0}. Under $\mathbb{Q}$, the joint distribution of (X_0, X_t) is Gaussian with mean zero and covariance:

$$\mathrm{Cov}^{\mathbb{Q}}\begin{pmatrix} X_0 \\ X_t \end{pmatrix} = \begin{pmatrix} C_0 & m_t C_0 \\ m_t C_0 & C_t \end{pmatrix}.$$

To evaluate (2), we consider the conditional distribution of X_0 given $X_t = x$ *under the reference measure* $\mathbb{Q}$. Using the Gaussian conditional formula, we can express the random variable X_0 conditioned on $X_t = x$ as:

$$(X_0 \mid X_t = x) \overset{d}{=} \widetilde{m}_t x + \widetilde{\sigma}_t y, \quad \text{with } y \sim \mu_{C_0}.$$

Here, $\widetilde{m}_t = m_t C_0 C_t^{-1}$ represents the backward mean shift, and $\widetilde{\sigma}_t = (I - m_t^2 C_0 C_t^{-1})^{1/2} = \sigma_t \sqrt{C C_t^{-1}}$ represents the reduction in variance due to the observation of X_t.
Substituting this back into (2), we obtain the integral representation:

$$\rho_t(x) = \int_{\mathbb{R}^\infty} \rho_0(\widetilde{m}_t x + \widetilde{\sigma}_t y)\, \mathrm{d}\mu_{C_0}(y).$$

This formula describes how the initial density profile ρ_0 is transported and smoothed by the dynamics to form ρ_t. □

Here, $\widetilde{m}_t$ and $\widetilde{\sigma}_t$ are diagonal:

$$\widetilde{m}_t = \mathrm{diag}\left\{m_t \frac{c_{0,1}}{c_{t,1}}, m_t \frac{c_{0,2}}{c_{t,2}}, \dots\right\},\ \widetilde{\sigma}_t = \mathrm{diag}\left\{\sigma_t \sqrt{\frac{c_1}{c_{t,1}}}, \sigma_t \sqrt{\frac{c_2}{c_{t,2}}}, \dots\right\}.$$

Using the representation of ρ_t given in the Lemma A.6, we derive the following reformulation of $\frac{\partial}{\partial x_i}\rho_t$, which will be used repeatedly later.

Lemma A.7 (the partial derivative of ρ_t)**.**

$$\frac{\partial}{\partial x_i}\rho_t = \sum_{l \in \mathbb{Z}_0^\infty} \langle \rho_0, \psi_l \rangle \frac{\widetilde{m}_{t,i}}{\widetilde{\sigma}_{t,i} c_{0,i}} \int_{\mathbb{R}^\infty} y_i \psi_l(\widetilde{m}_t x + \widetilde{\sigma}_t y) \mathrm{d}\mu_{C_0}(y)$$

Proof.

$$\begin{aligned}\frac{\partial}{\partial x_i}\rho_t &= \sum_{l \in \mathbb{Z}_0^\infty} \langle \rho_0, \psi_l \rangle \int_{\mathbb{R}^\infty} \partial_{x_i} \psi_l(\widetilde{m}_t x + \widetilde{\sigma}_t y) \mathrm{d}\mu_{C_0}(y) \\ &= \sum_{l \in \mathbb{Z}_0^\infty} \langle \rho_0, \psi_l \rangle \widetilde{m}_{t,i} \int_{\mathbb{R}^\infty} (\partial_{x_i} \psi_l)(\widetilde{m}_t x + \widetilde{\sigma}_t y) \mathrm{d}\mu_{C_0}(y)\end{aligned}$$

$$= \sum_{l\in\mathbb{Z}_0^\infty} \langle \rho_0, \psi_l \rangle \widetilde{m}_{t,i} \int_{\mathbb{R}} (\partial_{x_i}\psi_{l_i})(\widetilde{m}_{t,i}x_i + \widetilde{\sigma}_{t,i}y_i) \mathrm{d}\mu_{C_0}(y_i) \prod_{j\neq i} \int_{\mathbb{R}} \psi_{l_j}(\widetilde{m}_{t,j}x_j + \widetilde{\sigma}_{t,j}y_j) \mathrm{d}\mu_{C_0}(y_j)$$

$$= \sum_{l\in\mathbb{Z}_0^\infty} \langle \rho_0, \psi_l \rangle \widetilde{m}_{t,i} \underbrace{\int_{\mathbb{R}} \frac{1}{\widetilde{\sigma}_{t,i}c_{0,i}} y_i \psi_{l_i}(\widetilde{m}_{t,i}x_i + \widetilde{\sigma}_{t,i}y_i) \mathrm{d}\mu_{C_0}(y_i)}_{\text{Integration by parts}} \prod_{j\neq i} \int_{\mathbb{R}} \psi_{l_j}(\widetilde{m}_{t,j}x_j + \widetilde{\sigma}_{t,j}y_j) \mathrm{d}\mu_{C_0}(y_j)$$

$$= \sum_{l\in\mathbb{Z}_0^\infty} \langle \rho_0, \psi_l \rangle \frac{\widetilde{m}_{t,i}}{\widetilde{\sigma}_{t,i}c_{0,i}} \int_{\mathbb{R}^\infty} y_i \psi_l(\widetilde{m}_t x + \widetilde{\sigma}_t y) \mathrm{d}\mu_{C_0}(y)$$

□

A.2 Contraction Property between $L^2(\mu_{C_t})$ and $L^2(\mu_{C_0})$

We demonstrate that the evolution operator $\mathcal{P}_t$ contracts the norm when mapping from the initial space $L^2(\mu_{C_0})$ to the current space $L^2(\mu_{C_t})$.

Let $f \in L^2(\mu_{C_0})$. The operator is defined by the conditional expectation under the reference measure $\mathbb{Q}$ (where $X_0 \sim \mu_{C_0}$ and $X_t \sim \mu_{C_t}$):

$$(\mathcal{P}_t f)(x) = \mathbb{E}^{\mathbb{Q}}\left[f(X_0) \mid X_t = x\right].$$

Note that the argument x on the left-hand side resides in the space at time t.

Lemma A.8 (L^2-Contractivity of the Evolution Operator)**.** *Let μ_{C_0} and μ_{C_t} denote the Gaussian measures associated with the process at time 0 and time t, respectively. For any function $f \in L^2(\mu_{C_0})$, the operator $\mathcal{P}_t$ maps $L^2(\mu_{C_0})$ to $L^2(\mu_{C_t})$ and satisfies the following contraction inequality:*

$$\|\mathcal{P}_t f\|_{L^2(\mu_{C_t})} \leq \|f\|_{L^2(\mu_{C_0})}.$$

Proof. We begin with the square of the norm in $L^2(\mu_{C_t})$. By definition:

$$\|\mathcal{P}_t f\|^2_{L^2(\mu_{C_t})} = \int_H |(\mathcal{P}_t f)(x)|^2 \, d\mu_{C_t}(x).$$

Substituting the definition of $\mathcal{P}_t$, we have:

$$\int_H \left|\mathbb{E}^{\mathbb{Q}}[f(X_0) \mid X_t = x]\right|^2 d\mu_{C_t}(x).$$

Applying Jensen's Inequality for conditional expectations ($|\mathbb{E}[Z|Y]|^2 \leq \mathbb{E}[|Z|^2|Y]$) pointwise to the integrand yields:

$$\left|\mathbb{E}^{\mathbb{Q}}[f(X_0) \mid X_t = x]\right|^2 \leq \mathbb{E}^{\mathbb{Q}}\left[|f(X_0)|^2 \mid X_t = x\right].$$

Integrating both sides with respect to the marginal measure $\mu_{C_t}(x)$, we obtain:

$$\|\mathcal{P}_t f\|^2_{L^2(\mu_{C_t})} \leq \int_H \mathbb{E}^{\mathbb{Q}}\left[|f(X_0)|^2 \mid X_t = x\right] d\mu_{C_t}(x).$$

By the *Law of Total Expectation* (or the definition of the marginal distribution), integrating the conditional expectation with respect to the marginal distribution of the conditioning variable (X_t) recovers the full expectation over the joint distribution, which simplifies to the expectation with respect to X_0:

$$\int_H \mathbb{E}^{\mathbb{Q}}\left[|f(X_0)|^2 \mid X_t = x\right] d\mu_{C_t}(x) = \mathbb{E}^{\mathbb{Q}}\left[|f(X_0)|^2\right].$$

Under the reference measure $\mathbb{Q}$, the initial variable X_0 follows the distribution μ_{C_0}. Therefore:

$$\mathbb{E}^{\mathbb{Q}}\left[|f(X_0)|^2\right] = \int_H |f(x)|^2 d\mu_{C_0}(x) = \|f\|^2_{L^2(\mu_{C_0})}.$$

Combining these results, we conclude:

$$\|\mathcal{P}_t f\|^2_{L^2(\mu_{C_t})} \leq \|f\|^2_{L^2(\mu_{C_0})}.$$

□

B SOME FEATURED QUANTITIES OF γ-SMOOTH SPACE

B.1 EFFECTIVE DIMENSION AND FREQUENCY COUNTS

Here, we introduce some featured quantities of γ-smooth space defined in Okumoto & Suzuki (2022).

For a given $\Gamma > 0$ and the smoothness $\gamma : \mathbb{N}_0^\infty \to \mathbb{R}_+$, define

$$I(\Gamma, \gamma) := \{i \in \mathbb{N} : \exists s \in \mathbb{N}_0^\infty, s_i \neq 0, \gamma(s) < \Gamma\},$$

and then the following quantities play an important role in our approximation error analysis.

Definition B.1 (Axial complexity and frequency direction complexity (Okumoto & Suzuki, 2022))**.** *The axial complexity is defined by*

$$d_{\max}(\Gamma, \gamma) := |I(\Gamma, \gamma)|.$$

Moreover, the frequency direction complexity is defined by

$$f_{\max}(\Gamma, \gamma) := \max_{s \in \mathbb{N}_0^\infty : \gamma(s) < \Gamma} \max_{i \in \mathbb{N}} s_i.$$

The axial complexity is used to evaluate how many components need to be extracted from a given infinite-dimensional sequence $x \in \mathbb{R}^\infty$ to achieve a particular approximation error, and the frequency complexity characterizes up to which frequency we require to approximate a target function with a particular error. Let $(x)_+ := \max\{x, 0\}$ and define

$$\alpha(\gamma) := \sup_{s \in \mathbb{N}_0^\infty} \frac{\sum_{i=1}^\infty s_i}{\gamma(s)}, \quad G(\Gamma, \gamma) := \sum_{s \in \mathbb{N}_0^\infty : \gamma(s) < \Gamma} 2^s.$$

These symbols are used in Appendix D to determine the neural network class.

For mixed smoothness $\gamma(s) = \sum_i a_i s_i$ and anisotropic smoothness $\gamma(s) = \max_i a_i s_i$, monotonicity of (a_i) gives

$$I(\Gamma, \gamma) = \{i : a_i < \Gamma\} = \{1, \dots, d_{\max}\},$$

with the empty-set convention when no coordinate satisfies $a_i < \Gamma$. The growth assumption $a_i = \Omega(i^\eta)$ implies

$$d_{\max} \lesssim \Gamma^{1/\eta}.$$

Under the strict truncation condition $\gamma(s) < \Gamma$, the frequency-level complexity is

$$f_{\max}(\Gamma, \gamma) = \left\lceil \frac{\Gamma}{a_1} \right\rceil - 1.$$

Indeed, every retained block satisfies $s_i < \Gamma / a_i \leq \Gamma / a_1$, and the largest admissible level is attained by a block supported on the first coordinate. Consequently,

$$|l_i| < 2^{f_{\max}}, \qquad 2^{f_{\max}} \asymp 2^{\Gamma / a_1}.$$

For mixed smoothness, $a^\dagger = a_1$ and

$$G(\Gamma, \gamma) \lesssim 2^{\Gamma / a_1}.$$

The constant depends on the fixed smoothness sequence; the associated infinite product is finite under $a_1 < a_2$ and $a_i = \Omega(i^\eta)$.

For anisotropic smoothness, direct counting gives

$$G(\Gamma, \gamma) = \prod_{i : a_i < \Gamma} \left(2^{\lceil \Gamma / a_i \rceil} - 1 \right).$$

For $x > 1$, put $k = \lceil x \rceil - 1 \geq 1$. Since $2^{k+1} - 1 \leq 3^k \leq 3^x$, we obtain

$$G(\Gamma, \gamma) \leq 3^{\Gamma \sum_{i : a_i < \Gamma} a_i^{-1}} \leq 2^{(\log_2 3) \Gamma / a^\dagger}.$$

See also Lemma 18 in Okumoto & Suzuki (2022) and the proof of Theorem 1 in Nishimura & Suzuki (2024). Moreover, $\alpha(\gamma) = 1 / a^\dagger$, where $a^\dagger$ is defined in Section 4.

B.2 Validity of Assumption 4.2

Here, we check the validity of assumption about the scale of $|\partial_{x_i}\rho_0|$.

Remark B.2 (Validity of Assumption 4.2)**.** *Let* $\rho_0 = \sum_{l\in\mathbb{Z}_0^\infty}\langle\rho_0,\psi_l\rangle\psi_l(x)$*. Then,*

$$\frac{\partial}{\partial x_j}\psi_{l_i}(x_i) = -\sqrt{2\pi}\, c_{0,i}^{-1/2}\exp\left(-\frac{x_i^2}{2c_{0,i}}\right) l_i\cdot\sqrt{2}\sin\left(2\pi l_i\Phi\left(\frac{x_i}{\sqrt{c_{0,i}}}\right)+\frac{\pi}{4}\right)\delta_{ij},$$

where δ_{ij} *denotes the Kronecker delta, i.e.,* $\delta_{ij} = \begin{cases}0 & (i\neq j),\\ 1 & (i=j).\end{cases}$

Therefore, we obtain

$$\frac{\partial}{\partial x_i}\rho_0(x) = -\sqrt{2\pi}\, c_{0,i}^{-1/2}\exp\left(-\frac{x_i^2}{2c_{0,i}}\right)\sum_{l\in\mathbb{Z}_0^\infty}\langle\rho_0,\psi_l\rangle l_i\tan\left(2\pi l_i\Phi\left(\frac{x_i}{\sqrt{c_{0,i}}}\right)+\frac{\pi}{4}\right)\psi_l(x).$$

From the above expression, we observe that the factor $c_{0,i}^{-1/2}$ appears due to differentiation. Moreover, since $\langle\rho_0,\psi_l\rangle$ decays as l_i increases, this supports the validity of the Assumption 4.2.

C Construction of Neural Networks

This section introduces fundamental tools on function approximation using neural networks. These lemmas play a central role in our approximation error analysis. Many of the lemmas below are based on Oko et al. (2023).

C.1 Composition and Combination Lemmas

We begin with lemmas that describe how to combine multiple neural networks. These lemmas are essential to realize a large neural network that approximates complicated functions.

Lemma C.1 (Concatenation of neural networks (Remark 13 of Nakada & Imaizumi (2020)))**.** *For any neural networks* $\phi^1,\dots,\phi^k$ *with* $\phi^i:\mathbb{R}^{d_i}\to\mathbb{R}^{d_{i+1}}$ *and* $\phi^i\in\Phi(L^i,W^i,S^i,B^i)$*, there exists a network* $\phi\in\Phi(L,W,S,B)$ *such that* $\phi(x)=\phi^k\circ\phi^{k-1}\circ\cdots\circ\phi^1(x)$ *for* $x\in\mathbb{R}^{d_1}$ *with*

$$L=\sum_{i=1}^k L^i,\qquad W\le 2\sum_{i=1}^k W^i,\qquad S\le 2\sum_{i=1}^k S^i,\qquad B\le\max_{1\le i\le k}B^i.$$

Lemma C.2 (Parallelization of neural networks (Oko et al., 2023))**.** *For any networks* $\phi^1,\dots,\phi^k$ *with* $\phi^i:\mathbb{R}^{d_i}\to\mathbb{R}^{d_i'}$ *and* $\phi^i\in\Phi(L^i,W^i,S^i,B^i)$*, there exists a network* $\phi\in\Phi(L,W,S,B)$ *such that*

$$\phi(x)=\left[(\phi^1(x_1))^\top,\dots,(\phi^k(x_k))^\top\right]^\top:\mathbb{R}^{d_1+\cdots+d_k}\to\mathbb{R}^{d_1'+\cdots+d_k'}$$

for $x=(x_1^\top\ \cdots\ x_k^\top)^\top$ *with*

$$L=\max_{1\le i\le k}L^i,\qquad W\le 2\sum_{i=1}^k W^i,\qquad S\le 2\sum_{i=1}^k\left(S^i+LW_L^i\right),\qquad B\le\max\left\{\max_{1\le i\le k}B^i,\,1\right\}.$$

Lemma C.3 (Clipping function Oko et al. (2023))**.** *For any* $a,b\in\mathbb{R}^d$ *with* $a_i\le b_i$ $(i=1,2,\dots,d)$*, there exists a clipping function* $\phi_{\mathrm{clip}}(x;a,b)\in\Phi(2,2d,7d,\max\{1,\max_{1\le i\le d}\max\{|a_i|,|b_i|\}\};d,d)$ *such that*

$$\phi_{\mathrm{clip}}(x;a,b)_i=\min\{b_i,\max\{x_i,a_i\}\}\ (i=1,2,\dots,d)$$

holds. When $a_i=c$ *and* $b_i=C$ *for all* i*, we sometimes denote* $\phi_{\mathrm{clip}}(x;a,b)$ *as* $\phi_{\mathrm{clip}}(x;c,C)$ *using scalar values* c *and* C*.*

Lemma C.4 (Summation function (Oko et al., 2023))**.** *For any networks* $\phi^1,\dots,\phi^k$ *with* $\phi^i:\mathbb{R}^{d_i}\to\mathbb{R}^d$ *and* $\phi^i\in\Phi(L^i,W^i,S^i,B^i)$*, there exists a network* $\phi\in\Phi(L,W,S,B)$ *such that*

$$\phi(x)=\sum_{i=1}^k\phi^i(x_i):\mathbb{R}^{d_1+\cdots+d_k}\to\mathbb{R}^d$$

for $x = (x_1^\top \ \cdots \ x_k^\top)^\top$ *with*

$$L = \max_{1\le i\le k} L^i + 1, \qquad W \le 4\sum_{i=1}^{k} W^i, \qquad S \le 4\sum_{i=1}^{k}\big(S^i + LW_L^i\big) + 2W_L, \qquad B \le \max\Big\{\max_{1\le i\le k} B^i,\, 1\Big\}.$$

C.2 Approximations of rational functions

The following lemmas describe neural network approximations of multiplication and reciprocal operations. We use these operations to combine approximations of the density ratio and its derivatives in the nonlinear score term $2C\nabla\rho_t/\rho_t$.

Lemma C.5 (Approximation of monomials (Oko et al., 2023)). *For $d \ge 2$, $C \ge 1$, $0 < \varepsilon_{\text{error}} \le 1$ and $\varepsilon > 0$, there exists a network $\phi_{\text{mult}}(x_1, \dots, x_d) \in \Phi(L, W, S, B)$ with*

$$L = \mathcal{O}\big(\log d(\log \varepsilon^{-1} + d\log C)\big), \qquad W = 48d, \qquad S = \mathcal{O}\big(d\log\varepsilon^{-1} + d\log C\big), \qquad B = C^d$$

such that $|\phi_{\text{mult}}(x)| \le C^d$, $\phi_{\text{mult}}(x_1, \dots, x_d) = 0$ if at least one of x_i is equal to zero, and

$$\Big|\phi_{\text{mult}}(x_1', x_2', \dots, x_d') - \prod_{i=1}^{d} x_i\Big| \le \varepsilon + dC^{d-1}\varepsilon_{\text{error}}, \qquad \forall x \in [-C, C]^d,\ \|x - x'\|_\infty \le \varepsilon_{\text{error}}.$$

Lemma C.6 (Approximating the reciprocal function (Oko et al., 2023)). *For a constant $R > 1$ and any $0 < \varepsilon < R^{-1}$, there exists a network $\phi_{\text{rec}}(x) \in \Phi(L, W, S, B)$ with*

$$L \le \mathcal{O}\big(\log^2\varepsilon^{-1}\big), \qquad W = \mathcal{O}\big(\log^3\varepsilon^{-1}\big), \qquad S = \mathcal{O}\big(\log^4\varepsilon^{-1}\big), \qquad B = \mathcal{O}\big(\varepsilon^{-2}\big)$$

such that

$$\Big|\phi_{\text{rec}}(x') - \frac{1}{x}\Big| \le \varepsilon + R^2|x' - x|, \qquad \forall x \in [R^{-1}, R],\ x' \in \mathbb{R}.$$

Lemma C.7 (Approximating the root function Oko et al. (2023)). *For any $0 < \varepsilon < 1$, there exists a network $\phi_{\text{root}}(x) \in \Phi(L, W, S, B)$ with*

$$L \le \mathcal{O}\big(\log^2\varepsilon^{-1}\big), \qquad W = \mathcal{O}\big(\log^3\varepsilon^{-1}\big), \qquad S = \mathcal{O}\big(\log^4\varepsilon^{-1}\big), \qquad B = \mathcal{O}\big(\varepsilon^{-1}\big)$$

such that

$$\Big|\phi_{\text{root}}(x') - \sqrt{x}\Big| \le \varepsilon + \varepsilon^{-\frac{1}{2}}|x' - x|, \qquad \forall x \in [\varepsilon, \varepsilon^{-1}],\ x' \in \mathbb{R}.$$

C.3 Approximation of m_t and σ_t

We begin with construction of sub-networks that approximate m_t and σ_t.

Under the noise-schedule assumptions in Section 2, Lemma B.1 of Oko et al. (2023) gives the following approximation bounds.

Lemma C.8 (Approximation of m_t and σ_t(Oko et al., 2023)). *Let $0 < \varepsilon < \frac{1}{2}$. Then, there exists a neural network $\phi_m(t) \in \Phi(L, W, S, B)$ that approximates m_t for all $t \ge 0$, within the additive error of ε, where*

$$L = \mathcal{O}(\log^2\varepsilon^{-1}), \quad W = \mathcal{O}(\log^2\varepsilon^{-1}), \quad S = \mathcal{O}(\log^3\varepsilon^{-1}), \quad B = \mathcal{O}(\log\varepsilon^{-1}).$$

Also, there exists a neural network $\phi_\sigma(t) \in \Phi(L, W, S, B)$ that approximates σ_t for all $t \ge \varepsilon$, within the additive error of ε, where

$$L \le \mathcal{O}(\log^2\varepsilon^{-1}), \quad W = \mathcal{O}(\log^2\varepsilon^{-1}), \quad S = \mathcal{O}(\log^3\varepsilon^{-1}), \quad B = \mathcal{O}(\log\varepsilon^{-1}).$$

C.4 Complexity of the clipped score class

For a function class $\mathcal{F}$ equipped with a metric d, let $\mathcal{N}(\mathcal{F}, \varepsilon, d)$ denote the minimum number of radius-ε balls, with centers in $\mathcal{F}$, required to cover $\mathcal{F}$. For vector-valued functions on a bounded domain D, define

$$d_{\infty,D}(f, g) := \sup_{z\in D}\max_i |f_i(z) - g_i(z)|.$$

Proposition C.9 (Covering numbers on bounded domains). *Let* $\Phi = \Phi(L, W, S, B; D_{\text{in}}, D_{\text{out}})$, *where* L *counts affine layers and* $L, S \geq 1$, *and set*

$$D_* := \max\{W, D_{\text{in}}, D_{\text{out}}, 1\}.$$

For $D \subset [-R, R]^{D_{\text{in}}}$, $R \geq 1$, *and* $0 < \varepsilon \leq 1$,

$$\log \mathcal{N}(\Phi, \varepsilon, d_{\infty,D}) \leq CSL \log\left(\frac{2(L+1)(D_*+1)(B \vee 1)(R+1)}{\varepsilon}\right),$$

where C *is a universal constant.*

Proof. Pad hidden layers to width D_*. There are at most $P_* \leq L(D_*+1)^2$ parameter positions, and at most $(P_*+1)^S$ supports with at most S nonzero parameters.
For a fixed support, discretize the parameters with mesh size η. The 1-Lipschitz property of ReLU and a layerwise induction bound the output perturbation by

$$CL(R+1)(D_*+1)^L(B \vee 1)^{L-1}\eta$$

in $d_{\infty,D}$. Choose the mesh so that this quantity is at most a fixed fraction of ε, and count the grid points over all supports. Taking logarithms gives the assertion. Centers can be chosen within the class by selecting one class representative from each occupied grid cell and adjusting the mesh by a fixed factor. □

For a diagonal positive definite matrix A, define

$$d_{A,D}(f, g) := \sup_{z \in D} \|f(z) - g(z)\|_A, \qquad \kappa_A := (\operatorname{Tr} A)^{1/2}.$$

Then

$$d_{A,D}(f, g) \leq \kappa_A d_{\infty,D}(f, g).$$

In particular,

$$\kappa_I = \sqrt{d_{\text{out}}}, \qquad \kappa_{C^{-1}} = \left(\sum_{i=1}^{d_{\text{out}}} c_i^{-1}\right)^{1/2}.$$

Coordinatewise clipping with fixed thresholds and zero-padding are nonexpansive in the coordinatewise uniform metric. Thus, separate covers of the unclipped gradient and linear branches induce a cover of the clipped class $\mathcal{S}$. Write

$$D_j := \max\{W_j, D_{\text{in},j}, D_{\text{out},j}, 1\}, \qquad j \in \{\text{grad}, \text{linear}\}.$$

For an input domain D contained in a box of radius $R \geq 1$ and $0 < \delta \leq 1$, we obtain

$$\log \mathcal{N}(\mathcal{S}, \delta, d_{A,D}) \leq C \sum_{j \in \{\text{grad},\text{linear}\}} S_j L_j \log\left(\frac{2(L_j+1)(D_j+1)(B_j \vee 1)(R+1)(1+\kappa_A)}{\delta}\right).$$

Only trainable parameters are counted. The clipping thresholds are fixed and do not introduce additional free parameters. This estimate is used on bounded input domains; the original ERM is not restricted to these domains.

D Proof of Theorem 4.4 : Approximation of score function $s(x, t)$

In this section, we prove Theorem 4.4. We begin by outlining the proof.

D.1 Proof Sketch

We approximate the first d_{out} coordinates of the score $s(x, t)$ by a neural network and set the remaining coordinates to zero. The truncation error incurred by ignoring the coordinates $d_{\text{out}} + 1, d_{\text{out}} + 2, \ldots$ is bounded by $\frac{r_0}{2-r_0} d_{\text{out}}^{1-\frac{2}{r_0}}$. For each $i \leq d_{\text{out}}$, we decompose

$$s^i(x, t) = 2c_i \partial_{x_i} \log \rho_t(x) - 2c_i c_{t,i}^{-1} x_i$$

into a gradient term and a term that is linear in x_i. Accordingly, we construct two subnetworks: $\phi_{\text{score,grad}}(x,t)$ to approximate the gradient component, and $\phi_{\text{score,linear}}(x_i,t)$ to approximate the linear component.

The gradient term admits the fractional representation $2c_i\partial_{x_i}\log\rho_t(x) = \dfrac{2c_i\partial_{x_i}\rho_t(x)}{\rho_t(x)}$. Since neural networks can implement multiplication as well as reciprocal-type operations (see Appendix C), it suffices to approximate the numerator and denominator separately by neural networks and then combine them to approximate the quotient.

Neural network approximation of ρ_t (Appendix E) To approximate $\rho_t := \mathcal{P}_t\rho_0$, we use the structure of ρ_0 : γ-*smoothness* (Assumption 4.1). As stated in Lemma A.6, letting $\widetilde{m}_t = m_t C_0 C_t^{-1}$ and $\widetilde{\sigma}_t = \sigma_t\sqrt{CC_t^{-1}}$, we may write ρ_t as a Gaussian convolution of ρ_0,

$$\begin{aligned}\rho_t(x) &= \int_{\mathbb{R}^\infty} \rho_0(\widetilde{m}_t x + \widetilde{\sigma}_t y)\,\mathrm{d}\mu_{C_0}(y)\\ &= \sum_{s\in\mathbb{N}_0^\infty}\sum_{l\in J(s)} \langle\rho_0,\psi_l\rangle \int_{\mathbb{R}^\infty}\psi_l(\widetilde{m}_t x + \widetilde{\sigma}_t y)\mathrm{d}\mu_{C_0}(y),\end{aligned}$$

that is, also as a superposition of convolutions of basis functions in the γ-smooth space.

Define the truncated expansion

$$R_\Gamma(\rho_0) := \sum_{s:\gamma(s)<\Gamma}\sum_{l\in J(s)}\langle\rho_0,\psi_l\rangle\psi_l.$$

Okumoto & Suzuki (2022) shows that one can approximate $R_\Gamma(\rho_0)$ by a neural network ϕ_ρ and ignore the remainder so that the approximation error $\|\rho_0 - \phi_\rho\|_2$ decays at rate $2^{-(1-v/a^\dagger)\Gamma}$.

We follow the same strategy for ρ_t: we approximate only the time-evolved truncated component $\mathcal{P}_t R_\Gamma(\rho_0)$ by a neural network. Moreover, there exists index $d_{\max}$ such that it holds $l_i = 0$ (which means $\psi_{l_i} = 1$) for all $i > d_{\max}$ for any basis element ψ_l with $\gamma(s) < \Gamma$ ($d_{\max}$ is officially defined in Definition B.1). Hence the integral in the basis convolution factorizes, and we can rewrite

$$\mathcal{P}_t R_\Gamma(\rho_0) = \sum_{s:\gamma(s)<\Gamma}\sum_{l\in J(s)}\langle\rho_0,\psi_l\rangle\prod_{i=1}^{d_{\max}}\int_{\mathbb{R}}\psi_{l_i}(\widetilde{m}_{t,i}x_i + \widetilde{\sigma}_{t,i}y_i)\mathrm{d}\mu_{C_0}(y_i).$$

For each coordinate, introduce the one-dimensional integral

$$\mathcal{I}_{l_i} := \int_{\mathbb{R}}\psi_{l_i}(\widetilde{m}_{t,i}x_i + \widetilde{\sigma}_{t,i}y_i)\mathrm{d}\mu_{C_0}(y_i).$$

By constructing a neural network $\phi_{\mathcal{I}_{l_i}}$ that approximates each $\mathcal{I}_{l_i}$, we can build a product network that approximates $\mathcal{P}_t\psi_l$. Multiplying by the coefficient $\langle\rho_0,\psi_l\rangle$ and summing over (s,l) yields a neural network approximation of $\rho_t(x)$.

Neural network approximation of $\partial_{x_i}\rho_t$ (Appendix G) The derivative $\partial_{x_i}\rho_t$ admits an analogous expansion:

$$\partial_{x_i}\rho_t = \sum_{s\in\mathbb{N}_0^\infty}\sum_{l\in J(s)}\langle\rho_0,\psi_l\rangle\frac{\widetilde{m}_{t,i}}{\widetilde{\sigma}_{t,i}c_{0,i}}\int_{\mathbb{R}} y_i\psi_{l_i}(\widetilde{m}_{t,i}x_i + \widetilde{\sigma}_{t,i}y_i)\mathrm{d}\mu_{C_0}(y_i)\prod_{j\neq i}\mathcal{I}_{l_j}.$$

All factors except the ith integral coincide with those in the construction of ρ_t, and thus the previously constructed subnetworks $\phi_{\mathcal{I}_{l_j}}$ can be reused. We approximate $\mathcal{J}_{l_i} := \int_{\mathbb{R}} y_i\psi_{l_i}(\widetilde{m}_{t,i}x_i + \widetilde{\sigma}_{t,i}y_i)\mathrm{d}\mu_{C_0}(y_i)$ by a neural network $\phi_{\mathcal{J}_{l_i}}$. Combining $\phi_{\mathcal{J}_{l_i}}$ with the product of the $\phi_{\mathcal{I}_{l_j}}$ then yields a network approximating $\partial_{x_i}\rho_t$.

Neural network approximation of the linear term To construct $\phi_{\text{score,linear}}(x_i,t)$, we approximate the scalar factor $c_i c_{t,i}^{-1}$ by a neural network and then multiply it by $-2x_i$. We can approximate m_t^2 from t by a neural network (Lemma C.8) and compose it with a reciprocal-approximation network to obtain a neural network approximation of $c_i c_{t,i}^{-1} = \{1 + m_t^2(c_{0,i}/c_i - 1)\}^{-1}$. Multiplication by x_i and the fixed factor -2 completes the construction of $\phi_{\text{score,linear}}(x_i,t)$.

D.2 DETAILED PROOF

As in (1), score function is defined as

$$s(x,t) := 2C\nabla \log \rho_t(x) - 2CC_t^{-1}x = (2c_i\partial_{x_i}\log\rho_t(x) - 2c_ic_{t,i}^{-1}x_i)_{i=1}^{\infty}.$$

The output of $s(x,t)$ is infinite-dimensional, so we set $\phi_{\text{score}} = (\phi^i_{\text{score}})_{i=1}^{\infty}$ as

$$\phi^i_{\text{score}}(x,t) := \begin{cases} \phi^i_{\text{score,grad}}(x,t) + \phi^i_{\text{score,linear}}(x_i,t) & i \le d_{\text{out}} \\ 0 & i > d_{\text{out}} \end{cases},$$

where $\phi^i_{\text{score,grad}}(x,t)$ and $\phi^i_{\text{score,linear}}(x_i,t)$ are neural networks that approximate $2c_i\partial_{x_i}\log\rho_t(x)$ and $-2c_ic_{t,i}^{-1}x_i$ each.

$\int_x \|\phi_{\text{score}}(x,t) - s(x,t)\|^2 \mathrm{d}P_t(x)$ can be decomposed to each dimension of score as follows:

$$\begin{aligned}
&\int_x \|2C\nabla\log\rho_t(x) - 2CC_t^{-1}x - \phi_{\text{score}}(x,t)\|^2 \rho_t(x)\mathrm{d}\mu_{C_t}(x) \\
&= \sum_{i=1}^{\infty}\int_x \left(2c_i\partial_{x_i}\log\rho_t(x) - 2c_ic_{t,i}^{-1}x_i - \phi^i_{\text{score}}(x,t)\right)^2 \rho_t(x)\mathrm{d}\mu_{C_t}(x) \\
&= \sum_{i=1}^{d_{\text{out}}}\int_x \left(2c_i\partial_{x_i}\log\rho_t(x) - 2c_ic_{t,i}^{-1}x_i - \phi^i_{\text{score}}(x,t)\right)^2 \rho_t(x)\mathrm{d}\mu_{C_t}(x) \\
&\quad + \sum_{i=d_{\text{out}}+1}^{\infty}\int_x \left(2c_i\partial_{x_i}\log\rho_t(x) - 2c_ic_{t,i}^{-1}x_i - \phi^i_{\text{score}}(x,t)\right)^2 \rho_t(x)\mathrm{d}\mu_{C_t}(x).
\end{aligned}$$

About $i > d_{\text{out}}$

$$\begin{aligned}
&\int_x \left(2c_i\partial_{x_i}\log\rho_t(x) - 2c_ic_{t,i}^{-1}x_i - \phi^i_{\text{score}}(x,t)\right)^2 \rho_t(x)\mathrm{d}\mu_{C_t}(x) \\
&= \int_x \left(2c_i\partial_{x_i}\rho_t(x)/\rho_t(x) - 2c_ic_{t,i}^{-1}x_i\right)^2 \rho_t(x)\mathrm{d}\mu_{C_t}(x) \\
&\le 2\int_x \left\{\left(2c_i\partial_{x_i}\rho_t(x)/\rho_t(x)\right)^2 + \left(2c_ic_{t,i}^{-1}x_i\right)^2\right\}\rho_t(x)\mathrm{d}\mu_{C_t}(x) \\
&\lesssim \int_x \left(c_i\partial_{x_i}\rho_t(x)/\rho_t(x)\right)^2 \rho_t(x)\mathrm{d}\mu_{C_t}(x) + \int_x \left(c_ic_{t,i}^{-1}x_i\right)^2 \rho_t(x)\mathrm{d}\mu_{C_t}(x) \\
&\lesssim \int_x \left(c_i\partial_{x_i}\rho_t(x)\right)^2 \mathrm{d}\mu_{C_t}(x) + \int_x c_i^2c_{t,i}^{-2}x_i^2\mathrm{d}\mu_{C_t}(x) \ \left(\|\rho_t - 1\|_\infty \le \frac{1}{2}\right) \\
&\lesssim \widetilde{m}_{t,i}^2c_i^2c_{0,i}^{-1} + c_i^2c_{t,i}^{-2}c_{t,i}^1 \qquad \left(\mathcal{O}(|\partial_{x_i}\rho_t(x)|) = \widetilde{m}_{t,i}/\sqrt{c_{0,i}}\right)
\end{aligned}$$

Hence,

$$\begin{aligned}
&\sum_{i=d_{\text{out}}+1}^{\infty}\int_x \left(2c_i\partial_{x_i}\log\rho_t(x) - 2c_ic_{t,i}^{-1}x_i - \phi^i_{\text{score}}(x,t)\right)^2 \rho_t(x)\mathrm{d}\mu_{C_t}(x) \\
&\lesssim \sum_{i=d_{\text{out}}+1}^{\infty}\left(\widetilde{m}_{t,i}^2c_i^2c_{0,i}^{-1} + c_i^2c_{t,i}^{-1}\right) \\
&\lesssim \sum_{i=d_{\text{out}}+1}^{\infty} i^{-\frac{2}{r_0}} \qquad (0 \le \widetilde{m}_{t,i}^2 = \mathcal{O}(1), \text{Assumption2.2}) \\
&\le \frac{r_0}{2-r_0}d_{\text{out}}^{1-\frac{2}{r_0}} \quad \text{(By simple integration calculation. See A.1.2 of Nishimura \& Suzuki (2024).)}.
\end{aligned}$$

About $i \le d_{\text{out}}$

$$\begin{aligned}
&\int_x \left(2c_i\partial_{x_i}\log\rho_t(x) - 2c_i c_{t,i}^{-1}x_i - \phi^i_{\text{score}}(x,t)\right)^2 \rho_t(x)\mathrm{d}\mu_{C_t}(x)\\
&= \int_x \left\{(2c_i\partial_{x_i}\log\rho_t(x) - \phi^i_{\text{score,grad}}(x,t)) + (-2c_i c_{t,i}^{-1}x_i - \phi^i_{\text{score,linear}}(x_i,t))\right\}^2 \rho_t(x)\mathrm{d}\mu_{C_t}(x)\\
&\le 2\int_x (2c_i\partial_{x_i}\log\rho_t(x) - \phi^i_{\text{score,grad}}(x,t))^2\rho_t(x)\mathrm{d}\mu_{C_t}(x) + 2\int_x (-2c_i c_{t,i}^{-1}x_i - \phi^i_{\text{score,linear}}(x_i,t))^2\rho_t(x)\mathrm{d}\mu_{C_t}(x)\\
&\lesssim \frac{1}{\sigma_t^2}c_i\Gamma\, 2^{-2(1-v/a^\dagger)\Gamma} + c_i\Gamma^{1/2}\, 2^{-2(1-v/a^\dagger)\Gamma}\\
&\lesssim \frac{1}{\sigma_t^2}c_i\Gamma\, 2^{-2(1-v/a^\dagger)\Gamma}.
\end{aligned}$$

The second last inequality is proved in the below subsection.

Hence,

$$\begin{aligned}
\sum_{i=1}^{d_{\text{out}}}\int_x \left(2c_i\partial_{x_i}\log\rho_t(x) - 2c_i c_{t,i}^{-1}x_i - \phi^i_{\text{score}}(x,t)\right)^2 \rho_t(x)\mathrm{d}\mu_{C_t}(x) &\lesssim \sum_{i=1}^{d_{\text{out}}}\frac{1}{\sigma_t^2}c_i\Gamma\, 2^{-2(1-v/a^\dagger)\Gamma}\\
&\lesssim \frac{1}{\sigma_t^2}\Gamma\, 2^{-2(1-v/a^\dagger)\Gamma}.
\end{aligned}$$

The last inequality is induced like $\sum_{i=1}^{d_{\text{out}}} i^{-2/r_0} \le 1 + \int_1^{d_{\text{out}}} x^{-2/r_0}\mathrm{d}x \le 2/(2-r_0) \sim 1$.

About $\phi^i_{\text{score,linear}}$ In this paragraph, $\phi_{\text{score,linear}}{}^i_{\text{raw}}$ denotes the linear approximation before the final output clipping. Since $\|\rho_t\|_\infty \le 3/2$, we have

$$\begin{aligned}
&\int \left(-2c_i c_{t,i}^{-1}x_i - \phi_{\text{score,linear}}{}^i_{\text{raw}}(x_i,t)\right)^2 \rho_t(x)\,\mathrm{d}\mu_{C_t}(x)\\
&\lesssim \int_{\mathbb{R}} \left(-2c_i c_{t,i}^{-1}x_i - \phi_{\text{score,linear}}{}^i_{\text{raw}}(x_i,t)\right)^2 \mathrm{d}\mu_{c_{t,i}}(x_i)\\
&= \int_{|x_i|<B_{x_i}} \left(-2c_i c_{t,i}^{-1}x_i - \phi_{\text{score,linear}}{}^i_{\text{raw}}(x_i,t)\right)^2 \mathrm{d}\mu_{c_{t,i}}(x_i)\\
&\quad + \int_{|x_i|\ge B_{x_i}} \left(-2c_i c_{t,i}^{-1}x_i - \phi_{\text{score,linear}}{}^i_{\text{raw}}(x_i,t)\right)^2 \mathrm{d}\mu_{c_{t,i}}(x_i). \qquad (3)
\end{aligned}$$

Instead of setting the output to zero outside the interval, we clip the input:

$$x_i^{\text{clip}} := \text{clip}_{[-B_{x_i},B_{x_i}]}(x_i).$$

This map is represented exactly by a continuous ReLU network. We construct $\phi_{\text{score,linear}}{}^i_{\text{raw}}$ so that

$$\left|\phi_{\text{score,linear}}{}^i_{\text{raw}}(x_i,t) + 2c_i c_{t,i}^{-1}x_i^{\text{clip}}\right| \le \varepsilon_{\text{linear}}$$

for all x_i and $t \in [\underline{T},\overline{T}]$. Since $c_i/c_{t,i} \le K$, the inner-domain error is at most $\varepsilon^2_{\text{linear}}$, while

$$\begin{aligned}
&\int_{|x_i|\ge B_{x_i}} \left(-2c_i c_{t,i}^{-1}x_i - \phi_{\text{score,linear}}{}^i_{\text{raw}}(x_i,t)\right)^2 \mathrm{d}\mu_{c_{t,i}}(x_i)\\
&\le 2\varepsilon^2_{\text{linear}} + 8K^2\int_{|x_i|\ge B_{x_i}} (|x_i| - B_{x_i})^2\,\mathrm{d}\mu_{c_{t,i}}(x_i)\\
&\le 2\varepsilon^2_{\text{linear}} + 8K^2\int_{|x_i|\ge B_{x_i}} x_i^2\,\mathrm{d}\mu_{c_{t,i}}(x_i).
\end{aligned}$$

By Remark A.3,

$$\int_{|x_i|\ge B_{x_i}} x_i^2\,\mathrm{d}\mu_{c_{t,i}}(x_i) \le \sqrt{\frac{2}{\pi}}\,c_{t,i}e^{-B^2_{x_i}/(2c_{t,i})}\left(\frac{B_{x_i}}{\sqrt{c_{t,i}}} + \frac{\sqrt{c_{t,i}}}{B_{x_i}}\right).$$

We next construct the required approximation. The time-dependent coefficient satisfies

$$\frac{c_i}{c_{t,i}} = \left[1 + \left(\frac{c_{0,i}}{c_i} - 1\right) m_t^2\right]^{-1} \in [K^{-1}, K].$$

Using Lemma C.8 and Lemmas C.5 and C.6, construct $\phi_{c_i/c_{t,i}}$ with error at most $\varepsilon_{\text{coef}}$. The denominator and reciprocal approximations are clipped to $[K^{-1}, K]$, so that

$$\phi_{c_i/c_{t,i}}(t) \in [K^{-1}, K].$$

Apply the two-input multiplication network on $[-1, 1]^2$ with error at most $\varepsilon_{\text{coef}}$, and define

$$\phi_{\text{score,linear}}{}^{i}_{\text{raw}}(x_i, t) := -2KB_{x_i}\, \phi_{\text{mult}}\left(\frac{x_i^{\text{clip}}}{B_{x_i}}, \frac{\phi_{c_i/c_{t,i}}(t)}{K}\right).$$

Taking

$$\varepsilon_{\text{coef}} := \frac{\varepsilon_{\text{linear}}}{8KB_{x_i}}$$

gives

$$\begin{aligned}\left|\phi_{\text{score,linear}}{}^{i}_{\text{raw}}(x_i, t) + 2c_i c_{t,i}^{-1} x_i^{\text{clip}}\right| &\le 2KB_{x_i}\varepsilon_{\text{coef}} + 2B_{x_i}\varepsilon_{\text{coef}} \\ &\le \varepsilon_{\text{linear}}.\end{aligned}$$

Choose the time-independent radius and accuracy as

$$B_{x_i} := A_{\text{linear}}\sqrt{c_i\Gamma}, \qquad \varepsilon_{\text{linear}} := \sqrt{c_i}\, 2^{-(1-v/a^\dagger)\Gamma},$$

where $\Gamma \ge 2$ and $A_{\text{linear}} \ge 1$ is a fixed constant satisfying

$$\frac{A_{\text{linear}}^2}{2K} \ge 2(1 - v/a^\dagger)\log 2.$$

Since $K^{-1}c_i \le c_{t,i} \le Kc_i$, (3) yields

$$\begin{aligned}&\int \left(-2c_i c_{t,i}^{-1} x_i - \phi_{\text{score,linear}}{}^{i}_{\text{raw}}(x_i, t)\right)^2 \rho_t(x)\, \mathrm{d}\mu_{C_t}(x) \\ &\lesssim \varepsilon_{\text{linear}}^2 + c_i\Gamma^{1/2}\exp\left(-\frac{A_{\text{linear}}^2\Gamma}{2K}\right) \\ &\lesssim c_i\, 2^{-2(1-v/a^\dagger)\Gamma} + c_i\Gamma^{1/2}2^{-2(1-v/a^\dagger)\Gamma} \\ &\lesssim c_i\Gamma^{1/2}2^{-2(1-v/a^\dagger)\Gamma}.\end{aligned}$$

Finally, we evaluate the network size. The chosen coefficient tolerance satisfies

$$\varepsilon_{\text{coef}} = \frac{2^{-(1-v/a^\dagger)\Gamma}}{8KA_{\text{linear}}\sqrt{\Gamma}}, \qquad \log \varepsilon_{\text{coef}}^{-1} = \mathcal{O}(\Gamma).$$

Including the time-coefficient approximation, reciprocal, multiplication, input clipping, and normalization, sufficient scalar-network budgets are

$$\begin{aligned}L &= \mathcal{O}(\Gamma^2), \qquad W = \mathcal{O}(\Gamma^3), \\ S &= \mathcal{O}(\Gamma^4), \qquad B \le C\max\left\{c_i^{-1/2}, \exp(C\Gamma)\right\}.\end{aligned}$$

The input normalization contributes the coefficient $B_{x_i}^{-1} \lesssim c_i^{-1/2}$. Fixed identity channels transmit the spatial input through the time subnetworks, and their costs are included in these budgets. Explicit signed ReLU interfaces are retained when composing subnetworks.

The scalar networks are needed for $i = 1, \dots, d_{\text{out}}$. Constructing them in parallel gives sufficient budgets for the underlying linear branch:

$$\begin{aligned} L_{\text{linear}} &= \mathcal{O}(\Gamma^2), \\ W_{\text{linear}} &= \mathcal{O}(d_{\text{out}}\Gamma^3), \\ S_{\text{linear}} &= \mathcal{O}(d_{\text{out}}\Gamma^4), \\ B_{\text{linear}} &\le C \max\left\{ \max_{i \le d_{\text{out}}} c_i^{-1/2}, \exp(C\Gamma) \right\}. \end{aligned}$$

The final output clipping prescribed in $\mathcal{S}$ is applied afterwards. Its additional approximation error is controlled by Lemma D.1, and its fixed layers are included in the score-network assembly.

About $\phi^i_{\text{score,grad}}$ We first construct the gradient approximation for $i \le d_{\max}$. The approximation error for omitted gradient coordinates is treated separately.

We clip the density-ratio approximation to its known range:

$$\phi^{\text{clip}}_{\rho_t}(x,t) := \text{clip}_{[1/2,3/2]}\big(\phi_{\rho_t}(x,t)\big).$$

Here,

$$\text{clip}_{[a,b]}(z) := a + \text{ReLU}(z-a) - \text{ReLU}(z-b)$$

is realized exactly by a fixed ReLU network. Since $1/2 \le \rho_t \le 3/2$,

$$|\phi^{\text{clip}}_{\rho_t}(x,t) - \rho_t(x)| \le |\phi_{\rho_t}(x,t) - \rho_t(x)|.$$

By Assumptions 2.2 and 4.2, choose a fixed constant $K_{\text{num}} > 0$ sufficiently large that

$$|2c_i \partial_{x_i} \rho_t(x)| \le K_{\text{num}} \sqrt{c_i}$$

uniformly in x, t, and i. Define

$$\phi^{i,\text{clip}}_{2c\nabla\rho_t}(x,t) := \text{clip}_{[-K_{\text{num}}\sqrt{c_i},\, K_{\text{num}}\sqrt{c_i}]}\big(\phi^i_{2c\nabla\rho_t}(x,t)\big).$$

This clipping does not increase the numerator approximation error.

Let ϕ_{rec} approximate the reciprocal function on $[1/2, 3/2]$ with uniform error at most ε_{rec}, and define

$$\phi^{\text{clip}}_{\text{rec}}(z) := \text{clip}_{[2/3,2]}\big(\phi_{\text{rec}}(z)\big).$$

For $z \in [1/2, 3/2]$, the true reciprocal $1/z$ belongs to $[2/3, 2]$, so this clipping does not increase the reciprocal approximation error.

The approximation before the final gradient-output clipping is

$$\phi_{\text{score,grad}^i_{\text{raw}}}(x,t) := \phi_{\text{mult}}\left(\phi^{i,\text{clip}}_{2c\nabla\rho_t}(x,t), \phi^{\text{clip}}_{\text{rec}}\big(\phi^{\text{clip}}_{\rho_t}(x,t)\big)\right).$$

The multiplication network is chosen to have uniform error at most $\varepsilon_{\text{mult},i}$ on $[-K_{\text{num}}, K_{\text{num}}] \times [2/3, 2]$. This fixed input domain contains all inputs to the multiplication network because $c_i \le 1$.

The reciprocal function is 4-Lipschitz on $[1/2, 3/2]$. Consequently,

$$\begin{aligned} \left|\phi^{\text{clip}}_{\text{rec}}\big(\phi^{\text{clip}}_{\rho_t}(x,t)\big) - \frac{1}{\rho_t(x)}\right| &\le \varepsilon_{\text{rec}} + \left|\frac{1}{\phi^{\text{clip}}_{\rho_t}(x,t)} - \frac{1}{\rho_t(x)}\right| \\ &\le \varepsilon_{\text{rec}} + 4|\phi_{\rho_t}(x,t) - \rho_t(x)|. \end{aligned}$$

Using the bounded reciprocal output and the numerator bound, we obtain

$$\begin{aligned} &\left|\phi_{\text{score,grad}^i_{\text{raw}}}(x,t) - \frac{2c_i\partial_{x_i}\rho_t(x)}{\rho_t(x)}\right| \\ &\quad\le \varepsilon_{\text{mult},i} + 2\left|\phi^{i,\text{clip}}_{2c\nabla\rho_t}(x,t) - 2c_i\partial_{x_i}\rho_t(x)\right| + K_{\text{num}}\sqrt{c_i}\left[\varepsilon_{\text{rec}} + 4|\phi_{\rho_t}(x,t) - \rho_t(x)|\right]. \end{aligned}$$

Squaring and integrating with respect to P_t yields

$$\int \left|\phi_{\mathrm{score,grad}^i_{\mathrm{raw}}}(x,t) - 2c_i\partial_{x_i}\log\rho_t(x)\right|^2 \mathrm{d}P_t(x)$$
$$\lesssim \varepsilon^2_{\mathrm{mult},i} + \left\|\phi^i_{2c\nabla\rho_t} - 2c_i\partial_{x_i}\rho_t\right\|^2_{L^2(P_t)} + c_i\varepsilon^2_{\mathrm{rec}} + c_i\|\phi_{\rho_t} - \rho_t\|^2_{L^2(P_t)}.$$

Here we used nonexpansiveness of the numerator clipping. Since $\rho_t \le 3/2$, we also have

$$\|\phi_{\rho_t} - \rho_t\|^2_{L^2(P_t)} \le \frac{3}{2}\|\phi_{\rho_t} - \rho_t\|^2_{L^2(\mu_{C_t})}.$$

Using (15) and Lemma E.1, together with $P_t \le (3/2)\mu_{C_t}$, the preceding estimate gives

$$\int \left|\phi_{\mathrm{score,grad}^i_{\mathrm{raw}}}(x,t) - 2c_i\partial_{x_i}\log\rho_t(x)\right|^2 \mathrm{d}P_t(x)$$
$$\lesssim \varepsilon^2_{\mathrm{mult},i} + c_i\Gamma\sigma_t^{-2}2^{-2(1-v/a^\dagger)\Gamma} + c_i\varepsilon^2_{\mathrm{rec}} + c_i 2^{-2(1-v/a^\dagger)\Gamma}.$$

Choose

$$\varepsilon_{\mathrm{mult},i} := \sqrt{c_i}\,2^{-(1-v/a^\dagger)\Gamma}, \qquad \varepsilon_{\mathrm{rec}} := 2^{-(1-v/a^\dagger)\Gamma}.$$

For sufficiently large Γ, using $\sigma_t^2 \le 1$, we conclude that

$$\int \left|\phi_{\mathrm{score,grad}^i_{\mathrm{raw}}}(x,t) - 2c_i\partial_{x_i}\log\rho_t(x)\right|^2 \mathrm{d}P_t(x) \lesssim c_i\Gamma\sigma_t^{-2}2^{-2(1-v/a^\dagger)\Gamma}.$$

The final gradient-output clipping is applied as in Lemma D.1. For each constructed gradient coordinate, the true relative score lies within its clipping interval. Thus this final clipping does not increase the approximation error. The error in omitted gradient coordinates is treated separately.

The fixed scalar clipping maps add only constant depth, width, and sparsity per scalar use, with uniformly bounded coefficients. The multiplication and reciprocal network budgets are evaluated at the accuracies specified above. In particular,

$$\log\varepsilon^{-1}_{\mathrm{mult},i} = (1 - v/a^\dagger)\Gamma\log 2 + \frac{1}{2}\log c_i^{-1}.$$

The coordinate dependence of the multiplication accuracy is therefore included in the network-size calculation. Under $d_{\max} \lesssim \Gamma^{1/\eta}$ and $c_i \simeq i^{-2/r_0}$, this additional logarithmic term is $\mathcal{O}(\log\Gamma)$ for $i \le d_{\max}$.

Clipping and approximation of the marginal score. The following lemma transfers the branch-wise approximation estimates to the clipped hypothesis class used in Appendix H.

Lemma D.1 (Clipped approximation and marginalization)**.** *Assume the standing density-ratio and covariance assumptions, $\Gamma \ge 1$, $n \ge 2$, $0 < \underline{T} < \overline{T}$, and $d_{\max} \le d_{\mathrm{out}}$. Write*

$$v^i(x,t) := 2c_i\partial_{x_i}\log\rho_t(x), \qquad a_i(x_i,t) := -2c_i c_{t,i}^{-1}x_i,$$

so that $s^i = v^i + a_i$.

Let g_Γ and ℓ_Γ be raw approximators in the underlying gradient and linear branch classes defining $\mathcal{S}$, with g_Γ extended by zero beyond $d_{\max}$, satisfying the following bounds for every $i \le d_{\mathrm{out}}$ and $t \in [\underline{T}, \overline{T}]$, with implicit constants independent of n, d_{out}, Γ, and t:

$$\mathbb{E}_{P_t}\left[|g^i_\Gamma(X_{t,[1:d_{\max}]},t) - v^i(X_t,t)|^2\right] \lesssim c_i\Gamma\sigma_t^{-2}2^{-2(1-v/a^\dagger)\Gamma}, \tag{4}$$

$$\mathbb{E}_{P_t}\left[|\ell^i_\Gamma(X_{t,[1:d_{\mathrm{out}}]},t) - a_i(X_{t,i},t)|^2\right] \lesssim c_i\Gamma^{1/2}2^{-2(1-v/a^\dagger)\Gamma}. \tag{5}$$

For $i > d_{\max}$, the first bound includes the error from setting the gradient output to zero.

Let $p > 0$ be fixed. Choose sufficiently large fixed clipping constants, allowing R_{linear} to depend on p. Using the thresholds and zero-padding map defined in Section 2, set

$$f_{\Gamma,n}(x,t) := E_{d_{\max},d_{\mathrm{out}}}\,\mathrm{clip}_{b_{\mathrm{grad}}}\big(g_\Gamma(x_{[1:d_{\max}]},t)\big) + \mathrm{clip}_{b_{\mathrm{linear}}}\big(\ell_\Gamma(x,t)\big),$$

where

$$b_{\mathrm{grad},i} = R_{\mathrm{grad}}\sqrt{c_i}\, i^{-1/2-\delta_0}, \qquad b_{\mathrm{linear},i} = R_{\mathrm{linear}}\sqrt{c_i \log(en)}.$$

Then $f_{\Gamma,n} \in \mathcal{S}$ and, for every $i \le d_{\mathrm{out}}$ and $t \in [\underline{T}, \overline{T}]$,

$$\mathbb{E}_{P_t}\left[|f^i_{\Gamma,n}(X_{t,[1:d_{\mathrm{out}}]}, t) - s^i(X_t, t)|^2\right] \lesssim c_i\left[\Gamma\sigma_t^{-2}2^{-2(1-v/a^\dagger)\Gamma} + n^{-p}\right].$$

Define

$$s_{d_{\mathrm{out}}}(x,t) := 2C_{[1:d_{\mathrm{out}}]}\nabla \log \tilde{p}_t(x), \qquad J_2 := \frac{1}{\overline{T}-\underline{T}}\int_{\underline{T}}^{\overline{T}} \sigma_t^{-2}\,\mathrm{d}t,$$

and assume $J_2 < \infty$. For $A \in \{I_{d_{\mathrm{out}}}, C^{-1}_{[1:d_{\mathrm{out}}]}\}$, it follows that

$$\inf_{f\in\mathcal{S}} \bar{R}_{\underline{T},\overline{T};A}(f, s_{d_{\mathrm{out}}}) \lesssim \mathrm{Tr}\big(AC_{[1:d_{\mathrm{out}}]}\big)\left[\Gamma J_2 2^{-2(1-v/a^\dagger)\Gamma} + n^{-p}\right].$$

The resulting bounds have implicit constants independent of n, d_{out}, Γ, and t. Dependence on p and the fixed distributional and clipping constants is allowed.

Proof. We first control the effect of clipping. For a scalar projection onto $[-b, b]$,

$$|\operatorname{clip}_b(z) - a| \le |z - a| \qquad \text{if } |a| \le b.$$

The relative-score bound gives

$$|v^i(x,t)| \lesssim \sqrt{c_i}\, i^{-1/2-\delta_0}.$$

Thus, for sufficiently large R_{grad}, gradient-output clipping does not increase the approximation error for $i \le d_{\max}$. The zero outputs beyond $d_{\max}$ remain unchanged, and their errors are included in (4).

For arbitrary $a, z \in \mathbb{R}$, nonexpansiveness of scalar projection also gives

$$|\operatorname{clip}_b(z) - a| \le |z - a| + (|a| - b)_+,$$

and therefore

$$|\operatorname{clip}_b(z) - a|^2 \le 2|z-a|^2 + 2(|a|-b)_+^2.$$

Under the reference Gaussian measure μ_{C_t}, the linear score coordinate $a_i(X_{t,i}, t)$ is centered Gaussian with variance

$$4c_i^2/c_{t,i} \le 4Kc_i,$$

where K is the covariance comparison constant. Since $P_t \le (3/2)\mu_{C_t}$,

$$\mathbb{E}_{P_t}\exp\left(\frac{|a_i(X_{t,i},t)|^2}{16Kc_i}\right) \le \frac{3}{2}\sqrt{2}.$$

Markov's inequality and integration of the tail yield

$$\begin{aligned}
&\mathbb{E}_{P_t}\left[(|a_i(X_{t,i},t)| - b)_+^2\right] \\
&\quad = 2\int_b^\infty (u-b)P_t(|a_i(X_{t,i},t)| > u)\,\mathrm{d}u \\
&\quad \lesssim c_i \exp\left(-\frac{b^2}{16Kc_i}\right).
\end{aligned}$$

Taking $b = R_{\mathrm{linear}}\sqrt{c_i\log(en)}$ and $R^2_{\mathrm{linear}} \ge 16Kp$ gives

$$\mathbb{E}_{P_t}\left[(|a_i(X_{t,i},t)| - b_{\mathrm{linear},i})_+^2\right] \lesssim c_i n^{-p}.$$

Combining the two branchwise errors, using $|u+v|^2 \le 2|u|^2 + 2|v|^2$, $\Gamma \ge 1$, and $\sigma_t^2 \le 1$, proves the coordinatewise approximation bound. Applying the prescribed clipping maps to these raw branch networks gives $f_{\Gamma,n} \in \mathcal{S}$ by the definition of the class.

We transfer this bound to the marginal score by conditional expectation. Since the reference covariance is diagonal,

$$\tilde{\rho}_t(z) = \int \rho_t(z,w)\,\mu_{C_{t,>d_{\mathrm{out}}}}(\mathrm{d}w),$$

where

$$\tilde{\rho}_t := \frac{\mathrm{d}\tilde{P}_t}{\mathrm{d}\mu_{C_{t,[1:d_{\mathrm{out}}]}}},$$

and $\mu_{C_{t,>d_{\mathrm{out}}}}$ denotes the reference Gaussian measure on the omitted coordinates. The coordinatewise derivative bounds justify differentiation under this integral. Using the conditional density of the omitted coordinates, we obtain

$$s^i_{d_{\mathrm{out}}}(X_{t,[1:d_{\mathrm{out}}]}, t) = \mathbb{E}_{P_t}\left[s^i(X_t, t) \mid X_{t,[1:d_{\mathrm{out}}]}\right], \qquad i \le d_{\mathrm{out}}.$$

Because $f_{\Gamma,n}$ depends only on the retained coordinates, conditional Jensen's inequality implies

$$\begin{aligned}
&\mathbb{E}_{\tilde{P}_t}\left[\|f_{\Gamma,n}(\tilde{X}_t, t) - s_{d_{\mathrm{out}}}(\tilde{X}_t, t)\|_A^2\right] \\
&\quad\le \mathbb{E}_{P_t}\left[\|f_{\Gamma,n}(X_{t,[1:d_{\mathrm{out}}]}, t) - s(X_t, t)_{[1:d_{\mathrm{out}}]}\|_A^2\right] \\
&\quad\lesssim \mathrm{Tr}\big(AC_{[1:d_{\mathrm{out}}]}\big)\left[\Gamma\sigma_t^{-2}2^{-2(1-v/a^\dagger)\Gamma} + n^{-p}\right].
\end{aligned}$$

Averaging over forward time $t \in [\underline{T}, \overline{T}]$ and using $f_{\Gamma,n} \in \mathcal{S}$ proves the assertion. □

Finally, we evaluate the score-network budgets using the density and derivative constructions. Assume $\Gamma \ge 2$ and $1 \le d_{\max} \le d_{\mathrm{out}}$.

Let K_ρ be the complexity quantity in Lemma E.1, and let K' be the corresponding quantity in the retained-derivative construction.

Density and numerator subnetworks. By (10), the density network has budgets

$$\begin{aligned}
L_\rho &= \mathcal{O}(K_\rho^2), \\
W_\rho &= \mathcal{O}(d_{\max} G K_\rho^3), \\
S_\rho &= \mathcal{O}(d_{\max} G K_\rho^4), \\
B_\rho^{\mathrm{net}} &\le C \max\left\{\max_{i \le d_{\max}} c_{0,i}^{-1/2}, \exp(CK_\rho)\right\}.
\end{aligned}$$

The retained-derivative vector network has sufficient budgets

$$\begin{aligned}
L_{\mathrm{num}} &= \mathcal{O}(K'^2), \\
W_{\mathrm{num}} &= \mathcal{O}(d_{\max} G K'^4), \\
S_{\mathrm{num}} &= \mathcal{O}(d_{\max} G K'^5), \\
B_{\mathrm{num}} &\le C \max\left\{\max_{i \le d_{\max}} c_{0,i}^{-1/2}, \exp(CK')\right\}.
\end{aligned}$$

Gradient branch. The gradient branch combines the numerator vector with the shared reciprocal of the clipped density approximation. The reciprocal accuracy is

$$\varepsilon_{\mathrm{rec}} = 2^{-(1-v/a^\dagger)\Gamma},$$

and the multiplication accuracy for coordinate i is

$$\varepsilon_{\mathrm{mult},i} = \sqrt{c_i}\, 2^{-(1-v/a^\dagger)\Gamma}.$$

Consequently,

$$\log \varepsilon_{\mathrm{mult},i}^{-1} = (1 - v/a^\dagger)\Gamma \log 2 + \frac{1}{2}\log c_i^{-1}.$$

Under covariance decay and $d_{\max} \lesssim \Gamma^{1/\eta}$, the additional coordinate term is $\mathcal{O}(\log \Gamma)$ for $i \le d_{\max}$.

The reciprocal network has depth $\mathcal{O}(\Gamma^2)$, width $\mathcal{O}(\Gamma^3)$, sparsity $\mathcal{O}(\Gamma^4)$, and logarithmic weight bound $\mathcal{O}(\Gamma)$. The coordinatewise multiplication networks and the fixed internal clipping maps are appended after aligning the numerator and reciprocal outputs.

The density approximation is computed once and shared across the output coordinates. Since K_ρ and K' dominate Γ up to fixed constants, the reciprocal, multiplication, and identity-layer costs are absorbed in the following sufficient budgets:

$$\begin{aligned}
L_{\mathrm{grad}} &= \mathcal{O}(K_\rho^2 + K'^2),\\
W_{\mathrm{grad}} &= \mathcal{O}\left(d_{\max} G (K_\rho^3 + K'^4)\right),\\
S_{\mathrm{grad}} &= \mathcal{O}\left(d_{\max} G (K_\rho^4 + K'^5)\right),\\
B_{\mathrm{grad}} &\le C \max\left\{ \max_{i \le d_{\max}} c_{0,i}^{-1/2}, \exp\big(C(K_\rho + K')\big) \right\}.
\end{aligned}$$

Here the gradient-branch budgets include the internal numerator, denominator, and reciprocal-output clipping. The final gradient-output clipping is counted below.

Final clipping and combination with the linear branch. The underlying linear branch constructed above has the following sufficient budgets:

$$\begin{aligned}
L_{\mathrm{linear}} &= \mathcal{O}(\Gamma^2),\\
W_{\mathrm{linear}} &= \mathcal{O}(d_{\mathrm{out}} \Gamma^3),\\
S_{\mathrm{linear}} &= \mathcal{O}(d_{\mathrm{out}} \Gamma^4),\\
B_{\mathrm{linear}} &= \mathcal{O}\left(\max_{i \le d_{\mathrm{out}}} c_i^{-1/2} \vee \Gamma\, 2^{2(1 - v/a^\dagger)\Gamma} \right).
\end{aligned}$$

These budgets include input clipping, time-coefficient approximation, multiplication, and the identity channels within the linear branch. The final output clipping and the depth alignment between the gradient and linear branches are counted below.

Append the fixed output clipping maps with thresholds

$$b_{\mathrm{grad},i} = R_{\mathrm{grad}} \sqrt{c_i}\, i^{-1/2-\delta_0}, \qquad b_{\mathrm{linear},i} = R_{\mathrm{linear}} \sqrt{c_i \log(en)}.$$

After zero-padding the gradient output, add the two branches. The fixed clipping and addition operations contribute constant depth and $\mathcal{O}(d_{\max} + d_{\mathrm{out}})$ width and sparsity.

Align unequal branch depths using signed identity channels. Sufficient single-network budgets satisfy

$$\begin{aligned}
L_{\mathrm{score}} &\le \max\{L_{\mathrm{grad}}, L_{\mathrm{linear}}\} + \mathcal{O}(1),\\
W_{\mathrm{score}} &\lesssim W_{\mathrm{grad}} + W_{\mathrm{linear}} + d_{\mathrm{out}},\\
S_{\mathrm{score}} &\lesssim S_{\mathrm{grad}} + S_{\mathrm{linear}} + d_{\mathrm{out}}\\
&\qquad + d_{\mathrm{out}}(L_{\mathrm{grad}} - L_{\mathrm{linear}})_+ + d_{\max}(L_{\mathrm{linear}} - L_{\mathrm{grad}})_+,\\
B_{\mathrm{score}} &\lesssim 1 \vee B_{\mathrm{grad}} \vee B_{\mathrm{linear}} \vee R_{\mathrm{grad}} \vee R_{\mathrm{linear}} \sqrt{\log(en)}.
\end{aligned}$$

If $L_{\mathrm{grad}} \ge L_{\mathrm{linear}}$, the sparsity bound becomes

$$S_{\mathrm{score}} \lesssim S_{\mathrm{grad}} + S_{\mathrm{linear}} + d_{\mathrm{out}}\big(1 + L_{\mathrm{grad}} - L_{\mathrm{linear}}\big).$$

Explicit signed ReLU interfaces are retained during composition, so adjacent affine matrices need not be merged. The identity and clipping coefficients are counted in the single-network sparsity. They do not introduce additional trainable parameters in the branchwise covering argument of Appendix H.

The clipped approximation errors are transferred to the marginal score by Lemma D.1.

E Approximation of ρ_t

In this section, we prove Lemma E.1, that is, ρ_t is approximated with the neural network ϕ_{ρ_t}.

Lemma E.1 (Approximation error of ρ_t by FNNs)**.** *Assume Assumptions 2.2, 4.1, 4.2, and 4.3. Let $v/a^\dagger < 1$, $\Gamma \ge 2$, and $0 < \underline{T} < \overline{T} < \infty$. The time approximations used in Lemma E.4 follow from Lemma C.8 under the standing noise-schedule assumptions.*

Let $d_{\max}$, $f_{\max}$, and G denote the complexity quantities associated with the truncation $\gamma(s) < \Gamma$, and suppose $d_{\max} \geq 1$. Choose ε as in (9), and set

$$K_\rho := 1 + d_{\max} + \log \varepsilon^{-1} + \log B_\rho + \max_{\substack{l \in \mathbb{Z} \\ 1 \leq |l| \leq 2^{f_{\max}}}} \Lambda_l(\varepsilon),$$

where Λ_l is defined in (8).

Then there exists $\phi_\rho \in \Phi(L, W, S, B; d_{\max} + 1, 1)$ with sufficient budgets

$$\begin{aligned} L &= \mathcal{O}(K_\rho^2), \\ W &= \mathcal{O}(d_{\max} G K_\rho^3), \\ S &= \mathcal{O}(d_{\max} G K_\rho^4), \\ B &\leq C \max\left\{ \max_{i \leq d_{\max}} c_{0,i}^{-1/2}, \exp(C K_\rho) \right\}, \end{aligned}$$

such that

$$\|\rho_t - \phi_\rho(\cdot, t)\|_{L^2(\mu_{C_t})} \lesssim 2^{-(1 - v/a^\dagger)\Gamma} \|\rho_0\|_{\mathcal{F}^\gamma_{p,q}}, \qquad t \in [\underline{T}, \overline{T}].$$

If the retained expansion contains only the constant basis function, its approximation is represented exactly by a constant network.

The proof is via two term decomposition as

$$\|\rho_t - \phi_\rho(t)\|_{L^2(R^\infty, \mu_{C_t})} \leq \underbrace{\|\rho_t - \mathcal{P}_t R_\Gamma(\rho_0)\|_{L^2(R^\infty, \mu_{C_t})}}_{\text{(I)}} + \underbrace{\|\mathcal{P}_t R_\Gamma(\rho_0) - \phi_\rho(t)\|_{L^2(R^\infty, \mu_{C_t})}}_{\text{(II)}}. \quad (6)$$

The first term (I) shows the error from truncation of the γ-smooth basis that $\gamma(s) \geq \Gamma$, and the second term (II) shows the error from approximation of the retained basis terms by a neural network.

E.1 The error bound of (I)

We can easily check

$$\|\psi_l\|_2 := \|\psi_l\|_{L^2(R^\infty, \mu_{C_0})} = \left\| \prod_{i=1}^{\infty} (\cos 2\pi l_i x_i - \sin 2\pi l_i x_i) \right\|_{L^2([0,1]^\infty, \lambda^\infty)},$$

so along with Okumoto & Suzuki (2022), we have that

$$\|\delta_s(f)\|_{L^2([0,1]^\infty, \lambda^\infty)} \leq 2^{vs} \|\delta_s(f)\|_{L^p([0,1]^\infty, \lambda^\infty)}$$

where $v = \left(\frac{1}{p} - \frac{1}{2}\right)_+$ and $2^{vs} = 2^{v \sum_{i=1}^\infty s_i}$ by using Theorem 1 of Nessel & Wilmes (1978).

We consider approximating ρ_t by $\mathcal{P}_t R_\Gamma(\rho_0)$. $R_\Gamma(\rho_0)$ is defined as

$$R_\Gamma(\rho_0) := \begin{cases} \sum_{s \in \mathbb{N}_0^\infty : \gamma(s) < \Gamma} \delta_s(\rho_0) & (1 \leq q \leq 2) \\ \sum_{s \in \mathbb{N}_0^\infty : \gamma'(s) < \Gamma} \delta_s(\rho_0) & (q > 2) \end{cases}$$

Then, we further approximate $\mathcal{P}_t R_\Gamma(\rho_0)$ by a fully connected neural network. For that purpose, we first analyze the approximation error $\|\rho_t - \mathcal{P}_t R_\Gamma(\rho_0)\|_{L^2(\mathbb{R}^\infty, \mu_{C_t})}$ by $\mathcal{P}_t R_\Gamma(\rho_0)$. This can be evaluated in the following lemma.

Lemma E.2 (Evaluation of remainder $\rho_t - \mathcal{P}_t R_\Gamma(\rho_0)$)**.** *Assume that $\gamma, \gamma' : \mathbb{N}_0^\infty \to \mathbb{R}_+$ satisfy $\gamma'(s) < \gamma(s), v\alpha(\gamma) < 1, v\alpha(\gamma') < 1$, we have that*

$$\|\rho_t - \mathcal{P}_t R_\Gamma(\rho_0)\|_{L^2(\mathbb{R}^\infty, \mu_{C_t})} \leq \begin{cases} 2^{-(1 - v\alpha)\Gamma} \|\rho_0\|_{\mathcal{F}^\gamma_{p,q}} & (1 \leq q \leq 2) \\ 2^{-(1 - v\alpha')\Gamma} \left[\sum_{s \in \mathbb{N}_0^\infty : \Gamma \leq \gamma'(s)} 2^{\frac{2q}{q-2}(\gamma'(s) - \gamma(s))} \right]^{1/2 - 1/q} \|\rho_0\|_{\mathcal{F}^\gamma_{p,q}} & (q > 2). \end{cases}$$

where $\alpha' := \alpha(\gamma')$.

Moreover, with the Assumption 4.3, it holds that

$$\|\rho_t - \mathcal{P}_t R_\Gamma(\rho_0)\|_{L^2(\mathbb{R}^\infty, \mu_{C_t})} \lesssim 2^{-\left(1 - v/a^\dagger\right)\Gamma} \|\rho_0\|_{\mathcal{F}^\gamma_{p,q}}.$$

Proof. By Lemma A.8, we have

$$\|\mathcal{P}_t(\rho_0 - R_\Gamma(\rho_0))\|_{L^2(\mathbb{R}^\infty,\mu_{C_t})} \le \|\rho_0 - R_\Gamma(\rho_0)\|_{L^2(\mathbb{R}^\infty,\mu_{C_0})}.$$

And from the proof of Theorem 7 of Okumoto & Suzuki (2022), we can get

$$\|\rho_0 - R_\Gamma(\rho_0)\|_{L^2(\mathbb{R}^\infty,\mu_{C_0})} \le \begin{cases} 2^{-(1-v\alpha)\Gamma}\|\rho_0\|_{\mathcal{F}^\gamma_{p,q}} & (1 \le q \le 2) \\ 2^{-(1-v\alpha')\Gamma}\left[\sum_{s\in\mathbb{N}_0^\infty:\Gamma\le\gamma'(s)} 2^{\frac{2q}{q-2}(\gamma'(s)-\gamma(s))}\right]^{1/2-1/q}\|\rho_0\|_{\mathcal{F}^\gamma_{p,q}} & (q > 2). \end{cases}$$

Furthermore, from the Theorem 9 of Okumoto & Suzuki (2022), we can get the bound for all $q \ge 1$.

$$\|\rho_0 - R_\Gamma(\rho_0)\|_{L^2(\mathbb{R}^\infty,\mu_{C_0})} \lesssim 2^{-\left(1-v/a^\dagger\right)\Gamma}\|\rho_0\|_{\mathcal{F}^\gamma_{p,q}}.$$

□

E.2 The error bound of (II)

In this section, we prove the below lemma.

Lemma E.3 (Approximation of $\mathcal{P}_t R_\Gamma(\rho_0)$ with a neural network)**.** *Under the assumptions and notation of Lemma E.1, choose ε as in (9) and use K_ρ defined there. Then there exists $\phi_\rho \in \Phi(L, W, S, B; d_{\max}+1, 1)$ such that*

$$\|\mathcal{P}_t R_\Gamma(\rho_0) - \phi_\rho(\cdot, t)\|_{L^2(\mu_{C_t})} \le 2^{-\Gamma}\|\rho_0\|_{\mathcal{F}^\gamma_{p,q}}, \qquad t \in [\underline{T}, \overline{T}],$$

with budgets

$$\begin{aligned} L &= \mathcal{O}(K_\rho^2), \\ W &= \mathcal{O}(d_{\max} G K_\rho^3), \\ S &= \mathcal{O}(d_{\max} G K_\rho^4), \\ B &\le C \max\left\{\max_{i\le d_{\max}} c_{0,i}^{-1/2}, \exp(CK_\rho)\right\}. \end{aligned}$$

If the retained expansion contains only the constant basis function, it is represented exactly by a constant network.

We explain in the case of $1 \le q \le 2$, and the case $q > 2$ is also proved in the same way. $\mathcal{P}_t R_\Gamma(\rho_0)(x)$ only has the basis that $\gamma(s) < \Gamma$ and then $l_{d_{\max}+1} = l_{d_{\max}+2} = \cdots = 0$, so this is independent from the input $(x_{d_{\max}+1}, x_{d_{\max}+2}, ...)$. Here we abbreviate $d_{\max}$ as d. By using Lemma A.6, $\mathcal{P}_t R_\Gamma(\rho_0)(x)$ is expanded as follows:

$$\begin{aligned} &\mathcal{P}_t R_\Gamma(\rho_0) \\ &= \sum_{\boldsymbol{s}:\gamma(\boldsymbol{s})<\Gamma}\sum_{\boldsymbol{l}\in J(\boldsymbol{s})}\langle\rho_0, \psi_{\boldsymbol{l}}\rangle \int_{\mathbb{R}^d} \underbrace{\psi_{\boldsymbol{l}}(\widetilde{m}_t x + \widetilde{\sigma}_t y)}_{d \text{ times product}} \mathrm{d}\mu_{C_0}(y) \\ &= \sum_{\boldsymbol{s}:\gamma(\boldsymbol{s})<\Gamma}\sum_{\boldsymbol{l}\in J(\boldsymbol{s})}\langle\rho_0, \psi_{\boldsymbol{l}}\rangle \prod_{i=1}^d \int_{\mathbb{R}} \psi_{l_i}(\widetilde{m}_{t,i} x_i + \widetilde{\sigma}_{t,i} y_i)\mathrm{d}\mu_{C_0}(y_i) \\ &= \sum_{\boldsymbol{s}:\gamma(\boldsymbol{s})<\Gamma}\sum_{\boldsymbol{l}\in J(\boldsymbol{s})}\langle\rho_0, \psi_{\boldsymbol{l}}\rangle \prod_{i=1}^d \int_{\mathbb{R}} \left(\cos 2\pi l_i \Phi\left(\frac{\widetilde{m}_{t,i} x_i + \widetilde{\sigma}_{t,i} y_i}{\sqrt{c_{0,i}}}\right) - \sin 2\pi l_i \Phi\left(\frac{\widetilde{m}_{t,i} x_i + \widetilde{\sigma}_{t,i} y_i}{\sqrt{c_{0,i}}}\right)\right) \frac{e^{-\frac{y_i^2}{2c_{0,i}}}}{\sqrt{2\pi c_{0,i}}}\mathrm{d}y_i, \end{aligned}$$

where $\langle f, g\rangle := \int_{\mathbb{R}^\infty} f(x)g(x)\mathrm{d}\mu_{C_0}$ denotes the inner product on μ_{C_0}.

We define $\mathcal{I}_{l_i}$ as

$$\mathcal{I}_{l_i}(x_i) := \int_{\mathbb{R}} \left(\cos 2\pi l_i \Phi\left(\frac{\widetilde{m}_{t,i} x_i + \widetilde{\sigma}_{t,i} y_i}{\sqrt{c_{0,i}}}\right) - \sin 2\pi l_i \Phi\left(\frac{\widetilde{m}_{t,i} x_i + \widetilde{\sigma}_{t,i} y_i}{\sqrt{c_{0,i}}}\right)\right) \frac{e^{-\frac{y_i^2}{2c_{0,i}}}}{\sqrt{2\pi c_{0,i}}}\mathrm{d}y_i, \tag{7}$$

and the error bound of (II) is derived from the approximation error of $\mathcal{I}_{l_i}$ with a neural network.

Lemma E.4 (ReLU approximation of $\mathcal{I}_l$)**.** *Assume covariance comparability with a fixed constant $K \geq 1$, and let $0 < c_0, c \leq 1$. Use the noise-schedule assumptions in Section 2, under which Lemma C.8 supplies the time approximations.*

Let $0 < \underline{T} < \overline{T} < \infty$ and $0 < \varepsilon < 1/2$. For $l \in \mathbb{Z} \setminus \{0\}$, define

$$\Lambda_l := 1 + \left(|l| + \log \varepsilon^{-1}\right) \log \frac{e + |l|}{\varepsilon} \log \left(e + \log \frac{e + |l|}{\varepsilon}\right) + \overline{\beta T} + \log(1 + \underline{T}^{-1}). \tag{8}$$

Then there exists a ReLU network

$$\phi_{\mathcal{I}_l} \in \Phi(L_{\mathcal{I}_l}, W_{\mathcal{I}_l}, S_{\mathcal{I}_l}, B_{\mathcal{I}_l}; 2, 1)$$

such that

$$\sup_{\substack{x \in \mathbb{R} \\ t \in [\underline{T}, \overline{T}]}} |\phi_{\mathcal{I}_l}(x,t) - \mathcal{I}_l(x,t)| \leq \varepsilon, \qquad \sup_{x,t} |\phi_{\mathcal{I}_l}(x,t)| \leq \sqrt{2}.$$

Sufficient network budgets are

$$\begin{aligned} L_{\mathcal{I}_l} &= \mathcal{O}(\Lambda_l^2), \\ W_{\mathcal{I}_l} &= \mathcal{O}(\Lambda_l^3), \\ S_{\mathcal{I}_l} &= \mathcal{O}(\Lambda_l^4), \\ B_{\mathcal{I}_l} &\leq C \max\left\{c_0^{-1/2}, \exp(C\Lambda_l)\right\}. \end{aligned}$$

The constants may depend on K, $\underline{\beta}$, $\overline{\beta}$, and the fixed constants in the arithmetic approximation lemmas, but not on $l, \varepsilon, c_0, c, \underline{T}$, or $\overline{T}$.

For $l = 0$, the factor $\mathcal{I}_0 = 1$ is represented exactly by a constant network.

The proof is given in Appendix F. The quantity $|l|$ in these budgets is the actual frequency index, not its dyadic frequency level.

Under Lemma E.4, each factor $\mathcal{I}_{l_i}$ admits a neural network approximation $\phi_{\mathcal{I}_{l_i}}$ satisfying

$$|\mathcal{I}_{l_i} - \phi_{\mathcal{I}_{l_i}}| \leq \varepsilon, \qquad |\phi_{\mathcal{I}_{l_i}}| \leq \sqrt{2}.$$

Here, $d = d_{\max}(\Gamma, \gamma)$, and all uniform approximation estimates are on $t \in [\underline{T}, \overline{T}]$.

For $d \geq 2$, Lemma C.5, with input radius $\sqrt{2}$ and approximation tolerance ε, gives a network $\phi_{\text{mult}} \in \Phi(L_{\text{mult}}, W_{\text{mult}}, S_{\text{mult}}, B_{\text{mult}})$ such that

$$\begin{aligned} &\left| \prod_{i=1}^{d} \mathcal{I}_{l_i} - \phi_{\text{mult}}\left(\phi_{\mathcal{I}_{l_1}}, \ldots, \phi_{\mathcal{I}_{l_d}}\right) \right| \\ &\quad \leq \varepsilon + d(\sqrt{2})^{d-1}\varepsilon \leq 2d(\sqrt{2})^{d-1}\varepsilon. \end{aligned}$$

Sufficient budgets are

$$L_{\text{mult}} = \mathcal{O}\left(\log(ed)(\log \varepsilon^{-1} + d)\right), \qquad W_{\text{mult}} = \mathcal{O}(d),$$

$$S_{\text{mult}} = \mathcal{O}\left(d(\log \varepsilon^{-1} + d)\right), \qquad B_{\text{mult}} = (\sqrt{2})^d.$$

For $d = 1$, use the identity map in place of ϕ_{mult}; the same error upper bound applies.

Define

$$\phi_\rho(x,t) := \sum_{s:\gamma(s)<\Gamma} \sum_{l \in J(s)} \langle \rho_0, \psi_l \rangle \phi_{\text{mult}}\left(\phi_{\mathcal{I}_{l_1}}(x_1,t), \ldots, \phi_{\mathcal{I}_{l_d}}(x_d,t)\right).$$

Then

$$\begin{aligned} \underbrace{\|\mathcal{P}_t R_\Gamma(\rho_0) - \phi_\rho(\cdot,t)\|_{L^2(\mu_{C_t})}}_{\text{(II)}} &\leq 2d(\sqrt{2})^{d-1} G(\Gamma,\gamma) \|\rho_0\|_{L^2(\mu_{C_0})} \varepsilon \\ &\leq 2d(\sqrt{2})^{d-1} G(\Gamma,\gamma) B_\rho \varepsilon, \end{aligned}$$

where $|\langle \rho_0, \psi_l \rangle| \le \|\rho_0\|_{L^2(\mu_{C_0})} \le B_\rho$, and $B_\rho \ge 1$ is fixed. Choose

$$\varepsilon := \min\left\{ \frac{1}{4}, \frac{2^{-\Gamma}\|\rho_0\|_{\mathcal{F}^\gamma_{p,q}}}{2d(\sqrt{2})^{d-1}G(\Gamma,\gamma)B_\rho} \right\}. \tag{9}$$

Then, from (6) and the initial truncation estimate, we obtain

$$\begin{aligned}
\|\rho_t - \phi_\rho(\cdot,t)\|_{L^2(\mu_{C_t})} &\le \underbrace{\|\rho_t - \mathcal{P}_t R_\Gamma(\rho_0)\|_{L^2(\mu_{C_t})}}_{\text{(I)}} + \underbrace{\|\mathcal{P}_t R_\Gamma(\rho_0) - \phi_\rho(\cdot,t)\|_{L^2(\mu_{C_t})}}_{\text{(II)}} \\
&\lesssim \left(2^{-(1-v/a^\dagger)\Gamma} + 2^{-\Gamma}\right)\|\rho_0\|_{\mathcal{F}^\gamma_{p,q}} \\
&\lesssim 2^{-(1-v/a^\dagger)\Gamma}\|\rho_0\|_{\mathcal{F}^\gamma_{p,q}}.
\end{aligned}$$

Finally, we evaluate the network size. Let $\Lambda_l(\varepsilon)$ be the quantity defined in (8), and set

$$K_\rho := 1 + d_{\max} + \log \varepsilon^{-1} + \log B_\rho + \max_{\substack{l\in\mathbb{Z} \\ 1\le|l|\le 2^{f_{\max}}}} \Lambda_l(\varepsilon).$$

The actual retained frequencies satisfy $|l_i| < 2^{f_{\max}}$. Lemma E.4 therefore gives uniform scalar-factor budgets

$$L_{\mathcal{I}_l} = \mathcal{O}(K_\rho^2), \qquad W_{\mathcal{I}_l} = \mathcal{O}(K_\rho^3), \qquad S_{\mathcal{I}_l} = \mathcal{O}(K_\rho^4),$$

and

$$B_{\mathcal{I}_l} \le C \max\left\{ \max_{i\le d_{\max}} c_{0,i}^{-1/2}, \exp(CK_\rho) \right\}.$$

Since ϕ_ρ is a linear combination of the product networks, using independent factor-network copies for each retained basis gives sufficient budgets

$$\begin{aligned}
L &\le \max_l L_{\mathcal{I}_l} + L_{\text{mult}} + \mathcal{O}(1) = \mathcal{O}(K_\rho^2), \\
W &\lesssim d_{\max}G(\Gamma,\gamma)\max_l W_{\mathcal{I}_l} + G(\Gamma,\gamma)W_{\text{mult}} \\
&= \mathcal{O}\left(d_{\max}G(\Gamma,\gamma)K_\rho^3\right), \\
S &\lesssim d_{\max}G(\Gamma,\gamma)\left(\max_l S_{\mathcal{I}_l} + \max_l L_{\mathcal{I}_l}\right) + G(\Gamma,\gamma)S_{\text{mult}} + G(\Gamma,\gamma) \\
&= \mathcal{O}\left(d_{\max}G(\Gamma,\gamma)K_\rho^4\right), \\
B &\lesssim \max_l B_{\mathcal{I}_l} \vee B_{\text{mult}} \vee B_\rho \vee 1 \\
&\le C \max\left\{ \max_{i\le d_{\max}} c_{0,i}^{-1/2}, \exp(CK_\rho) \right\}.
\end{aligned} \tag{10}$$

The additional depth term in the sparsity estimate accounts for fixed identity layers aligning the scalar-factor outputs. Explicit signed ReLU interfaces are retained when composing subnetworks, so adjacent affine weight matrices need not be multiplied. Thus

$$\phi_\rho \in \Phi(L, W, S, B; d_{\max} + 1, 1).$$

F ReLU APPROXIMATION OF THE NORMALIZED ONE-DIMENSIONAL FACTOR

In this section, we omit the coordinate index i. We consider

$$\mathcal{I}_l(x,t) := \int_{\mathbb{R}} g_l\left(\Phi\left(\frac{\widetilde{m}_t x + \widetilde{\sigma}_t y}{\sqrt{c_0}}\right)\right)\varphi_{c_0}(y)\,\mathrm{d}y, \qquad g_l(u) := \cos(2\pi l u) - \sin(2\pi l u),$$

where Φ is the standard Gaussian distribution function and φ_{c_0} is the centered Gaussian density with variance c_0. Recall that

$$\begin{aligned}
&\widetilde{m}_t = \frac{m_t c_0}{c_t}, \qquad \widetilde{\sigma}_t = \sigma_t\sqrt{\frac{c}{c_t}}, \\
&c_t = m_t^2 c_0 + \sigma_t^2 c, \qquad m_t = \exp\left(-\int_0^t \beta_s\,\mathrm{d}s\right), \qquad \sigma_t^2 = 1 - m_t^2.
\end{aligned}$$

We work on $t \in [\underline{T}, \overline{T}]$, where $0 < \underline{T} < \overline{T} < \infty$. Let $K \geq 1$ be a fixed covariance comparison constant:

$$K^{-1} \leq c_0/c \leq K.$$

Then

$$K^{-1} \leq c_t/c \leq K, \qquad 0 < \widetilde{m}_t \leq K, \qquad 0 \leq \widetilde{\sigma}_t \leq 1.$$

If $l = 0$, then $\mathcal{I}_0(x,t) = 1$ and no approximation is needed. Henceforth, let $l \in \mathbb{Z} \setminus \{0\}$ and $0 < \varepsilon < 1/2$. We first construct an inner polynomial approximation, then implement it by ReLU networks, and finally join it continuously to the outer approximation.

F.1 Outer-domain approximation

Define

$$H_l := \sqrt{8 \log\left(\frac{3\sqrt{2\pi}|l|}{\varepsilon}\right)},$$

and

$$B_x(t) := \frac{\sqrt{c_0}}{\widetilde{m}_t} H_l.$$

Lemma F.1 (Outer-domain approximation)**.** *For every $t \in [\underline{T}, \overline{T}]$,*

$$\sup_{|x| \geq B_x(t)} |\mathcal{I}_l(x,t) - 1| \leq \varepsilon.$$

The constant function 1 is represented exactly by a ReLU network with constant depth, width, sparsity, and weight bound.

Proof. Since l is an integer, $g_l(0) = g_l(1) = 1$, and

$$\|g_l'\|_\infty \leq 2\sqrt{2}\pi|l|.$$

Therefore, for $u \in [0,1]$,

$$|g_l(u) - 1| \leq 2\sqrt{2}\pi|l| \min\{u, 1-u\}.$$

Set

$$X := \frac{\widetilde{m}_t x}{\sqrt{c_0}}, \qquad Z_0 \sim \mathcal{N}(0,1).$$

Then

$$|\mathcal{I}_l(x,t) - 1| \leq 2\sqrt{2}\pi|l| \, \mathbb{E}\Phi(-|X + \widetilde{\sigma}_t Z_0|).$$

On $\{|Z_0| \leq |X|/2\}$, the condition $\widetilde{\sigma}_t \leq 1$ gives $|X + \widetilde{\sigma}_t Z_0| \geq |X|/2$. Hence

$$\mathbb{E}\Phi(-|X + \widetilde{\sigma}_t Z_0|) \leq 3\Phi(-|X|/2).$$

By Mills' inequality,

$$|\mathcal{I}_l(x,t) - 1| \leq \frac{12\sqrt{\pi}|l|}{|X|} e^{-|X|^2/8}.$$

For $|X| \geq H_l$, the right-hand side is at most

$$\frac{\varepsilon}{\sqrt{\log(3\sqrt{2\pi}|l|/\varepsilon)}} \leq \varepsilon.$$

This proves the assertion. □

F.2 Polynomial approximation on an enlarged inner domain

For continuous joining, the inner approximation is constructed on the enlarged domain

$$\left|\frac{\widetilde{m}_t x}{\sqrt{c_0}}\right| \leq H_l + 2, \qquad t \in [\underline{T}, \overline{T}].$$

The overlap with the outer domain will be used below.

F.2.1 Step 1: Truncation of the Gaussian integral

Choose

$$A := \sqrt{2c_0 \log(24/\varepsilon)}, \tag{11}$$

and define

$$\mathcal{I}_l^{(A)}(x,t) := \int_{-A}^{A} g_l\left(\Phi\left(\frac{\widetilde{m}_t x + \widetilde{\sigma}_t y}{\sqrt{c_0}}\right)\right) \varphi_{c_0}(y)\,\mathrm{d}y.$$

Lemma F.2 (Gaussian truncation error)**.** *For every $A > 0$,*

$$\sup_{x,t} |\mathcal{I}_l(x,t) - \mathcal{I}_l^{(A)}(x,t)| \le \frac{2\sqrt{c_0}}{\sqrt{\pi}A} e^{-A^2/(2c_0)}. \tag{12}$$

With the choice (11)*,*

$$\sup_{x,t} |\mathcal{I}_l(x,t) - \mathcal{I}_l^{(A)}(x,t)| < \varepsilon/3.$$

Proof. The bound $\|g_l\|_\infty \le \sqrt{2}$ gives

$$|\mathcal{I}_l - \mathcal{I}_l^{(A)}| \le \sqrt{2}\,\mathbb{P}\left(|Z_0| > A/\sqrt{c_0}\right), \qquad Z_0 \sim \mathcal{N}(0,1).$$

Mills' inequality yields (12). Alternatively, the Gaussian tail bound gives

$$|\mathcal{I}_l - \mathcal{I}_l^{(A)}| \le 2\sqrt{2} e^{-A^2/(2c_0)} = \frac{\sqrt{2}}{12}\varepsilon < \varepsilon/3.$$

□

Set

$$Z := H_l + 2 + \frac{A}{\sqrt{c_0}}.$$

On the enlarged inner domain and for $|y| \le A$,

$$\left|\frac{\widetilde{m}_t x + \widetilde{\sigma}_t y}{\sqrt{c_0}}\right| \le Z.$$

In particular,

$$Z^2 = \mathcal{O}\left(\log \frac{e + |l|}{\varepsilon}\right).$$

F.2.2 Step 2: Polynomialization of the integrand

Write

$$z(x,y,t) := \frac{\widetilde{m}_t x + \widetilde{\sigma}_t y}{\sqrt{c_0}}.$$

Approximation of the Gaussian distribution function. For an integer $R \ge 1$, define

$$p_R(z) := \sum_{r=0}^{R-1} \frac{(-1)^r}{2^r r!} z^{2r},$$

$$\Phi_R(z) := \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_0^z p_R(w)\,\mathrm{d}w.$$

Then Φ_R has degree $2R-1$.

Lemma F.3 (Taylor remainder for Φ)**.** *For $Z > 0$ and $R \ge 1$,*

$$\sup_{|z| \le Z} |\Phi(z) - \Phi_R(z)| \le \frac{Z}{2\pi\sqrt{R}} \left(\frac{eZ^2}{2R}\right)^R =: \delta_\Phi(R; Z).$$

Proof. Taylor's theorem for e^{-u} on $u \geq 0$ gives

$$\left| e^{-z^2/2} - \sum_{r=0}^{R-1} \frac{(-1)^r}{r!} \left(\frac{z^2}{2}\right)^r \right| \leq \frac{(z^2/2)^R}{R!}.$$

Integrating from 0 to z yields

$$|\Phi(z) - \Phi_R(z)| \leq \frac{Z}{\sqrt{2\pi}} \frac{(Z^2/2)^R}{R!}.$$

The stated bound follows from Stirling's inequality. □

Approximation of the trigonometric factor. For $N \geq 1$, define

$$\cos_N(w) := \sum_{n=0}^{N-1} \frac{(-1)^n}{(2n)!} w^{2n}, \qquad \sin_N(w) := \sum_{n=0}^{N-1} \frac{(-1)^n}{(2n+1)!} w^{2n+1}.$$

Since $l \in \mathbb{Z}$,

$$g_l(u) = (-1)^l \left[\cos\left(2\pi l(u - \tfrac{1}{2})\right) - \sin\left(2\pi l(u - \tfrac{1}{2})\right)\right].$$

Define

$$g_{l,N}^{(\delta)}(u) := (-1)^l \left[\cos_N\left(2\pi l(u - \tfrac{1}{2})\right) - \sin_N\left(2\pi l(u - \tfrac{1}{2})\right)\right].$$

The superscript δ specifies the approximation interval; the polynomial formula does not depend on it.

Lemma F.4 (Taylor remainder for g_l)**.** *Let $\delta \geq 0$ and $\Lambda = 2\pi|l|(1/2 + \delta)$. Then*

$$\sup_{u \in [-\delta, 1+\delta]} |g_l(u) - g_{l,N}^{(\delta)}(u)| \leq \frac{\Lambda^{2N}}{(2N)!} + \frac{\Lambda^{2N+1}}{(2N+1)!}.$$

If $2N + 1 \geq \Lambda$, this is at most

$$\frac{1}{\sqrt{\pi N}} \left(\frac{e\pi|l|(1+2\delta)}{2N}\right)^{2N} =: \delta_g(N; l, \delta).$$

Proof. Apply the Taylor remainder formulas for sine and cosine on $[-\Lambda, \Lambda]$. If $2N + 1 \geq \Lambda$, the second remainder is at most the first. Stirling's inequality gives the final bound. □

Approximation of the Gaussian weight. Define

$$q_M(y) := \frac{1}{\sqrt{2\pi c_0}} \sum_{m=0}^{M-1} \frac{(-1)^m}{m!} \left(\frac{y^2}{2c_0}\right)^m.$$

Lemma F.5 (Taylor remainder for the Gaussian weight)**.** *For $A > 0$ and $M \geq 1$,*

$$\sup_{|y| \leq A} |\varphi_{c_0}(y) - q_M(y)| \leq \frac{1}{2\pi\sqrt{c_0 M}} \left(\frac{eA^2}{2c_0 M}\right)^M =: \delta_\varphi(M; A).$$

Proof. Taylor's theorem applied to e^{-u}, with $u = y^2/(2c_0)$, bounds the error by

$$\frac{1}{\sqrt{2\pi c_0}} \frac{(A^2/(2c_0))^M}{M!}.$$

Apply Stirling's inequality. □

Polynomialization error and degree selection. Set $\delta_\Phi = \delta_\Phi(R; Z)$ and define

$$\mathcal{I}_l^{(A),\mathrm{poly}}(x, t) := \int_{-A}^{A} g_{l,N}^{(\delta_\Phi)}\big(\Phi_R(z(x, y, t))\big) q_M(y)\, \mathrm{d}y.$$

Lemma F.6 (Polynomialization error)**.** *There is a universal constant $C_* > 0$ such that the choices*

$$R := \left\lceil C_* \log \frac{e + |l|}{\varepsilon} \right\rceil,$$
$$N := \left\lceil C_* \left(|l| + \log \varepsilon^{-1} \right) \right\rceil,$$
$$M := \left\lceil C_* \log \varepsilon^{-1} \right\rceil$$

give

$$\sup_{\substack{t \in [\underline{T}, \overline{T}] \\ |\widetilde{m}_t x / \sqrt{c_0}| \le H_l + 2}} |\mathcal{I}_l^{(A)}(x,t) - \mathcal{I}_l^{(A),\mathrm{poly}}(x,t)| \le \varepsilon/3.$$

Proof. Choose C_* large enough that $\delta_\Phi \le 1/2$, $2N + 1 \ge 2\pi|l|(1/2 + \delta_\Phi)$, and $\delta_g \le \sqrt{2}$. Since $\Phi_R(z) \in [-\delta_\Phi, 1 + \delta_\Phi]$, adding and subtracting intermediate terms yields

$$|\mathcal{I}_l^{(A)} - \mathcal{I}_l^{(A),\mathrm{poly}}| \le 2\sqrt{2}\pi|l|\delta_\Phi(R;Z) + \delta_g(N;l,\delta_\Phi) + 4\sqrt{2}A\,\delta_\varphi(M;A).$$

For the last term, we used

$$|g_{l,N}^{(\delta_\Phi)}(\Phi_R(z))| \le \sqrt{2} + \delta_g \le 2\sqrt{2}.$$

The bound on Z allows us to choose $R \ge e^2 Z^2/2$. Then $|l|\delta_\Phi(R;Z) \lesssim |l|e^{-R}$. Similarly, $N \ge e^2\pi|l|$ gives $\delta_g(N;l,\delta_\Phi) \lesssim e^{-2N}$. Finally,

$$\frac{A^2}{2c_0} = \log(24/\varepsilon),$$

so the stated choice of M gives

$$A\delta_\varphi(M;A) \lesssim \frac{A}{\sqrt{c_0 M}} e^{-M} \lesssim e^{-M}.$$

Increasing the universal constant C_* makes each of the three terms at most $\varepsilon/9$. □

Analytic integration and polynomial coefficients. Introduce the exact normalized variables

$$X(x,t) := \frac{\widetilde{m}_t x}{\sqrt{c_0}}, \qquad Y(t) := \frac{\widetilde{\sigma}_t A}{\sqrt{c_0}}.$$

Changing variables $y = Au$ gives

$$\mathcal{I}_l^{(A),\mathrm{poly}}(x,t) = P(X(x,t), Y(t)),$$
$$P(X,Y) := \frac{A}{\sqrt{2\pi c_0}} \int_{-1}^{1} g_{l,N}^{(\delta_\Phi)}\big(\Phi_R(X + Yu)\big) \times \sum_{m=0}^{M-1} \frac{(-1)^m}{m!} \left(\frac{A^2}{2c_0}\right)^m u^{2m}\,\mathrm{d}u.$$

The integrand is a polynomial in u of degree at most

$$K_y := (2N-1)(2R-1) + 2(M-1).$$

Its degree in (X,Y) is at most

$$K_x := (2N-1)(2R-1).$$

Termwise integration is exact, using

$$\int_{-1}^{1} u^j\,\mathrm{d}u = \begin{cases} 0, & j \text{ odd}, \\ 2/(j+1), & j \text{ even}. \end{cases}$$

Thus

$$P(X,Y) = \sum_{\substack{k,j \ge 0 \\ k+j \le K_x}} a_{k,j} X^k Y^j,$$

with at most $(K_x+1)(K_x+2)/2$ possible monomials. The degree satisfies

$$K_x = \mathcal{O}\left(\left(|l| + \log \varepsilon^{-1}\right) \log \frac{e+|l|}{\varepsilon}\right).$$

For a polynomial Q, denote its coefficient absolute sum by $\|Q\|_{\mathrm{coef},1}$. This quantity is submultiplicative. Write

$$\Phi_R(z) - \frac{1}{2} = \sum_{r=0}^{R-1} \alpha_r z^{2r+1}, \qquad \alpha_r = \frac{(-1)^r}{\sqrt{2\pi}\, 2^r r!(2r+1)}.$$

Since $\|X + Yu\|_{\mathrm{coef},1} = 2$,

$$\begin{aligned}\left\|\Phi_R(X+Yu) - \frac{1}{2}\right\|_{\mathrm{coef},1} &\le \sum_{r=0}^{R-1} |\alpha_r| 2^{2r+1} \\ &\le \frac{2}{\sqrt{2\pi}} \sum_{r=0}^{\infty} \frac{2^r}{r!} = \frac{2e^2}{\sqrt{2\pi}}.\end{aligned}$$

Retaining the factorial denominators in the trigonometric Taylor polynomial therefore gives

$$\left\| g_{l,N}^{(\delta_\Phi)}(\Phi_R(X+Yu)) \right\|_{\mathrm{coef},1} \le e^{C|l|}.$$

The Gaussian-weight polynomial has coefficient absolute sum at most $e^{A^2/(2c_0)}$. Since integration in u multiplies each coefficient by a number of absolute value at most 2,

$$\sum_{k,j} |a_{k,j}| \le C \frac{A}{\sqrt{c_0}} \exp\left(C|l| + \frac{A^2}{2c_0}\right).$$

Consequently,

$$\log\left(1 + \sum_{k,j} |a_{k,j}|\right) = \mathcal{O}\left(|l| + \log \varepsilon^{-1}\right).$$

F.2.3 Network construction for polynomial approximation

We implement P on the enlarged rectangle $[-Z-1, Z+1]^2$. Define

$$b_{k,j} := a_{k,j}(Z+1)^{k+j}.$$

Then

$$P(X,Y) = \sum_{k+j \le K_x} b_{k,j} \left(\frac{X}{Z+1}\right)^k \left(\frac{Y}{Z+1}\right)^j,$$

and

$$\begin{aligned}\log\left(1 + \sum_{k,j} |b_{k,j}|\right) &\le K_x \log(Z+1) + \log\left(1 + \sum_{k,j} |a_{k,j}|\right) \\ &= \mathcal{O}\left(K_x \log(Z+1) + |l| + \log \varepsilon^{-1}\right). \end{aligned} \tag{13}$$

On this rectangle,

$$|\partial_X P| + |\partial_Y P| \le \frac{K_x}{Z+1} \sum_{k,j} |b_{k,j}|.$$

Choose the input tolerance

$$\varepsilon_{\mathrm{input}} := \min\left\{\frac{1}{4}, \frac{\varepsilon(Z+1)}{9K_x(1 + \sum_{k,j} |b_{k,j}|)}\right\}.$$

For input pairs in the implementation rectangle, perturbing each of the two inputs by at most $\varepsilon_{\mathrm{input}}$ changes P by at most $\varepsilon/9$.

Subnet 1: Approximation of the time-dependent coefficients. Set

$$B_x^* := \sup_{t\in[\underline{T},\overline{T}]} \frac{\sqrt{c_0}}{\widetilde{m}_t}(H_l+2).$$

The input subnetworks are constructed on $[-B_x^*, B_x^*] \times [\underline{T}, \overline{T}]$. Choose

$$\varepsilon_1 := \frac{\varepsilon_{\text{input}}}{2\max\{1, B_x^*/\sqrt{c_0}, A/\sqrt{c_0}\}}.$$

Recall

$$\widetilde{m}_t = \frac{(c_0/c)m_t}{1+(c_0/c-1)m_t^2}, \qquad \widetilde{\sigma}_t = \sigma_t\left(1+(c_0/c-1)m_t^2\right)^{-1/2}.$$

Approximate m_t and σ_t to accuracy $c_K\varepsilon_1$, where $c_K > 0$ is sufficiently small and depends only on K, and clip their approximations to $[0,1]$.

Use a multiplication network to approximate m_t^2. Form the denominator approximation and clip it to $[K^{-1}, K]$ before applying a reciprocal network. Clip the reciprocal output to $[K^{-1}, K]$. Apply a square-root network on that same interval and multiplication networks to obtain the two coefficient approximations.

All arithmetic errors are chosen at most $c_K\varepsilon_1$. The reciprocal and square-root maps are Lipschitz on these fixed intervals. Final clipping to $[0, K]$ and $[0, 1]$, respectively, therefore gives

$$\sup_{t\in[\underline{T},\overline{T}]} |\phi_{\widetilde{m}}(t) - \widetilde{m}_t| \le \varepsilon_1,$$

and

$$\sup_{t\in[\underline{T},\overline{T}]} |\phi_{\widetilde{\sigma}}(t) - \widetilde{\sigma}_t| \le \varepsilon_1.$$

We now evaluate the budgets of the joint two-output coefficient network using Lemma C.8 and Lemmas C.5, C.6, and C.7. All implicit constants in this calculation may depend on the fixed covariance comparison constant K.

Choose the time-network accuracy

$$\epsilon_{\text{time}} := \min\{c_K\varepsilon_1, \underline{T}, 1/4\}.$$

Then the approximation guarantee for σ_t applies throughout $[\underline{T}, \overline{T}]$, and both time-coefficient errors are at most $c_K\varepsilon_1$. Moreover,

$$\log \epsilon_{\text{time}}^{-1} \lesssim_K 1 + \log \varepsilon_1^{-1} + \log(1+\underline{T}^{-1}).$$

For the reciprocal and square-root operations, all inputs are clipped to the fixed interval $[K^{-1}, K]$. The reciprocal lemma can be applied with a fixed interval parameter greater than $\max\{K, 1\}$. For the square-root approximation, choose its tolerance ϵ_{root} at most $\min\{c_K\varepsilon_1, (2K)^{-1}, 1/4\}$. Then $[K^{-1}, K]$ is contained in its approximation interval.

If $z, z' \in [K^{-1}, K]$, the exact-input guarantee of Lemma C.7 and the Lipschitz continuity of the square root on this interval give

$$\begin{aligned}|\phi_{\text{root}}(z') - \sqrt{z}| &\le |\phi_{\text{root}}(z') - \sqrt{z'}| + |\sqrt{z'} - \sqrt{z}| \\ &\le \epsilon_{\text{root}} + \frac{\sqrt{K}}{2}|z'-z|.\end{aligned}$$

Thus the preceding approximation errors need only be controlled to a fixed K-dependent multiple of ε_1. There is no need to use the larger $\epsilon_{\text{root}}^{-1/2}$ perturbation coefficient on this fixed interval.

The construction uses a fixed number of reciprocal, square-root, and two-input multiplication networks. Their input ranges are bounded by constants depending only on K. Evaluating the arithmetic lemmas at tolerances of order ε_1 gives sufficient arithmetic budgets

$$L = \mathcal{O}\big((1+\log\varepsilon_1^{-1})^2\big), \qquad W = \mathcal{O}\big((1+\log\varepsilon_1^{-1})^3\big),$$

$$S = \mathcal{O}\big((1+\log \varepsilon_1^{-1})^4\big), \qquad B = \mathcal{O}(\varepsilon_1^{-2}).$$

Combining the time networks with these arithmetic networks, sufficient budgets for Subnet 1 are

$$\begin{aligned} L_1 &= \mathcal{O}\left(\left[1+\log \varepsilon_1^{-1} + \log(1+\underline{T}^{-1})\right]^2\right), \\ W_1 &= \mathcal{O}\left(\left[1+\log \varepsilon_1^{-1} + \log(1+\underline{T}^{-1})\right]^3\right), \\ S_1 &= \mathcal{O}\left(\left[1+\log \varepsilon_1^{-1} + \log(1+\underline{T}^{-1})\right]^4\right), \\ B_1 &= \mathcal{O}\left(\varepsilon_1^{-2} + 1 + \log(1+\underline{T}^{-1})\right). \end{aligned} \tag{14}$$

These budgets include the fixed clipping operations and the identity channels needed to preserve the finitely many intermediate scalar outputs. Subnetworks can be composed through signed ReLU interfaces, rather than by merging adjacent affine matrices, so the displayed weight bound is preserved up to a fixed multiplicative constant.

Subnet 2: Construction of the normalized inputs. For $|x| \le B_x^*$, both x/B_x^* and $\phi_{\widetilde{m}}(t)/K$ belong to $[-1,1]$. Choose

$$\varepsilon_{\mathrm{mult}} := \frac{\varepsilon_{\mathrm{input}}}{2\max\{1, KB_x^*/\sqrt{c_0}\}}.$$

Apply Lemma C.5 with two inputs, input radius 1, and tolerance $\varepsilon_{\mathrm{mult}}$. Only its exact-input approximation guarantee is used. Define the raw normalized-input approximation by

$$\phi_X(x,t) := \frac{KB_x^*}{\sqrt{c_0}}\phi_{\mathrm{mult}}\left(\frac{x}{B_x^*}, \frac{\phi_{\widetilde{m}}(t)}{K}\right),$$

and

$$\phi_Y(t) := \frac{A}{\sqrt{c_0}}\phi_{\widetilde{\sigma}}(t).$$

Then, throughout the input box,

$$\begin{aligned} |\phi_X(x,t) - X(x,t)| &\le \frac{KB_x^*}{\sqrt{c_0}}\varepsilon_{\mathrm{mult}} + \frac{B_x^*}{\sqrt{c_0}}\varepsilon_1 \le \varepsilon_{\mathrm{input}}, \\ |\phi_Y(t) - Y(t)| &\le \frac{A}{\sqrt{c_0}}\varepsilon_1 \le \varepsilon_{\mathrm{input}}/2. \end{aligned}$$

Let

$$\widehat{X} := \mathrm{clip}_{[-Z-1,Z+1]}(\phi_X), \qquad \widehat{Y} := \mathrm{clip}_{[-Z-1,Z+1]}(\phi_Y).$$

On the enlarged inner domain, $|X|, |Y| \le Z$, so these clipping operations do not increase the approximation errors.

The additional two-input multiplication has budgets

$$L_2 = \mathcal{O}(1+\log \varepsilon_{\mathrm{mult}}^{-1}), \qquad W_2 = \mathcal{O}(1), \qquad S_2 = \mathcal{O}(1+\log \varepsilon_{\mathrm{mult}}^{-1}).$$

Including affine scaling and clipping, a sufficient weight bound for this stage is

$$B_2 \lesssim 1 \vee (B_x^*)^{-1} \vee \frac{KB_x^*}{\sqrt{c_0}} \vee \frac{A}{\sqrt{c_0}} \vee (Z+1).$$

The normalized state input is transmitted through Subnet 1 using a fixed signed identity channel. The corresponding sparsity cost is $\mathcal{O}(L_1)$.

Subnet 3: Implementation of the polynomial. Set

$$\varepsilon_{\mathrm{poly}} := \varepsilon/9, \qquad \varepsilon_{\mathrm{mono}} := \frac{\varepsilon_{\mathrm{poly}}}{1+\sum_{k,j}|b_{k,j}|}.$$

For $m = k + j \ge 2$, apply Lemma C.5 with m inputs, input radius 1, and tolerance $\varepsilon_{\mathrm{mono}}$. Feed k copies of $X/(Z+1)$ and j copies of $Y/(Z+1)$ to this network. The resulting monomial approximation has error at most $\varepsilon_{\mathrm{mono}}$ and budgets

$$\begin{aligned} L_{k,j} &= \mathcal{O}\left(\log(em)\log \varepsilon_{\mathrm{mono}}^{-1}\right), \\ W_{k,j} &= \mathcal{O}(m), \qquad S_{k,j} = \mathcal{O}\left(m \log \varepsilon_{\mathrm{mono}}^{-1}\right). \end{aligned}$$

Constant and degree-one monomials are represented exactly.

Run these subnetworks in parallel, align their scalar outputs, and sum them with coefficients $b_{k,j}$. This gives a network ϕ_{poly} satisfying

$$\sup_{|X|,|Y|\le Z+1} |\phi_{\mathrm{poly}}(X,Y) - P(X,Y)| \le \varepsilon_{\mathrm{poly}}.$$

There are $\mathcal{O}(K_x^2)$ monomials and

$$\sum_{\substack{k,j\ge 0\\ 2\le k+j\le K_x}} (k+j) = \mathcal{O}(K_x^3).$$

Including fixed identity layers and the final affine sum, sufficient budgets are

$$\begin{aligned} L_3 &= \mathcal{O}\left(\log(eK_x)\log\varepsilon_{\mathrm{mono}}^{-1}\right),\\ W_3 &= \mathcal{O}(K_x^3),\\ S_3 &= \mathcal{O}\left(K_x^3\log\varepsilon_{\mathrm{mono}}^{-1}\right),\\ B_3 &\lesssim 1+\max_{k,j}|b_{k,j}|. \end{aligned}$$

Here the scalar identity-padding cost is bounded by $\mathcal{O}(K_x^2\log(eK_x)\log\varepsilon_{\mathrm{mono}}^{-1})$ and is absorbed in the displayed sparsity bound.

By (13),

$$\log\varepsilon_{\mathrm{mono}}^{-1} = \mathcal{O}\left(\log\varepsilon^{-1} + K_x\log(Z+1) + |l|\right).$$

Using the selected degrees, we obtain the sufficient bounds

$$\begin{aligned} L_3 &= \mathcal{O}\left(\log(eK_x)\left[\log\varepsilon^{-1} + K_x\log(e+Z)\right]\right),\\ W_3 &= \mathcal{O}(K_x^3),\\ S_3 &= \mathcal{O}\left(K_x^3\left[\log\varepsilon^{-1} + K_x\log(e+Z)\right]\right),\\ \log(B_3\vee e) &= \mathcal{O}\left(\log\varepsilon^{-1} + K_x\log(e+Z)\right). \end{aligned}$$

These bounds follow from the explicit monomial construction.

Overall inner approximation. Define

$$\phi_{\mathcal{I}_l}^{\mathrm{in}}(x,t) := \phi_{\mathrm{poly}}(\widehat{X}(x,t),\widehat{Y}(t)).$$

On the enlarged inner domain, the input-accuracy choice and the derivative bound for P imply

$$|P(X,Y) - P(\widehat{X},\widehat{Y})| \le \varepsilon/9.$$

Therefore,

$$\begin{aligned} |\mathcal{I}_l - \phi_{\mathcal{I}_l}^{\mathrm{in}}| &\le |\mathcal{I}_l - \mathcal{I}_l^{(A)}| + |\mathcal{I}_l^{(A)} - P(X,Y)|\\ &\quad + |P(X,Y) - P(\widehat{X},\widehat{Y})| + |P(\widehat{X},\widehat{Y}) - \phi_{\mathrm{poly}}(\widehat{X},\widehat{Y})|\\ &\le \varepsilon/3 + \varepsilon/3 + \varepsilon/9 + \varepsilon/9 < \varepsilon. \end{aligned}$$

Including transmission of intermediate scalar outputs, sufficient inner-network budgets satisfy

$$\begin{aligned} L_{\mathrm{in}} &\le L_1 + L_2 + L_3 + \mathcal{O}(1),\\ W_{\mathrm{in}} &\lesssim W_1 + W_2 + W_3 + 1,\\ S_{\mathrm{in}} &\lesssim S_1 + S_2 + S_3 + L_1 + L_2 + L_3 + 1,\\ B_{\mathrm{in}} &\lesssim 1 \vee B_1 \vee B_2 \vee B_3. \end{aligned}$$

These composition bounds can be implemented using explicit identity and activation interfaces between subnetworks.

Dependence on the time endpoints. Since $\widetilde{m}_t \geq m_t/K$ and $m_t \geq e^{-\overline{\beta T}}$,

$$\frac{B_x^*}{\sqrt{c_0}} \leq K e^{\overline{\beta T}}(H_l + 2).$$

Consequently,

$$\log \varepsilon_1^{-1} + \log \varepsilon_{\text{mult}}^{-1} \lesssim_K 1 + \log \varepsilon_{\text{input}}^{-1} + \overline{\beta T} + \log\left(1 + \log \frac{e + |l|}{\varepsilon}\right).$$

Also,

$$\log \varepsilon_{\text{input}}^{-1} = \mathcal{O}\left(\log \varepsilon^{-1} + \log(eK_x) + K_x \log(Z+1) + |l|\right).$$

F.3 Continuous joining and global approximation

We join the inner approximation to the outer constant without using discontinuous indicator functions.

First, clip the spatial input exactly:

$$x^{\text{clip}} := \text{clip}_{[-B_x^*, B_x^*]}(x).$$

In this subsection, evaluate the previously constructed input subnetworks at (x^{clip}, t). Define

$$\widehat{X}(x,t) := \text{clip}_{[-Z-1, Z+1]}\big(\phi_X(x^{\text{clip}}, t)\big).$$

Use the exact ReLU cutoff

$$\chi_l(z) := 2\text{ReLU}(H_l + 1 - |z|) - 2\text{ReLU}(H_l + 1/2 - |z|),$$

where $|z| = \text{ReLU}(z) + \text{ReLU}(-z)$. It satisfies

$$0 \leq \chi_l \leq 1,$$
$$\chi_l(z) = 1 \quad \text{for } |z| \leq H_l + 1/2, \qquad \chi_l(z) = 0 \quad \text{for } |z| \geq H_l + 1.$$

Let

$$F_{\text{in}}(x,t) := \text{clip}_{[-\sqrt{2}, \sqrt{2}]}\big(\phi_{\mathcal{I}_l}^{\text{in}}(x^{\text{clip}}, t)\big).$$

On the inner approximation domain, clipping does not increase the error.

Consider the ideal blend

$$F(x,t) := 1 + \chi_l(\widehat{X}(x,t))\big(F_{\text{in}}(x,t) - 1\big).$$

If $|x| > B_x^*$, then

$$|X(x^{\text{clip}}, t)| = \frac{\widetilde{m}_t B_x^*}{\sqrt{c_0}} \geq H_l + 2.$$

The raw input approximation has error at most $1/4$ throughout the input box. Since the output-clipping threshold $Z + 1$ is larger than $H_l + 1$, this implies

$$|\widehat{X}(x,t)| \geq H_l + 1.$$

Hence the cutoff is zero outside the box. The original input also satisfies $|X(x,t)| \geq H_l + 2$, so the outer approximation is valid there.

If $\chi_l(\widehat{X}(x,t)) > 0$, then $|x| \leq B_x^*$ and $|\widehat{X}(x,t)| < H_l + 1$. At such a point the output clipping of ϕ_X cannot be saturated, so

$$|X(x,t)| \leq H_l + 1 + \frac{1}{4} < H_l + 2.$$

The inner approximation bound therefore gives:

$$|F_{\text{in}}(x,t) - \mathcal{I}_l(x,t)| \leq \varepsilon.$$

If $\chi_l(\widehat{X}(x,t)) < 1$ and $|x| \leq B_x^*$, then $|\widehat{X}(x,t)| > H_l + 1/2$. Output clipping cannot increase the absolute value, and the raw approximation error is at most $1/4$. Thus

$$|X(x,t)| > H_l + 1/4 > H_l,$$

so the outer approximation is valid. It was already verified outside the input box. Consequently, on the region where the outer contribution is nonzero,

$$|\mathcal{I}_l(x,t) - 1| \le \varepsilon.$$

The blend therefore satisfies, globally in x,

$$|F(x,t) - \mathcal{I}_l(x,t)| \le \chi_l(\widehat{X})|F_{\text{in}} - \mathcal{I}_l| + (1 - \chi_l(\widehat{X}))|1 - \mathcal{I}_l| \le \varepsilon.$$

To implement the blend, apply Lemma C.5 to the two inputs

$$\chi_l(\widehat{X}(x,t)), \qquad \frac{F_{\text{in}}(x,t) - 1}{3},$$

which belong to $[-1,1]$. Let $\phi^{\text{join}}_{\text{mult}}$ have approximation error at most $\varepsilon/3$, and define

$$\phi_{\mathcal{I}_l}(x,t) := \text{clip}_{[-\sqrt{2},\sqrt{2}]}\left(1 + 3\phi^{\text{join}}_{\text{mult}}\left(\chi_l(\widehat{X}(x,t)), \frac{F_{\text{in}}(x,t) - 1}{3}\right)\right).$$

Then

$$\sup_{\substack{x\in\mathbb{R}\\ t\in[\underline{T},\overline{T}]}} |\phi_{\mathcal{I}_l}(x,t) - \mathcal{I}_l(x,t)| \le 2\varepsilon.$$

Replacing the construction tolerance by $\varepsilon/2$ yields final error at most ε without changing the stated complexity orders.

Global network budgets. We now substitute the Subnet 1 budgets (14) into the composition and continuous-joining estimates. Let Λ_l be defined by (8).

The selected polynomial degrees and the input-domain bound satisfy

$$K_x \lesssim \left(|l| + \log \varepsilon^{-1}\right) \log \frac{e + |l|}{\varepsilon},$$

and

$$Z^2 \lesssim \log \frac{e + |l|}{\varepsilon}.$$

Consequently,

$$K_x \le C\Lambda_l, \qquad K_x \log(e + Z) \le C\Lambda_l, \qquad \log \varepsilon^{-1} \le C\Lambda_l.$$

By the coefficient and input-accuracy estimates,

$$\log \varepsilon^{-1}_{\text{mono}} \lesssim \Lambda_l,$$
$$\log \varepsilon^{-1}_{\text{input}} \lesssim \Lambda_l.$$

The time-horizon estimate further gives

$$\log \varepsilon_1^{-1} + \log \varepsilon^{-1}_{\text{mult}} + \log(1 + \underline{T}^{-1}) \lesssim_K \Lambda_l.$$

Applying (14), we therefore obtain

$$L_1 = \mathcal{O}(\Lambda_l^2), \qquad W_1 = \mathcal{O}(\Lambda_l^3),$$
$$S_1 = \mathcal{O}(\Lambda_l^4), \qquad \log(B_1 \vee e) = \mathcal{O}(\Lambda_l).$$

The additional multiplication in Subnet 2 satisfies

$$L_2 = \mathcal{O}(\Lambda_l), \qquad W_2 = \mathcal{O}(1), \qquad S_2 = \mathcal{O}(\Lambda_l).$$

For Subnet 3, the monomial construction gives

$$L_3 = \mathcal{O}\left(\log(eK_x) \log \varepsilon^{-1}_{\text{mono}}\right) = \mathcal{O}\left(\Lambda_l \log(e\Lambda_l)\right),$$
$$W_3 = \mathcal{O}(K_x^3) = \mathcal{O}(\Lambda_l^3),$$
$$S_3 = \mathcal{O}\left(K_x^3 \log \varepsilon^{-1}_{\text{mono}}\right) = \mathcal{O}(\Lambda_l^4),$$
$$\log(B_3 \vee e) = \mathcal{O}(\Lambda_l).$$

The affine normalizations and continuous joining have the following costs. Since $\widetilde{m}_t \le K$,

$$B_x^* \ge \frac{\sqrt{c_0}}{K}(H_l + 2).$$

Consequently,

$$(B_x^*)^{-1} \le \frac{K}{\sqrt{c_0}(H_l+2)} \le \frac{K}{2}\, c_0^{-1/2}.$$

On the other hand,

$$\frac{B_x^*}{\sqrt{c_0}} \le K e^{\overline{\beta T}}(H_l + 2).$$

Together with $c_0 \le 1$, these estimates bound all affine normalization, output-scaling, and spatial clipping coefficients by

$$C \max\left\{c_0^{-1/2}, \exp(C\Lambda_l)\right\}.$$

The coefficients of the cutoff and normalized-input clipping maps obey the same bound.

The continuous joining uses a two-input multiplication network of depth and sparsity $\mathcal{O}(1 + \log \varepsilon^{-1})$ and constant width. The cutoff value is transmitted through the remaining inner layers by a fixed identity channel. This adds constant width and $\mathcal{O}(L_{\text{in}})$ nonzero parameters.

Combining the subnetwork budgets and these fixed transmission layers yields

$$\begin{aligned} L_{\mathcal{I}_l} &= \mathcal{O}(\Lambda_l^2), \\ W_{\mathcal{I}_l} &= \mathcal{O}(\Lambda_l^3), \\ S_{\mathcal{I}_l} &= \mathcal{O}(\Lambda_l^4), \\ B_{\mathcal{I}_l} &\le C \max\left\{c_0^{-1/2}, \exp(C\Lambda_l)\right\}. \end{aligned}$$

When composing subnetworks, we retain explicit signed ReLU interfaces rather than merging adjacent affine maps. Thus the weight bounds are combined by a maximum, up to fixed factors, rather than by an unnecessary product of matrix bounds.

The continuous-joining argument gives approximation error at most 2ε at construction tolerance ε. Running the construction with tolerance $\varepsilon/2$ gives final error at most ε. Since

$$\Lambda_l(\varepsilon/2) \lesssim \Lambda_l(\varepsilon),$$

this adjustment does not change the displayed budget orders. The final output clipping guarantees $|\phi_{\mathcal{I}_l}| \le \sqrt{2}$ on the entire input space. This proves Lemma E.4.

G Evaluation of $\int_x \left|\phi^i_{2c\nabla\rho_t} - 2c_i\partial_{x_i}\rho_t\right|^2 \rho_t(x)\mathrm{d}\mu_{C_t}(x)$

$\partial_{x_i}\rho_t$ denotes $\dfrac{\partial \rho_t}{\partial x_i}$.

The integrand is evaluated as follows:

$$\left|\phi^i_{2c\nabla\rho_t} - 2c_i\partial_{x_i}\rho_t\right| \le \left|\phi^i_{2c\nabla\rho_t} - 2c_i\partial_{x_i}\mathcal{P}_t R_\Gamma(\rho_0)\right| + \left|2c_i\partial_{x_i}\mathcal{P}_t(R_\Gamma(\rho_0) - \rho_0)\right|.$$

By (16) and (17) below,

$$\begin{aligned} \int_x \left|\phi^i_{2c\nabla\rho_t} - 2c_i\partial_{x_i}\rho_t\right|^2 \mathrm{d}P_t(x) &\lesssim \int_x \left|\phi^i_{2c\nabla\rho_t} - 2c_i\partial_{x_i}\mathcal{P}_t R_\Gamma(\rho_0)\right|^2 \mathrm{d}P_t(x) \\ &\quad + \int_x \left|2c_i\partial_{x_i}\mathcal{P}_t(\rho_0 - R_\Gamma(\rho_0))\right|^2 \mathrm{d}P_t(x) \\ &\lesssim \frac{m_t^2}{\sigma_t^2} c_i\Gamma 2^{-2(1-v/a^\dagger)\Gamma} + c_i\Gamma 2^{-2(1-v/a^\dagger)\Gamma} \\ &\lesssim \frac{1}{\sigma_t^2} c_i\Gamma 2^{-2(1-v/a^\dagger)\Gamma}. \end{aligned} \tag{15}$$

G.1 Evaluation of the derivative of the truncation remainder

We first establish an L^2 smoothing estimate for $\mathcal{P}_t$. This estimate does not require an L^∞ bound on the truncation remainder.

Lemma G.1 (Coordinatewise Gaussian smoothing)**.** *Let $t > 0$ satisfy $\sigma_t > 0$, and let $f \in L^2(\mu_{C_0})$. Then the coordinatewise weak derivative satisfies*

$$\|2c_i\partial_{x_i}\mathcal{P}_t f\|^2_{L^2(\mu_{C_t})} \le \frac{4m_t^2 c_i c_{0,i}}{\sigma_t^2 c_{t,i}}\|f\|^2_{L^2(\mu_{C_0})}.$$

Consequently, under covariance comparability,

$$\|2c_i\partial_{x_i}\mathcal{P}_t f\|^2_{L^2(\mu_{C_t})} \lesssim \frac{m_t^2}{\sigma_t^2}c_i\|f\|^2_{L^2(\mu_{C_0})}.$$

If $P_t \le (3/2)\mu_{C_t}$, the same upper bound, up to a fixed constant, holds in $L^2(P_t)$.

Proof. First suppose that f is a smooth cylindrical function. By the Gaussian conditional representation,

$$\mathcal{P}_t f(x) = \int f(\widetilde{m}_t x + \widetilde{\sigma}_t y)\,\mathrm{d}\mu_{C_0}(y),$$

where

$$\widetilde{m}_{t,i} = \frac{m_t c_{0,i}}{c_{t,i}}, \qquad \widetilde{\sigma}^2_{t,i} = \frac{\sigma_t^2 c_i}{c_{t,i}}.$$

Gaussian integration by parts gives

$$\partial_{x_i}\mathcal{P}_t f(x) = \frac{\widetilde{m}_{t,i}}{\widetilde{\sigma}_{t,i}c_{0,i}}\int y_i f(\widetilde{m}_t x + \widetilde{\sigma}_t y)\,\mathrm{d}\mu_{C_0}(y).$$

Applying Cauchy–Schwarz inside this integral and using $\int y_i^2\,\mathrm{d}\mu_{C_0}(y) = c_{0,i}$, we obtain

$$|\partial_{x_i}\mathcal{P}_t f(x)|^2 \le \frac{\widetilde{m}^2_{t,i}}{\widetilde{\sigma}^2_{t,i}c_{0,i}}\int |f(\widetilde{m}_t x + \widetilde{\sigma}_t y)|^2\,\mathrm{d}\mu_{C_0}(y).$$

Let $X \sim \mu_{C_t}$ and $Y \sim \mu_{C_0}$ be independent. The centered Gaussian random variable $\widetilde{m}_t X + \widetilde{\sigma}_t Y$ has covariance C_0, because, coordinatewise,

$$\widetilde{m}^2_{t,i}c_{t,i} + \widetilde{\sigma}^2_{t,i}c_{0,i} = \frac{m_t^2 c_{0,i}^2 + \sigma_t^2 c_i c_{0,i}}{c_{t,i}} = c_{0,i}.$$

Thus

$$\int\int |f(\widetilde{m}_t x + \widetilde{\sigma}_t y)|^2\,\mathrm{d}\mu_{C_0}(y)\,\mathrm{d}\mu_{C_t}(x) = \|f\|^2_{L^2(\mu_{C_0})}.$$

Integrating the preceding derivative bound yields

$$\begin{aligned}\|2c_i\partial_{x_i}\mathcal{P}_t f\|^2_{L^2(\mu_{C_t})} &\le \frac{4c_i^2\widetilde{m}^2_{t,i}}{\widetilde{\sigma}^2_{t,i}c_{0,i}}\|f\|^2_{L^2(\mu_{C_0})} \\ &= \frac{4m_t^2 c_i c_{0,i}}{\sigma_t^2 c_{t,i}}\|f\|^2_{L^2(\mu_{C_0})}.\end{aligned}$$

Covariance comparability bounds $c_{0,i}/c_{t,i}$ uniformly in i and t.

For general $f \in L^2(\mu_{C_0})$, approximate f by smooth cylindrical functions. Contractivity of $\mathcal{P}_t$ and the derivative estimate show convergence of both the functions and their coordinate derivatives in the respective Gaussian L^2 spaces. Closedness of the coordinatewise weak derivative gives the asserted estimate for f.

Finally, $P_t \le (3/2)\mu_{C_t}$ implies

$$\|u\|^2_{L^2(P_t)} \le \frac{3}{2}\|u\|^2_{L^2(\mu_{C_t})},$$

which proves the last assertion. □

Apply Lemma G.1 to $f = \rho_0 - R_\Gamma(\rho_0)$. We obtain

$$\int |2c_i \partial_{x_i} \mathcal{P}_t(\rho_0 - R_\Gamma(\rho_0))(x)|^2 \, \mathrm{d}P_t(x) \lesssim \frac{m_t^2}{\sigma_t^2} c_i \|\rho_0 - R_\Gamma(\rho_0)\|^2_{L^2(\mu_{C_0})}.$$

Using the truncation bound

$$\|\rho_0 - R_\Gamma(\rho_0)\|_{L^2(\mu_{C_0})} \lesssim 2^{-(1-v/a^\dagger)\Gamma} \|\rho_0\|_{\mathcal{F}^\gamma_{p,q}},$$

and the radius-R_ρ ball assumption, with $R_\rho = \mathcal{O}(1)$ absorbed into the implicit constant, we obtain

$$\int |2c_i \partial_{x_i} \mathcal{P}_t(\rho_0 - R_\Gamma(\rho_0))(x)|^2 \, \mathrm{d}P_t(x) \lesssim \frac{m_t^2}{\sigma_t^2} c_i \, 2^{-2(1-v/a^\dagger)\Gamma}. \tag{16}$$

In particular, no uniform bound on $\rho_0 - R_\Gamma(\rho_0)$ outside a truncated input domain is needed.

G.2 Evaluation of $\int_x \left|\phi^i_{2c\nabla\rho_t} - 2c_i \partial_{x_i} \mathcal{P}_t R_\Gamma(\rho_0)\right|^2 \mathrm{d}P_t(x)$

Note that the input of $\partial_{x_i} \mathcal{P}_t R_\Gamma(\rho_0)$ is reduced to $d_{\max}$ dimensional, that means $2c_i \partial_{x_i} \mathcal{P}_t R_\Gamma(\rho_0) = 0$ and the output of $\phi^i_{2c\nabla\rho_t}$ is 0 for $d_{\max} < i \le d_{\text{out}}$. So note that $\phi^i_{2c\nabla\rho_t}$ uses only $(x_i)_{i=1}^{d_{\max}}$ for the input. Here, d denotes $d_{\max}$ for simplicity.

By Lemma A.7,

$$\begin{aligned}
2c_{0,i} \partial_{x_i} \mathcal{P}_t R_\Gamma(\rho_0) &= \sum_{s:\gamma(s)<\Gamma} \sum_{l \in J(s)} \langle \rho_0, \psi_l \rangle \frac{2\widetilde{m}_{t,i}}{\widetilde{\sigma}_{t,i}} \int_{\mathbb{R}^d} y_i \psi_l(\widetilde{m}_t x + \widetilde{\sigma}_t y) \mathrm{d}\mu_{C_0}(y) \\
&= \frac{2\widetilde{m}_{t,i}}{\widetilde{\sigma}_{t,i}} \sum_{s:\gamma(s)<\Gamma} \sum_{l \in J(s)} \langle \rho_0, \psi_l \rangle \int_{\mathbb{R}} y_i g_{l_i} \left(\Phi \left(\frac{\widetilde{m}_{t,i} x_i + \widetilde{\sigma}_{t,i} y_i}{\sqrt{c_{0,i}}} \right) \right) \varphi_{c_{0,i}}(y_i) \mathrm{d}y_i \prod_{j \neq i}^{d} \mathcal{I}_{l_j},
\end{aligned}$$

using $\mathcal{I}_{l_j}$ defined in (7) and it is approximated by a neural network $\phi_{\mathcal{I}_{l_i}}$.

Define

$$\mathcal{J}_{l_i}(x_i, t) := \int_{\mathbb{R}} y_i \, g_{l_i} \left(\Phi \left(\frac{\widetilde{m}_{t,i} x_i + \widetilde{\sigma}_{t,i} y_i}{\sqrt{c_{0,i}}} \right) \right) \varphi_{c_{0,i}}(y_i) \, \mathrm{d}y_i.$$

We approximate this factor at its natural scale $\sqrt{c_{0,i}}$.

Lemma G.2 (ReLU approximation of $\mathcal{J}_l$)**.** *Assume the same covariance conditions as in Lemma E.4 and the noise-schedule assumptions in Section 2. Let $0 < \underline{T} < \overline{T} < \infty$, $0 < c_0, c \le 1$, and $0 < \varepsilon < 1/2$. For $l \in \mathbb{Z} \setminus \{0\}$, let Λ_l be defined by (8) with tolerance ε.*

Then there exists a ReLU network

$$\phi_{\mathcal{J}_l} \in \Phi(L_{\mathcal{J}_l}, W_{\mathcal{J}_l}, S_{\mathcal{J}_l}, B_{\mathcal{J}_l}; 2, 1)$$

such that

$$\sup_{\substack{x \in \mathbb{R} \\ t \in [\underline{T}, \overline{T}]}} |\phi_{\mathcal{J}_l}(x,t) - \mathcal{J}_l(x,t)| \le \sqrt{c_0}\, \varepsilon,$$

and

$$\sup_{x,t} |\phi_{\mathcal{J}_l}(x,t)| \le 2\sqrt{c_0}.$$

Sufficient budgets are

$$\begin{aligned}
L_{\mathcal{J}_l} &= \mathcal{O}(\Lambda_l^2), \\
W_{\mathcal{J}_l} &= \mathcal{O}(\Lambda_l^3), \\
S_{\mathcal{J}_l} &= \mathcal{O}(\Lambda_l^4), \\
B_{\mathcal{J}_l} &\le C \max\left\{ c_0^{-1/2}, \exp(C\Lambda_l) \right\}.
\end{aligned}$$

The constants have the same allowed dependencies as in Lemma E.4. For $l = 0$, $\mathcal{J}_0 = 0$ is represented exactly.

Proof. We adapt the construction in Appendix F. The argument below gives normalized error at most 2ε. Applying it with tolerance $\varepsilon/2$ gives the stated result.

Normalization and boundedness. Changing variables $y = \sqrt{c_0}z$, write

$$\frac{\mathcal{J}_l(x,t)}{\sqrt{c_0}} = \int_{\mathbb{R}} z\, g_l\left(\Phi\left(\frac{\widetilde{m}_t x}{\sqrt{c_0}} + \widetilde{\sigma}_t z\right)\right)\varphi_1(z)\,\mathrm{d}z.$$

Since $\|g_l\|_\infty \le \sqrt{2}$,

$$\left|\frac{\mathcal{J}_l(x,t)}{\sqrt{c_0}}\right| \le \sqrt{2}\,\mathbb{E}|Z_0| = \frac{2}{\sqrt{\pi}} < 2, \qquad Z_0 \sim \mathcal{N}(0,1).$$

Outer approximation. Set

$$X := \frac{\widetilde{m}_t x}{\sqrt{c_0}}.$$

Using $\mathbb{E}Z_0 = 0$ and $g_l(0) = g_l(1) = 1$,

$$\begin{aligned}\left|\frac{\mathcal{J}_l(x,t)}{\sqrt{c_0}}\right| &= \left|\mathbb{E}\left[Z_0\big(g_l(\Phi(X+\widetilde{\sigma}_t Z_0)) - 1\big)\right]\right| \\ &\le 2\sqrt{2}\pi|l|\,\mathbb{E}\left[|Z_0|\Phi(-|X+\widetilde{\sigma}_t Z_0|)\right].\end{aligned}$$

By Cauchy–Schwarz and $0 \le \Phi(-|z|) \le 1$,

$$\mathbb{E}\left[|Z_0|\Phi(-|X+\widetilde{\sigma}_t Z_0|)\right] \le \left[\mathbb{E}\Phi(-|X+\widetilde{\sigma}_t Z_0|)\right]^{1/2}.$$

Since $\widetilde{\sigma}_t \le 1$, splitting according to $|Z_0| \le |X|/2$ gives

$$\mathbb{E}\Phi(-|X+\widetilde{\sigma}_t Z_0|) \le 3\Phi(-|X|/2) \le 3e^{-|X|^2/8}.$$

Consequently,

$$\left|\frac{\mathcal{J}_l(x,t)}{\sqrt{c_0}}\right| \le 2\sqrt{6}\pi|l|e^{-|X|^2/16}.$$

Choose

$$H_l^{\mathcal{J}} := \sqrt{16\log\left(\frac{48\pi|l|}{\varepsilon}\right)}.$$

On $|X| \ge H_l^{\mathcal{J}}$, the constant 0 approximates the normalized factor with error at most ε.

Integral truncation and polynomialization. Take

$$A := \sqrt{2c_0\log(24/\varepsilon)}.$$

The normalized truncation error is bounded by

$$\begin{aligned}&\frac{1}{\sqrt{c_0}}\left|\int_{|y|>A} y\, g_l\left(\Phi\left(\frac{\widetilde{m}_t x + \widetilde{\sigma}_t y}{\sqrt{c_0}}\right)\right)\varphi_{c_0}(y)\,\mathrm{d}y\right| \\ &\le \sqrt{2}\,\mathbb{E}\left[|Z_0|\mathbf{1}_{\{|Z_0|>A/\sqrt{c_0}\}}\right] = \frac{2}{\sqrt{\pi}}e^{-A^2/(2c_0)} < \varepsilon/3.\end{aligned}$$

Construct the inner approximation on $|X| \le H_l^{\mathcal{J}} + 2$. Define

$$Y := \frac{\widetilde{\sigma}_t A}{\sqrt{c_0}}, \qquad Z_{\mathcal{J}} := H_l^{\mathcal{J}} + 2 + \frac{A}{\sqrt{c_0}}.$$

Then

$$Z_{\mathcal{J}}^2 = \mathcal{O}\left(\log\frac{e+|l|}{\varepsilon}\right).$$

Use the polynomials Φ_R, $g_{l,N}^{(\delta_\Phi)}$, and q_M from Appendix F, with $\delta_\Phi = \delta_\Phi(R; Z_{\mathcal{J}})$. The normalized polynomialization error is at most

$$2\sqrt{2}\pi|l|\,\delta_\Phi + \delta_g(N; l, \delta_\Phi) + 2\sqrt{2}\frac{A^2}{\sqrt{c_0}}\,\delta_\varphi(M; A).$$

For the first two terms, we used

$$\int_{-A}^{A} \frac{|y|}{\sqrt{c_0}} \varphi_{c_0}(y)\,\mathrm{d}y \le \mathbb{E}|Z_0| < 1.$$

For the last term, we used $|g_{l,N}^{(\delta_\Phi)}(\Phi_R)| \le 2\sqrt{2}$ and

$$\frac{1}{\sqrt{c_0}} \int_{-A}^{A} |y|\,\mathrm{d}y = \frac{A^2}{\sqrt{c_0}}.$$

Sufficiently large fixed constants in the choices

$$\begin{aligned} R &= \mathcal{O}\left(\log \frac{e+|l|}{\varepsilon}\right), \\ N &= \mathcal{O}\left(|l| + \log \varepsilon^{-1}\right), \\ M &= \mathcal{O}(\log \varepsilon^{-1}) \end{aligned}$$

make this error at most $\varepsilon/3$. Indeed, the weight-error term is bounded by a fixed polynomial in $\log \varepsilon^{-1}$ times e^{-M} once $M \ge e^2 A^2/(2c_0)$.

Polynomial degree and coefficients. After the substitution $y = Au$, the normalized polynomial approximation is

$$\begin{aligned} P_{\mathcal{J}}(X,Y) := \frac{A^2}{\sqrt{2\pi}\,c_0} \int_{-1}^{1} u\, g_{l,N}^{(\delta_\Phi)}\big(\Phi_R(X+Yu)\big) \\ \times \sum_{m=0}^{M-1} \frac{(-1)^m}{m!} \left(\frac{A^2}{2c_0}\right)^m u^{2m}\,\mathrm{d}u. \end{aligned}$$

The additional factor u does not increase the degree in (X,Y). Thus

$$P_{\mathcal{J}}(X,Y) = \sum_{k+j \le K_x} a_{k,j}^{\mathcal{J}} X^k Y^j, \qquad K_x = (2N-1)(2R-1).$$

Integration is exact, using

$$\int_{-1}^{1} u^{r+1}\,\mathrm{d}u = \begin{cases} 0, & r \text{ even}, \\ 2/(r+2), & r \text{ odd}. \end{cases}$$

The coefficient absolute-sum argument used for $\mathcal{I}_l$ gives

$$\sum_{k,j} |a_{k,j}^{\mathcal{J}}| \le C \frac{A^2}{c_0} \exp\left(C|l| + \frac{A^2}{2c_0}\right).$$

Consequently,

$$\log\left(1 + \sum_{k,j} |a_{k,j}^{\mathcal{J}}|\right) = \mathcal{O}\left(|l| + \log \varepsilon^{-1}\right).$$

The degree and logarithmic coefficient bounds therefore have the same orders as for $\mathcal{I}_l$.

Network construction and continuous joining. Apply the normalized monomial construction of Appendix F to $P_{\mathcal{J}}$ on $[-Z_{\mathcal{J}} - 1, Z_{\mathcal{J}} + 1]^2$. Choose the polynomial-network error at most $\varepsilon/9$ and the normalized-input errors small enough that their contribution is at most $\varepsilon/9$. Also impose an input-error upper bound of $1/4$.
The time and input networks are constructed on the spatial box of radius

$$B_{x,\mathcal{J}}^{*} := \sup_{t \in [\underline{T}, \overline{T}]} \frac{\sqrt{c_0}}{\widetilde{m}_t} (H_l^{\mathcal{J}} + 2).$$

Exactly as for $\mathcal{I}_l$, this produces an inner approximation to $\mathcal{J}_l/\sqrt{c_0}$ with error less than ε on the enlarged inner domain. Clip its output to $[-2, 2]$.

For global extension, first clip the spatial input to $[-B^*_{x,\mathcal{J}}, B^*_{x,\mathcal{J}}]$. Let $\widehat{X}_{\mathcal{J}}$ denote the corresponding approximate normalized input, clipped to $[-Z_{\mathcal{J}} - 1, Z_{\mathcal{J}} + 1]$. Use the exact ReLU cutoff

$$\chi_{\mathcal{J}}(z) := 2\mathrm{ReLU}(H_l^{\mathcal{J}} + 1 - |z|) - 2\mathrm{ReLU}(H_l^{\mathcal{J}} + 1/2 - |z|).$$

Denote the clipped normalized inner output by $F_{\mathrm{in}}^{\mathcal{J}}$. The ideal blend is

$$F^{\mathcal{J}}(x,t) := \chi_{\mathcal{J}}(\widehat{X}_{\mathcal{J}}(x,t)) F_{\mathrm{in}}^{\mathcal{J}}(x,t).$$

The same overlap argument as for $\mathcal{I}_l$ shows that the inner approximation is valid wherever the cutoff is positive, and the outer approximation by zero is valid wherever the cutoff is less than one. Indeed, the normalized-input error is at most $1/4$, and the inner domain extends two units beyond the outer threshold. Thus

$$\sup_{x,t} \left| F^{\mathcal{J}}(x,t) - \frac{\mathcal{J}_l(x,t)}{\sqrt{c_0}} \right| \le \varepsilon.$$

Implement the blend by applying the two-input multiplication network to

$$\chi_{\mathcal{J}}(\widehat{X}_{\mathcal{J}}), \qquad F_{\mathrm{in}}^{\mathcal{J}}/2,$$

with multiplication error at most $\varepsilon/2$, and scale its output by 2. Final clipping to $[-2, 2]$ gives normalized error at most 2ε.

The polynomial degree, coefficient bound, spatial normalization, and time-approximation accuracies have the same orders as in Lemma E.4. Substitution into the same subnetwork budget calculation therefore yields

$$L = \mathcal{O}(\Lambda_l^2), \qquad W = \mathcal{O}(\Lambda_l^3), \qquad S = \mathcal{O}(\Lambda_l^4),$$

and

$$B \le C \max\{c_0^{-1/2}, \exp(C\Lambda_l)\}.$$

These budgets include the continuous-joining network and the fixed identity channels.

Run the construction at tolerance $\varepsilon/2$ and multiply the final output by $\sqrt{c_0}$. The tolerance change does not alter the budget orders. Since $c_0 \le 1$, the final scaling does not enlarge the weight bound. The approximation error is at most $\sqrt{c_0}\varepsilon$, and the output is bounded by $2\sqrt{c_0}$. □

The derivative construction uses the prefactor

$$\frac{2c_i \widetilde{m}_{t,i}}{c_{0,i} \widetilde{\sigma}_{t,i}}.$$

Its product with the natural scale of $\mathcal{J}_{l_i}$ satisfies

$$\frac{2c_i \widetilde{m}_{t,i}}{c_{0,i} \widetilde{\sigma}_{t,i}} \sqrt{c_{0,i}} = \frac{2m_t \sqrt{c_i c_{0,i}}}{\sigma_t \sqrt{c_{t,i}}} \lesssim K \frac{m_t \sqrt{c_i}}{\sigma_t}.$$

Thus the normalized factor accuracy is allocated after accounting for this amplification, the remaining product factors, and the basis coefficients.

To obtain the retained-derivative approximation error in (17), under the radius-R_ρ ball assumption on ρ_0, with the fixed $R_\rho = \mathcal{O}(1)$ absorbed into the implicit constants, take

$$\varepsilon := \min\left\{ \frac{1}{4}, \frac{\sigma_{\underline{T}} \sqrt{\Gamma}\, 2^{-(1-v/a^\dagger)\Gamma}}{8\sqrt{K}\, 2^{d_{\max}} G B_\rho (d_{\max} + 2)} \right\},$$

where G denotes the number of retained basis functions and $B_\rho \ge 1 \vee \|\rho_0\|_{L^2(\mu_{C_0})}$. Using Lemmas E.4 and G.2, approximate $\mathcal{I}_{l_j}$ and $\mathcal{J}_{l_i}/\sqrt{c_{0,i}}$ at accuracy ε. The derivative prefactor satisfies

$$\frac{2c_i \widetilde{m}_{t,i}}{c_{0,i} \widetilde{\sigma}_{t,i}} \sqrt{c_{0,i}} = \frac{2m_t \sqrt{c_i c_{0,i}}}{\sigma_t \sqrt{c_{t,i}}} \le \frac{2\sqrt{K c_i}}{\sigma_{\underline{T}}}.$$

Normalize this coefficient to $[0, 1]$ and approximate it and the normalized products at accuracy ε. The product-error bound and summation over the retained basis yield the required coordinatewise error.

Let $\Lambda_l(\varepsilon)$ denote the quantity in (8), and set

$$K' := 1 + d_{\max} + \log(1 + \sigma_{\underline{T}}^{-1}) + \max_{\substack{l \in \mathbb{Z} \\ 1 \le |l| \le 2^{f_{\max}}}} \Lambda_l(\varepsilon).$$

Then there exists a vector-valued network

$$\phi_{2c\nabla\rho_t} \in \Phi(L, W, S, B; d_{\max} + 1, d_{\max})$$

such that, for $i \le d_{\max}$ and $t \in [\underline{T}, \overline{T}]$,

$$\int \left| \phi^i_{2c\nabla\rho_t}(x,t) - 2c_i \partial_{x_i} \mathcal{P}_t R_\Gamma(\rho_0)(x) \right|^2 \mathrm{d}P_t(x) \lesssim c_i \Gamma \, 2^{-2(1-v/a^\dagger)\Gamma}. \tag{17}$$

Using independent copies for each retained basis term and output coordinate, sufficient budgets are

$$\begin{aligned} L &= \mathcal{O}(K'^2), \\ W &= \mathcal{O}(d_{\max} G K'^4), \\ S &= \mathcal{O}(d_{\max} G K'^5), \\ B &\le C \max\left\{ \max_{j \le d_{\max}} c_{0,j}^{-1/2}, \exp(CK') \right\}. \end{aligned}$$

These budgets include the time-coefficient approximation, product networks, coefficient summation, and depth alignment. To approximate σ_t^{-1}, use a dyadic partition of $[\sigma_{\underline{T}}, 1]$ and rescale each interval to a fixed interval before applying Lemma C.6. This contributes at most depth $\mathcal{O}(K'^2)$, width $\mathcal{O}(K'^4)$, sparsity $\mathcal{O}(K'^5)$, and logarithmic weight bound $\mathcal{O}(K')$ per coefficient network. These bounds use independent network copies for different basis terms.

H Proof of Theorem 5.1 : The estimation error bound

H.1 Wasserstein-1 bounds

Let $\tilde{X}_t$ be the top d_{out}-dimensional marginal of X_t. Then, its forward process is given by

$$\tilde{X}_0 \sim \tilde{P}_0, \quad \mathrm{d}\tilde{X}_t = -\beta_t \tilde{X}_t \, \mathrm{d}t + \sqrt{2\beta_t C_{[1:d_{\text{out}}]}} \, \mathrm{d}B_t,$$

where $\tilde{P}_t$ is the marginal distribution of P_t. By abuse of notation, we also use the same notation C for $C_{[1:d_{\text{out}}]}$ in finite-dimensional expressions. In expressions involving the full Hilbert space, C retains its original meaning. Let $\tilde{p}_t$ be the density of $\tilde{P}_t$. Then, its reverse process is given by

$$\tilde{Y}_0 \sim \tilde{P}_{\overline{T}}, \quad \mathrm{d}\tilde{Y}_t = \beta_{\overline{T}-t} \left(\tilde{Y}_t + 2C_{[1:d_{\text{out}}]} \nabla \log \tilde{p}_{\overline{T}-t}(\tilde{Y}_t) \right) \mathrm{d}t + \sqrt{2\beta_{\overline{T}-t} C_{[1:d_{\text{out}}]}} \, \mathrm{d}B_t.$$

Let $\iota_{d_{\text{out}}} : \mathbb{R}^{d_{\text{out}}} \to H$ denote the zero-padding embedding along the chosen orthonormal basis. Whenever a finite-dimensional distribution is compared with a distribution on H, we identify it with its pushforward under $\iota_{d_{\text{out}}}$. In particular, $\|\iota_{d_{\text{out}}} y\|_H = \|y\|_2$.

By the density-ratio bound, covariance comparability, and the eigenvalue decay assumption, the natural coordinate coupling gives

$$\mathcal{W}_2^2(P_0, \tilde{P}_0) \le \mathbb{E}\left[\sum_{i > d_{\text{out}}} |\langle X_0, e_i \rangle_H|^2 \right] \lesssim \sum_{i > d_{\text{out}}} c_i \lesssim d_{\text{out}}^{1-2/r_0}.$$

Moreover, the natural coupling of the forward OU process gives, for $0 < \underline{T} \le 1$, using $\beta_t \le \overline{\beta}$,

$$\mathcal{W}_2^2(\tilde{P}_0, \tilde{P}_{\underline{T}}) \lesssim \underline{T}.$$

Consequently,

$$\mathcal{W}_1^2(P_0, \tilde{P}_{\underline{T}}) \le \mathcal{W}_2^2(P_0, \tilde{P}_{\underline{T}}) \lesssim \mathcal{W}_2^2(\tilde{P}_0, \tilde{P}_{\underline{T}}) + \mathcal{W}_2^2(P_0, \tilde{P}_0) \lesssim \underline{T} + d_{\text{out}}^{1-2/r_0}.$$

We keep the initialization error explicit by writing

$$K_{\text{init}} := \mathrm{KL}\left(\tilde{P}_{\overline{T}} \,\middle\|\, \mathcal{N}(0, C_{[1:d_{\text{out}}]})\right).$$

For the time horizon chosen below, the Gaussian initialization estimate gives $K_{\text{init}} = \mathcal{O}(n^{-3})$.

We use forward time as the time argument of all score functions and risks:

$$\bar{R}_{T_0,T_1;A}(F,F') := \frac{1}{T_1 - T_0}\int_{T_0}^{T_1} \mathbb{E}_{x\sim\tilde{P}_t}\left[\|F(x,t) - F'(x,t)\|_A^2\right] \mathrm{d}t.$$

Here and below, $C = C_{[1:d_{\text{out}}]}$ in finite-dimensional expressions.

Conditionally on the training data, the learned reverse process is

$$\begin{aligned}
&\hat{Y}_0^{(d_{\text{out}})} \sim \mathcal{N}(0, C),\\
&\mathrm{d}\hat{Y}_u^{(d_{\text{out}})} = \beta_{\overline{T}-u}\left[\hat{Y}_u^{(d_{\text{out}})} + \hat{s}(\hat{Y}_u^{(d_{\text{out}})}, \overline{T} - u)\right] \mathrm{d}u + \sqrt{2\beta_{\overline{T}-u}C}\,\mathrm{d}B_u,
\end{aligned}$$

for $u \in [0, \overline{T} - \underline{T}]$. Let $\hat{Q}_u$ denote its conditional law.

By the standing noise-schedule assumptions in Section 2, $\underline{\beta} \le \beta_t \le \overline{\beta}$ on $[\underline{T}, \overline{T}]$.

Applicability of the localized Girsanov bound. Fix the training data, $d_{\text{out}} < \infty$, and $0 < \underline{T} < \overline{T}$. The clipped score $\hat{s}$ is a finite ReLU network followed by fixed coordinatewise clipping maps. It is therefore globally Lipschitz in the spatial input, uniformly over the compact time interval, and has bounded output. The learned reverse drift

$$x \mapsto \beta_{\overline{T}-u}\big(x + \hat{s}(x, \overline{T} - u)\big)$$

has linear growth and defines a nonexplosive finite-dimensional SDE.

For the true marginal score, write

$$s_{d_{\text{out}}}(x,t) = -2CC^{-1}_{t,[1:d_{\text{out}}]}x + v_t(x).$$

Covariance comparability bounds the linear coefficient, and the relative-score bounds give

$$\sup_{x,t}\|v_t(x)\|_{C^{-1}} < \infty.$$

The constants in this existence argument may depend on d_{out} and $\underline{T}$.

Since $\sigma_{\underline{T}} > 0$, finite-dimensional Gaussian smoothing makes the marginal density positive and smooth in the spatial variable on $[\underline{T}, \overline{T}]$. Together with the density-ratio bounds, this gives the local spatial regularity needed for uniqueness of the true reverse dynamics and the stopped changes of drift. The true reverse drift has linear growth by the displayed score decomposition.

The drift-change energy is finite, as the following calculation shows. The two reverse processes have diffusion matrix

$$\Sigma_u = \sqrt{2\beta_{\overline{T}-u}C}.$$

For $t = \overline{T} - u$, the squared norm of the normalized drift difference is

$$\left\|\Sigma_u^{-1}\beta_t\big(s_{d_{\text{out}}}(x,t) - \hat{s}(x,t)\big)\right\|_2^2 = \frac{\beta_t}{2}\|s_{d_{\text{out}}}(x,t) - \hat{s}(x,t)\|_{C^{-1}}^2.$$

Thus no inverse power of β_t remains in the energy integrand.

Gaussian domination and covariance comparability give

$$\sup_{t\in[\underline{T},\overline{T}]}\mathbb{E}_{\tilde{P}_t}[\|X\|_{C^{-1}}^2] \lesssim d_{\text{out}}.$$

Using the true score decomposition and the clipped output envelope,

$$\sup_{t\in[\underline{T},\overline{T}]}\mathbb{E}_{\tilde{P}_t}[\|s_{d_{\text{out}}}(X,t) - \hat{s}(X,t)\|_{C^{-1}}^2] \lesssim 1 + d_{\text{out}} + F_{C^{-1}}^2 < \infty.$$

Since $\beta_t \le \overline{\beta}$ and the time interval is finite, the required energy is finite.

Finally, $\mathrm{KL}(\tilde{P}_{\overline{T}} \| \mathcal{N}(0, C)) < \infty$. Indeed, $\tilde{P}_{\overline{T}}$ is dominated by a fixed multiple of $\mathcal{N}(0, C_{\overline{T},[1:d_{\mathrm{out}}]})$, has finite second moment, and the two finite-dimensional Gaussian covariances are positive definite. Proposition A.4 therefore applies conditionally on the training data.

By Proposition A.4,

$$\begin{aligned}\mathrm{KL}\left(\tilde{P}_{\underline{T}} \middle\| \hat{Q}_{\overline{T}-\underline{T}}\right) &\le K_{\mathrm{init}} + \frac{1}{4}\int_{\underline{T}}^{\overline{T}} \beta_t \mathbb{E}_{x\sim\tilde{P}_t}\left[\|\hat{s}(x,t) - s_{d_{\mathrm{out}}}(x,t)\|^2_{C^{-1}}\right] \mathrm{d}t \\ &\le K_{\mathrm{init}} + \frac{\overline{\beta}(\overline{T}-\underline{T})}{4} \bar{R}_{\underline{T},\overline{T};C^{-1}}(\hat{s}, s_{d_{\mathrm{out}}}).\end{aligned}$$

We retain $\hat{Q}_{\overline{T}-\underline{T}}$ for the unprocessed output distribution of the learned reverse process. For $R > 0$, define

$$\Pi_R(y) := \begin{cases} y, & \|y\|_2 \le R, \\ \dfrac{R}{\|y\|_2} y, & \|y\|_2 > R. \end{cases}$$

This projection agrees, after zero-padding, with the projection onto the radius-R ball in H. Set

$$R_n = K_{\mathrm{proj}}\sqrt{\log n}, \qquad \hat{Q}^{\mathrm{proj}}_{\overline{T}-\underline{T}} := (\Pi_{R_n})_{\#}\hat{Q}_{\overline{T}-\underline{T}},$$

where $n \ge 2$ and K_{proj} is a sufficiently large fixed constant specified below.

Lemma H.1. *Suppose that $\|\rho_0 - 1\|_{L^\infty(\mu_{C_0})} \le 1/2$ and that C_0 and C are trace-class covariance operators. For a sufficiently large fixed K_{proj}, it holds that*

$$\mathcal{W}_1^2\left(\tilde{P}_{\underline{T}}, \hat{Q}^{\mathrm{proj}}_{\overline{T}-\underline{T}}\right) \lesssim n^{-2} + \log n \ \mathrm{KL}\left(\tilde{P}_{\underline{T}} \middle\| \hat{Q}_{\overline{T}-\underline{T}}\right).$$

The implicit constant is independent of n and d_{out}.

Proof. We first establish a general projection bound. Throughout this proof, total variation is normalized as

$$\mathrm{TV}(P,Q) := \sup_E |P(E) - Q(E)|.$$

With this convention, Pinsker's inequality (Lemma A.5) gives

$$2\,\mathrm{TV}(P,Q)^2 \le \mathrm{KL}(P\|Q).$$

Let P, Q be probability measures on $\mathbb{R}^{d_{\mathrm{out}}}$, with P having a finite first moment, and write

$$b_P(R) := \mathbb{E}_{X\sim P}\left[(\|X\|_2 - R)_+\right].$$

The coupling $(X, \Pi_R X)$ gives

$$\mathcal{W}_1(P, (\Pi_R)_{\#}P) \le \mathbb{E}_{X\sim P}\|X - \Pi_R X\| = b_P(R).$$

Both $(\Pi_R)_{\#}P$ and $(\Pi_R)_{\#}Q$ are supported on a ball of diameter $2R$. Consequently, by the contraction of total variation under measurable maps,

$$\begin{aligned}\mathcal{W}_1\left((\Pi_R)_{\#}P, (\Pi_R)_{\#}Q\right) &\le 2R\,\mathrm{TV}\left((\Pi_R)_{\#}P, (\Pi_R)_{\#}Q\right) \\ &\le 2R\,\mathrm{TV}(P,Q).\end{aligned}$$

The triangle inequality therefore yields

$$\mathcal{W}_1(P, (\Pi_R)_{\#}Q) \le b_P(R) + 2R\,\mathrm{TV}(P,Q).$$

Squaring this bound and applying Pinsker's inequality, we obtain

$$\begin{aligned}\mathcal{W}_1^2(P, (\Pi_R)_{\#}Q) &\le 2b_P(R)^2 + 8R^2\mathrm{TV}(P,Q)^2 \\ &\le 2b_P(R)^2 + 4R^2\mathrm{KL}(P\|Q).\end{aligned}$$

If the relative entropy is infinite, the last inequality is understood in the extended sense.

For $P = \tilde{P}_{\underline{T}}$, Gaussian domination gives the following bound on $b_P(R)$. Under the reference Gaussian forward process, the density ratio satisfies

$$\rho_t(x) = \mathbb{E}^{\mathrm{ref}}\left[\rho_0(X_0) \mid X_t = x\right].$$

Thus the assumption on ρ_0 implies

$$\frac{1}{2} \le \rho_t \le \frac{3}{2},$$

and, after taking the first d_{out} coordinates,

$$\tilde{P}_t \le \frac{3}{2}\,\mathcal{N}(0, C_{t,[1:d_{\mathrm{out}}]})$$

as measures.
In the following definitions, C_0 and C denote the full covariance operators. Set

$$V := \max\{\mathrm{Tr}(C_0), \mathrm{Tr}(C)\}, \qquad \Lambda := \max\{\|C_0\|_{\mathrm{op}}, \|C\|_{\mathrm{op}}\}.$$

Since $C_t = m_t^2 C_0 + (1 - m_t^2)C$, where $m_t = \exp(-\int_0^t \beta_u \,\mathrm{d}u)$, we have

$$\mathrm{Tr}(C_{t,[1:d_{\mathrm{out}}]}) \le V, \qquad \|C_{t,[1:d_{\mathrm{out}}]}\|_{\mathrm{op}} \le \Lambda.$$

Choose a fixed $a > 0$ such that $2a\Lambda \le 1/2$. Since $0 \le c_{t,i} \le \Lambda$, the Gaussian exponential-moment formula and $-\log(1-u) \le 2u$ for $0 \le u \le 1/2$ imply

$$\begin{aligned}
\mathbb{E}_{X\sim\tilde{P}_t}\left[\exp(a\|X\|^2)\right] &\le \frac{3}{2}\prod_{i=1}^{d_{\mathrm{out}}}(1-2ac_{t,i})^{-1/2} \\
&\le \frac{3}{2}\exp\left(2a\sum_{i=1}^{d_{\mathrm{out}}} c_{t,i}\right) \\
&\le \frac{3}{2}\exp(2aV) =: M.
\end{aligned}$$

In particular, a and M can be chosen independently of t and d_{out}. Markov's inequality gives

$$\tilde{P}_t(\|X\| > u) \le M\exp(-au^2).$$

Consequently, for every $R > 0$,

$$\begin{aligned}
b_{\tilde{P}_t}(R) &= \int_R^\infty \tilde{P}_t(\|X\| > u)\,\mathrm{d}u \\
&\le M\int_R^\infty \exp(-au^2)\,\mathrm{d}u \le \frac{M}{2aR}\exp(-aR^2).
\end{aligned}$$

Choose $K_{\mathrm{proj}} > 0$ such that $aK_{\mathrm{proj}}^2 \ge 2$, and set $R_n = K_{\mathrm{proj}}\sqrt{\log n}$. For $n \ge 2$,

$$b_{\tilde{P}_t}(R_n)^2 \lesssim \frac{n^{-4}}{\log n} \lesssim n^{-2}.$$

Applying the general projection bound with $P = \tilde{P}_{\underline{T}}$ and $Q = \hat{Q}_{\overline{T}-\underline{T}}$ yields

$$\mathcal{W}_1^2\left(\tilde{P}_{\underline{T}}, \hat{Q}^{\mathrm{proj}}_{\overline{T}-\underline{T}}\right) \lesssim n^{-2} + \log n\ \mathrm{KL}\left(\tilde{P}_{\underline{T}} \,\middle\|\, \hat{Q}_{\overline{T}-\underline{T}}\right).$$

The same exponential-moment bound also holds for P_0 on H, by the density-ratio bound and the Gaussian exponential-moment formula.
Hence

$$\begin{aligned}
\mathbb{E}_{X_0\sim P_0}\left[(\|X_0\|_H - R)_+^2\right] &= 2\int_R^\infty (u-R)P_0(\|X_0\|_H > u)\,\mathrm{d}u \\
&\le 2M\int_R^\infty u\exp(-au^2)\,\mathrm{d}u = \frac{M}{a}\exp(-aR^2).
\end{aligned}$$

In particular,

$$\mathbb{E}_{X_0\sim P_0}\left[(\|X_0\|_H - R_n)_+^2\right] \lesssim n^{-2}.$$

□

Combining the preceding estimates gives the following decomposition.

Lemma H.2 (Reformulation of Lemma 5.2)**.** *Under the standing assumptions, it holds that*

$$\mathbb{E}_n\left[\mathcal{W}_1^2\left(P_0, \hat{Q}^{\mathrm{proj}}_{\overline{T}-\underline{T}}\right)\right] \lesssim \underline{T} + d_{\mathrm{out}}^{1-2/r_0} + n^{-2} + \log n\left[\overline{\beta}(\overline{T}-\underline{T})\mathbb{E}_n\left[\bar{R}_{\underline{T},\overline{T};C^{-1}}(\hat{s}, s_{d_{\mathrm{out}}})\right] + K_{\mathrm{init}}\right].$$

Proof. Conditionally on the training data, the triangle inequality gives

$$\mathcal{W}_1^2\left(P_0, \hat{Q}^{\mathrm{proj}}_{\overline{T}-\underline{T}}\right) \le 2\mathcal{W}_1^2(P_0, \tilde{P}_{\underline{T}}) + 2\mathcal{W}_1^2\left(\tilde{P}_{\underline{T}}, \hat{Q}^{\mathrm{proj}}_{\overline{T}-\underline{T}}\right).$$

Apply the early-stopping and coordinate-truncation bounds to the first term, and Lemma H.1 together with the Girsanov bound to the second term. Taking expectation over the training data proves the assertion. □

For completeness, nonexpansiveness of the Hilbert-ball projection and the triangle inequality also give the following general $\mathcal{W}_2$ estimate.

$$\mathcal{W}_2^2\left(P_0, \hat{Q}^{\mathrm{proj}}_{\overline{T}-\underline{T}}\right) \le 2\mathcal{W}_2^2\left(P_0, \hat{Q}_{\overline{T}-\underline{T}}\right) + 2\mathbb{E}_{X_0\sim P_0}\left[(\|X_0\|_H - R_n)_+^2\right].$$

Therefore,

$$\mathbb{E}_n[\mathcal{W}_2^2(P_0, \hat{Q}^{\mathrm{proj}}_{\overline{T}-\underline{T}})] \le 2\mathbb{E}_n[\mathcal{W}_2^2(P_0, \hat{Q}_{\overline{T}-\underline{T}})] + \mathcal{O}(n^{-2}).$$

Applied to the modified estimator in Section H.2, this inequality transfers (31) to its terminally projected law, with an additional $\mathcal{O}(n^{-2})$ term.

Population denoising risk and loss contrasts. Throughout the oracle argument, write

$$C = C_{[1:d_{\mathrm{out}}]}, \qquad A \in \{I_{d_{\mathrm{out}}}, C^{-1}\}, \qquad \Delta T := \overline{T} - \underline{T}.$$

Assume $0 < \underline{T} < \overline{T}$. The conditional denoising target is

$$h_t(x, x_0) := 2C\nabla_x \log \tilde{p}_t(x \mid x_0) = -\frac{2}{\sigma_t^2}\left(x - m_t x_{0,[1:d_{\mathrm{out}}]}\right).$$

Under the conditional representation

$$\tilde{X}_t = m_t x_{0,[1:d_{\mathrm{out}}]} + \sigma_t C^{1/2} Z, \qquad Z \sim \mathcal{N}(0, I_{d_{\mathrm{out}}}),$$

we have

$$h_t(\tilde{X}_t, x_0) = -\frac{2}{\sigma_t} C^{1/2} Z.$$

Define the integrated loss contrast

$$\ell_{A,f}(x_0) := \frac{1}{\Delta T}\int_{\underline{T}}^{\overline{T}} \mathbb{E}_{\tilde{X}_t \mid X_0 = x_0}\left[\|f(\tilde{X}_t, t)\|_A^2 - 2\langle f(\tilde{X}_t, t), h_t(\tilde{X}_t, x_0)\rangle_A\right] \mathrm{d}t.$$

Removing the target-square term does not change the empirical minimizer. Thus the estimator associated with A minimizes

$$\frac{1}{n}\sum_{i=1}^n \ell_{A,f}(X_{0,i}) \qquad \text{over } f \in \mathcal{S}.$$

Let

$$R_A(f) := \bar{R}_{\underline{T},\overline{T};A}(f, s_{d_{\mathrm{out}}}).$$

The conditional-score identity gives

$$\mathbb{E}\left[h_t(\tilde{X}_t, X_0) \mid \tilde{X}_t\right] = s_{d_{\mathrm{out}}}(\tilde{X}_t, t).$$

Consequently,

$$\mathbb{E}\left[\ell_{A,f}(X_0) - \ell_{A,s_{d_{\mathrm{out}}}}(X_0)\right] = R_A(f).$$

The expectation here is over the initial observation. The time and perturbation variables have already been integrated within $\ell_{A,f}$.

For the variance analysis, set

$$g_f(x_0) := \ell_{A,f}(x_0) - \ell_{A,s_{d_{\text{out}}}}(x_0),$$

and

$$r_f(x_0) := \frac{1}{\Delta T}\int_{\underline{T}}^{\overline{T}} \mathbb{E}_{\tilde{X}_t|X_0=x_0}\left[\|f(\tilde{X}_t,t) - s_{d_{\text{out}}}(\tilde{X}_t,t)\|_A^2\right] \mathrm{d}t.$$

Then $\mathbb{E}g_f(X_0) = \mathbb{E}r_f(X_0) = R_A(f)$, although $g_f(x_0)$ and $r_f(x_0)$ need not coincide pointwise.

An oracle inequality. Write the marginal score as

$$s_{d_{\text{out}}}(x,t) = -D_t x + v_t(x), \qquad D_t := 2CC^{-1}_{t,[1:d_{\text{out}}]},$$

where v_t is the relative-score component. Covariance comparability implies that the diagonal operators D_t are uniformly bounded in both the Euclidean and C^{-1}-weighted geometries.

Let

$$\tilde{\rho}_t := \frac{\mathrm{d}\tilde{P}_t}{\mathrm{d}\mu_{C_{t,[1:d_{\text{out}}]}}}, \qquad v_t(x) := 2C\nabla \log \tilde{\rho}_t(x).$$

By the conditional-score identity and the diagonal Gaussian reference structure,

$$v_t^i(X_{t,[1:d_{\text{out}}]}) = \mathbb{E}_{P_t}\left[2c_i\partial_{x_i}\log \rho_t(X_t) \,\middle|\, X_{t,[1:d_{\text{out}}]}\right].$$

The coordinatewise relative-score bound therefore implies

$$|v_t^i(x)| \lesssim \sqrt{c_i}\, i^{-1/2-\delta_0}.$$

Consequently, for the corresponding versions of these functions,

$$G_A^2 := \sup_{\substack{x\in\mathbb{R}^{d_{\text{out}}}\\ t\in[\underline{T},\overline{T}]}} \|v_t(x)\|_A^2 \lesssim 1, \qquad A \in \{I_{d_{\text{out}}}, C^{-1}\}.$$

Define

$$J_2 := \frac{1}{\Delta T}\int_{\underline{T}}^{\overline{T}} \frac{\mathrm{d}t}{\sigma_t^2}, \qquad H_A := 2\sqrt{\mathrm{Tr}(AC)J_2}.$$

In particular,

$$\frac{1}{\Delta T}\int_{\underline{T}}^{\overline{T}} \mathbb{E}\left[\|h_t(\tilde{X}_t,x_0)\|_A^2 \mid X_0 = x_0\right] \mathrm{d}t = H_A^2.$$

Proposition H.3 (Oracle inequality)**.** *Assume Assumptions 2.1, 2.2, and 4.2. Let $0 < \underline{T} < \overline{T}$ and assume $J_2 < \infty$. Use the finite-dimensional marginal score decomposition and the uniform relative-score bounds stated above.*

Let $X_{0,1},\dots,X_{0,n}$ be independent observations with common law P_0. Let $\hat{s}_A$ be a measurable empirical risk minimizer over the clipped class $\mathcal{S}$ for the integrated denoising loss associated with $A \in \{I_{d_{\text{out}}}, C^{-1}\}$. All distributional constants and clipping constants are fixed independently of n and d_{out}.

Set

$$F_A := \sup_{f\in\mathcal{S}}\sup_{x,t}\|f(x,t)\|_A.$$

For $U, V \geq 1$, define

$$R_x := \sqrt{K}\,U + V, \qquad D_{U,V} := [-R_x, R_x]^{d_{\text{out}}} \times [\underline{T},\overline{T}],$$

where K is the covariance comparison constant, and set

$$p_0(U) := \min\{1, 3d_{\text{out}}e^{-U^2/2}\}, \qquad p_Z(V) := \min\{1, 2d_{\text{out}}e^{-V^2/2}\}.$$

Write

$$B_A := 1 + F_A^2 + G_A^2 + \mathrm{Tr}(AC)(1 + U^2 + J_2),$$

and, for $0 < \delta \le 1$,

$$\mathcal{N}_A := \mathcal{N}\left(\mathcal{S}, \delta, d_{A,D_{U,V}}\right).$$

Then

$$\mathbb{E}_n[R_A(\hat{s}_A)] \le 3 \inf_{f \in \mathcal{S}} R_A(f) + C'\left[\frac{B_A \log(2\mathcal{N}_A)}{n} + (F_A + H_A)\delta + B_A\left(\sqrt{p_0(U)} + \sqrt{p_Z(V)}\right)\right].$$

The constant C' is independent of n, d_{out}, $\underline{T}$, and $\overline{T}$, with dependence on the fixed distributional constants allowed.

Proof. Let $P = P_0$, and let P_n be the empirical measure of the initial samples. We divide the argument into four steps.

Step 1: Conditional second moments and localization. Define

$$E_U := \left\{|X_{0,i}| \le \sqrt{c_{0,i}}\, U \text{ for all } i \le d_{\mathrm{out}}\right\}.$$

Gaussian domination and a union bound give

$$P(E_U^c) \le p_0(U).$$

For $x_0 \in E_U$,

$$\|x_{0,[1:d_{\mathrm{out}}]}\|_A^2 \le KU^2\,\mathrm{Tr}(AC).$$

The conditional Gaussian representation and the score decomposition imply, for arbitrary x_0,

$$\frac{1}{\Delta T}\int_{\underline{T}}^{\overline{T}} \mathbb{E}\left[\|s_{d_{\mathrm{out}}}(\tilde{X}_t, t)\|_A^2 \mid X_0 = x_0\right] \mathrm{d}t \le C'\left[\|x_{0,[1:d_{\mathrm{out}}]}\|_A^2 + \mathrm{Tr}(AC) + G_A^2\right].$$

Consequently, on E_U,

$$r_f(x_0) \le C' B_A \qquad (f \in \mathcal{S}).$$

Expanding the loss difference gives

$$g_f(x_0) = \mathbb{E}_{t,\tilde{X}_t|x_0}\left[\langle f - s_{d_{\mathrm{out}}}, f + s_{d_{\mathrm{out}}} - 2h_t\rangle_A\right],$$

where time is uniform on $[\underline{T}, \overline{T}]$. Cauchy–Schwarz over time and perturbations yields

$$|g_f(x_0)|^2 \le r_f(x_0)\,\mathbb{E}_{t,\tilde{X}_t|x_0}\|f + s_{d_{\mathrm{out}}} - 2h_t\|_A^2.$$

The second factor is at most $C'B_A$ on E_U. Therefore, for

$$g_f^U := g_f \mathbf{1}_{E_U},$$

we obtain

$$|g_f^U| \le C' B_A, \qquad P[(g_f^U)^2] \le C' B_A R_A(f).$$

This is the variance bound used below.
We also need to compare the truncated and original means. For arbitrary x_0, the same moment estimates give

$$|g_f(x_0)| \le C'\left[B_A + \|x_{0,[1:d_{\mathrm{out}}]}\|_A^2\right].$$

Gaussian domination and covariance comparability imply

$$\mathbb{E}[\|X_{0,[1:d_{\mathrm{out}}]}\|_A^4] \le C'\{\mathrm{Tr}(AC)\}^2.$$

Hence, uniformly over $f \in \mathcal{S}$,

$$|Pg_f^U - R_A(f)| \le C' B_A \sqrt{p_0(U)}.$$

Step 2: Approximation by a finite cover. Choose a δ-net $\{f_1, \dots, f_N\} \subset \mathcal{S}$ on $D_{U,V}$ in the metric $d_{A,D_{U,V}}$, where $N = \mathcal{N}_A$. For every $f \in \mathcal{S}$, choose f_j with $d_{A,D_{U,V}}(f, f_j) \le \delta$.
If $x_0 \in E_U$ and $\max_i |Z_i| \le V$, then

$$|\tilde{X}_{t,i}| \le \sqrt{c_i}(\sqrt{K}\,U + V) \le R_x.$$

The complementary Gaussian event has probability at most $p_Z(V)$. Using the output bound and Cauchy–Schwarz on this complementary event gives, uniformly on E_U,

$$|\ell_{A,f} - \ell_{A,f_j}| \le e_A,$$

where

$$e_A := C' \left[(F_A + H_A)\delta + F_A^2 p_Z(V) + F_A H_A \sqrt{p_Z(V)} \right].$$

Since $F_A^2 + H_A^2 \lesssim B_A$,

$$e_A \le C' \left[(F_A + H_A)\delta + B_A \sqrt{p_Z(V)} \right].$$

For all initial observations,

$$|\ell_{A,f}(x_0)| \le M_A := F_A^2 + 2F_A H_A \lesssim B_A.$$

It follows that

$$|R_A(f) - R_A(f_j)| = |P(\ell_{A,f} - \ell_{A,f_j})| \le e_A + 2M_A p_0(U).$$

Moreover,

$$|g_f^U - g_{f_j}^U| \le e_A.$$

Step 3: Bernstein's inequality on the cover. For each net element, Step 1 gives

$$|g_{f_j}^U| \le C' B_A, \qquad \mathrm{Var}(g_{f_j}^U) \le C' B_A R_A(f_j).$$

Bernstein's inequality, a union bound, and Young's inequality imply that, for every $x \ge 0$, with probability at least $1 - e^{-x}$, simultaneously for all $j \le N$,

$$|(P_n - P)g_{f_j}^U| \le \frac{1}{4} R_A(f_j) + \frac{C' B_A(\log(2N) + x)}{n}.$$

Define

$$Z_n := \max_{j \le N} \left(|(P_n - P)g_{f_j}^U| - \frac{1}{4} R_A(f_j) \right)_+.$$

Integration of the preceding tail bound yields

$$\mathbb{E}_n[Z_n] \le \frac{C' B_A \log(2N)}{n}.$$

The cover approximation from Step 2 then gives, for every $f \in \mathcal{S}$,

$$|(P_n - P)g_f^U| \le \frac{1}{4} R_A(f) + Z_n + C' e_A + C' B_A p_0(U).$$

Step 4: The original ERM and the oracle bound. The estimator minimizes the original integrated empirical loss, not its truncated version. For every fixed comparator $f \in \mathcal{S}$,

$$P_n(g_{\hat{s}_A} - g_f) \le 0.$$

Since $g_{\hat{s}_A} - g_f = \ell_{A,\hat{s}_A} - \ell_{A,f}$, the global bound on the contrasts gives

$$P_n g_{\hat{s}_A}^U \le P_n g_f^U + 2M_A P_n(E_U^c).$$

Using the mean-truncation bound from Step 1 and the deviation bound from Step 3 on both sides, we obtain

$$\frac{3}{4} R_A(\hat{s}_A) \le \frac{5}{4} R_A(f) + 2Z_n + C' e_A + C' B_A \sqrt{p_0(U)} + 2M_A P_n(E_U^c).$$

Taking expectations, using $\mathbb{E}_n[P_n(E_U^c)] \le p_0(U)$, and then taking the infimum over $f \in \mathcal{S}$ proves the assertion. □

Dependence on time, dimension, and covering accuracy. Under $\underline{\beta} \le \beta_t \le \overline{\beta}$, $\sigma_t^2 \asymp \min\{t, 1\}$. Therefore,

$$J_2 \lesssim 1 + \frac{\log(\overline{T}/\underline{T})}{\overline{T} - \underline{T}}.$$

In particular, when $\overline{T} \asymp \log n$ and $\underline{T} = n^{-\alpha}$ for a fixed $\alpha > 0$, we have $J_2 = \mathcal{O}(1)$. The variance calculation thus involves a time average of σ_t^{-2}, rather than its maximum $\sigma_{\underline{T}}^{-2}$.

Take

$$U = V = \sqrt{2q \log(end_{\text{out}})}$$

for a sufficiently large fixed constant q. Then

$$\sqrt{p_0(U)} + \sqrt{p_Z(V)} \lesssim d_{\text{out}}^{1/2} (end_{\text{out}})^{-q/2}.$$

The covering number is evaluated on $D_{U,V}$ using Proposition C.9 with

$$R = \max\{1, \sqrt{K}\, U + V, \overline{T}\}.$$

Thus both the spatial localization radius and the time interval enter the entropy bound only logarithmically.

For the clipped class,

$$F_I^2 \lesssim 1 + \log(en), \qquad F_{C^{-1}}^2 \lesssim 1 + d_{\text{out}} \log(en),$$

and $G_I^2, G_{C^{-1}}^2 \lesssim 1$. Since

$$\operatorname{Tr}(IC) \le \operatorname{Tr}(C_{\text{full}}), \qquad \operatorname{Tr}(C^{-1}C) = d_{\text{out}},$$

the variance coefficients satisfy, when $J_2 = \mathcal{O}(1)$,

$$B_I \lesssim 1 + \log(end_{\text{out}}), \qquad B_{C^{-1}} \lesssim 1 + d_{\text{out}} \log(end_{\text{out}}).$$

Here C_{full} denotes the covariance operator on H.

For instance, the choice

$$\delta := \frac{n^{-3}}{1 + F_A + H_A}$$

makes the covering approximation term at most n^{-3}. When the truncation dimension grows polynomially in n, a sufficiently large fixed q also makes the tail remainder negligible at any prescribed fixed polynomial order.

Consequently, the statistical contribution has the form

$$\frac{B_A}{n} \log \mathcal{N}\left(\mathcal{S}, \delta, d_{A, D_{U,V}}\right),$$

rather than a square-root empirical-process bound. In particular, up to logarithmic factors, the coefficient outside the entropy is dimension independent for $A = I$ and proportional to d_{out} for $A = C^{-1}$.

Scale of F_A. We evaluate the output envelope of the clipped hypothesis class $\mathcal{S}$ defined in Section 2. For a positive definite matrix A, write

$$F_A := \sup_{f \in \mathcal{S}} \sup_{\substack{x \in \mathbb{R}^{d_{\text{out}}} \\ t \in [\underline{T}, \overline{T}]}} \|f(x,t)\|_A.$$

Every $f = \phi_{\text{score,grad}} + \phi_{\text{score,linear}} \in \mathcal{S}$ satisfies

$$|\phi^i_{\text{score,grad}}(x,t)| \le R_{\text{grad}} \sqrt{c_i}\, i^{-1/2-\delta_0} \quad (i \le d_{\max}),$$

with $\phi^i_{\text{score,grad}} = 0$ for $i > d_{\max}$, and

$$|\phi^i_{\text{score,linear}}(x,t)| \le R_{\text{linear}} \sqrt{c_i}\, u_n \quad (i \le d_{\text{out}}), \qquad u_n := \sqrt{\log(en)}.$$

These are constraints on every candidate in the hypothesis class, rather than bounds for a particular approximation network.

Using $\|a+b\|_A^2 \le 2\|a\|_A^2 + 2\|b\|_A^2$, we obtain

$$\begin{aligned} F_I^2 &\le 2R_{\text{grad}}^2 \sum_{i=1}^{d_{\max}} c_i i^{-1-2\delta_0} + 2R_{\text{linear}}^2 u_n^2 \sum_{i=1}^{d_{\text{out}}} c_i \\ &\lesssim 1 + u_n^2. \end{aligned}$$

Similarly, with $C = C_{[1:d_{\text{out}}]}$,

$$\begin{aligned} F_{C^{-1}}^2 &\le 2R_{\text{grad}}^2 \sum_{i=1}^{d_{\max}} i^{-1-2\delta_0} + 2R_{\text{linear}}^2 d_{\text{out}} u_n^2 \\ &\lesssim 1 + d_{\text{out}} u_n^2. \end{aligned}$$

Consequently,

$$F_I^2 \lesssim 1 + \log(en), \qquad F_{C^{-1}}^2 \lesssim 1 + d_{\text{out}} \log(en).$$

The implicit constants are independent of n and d_{out}. These bounds control the network outputs, not the entire denoising loss, whose conditional-score target is unbounded.

Proof of each case. We use the structured clipped class $\mathcal{S}$ rather than the full class of networks with the single-network budgets in the approximation theorem. As explained in Appendix D, representing the two branches as a single sequential network requires additional identity layers for depth alignment. These fixed layers contribute to the single-network sparsity, but do not introduce additional free parameters.

Accordingly, the covering numbers below are obtained by covering the underlying gradient and linear branches separately and then applying their fixed output clipping, zero-padding, and addition. The internal clipping operations used in constructing a gradient approximator are included in the gradient-branch budgets.

The following calculations use the branchwise approximation bounds in Lemma D.1 and the stated budgets of the underlying gradient and linear networks. Throughout the finite-dimensional calculations below, $C = C_{[1:d_{\text{out}}]}$ and $s_{d_{\text{out}}}(x,t) = 2C\nabla \log \tilde{p}_t(x)$.

For Cases (i) and (ii), take

$$\underline{T} = n^{-2}, \qquad \overline{T} = L_T \log n,$$

where $L_T > 0$ is a sufficiently large fixed constant. Under $\underline{\beta} \le \beta_t \le \overline{\beta}$,

$$J_2 = \frac{1}{\overline{T} - \underline{T}} \int_{\underline{T}}^{\overline{T}} \sigma_t^{-2}\, \mathrm{d}t \lesssim 1.$$

All subsequent asymptotic statements are for sufficiently large n.

In Proposition H.3, choose

$$U = V = \sqrt{24 \log(e n d_{\text{out}})}, \qquad \delta_A := \frac{n^{-4}}{1 + F_A + H_A},$$

and write

$$\mathcal{N}_A := \mathcal{N}\left(\mathcal{S}, \delta_A, d_{A, D_{U,V}}\right).$$

For the dimension choices below, $d_{\text{out}} \lesssim n$. The envelope estimates and $J_2 \lesssim 1$ give

$$B_I \lesssim \log n, \qquad B_{C^{-1}} \lesssim d_{\text{out}} \log n.$$

Moreover,

$$\sqrt{p_0(U)} + \sqrt{p_Z(V)} \lesssim d_{\text{out}}^{1/2} (e n d_{\text{out}})^{-6}.$$

Consequently, the covering and tail remainders in Proposition H.3 are $\mathcal{O}(n^{-3})$ for both choices of A.

Choose the fixed linear clipping constant so that Lemma D.1 holds with $p = 4$. The additional clipping errors are then $\mathcal{O}(n^{-4})$ in the unweighted geometry and $\mathcal{O}(d_{\text{out}} n^{-4})$ in the weighted geometry.

We also evaluate the initialization error. By convexity of relative entropy in its first argument,

$$\begin{aligned} K_{\mathrm{init}} &= \mathrm{KL}\left(\tilde{P}_{\overline{T}} \,\middle\|\, \mathcal{N}(0, C)\right) \\ &\le \mathbb{E}_{X_0}[\mathrm{KL}\left(\mathcal{N}\left(m_{\overline{T}} X_{0,[1:d_{\mathrm{out}}]}, {\sigma_{\overline{T}}}^2 C\right) \,\middle\|\, \mathcal{N}(0, C)\right)] \\ &= \frac{1}{2}\left[m_{\overline{T}}^2 \mathbb{E}[\|X_{0,[1:d_{\mathrm{out}}]}\|_{C^{-1}}^2] + d_{\mathrm{out}}\left({\sigma_{\overline{T}}}^2 - 1 - \log {\sigma_{\overline{T}}}^2\right)\right]. \end{aligned}$$

Gaussian domination and covariance comparability imply

$$\mathbb{E}[\|X_{0,[1:d_{\mathrm{out}}]}\|_{C^{-1}}^2] \lesssim d_{\mathrm{out}}.$$

For $m_{\overline{T}}^2 \le 1/2$, it follows that

$$K_{\mathrm{init}} \lesssim d_{\mathrm{out}} m_{\overline{T}}^2 \lesssim d_{\mathrm{out}} e^{-2\underline{\beta}\overline{T}},$$

where $\int_0^t \beta_s \,\mathrm{d}s \ge \underline{\beta} t$. Taking L_T sufficiently large therefore ensures $K_{\mathrm{init}} = \mathcal{O}(n^{-3})$.

We next make the dependence of the network budgets explicit. Suppose that $\Gamma \asymp \log n$, d_{out} grows polynomially in n, and the retained expansion satisfies

$$d_{\max} \lesssim \Gamma^{1/\eta}, \qquad f_{\max} \lesssim \Gamma, \qquad \log(eG) \lesssim \Gamma^{\max\{1,1/\eta\}}.$$

These bounds concern the actual retained expansion.

For the density and derivative constructions, the chosen factor accuracies satisfy

$$\log \varepsilon^{-1} \lesssim \Gamma + d_{\max} + \log(eG) + \log(1 + \sigma_{\underline{T}}^{-1}).$$

Here the density-ratio norm in (9) is bounded above and away from zero under the standing assumptions. With $\underline{T} = n^{-2}$, $\overline{T} = L_T \log n$, and $\sigma_t^2 \asymp \min\{t, 1\}$, it follows that

$$\log \varepsilon^{-1} \lesssim \Gamma^{\max\{1,1/\eta\}}.$$

Substituting the actual-frequency bound $|l| \le 2^{f_{\max}}$ into (8) therefore gives

$$K_\rho + K' \lesssim \left(2^{f_{\max}} + \Gamma^{\max\{1,1/\eta\}}\right) \Gamma^{\max\{1,1/\eta\}} \log(e\Gamma).$$

The time-endpoint terms are absorbed by this bound, because they are $\mathcal{O}(\log n)$.

For both choices $A \in \{I_{d_{\mathrm{out}}}, C^{-1}\}$, the covering accuracy and localization parameters chosen above satisfy

$$\log \delta_A^{-1} \lesssim \log n, \qquad \log(1 + \kappa_A) \lesssim \log n, \qquad \log(R + 1) \lesssim \log\log n.$$

Covariance decay also gives

$$\max_{i \le d_{\max}} \log(c_{0,i}^{-1/2}) \lesssim \log(e + d_{\max}).$$

Consequently, (21) implies

$$\begin{aligned} \log(2\mathcal{N}_A) &\lesssim d_{\max} G (K_\rho + K')^8 + d_{\mathrm{out}} \Gamma^6 \log n \\ &\lesssim G\left(2^{f_{\max}} + \Gamma^{\max\{1,1/\eta\}}\right)^8 \Gamma^{1/\eta + 8\max\{1,1/\eta\}} [\log(e\Gamma)]^8 + d_{\mathrm{out}} \Gamma^6 \log n. \end{aligned} \tag{18}$$

We now substitute the counting and frequency bounds for the retained expansion. Since

$$f_{\max} = \left\lceil \frac{\Gamma}{a_1} \right\rceil - 1,$$

we have, for sufficiently large Γ,

$$\left(2^{f_{\max}} + \Gamma^{\max\{1,1/\eta\}}\right)^8 \lesssim 2^{8\Gamma/a_1}.$$

Using the counting bounds for mixed and anisotropic smoothness, define

$$\kappa := \begin{cases} 9, & \text{for mixed smoothness,} \\ \log_2 3 + 8a^\dagger/a_1, & \text{for anisotropic smoothness.} \end{cases}$$

Then (18) yields

$$\log(2\mathcal{N}_A) \lesssim 2^{\kappa\Gamma/a^\dagger}\Gamma^{1/\eta+8\max\{1,1/\eta\}}[\log(e\Gamma)]^8 + d_{\text{out}}\Gamma^6\log n, \quad A \in \{I_{d_{\text{out}}}, C^{-1}\}. \tag{19}$$

To balance the approximation and gradient-branch statistical terms, including their logarithmic factors, choose

$$\Gamma := \frac{a^\dagger}{2(a^\dagger - v) + \kappa}\Big[\log_2 n - \big(1/\eta + 8\max\{1,1/\eta\}\big)\log_2(\log n) - 8\log_2(\log\log n)\Big]. \tag{20}$$

For sufficiently large n, this choice satisfies $\Gamma \geq 2$ and $\Gamma \asymp \log n$. Write

$$\nu := \frac{2(a^\dagger - v)}{2(a^\dagger - v) + \kappa}.$$

Then

$$2^{-2(1-v/a^\dagger)\Gamma} = n^{-\nu}(\log n)^{(1/\eta+8\max\{1,1/\eta\})\nu}(\log\log n)^{8\nu},$$
$$\frac{2^{\kappa\Gamma/a^\dagger}}{n} = n^{-\nu}(\log n)^{-(1/\eta+8\max\{1,1/\eta\})(1-\nu)}(\log\log n)^{-8(1-\nu)}.$$

All subsequent bounds use this choice of Γ.

Case (i): When $\hat{s}$ is obtained by Option 1.

By Proposition H.3, we obtain

$$\mathbb{E}_n\left[\bar{R}_{\underline{T},\overline{T};I}(\hat{s}, s_{d_{\text{out}}})\right] \leq 3\inf_{f\in\mathcal{S}}\bar{R}_{\underline{T},\overline{T};I}(f, s_{d_{\text{out}}}) + C'\frac{B_I\log(2\mathcal{N}_I)}{n} + \mathcal{O}(n^{-3}).$$

Since $c_i \simeq i^{-2/r_0}$,

$$\|z\|^2_{C^{-1}} \lesssim d_{\text{out}}^{2/r_0}\|z\|^2 \qquad (z \in \mathbb{R}^{d_{\text{out}}}).$$

Therefore,

$$\mathbb{E}_n\left[\bar{R}_{\underline{T},\overline{T};C^{-1}}(\hat{s}, s_{d_{\text{out}}})\right] \lesssim d_{\text{out}}^{2/r_0}\left[\inf_{f\in\mathcal{S}}\bar{R}_{\underline{T},\overline{T};I}(f, s_{d_{\text{out}}}) + \frac{B_I\log(2\mathcal{N}_I)}{n} + n^{-3}\right].$$

The approximation term is bounded by Lemma D.1:

$$d_{\text{out}}^{2/r_0}\inf_{f\in\mathcal{S}}\bar{R}_{\underline{T},\overline{T};I}(f, s_{d_{\text{out}}}) \lesssim d_{\text{out}}^{2/r_0}\left[\Gamma J_2 2^{-2(1-v/a^\dagger)\Gamma} + n^{-4}\right] \lesssim d_{\text{out}}^{2/r_0}\left[\Gamma 2^{-2(1-v/a^\dagger)\Gamma} + n^{-4}\right].$$

Next, we evaluate the entropy term using Proposition C.9 and the branch budgets in Appendix D. We cover the underlying gradient and linear branches separately and then apply their fixed output clipping, zero-padding, and addition.

Write

$$R := \max\{1, \sqrt{K}\,U + V, \overline{T}\}, \qquad \kappa_A := (\operatorname{Tr} A)^{1/2}.$$

For each branch, a coordinatewise uniform cover at accuracy $\delta_A/[2(1+\kappa_A)]$ induces, after combination, a cover of $\mathcal{S}$ at accuracy δ_A in $d_{A,D_{U,V}}$.

The gradient-branch budgets satisfy

$$L_{\text{grad}} \lesssim K_\rho^2 + K'^2,$$
$$W_{\text{grad}} \lesssim d_{\max}G(K_\rho^3 + K'^4),$$
$$S_{\text{grad}} \lesssim d_{\max}G(K_\rho^4 + K'^5),$$
$$\log(B_{\text{grad}} \vee e) \lesssim K_\rho + K' + \max_{i\leq d_{\max}}\log(c_{0,i}^{-1/2}).$$

The linear-branch budgets satisfy

$$L_{\mathrm{linear}} \lesssim \Gamma^2, \qquad W_{\mathrm{linear}} \lesssim d_{\mathrm{out}}\Gamma^3,$$
$$S_{\mathrm{linear}} \lesssim d_{\mathrm{out}}\Gamma^4, \qquad \log(B_{\mathrm{linear}} \vee e) \lesssim \Gamma + \log(e + d_{\mathrm{out}}),$$

where covariance decay is used in the last inequality.

Applying the covering-number bound to each branch and adding their logarithmic covering numbers gives, for $A \in \{I_{d_{\mathrm{out}}}, C^{-1}\}$,

$$\begin{aligned}\log(2\mathcal{N}_A) &\lesssim d_{\max} G (K_\rho^4 + K'^5)(K_\rho^2 + K'^2) \\ &\quad \times \Big[K_\rho + K' + \log \delta_A^{-1} + \log(e + d_{\mathrm{out}} + G) + \log(R+1) \\ &\qquad + \log(1+\kappa_A) + \max_{i \le d_{\max}} \log(c_{0,i}^{-1/2})\Big] \\ &\quad + d_{\mathrm{out}}\Gamma^6 \Big[\Gamma + \log \delta_A^{-1} + \log(e + d_{\mathrm{out}}) + \log(R+1) + \log(1+\kappa_A)\Big]. \end{aligned} \tag{21}$$

Here, G, $d_{\max}$, K_ρ, and K' correspond to the actual retained expansion and the approximation accuracies used in the branch constructions. The actual-frequency and time-endpoint dependencies in K_ρ and K' are retained.

The identity layers used to represent the two branches as a single sequential network do not introduce additional free parameters into this branchwise covering argument.

For Option 1, $\kappa_I = \sqrt{d_{\mathrm{out}}}$. Using $B_I \lesssim \log n$, the statistical contribution after conversion to the weighted score geometry satisfies

$$d_{\mathrm{out}}^{2/r_0} \frac{B_I \log(2\mathcal{N}_I)}{n} \lesssim \frac{d_{\mathrm{out}}^{2/r_0} \log n}{n} \log(2\mathcal{N}_I),$$

where $\log(2\mathcal{N}_I)$ is evaluated by (21).

Combining Proposition H.3, Lemma D.1, and (19), we obtain

$$\begin{aligned}&\mathbb{E}_n\left[\bar{R}_{\underline{T},\overline{T};C^{-1}}(\hat{s}, s_{d_{\mathrm{out}}})\right] \\ &\lesssim d_{\mathrm{out}}^{2/r_0}\left[\Gamma\, 2^{-2(1-v/a^\dagger)\Gamma} + \frac{2^{\kappa\Gamma/a^\dagger}}{n}\Gamma^{1/\eta+8\max\{1,1/\eta\}}[\log(e\Gamma)]^8 \log n + \frac{d_{\mathrm{out}}\Gamma^6(\log n)^2}{n} + n^{-3}\right].\end{aligned}$$

Here d_{out}^{2/r_0} converts the unweighted score risk into the C^{-1}-weighted risk.

By Lemma H.2, the specified time endpoints, and the initialization estimate,

$$\begin{aligned}&\mathbb{E}_n\left[\mathcal{W}_1^2\left(P_0, \hat{Q}^{\mathrm{proj}}_{\overline{T}-\underline{T}}\right)\right] \\ &\lesssim d_{\mathrm{out}}^{1-2/r_0} + n^{-2} + (\log n)^2 \mathbb{E}_n\left[\bar{R}_{\underline{T},\overline{T};C^{-1}}(\hat{s}, s_{d_{\mathrm{out}}})\right] \\ &\lesssim d_{\mathrm{out}}^{1-2/r_0} + d_{\mathrm{out}}^{2/r_0}\Gamma 2^{-2(1-v/a^\dagger)\Gamma}(\log n)^2 + \frac{d_{\mathrm{out}}^{2/r_0}}{n} 2^{\kappa\Gamma/a^\dagger}\Gamma^{1/\eta+8\max\{1,1/\eta\}}[\log(e\Gamma)]^8(\log n)^3 \\ &\quad + \frac{d_{\mathrm{out}}^{1+2/r_0}\Gamma^6(\log n)^4}{n} + d_{\mathrm{out}}^{2/r_0} n^{-3}(\log n)^2 + n^{-2}.\end{aligned}$$

The factor $(\log n)^2$ multiplying the score risk comes from time integration and terminal projection.

Substituting (20) gives

$$\begin{aligned}&\mathbb{E}_n\left[\mathcal{W}_1^2\left(P_0, \hat{Q}^{\mathrm{proj}}_{\overline{T}-\underline{T}}\right)\right] \\ &\lesssim d_{\mathrm{out}}^{1-2/r_0} + d_{\mathrm{out}}^{2/r_0} n^{-\nu}(\log n)^{3+(1/\eta+8\max\{1,1/\eta\})\nu}(\log\log n)^{8\nu} \\ &\quad + \frac{d_{\mathrm{out}}^{1+2/r_0}(\log n)^{10}}{n} + d_{\mathrm{out}}^{2/r_0} n^{-3}(\log n)^2 + n^{-2}.\end{aligned}$$

Choose

$$d_{\text{out}} := \left\lceil \min\left\{ n^{\frac{r_0\nu}{4-r_0}} (\log n)^{-\frac{r_0}{4-r_0}[3+(1/\eta+8\max\{1,1/\eta\})\nu]} (\log\log n)^{-\frac{8r_0\nu}{4-r_0}}, n^{r_0/4} (\log n)^{-10r_0/4} \right\} \right\rceil.$$

Both quantities inside the minimum grow polynomially in n, up to logarithmic factors. Hence $d_{\text{out}} \geq d_{\max}$ for sufficiently large n.

The first quantity balances coordinate truncation with the approximation and gradient-branch statistical terms. The second balances coordinate truncation with the linear-branch statistical term. Since $d_{\text{out}} \lesssim n^{r_0/4}$, the remaining $d_{\text{out}}^{2/r_0} n^{-3} (\log n)^2$ term is negligible. Consequently,

$$\begin{aligned} &\mathbb{E}_n \left[\mathcal{W}_1^2 \left(P_0, \hat{Q}^{\text{proj}}_{\overline{T}-\underline{T}} \right) \right] \\ &\quad = \mathcal{O}\Bigg(n^{-\frac{(2-r_0)\nu}{4-r_0}} (\log n)^{\frac{2-r_0}{4-r_0}[3+(1/\eta+8\max\{1,1/\eta\})\nu]} (\log\log n)^{\frac{8(2-r_0)\nu}{4-r_0}} \\ &\qquad\qquad + n^{-\frac{2-r_0}{4}} (\log n)^{\frac{10(2-r_0)}{4}} \Bigg). \end{aligned} \tag{22}$$

Thus the polynomial convergence exponent is

$$\min\left\{ \frac{2-r_0}{4-r_0} \frac{2(a^\dagger - v)}{2(a^\dagger - v)+\kappa}, \frac{2-r_0}{4} \right\},$$

with the logarithmic factors retained explicitly in (22).

Case (ii): When $\hat{s}$ is obtained by Option 2.

By Proposition H.3, we obtain

$$\mathbb{E}_n \left[\bar{R}_{\underline{T},\overline{T};C^{-1}}(\hat{s}, s_{d_{\text{out}}}) \right] \leq 3 \inf_{f \in \mathcal{S}} \bar{R}_{\underline{T},\overline{T};C^{-1}}(f, s_{d_{\text{out}}}) + C' \frac{B_{C^{-1}} \log(2\mathcal{N}_{C^{-1}})}{n} + \mathcal{O}(n^{-3}).$$

The approximation term is bounded by Lemma D.1:

$$\inf_{f \in \mathcal{S}} \bar{R}_{\underline{T},\overline{T};C^{-1}}(f, s_{d_{\text{out}}}) \lesssim d_{\text{out}} \left[\Gamma J_2 2^{-2(1-v/a^\dagger)\Gamma} + n^{-4} \right] \lesssim d_{\text{out}} \left[\Gamma 2^{-2(1-v/a^\dagger)\Gamma} + n^{-4} \right].$$

For the entropy term, the weighted covering metric satisfies

$$d_{C^{-1},D}(f,g) \leq \left(\sum_{i=1}^{d_{\text{out}}} c_i^{-1} \right)^{1/2} d_{\infty,D}(f,g).$$

By covariance decay,

$$\kappa_{C^{-1}} = \left(\sum_{i=1}^{d_{\text{out}}} c_i^{-1} \right)^{1/2} \lesssim d_{\text{out}}^{1/2+1/r_0},$$

and hence

$$\log(1+\kappa_{C^{-1}}) \lesssim \log(e + d_{\text{out}}).$$

Applying (21) with $A = C^{-1}$ yields

$$\begin{aligned} \log(2\mathcal{N}_{C^{-1}}) &\lesssim d_{\max} G (K_\rho^4 + K'^5)(K_\rho^2 + K'^2) \\ &\quad \times \Big[K_\rho + K' + \log \delta_{C^{-1}}^{-1} + \log(e + d_{\text{out}} + G) + \log(R+1) \\ &\qquad + \max_{i \leq d_{\max}} \log(c_{0,i}^{-1/2}) \Big] \\ &\quad + d_{\text{out}} \Gamma^6 \Big[\Gamma + \log \delta_{C^{-1}}^{-1} + \log(e + d_{\text{out}}) + \log(R+1) \Big]. \end{aligned}$$

The additional change of covering accuracy required by the weighted norm enters only logarithmically.

Using $B_{C^{-1}} \lesssim d_{\text{out}} \log n$, the statistical contribution satisfies

$$\frac{B_{C^{-1}} \log(2\mathcal{N}_{C^{-1}})}{n} \lesssim \frac{d_{\text{out}} \log n}{n} \log(2\mathcal{N}_{C^{-1}}).$$

The entropy is evaluated with the branch budgets, retaining the actual-frequency dependence in K_ρ and K'.

Combining Proposition H.3, Lemma D.1, and (19), we obtain

$$\begin{aligned}
&\mathbb{E}_n\left[\bar{R}_{\underline{T},\overline{T};C^{-1}}(\hat{s}, s_{d_{\text{out}}})\right] \\
&\lesssim d_{\text{out}}\Gamma\, 2^{-2(1-v/a^\dagger)\Gamma} + \frac{d_{\text{out}}}{n} 2^{\kappa\Gamma/a^\dagger}\Gamma^{1/\eta+8\max\{1,1/\eta\}}[\log(e\Gamma)]^8 \log n + \frac{d_{\text{out}}^2\Gamma^6(\log n)^2}{n} + n^{-3}.
\end{aligned}$$

The clipping remainder $d_{\text{out}} n^{-4}$ is absorbed into n^{-3} because $d_{\text{out}} \lesssim n$.

Applying Lemma H.2 yields

$$\begin{aligned}
&\mathbb{E}_n\left[\mathcal{W}_1^2\left(P_0, \hat{Q}^{\text{proj}}_{\overline{T}-\underline{T}}\right)\right] \\
&\lesssim d_{\text{out}}^{1-2/r_0} + n^{-2} + (\log n)^2 \mathbb{E}_n\left[\bar{R}_{\underline{T},\overline{T};C^{-1}}(\hat{s}, s_{d_{\text{out}}})\right] \\
&\lesssim d_{\text{out}}^{1-2/r_0} + d_{\text{out}}\Gamma 2^{-2(1-v/a^\dagger)\Gamma}(\log n)^2 \\
&\quad + \frac{d_{\text{out}}}{n} 2^{\kappa\Gamma/a^\dagger}\Gamma^{1/\eta+8\max\{1,1/\eta\}}[\log(e\Gamma)]^8(\log n)^3 + \frac{d_{\text{out}}^2\Gamma^6(\log n)^4}{n} + n^{-2}.
\end{aligned}$$

Substituting (20) gives

$$\begin{aligned}
&\mathbb{E}_n\left[\mathcal{W}_1^2\left(P_0, \hat{Q}^{\text{proj}}_{\overline{T}-\underline{T}}\right)\right] \\
&\lesssim d_{\text{out}}^{1-2/r_0} + d_{\text{out}} n^{-\nu}(\log n)^{3+(1/\eta+8\max\{1,1/\eta\})\nu}(\log\log n)^{8\nu} + \frac{d_{\text{out}}^2(\log n)^{10}}{n} + n^{-2}.
\end{aligned}$$

Choose

$$d_{\text{out}} := \left\lceil \min\left\{ n^{r_0\nu/2}(\log n)^{-\frac{r_0}{2}[3+(1/\eta+8\max\{1,1/\eta\})\nu]}(\log\log n)^{-4r_0\nu}, n^{\frac{r_0}{2+r_0}}(\log n)^{-\frac{10r_0}{2+r_0}} \right\} \right\rceil.$$

Again, $d_{\text{out}} \geq d_{\max}$ for sufficiently large n. Balancing coordinate truncation with the approximation and statistical terms gives

$$\begin{aligned}
&\mathbb{E}_n\left[\mathcal{W}_1^2\left(P_0, \hat{Q}^{\text{proj}}_{\overline{T}-\underline{T}}\right)\right] \\
&= \mathcal{O}\left(n^{-\frac{(2-r_0)\nu}{2}}(\log n)^{\frac{2-r_0}{2}[3+(1/\eta+8\max\{1,1/\eta\})\nu]}(\log\log n)^{4(2-r_0)\nu} + n^{-\frac{2-r_0}{2+r_0}}(\log n)^{\frac{10(2-r_0)}{2+r_0}} \right).
\end{aligned} \tag{23}$$

Thus the polynomial convergence exponent is

$$\min\left\{ \frac{(2-r_0)(a^\dagger - v)}{2(a^\dagger - v) + \kappa}, \frac{2-r_0}{2+r_0} \right\},$$

with the logarithmic factors retained explicitly in (23).

H.2 Wasserstein-2 bounds

In addition to Cases (i) and (ii) of Theorem 5.1, we consider two $\mathcal{W}_2$ bounds in this appendix. We call the analysis of a modified Option 1 estimator under the Gaussian-corrected Lipschitz condition of Wang & Wang (2024) *Case (iii)*. We call the projection-and-relative-entropy analysis of the original Options 1 and 2 *Case (iv)*. Case (iv) introduces no new estimator and does not use that Wasserstein stability result or its additional regularity assumptions. All scores below use forward time. We continue to write $C = C_{[1:d_{\text{out}}]}$ and $C_t = C_{t,[1:d_{\text{out}}]}$, and identify finite-dimensional laws with their zero-padding embeddings into H. The bounds concern the continuous-time reverse SDE.

H.2.1 Case (iii): estimating the residual after removing the stationary linear score

Let $\mathcal{S}_{\text{raw}}$ be the structured two-branch class used to construct $\mathcal{S}$, with its final gradient and linear output clipping removed. Internal operations in the branch networks, their budgets, and zero-padding are unchanged. Write its elements as F_θ. For $n \geq 2$, choose $B_n = K_{\text{res}}\sqrt{\log(en)}$, where K_{res} is a sufficiently large fixed constant. Using the radial projection Π_R defined in Section 2, set

$$g_\theta(x,t) := \Pi_{m_t B_n}\big(F_\theta(x,t) + 2x\big), \qquad \tilde{s}_\theta(x,t) := g_\theta(x,t) - 2x, \tag{24}$$
$$\mathcal{S}_{\text{res}} := \{\tilde{s}_\theta : F_\theta \in \mathcal{S}_{\text{raw}}\}.$$

Thus the learned residual targets $r_t(x) := s_{d_{\text{out}}}(x,t) + 2x$. The stationary linear score $-2x$ is added outside the clipping. The radius B_n for residual clipping is distinct from the terminal projection radius R_n.

For comparison with an input/output time-scaled representation, define

$$f_\theta(z,t) := m_t^{-1}\big[F_\theta(z/m_t, t) + 2z/m_t\big].$$

Since $m_t > 0$ on the finite time interval, this gives exactly

$$\tilde{s}_\theta(x,t) = -2x + m_t \Pi_{B_n}\big(f_\theta(m_t x, t)\big). \tag{25}$$

Here f_θ is a fixed reparameterization of the raw branch networks, not an additional unrestricted network class. The deterministic time scaling and linear maps introduce no trainable parameters. In particular, the final clipping of the original full-score class is not applied before this transformation.

Define $\hat{s}^{(iii)}$ to be a measurable minimizer of the Option 1 integrated denoising loss over $\mathcal{S}_{\text{res}}$, and let $\hat{Q}^{(iii)}$ denote its reverse-SDE output law. Equivalently, minimize the integrated squared loss for g_θ with target $h_t(x, x_0) + 2x$, where h_t is the conditional score defined above. This is training over the modified class, not postprocessing an empirical minimizer from the original class.

Assumption H.4 (Additional regularity for Case (iii))**.** *In addition to the standing assumptions, the following hold, with constants independent of n and d_{out}.*

1. *The finite-dimensional marginal initial density ratio $\tilde{\rho}_0 = \mathrm{d}\tilde{P}_0/\mathrm{d}\mu_{C_{0,[1:d_{\text{out}}]}}$ is twice continuously differentiable, and*

$$\sup_{d_{\text{out}}} \|\sqrt{C}\nabla \log \tilde{\rho}_0\|_\infty < \infty, \qquad \sup_{d_{\text{out}}} \|C\nabla^2 \log \tilde{\rho}_0\|_\infty < \infty.$$

   The matrix norm here is the Euclidean operator norm.

2. *Almost surely over the training data, the fitted estimator satisfies*

$$\big\|[\hat{s}^{(iii)}(x,t) + 2CC_t^{-1}x] - [\hat{s}^{(iii)}(y,t) + 2CC_t^{-1}y]\big\| \leq L_{\text{est}} m_t^2 \|x - y\| \tag{26}$$

   for all $x, y \in \mathbb{R}^{d_{\text{out}}}$ and $t \in [\underline{T}, \overline{T}]$, with a fixed deterministic L_{est}.

The first condition is the Gaussian-tail regularity used in Corollary 3.2 of Wang & Wang (2024), applied to the marginal density. With the time change $u = 2\int_0^t \beta_s \,\mathrm{d}s$, its bounds give

$$v_t := s_{d_{\text{out}}}(\cdot,t) + 2CC_t^{-1}(\cdot), \qquad \|v_t\|_\infty \leq L_{\text{true}} m_t, \qquad \text{Lip}(v_t) \leq L_{\text{true}} m_t^2. \tag{27}$$

The factor 2 in our score convention is absorbed into L_{true}. The second condition is imposed on the fitted estimator separately from empirical risk minimization. No such condition is required for the projection-based bound below.

To see the role of the construction, covariance comparability gives

$$CC_t^{-1} - I = m_t^2 (C - C_0)C_t^{-1}, \qquad \|CC_t^{-1} - I\|_{\text{op}} \leq K_{\text{cov}} m_t^2. \tag{28}$$

Hence $\tilde{s}_\theta + 2CC_t^{-1}x = g_\theta + 2(CC_t^{-1} - I)x$ has no nondecaying linear remainder. A uniform bound $\text{Lip}_z\, f_\theta(\cdot,t) \leq L_f$ would imply (26) with $L_{\text{est}} = L_f + 2K_{\text{cov}}$; the assumption itself does not require this particular sufficient condition. Moreover,

$$\|\tilde{s}_\theta(x,t)\| \leq 2\|x\| + m_t B_n, \qquad \langle x, x + \tilde{s}_\theta(x,t)\rangle \leq -\|x\|^2 + m_t B_n \|x\|.$$

Thus the full score has linear growth. The Gaussian-corrected estimated score need not be uniformly bounded in space: Assumption 4 of Wang & Wang (2024) requires its Lipschitz bound, whereas the uniform value bound in their Corollary 3.2 concerns the true corrected score.

Covering numbers. For every fixed (x,t), Euclidean projection is nonexpansive, so

$$\|\tilde{s}_\theta(x,t) - \tilde{s}_{\theta'}(x,t)\| \le \|F_\theta(x,t) - F_{\theta'}(x,t)\|.$$

Consequently, on every bounded input domain D,

$$\mathcal{N}(\mathcal{S}_{\mathrm{res}}, \delta, d_{I,D}) \le \mathcal{N}(\mathcal{S}_{\mathrm{raw}}, \delta, d_{I,D}).$$

Covering the two raw branches separately as in (21) therefore gives the same upper bound (19) for this class. In particular, for polynomial covering accuracy and $\Gamma \asymp \log n$,

$$\log(2\mathcal{N}_I^{\mathrm{res}}) \lesssim 2^{\kappa\Gamma/a^\dagger}\Gamma^{1/\eta+8\max\{1,1/\eta\}}\log^8(e\Gamma) + d_{\mathrm{out}}\Gamma^6\log n. \tag{29}$$

Here κ has the same mixed/anisotropic values as in Theorem 5.1. The factors m_t^{-1} in the auxiliary representation (25) do not enlarge this cover: the contraction is applied directly to (24).

Approximation and the oracle bound. The raw approximation bounds (4)–(5), summed over coordinates and marginalized by conditional Jensen's inequality, provide a comparator $F_\Gamma \in \mathcal{S}_{\mathrm{raw}}$ with

$$\bar{R}_{\underline{T},\overline{T};I}(F_\Gamma, s_{d_{\mathrm{out}}}) \lesssim \Gamma J_2\, 2^{-2(1-v/a^\dagger)\Gamma}.$$

By (27) and (28),

$$\|r_t(x)\| \le L_{\mathrm{true}} m_t + 2K_{\mathrm{cov}} m_t^2 \|x\|.$$

The dimension-uniform Gaussian tail estimate established in Lemma H.1 implies that, for any fixed $p > 0$, a sufficiently large K_{res} ensures

$$\sup_{t\in[\underline{T},\overline{T}]} \mathbb{E}_{\tilde{P}_t}[(\|r_t(X_t)\| - m_t B_n)_+^2] \lesssim n^{-p}.$$

Nonexpansiveness and the triangle inequality now yield

$$\inf_{f\in\mathcal{S}_{\mathrm{res}}} \bar{R}_{\underline{T},\overline{T};I}(f, s_{d_{\mathrm{out}}}) \lesssim \Gamma J_2\, 2^{-2(1-v/a^\dagger)\Gamma} + n^{-p}.$$

For clarity, the oracle argument is applied to the bounded residual g_θ, not to the linearly growing full score. Writing $k_t = h_t + 2x$, the forward representation gives

$$k_t(\tilde{X}_t, x_0) = 2m_t x_{0,[1:d_{\mathrm{out}}]} - 2m_t^2\sigma_t^{-1}C^{1/2}Z.$$

Its conditional time-averaged second moment is at most $C'(\|x_{0,[1:d_{\mathrm{out}}]}\|^2 + \mathrm{Tr}(C)J_2)$. The conditional mean given $\tilde{X}_t$ is r_t. Let $\ell_{I,g}^{\mathrm{res}}$ denote the integrated loss contrast $\ell_{A,f}$ defined above, with $A = I$, $f = g$, and the target h_t replaced by $k_t = h_t + 2x$. The residual class has envelope B_n, while $r_t = v_t + 2(I - CC_t^{-1})x$ has a uniformly bounded linear coefficient and bounded nonlinear part. Thus the loss contrasts, localization, and Bernstein argument in Proposition H.3 apply with

$$B_I^{\mathrm{res}} \lesssim 1 + B_n^2 + U^2 + J_2.$$

We adapt Steps 1–4 in the proof of Proposition H.3 to the residual loss. On the localization event E_U, define

$$H_{\mathrm{res},U}^2 := \sup_{x_0\in E_U} \frac{1}{\Delta T}\int_{\underline{T}}^{\overline{T}} \mathbb{E}\Big[\|k_t(\tilde{X}_t, x_0)\|^2 \mid X_0 = x_0\Big]\,\mathrm{d}t.$$

The conditional moment bound above and $\|x_{0,[1:d_{\mathrm{out}}]}\|^2 \le KU^2\,\mathrm{Tr}(C)$ on E_U give $H_{\mathrm{res},U}^2 \lesssim U^2 + J_2$. Consequently, the finite-cover comparison in Step 2 uses the local error bound

$$e_{\mathrm{res}} \lesssim (B_n + H_{\mathrm{res},U})\delta + B_n^2 p_Z(V) + B_n H_{\mathrm{res},U}\sqrt{p_Z(V)}.$$

For arbitrary x_0, the residual loss contrast satisfies

$$|\ell_{I,g}^{\mathrm{res}}(x_0)| \lesssim B_n^2 + B_n\left(\|x_{0,[1:d_{\mathrm{out}}]}\| + \sqrt{\mathrm{Tr}(C)J_2}\right).$$

Cauchy–Schwarz and Gaussian domination therefore yield

$$\sup_g \mathbb{E}\big[|\ell_{I,g}^{\mathrm{res}}(X_0)|\mathbf{1}_{E_U^c}\big] \lesssim B_I^{\mathrm{res}}\sqrt{p_0(U)}.$$

This estimate controls the contribution from E_U^c both in the population-risk comparison with a cover center in Step 2 and, after taking training-data expectations, in the comparison of the original and localized empirical losses in Step 4. The localized moment and variance estimates in Step 1 and the Bernstein argument in Step 3 apply with B_I^{res} in place of B_A.

With $\underline{T} = n^{-2}$, $\overline{T} = L_T \log n$, the localization choices used above, and $p \geq 4$, we obtain

$$\mathbb{E}_n[\bar{R}_{\underline{T},\overline{T};I}(\hat{s}^{(iii)}, s_{d_{\mathrm{out}}})] \lesssim \Gamma\, 2^{-2(1-v/a^\dagger)\Gamma} + \frac{\log n}{n} \log(2\mathcal{N}_I^{\mathrm{res}}) + n^{-3}. \tag{30}$$

The conditions on the raw comparators are exactly those used in the preceding approximation and estimation bounds.

Reference result and constants. We record Theorem 3.3 of Wang & Wang (2024) in its original normalization. In this paragraph only, the forward drift is $-x/2$, the diffusion covariance is C, and the score is $s_{\mathrm{ref}}(u,x) = C\nabla \log p_u(x)$. Write $p_0(x) \propto \exp(-x^\top C_0^{-1}x/2 + h(x))$ and $\bar{C}_u = e^{-u}C_0 + (1-e^{-u})C$. Their constants are

$$\begin{aligned} K &:= \max\{1, \|C_0 C^{-1}\|_{\mathrm{op}}\}, \\ L_0 &:= K^2\big(\|C\nabla^2 h\|_\infty + \|\sqrt{C}\nabla h\|_\infty^2\big), \\ L_1 &:= K\|C\|_{\mathrm{op}}^{1/2}\|\sqrt{C}\nabla h\|_\infty, \qquad L_2 := \max\{\|I - CC_0^{-1}\|_{\mathrm{op}}, \|C_0C^{-1} - I\|_{\mathrm{op}}\}. \end{aligned}$$

Their Corollary 3.2 bounds the spatial derivative and value of $s_{\mathrm{ref}}(u,x) + C\bar{C}_u^{-1}x$ by $L_0 e^{-u}$ and $L_1 e^{-u/2}$, respectively. Thus the covariance-mismatch constant entering the stability bound is L_2 in their notation.

Proposition H.5 (Wang & Wang, 2024, Theorem 3.3)**.** *Under Assumptions 1–4 of the cited theorem, let $M_2 = \mathbb{E}_{p_0}[\|X\|^2]$, and let ε^2 bound the time-averaged squared score error at the discretization points. The estimated Gaussian-corrected score has derivative bound $L_0 e^{-u}$ at those points. For their scheme (5), initialized from $\mathcal{N}(0,C)$, with horizon T_{ref} and sufficiently small maximum step size Δ_{ref}, its terminal law Q_{ref} satisfies*

$$\mathcal{W}_2^2(p_0, Q_{\mathrm{ref}}) \leq K_1\left[e^{-2T_{\mathrm{ref}}}(M_2 + \operatorname{Tr} C) + \varepsilon^2 T_{\mathrm{ref}} + \operatorname{Tr}(C)\Delta_{\mathrm{ref}}^2\right],$$
$$K_1 := e^{4+3(L_0+L_2)} \max\{(e+1)e, e^3(L_0+L_2)^2\}.$$

In our convention, $u = 2\int_0^t \beta_s\, \mathrm{d}s$ and the score is twice s_{ref}. The next proposition supplies the continuous-time, early-stopped marginal bound directly in this convention; its implicit constant is not identified with the explicit K_1 above.

Continuous-time Wasserstein stability. The following is the continuous-time coupling bound underlying Wang & Wang (2024), stated with our score normalization, finite-dimensional marginals, and early stopping.

Proposition H.6 (Stability for the modified score)**.** *Assume the standing covariance and moment conditions, $0 < \underline{T} \leq 1$, $\overline{T} > \underline{T}$, and $0 < \underline{\beta} \leq \beta_t \leq \overline{\beta} < \infty$. If the fitted modified score satisfies (26), then*

$$\mathbb{E}_n[\mathcal{W}_2^2(P_0, \hat{Q}^{(iii)}_{\overline{T}-\underline{T}})] \lesssim \underline{T} + d_{\mathrm{out}}^{1-2/r_0} + m_{\overline{T}}^2 + (\overline{T} - \underline{T})\mathbb{E}_n[\bar{R}_{\underline{T},\overline{T};I}(\hat{s}^{(iii)}, s_{d_{\mathrm{out}}})]. \tag{31}$$

The implicit constant is independent of n, d_{out}, and $\overline{T}$. The same bound with an additional $O(n^{-2})$ term holds for the terminally projected law.

Proof. Condition on the training data and synchronously couple the true reverse process Y_u and the learned process $\hat{Y}_u$. Their initial laws are $\tilde{P}_{\overline{T}}$ and $\mathcal{N}(0,C)$. For $t = \overline{T} - u$, the corrected Lipschitz condition and (28) give the one-sided bound

$$\langle x - y, (x + \hat{s}^{(iii)}(x,t)) - (y + \hat{s}^{(iii)}(y,t))\rangle \leq (-1 + Am_t^2)\|x-y\|^2, \quad A = L_{\mathrm{est}} + 2K_{\mathrm{cov}}.$$

The difference process therefore satisfies the scalar comparison inequality with forcing $\beta_t\|\hat{s}^{(iii)}(Y_u,t) - s_{d_{\mathrm{out}}}(Y_u,t)\|$. Since $\int_0^{\overline{T}} \beta_t m_t^2\, \mathrm{d}t \leq 1/2$, its propagation factor from v to

u is bounded by

$$e^{A/2}\exp\left(-\int_v^u \beta_{\overline{T}-w}\,\mathrm{d}w\right).$$

The integral of this exponential kernel against $\beta_{\overline{T}-v}\,\mathrm{d}v$ is at most one. Cauchy–Schwarz with that measure consequently gives

$$\mathbb{E}[\|Y_{\overline{T}-\underline{T}}-\hat{Y}_{\overline{T}-\underline{T}}\|^2]\lesssim \mathcal{W}_2^2(\tilde{P}_{\overline{T}},\mathcal{N}(0,C))+\int_{\underline{T}}^{\overline{T}}\beta_t\mathbb{E}_{\tilde{P}_t}[\|\hat{s}^{(iii)}(X_t,t)-s_{d_{\mathrm{out}}}(X_t,t)\|^2]\,\mathrm{d}t.$$

The forward OU coupling bounds the initialization term by $Cm_{\overline{T}}^2$ using the dimension-uniform second moment and $1-\sigma_t\le m_t^2$. The early-stopping and coordinate-coupling bounds established at the beginning of this appendix add $C(\underline{T}+d_{\mathrm{out}}^{1-2/r_0})$. Taking training-data expectations proves (31). Nonexpansiveness of terminal projection and the true-law tail bound give the final assertion. □

Theorem H.7 (Case (iii): rate under corrected Lipschitz regularity)**.** *Assume the standing hypotheses and approximation conditions used for Theorem 5.1, Assumption H.4, the noise-schedule assumptions in Section 2, and the time-endpoint conditions above. Use* Γ *from* (20)*, and recall*

$$\nu=\frac{2(a^\dagger-v)}{2(a^\dagger-v)+\kappa}.$$

Choose

$$d_{\mathrm{out}}\asymp n^{r_0/2},\qquad \underline{T}=n^{-2},\qquad \overline{T}=L_T\log n,$$

with sufficiently large fixed L_T*. Then*

$$\mathbb{E}_n[\mathcal{W}_2^2(P_0,\hat{Q}^{(iii)}_{\overline{T}-\underline{T}})]=\widetilde{O}\left(n^{-\min\{\nu,(2-r_0)/2\}}\right).$$

The terminally projected generator has the same rate. Here $\widetilde{O}$ *suppresses only powers of* $\log n$ *and* $\log\log n$*, with the distributional and regularity constants fixed.*

Proof. Combining (29), (30), and (31) gives

$$\mathbb{E}_n[\mathcal{W}_2^2(P_0,\hat{Q}^{(iii)}_{\overline{T}-\underline{T}})]\lesssim d_{\mathrm{out}}^{1-2/r_0}+n^{-2}$$
$$+n^{-\nu}(\log n)^{2+(1/\eta+8\max\{1,1/\eta\})\nu}(\log\log n)^{8\nu}+\frac{d_{\mathrm{out}}(\log n)^9}{n}.$$

Here $m_{\overline{T}}^2$ is absorbed by n^{-2} by taking L_T sufficiently large, using $\beta_t\ge\underline{\beta}$. For $d_{\mathrm{out}}\asymp n^{r_0/2}$, both $d_{\mathrm{out}}^{1-2/r_0}$ and d_{out}/n have polynomial order $n^{-(2-r_0)/2}$. This proves the assertion. □

H.2.2 Case (iv): a projection-based bound without corrected Lipschitz regularity

Case (iv) uses the original estimators of Options 1 and 2. It does not require Assumption H.4 or the residual-score modification.

Proposition H.8 (Terminal projection and relative entropy)**.** *Under the standing tail and finite-dimensional Girsanov conditions, let* $\hat{s}$ *be either original estimator, with output law* $\hat{Q}$*. For* $R_n=K_{\mathrm{proj}}\sqrt{\log n}$ *with sufficiently large fixed* K_{proj}*,*

$$\mathbb{E}_n[\mathcal{W}_2^2(P_0,\hat{Q}^{\mathrm{proj}}_{\overline{T}-\underline{T}})]\lesssim \underline{T}+d_{\mathrm{out}}^{1-2/r_0}+n^{-2}+\log n\left[K_{\mathrm{init}}+\overline{\beta}(\overline{T}-\underline{T})\mathbb{E}_n[\bar{R}_{\underline{T},\overline{T};C^{-1}}(\hat{s},s_{d_{\mathrm{out}}})]\right]^{1/2}. \tag{32}$$

Proof. For probability measures P,Q on $\mathbb{R}^{d_{\mathrm{out}}}$, maximal coupling on the radius-R ball gives

$$\mathcal{W}_2^2((\Pi_R)_\#P,(\Pi_R)_\#Q)\le 4R^2\,\mathrm{TV}((\Pi_R)_\#P,(\Pi_R)_\#Q)\le 4R^2\,\mathrm{TV}(P,Q).$$

Consequently, Pinsker's inequality and the triangle inequality imply

$$\mathcal{W}_2^2(P,(\Pi_R)_\#Q)\le 2\mathbb{E}_P[(\|X\|-R)_+^2]+8R^2\sqrt{\mathrm{KL}(P\|Q)/2}.$$

Apply this with $P = \tilde{P}_{\underline{T}}$ and $Q = \hat{Q}_{\overline{T}-\underline{T}}$. The Gaussian tail estimate makes the first term $O(n^{-2})$. Use Girsanov's bound for the KL term and Jensen's inequality for training-data expectation, then add the early-stopping and coordinate-truncation errors. This proves the claim. □

Retain Γ from (20) and the same time endpoints, but choose the output dimension for this bound. The preceding oracle estimates give, up to logarithmic factors,

$$\mathbb{E}_n[\bar{R}_{\underline{T},\overline{T};C^{-1}}(\hat{s}, s_{d_{\mathrm{out}}})] \lesssim \begin{cases} d_{\mathrm{out}}^{2/r_0}(n^{-\nu} + d_{\mathrm{out}}/n + n^{-3}), & \text{Option 1}, \\ d_{\mathrm{out}} n^{-\nu} + d_{\mathrm{out}}^2/n + n^{-3}, & \text{Option 2}. \end{cases}$$

Balancing these terms in (32) yields

$$\mathbb{E}_n[\mathcal{W}_2^2(P_0, \hat{Q}^{\mathrm{proj}}_{\overline{T}-\underline{T}})] = \begin{cases} \widetilde{O}\left(n^{-\min\{\frac{(2-r_0)\nu}{2(3-r_0)}, \frac{2-r_0}{6-r_0}\}}\right), & \text{Option 1}, \\ \widetilde{O}\left(n^{-\min\{\frac{(2-r_0)\nu}{4-r_0}, \frac{2-r_0}{4}\}}\right), & \text{Option 2}. \end{cases}$$

Indeed, the respective choices are

$$d_{\mathrm{out}} \asymp \begin{cases} n^{\min\{r_0\nu/[2(3-r_0)],\, r_0/(6-r_0)\}}, & \text{Option 1}, \\ n^{\min\{r_0\nu/(4-r_0),\, r_0/4\}}, & \text{Option 2}. \end{cases}$$

Both dominate $d_{\max}$ for sufficiently large n. These are rates for the squared Wasserstein distance; taking another square root gives the corresponding root-mean-square $\mathcal{W}_2$ rates. In particular, the entropy argument controls $\mathcal{W}_2$ through the fourth root of KL, whereas Case (iii) controls $\mathcal{W}_2^2$ directly by the unweighted score risk.