A STATISTICAL THEORY OF FREQUENCY RECOVERY IN OVERPARAMETERIZED DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models have achieved broad success in image, video, and sequence generation, yet they often learn global layouts and temporal trends more faithfully than fine local details. Existing theory establishes diffusion models as powerful distribution estimators, but largely concerns full distribution estimation rates, leaving unclear how data properties and network capacity govern accuracy at different resolutions. We address this gap by developing a statistical theory of frequency recovery in overparameterized diffusion models, using the discrete Fourier transform to describe data at different resolutions. For a prescribed frequency cutoff, we establish approximation and estimation guarantees for the score function associated with the low-frequency component of the data distribution. We further prove a distribution recovery bound separating a statistical error, whose rate depends on an effective dimension determined by the number of retained frequencies, from a remainder reflecting dependence between low and high frequencies and the spectral energy of the latter. Crucially, our guarantees use weight matrix norm constraints tailored to the frequency cutoff and, at fixed depth and width, remain unchanged as the number of available parallel networks grows. Synthetic and CIFAR-10 experiments illustrate how sample size and weight decay affect frequency recovery.

1 INTRODUCTION

Diffusion models have achieved strong performance in a broad range of applications, including image synthesis and editing (Song & Ermon, 2019; Ho et al., 2020; Dhariwal & Nichol, 2021; Ho et al., 2022; Rombach et al., 2022; Saharia et al., 2022), protein design and generation (Watson et al., 2023; Ingraham et al., 2023; Yim et al., 2023; Lin & AlQuraishi, 2023), and robot control and manipulation (Chi et al., 2025; Janner et al., 2022; Wang et al., 2022; Ze et al., 2024). In these applications, diffusion models seek to learn high-dimensional data distributions from observed samples. They estimate the score functions of noise-perturbed data distributions and use these estimates to guide the reverse diffusion process for generating new samples.

Learning a complex, high-dimensional distribution from finite data is statistically challenging. A growing body of theory establishes diffusion models as distribution estimators that attain minimax optimal rates (up to logarithmic factors) under smoothness assumptions (Oko et al., 2023; Zhang et al., 2024a; Cai & Li, 2025; Stéphanovitch et al., 2025). These rates nevertheless suffer from the curse of dimensionality: the sample size required for a fixed accuracy can grow exponentially with the ambient dimension. Structural assumptions offer a way to reduce this difficulty. For data supported on a low-dimensional linear subspace, Oko et al. (2023) and Chen et al. (2023a) derive rates whose sample-size exponents depend on the subspace dimension. Further results extend to distributions supported on low-dimensional manifolds and with densities satisfying low-dimensional factorizations (Tang & Yang, 2024; Kwon et al., 2026).

These guarantees show how low-dimensional structure can ease distribution estimation, but still exhibit notable limitations. On the one hand, exact support on a subspace or manifold excludes variation outside that structure, even when such variation is small. It is therefore important to understand what can be recovered when the data retain variation beyond a low-dimensional representation. On the other hand, empirical studies suggest that learning and generation accuracy can differ substantially across resolutions, demonstrating that a convergence rate for the full distribution alone does not characterize recovery accuracy at different resolutions. In image generation, empirical studies

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.

reveal frequency-dependent recovery in diffusion models: global structure and low-frequency components are typically recovered more readily, whereas high-frequency textures and fine details can be more difficult to capture (Yang et al., 2023; Qian et al., 2024; Huang et al., 2024; Yang et al., 2025). For time series, diffusion models can effectively capture long-term trends and low-frequency components while struggling with high-frequency variations and abrupt changes; frequency-domain modeling can substantially improve distributional fidelity (Galib et al., 2024; Crabbé et al., 2024; Jeong et al., 2025). These observations motivate a more detailed account of what diffusion models recover at each resolution, beyond a single rate for the full distribution.

This paper makes a step toward establishing finer-grained statistical properties of diffusion models. We adopt the Discrete Fourier Transform (DFT) to describe distributional information across resolutions and investigate following fundamental questions:

How well do diffusion models recover data distributions across different frequency ranges?
How do the corresponding statistical rates vary across frequencies?

To answer these questions, we introduce a frequency cutoff w and study the distribution retaining frequencies up to that cutoff. Unlike existing theoretical works (Oko et al., 2023; Chen et al., 2022; Tang & Yang, 2024; Kwon et al., 2026; Wang et al., 2026; Benita et al., 2026), we consider denoising score matching with overparameterized score networks, using weight matrix norm constraints to control statistical complexity. Under smoothness and spectral energy assumptions, we establish guarantees for score approximation, score estimation, and distribution recovery of the retained frequencies. This result connects data frequency properties, overparameterized score network, and generalization theories for diffusion models. Our main contributions are summarized as follows.

• **Score approximation (Theorem 4.2).** We show that, within one sufficiently overparameterized ReLU architecture, for every frequency cutoff up to a prescribed maximum there exists a cutoff-specific network that approximates the score function. The construction admits a mixture-of-experts interpretation, meaning that for different cutoffs, the network maintains a sparse set of active branches. Furthermore, we derive norm bounds tailored to the frequency cutoff that do not grow with the number of parallel networks. This differs from the existing results for data supported on a fixed low-dimensional structure: our network architecture supports a range of dimensions determined by the retained frequencies without assuming low-dimensional support for the data.

• **Score estimation (Theorem 5.1).** We establish a low frequency generalization bound of the empirical risk minimizer of the denoising score matching over the network architecture from the approximation theory. We separate a statistical error governed by sample size, the number of retained frequencies, data smoothness, and network depth from remainders measuring cross-frequency dependence and spectral energy above the cutoff. Under practical spectral decay, the bound quantifies a balance: retaining more frequencies reduces the remainders but makes statistical estimation harder.

• **Distribution recovery (Proposition 5.4) and numerical validation.** We transfer the score-estimation bound to a KL guarantee for the low-frequency projection of the generated distribution, separating recovery of chosen frequencies from full-distribution estimation. As network depth increases, the statistical term approaches the minimax lower-bound rate for estimating the projected distribution, up to logarithmic factors. Simulations and CIFAR-10 experiments validate these trends across sample sizes and norm constraints.

2 Related Work

Frequency-aware diffusion modeling. Diffusion models have emerged as a major class of generative models built upon forward and reverse stochastic processes (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Song et al., 2020a;b; Ho et al., 2020). They have achieved strong generative performance across a wide range of applications, particularly in image synthesis (Dhariwal & Nichol, 2021; Ho et al., 2022; Rombach et al., 2022; Saharia et al., 2022). Beyond improving overall generation quality, several works explicitly study frequency structure in diffusion models. Yang et al. (2023) analyze frequency-dependent denoising behavior and the loss of high-frequency details in compact models. FreeU, FouriScale, and FreCaS introduce frequency-aware feature modulation or sampling strategies for image generation (Si et al., 2024; Huang et al., 2024; Zhang et al., 2024b). Related ideas have also been explored in video diffusion through frequency-aware initialization and feature aggregation (Wu et al., 2024; Lu et al., 2024). Other approaches formulate diffusion directly in the Fourier domain (Crabbé et al., 2024).

Statistical theory of diffusion models. From a statistical perspective, existing work studies both the estimation of score functions from finite data and the resulting accuracy of the generated distribution. Neural score approximations have been developed using local Taylor expansions and diffused basis functions (Chen et al., 2023a; Oko et al., 2023; Fu et al., 2024), with results on score estimation (Wibisono et al., 2024; Dou et al., 2024; Gupta et al., 2024) and weaker assumptions on the underlying density (Zhang et al., 2024a). Structural assumptions can improve statistical rates, including low-dimensional subspaces or manifolds (Chen et al., 2023a; Tang & Yang, 2024; Azangulov et al., 2024; Zhang et al., 2026) and factorizable densities (Kwon et al., 2026; Fan et al., 2025). Sampling theory further translates score-estimation error into distributional guarantees for reverse SDEs and probability-flow methods (Chen et al., 2022; Benton et al., 2024; Chen et al., 2023b).

Recent theory has also characterized frequency-dependent behavior in diffusion models. Under the setting of linear denoising network, Wang & Pehlevan (2026) show that different spectral modes can be learned at different rates, and Wang et al. (2026) characterize how finite-sample errors vary across spectral directions. Benita et al. (2026) analyze frequency-dependent sampling dynamics under Gaussian data assumptions and use the resulting characterization for noise-schedule design.

3 Diffusion Models, Frequency Hierarchy, and Score Networks

We introduce continuous-time diffusion models and data distributions with spectral energy patterns.

3.1 Forward and Backward Processes and Score Estimation

A diffusion model adds noise to the data and learns the "score function" to reverse the noise corruption for generating new samples. The noise corruption is described by a forward stochastic process:

$$\mathrm{d}X_t = -\frac{1}{2}X_t \mathrm{d}t + \mathrm{d}W_t, \quad X_0 \sim P_{\mathrm{data}}, \quad t \in [0, T]. \tag{3.1}$$

Here, W_t is standard Brownian motion in $\mathbb{R}^d$, independent of X_0, P_{data} is a data distribution, and T is a terminal time. At time t, X_t can be represented as $X_t = \alpha_t X_0 + \sigma_t Z$ where $Z \sim \mathsf{N}(0, I)$ independent of X_0. The time-dependent coefficients are $\alpha_t = \exp(-\frac{t}{2})$ and $\sigma_t^2 = 1 - \alpha_t^2$. We denote P_t as the marginal distribution of X_t and let p_t denote its density.

The forward process (3.1) has a time-reversal under mild conditions (Anderson, 1982; Haussmann & Pardoux, 1986), which is known as the backward process:

$$\mathrm{d}X_t^{\leftarrow} = \left[\frac{1}{2}X_t^{\leftarrow} + \nabla \log p_{T-t}(X_t^{\leftarrow})\right]\mathrm{d}t + \mathrm{d}\overline{W}_t, \quad X_0^{\leftarrow} \sim P_T, \quad t \in [0, T]. \tag{3.2}$$

Here, $\overline{W}_t$ is another Brownian motion and $\nabla \log p_t$ is the score function. To generate new samples, we discretize the backward process (3.2) and make two further approximations. First, we replace P_T by the standard Gaussian distribution $\mathsf{N}(0, I)$. Second, we estimate the unknown score function $\nabla \log p_t$ using finite training samples.

Given a data set $\mathcal{D}_n = \{X^{(1)}, \dots X^{(n)}\}$ with i.i.d. samples from P_{data}, the score estimation is to optimize a quadratic loss:

$$\widehat{s} \in \operatorname*{arg\,min}_{s \in \mathcal{F}} \widehat{\mathcal{L}}(s; \mathcal{D}_n) = \frac{1}{n}\sum_{i=1}^{n}\int_{t_0}^{T} \mathbb{E}_{X_t \sim \mathsf{N}(\alpha_t X^{(i)}, \sigma_t^2 I)} \left\| s(X_t, t) + \frac{X_t - \alpha_t X^{(i)}}{\sigma_t^2} \right\|_2^2 \mathrm{d}t, \tag{3.3}$$

where $\mathcal{F}$ is a class of functions and t_0 is an early-stopping time. We will introduce our neural network parameterization of $\mathcal{F}$ slightly later.

3.2 Data Distribution and Fourier Frequency Hierarchy

To describe the data frequency content, we adopt the discrete Fourier transform (DFT). Let $\mathsf{F} \in \mathbb{C}^{d\times d}$ with $\mathsf{F}_{k+1,\ell+1} = d^{-1/2}\exp(-2\pi\sqrt{-1}k\ell/d)$ for $k, \ell = 0, \dots, d-1$. Then the DFT of a data point $x \in \mathbb{R}^d$ is denoted as $\mathsf{F}x \in \mathbb{C}^d$. The inverse DFT is simply the conjugate transpose F^*.

Consider a cut-off frequency $w \in \{0, 1, \dots, \lfloor d/2 \rfloor\}$. We denote $[\Lambda_w]_{k+1,k+1} = \mathbb{1}\{\min(k, d-k) \le w\}$ as a diagonal mask matrix. The corresponding low-frequency projection of a data point x at frequency w is $x_w = \Pi_w x$ with $\Pi_w = \mathsf{F}^*\Lambda_w\mathsf{F}$. Note that Π_w is a real orthogonal projector, and we denote its rank as $d_{\mathrm{eff}}(w) = \mathrm{rank}(\Pi_w) = \min\{d, 2w+1\}$.

In practical applications, data distributions often exhibit pronounced structure in the frequency domain. For example, natural images have approximately inverse-square spatial power spectra (Field, 1987; Ruderman & Bialek, 1993), while natural videos exhibit non-separable coupling and scaling between spatial and temporal frequencies (Dong, 1996). These observations motivate a frequency-resolved description of the data distribution.

3.3 Score Network Architecture

We close this section by introducing the score network architecture $\mathcal{F}$. Let $y = [x^\top, t]^\top \in \mathbb{R}^{d+1}$. We focus on the ReLU activation function and denote an ℓ-layer network as

$$f(y) = A_\ell \cdot \mathrm{ReLU}\left(A_{\ell-1} \cdot \mathrm{ReLU}\left(\dots \mathrm{ReLU}(A_1 y + b_1)\dots\right) + b_{\ell-1}\right) + b_\ell, \tag{3.4}$$

where $\{A_j, b_j\}_{j=1}^{\ell}$ are trainable weight parameters of compatible sizes. Mimicking the mixture of experts structure, we consider $\mathcal{F}$ consisting of M parallel ReLU networks in the form (3.4) with shared output layers. More precisely, we denote A_j^m and b_j^m as weight parameters in each parallel network f_m and A_j^{out} and b_j^{out} as parameters in the output layers f_{out}. Then we define class $\mathcal{F}$ with controllable hyperparameters as

$$\begin{aligned}\mathcal{F}(M, L, W, B, B_{\mathrm{out}}, S_t) = \Big\{ & g(y) = f_{\mathrm{out}}([f_1(y)^\top, \dots, f_M(y)^\top]^\top) : f_{\mathrm{out}}, f_m \text{ in the form (3.4)}, \\ & \text{each with depth } L \text{ and width } W, \quad \sup_x \|g(y = [x^\top, t]^\top)\|_2 \le S_t, \\ & \sum_m \sum_j (\|A_j^m\|_{\mathrm{F}}^2 + \|b_j^m\|_2^2) \le B, \ \ \sum_j (\|A_j^{\mathrm{out}}\|_{\mathrm{F}}^2 + \|b_j^{\mathrm{out}}\|_2^2) \le B_{\mathrm{out}} \Big\}.\end{aligned} \tag{3.5}$$

The norm constraints B and B_{out} help to restrict the complexity of the network class, even as the number of parallel networks grows large. Together with the output magnitude bound S_t, they are critical for the established score estimation rates (Theorem 5.1) to circumvent the naïve dependence on the full network size. In practice, controlling the norm bounds of a neural network can be realized by a weight decay. In Section 6, we therefore use weight decay as a practical surrogate for norm control and examine its effect on frequency recovery.

4 Overparameterized Networks Approximate Score Functions

To study how networks represent different frequency components, we consider the projected data law $P_{\mathrm{data}}^{(w)} = (\Pi_w)_\sharp P_{\mathrm{data}}$. The rank of Π_w is $d_{\mathrm{eff}}(w)$ and increasing w retains more frequencies from the original data distribution. When $w < \lfloor d/2 \rfloor$, Π_w is not full-rank and $P_{\mathrm{data}}^{(w)}$ is indeed supported on a low-dimensional subspace. Initializing the forward process (3.1) from $P_{\mathrm{data}}^{(w)}$ gives the density $p_t^{(w)}$ in $\mathbb{R}^d$ for $t > 0$. Correspondingly, the score function is $\nabla \log p_t^{(w)}$ and this section investigates using neural networks to approximate such a score function.

Our goal is to use one network architecture for every frequency cutoff up to a fixed value $w_{\max}$. Existing quantitative score-approximation results typically study the full score function (Oko et al., 2023; Fu et al., 2024) or assume a fixed low-dimensional structure (Chen et al., 2023a; Tang & Yang, 2024). Here, the network architecture must accommodate the highest retained frequency. We therefore characterize a sufficient network size according to $w_{\max}$ with lower bounds on the width, depth and the number of parallel structures. Further, we identify weight matrix norm bounds tailored to each cutoff for controlling active parallel structures when the network is overparameterized. Importantly, the norm bounds control the complexity relevant to score estimation from data, which we establish in Section 5.1. Before stating the result, we impose the following assumption on the data distribution.

Assumption 4.1 (*Light-tailed data distribution and Hölder regularity*). The data distribution P_{data} has a density p_{data} with respect to the Lebesgue measure in $\mathbb{R}^d$. There exists a constant $C > 0$ and a function f such that

$$p_{\mathrm{data}}(x) = f(x) \cdot (C/2\pi)^{d/2} \exp(-C\|x\|_2^2/2).$$

Moreover, f is lower bounded away from 0, i.e., $\inf_x f(x) \ge c$, and belongs to a bounded Hölder ball, i.e., $f \in \mathcal{H}^\beta$ with Hölder index $\beta > 0$ and $\|f\|_{\mathcal{H}^\beta} \le H_\beta$, for constants $c, H_\beta > 0$.

The density condition on P_{data} resembles the light-tailed assumption commonly adopted in the existing literature on the theoretical foundations of diffusion models. The lower bound on f ensures a positive probability mass coverage to the whole space, which is useful to control the magnitude of the score function. The density is also Hölder continuous, following the standard assumption in nonparametric estimation (Tsybakov & Zaiats, 2009). We provide a self-contained definition of Hölder functions and their Hölder norm in Appendix E.1. The following theorem establishes approximation guarantees.

Theorem 4.2 (Score approximation using overparameterized networks)**.** Suppose Assumption 4.1 holds. Fix an early-stopping time $t_0 \in (0, T)$, a maximal frequency cutoff $w_{\max} \in \{0, \dots, \lfloor d/2 \rfloor\}$, and an approximation error tolerance $\epsilon \in (0, e^{-2})$. Choose the integers M, L, W to satisfy

$$M \geq C_M \epsilon^{-d_{\text{eff}}(w_{\max})/2\beta} \big(\log \frac{d}{\epsilon}\big)^{d_{\text{eff}}(w_{\max})/2}, \quad L \geq C_L \big(\log \frac{d}{\sigma_{t_0}^2 \epsilon}\big)^3, \quad W \geq C_W d,$$

where C_M, C_L, C_W depend on $\beta, c, C, d_{\text{eff}}(w_{\max}), H_\beta$ and T. Then, for any $w \in \{0, \dots, w_{\max}\}$, there exists a function $\bar{s}_w \in \mathcal{F}(M, L, W, B, B_{\text{out}}, S_t = \infty)$ such that, for any $t \in [t_0, T]$, we have

$$\mathbb{E}_{X_t \sim P_t} \Big[\big\| \Pi_w \big(\bar{s}_w(X_t, t) - \nabla \log p_t^{(w)}(X_t) \big) \big\|_2^2 \Big] \leq \sigma_t^{-2} \epsilon. \tag{4.1}$$

The norm constraints are

$$B = \widetilde{\mathcal{O}}\big(L(w+1)^2 (\epsilon^{-d_{\text{eff}}(w)(1+1/L)/2\beta} + d)\big) \quad \text{and} \quad B_{\text{out}} = \mathcal{O}(Ld(w+1)^2). \tag{4.2}$$

The proof is deferred to Appendix A. We discuss several interpretations.

Projected low-frequency score approximation error. Theorem 4.2 states the approximation error under the frequency projection Π_w. Whenever $t > 0$, the density $p_t^{(w)}$ is supported on the full space because of the standard Gaussian noise corruption even if the initial distribution is only supported on a low-frequency subspace. For any $x \in \mathbb{R}^d$, we have

$$\nabla \log p_t^{(w)}(x) = \nabla \log p_t^{(w)}(\Pi_w x) - \sigma_t^{-2}(I - \Pi_w)x.$$

The display above separates an orthogonal component and a low-frequency component. Applying Π_w to $\nabla \log p_t^{(w)}$ removes the influence of the orthogonal one and keeps the approximation error focused on the low-frequency signal. Furthermore, let $U_w \in \mathbb{R}^{d \times d_{\text{eff}}(w)}$ satisfy $U_w U_w^\top = \Pi_w$ and $U_w^\top U_w = I$. The projected score $\Pi_w \nabla \log p_t^{(w)}(x)$ depends only on $U_w^\top x$, which reproduces the score decomposition for linear subspace data in Chen et al. (2023a). In our network construction, $\bar{s}_w$ only depends on $U_w^\top x$. Thus, (4.1) is identical to replace the marginal distribution P_t by $X_t \sim P_t^{(w)}$. We retain P_t for an easier connection to the estimation guarantee in Section 5.1.

Network size and active parallel networks. A sufficient network size is chosen to accommodate the most demanding frequency cutoff $w_{\max}$. At fixed accuracy, the number of parallel networks depends ϵ, $d_{\text{eff}}(w_{\max})$ and smoothness β. The width and depth of each parallel network grow at most poly-logarithmically in ϵ. Thus, the total number of weight parameters in the sufficient construction matches the familiar neural network universal approximation rate (Oko et al., 2023; Fu et al., 2024) when $d_{\text{eff}}(w_{\max})$ is treated as the data dimension.

When a cutoff frequency w is smaller than $w_{\max}$, our constructive proof yields a sparse active parallel network structure analogous to mixture of experts. Specifically, we rewrite the score function $\nabla \log p_t^{(w)}$ as a fraction $\nabla p_t^{(w)} / p_t^{(w)}$. On a small cube in $\mathbb{R}^d$, each parallel network implements a local Taylor polynomial for both approximating the denominator and the numerator, which acts as an expert. Then the shared output layer combines these local polynomials to evaluate the ratio. The total number of small cubes exponentially scales with $d_{\text{eff}}(w)$. Therefore, those constructed parallel networks form an active part among all parallel structures; the inactive part has zero output.

Norm bounds and weight decay. The weight matrix norm bounds are adaptive to the cutoff frequency, but independent of the number of parallel networks. This is important to control the statistical complexity of $\mathcal{F}$ so that Section 5.1 establishes an estimation rate independent of M. In practice, directly imposing norm constraints can complicate the optimization procedure. Instead, we can replace the norm constraints by weight-decay penalties:

$$\underset{s \in \mathcal{F}(M, L, W, \infty, \infty, S_t)}{\arg\min} \widehat{\mathcal{L}}(s; \mathcal{D}_n) + \lambda \sum_m \sum_j \big(\|A_j^m\|_{\text{F}}^2 + \|b_j^m\|_2^2 \big) + \lambda_{\text{out}} \sum_j \big(\|A_j^{\text{out}}\|_{\text{F}}^2 + \|b_j^{\text{out}}\|_2^2 \big), \tag{4.3}$$

where $\lambda, \lambda_{\text{out}} > 0$ are tuning parameters. Proposition E.11 in Appendix E.5 gives explicit penalty coefficients for attaining norm budgets B and B_{out}. In Section 6 we use weight decay for training.

5 Score Estimation and Distribution Recovery

We now study score estimation from finite data and the accuracy of the generated distribution using the learned score function. Section 5.1 shows how sample size, data regularity, and the norm bounds from Theorem 4.2 control estimation of the low-frequency score. Section 5.2 then bounds the error in recovering the distribution of those frequencies.

5.1 Score Estimation for Low Frequencies

Recall that $\widehat{s}$ minimizes the empirical denoising loss $\widehat{\mathcal{L}}(s; \mathcal{D}_n)$ over $\mathcal{F}$ in (3.3). We measure its low-frequency error by

$$\mathcal{R}_w(\widehat{s}) = \int_{t_0}^{T} \mathbb{E}_{X_t \sim P_t} \left[\left\| \Pi_w \left(\widehat{s}(X_t, t) - \nabla \log p_t^{(w)}(X_t) \right) \right\|_2^2 \right] \mathrm{d}t. \tag{5.1}$$

The denoising score matching loss measures error in all frequencies, whereas $\mathcal{R}_w$ compares only the low-frequency output with the score of the low-frequency distribution. When the low- and high-frequency components of the data are independent, the low-frequency part of the full score agrees with $\Pi_w \nabla \log p_t^{(w)}$. Dependence between these components can make the two targets differ, because the full score also uses information from the high-frequency coordinates. The following theorem bounds the error in estimating the projected target by a statistical term and two remainder terms associated with dependence across frequencies and energy above the cutoff.

Theorem 5.1 (Score estimation guarantee)**.** Suppose Assumption 4.1 holds. For a sufficiently large n, choose $t_0 = 1/n$, $T = \log n$, and $L = C_L(\log(dn))^3$ for the network architecture $\mathcal{F}$ in Theorem 4.2. For a given cutoff frequency w, set $\gamma_L = \frac{2\beta(L-1)}{(2\beta + d_{\text{eff}}(w))(L+1)}$. Instantiating $M, W, B_w, B_{\text{out},w}$ with $\epsilon = e^{-2} \min\{1, (d^4/n)^{\gamma_L}\}$ and $S_t = \frac{\sqrt{d}}{\sigma_t} \mathrm{polylog}(dn)$, then it holds that

$$\mathbb{E}_{\mathcal{D}_n}[\mathcal{R}_w(\widehat{s})] = \underbrace{\widetilde{\mathcal{O}}\left(L^2 d^5 n^{-\gamma_L}\right)}_{\text{Statistical rate}} + \underbrace{\mathcal{O}(\mathcal{I}(P_{\text{data}}^{(w)}, P_{\text{data}}^{(>w)}))}_{\text{Cross-frequency mutual information}} + \underbrace{\mathcal{O}\left(\left(\mathbb{E}_{X \sim P_{\text{data}}} \|(I - \Pi_w)X\|_2^2\right)^{\beta \wedge 1}\right)}_{\text{Spectral energy above the cutoff}}. \tag{5.2}$$

Here, $P_{\text{data}}^{(>w)} = (I - \Pi_w)_\sharp P_{\text{data}}$ and the mutual information $\mathcal{I}(P_{\text{data}}^{(w)}, P_{\text{data}}^{(>w)}) = \mathrm{KL}(P_{\text{data}}, P_{\text{data}}^{(w)} \otimes P_{\text{data}}^{(>w)})$, where $P_{\text{data}}^{(w)} \otimes P_{\text{data}}^{(>w)}$ denotes the Cartesian product of the low- and high-frequency marginal distributions.

The proof is deferred to Appendix B.2. Several interpretations are in turn.

Statistical estimation rate. The statistical rate in Theorem 5.1 scales as $n^{-\gamma_L}$, which is governed by the number of retained frequencies. The exponent γ_L is independent of the ambient dimension d and depends weakly on network depth: Deeper networks lead to slightly faster rates, but increasing the number of available parallel networks does not degrade the rate. When L is exceedingly large, we have $\gamma_L \to 2\beta/(2\beta + d_{\text{eff}}(w))$, which recovers the nonparametric estimation rate for low-dimensional data (Tsybakov & Zaiats, 2009; Oko et al., 2023). Our explicit depth dependence complements existing minimax analyses of nonparametric score estimation (Wibisono et al., 2024; Dou et al., 2024).

Influence of high frequencies. The two additional terms in (5.2) arise because denoising score matching targets the full data score function, whereas our risk compares the low-frequency output with $\Pi_w \nabla \log p_t^{(w)}$. The mutual information term controls the discrepancy caused by cross-frequency dependency. The spectral energy remainder is caused by discarding high frequencies and depends on the Hölder smoothness. For a generic data distribution, quantitative bounds on the remainders are difficult. However, practical data often exhibit sharp spectral energy decay (Field, 1987; Ruderman & Bialek, 1993). We study the following two common patterns.

Assumption 5.2 . There exist constants $\kappa, \iota, \rho > 0$ such that the spectral energy of P_{data} satisfies

$$\mathbb{E}_{X \sim P_{\text{data}}}[\|(I - \Pi_w)X\|_2^2] \le \begin{cases} \kappa(1+w)^{-2\iota}, & \text{(Polynomial decay)} \\ \kappa e^{-2\rho w}. & \text{(Exponential decay)} \end{cases} \tag{5.3}$$

The following corollary bounds the remainders in Theorem 5.1.

Corollary 5.3. Under the setup of Theorem 5.1 and Assumption 5.2, it holds that

$$\mathbb{E}_{\mathcal{D}_n}[\mathcal{R}_w(\widehat{s})] = \widetilde{\mathcal{O}}\left(L^2 d^5 n^{-\gamma_L}\right) + \begin{cases} \mathcal{O}(\kappa^{1\wedge\beta}(1+w)^{-2\iota(1\wedge\beta)}), & \text{(Polynomial decay)} \\ \mathcal{O}(\kappa^{1\wedge\beta} e^{-2\rho w(1\wedge\beta)}), & \text{(Exponential decay)} \end{cases}.$$

The proof is deferred to Appendix B.3. For fixed sample size, the frequency cutoff w indicates a trade-off. Increasing w reduces the remainders, but increases the effective dimension $d_{\text{eff}}(w)$ in the statistical rate. Thus, a theoretical optimal cutoff can be chosen by minimizing the upper bound in Corollary 5.3. For exponential decay, the order of the optimal cutoff is $w_n^* = \widetilde{\mathcal{O}}(\min\{w_{\max}, \sqrt{\log n}\})$. For polynomial decay, the order becomes $\widetilde{\mathcal{O}}(\min\{w_{\max}, \log n/\log\log n\})$. As the sample size increases, the optimal cutoff increases until $w_{\max}$. This says with more information from the training data, more frequencies can be learned.

5.2 Distribution Recovery

This section establishes distribution estimation guarantees for low-frequency components. The backward process with the learned score function $\widehat{s}$ is

$$\mathrm{d}\widehat{X}_t^{\leftarrow} = \left[\widehat{X}_t^{\leftarrow}/2 + \widehat{s}(\widehat{X}_t^{\leftarrow}, T-t)\right]\mathrm{d}t + \mathrm{d}\overline{W}_t \quad \text{with} \quad \widehat{X}_0^{\leftarrow} \sim \mathsf{N}(0, I).$$

We denote the generated distribution as $\widehat{P}$, which is the marginal distribution of $\widehat{X}_{T-t_0}^{\leftarrow}$ because of early-stopping. Such a sampling process evolves in the full $\mathbb{R}^d$ space, yet our analysis focuses again on the low-frequency component. To ease the presentation, we adopt the continuous-time dynamics in the backward process without discretization, whose analysis has been well established in the existing sampling literature for diffusion models (Chen et al., 2022). The following proposition quantifies the accuracy of the low-frequency pushforward distribution $(\Pi_w)_\sharp \widehat{P}$.

Proposition 5.4 (Low frequency distribution recovery)**.** Suppose Assumption 4.1 holds. For sufficiently large n, set $t_0 = n^{-2}/(8\sqrt{d}(1+C+C^{-1}))$ and $T = t_0 + \log(2en)$, with all other settings as in Theorem 5.1. Then it holds that

$$\mathbb{E}_{\mathcal{D}_n}\left[\mathrm{KL}(P_{\text{data}}^{(w)}, (\Pi_w)_\sharp \widehat{P})\right] = \widetilde{\mathcal{O}}(L^2 d^5 n^{-\gamma_L} + \mathcal{I}(P_{\text{data}}^{(w)}, P_{\text{data}}^{(>w)}) + \left(\mathbb{E}_{X\sim P_{\text{data}}}\|(I-\Pi_w)X\|_2^2\right)^{\beta\wedge 1}).$$

The proof is deferred to Appendix C. The distribution estimation error inherits the score estimation bound. Despite the two remainders, the statistical term matches the minimax lower bound (up to logarithmic factors) when depth L approaches infinity. More precisely, for data distributions satisfying Assumption 4.1 and any estimators $\widehat{P}$, we have

$$\inf_{\widehat{P}} \sup_{P_{\text{data}}} \mathbb{E}_{\mathcal{D}_n}\left[\mathrm{KL}(P_{\text{data}}^{(w)}, (\Pi_w)_\sharp \widehat{P})\right] \gtrsim n^{-2\beta/(2\beta + d_{\text{eff}}(w))}. \tag{5.4}$$

See a proof in Appendix C.1. Here, the supremum is over the worst case distribution, and the infimum includes all possible estimators based on the data set $\mathcal{D}_n$. The mismatch between our guarantees and the minimax lower bound arises from our training procedure. The standard denoising score matching in (3.3) includes all the frequencies, whereas the optimal estimator $\widehat{P}$ is obtained by projecting the data points under Π_w.

6 Experiments

We examine how sample size and norm control affect frequency recovery within fixed network architectures, using both a controlled synthetic distribution and CIFAR-10 (Krizhevsky et al., 2009). In both settings, we vary sample size and use weight decay as a practical surrogate for norm control, evaluating recovery across multiple frequency cutoffs.

6.1 Frequency recovery in simulations

We use a fixed synthetic non-Gaussian distribution on $\mathbb{R}^{15}$ satisfying Assumptions 4.1 and 5.2; its construction and verification are given in Appendix D.1. A cutoff $w \in \{0, \ldots, 7\}$ retains $d_{\text{eff}}(w) = 1 + 2w$ Fourier coordinates. The Fourier coordinates are independent, and the score is analytically available, allowing direct test-set evaluation of R_w. For each (n, λ), we train a full-input parallel ReLU score network g_θ with the same architecture using the weight-decay-regularized DSM objective in Eq. (4.3), with $\lambda_{\text{out}} = \lambda$ for the branch and output parameters. Implementation details are given in Appendix D.2.

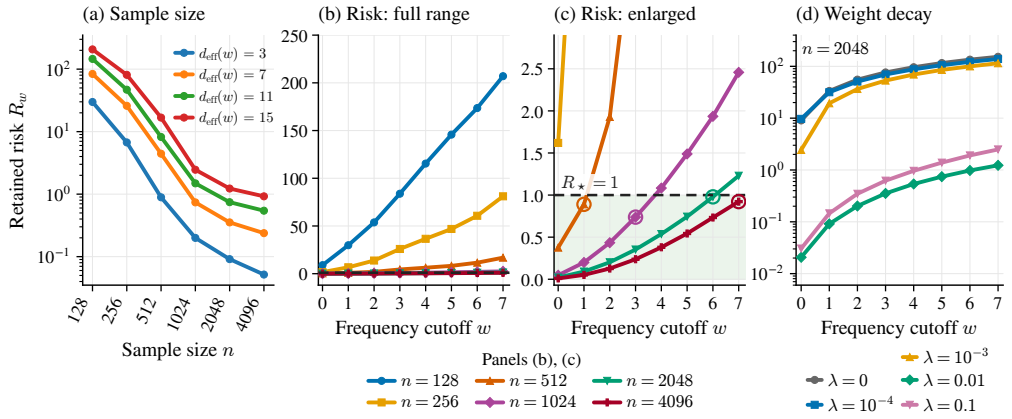


Figure 1: **Sample size, regularization, and frequency recovery.** (a) R_w versus n for $d_{\text{eff}}(w) = 3, 7, 11, 15$ (log–log axes). (b) R_w versus cutoff w for different sample sizes. (c) Enlarged view of (b); shading denotes $R_w \leq 1$, and rings mark the largest empirically learnable cutoffs at this tolerance. (d) Weight-decay comparison at $n = 2048$ (log y-axis). Panels (a–c) use $\lambda = 0.01$. Points show means across five seeds.

Sample size improves low-frequency score estimation. We fix $\lambda = 0.01$ and vary n over powers of two from 128 to 4096. Figure 1(a) shows that R_w decreases with n at all retained dimensions $d_{\text{eff}}(w) = 3, 7, 11, 15$. The log–log curves show an tendency toward shallower declines at larger $d_{\text{eff}}(w)$. This pattern is qualitatively consistent with the dimension dependence of the statistical term $n^{-\gamma_L}$ in Theorem 5.1, where $\gamma_L = 2\beta(L-1)/[(2\beta + d_{\text{eff}}(w))(L+1)]$ decreases with $d_{\text{eff}}(w)$.

More data expand the empirically learnable frequency range. With $\lambda = 0.01$ fixed, Figure 1(b,c) compares the same networks across frequency cutoffs. At tolerance $R_\star$, we call the frequency band up to w empirically learnable if $R_w \leq R_\star$; otherwise, the evaluated models have not reached the target accuracy for that band. Using $R_\star = 1$ as an illustrative post-hoc tolerance, panel (c) marks the largest qualifying cutoffs as $1, 3, 6, 7$ for $n = 512, 1024, 2048, 4096$, respectively. Neither $n = 128$ nor $n = 256$ meets the criterion even at $w = 0$. Thus, more data allow a broader frequency range to meet the same risk requirement without changing the architecture.

Weight decay probes the effect of norm control on frequency recovery. We fix $n = 2048$ and vary the weight-decay coefficient $\lambda \in \{0, 10^{-4}, 10^{-3}, 0.01, 0.1\}$. Figure 1(d) shows that $\lambda = 0.01$ attains the lowest mean R_w at every cutoff among the tested coefficients, while increasing it to 0.1 worsens accuracy. The comparison illustrates how quadratic parameter regularization affects frequency recovery within the fixed architecture: stronger weight decay is not uniformly better.

Together, simulations show that more data expand the empirically learnable frequency range, while weight-decay strength substantially changes the score accuracy attained by the fixed architecture.

6.2 Frequency recovery on CIFAR-10

We train fixed-architecture unconditional DDPM++ models using the Score-SDE implementation (Song et al., 2020b). Models are trained from scratch on nested, class-balanced subsets of $n \in \{5000, 10000, 20000, 40000\}$ CIFAR-10 images, using noise-prediction DSM with weight-decay term $\lambda\|\theta\|_2^2$ over all trainable parameters. Training and sampling details are given in Appendix D.3.

Let x_i and $\widehat{x}_i$ denote real and generated CIFAR-10 images, respectively. For a cutoff w, let Π_w denote the channelwise Fourier projection retaining spatial modes satisfying $\max\{|k_x|, |k_y|\} \leq w$. We define Fourier-projected FID (pFID) as

$$\text{pFID}_w = \text{FID}\big(\{\Pi_w x_i\}_{i=1}^N, \{\Pi_w \widehat{x}_i\}_{i=1}^N\big), \qquad N = 10000, \tag{6.1}$$

where the same FID preprocessing is applied to both projected image sets (Heusel et al., 2017). At $w = 16$, pFID_{16} equals ordinary FID. Implementation details are given in Appendix D.4.

Spectral tail energy decays rapidly with the cutoff. Figure 2(a) reports the empirical mean of $\|(I - \Pi_w)x\|_2^2$ over all 60000 CIFAR-10 images. The tail energy falls from 23.34 at $w = 4$ to 1.79 at $w = 12$. This empirical spectral energy decay is consistent with Assumption 5.2.

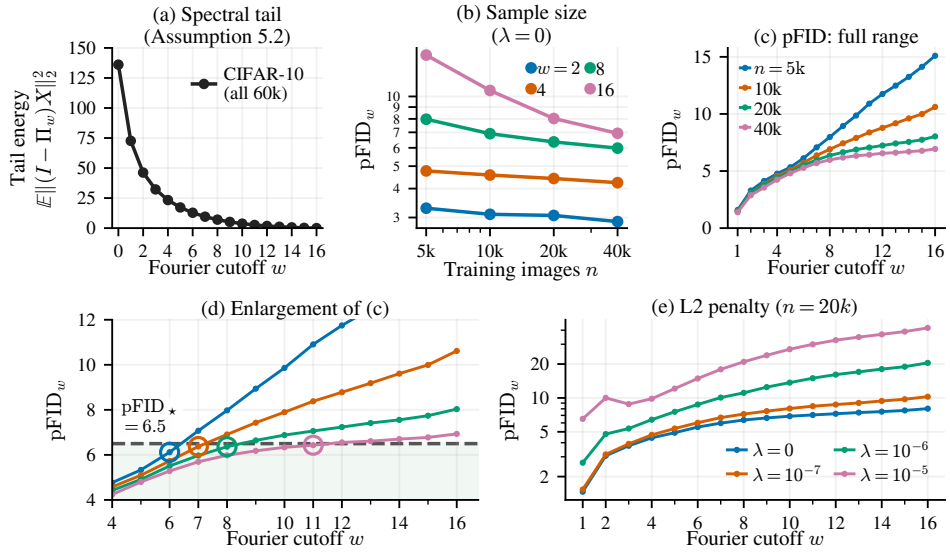


Figure 2: **Frequency-resolved recovery on CIFAR-10.** (a) Empirical spectral tail energy over all 60000 images. (b) pFID_w versus n for $w = 2, 4, 8, 16$ (log–log axes). (c) pFID_w versus cutoff w. (d) Enlarged view of (c); shading denotes $\mathrm{pFID}_w \leq 6.5$, and rings mark the largest cutoffs satisfying this criterion at all lower tested cutoffs as well. (e) Weight-decay comparison at $n = 20000$ (log y-axis). Panels (b–d) use $\lambda = 0$.

Larger datasets extend the empirically recovered frequency range. With $\lambda = 0$, pFID decreases with n at every evaluated cutoff, although the magnitude of improvement varies with w (Figure 2(b,c)). Figure 2(d) summarizes the recovered range using the illustrative tolerance $\mathrm{pFID}_\star = 6.5$, chosen post hoc for visualization. Rings mark the largest tested cutoff satisfying $\mathrm{pFID}_w \leq \mathrm{pFID}_\star$. Larger datasets reach higher qualifying cutoffs, allowing a broader frequency range to meet the same pFID criterion with the same architecture. This parallels the expansion of the score-risk frontier in Section 6.1.

Weight decay has a setting-dependent effect on frequency recovery. At $n = 20000$, we vary the weight-decay coefficient $\lambda \in \{0, 10^{-7}, 10^{-6}, 10^{-5}\}$ (Figure 2(e)). The model with $\lambda = 0$ achieves the lowest pFID at every evaluated cutoff, and each tested increase in λ raises pFID. Unlike the simulation study, where an intermediate coefficient performs best, none of the tested positive coefficients improves on the CIFAR-10 model without added weight decay.

Overall, the CIFAR-10 experiments show that more data improve projected generation quality and extend the frequency range meeting a common pFID criterion, whereas added weight decay does not improve the tested models. Together with the simulations, these results highlight a shared sample-size effect on frequency recovery but a setting-dependent choice of weight-decay strength.

7 Conclusion

We developed a statistical theory of frequency recovery in overparameterized diffusion models, connecting score approximation and estimation to low-frequency distribution recovery. Our bounds separate a statistical term, whose sample-size exponent depends on the number of retained frequencies, data smoothness, and network depth, from remainders reflecting cross-frequency dependence and discarded spectral energy. Under spectral decay, retaining more frequencies reduces the remainder bound but slows the decay of the statistical term with sample size. Cutoff-dependent norm constraints keep these guarantees uniform in the number of available parallel networks at the prescribed depth and width. Synthetic and CIFAR-10 experiments show that larger datasets expand the frequency range meeting their respective empirical accuracy criteria, while the effect of weight decay is setting dependent. Together, these results provide a frequency-resolved account of what diffusion models recover from finite data, beyond a single estimation rate for the full distribution.

AI USE DISCLOSURE

Generative AI was used as a supporting tool during manuscript development, including discussions of experimental choices, refinement of technical writing, code-related assistance, and preparation of figures. All AI-generated or AI-edited content was independently checked by the authors, who remain fully responsible for the correctness of the analysis, references, proofs, experiments, and final manuscript.

REPRODUCIBILITY STATEMENT

The assumptions underlying our analysis are stated in Sections 4–5, with proofs of the main results in Appendices A–C and auxiliary lemmas in Appendix E. Appendices D and F document the data construction and preprocessing, network architectures, training and sampling configurations, checkpoint selection, and frequency-resolved evaluation procedures for the synthetic and CIFAR-10 experiments. We plan to release the code in the future.

REFERENCES

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

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

Roi Benita, Miki Elad, and Joseph Keshet. Spectral analysis of diffusion models with application to schedule design. *Advances in Neural Information Processing Systems*, 38:2073–2127, 2026.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *International Conference on Learning Representations*, volume 2024, pp. 36916–36936, 2024.

Changxiao Cai and Gen Li. Minimax optimality of the probability flow ode for diffusion models. *arXiv preprint arXiv:2503.09583*, 2025.

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, 2023a.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*, 2022.

Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ode is provably fast. *Advances in Neural Information Processing Systems*, 36:68552–68575, 2023b.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

Jonathan Crabbé, Nicolas Huynh, Jan Stanczuk, and Mihaela Van Der Schaar. Time series diffusion in the frequency domain. *arXiv preprint arXiv:2402.05933*, 2024.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Dawei Dong. Spatiotemporal coupling and scaling of natural images and human visual sensitivities. *Advances in neural information processing systems*, 9, 1996.

Zehao Dou, Subhodh Kotekal, Zhehao Xu, and Harrison H Zhou. From optimal score matching to optimal sampling. *arXiv preprint arXiv:2409.07032*, 2024.

Jianqing Fan, Yihong Gu, and Ximing Li. Optimal estimation of a factorizable density using diffusion models with relu neural networks. *arXiv preprint arXiv:2510.03994*, 2025.

David J Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12):2379–2394, 1987.

Simon Foucart, Alain Pajor, Holger Rauhut, and Tino Ullrich. The Gelfand widths of ℓ_p-balls for $0 < p \leq 1$. *Journal of Complexity*, 26(6):629–640, 2010. doi: 10.1016/j.jco.2010.04.004.

Hengyu Fu, Zhuoran Yang, Mengdi Wang, and Minshuo Chen. Unveil conditional diffusion models with classifier-free guidance: A sharp statistical theory. *arXiv preprint arXiv:2403.11968*, 2024.

Asadullah Hill Galib, Pang-Ning Tan, and Lifeng Luo. Fide: Frequency-inflated conditional diffusion model for extreme-aware time series generation. *Advances in Neural Information Processing Systems*, 37:114434–114457, 2024.

Robert M. Gray. *Entropy and Information Theory*. Springer-Verlag, 1 edition, 1990. URL https://ee.stanford.edu/~gray/it.pdf. Corrected first edition, June 2023.

Shivam Gupta, Aditya Parulekar, Eric Price, and Zhiyang Xun. Improved sample complexity bounds for diffusion model training. *Advances in Neural Information Processing Systems*, 37:40976–41012, 2024.

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

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022.

Linjiang Huang, Rongyao Fang, Aiping Zhang, Guanglu Song, Si Liu, Yu Liu, and Hongsheng Li. Fouriscale: A frequency perspective on training-free high-resolution image synthesis. In *European conference on computer vision*, pp. 196–212. Springer, 2024.

John B Ingraham, Max Baranov, Zak Costello, Karl W Barber, Wujie Wang, Ahmed Ismail, Vincent Frappier, Dana M Lord, Christopher Ng-Thow-Hing, Erik R Van Vlack, et al. Illuminating protein space with a programmable generative model. *Nature*, 623(7989):1070–1078, 2023.

Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*, 2022.

Seungwoo Jeong, Junghyo Sohn, Jaehyun Jeon, and Heung-Il Suk. Frequency-conditioned diffusion models for time series generation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 1114–1123, 2025.

Raymond Kan and Cesare Robotti. On moments of folded and truncated multivariate normal distributions. *Journal of Computational and Graphical Statistics*, 26(4):930–934, 2017. doi: 10.1080/10618600.2017.1322092.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Hyeok Kyu Kwon, Dongha Kim, Ilsang Ohn, and Minwoo Chae. Nonparametric estimation of a factorizable density using diffusion models. *Journal of Machine Learning Research*, 27(22): 1–125, 2026.

Yeqing Lin and Mohammed AlQuraishi. Generating novel, designable, and diverse protein structures by equivariantly diffusing oriented residue clouds. *arXiv preprint arXiv:2301.12485*, 2023.

Yu Lu, Yuanzhi Liang, Linchao Zhu, and Yi Yang. Freelong: Training-free long video generation with spectralblend temporal attention. *Advances in Neural Information Processing Systems*, 37: 131434–131455, 2024.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on learning theory*, pp. 1376–1401. PMLR, 2015.

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.

Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

Yurui Qian, Qi Cai, Yingwei Pan, Yehao Li, Ting Yao, Qibin Sun, and Tao Mei. Boosting diffusion models with moving average sampling in frequency domain. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8911–8920. IEEE, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Daniel Ruderman and William Bialek. Statistics of natural images: Scaling in the woods. *Advances in neural information processing systems*, 6, 1993.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4733–4743. IEEE, 2024.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

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

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Arthur Stéphanovitch, Eddie Aamari, and Clément Levrard. Generalization bounds for score-based generative models: a synthetic proof. *arXiv preprint arXiv:2507.04794*, 2025.

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.

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

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Binxu Wang and Cengiz Pehlevan. An analytical theory of spectral bias in the learning dynamics of diffusion models. *Advances in Neural Information Processing Systems*, 38:95865–95963, 2026.

Binxu Wang, Jacob Zavatone-Veth, and Cengiz Pehlevan. A random matrix theory perspective on the consistency of diffusion models. *arXiv preprint arXiv:2602.02908*, 2026.

Zhendong Wang, Jonathan J Hunt, and Mingyuan Zhou. Diffusion policies as an expressive policy class for offline reinforcement learning. *arXiv preprint arXiv:2208.06193*, 2022.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Andre Wibisono, Yihong Wu, and Kaylee Yingxi Yang. Optimal score estimation via empirical bayes smoothing. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 4958–4991. PMLR, 2024.

Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. Freeinit: Bridging initialization gap in video diffusion models. In *European conference on computer vision*, pp. 378–394. Springer, 2024.

Haosen Yang, Adrian Bulat, Isma Hadji, Hai X Pham, Xiatian Zhu, Georgios Tzimiropoulos, and Brais Martinez. Fam diffusion: Frequency and attention modulation for high-resolution image generation with stable diffusion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2459–2468. IEEE, 2025.

Xingyi Yang, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Diffusion probabilistic model made slim. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22552–22562. IEEE, 2023.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural networks*, 94: 103–114, 2017.

Jason Yim, Brian L Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. Se (3) diffusion model with application to protein backbone generation. *arXiv preprint arXiv:2302.02277*, 2023.

Hiroaki Yoshida. A dissipation of relative entropy by diffusion flows. *Entropy*, 19(1):9, 2017. doi: 10.3390/e19010009.

Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. *arXiv preprint arXiv:2403.03954*, 2024.

Kaihong Zhang, Caitlyn H Yin, Feng Liang, and Jingbo Liu. Minimax optimality of score-based diffusion models: Beyond the density lower bound assumptions. *arXiv preprint arXiv:2402.15602*, 2024a.

Zhengqiang Zhang, Ruihuang Li, and Lei Zhang. Frecas: Efficient higher-resolution image generation via frequency-aware cascaded sampling. *arXiv preprint arXiv:2410.18410*, 2024b.

Zixuan Zhang, Kaiqi Zhang, Minshuo Chen, Yuma Takeda, Mengdi Wang, Tuo Zhao, and Yu-Xiang Wang. Nonparametric classification on low dimensional manifolds using overparameterized convolutional residual networks. *Advances in Neural Information Processing Systems*, 37:65738–65764, 2024c.

Zixuan Zhang, Kaixuan Huang, Tuo Zhao, Mengdi Wang, and Minshuo Chen. Diffusion model for manifold data: Score decomposition, curvature, and statistical complexity. *arXiv preprint arXiv:2603.20645*, 2026.

A PROOF OF THEOREM 4.2

We prove the following stronger version of Theorem 4.2. In addition to the projected score, the constructed network approximates a Gaussian score in the discarded frequencies and has a finite output bound. The statistical proof uses both properties.

We use the following quantity throughout the proof:

$$\mathsf{K}_\beta = 4 + \beta + \beta^{-1} + T + C + C^{-1} + c^{-1} + 2H_\beta. \tag{A.1}$$

Theorem A.1. In the setting of Theorem 4.2, for each $w \le w_{\max}$, there is a network $g_\theta \in \mathcal{F}(M, L, W, B_w, B_{\mathrm{out},w}, S_t)$ with

$$S_t = \frac{64\sqrt{d}}{\sigma_t}\mathsf{K}_\beta^2\sqrt{\log(ed/\epsilon)}, \tag{A.2}$$

which satisfies, simultaneously for $t \in [t_0, T]$,

$$\mathbb{E}_{P_t}\left[\left\| g_\theta(X_t, t) - \Pi_w \nabla \log p_t^{(w)}(X_t) + \frac{C(I_d - \Pi_w)X_t}{\alpha_t^2 + C\sigma_t^2}\right\|_2^2\right] \le \frac{\epsilon}{\sigma_t^2}.$$

Theorem A.1 immediately implies Theorem 4.2 since the $(I_d - \Pi_w)$ term vanishes under the projection Π_w. We present the proof in seven major steps.

- Step 1 derives the density and score formulas from the Gaussian transition kernel.
- Step 2 bounds the functions in the score formula.
- Step 3 bounds the error from spatial domain truncation.
- Step 4 uses Taylor polynomials inside the Gaussian integrals and bounds the approximation errors in the score numerator and denominator.
- Step 5 constructs parallel branches for the polynomial terms and a shared network for their ratio.
- Step 6 combines the errors under P_t and bounds the errors in the low- and high-frequency components separately.
- Step 7 verifies the network size and norm bounds stated in the theorem.

• Step 1: derive the density and score formulas.

Fix $w \le w_{\max}$ and choose V_w so that $[U_w, V_w]$ is orthogonal. Write ϕ_{r,s^2} for the density of $\mathsf{N}(0, s^2 I_r)$ and let $G_r \sim \mathsf{N}(0, I_r)$. When $d_{\mathrm{eff}}(w) = d$, the discarded Gaussian factor is one.

⋆ *Factor the density.* Define

$$f_w(u) := \int f(U_w u + V_w v)\phi_{d-d_{\mathrm{eff}}(w),C^{-1}}(v)\,dv.$$

Using the Gaussian transition density of the forward process, we can express the diffused density as

$$p_t^{(w)}(x) = \phi_{d-d_{\mathrm{eff}}(w),\sigma_t^2}(V_w^\top x)\int \phi_{d_{\mathrm{eff}}(w),\sigma_t^2}(U_w^\top x - \alpha_t u) f_w(u)\phi_{d_{\mathrm{eff}}(w),C^{-1}}(u)\,du.$$

The two Gaussian factors satisfy the following identity, obtained by completing the square:

$$\begin{aligned}\phi_{d_{\mathrm{eff}}(w),C^{-1}}(u)\phi_{d_{\mathrm{eff}}(w),\sigma_t^2}(U_w^\top x - \alpha_t u) &= \phi_{d_{\mathrm{eff}}(w),(\alpha_t^2+C\sigma_t^2)/C}(U_w^\top x)\\ &\quad\times \phi_{d_{\mathrm{eff}}(w),\sigma_t^2/(\alpha_t^2+C\sigma_t^2)}\left(u - \frac{\alpha_t U_w^\top x}{\alpha_t^2 + C\sigma_t^2}\right).\end{aligned} \tag{A.3}$$

Consequently, we have that

$$p_t^{(w)}(x) = f_{w,t}(U_w^\top x)\phi_{d_{\mathrm{eff}}(w),(\alpha_t^2+C\sigma_t^2)/C}(U_w^\top x)\phi_{d-d_{\mathrm{eff}}(w),\sigma_t^2}(V_w^\top x),$$

where, for $u \in \mathbb{R}^{d_{\mathrm{eff}}(w)}$, we define

$$f_{w,t}(u) := \mathbb{E}\left[f_w\left(\frac{\alpha_t u}{\alpha_t^2 + C\sigma_t^2} + \frac{\sigma_t G_{d_{\mathrm{eff}}(w)}}{\sqrt{\alpha_t^2 + C\sigma_t^2}}\right)\right].$$

⋆ *Express the approximation target.* Here $\nabla f_{w,t}$ denotes the gradient with respect to its $d_{\text{eff}}(w)$-dimensional argument. Taking the gradient of the logarithm of the density, we find

$$\nabla \log p_t^{(w)}(x) = -\frac{C}{\alpha_t^2 + C\sigma_t^2}\Pi_w x - \frac{(I_d - \Pi_w)x}{\sigma_t^2} + U_w \frac{\nabla f_{w,t}(U_w^\top x)}{f_{w,t}(U_w^\top x)}.$$

Define $q_t(x) = f_{w,t}(U_w^\top x)\phi_{d,(\alpha_t^2+C\sigma_t^2)/C}(x)$. Its score is the target in Theorem A.1:

$$\begin{aligned}\nabla \log q_t(x) &= \Pi_w \nabla \log p_t^{(w)}(x) - \frac{C}{\alpha_t^2 + C\sigma_t^2}(I_d - \Pi_w)x \\ &= -\frac{C}{\alpha_t^2 + C\sigma_t^2}x + U_w \frac{\nabla f_{w,t}(U_w^\top x)}{f_{w,t}(U_w^\top x)}.\end{aligned} \tag{A.4}$$

By differentiating the Gaussian convolution kernel, we can write the gradient of $f_{w,t}$ as

$$\nabla f_{w,t}(u) = \frac{\alpha_t}{\sigma_t\sqrt{\alpha_t^2 + C\sigma_t^2}}\mathbb{E}\left[G_{d_{\text{eff}}(w)} f_w\left(\frac{\alpha_t u}{\alpha_t^2 + C\sigma_t^2} + \frac{\sigma_t G_{d_{\text{eff}}(w)}}{\sqrt{\alpha_t^2 + C\sigma_t^2}}\right)\right]. \tag{A.5}$$

The boundedness of f_w justifies differentiation for $t > 0$, including when $\beta \le 1$.

• **Step 2: bound f_w, $f_{w,t}$, and $\sigma_t \nabla f_{w,t}$.**

The Hölder norm and Gaussian normalization imply $c \le 1 \le H_\beta$. Integration with respect to v in the definition of f_w preserves the Hölder norm bound, so we have

$$\|f_w\|_{\mathcal{H}^\beta} \le H_\beta.$$

Moreover, the scaled gradient satisfies $\|\sigma_t \nabla f_{w,t}(U_w^\top x)\|_2 \le H_\beta$, and $c \le f_{w,t} \le H_\beta$. Appendix E.1 defines the derivative convention and proves these claims in (E.2), (E.3), and (E.4). The Hölder bound on f_w controls the Taylor approximation error in Step 4, and its Taylor coefficients do not depend on t.

• **Step 3: bound the error from spatial domain truncation.**

For $R \ge 1$, define

$$\mathcal{C}_R := \{x : \|U_w^\top x\|_\infty \le R,\ \|V_w^\top x\|_\infty \le R\}.$$

Omit the second condition when $d_{\text{eff}}(w) = d$. We convolve the density upper bound in Assumption 4.1 with the Gaussian transition kernel to obtain

$$p_t(x) \le H_\beta\, \phi_{d,(\alpha_t^2+C\sigma_t^2)/C}(x).$$

Since $(\alpha_t^2 + C\sigma_t^2)/C \le \max\{1, C^{-1}\}$, integrating this density bound over $\mathcal{C}_R^c$ and applying the Gaussian union bound, we obtain

$$\mathbb{P}_{P_t}(X_t \notin \mathcal{C}_R) \le 2dH_\beta \exp\left(-\frac{R^2}{2\max\{1, C^{-1}\}}\right).$$

Using the moment identity $\mathbb{E}[\|G_d\|_2^4] = d(d+2)$, we apply Cauchy–Schwarz to bound the weighted tail integral as follows:

$$\begin{aligned}&\mathbb{E}_{P_t}\left[(1 + \|X_t\|_2^2 + R^2)\,\mathbb{1}\{X_t \notin \mathcal{C}_R\}\right] \\ &\le H_\beta\sqrt{2d}\{1 + R^2 + \sqrt{d(d+2)}\max\{1, C^{-1}\}\}\exp\left(-\frac{R^2}{4\max\{1, C^{-1}\}}\right).\end{aligned} \tag{A.6}$$

By (A.4) and (E.4), we obtain

$$\sigma_t\|\nabla \log q_t(x)\|_2 \le \max\{1, C\}\|x\|_2 + H_\beta/c.$$

Step 5 constructs a network with the global bound

$$\sigma_t\|g_\theta(x,t)\|_2 \le 4\left[\sqrt{d_{\text{eff}}(w)}\left(1 + \frac{4H_\beta}{c}\right) + \sqrt{d}R(1 + \max\{1, C\})\right]. \tag{A.7}$$

The squared triangle inequality and (A.6) then imply

$$\begin{aligned}&\sigma_t^2\mathbb{E}_{P_t}\left[\|g_\theta(X_t,t) - \nabla \log q_t(X_t)\|_2^2\,\mathbb{1}\{X_t \notin \mathcal{C}_R\}\right] \\ &\le 128H_\beta d\{1 + R^2 + \sqrt{d(d+2)}\max\{1, C^{-1}\}\}\exp\left(-\frac{R^2}{4\max\{1, C^{-1}\}}\right) \\ &\quad\times\left[1 + d(1 + \max\{1, C\})^2 + d_{\text{eff}}(w)\left(1 + \frac{4H_\beta}{c}\right)^2 + \frac{H_\beta^2}{c^2}\right].\end{aligned} \tag{A.8}$$

Choose

$$R_\epsilon^2 := 16\max\{1, C^{-1}\}\log\Bigg(\frac{2^{12}edH_\beta[1+d(d+2)\max\{1,C^{-1}\}^2]}{\epsilon} \times \left[1+d(1+\max\{1,C\})^2 + d_{\text{eff}}(w)\left(1+\frac{4H_\beta}{c}\right)^2 + \frac{H_\beta^2}{c^2}\right]\Bigg).$$

Denote the positive quantity inside this logarithm by Ξ_ϵ. It satisfies $\Xi_\epsilon \geq e^3$ and $e^{-R_\epsilon^2/(4\max\{1,C^{-1}\})} = \Xi_\epsilon^{-4}$. We also have that

$$1 + R_\epsilon^2 + \sqrt{d(d+2)}\max\{1, C^{-1}\} \leq [1+d(d+2)\max\{1,C^{-1}\}^2](1+16\log\Xi_\epsilon).$$

Substituting this estimate into (A.8), we obtain

$$\sup_{t\in[t_0,T]} \sigma_t^2 \mathbb{E}_{P_t}\left[\|g_\theta(X_t,t) - \nabla\log q_t(X_t)\|_2^2\, \mathbb{1}\{X_t \notin \mathcal{C}_{R_\epsilon}\}\right] \leq \frac{\epsilon}{32e}(1+16\log\Xi_\epsilon)\Xi_\epsilon^{-3} \leq \frac{\epsilon}{2}.$$

The last inequality uses the fact that $(1+16\log z)z^{-3}$ decreases on $[e^3,\infty)$ and equals $49e^{-9}$ at $z = e^3$.

To check the output cap, we first bound the logarithmic factor using the definition of K_β:

$$\Xi_\epsilon \leq 2^{22}d^4\mathsf{K}_\beta^7/\epsilon, \qquad \log\Xi_\epsilon \leq 4\mathsf{K}_\beta\log(ed/\epsilon).$$

Thus, we obtain

$$R_\epsilon \leq 8\mathsf{K}_\beta\sqrt{\log(ed/\epsilon)}. \tag{A.9}$$

Using $1 + 4H_\beta/c \leq 5\mathsf{K}_\beta^2$ and $1+\max\{1,C\} \leq \mathsf{K}_\beta$ in (A.7), we conclude that

$$\begin{aligned} 4\left[\sqrt{d_{\text{eff}}(w)}\left(1+\frac{4H_\beta}{c}\right) + \sqrt{d}R_\epsilon(1+\max\{1,C\})\right] &\leq 44\sqrt{d}\,\mathsf{K}_\beta^2\sqrt{\log(ed/\epsilon)} \\ &\leq \sigma_t S_t. \end{aligned} \tag{A.10}$$

Here $20/\sqrt{3} + 32 < 44 < 64$.

• Step 4: approximate f_w before Gaussian convolution.

We approximate f_w once, with time-independent Taylor coefficients, and insert this approximation into the integral formulas for $f_{w,t}$ and $\sigma_t\nabla f_{w,t}$. We then bound the errors in both the denominator and the normalized vector numerator.

⋆ *Choose the Taylor approximation domain and mesh.* Choose

$$\delta = \frac{\sqrt{\epsilon}}{2^{12}(1+c^{-1}+H_\beta c^{-2})}, \qquad h = \min\left\{1, \frac{1}{\sqrt{d_{\text{eff}}(w)}}\left(\frac{\delta}{1+H_\beta}\right)^{1/\beta}\right\}. \tag{A.11}$$

In particular, $\delta \leq c/2^{12}$, because $\epsilon \leq 1$. For x with $\|U_w^\top x\|_\infty \leq R_\epsilon$, the Gaussian argument in the expectation defining $f_{w,t}$ has coordinate means of absolute value at most $\max\{1,C^{-1}\}R_\epsilon$ and coordinate variances at most $\max\{1,C^{-1}\}$. We therefore set

$$R_{\text{clean}} = \max\{1,C^{-1}\}R_\epsilon + \sqrt{2\max\{1,C^{-1}\}\log(8d_{\text{eff}}(w)H_\beta^2/\delta^2)}. \tag{A.12}$$

By the Gaussian union bound, the tail probability for these inputs satisfies the following estimate uniformly over all $t \in [t_0, T]$:

$$\begin{aligned} &\mathbb{P}\left\{\frac{\alpha_t U_w^\top x}{\alpha_t^2 + C\sigma_t^2} + \frac{\sigma_t G_{d_{\text{eff}}(w)}}{\sqrt{\alpha_t^2+C\sigma_t^2}} \notin [-R_{\text{clean}}, R_{\text{clean}}]^{d_{\text{eff}}(w)}\right\} \\ &\quad \leq 2d_{\text{eff}}(w)\exp\{-\log(8d_{\text{eff}}(w)H_\beta^2/\delta^2)\} = \frac{\delta^2}{4H_\beta^2}. \end{aligned} \tag{A.13}$$

This radius is independent of t_0.

Partition each coordinate interval $[-R_{\text{clean}}, R_{\text{clean}}]$ into $K_{\text{grid}} = \lceil 2R_{\text{clean}}/h\rceil$ intervals of equal length $h' = 2R_{\text{clean}}/K_{\text{grid}} \leq h$. Use half-open intervals, with the last one closed at its right endpoint, so that each point belongs to exactly one cell. Denote the resulting $N = K_{\text{grid}}^{d_{\text{eff}}(w)}$ cells by Q_z, indexed by their midpoints z. Every $u \in Q_z$ satisfies $\|u-z\|_2 \leq \sqrt{d_{\text{eff}}(w)}\,h'/2 \leq \sqrt{d_{\text{eff}}(w)}\,h$.

⋆ *Apply Taylor's theorem before Gaussian convolution.* Let $k = \lceil \beta \rceil - 1$. On each cell Q_z, take the Taylor polynomial

$$\mathcal{T}_z(u) = \sum_{|r| \le k} \frac{\partial^r f_w(z)}{r!}(u - z)^r.$$

Here r is a multi-index and ∂^r is the corresponding coordinate derivative. By Taylor's theorem and the Hölder bound in Step 2, the error on each cell satisfies

$$\sup_{u \in Q_z} |\mathcal{T}_z(u) - f_w(u)| \le H_\beta(\sqrt{d_{\text{eff}}(w)}h)^\beta \le \delta.$$

For $k = 0$, this is the Hölder continuity bound.

Define the approximation on all of $\mathbb{R}^{d_{\text{eff}}(w)}$ by

$$\widetilde{f}_w(u) = c + \sum_z \mathbb{1}\{u \in Q_z\}\{\mathcal{T}_z(u) - c\}. \tag{A.14}$$

For $u \in [-R_{\text{clean}}, R_{\text{clean}}]^{d_{\text{eff}}(w)}$, exactly one summand is active, so $\widetilde{f}_w(u)$ equals the Taylor polynomial on the cell containing u. Outside this cube, $\widetilde{f}_w(u) = c$. Consequently $c - \delta \le \widetilde{f}_w \le H_\beta + \delta$ everywhere. The network will approximate the Gaussian integrals of these terms, so it need not represent the discontinuous cell indicators.

⋆ *Bound the errors in the numerator and denominator.* For fixed x, t with $\|U_w^\top x\|_\infty \le R_\epsilon$, write the error as

$$e(G_{d_{\text{eff}}(w)}) = (\widetilde{f}_w - f_w)\left(\frac{\alpha_t U_w^\top x}{\alpha_t^2 + C\sigma_t^2} + \frac{\sigma_t G_{d_{\text{eff}}(w)}}{\sqrt{\alpha_t^2 + C\sigma_t^2}}\right).$$

On the event in (A.13), $|e| \le H_\beta$; on its complement, $|e| \le \delta$. Together with the tail bound (A.13), these pointwise estimates imply

$$\mathbb{E}\big[|e(G_{d_{\text{eff}}(w)})|^2\big] \le \delta^2 + H_\beta^2 \frac{\delta^2}{4H_\beta^2} = \frac{5}{4}\delta^2.$$

Define $\widetilde{f}_{w,t}$ by convolving $\widetilde{f}_w$ with the same kernel as $f_{w,t}$. Differentiating this convolution, we obtain

$$\widetilde{f}_{w,t}(U_w^\top x) = \mathbb{E}\left[\widetilde{f}_w\left(\frac{\alpha_t U_w^\top x}{\alpha_t^2 + C\sigma_t^2} + \frac{\sigma_t G_{d_{\text{eff}}(w)}}{\sqrt{\alpha_t^2 + C\sigma_t^2}}\right)\right],$$

$$\sigma_t \nabla \widetilde{f}_{w,t}(U_w^\top x) = \frac{\alpha_t}{\sqrt{\alpha_t^2 + C\sigma_t^2}}\,\mathbb{E}\left[G_{d_{\text{eff}}(w)}\widetilde{f}_w\left(\frac{\alpha_t U_w^\top x}{\alpha_t^2 + C\sigma_t^2} + \frac{\sigma_t G_{d_{\text{eff}}(w)}}{\sqrt{\alpha_t^2 + C\sigma_t^2}}\right)\right].$$

Cauchy–Schwarz bounds the scalar error by 2δ. For the vector error, we apply Cauchy–Schwarz to the inner product with each unit vector:

$$\begin{aligned}\|\sigma_t \nabla \widetilde{f}_{w,t} - \sigma_t \nabla f_{w,t}\|_2 &= \frac{\alpha_t}{\sqrt{\alpha_t^2 + C\sigma_t^2}} \sup_{\|q\|_2 = 1} |\mathbb{E}[(q^\top G_{d_{\text{eff}}(w)})e(G_{d_{\text{eff}}(w)})]| \\ &\le \sup_{\|q\|_2 = 1} \{\mathbb{E}[(q^\top G_{d_{\text{eff}}(w)})^2]\}^{1/2}\{\mathbb{E}[e(G_{d_{\text{eff}}(w)})^2]\}^{1/2} \le 2\delta.\end{aligned}$$

The first inequality uses $\alpha_t/\sqrt{\alpha_t^2 + C\sigma_t^2} \le 1$. The second moment of $q^\top G_{d_{\text{eff}}(w)}$ is one for every unit vector q, so this bound has no additional dimension factor.

⋆ *Express each cell contribution as Gaussian integrals.* Write $[b_{z,i}^-, b_{z,i}^+]$ for the ith interval of Q_z. Endpoint conventions do not affect its Gaussian integral. Define

$$\begin{aligned}T_{z,i,r}((U_w^\top x)_i, t) &= \int_{b_{z,i}^-}^{b_{z,i}^+} (u - z_i)^r \phi_{1,\sigma_t^2/(\alpha_t^2 + C\sigma_t^2)}\left(u - \frac{\alpha_t (U_w^\top x)_i}{\alpha_t^2 + C\sigma_t^2}\right) du, \\ J_{z,i,r}((U_w^\top x)_i, t) &= \frac{\sqrt{\alpha_t^2 + C\sigma_t^2}}{\sigma_t} \int_{b_{z,i}^-}^{b_{z,i}^+} \left(u - \frac{\alpha_t (U_w^\top x)_i}{\alpha_t^2 + C\sigma_t^2}\right)(u - z_i)^r \\ &\quad \times \phi_{1,\sigma_t^2/(\alpha_t^2 + C\sigma_t^2)}\left(u - \frac{\alpha_t (U_w^\top x)_i}{\alpha_t^2 + C\sigma_t^2}\right) du.\end{aligned} \tag{A.15}$$

Let $b_{z,r} = \partial^r f_w(z)/r! - c\,\mathbb{1}\{r = 0\}$. These are the fixed coefficients of $\mathcal{T}_z - c$; in particular, $|b_{z,r}| \le H_\beta + c$. Since the Gaussian density factors across the $d_{\text{eff}}(w)$ coordinates, the denominator and numerator contributions of each cell can be written as

$$\begin{aligned} d_z(U_w^\top x, t) &= \sum_{|r|\le k} b_{z,r} \prod_{i=1}^{d_{\text{eff}}(w)} T_{z,i,r_i}((U_w^\top x)_i, t), \\ u_{z,i}(U_w^\top x, t) &= \frac{\alpha_t}{\sqrt{\alpha_t^2 + C\sigma_t^2}} \sum_{|r|\le k} b_{z,r} J_{z,i,r_i}((U_w^\top x)_i, t) \prod_{j\ne i} T_{z,j,r_j}((U_w^\top x)_j, t). \end{aligned} \tag{A.16}$$

Write $u_z = (u_{z,1}, \dots, u_{z,d_{\text{eff}}(w)})^\top$. The expression for d_z is the integral of $(\mathcal{T}_z - c)\,\mathbb{1}_{Q_z}$ against the Gaussian kernel. For $u_{z,i}$, insert the ith standardized Gaussian coordinate and multiply the integral by $\alpha_t/\sqrt{\alpha_t^2 + C\sigma_t^2}$. Summing the cell contributions in (A.14), we obtain

$$\widetilde{f}_{w,t} = c + \sum_z d_z, \qquad \sigma_t \nabla \widetilde{f}_{w,t} = \sum_z u_z.$$

For the numerator, the background term vanishes because $\mathbb{E}[G_{d_{\text{eff}}(w)}] = 0$. After convolution, many cells contribute at the same input; their contributions are summed before the output network forms the ratio.

The approximation of f_w by Taylor polynomials before integration against the Gaussian kernel follows the construction in Fu et al. (2024, Lemma B.4 and equations (B.13)–(B.27)) and Oko et al. (2023, equations (31)–(32)). Here the interval integrals also allow us to bound the network depth and parameter norms explicitly in Step 7.

• Step 5: construct the parallel branches and shared output network.

We construct one branch for each Taylor cell, with outputs that approximate its denominator and normalized numerator contributions. The shared output network sums these contributions, divides the two sums, and adds the explicit Gaussian part of the score function.

⋆ *Bound the spatial inputs before approximating products.* Each cell branch receives the coordinate-wise clipped input

$$\text{Clip}_{[-R_\epsilon, R_\epsilon]^{d_{\text{eff}}(w)}}(U_w^\top x).$$

The projection is an exact affine map and clipping is an exact ReLU operation. These inputs are bounded for every x and equal $U_w^\top x$ whenever $\|U_w^\top x\|_\infty \le R_\epsilon$.

⋆ *Approximate the Gaussian integrals for one cell.* Apply Lemma E.6 to the cells and accuracies from Step 4. Its hypotheses hold because $|b_{z,r}| \le H_\beta + c$ and $|u - z_i| \le h'/2 \le 1$ on each cell. It produces branch outputs $(\bar{d}_z, \bar{u}_z)$ such that

$$|\bar{d}_z - d_z| \le \frac{\delta}{2N}, \qquad \|\bar{u}_z - u_z\|_2 \le \frac{\delta}{2N}$$

uniformly on the clipped input domain and $[t_0, T]$. Each branch has width at most $4d_{\text{eff}}(w) + 106$ and coefficients bounded by four. Step 7 checks its depth and parameter norms using (E.11).

⋆ *Sum the cell contributions before forming a ratio.* The constant background in the Taylor approximation contributes c to the denominator and zero to the normalized numerator, because $\mathbb{E}[G_{d_{\text{eff}}(w)}] = 0$. The shared network defines the numerator output $\widehat{u}_{w,t}$ and the clipped denominator by

$$\widehat{f}^{\text{raw}}_{w,t} = c + \sum_z \bar{d}_z, \qquad \widehat{u}_{w,t} = \sum_z \bar{u}_z, \qquad \widehat{f}_{w,t} = \text{Clip}_{[c/2, 2H_\beta]}(\widehat{f}^{\text{raw}}_{w,t}).$$

Here "raw" denotes the value before clipping, and $\widehat{u}_{w,t}$ approximates $\sigma_t \nabla f_{w,t}$. All quantities in the following calculation are evaluated at the clipped input specified above. We sum (E.10) over the N cells to bound the errors in the aggregated denominator and numerator:

$$|\widehat{f}^{\text{raw}}_{w,t} - \widetilde{f}_{w,t}| \le \delta/2, \qquad \|\widehat{u}_{w,t} - \sigma_t \nabla \widetilde{f}_{w,t}\|_2 \le \delta/2. \tag{A.17}$$

In Step 4, we established $c - \delta \le \widetilde{f}_{w,t} \le H_\beta + \delta$. Together with $\delta \le c/2^{12}$, this shows that the raw denominator is positive. The clip nevertheless remains part of the shared network: it ensures that its reciprocal is evaluated only on $[c/2, 2H_\beta]$. The numerator is implemented directly from the moments $J_{z,i,r}$; it is not obtained by differentiating the ReLU denominator.

⋆ *Construct the branch for the Gaussian term.* The cell branches supply the non-Gaussian correction to the score. One additional branch supplies its explicit linear Gaussian part and the coefficient σ_t^{-1} needed to undo the numerator normalization. The clipped Gaussian term is

$$G_t(x) = -\frac{C}{\alpha_t^2 + C\sigma_t^2}\left\{U_w \operatorname{Clip}_{[-R_\epsilon,R_\epsilon]^{d_{\text{eff}}(w)}}(U_w^\top x) + V_w \operatorname{Clip}_{[-R_\epsilon,R_\epsilon]^{d-d_{\text{eff}}(w)}}(V_w^\top x)\right\}.$$

On $\mathcal{C}_{R_\epsilon}$, this equals $-(C/(\alpha_t^2 + C\sigma_t^2))x$, the linear part of (A.4). Both coordinate blocks use the same coefficient and fixed clipping radius. Let K be as in (E.9), and set

$$\eta_G = \frac{\delta}{2^{16}(d+1)(1+R_\epsilon)K^2(1+\sigma_{t_0}^{-2})^2}. \tag{A.18}$$

Lemma E.5 approximates σ_t^{-1} and $C/(\alpha_t^2 + C\sigma_t^2)$ to absolute accuracy η_G. For each coordinate use Mult with error η_G and arguments bounded by $4K(1+R_\epsilon)(1+\sigma_{t_0}^{-2})$. The fixed clip is implemented exactly by

$$\operatorname{Clip}_{[-R_\epsilon,R_\epsilon]}(z) = -R_\epsilon + \operatorname{ReLU}(z + R_\epsilon) - \operatorname{ReLU}(z - R_\epsilon).$$

Denote the outputs by $(\bar{G}_t, \bar{s}_t)$, where $\bar{s}_t$ approximates σ_t^{-1}. Each clipped coordinate has magnitude at most R_ϵ, so the coefficient and multiplication errors total at most $(R_\epsilon + 1)\eta_G$ per coordinate. The products are computed in the $[U_w, V_w]$ coordinates before the final exact orthogonal map, so the retained Gaussian output depends only on $U_w^\top x, t$. The orthogonal map $[U_w, V_w]$ preserves the Euclidean norm, so the Gaussian approximation error satisfies

$$\sup_{x,t} \|\bar{G}_t(x) - G_t(x)\|_2 \le \sqrt{d}(R_\epsilon + 1)\eta_G \le \delta/16. \tag{A.19}$$

These bounds hold for every spatial input because the inputs are clipped before multiplication. Also $\delta/16 \le \sqrt{\epsilon}/2^{16}$ by (A.11).

⋆ *Sum the branch outputs.* The cell outputs contain $d_{\text{eff}}(w) + 1$ coordinates. A linear map sums their denominator coordinates into one coordinate and their numerator vectors into the first $d_{\text{eff}}(w)$ coordinates of a d-coordinate block. The remaining coordinates of that block are zero. The Gaussian branch fills another d-coordinate block and the final scalar coordinate. Denote the result by h_θ. Thus, we have

$$h_\theta(x,t) = \left(\sum_z \bar{d}_z, \quad (\widehat{u}_{w,t}^\top, 0_{d-d_{\text{eff}}(w)}^\top), \quad (\bar{G}_t)^\top, \bar{s}_t\right)^\top \in \mathbb{R}^{2d+2}.$$

The shared network adds the constant c to the first coordinate to form the raw denominator. This addition and the matrix that sums the branch outputs are included in B_{out}. The resulting vector has $2d + 2$ coordinates, regardless of the number of Taylor cells.

⋆ *Divide the summed numerator by the summed denominator.* After summing the cell contributions, the shared network forms their ratio. The clipped ratio is

$$\widehat{q}_{w,t} = \operatorname{Clip}_{[-\frac{2H_\beta}{c}, \frac{2H_\beta}{c}]^{d_{\text{eff}}(w)}}\left(\frac{\widehat{u}_{w,t}}{\widehat{f}_{w,t}}\right). \tag{A.20}$$

First clip the numerator coordinates to $[-2H_\beta, 2H_\beta]$. By Step 4 and (A.17), this does not change the sum of the branch outputs. It provides a bounded input interval for the reciprocal and multiplication operations in the shared network. Choose their common accuracy to be

$$\eta_H = \frac{\delta}{2^{16}(d+1)(1+H_\beta+c^{-1})^2(1+\sigma_{t_0}^{-2})}. \tag{A.21}$$

Both multiplication stages use the domain $[-R_H, R_H]^2$, where $R_H = 4(1+H_\beta+c^{-1})^2(1+\sigma_{t_0}^{-2})$. It contains the clipped numerator, reciprocal, ratio, and inverse-noise inputs, including their specified approximation errors. To approximate the reciprocal of a denominator value $r \in [c/2, 2H_\beta]$, put $z = 1 - r/(2H_\beta)$, so $0 \le z \le 1 - c/(4H_\beta)$. We express the reciprocal as a convergent geometric series:

$$\frac{1}{r} = \frac{1}{2H_\beta}\sum_{j=0}^{\infty} z^j.$$

Truncate at $q = \lceil (4H_\beta/c)\log(8/(c\eta_H))\rceil$. The tail is at most $(2/c)(1 - c/(4H_\beta))^{q+1} \le \eta_H/2$. Evaluate the finite polynomial using ReLU multiplication, to additional error $\eta_H/2$, by Lemma E.3.

Denote this approximation by $\overline{r^{-1}}$ and clip its output to $[0, 2/c]$. The clip cannot increase its error from r^{-1}. The shared network then computes the final vector output coordinatewise as

$$g_\theta(x,t) = \bar{G}_t(x) + U_w \left(\mathrm{Mult}\Big(\bar{s}_t, \mathrm{Clip}_{[-\frac{2H_\beta}{c}, \frac{2H_\beta}{c}]} \{ \mathrm{Mult}(\widehat{u}_{w,t,i}, \overline{r^{-1}}(\widehat{f}_{w,t})) \} \Big) \right)_{i=1}^{d_{\mathrm{eff}}(w)}.$$

All scalar arithmetic for the non-Gaussian correction precedes the final map U_w, so its output and arithmetic errors lie in the retained subspace.

⋆ *Bound the error and size of the shared network output.* We first bound the error from implementing the clipped ratio (A.20). Step 6 bounds the resulting error from the target score. The reciprocal and first multiplication contribute coordinate error at most $(2H_\beta + 1)\eta_H$. The clip is nonexpansive, the last product contributes another η_H, and $0 \le \bar{s}_t \le (1+\eta_G)/\sigma_t$. The retained arithmetic error is therefore at most

$$\sqrt{d_{\mathrm{eff}}(w)} \left\{ 2\sigma_{t_0}^{-1}(2H_\beta + 1)\eta_H + \frac{2H_\beta}{c}\eta_G + \eta_H \right\} \le \frac{\delta}{32}.$$

This follows by substituting η_G, η_H and using $\sigma_{t_0}^{-1} \le 1 + \sigma_{t_0}^{-2}$, $2H_\beta/c \le K^2/4$, and $\sqrt{d_{\mathrm{eff}}(w)} \le d+1$. Combining this estimate with the Gaussian branch error $\delta/16$ from (A.19), we obtain

$$\sup_{x,t} \left\| g_\theta(x,t) - G_t(x) - \sigma_t^{-1} U_w \widehat{q}_{w,t} \Big(\mathrm{Clip}_{[-R_\epsilon, R_\epsilon]^{d_{\mathrm{eff}}(w)}}(U_w^\top x) \Big) \right\|_2 \le \delta/8. \tag{A.22}$$

Finally, the bound $\bar{s}_t \le (1+\eta_G)/\sigma_t$ and the ratio clip ensure that the network output satisfies the following bound for all spatial inputs:

$$\sigma_t \sup_x \|g_\theta(x,t)\|_2 \le 4 \left\{ \sqrt{d_{\mathrm{eff}}(w)} \left(1 + \frac{4H_\beta}{c} \right) + \sqrt{d} R_\epsilon (1 + \max(1, C)) \right\}. \tag{A.23}$$

Indeed, the ratio coordinates are bounded by $\frac{2H_\beta}{c}$, the Gaussian inputs are clipped at R_ϵ, and the arithmetic errors satisfy (A.19) and (A.22). This proves the bound (A.7) used in Step 3.

• Step 6: add the approximation errors under the required law.

We combine the Taylor approximation, network, and tail errors under P_t. We then prove a sharper bound on the high-frequency error and show that $\Pi_w g_\theta(x,t)$ depends only on $U_w^\top x$ and t, as needed in the statistical proof.

⋆ *Control the retained ratio on the central domain.* The denominator and numerator errors in Steps 4–5 are each at most 3δ. Clipping preserves these bounds because the targets lie in the clipping intervals. Since $\widehat{f}_{w,t} \ge c/2$, $f_{w,t} \ge c$, and $\|\sigma_t \nabla f_{w,t}\|_2 \le H_\beta$, we subtract the two quotients and bound their difference as follows:

$$\begin{aligned} \left\| \frac{\widehat{u}_{w,t}}{\widehat{f}_{w,t}} - \frac{\sigma_t \nabla f_{w,t}}{f_{w,t}} \right\|_2 &\le \frac{2}{c} \|\widehat{u}_{w,t} - \sigma_t \nabla f_{w,t}\|_2 + \frac{2H_\beta}{c^2} |\widehat{f}_{w,t} - f_{w,t}| \\ &\le 6\delta(c^{-1} + H_\beta c^{-2}) \le \frac{\sqrt{\epsilon}}{512}. \end{aligned}$$

The last inequality follows from (A.11). The ratio clip also preserves the error because $\|\sigma_t \nabla f_{w,t}/f_{w,t}\|_\infty \le H_\beta/c$.

The shared output network forms the reciprocal and final products using the approximation to σ_t^{-1} supplied by the Gaussian branch. The resulting error in the retained correction is at most $\sqrt{\epsilon}/(16\sigma_t)$ by Step 5. The Gaussian-branch error is at most $\sqrt{\epsilon}/(16\sigma_t)$ on $\mathcal{C}_{R_\epsilon}$. Multiplication by the final isometry U_w preserves Euclidean errors. Therefore, we obtain

$$\sup_{x \in \mathcal{C}_{R_\epsilon}} \sigma_t \|g_\theta(x,t) - \nabla \log q_t(x)\|_2 \le \frac{\sqrt{\epsilon}}{4}. \tag{A.24}$$

The three errors fit this bound because $1/512 + 1/16 + 1/16 < 1/4$.

⋆ *Integrate the central and tail parts separately.* The constructed network satisfies the global bound (A.7). Step 3 consequently bounds the normalized mean squared error off the domain $\mathcal{C}_{R_\epsilon}$ by $\epsilon/2$. Equation (A.24) bounds its mean on this domain by $\epsilon/16$. Combining the two parts, we conclude that

$$\sigma_t^2 \mathbb{E}_{P_t} \left[\|g_\theta(X_t,t) - \nabla \log q_t(X_t)\|_2^2 \right] \le \epsilon/16 + \epsilon/2 \le \epsilon.$$

This proves the error claim of Theorem A.1 once Step 7 verifies membership in its network class. The corresponding retained-score bound follows by projection.

⋆ *Bound the retained and discarded errors separately.* Every operation used to compute the retained output receives only $U_w^\top x$ and t; the final multiplication by U_w is exact. The retained Gaussian part also depends only on $U_w^\top x, t$. Thus $\Pi_w g_\theta(x,t)$ is a function of these variables alone. The fitted network class in the statistical theorem remains unrestricted in its spatial inputs.

Projecting an error cannot increase its Euclidean norm. By the central and tail estimates above, the retained error satisfies

$$\mathbb{E}_{P_t}\left[\left\|\Pi_w g_\theta(X_t,t) + \frac{C}{\alpha_t^2 + C\sigma_t^2}\Pi_w X_t - U_w \nabla \log f_{w,t}(U_w^\top X_t)\right\|_2^2\right] \le \frac{9\epsilon}{16\sigma_t^2}. \tag{A.25}$$

The discarded part needs no Taylor approximation. It consists only of the Gaussian branch, so we can bound its error more sharply.

For a centered Gaussian coordinate Y of variance at most $\max\{1, C^{-1}\}$, we apply Cauchy–Schwarz and the Gaussian tail bound to obtain

$$\mathbb{E}[Y^2 \mathbb{1}_{\{|Y|>R_\epsilon\}}] \le (\mathbb{E}[Y^4])^{1/2}\mathbb{P}(|Y|>R_\epsilon)^{1/2} \le \sqrt{6}\,\max\{1,C^{-1}\}e^{-R_\epsilon^2/(4\max\{1,C^{-1}\})}.$$

Recall the density bound $p_t(x) \le H_\beta \phi_{d,(\alpha_t^2+C\sigma_t^2)/C}(x)$ established in Step 3. Clipping displaces a coordinate by at most its absolute value on the tail. Summing the preceding bound over the coordinates of $V_w^\top X_t$, we obtain

$$\mathbb{E}_{P_t}\left[\left\|V_w^\top X_t - \mathrm{Clip}_{[-R_\epsilon,R_\epsilon]^{d-d_{\mathrm{eff}}(w)}}(V_w^\top X_t)\right\|_2^2\right] \le \sqrt{6}\,H_\beta d \max\{1,C^{-1}\}e^{-R_\epsilon^2/(4\max\{1,C^{-1}\})}.$$

Step 5 bounds the Gaussian arithmetic error by $\delta/16 \le \sqrt{\epsilon}/2^{16}$. Using $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$, $\sigma_t \le 1$, and $C/(\alpha_t^2 + C\sigma_t^2) \le \max(1, C)$, we can bound the discarded error by

$$\begin{aligned}&\sigma_t^2 \mathbb{E}_{P_t}\left[\left\|V_w^\top g_\theta(X_t,t) + \frac{C}{\alpha_t^2+C\sigma_t^2}V_w^\top X_t\right\|_2^2\right]\\ &\quad\le \frac{\epsilon}{2^{25}} + 2\sqrt{6}\,H_\beta d \max(1,C)^2 \max\{1,C^{-1}\}e^{-R_\epsilon^2/(4\max\{1,C^{-1}\})}.\end{aligned}$$

The bracketed factors in the definition of Ξ_ϵ dominate $\max\{1, C^{-1}\}$ and $\max(1, C)^2$. Hence $H_\beta d \max(1,C)^2 \max\{1,C^{-1}\} \le \epsilon\Xi_\epsilon/(2^{12}e)$. Substitute $e^{-R_\epsilon^2/(4\max\{1,C^{-1}\})} = \Xi_\epsilon^{-4}$ and $\Xi_\epsilon \ge e^3$: the second term is at most $\epsilon/128$, and so is the first. We obtain

$$\mathbb{E}_{P_t}\left[\left\|V_w^\top g_\theta(X_t,t) + \frac{C}{\alpha_t^2+C\sigma_t^2}V_w^\top X_t\right\|_2^2\right] \le \frac{\epsilon}{64\sigma_t^2}. \tag{A.26}$$

This quantity is zero when $d_{\mathrm{eff}}(w) = d$. Finally, integrating the total approximation bound and applying (B.1), we obtain

$$\int_{t_0}^{T} \mathbb{E}_{P_t}\left[\|g_\theta(X_t,t) - \nabla \log q_t(X_t)\|_2^2\right] dt \le \epsilon \log\left(\frac{e^T-1}{e^{t_0}-1}\right).$$

• Step 7: verify the network size and norm bounds.

It remains to check that the network from Steps 5–6 belongs to the prescribed class. We count its branches and bound their depth and width, then rescale the branches to satisfy the two norm constraints.

⋆ *Choose the network size and norm bounds.* We verify that the following choices suffice for every $w \le w_{\max}$:

$$\begin{aligned}M &\ge 1 + \Big\lceil 64(d_{\mathrm{eff}}(w_{\max}) + \lceil\beta\rceil + 2)^{64(d_{\mathrm{eff}}(w_{\max})+\lceil\beta\rceil+1)}\epsilon^{-d_{\mathrm{eff}}(w_{\max})/(2\beta)}\\ &\qquad\times [\log(ed/\epsilon)]^{d_{\mathrm{eff}}(w_{\max})/2}\mathsf{K}_\beta^{64(d_{\mathrm{eff}}(w_{\max})+\lceil\beta\rceil+1)(1+\beta^{-1})}\Big\rceil, \qquad (A.27)\\ L &\ge 64(d_{\mathrm{eff}}(w_{\max}) + \lceil\beta\rceil + 2)^{64(\lceil\beta\rceil+2)}\mathsf{K}_\beta^{64(\lceil\beta\rceil+2)(1+\beta^{-1})}\\ &\qquad\times \left[\log\left(\frac{ed(1+\sigma_{t_0}^{-2})\mathsf{K}_\beta}{\epsilon}\right)\right]^3, \qquad (A.28)\\ W &\ge 64(d + d_{\mathrm{eff}}(w_{\max}) + \lceil\beta\rceil + 2). \qquad (A.29)\end{aligned}$$

Let $M_\star(\epsilon, w)$ denote the right-hand side of (A.27) with $w_{\max}$ replaced by w. For the chosen depth L, take the norm bounds

$$B_w = 2^{10} L\big\{(d_{\text{eff}}(w)+32)^2[M_\star(\epsilon,w)-1]^{1+1/L} + 16d\big\}, \tag{A.30}$$

$$B_{\text{out},w} = 4L\big\{d + 2^{10}(d_{\text{eff}}(w)+32)^2\big\}. \tag{A.31}$$

$\star$ *Count the active branches.* The definitions of h and R_{clean}, together with (A.9), imply

$$h^{-1} \le \sqrt{d_{\text{eff}}(w)}\, 2^{15/\beta} \mathsf{K}_\beta^{4/\beta} \epsilon^{-1/(2\beta)}, \qquad R_{\text{clean}} \le 16\mathsf{K}_\beta^2 \sqrt{\log(ed/\epsilon)}.$$

Since $R_{\text{clean}} \ge 1$ and $h \le 1$, the grid has at most $(3R_{\text{clean}}/h)^{d_{\text{eff}}(w)}$ cells. Hence the number of cells satisfies

$$N \le 48^{d_{\text{eff}}(w)} d_{\text{eff}}(w)^{d_{\text{eff}}(w)/2} 2^{15 d_{\text{eff}}(w)/\beta} \mathsf{K}_\beta^{(2+4/\beta)d_{\text{eff}}(w)} [\log(ed/\epsilon)]^{d_{\text{eff}}(w)/2} \epsilon^{-d_{\text{eff}}(w)/(2\beta)}.$$

Using $48\sqrt{d_{\text{eff}}(w)} \le (d_{\text{eff}}(w)+k+3)^{7/2}$ and $2^{15} \le \mathsf{K}_\beta^6$, we obtain $N \le M_\star(\epsilon, w) - 1$. Including the Gaussian branch, the construction therefore uses at most $M_\star(\epsilon, w)$ branches.

$\star$ *Bound the depth, width, and layer norms.* Use K, Q, Λ from (E.9) and (E.11). The cell-branch depth is bounded by Lemma E.6. For the other components, the accuracy choices (A.18) and (A.21) satisfy

$$\log\left(\frac{eK(1+\sigma_{t_0}^{-2})}{\eta_G}\right) \le 16\Lambda, \qquad \log(eK/\eta_H) \le 16\Lambda.$$

By (E.7), the Gaussian branch uses at most $2^{12} c_0^8 K^4 \Lambda^3$ layers for the time coefficients and $2^{12}\Lambda$ further layers for parallel coordinate products, clipping, and scaling. The shared reciprocal has at most $c_0 K^2 \log(eK/\eta_H)$ Horner stages, with 24 scalar registers and intermediate magnitudes at most $4K^2$. Applying Lemma E.3 with error $\eta_H/2$ bounds its depth by $2^{19} c_0^2 K^4 \Lambda^3$; the remaining products, clips, and final affine map add at most $2^{12}\Lambda$ layers. Scalar operations are evaluated successively, while products on distinct spatial coordinates are computed in parallel. Every component has depth at most

$$L_* = c_0^{20}(Q+1)^2 K^{20} \Lambda^3, \qquad c_0 = 2^{14}.$$

To compare with (A.28), put $\Lambda_0 = \log(ed(1+\sigma_{t_0}^{-2})\mathsf{K}_\beta/\epsilon)$. The cell count and the definitions of the radii and δ imply

$$K \le (d_{\text{eff}}(w)+k+3)\mathsf{K}_\beta, \qquad \Lambda \le 16(d_{\text{eff}}(w)+k+3)(1+\beta^{-1})\Lambda_0.$$

Substituting these bounds and $Q+1 \le 2(d_{\text{eff}}(w)+k+3)^k$ into L_*, and using $1+\beta^{-1} \le \mathsf{K}_\beta$, we obtain

$$2L_* \le 2^{295}(d_{\text{eff}}(w)+k+3)^{2k+23} \mathsf{K}_\beta^{23} \Lambda_0^3 \le L.$$

For the last inequality, each exponent in (A.28) exceeds the corresponding exponent here by at least 169, and $64 \cdot 4^{169} \ge 2^{295}$.

The widths and squared parameter norms before rescaling are bounded as follows. The norm bound includes the biases and any identity padding.

Component	Width	Squared parameter norm per layer
Cell branch	$4d_{\text{eff}}(w) + 106$	$512(d_{\text{eff}}(w)+32)^2$
Gaussian branch	$10d + 60$	$11360d + 58560$
Shared network	$2d + 10d_{\text{eff}}(w) + 64$	$2d + 3200(d_{\text{eff}}(w)+32)^2$

All arithmetic weights and biases have magnitude at most four. For a cell layer, the squared norm is at most $32(4d_{\text{eff}}(w)+106)^2 \le 512(d_{\text{eff}}(w)+32)^2$. Its input projection has squared norm $2(d_{\text{eff}}(w)+1)$ and satisfies the same bound. In the Gaussian branch, each of the d coordinate blocks has width ten and reads only its own block and sixty shared scalar coordinates; the scalar block reads no spatial coordinates. These zero connections are preserved by the construction in Lemma E.3. The squared norm is therefore at most

$$16\{d[10(10+60)+10] + 60^2 + 60\} = 11360d + 58560.$$

The separate orthogonal input and output maps have squared norms at most $4(d+1)$, including signed coordinates, and satisfy this bound as well. The shared network copies the Gaussian coordinates through $2d$ unit weights, without connections to its arithmetic block of width at most

$10(d_{\text{eff}}(w)+32)$. Its final matrix $[I_d, -I_d, U_w, -U_w]$ has squared Frobenius norm $2(d+d_{\text{eff}}(w))$. This verifies its layer norm bound in the table. Every listed width is at most $32(d+d_{\text{eff}}(w)+k+3) \le W/2$, leaving room for signed identity padding and one constant coordinate.

⋆ *Rescale the branches and include their sum in the shared network.* Pad each branch to depth L. Apply Lemma E.7 with $q = \sqrt{N}$ to each cell branch, and divide its aggregation block by $\sqrt{N}$. This preserves the function and multiplies the squared norm bound, including the added constant coordinate, by $N^{1/L}$. Write A_j^z, b_j^z for the cell-branch parameters and $A_j^{\mathrm{G}}, b_j^{\mathrm{G}}$ for the Gaussian-branch parameters. Applying the same lemma with $q=1$ to the Gaussian branch, we obtain

$$\begin{aligned}&\sum_{z,j}(\|A_j^z\|_F^2 + \|b_j^z\|_2^2) + \sum_j(\|A_j^{\mathrm{G}}\|_F^2 + \|b_j^{\mathrm{G}}\|_2^2)\\ &\quad\le L\{513(d_{\text{eff}}(w)+32)^2 N^{1+1/L} + 11360d + 58561\}\\ &\quad\le 2^{10}L\{(d_{\text{eff}}(w)+32)^2 N^{1+1/L} + 16d\}.\end{aligned} \tag{A.32}$$

The last inequality uses $N \ge 1$ and $58561 \le 54(d_{\text{eff}}(w)+32)^2$. The matrix C_{out} that sums the rescaled outputs has one block of squared norm $(d_{\text{eff}}(w)+1)/N$ per cell and one Gaussian block of squared norm $d+1$. Its squared Frobenius norm is thus

$$\|C_{\text{out}}\|_F^2 = N\frac{d_{\text{eff}}(w)+1}{N} + d + 1 = d + d_{\text{eff}}(w) + 2 \le 2d+2.$$

To include this map in f_{out}, take $[C_{\text{out}}^\top, -C_{\text{out}}^\top]^\top$ as its first matrix with zero bias, apply ReLU, and replace the next matrix A by $[A, -A]$. This preserves the function and doubles only the latter matrix's squared norm. The first layer has width at most $4d+4 \le W$. Since $2L_* \le L$, identity layers can copy the nonnegative hidden coordinates before the final affine map to reach depth L. The table bounds these padding layers as well. The remaining $L-1$ layers and the extra contribution from doubling one matrix therefore have total squared norm at most $L\{2d+3200(d_{\text{eff}}(w)+32)^2\}$. Consequently, we have

$$\begin{aligned}&\sum_{j=1}^{L}(\|A_j^{\text{out}}\|_F^2 + \|b_j^{\text{out}}\|_2^2)\\ &\quad\le 2(d+d_{\text{eff}}(w)+2) + L\{2d+3200(d_{\text{eff}}(w)+32)^2\}\\ &\quad\le 4L\{d+2^{10}(d_{\text{eff}}(w)+32)^2\}.\end{aligned} \tag{A.33}$$

Together with $N \le M_\star(\epsilon, w) - 1$, (A.32) and (A.33) verify (A.30)–(A.31). Set all parameters of unused branches to zero, without adding constant coordinates or identity layers, and add zero rows and columns to reach the prescribed width. This changes neither the function nor the norm bounds. The output cap follows from (A.23) and (A.10), proving Theorem A.1.

⋆ *Deduce the projected approximation and the stated orders.* Apply Π_w to the final affine matrix and bias to obtain $\bar{s}_w = \Pi_w g_\theta$. This preserves the architecture and does not increase either norm bound or the output cap. The resulting function depends only on $(U_w^\top x, t)$. Since $\Pi_w \nabla \log q_t = \Pi_w \nabla \log p_t^{(w)}$, (A.25) proves (4.1). The expectation is unchanged under $P_t^{(w)}$, which has the same retained marginal as P_t.

For fixed $\beta, c, C, H_\beta, T, w_{\max}$, the explicit choices have the orders stated in Theorem 4.2. Indeed, their logarithms satisfy

$$\log(ed/\epsilon) \le \tfrac{3}{2}\log(d/\epsilon), \qquad \log\left(\frac{ed(1+\sigma_{t_0}^{-2})\mathsf{K}_\beta}{\epsilon}\right) \le \left(1+\tfrac{1}{2}\log(2e\mathsf{K}_\beta)\right)\log\left(\frac{d}{\sigma_{t_0}^2\epsilon}\right).$$

The ceiling in $M_\star - 1$ increases its argument by at most a factor of two. Substituting into (A.30) and using $(d_{\text{eff}}(w)+32)^2 \le 1089(1+w)^2$, we conclude that

$$B_w = \widetilde{\mathcal{O}}\Big(L(1+w)^2\{\epsilon^{-d_{\text{eff}}(w)(1+1/L)/(2\beta)} + d\}\Big), \qquad B_{\text{out},w} = \mathcal{O}\big(Ld(1+w)^2\big).$$

The suppressed factors are powers of $\log(ed/\epsilon)$ and constants depending only on the fixed parameters listed above.

B Proof of Theorem 5.1

We prove a stronger score estimation result that controls the full-score error for general training times $0 < t_0 < T$. This result is also used in the sampling proof. We then derive the low-frequency guarantee in Theorem 5.1 by comparing the full score with the low-frequency target.

B.1 A FULL-SCORE BOUND FOR GENERAL TRAINING TIMES

Under Assumption 4.1, fix $0 < t_0 < T$. We use the orthogonal coordinates $[U_w, V_w]$ from Appendix A and the constants in (A.1). For a score function h, define its full-score risk by

$$\mathcal{R}(h) := \int_{t_0}^{T} \mathbb{E}_{P_t}\left[\|h(X_t,t) - \nabla \log p_t(X_t)\|_2^2\right] dt.$$

The difference between the full score and the approximation target in Theorem A.1 has two orthogonal parts. We denote their integrated squared norms by

$$\mathcal{E}_{\mathrm{dep}} := \int_{t_0}^{T} \mathbb{E}_{P_t}\left[\|\Pi_w\{\nabla \log p_t(X_t) - \nabla \log p_t^{(w)}(X_t)\}\|_2^2\right] dt,$$

$$\mathcal{E}_{\mathrm{cutoff}} := \int_{t_0}^{T} \mathbb{E}_{P_t}\left[\left\|(I_d - \Pi_w)\left\{\nabla \log p_t(X_t) + \frac{CX_t}{\alpha_t^2 + C\sigma_t^2}\right\}\right\|_2^2\right] dt.$$

Both terms vanish when $d_{\mathrm{eff}}(w) = d$. For $t \geq 0$, define

$$I_w(t) := \mathcal{I}((\Pi_w)_\sharp P_t, (I_d - \Pi_w)_\sharp P_t)\,,$$
$$K_w(t) := \mathrm{KL}\big(\mathrm{Law}(V_w^\top X_t) \,\big\|\, \mathsf{N}\big(0, (\alpha_t^2/C + \sigma_t^2) I_{d-d_{\mathrm{eff}}(w)}\big)\big)\,.$$

The joint law in this mutual information is induced by the two projections of the same $X_t \sim P_t$. Expressing these projections in the orthonormal bases U_w, V_w leaves the mutual information unchanged. At $t = 0$, we recover the mutual information in Theorem 5.1:

$$I_w(0) = \mathcal{I}(P_{\mathrm{data}}^{(w)}, P_{\mathrm{data}}^{(>w)}).$$

Both $I_w(t)$ and $K_w(t)$ are zero when the discarded block is empty. The inverse-noise integral is

$$J := \int_{t_0}^{T} \sigma_t^{-2}\, dt = \log\left(\frac{e^T - 1}{e^{t_0} - 1}\right). \tag{B.1}$$

Recall that $M_\star(\epsilon, w)$ is the right-hand side of (A.27), with $w_{\max}$ replaced by w. For $n \geq 2$ and the prescribed $w_{\max}$, choose the capacities from (A.27)–(A.29) at accuracy e^{-2}/n. Take the smallest width in that bound and a depth satisfying $L \leq \sqrt{n}$. Thus the architecture is common to all $w \leq w_{\max}$, with

$$\begin{aligned} M &\geq M_\star(e^{-2}/n, w_{\max}), \qquad L \leq \sqrt{n}, \\ W &= 64(d + d_{\mathrm{eff}}(w_{\max}) + \lceil \beta \rceil + 2). \end{aligned} \tag{B.2}$$

For each fixed $w \leq w_{\max}$, use the exponent γ_L from Theorem 5.1 and choose

$$\epsilon_n = e^{-2} \min\{1, (d^4/n)^{\gamma_L}\}. \tag{B.3}$$

Use the local norm budgets in (A.30)–(A.31), evaluated at ϵ_n and the actual depth L:

$$B = B_w, \qquad B_{\mathrm{out}} = B_{\mathrm{out},w}. \tag{B.4}$$

Take the output cap in (A.2) at this same accuracy. Let $\widehat{s}$ be a measurable global minimizer of (3.3) over this class, with exact conditional expectation over the corruption noise and time integration. The architecture and norm budgets are chosen before observing the sample. Define the statistical error bound by

$$\begin{aligned} \mathsf{Est}_n(w) := {} & 2^{19}(d_{\mathrm{eff}}(w) + \lceil \beta \rceil + 2)^{68(d_{\mathrm{eff}}(w) + \lceil \beta \rceil + 1)} \\ & \times \mathsf{K}_\beta^{68(d_{\mathrm{eff}}(w) + \lceil \beta \rceil + 1)(1+\beta^{-1})} JL^2 \\ & \times \left[\log\big(edn(1 + \sigma_{t_0}^{-2})\mathsf{K}_\beta\big)\right]^{d_{\mathrm{eff}}(w)/2+3} \{d^{4\gamma_L} n^{-\gamma_L} + d^5/n\}. \end{aligned} \tag{B.5}$$

The chosen width satisfies $W \leq 80d(d_{\mathrm{eff}}(w) + \lceil \beta \rceil + 2)$ and $W \leq 64d\mathsf{K}_\beta$, as required in the covering proof. Since $\epsilon_n \geq e^{-2}/n$, the common architecture accommodates the local approximation at every $w \leq w_{\max}$.

Theorem B.1. For the estimator and deterministic choices above, it holds that

$$\begin{aligned} \mathbb{E}_{\mathcal{D}_n}[\mathcal{R}(\widehat{s})] &\leq \mathsf{Est}_n(w) + 3\mathcal{E}_{\mathrm{dep}} + 4\mathcal{E}_{\mathrm{cutoff}} \\ &\leq \mathsf{Est}_n(w) + 8\{I_w(t_0) - I_w(T) + K_w(t_0) - K_w(T)\} \\ &\leq \mathsf{Est}_n(w) + 8I_w(0) + 8K_w(0). \end{aligned} \tag{B.6}$$

The bound holds uniformly over the available number M above its prescribed minimum, at the stated depth, width, budgets, and output cap.

To obtain the projected-score bound in Theorem 5.1, we project the full-score error and account for the difference between the two score targets using $\mathcal{E}_{\mathrm{dep}}$. We then set $t_0 = 1/n$ and $T = \log n$ and bound $K_w(0)$ by the discarded spectral energy. These steps are given in Appendix B.3. The proof of the full-score bound has five major steps:

- Step 1 compares the empirical minimizer with a fixed approximation network.
- Step 2 replaces the empirical supremum by a maximum over a finite cover.
- Step 3 uses Bernstein's inequality to bound the maximum empirical fluctuation over the finite cover.
- Step 4 chooses the approximation accuracy and bounds the estimation error.
- Step 5 bounds the dependence and cutoff errors by relative entropy and spectral energy.

B.2 Proof of the full-score estimate

• Step 1: compare the fitted and approximation networks.

Put $H = \sigma_t S_t$, which is independent of t for the prescribed cap. For a fixed network h, subtract the zero-network loss from its conditional denoising loss and write

$$\ell_h^\circ(x) := \int_{t_0}^{T} \mathbb{E}_Z \left[\|h(\alpha_t x + \sigma_t Z, t)\|_2^2 + \frac{2}{\sigma_t} \langle h(\alpha_t x + \sigma_t Z, t), Z \rangle \right] dt. \tag{B.7}$$

Write $Pv = \mathbb{E}_{X_0 \sim P_{\mathrm{data}}}[v(X_0)]$ and $P_n v = n^{-1} \sum_{i=1}^n v(X^{(i)})$. Then $\widehat{\mathcal{L}}(h; \mathcal{D}_n) = P_n \ell_h^\circ + dJ$. For a deterministic comparator $g \in \mathcal{F}$, set $\Delta_h = \ell_h^\circ - \ell_g^\circ$ and

$$\mathcal{E}_{\mathrm{stat}} := \sup_{h \in \mathcal{F}} \left[(P - P_n)\Delta_h - \tfrac{1}{2}\{\mathcal{R}(h) + \mathcal{R}(g)\} \right]_+ . \tag{B.8}$$

Expanding the losses using the Gaussian conditional-score identity $\mathbb{E}[-Z/\sigma_t \mid X_t] = \nabla \log p_t(X_t)$, we find that

$$P\Delta_h = \mathcal{R}(h) - \mathcal{R}(g). \tag{B.9}$$

These terms are finite because conditional Jensen's inequality and the output cap give $\mathcal{R}(h) \le 2J(H^2 + d)$. Since $P_n \Delta_{\widehat{s}} \le 0$, we obtain

$$\begin{aligned} \mathcal{R}(\widehat{s}) - \mathcal{R}(g) &\le (P - P_n)\Delta_{\widehat{s}} \\ &\le \tfrac{1}{2}\{\mathcal{R}(\widehat{s}) + \mathcal{R}(g)\} + \mathcal{E}_{\mathrm{stat}}. \end{aligned}$$

Rearranging proves the empirical oracle inequality

$$\mathcal{R}(\widehat{s}) \le 3\mathcal{R}(g) + 2\mathcal{E}_{\mathrm{stat}}. \tag{B.10}$$

• Step 2: reduce to a finite cover.

⋆ *Choose the input domain.* For $n \ge 2$, set

$$R_X = C^{-1/2}\{\sqrt{d} + \sqrt{2\log(H_\beta n^3)}\}, \qquad R_Z = \sqrt{d} + \sqrt{6\log n}.$$

For $G \sim \mathsf{N}(0, I_d)$, the exponential Markov inequality with $\lambda = u/\{2(\sqrt{d} + u)\}$ gives $\Pr\{\|G\| \ge \sqrt{d} + u\} \le e^{-u^2/2}$. Integrating $p_{\mathrm{data}} \le H_\beta \phi_{d,1/C}$ outside the radius R_X, we bound the two tail probabilities as follows:

$$\Pr\{\|X_0\| > R_X\} \le n^{-3}, \qquad \Pr\{\|Z\| > R_Z\} \le n^{-3}. \tag{B.11}$$

By Cauchy–Schwarz, the first noise moment on the omitted event satisfies

$$\mathbb{E}[\|Z\| \, \mathbb{1}_{\{\|Z\| > R_Z\}}] \le \sqrt{d}\, n^{-3/2}. \tag{B.12}$$

On the event $\{\|X_0\| \le R_X,\ \|Z\| \le R_Z\}$, the bounds $\alpha_t, \sigma_t \le 1$ imply $\|\alpha_t X_0 + \sigma_t Z\| \le R_X + R_Z$ for every $t \in [t_0, T]$. Thus all network inputs on this event belong to

$$\mathcal{K}_R := \{(x,t) : t_0 \le t \le T,\ \|x\|_2^2 + t^2 \le R^2\}, \qquad R := \max\{1, \sqrt{(R_X + R_Z)^2 + T^2}\}.$$

⋆ *Choose representatives in the fitted class.* Apply Lemma E.9 on $\mathcal{K}_R$ at radius $\delta/2$. In each covering ball that meets $\mathcal{F}$, choose one member of $\mathcal{F}$. By the triangle inequality, these representatives form a deterministic δ-cover $h_1, \dots, h_{N_\delta} \in \mathcal{F}$, whose size satisfies

$$\log N_\delta \le \mathcal{H}_\delta := \mathcal{H}_R(\delta/2).$$

Each representative satisfies the output cap, and the bound in (E.17) is independent of the available branch count.

⋆ *Bound the change in the loss.* The centered loss is bounded at every clean input, including those outside the chosen domain. Apply the output cap in (B.7) and use $\mathbb{E}[\|Z\|] \le \sqrt{d}$ to obtain

$$|\ell_h^\circ(x)| \le \int_{t_0}^{T} \frac{H^2 + 2H\mathbb{E}[\|Z\|]}{\sigma_t^2}\, dt \le J(H^2 + 2H\sqrt{d}). \tag{B.13}$$

Choose the representative h_j of a candidate h. On $\mathcal{K}_R$, the representative satisfies $\|\sigma_t(h - h_j)\|_2 \le \delta$, because $\sigma_t \le 1$. Put $h_t = h(\alpha_t x + \sigma_t Z, t)$ and define $h_{j,t}$ in the same way. Factoring the difference of squared losses, we obtain

$$\left| \left\| h_t + \frac{Z}{\sigma_t} \right\|_2^2 - \left\| h_{j,t} + \frac{Z}{\sigma_t} \right\|_2^2 \right| \le \frac{2(H + \|Z\|)}{\sigma_t^2} \|\sigma_t(h_t - h_{j,t})\|_2.$$

For $\|x\| \le R_X$, split the expectation over $Z \sim \mathsf{N}(0, I_d)$ into $\|Z\| \le R_Z$ and its complement. The first contribution is at most $2J(H + R_Z)\delta$. On the complement, both networks satisfy the output cap, so $\|\sigma_t(h_t - h_{j,t})\| \le 2H$; equations (B.11) and (B.12) bound the second contribution by $4HJ(Hn^{-3} + \sqrt{d}\, n^{-3/2})$. For $\|x\| > R_X$, apply (B.13) to both losses. Together these estimates prove, uniformly over the candidate networks,

$$|\ell_h^\circ(x) - \ell_{h_j}^\circ(x)| \le e_\delta(x),$$

where the deterministic upper bound is

$$\begin{aligned} e_\delta(x) := {} & 2J(H + R_Z)\delta + 4HJ\{Hn^{-3} + \sqrt{d}\, n^{-3/2}\} \\ & + 2J(H^2 + 2H\sqrt{d})\, \mathbb{1}_{\{\|x\| > R_X\}}. \end{aligned}$$

⋆ *Replace the empirical supremum by a finite maximum.* By (B.9), the difference between the expressions for h and h_j inside (B.8) satisfies

$$\tfrac{1}{2}P(\ell_h^\circ - \ell_{h_j}^\circ) - P_n(\ell_h^\circ - \ell_{h_j}^\circ) \le \tfrac{1}{2}Pe_\delta + P_n e_\delta.$$

The right side is nonnegative and independent of the candidate h. Taking positive parts and then the supremum, we bound the empirical deviation by a finite maximum and the approximation error:

$$\begin{aligned} \mathcal{E}_{\mathrm{stat}} \le {} & \max_{1 \le j \le N_\delta} \left[(P - P_n)(\ell_{h_j}^\circ - \ell_g^\circ) - \tfrac{1}{2}\{\mathcal{R}(h_j) + \mathcal{R}(g)\} \right]_+ \\ & + \tfrac{1}{2}Pe_\delta + P_n e_\delta. \end{aligned} \tag{B.14}$$

Since $\mathbb{E}_{\mathcal{D}_n}[P_n e_\delta] = Pe_\delta$, the contribution of the last two terms to $2\mathbb{E}_{\mathcal{D}_n}[\mathcal{E}_{\mathrm{stat}}]$ is $3Pe_\delta$. In particular, we have that

$$3Pe_\delta \le 6J(H + R_Z)\delta + 24J(H + \sqrt{d})^2 n^{-3/2}. \tag{B.15}$$

• Step 3: bound the empirical fluctuations.

Fix a representative h_j and put $\Delta_j = \ell_{h_j}^\circ - \ell_g^\circ$ and $V = J(H + \sqrt{d})^2$. The loss difference satisfies $|\Delta_j(x)| \le V$: by the output cap, both normalized squared losses lie in $[0, (H + \|Z\|)^2]$ before averaging over Z. Consequently, for the independent centered variables $Y_i = P\Delta_j - \Delta_j(X^{(i)})$, we have $|Y_i| \le 2V$.

To bound their variance, abbreviate $h_{j,t} = h_j(\alpha_t x + \sigma_t Z, t)$ and $g_t = g(\alpha_t x + \sigma_t Z, t)$. Apply Cauchy–Schwarz in (t, Z) to the factored loss difference to obtain

$$\begin{aligned} |\Delta_j(x)|^2 &\le \left(\int_{t_0}^{T} \mathbb{E}_Z[\|h_{j,t} - g_t\|^2] dt \right) \left(\int_{t_0}^{T} \mathbb{E}_Z[\|h_{j,t} + g_t + 2Z/\sigma_t\|^2] dt \right) \\ &\le 4V \int_{t_0}^{T} \mathbb{E}_Z[\|h_{j,t} - g_t\|^2] dt. \end{aligned}$$

After averaging over X_0 and inserting the true score between the two networks, we obtain

$$\mathrm{Var}(\Delta_j(X_0)) \le 8V\{\mathcal{R}(h_j) + \mathcal{R}(g)\}.$$

By Bernstein's inequality, the empirical deviation satisfies, with probability at least $1 - e^{-u}$,

$$(P - P_n)\Delta_j \le \sqrt{\frac{16V\{\mathcal{R}(h_j) + \mathcal{R}(g)\}u}{n}} + \frac{2Vu}{3n}$$
$$\le \tfrac{1}{2}\{\mathcal{R}(h_j) + \mathcal{R}(g)\} + \frac{26Vu}{3n},$$

where the second inequality is Young's inequality. Replace u by $u + \log N_\delta$, take a union bound over the representatives, and integrate the resulting tail. The expected maximum satisfies

$$\mathbb{E}_{\mathcal{D}_n}\left[\max_j \left[(P - P_n)\Delta_j - \tfrac{1}{2}\{\mathcal{R}(h_j) + \mathcal{R}(g)\}\right]_+\right] \le \frac{26V(\log N_\delta + 1)}{3n}.$$

Together with (B.14) and (B.15), this proves

$$2\mathbb{E}_{\mathcal{D}_n}[\mathcal{E}_{\mathrm{stat}}] \le \frac{52J(H + \sqrt{d})^2(\mathcal{H}_\delta + 1)}{3n} + 6J(H + R_Z)\delta + 24J(H + \sqrt{d})^2 n^{-3/2}. \tag{B.16}$$

• Step 4: choose the approximation accuracy.

⋆ *Use the approximation network.* Throughout this step, write $\epsilon = \epsilon_n$, with ϵ_n defined in (B.3). Choose $g = g_\epsilon$, the deterministic network constructed in Theorem A.1. Its integrated squared error against $\nabla \log q_t$ is at most ϵJ, and its discarded error is at most $\epsilon J/64$ by (A.26).

Differentiating the Gaussian convolution in the retained coordinates, we find that

$$\Pi_w\{\nabla \log p_t(X_t) - \nabla \log p_t^{(w)}(X_t)\} = \frac{\alpha_t}{\sigma_t^2}\{\mathbb{E}[\Pi_w X_0 \mid X_t] - \mathbb{E}[\Pi_w X_0 \mid \Pi_w X_t]\}. \tag{B.17}$$

By orthogonality of the low- and high-frequency subspaces, the squared score discrepancy decomposes as follows:

$$\int_{t_0}^{T} \mathbb{E}_{P_t}\left[\|\nabla \log p_t(X_t) - \nabla \log q_t(X_t)\|_2^2\right] dt = \mathcal{E}_{\mathrm{dep}} + \mathcal{E}_{\mathrm{cutoff}}. \tag{B.18}$$

The conditional expectation of the right side of (B.17) given $\Pi_w X_t$ is zero. Since the retained output of g_ϵ depends only on $(U_w^\top X_t, t)$, its retained approximation error has zero expected inner product with this residual. The discarded cross term is at most $2\sqrt{(\epsilon J/64)\mathcal{E}_{\mathrm{cutoff}}}$. Therefore, we obtain

$$\mathcal{R}(g_\epsilon) \le \epsilon J + \mathcal{E}_{\mathrm{dep}} + \mathcal{E}_{\mathrm{cutoff}} + \frac{1}{4}\sqrt{\epsilon J \mathcal{E}_{\mathrm{cutoff}}}.$$

Substitute into (B.10) and use $\frac{3}{4}\sqrt{\epsilon J \mathcal{E}_{\mathrm{cutoff}}} \le 3\epsilon J + \frac{3}{64}\mathcal{E}_{\mathrm{cutoff}}$ to obtain

$$\mathcal{R}(\widehat{s}) \le 6\epsilon J + 2\mathcal{E}_{\mathrm{stat}} + 3\mathcal{E}_{\mathrm{dep}} + 4\mathcal{E}_{\mathrm{cutoff}}. \tag{B.19}$$

⋆ *Insert the norm budgets.* The zero-padded comparator is admissible for every $M \ge M_\star(\epsilon, w)$. Lemma E.9 covers the union over all finite M at the prescribed $L, W, B, B_{\mathrm{out}}$. Thus (B.16) applies to each enlarged fitted class with the same deterministic budgets.

Within this step, write $A = d_{\mathrm{eff}}(w) + \lceil\beta\rceil + 2$ and $N = M_\star(\epsilon, w) - 1$. We also distinguish the logarithm used in the statistical bounds from the one used in the depth condition by setting

$$\Lambda = \log\left(\frac{edn(1 + \sigma_{t_0}^{-2})\mathsf{K}_\beta}{\epsilon}\right), \qquad \lambda_0 = \log\left(\frac{ed(1 + \sigma_{t_0}^{-2})\mathsf{K}_\beta}{\epsilon}\right)$$

so that $\Lambda = \lambda_0 + \log n$.

From the definition of $M_\star$ in (A.27) and the common width in (B.2), we obtain the following bounds on the branch count, width, and input radius:

$$N \le 128A^{64(A-1)}\mathsf{K}_\beta^{64(A-1)(1+\beta^{-1})} \times [\log(ed/\epsilon)]^{d_{\mathrm{eff}}(w)/2}\epsilon^{-d_{\mathrm{eff}}(w)/(2\beta)},$$
$$W \le 64(2d + \lceil\beta\rceil + 2), \qquad R \le 10\mathsf{K}_\beta\sqrt{d\Lambda}. \tag{B.20}$$

Here $\mathsf{K}_\beta \geq 9$ by normalization, and the radius bound uses $C^{-1/2}, H_\beta, T \leq \mathsf{K}_\beta$ and $\log(H_\beta n^3) \leq 3\Lambda$. The depth condition (A.28) ensures $L-1 \geq 1024A(1+\beta^{-1})\lambda_0$, hence $L-1 \geq \max\{d_{\text{eff}}(w), 512, 128\lambda_0\}$.

Choose $\delta = \epsilon/(1+H+R_Z)$ in (B.16). The term $6J(H+R_Z)\delta$ is then at most $6J\epsilon$. The logarithms in the covering bound satisfy

$$\begin{aligned} Q &:= \log(2R) + \log(2L) + \log(1/\delta) \\ &\quad + \log\{2^{16}W^2(d+(d_{\text{eff}}(w)+32)^2)((d_{\text{eff}}(w)+32)^2N^2+d)\} \\ &\leq 448A(1+\beta^{-1})\Lambda, \qquad L-1 \geq 1024A(1+\beta^{-1})\lambda_0. \end{aligned} \tag{B.21}$$

Indeed, (B.20) implies $\log N \leq 192A(1+\beta^{-1})\Lambda$ and $Q \leq 2\log N + 40\Lambda$. The remaining logarithms are bounded using $\log A, \log \mathsf{K}_\beta, \log\Lambda \leq \Lambda$, $\log L \leq \frac{1}{2}\log n$, and $1+H+R_Z \leq 128\sqrt{d}\mathsf{K}_\beta^2\sqrt{\Lambda}$.

⋆ *Bound the number of retained branches.* We bound the contributions from the local cells and the ambient coordinates separately when taking powers of order $1/(L-1)$. With V_e, O_e, A_h defined in (E.16), the number of branches retained in the internal cover at accuracy δ is

$$K_\delta = \max\left\{1, \left\lceil (32RA_hV_e/\delta)^{2/(L-1)} \right\rceil\right\}.$$

The factor 32 accounts for the external radius $\delta/2$ used to obtain an internal cover. The prescribed caps have the forms $B/L = 2^{10}(d_{\text{eff}}(w)+32)^2N^{1+1/L} + 2^{14}d$ and $B_{\text{out}}/L = 4d + 2^{12}(d_{\text{eff}}(w)+32)^2$. Apply $(x+y)^{1+1/(L-1)} \leq 2^{1/(L-1)}(x^{1+1/(L-1)}+y^{1+1/(L-1)})$ to B/L and use $1+B_{\text{out}}/L \leq 2^{12}(d+(d_{\text{eff}}(w)+32)^2)$. The expression inside the ceiling is at least 64, so rounding increases it by a factor of at most $65/64$. Substituting these bounds into the definition of K_δ, we obtain

$$\begin{aligned} K_\delta \leq \frac{65}{64}\, &2^{22}(d+(d_{\text{eff}}(w)+32)^2) \\ &\times \{(d_{\text{eff}}(w)+32)^2N^{1+2/(L-1)} + 16d\} \\ &\times \left\{2^{49}L(d+(d_{\text{eff}}(w)+32)^2)^3(R/\delta)^2\right\}^{1/(L-1)}. \end{aligned}$$

By the radius and range bounds, the squared ratio satisfies $(R/\delta)^2 \leq 2^{21}\mathsf{K}_\beta^6 d^2\Lambda^2\epsilon^{-2}$. Using $(d+(d_{\text{eff}}(w)+32)^2) \leq 2^7dA^2$, we obtain

$$\begin{aligned} 2^{49}(d+(d_{\text{eff}}(w)+32)^2)^3(R/\delta)^2 &\leq 2^{91}A^6\mathsf{K}_\beta^6 d^5\Lambda^2\epsilon^{-2}, \\ \log(2^{91}A^6\mathsf{K}_\beta^6 d^5) &\leq 33\lambda_0, \\ \{2^{49}L(d+(d_{\text{eff}}(w)+32)^2)^3(R/\delta)^2\}^{1/(L-1)} &\leq e^{35/128}\Lambda^{2/(L-1)}\epsilon^{-2/(L-1)} < \tfrac{4}{3}\Lambda\epsilon^{-2/(L-1)}. \end{aligned}$$

These bounds use $d+(d_{\text{eff}}(w)+32)^2 \leq 2^7dA^2$, $\log A, \log\mathsf{K}_\beta, \log d \leq \lambda_0$, $L-1 \geq 128\lambda_0$, and $\log L/(L-1) \leq 1/64$.

We retain the accuracy power of N while controlling its prefactor. The logarithm of the prefactor multiplying $\epsilon^{-d_{\text{eff}}(w)/(2\beta)}$ in (B.20) is at most $192A(1+\beta^{-1})\lambda_0$. Consequently its $2/(L-1)$ power is at most $e^{3/8} < 2$, by the lower bound on $(L-1)$ in (B.21). Hence, we have that

$$N^{1+2/(L-1)} \leq 256A^{64(A-1)}\mathsf{K}_\beta^{64(A-1)(1+\beta^{-1})}\Lambda^{d_{\text{eff}}(w)/2}\epsilon^{-\frac{d_{\text{eff}}(w)}{2\beta}(1+2/(L-1))}.$$

Finally, using $(d_{\text{eff}}(w)+32) \leq 9A$ and $A \geq 4$, we find that $(d+(d_{\text{eff}}(w)+32)^2)(d_{\text{eff}}(w)+32)^2 \leq (81/16+6561)dA^4 < 2^{13}dA^4$ and $(d+(d_{\text{eff}}(w)+32)^2)d \leq (81+1/16)d^2A^2 < 2^7d^2A^2$. For the ambient-coordinate term only, use $\epsilon^{-2/(L-1)} \leq \exp(2\lambda_0/(L-1)) \leq 2$. The resulting bound on the number of branches used in the cover is

$$\begin{aligned} K_\delta &\leq 2^{44}A^{64(A-1)+4}\mathsf{K}_\beta^{64(A-1)(1+\beta^{-1})} \\ &\quad \times \Lambda^{d_{\text{eff}}(w)/2+1}\{d\epsilon^{-\kappa_L} + d^2\}, \\ \kappa_L &:= \frac{d_{\text{eff}}(w)}{2\beta} + \frac{d_{\text{eff}}(w)/\beta+2}{L-1}. \end{aligned} \tag{B.22}$$

⋆ *Bound the covering number.* For the prescribed norm budgets and $W \geq 64$, the product satisfies

$$(1+B/L)(1+B_{\text{out}}/L) \leq 2^{16}W^2(d+(d_{\text{eff}}(w)+32)^2)((d_{\text{eff}}(w)+32)^2N^2+d).$$

Thus $\log(A_h O_e) \le (L+1)Q/2$ in (E.16). The input radius of the shared output network satisfies $R_h \le 2RO_e$ and $\sqrt{1+R_h^2} \le 3RO_e$. The logarithmic factors for the offset, branch functions, and shared output network in (E.17), evaluated at $\delta/2$, are therefore at most $(L+3)Q/2 \le LQ$. For the amplitude term, use $V_e \le O_e$ and the definition of K_δ to bound its logarithm:

$$\begin{aligned}\log(1+32RK_\delta A_h V_e/\delta) &\le 2\log 2 + (1+2/(L-1))\log(32RA_hV_e/\delta)\\ &\le \{1+(1+2/(L-1))(L+3)/2\}Q \le LQ,\end{aligned}$$

where the last inequality holds for $L \ge 9$.

The exact parameter counts in the composition cover simplify to $d_e = LW^2 + (L+d)W$ and $P_h = LW^2 + LW$. Since $K_\delta \ge 1$, $d \le W$, and $L, W \ge 64$, the covering entropy, including the extra 1 from Bernstein's inequality, satisfies

$$\mathcal{H}_\delta + 1 \le K_\delta LQ(d_e + P_h + W + 3) \le \frac{33}{16} K_\delta L^2 W^2 Q.$$

To incorporate the loss range, use the prescribed width and output cap:

$$W^2 \le 6400 d^2 A^2, \qquad H^2 + d \le 4160 d \mathsf{K}_\beta^4 \Lambda.$$

The width bound uses $2d + \lceil \beta \rceil + 2 \le (5/4)dA$. The prescribed cap also satisfies $H \ge 64\sqrt{d}$, hence $(H+\sqrt{d})^2 \le (33/32)(H^2+d)$. The leading coefficient in (B.16) is therefore at most $18J(H^2+d)$, since $(52/3)(33/32) < 18$. Combining these estimates with $Q \le 448 A \mathsf{K}_\beta \Lambda$, we obtain the following bound for the entropy term in Bernstein's inequality:

$$18(H^2+d)(\mathcal{H}_\delta + 1) \le 27 \cdot 2^{34} A^3 \mathsf{K}_\beta^5 \Lambda^2 K_\delta L^2 d^3. \tag{B.23}$$

⋆ *Combine the error bounds.* Substitute (B.23) into (B.16), with $\delta = \epsilon/(1+H+R_Z)$, and use (B.22). The leading coefficient is $27 \cdot 2^{78} = (27/32)2^{83}$. The approximation error in the oracle inequality, the error from replacing the class by its cover, and the Gaussian tail satisfy

$$\begin{aligned}&6\epsilon J + 6J(H+R_Z)\delta \le 12\epsilon J,\\ &24J(H+\sqrt{d})^2 n^{-3/2} \le 102400 J d \mathsf{K}_\beta^4 \Lambda n^{-3/2}.\end{aligned}$$

Since $dn^{-3/2} \le L^2 d^5/n$, the tail is bounded by the unused $5/32$ of the coefficient 2^{83} below. Hence, we obtain

$$\begin{aligned}6\epsilon J + 2\mathbb{E}_{\mathcal{D}_n}[\mathcal{E}_{\mathrm{stat}}] \le\ & 2^{83} A^{64(A-1)+7}\\ &\times \mathsf{K}_\beta^{64(A-1)(1+\beta^{-1})+5} \Lambda^{d_{\mathrm{eff}}(w)/2+3} J\\ &\times \left\{ \frac{L^2}{n}(d^4 \epsilon^{-\kappa_L} + d^5) + \epsilon \right\}.\end{aligned} \tag{B.24}$$

⋆ *Evaluate the bound at the chosen accuracy.* The reciprocal of one plus the statistical accuracy exponent is

$$\frac{1}{1+\kappa_L} = \frac{2\beta(L-1)}{(2\beta + d_{\mathrm{eff}}(w))(L+1)} = \gamma_L, \qquad 0 < \gamma_L < 1.$$

Write $\Lambda_n = \log(edn(1+\sigma_{t_0}^{-2})\mathsf{K}_\beta)$ for the logarithm in (B.5). First suppose $n \ge d^4$. By (B.3), we have $\epsilon = e^{-2} d^{4\gamma_L} n^{-\gamma_L}$. The two statistical terms satisfy

$$\frac{L^2 d^4}{n} \epsilon^{-\kappa_L} = e^{2\kappa_L} L^2 d^{4\gamma_L} n^{-\gamma_L}, \quad \epsilon \le e^{-2} d^{4\gamma_L} n^{-\gamma_L}.$$

Then $\Lambda \le \Lambda_n + 2 + \gamma_L \log n \le 2\Lambda_n$, because $\mathsf{K}_\beta \ge 6$ gives $\Lambda_n \ge 2 + \log n$. The expression in braces in (B.24) is at most $2e^{2\kappa_L} L^2 \{d^{4\gamma_L} n^{-\gamma_L} + d^5/n\}$. Replacing Λ by $2\Lambda_n$ adds the factor $2^{d_{\mathrm{eff}}(w)/2+4} e^{2\kappa_L}$ in total. The following numerical comparison absorbs it into the coefficient in (B.5).

By the depth bound in (B.21), the accuracy exponent satisfies

$$\kappa_L - \frac{d_{\mathrm{eff}}(w)}{2\beta} = \frac{d_{\mathrm{eff}}(w)/\beta + 2}{L-1} \le \frac{1}{1024\lambda_0} \le \frac{1}{64},$$

where $d_{\mathrm{eff}}(w)/\beta + 2 \le A(1+\beta^{-1})$. The required numerical margin is

$$\frac{A^{4(A-1)-7} \mathsf{K}_\beta^{4(A-1)(1+\beta^{-1})-5}}{2^{d_{\mathrm{eff}}(w)/2+4} e^{2\kappa_L}} > 2^{64}.$$

To check the numerical bound, use $2\kappa_L \le d_{\text{eff}}(w)/\beta + 1/32$ and $\mathsf{K}_\beta \ge 9$. The resulting lower bound increases with $d_{\text{eff}}(w)$. At $d_{\text{eff}}(w) = 1$, the minimum occurs at $\beta = 1$ when $\beta \le 1$; when $\beta \ge 1$, use $(\lceil\beta\rceil + 2)(1 + \beta^{-1}) \ge 6$. In both cases the ratio is at least $4^5 9^{19}/(2^{9/2} e^{33/32}) > 2^{64}$.

The numerator of this ratio contains the additional powers of A and K_β in (B.5) relative to (B.24). The resulting numerical coefficient is $2^{83}/2^{64} = 2^{19}$.

⋆ *Handle the remaining sample sizes.* If $n < d^4$, then $\epsilon = e^{-2}$. Applying the loss range in (B.13) to the empirical deviation in Step 1, we obtain $2\mathbb{E}_{\mathcal{D}_n}[\mathcal{E}_{\text{stat}}] \le 8J(H^2 + 2H\sqrt{d})$: each loss has that range, and the risks subtracted in (B.8) are nonnegative. This is at most $36864 J d \mathsf{K}_\beta^4 \Lambda$. Adding $6\epsilon J = 6e^{-2}J$ bounds the joint quantity by $40960 J d \mathsf{K}_\beta^4 \Lambda$. Since $d < d^5/n$ and $\Lambda = \Lambda_n + 2 \le 2\Lambda_n$, the numerical coefficient needed in this case is at most $81920 < 2^{19}$. The stated problem-dependent and logarithmic powers dominate $\mathsf{K}_\beta^4 \Lambda_n$, so the same bound holds in this case.

The same estimation bound therefore holds in both cases:

$$6\epsilon_n J + 2\mathbb{E}_{\mathcal{D}_n}[\mathcal{E}_{\text{stat}}] \le \mathsf{Est}_n(w), \tag{B.25}$$

where (B.5) specifies the full expression, including the coefficient 2^{19}. It remains to bound the two score residuals.

• Step 5: bound the dependence and cutoff errors.

Let Q_t have the density q_t from Appendix A. By Proposition E.10, the sum of the score residuals is twice the decrease in relative entropy:

$$\mathcal{E}_{\text{dep}} + \mathcal{E}_{\text{cutoff}} = 2\{I_w(t_0) - I_w(T) + K_w(t_0) - K_w(T)\} \le 2\,\mathrm{KL}(P_0 \| Q_0).$$

Together with (B.19) and (B.25), this proves (B.6). We next bound $\mathrm{KL}(P_0\|Q_0)$ by the discarded spectral energy.

Let $\gamma_C = \mathsf{N}(0, C^{-1} I_d)$. The density of Q_0 relative to γ_C is $f_w(U_w^\top x)$. The second derivative of $u \log u$ is at most $1/c$ on $[c, H_\beta]$. Applying Taylor's inequality, we bound the initial relative entropy as follows:

$$2\,\mathrm{KL}(P_0\|Q_0) \le \frac{1}{c} \int \{f(x) - f_w(U_w^\top x)\}^2 \, d\gamma_C(x).$$

The linear term vanishes because, for each fixed u, the integral of $f(U_w u + V_w v) - f_w(u)$ with respect to $\mathsf{N}(0, C^{-1} I_{d - d_{\text{eff}}(w)})$ in v is zero. For independent $Z, Z' \sim \mathsf{N}(0, C^{-1} I_{d-d_{\text{eff}}(w)})$, apply the conditional variance identity and the Hölder bound of order $\beta \wedge 1$ to obtain

$$\begin{aligned}\int \{f(x) - f_w(U_w^\top x)\}^2 \, d\gamma_C(x) &\le \tfrac{1}{2} H_\beta^2 \mathbb{E}[\|Z - Z'\|_2^{2(\beta\wedge 1)}] \\ &\le 2^{(\beta\wedge 1)-1} H_\beta^2 \{(d - d_{\text{eff}}(w))/C\}^{\beta\wedge 1}.\end{aligned}$$

For $\beta > 1$, this bound follows from the bounded first total derivative. The lower density bound $f \ge c$ implies $\mathbb{E}[\|(I_d - \Pi_w)X_0\|_2^2] \ge c(d - d_{\text{eff}}(w))/C$. Substituting this inequality into the preceding bound proves

$$\mathcal{E}_{\text{dep}} + \mathcal{E}_{\text{cutoff}} \le \frac{2^{(\beta\wedge 1)-1} H_\beta^2}{c^{1+(\beta\wedge 1)}} \left(\mathbb{E}[\|(I_d - \Pi_w)X_0\|_2^2]\right)^{\beta\wedge 1}.$$

The same calculation and $\mathrm{KL}(P_0\|Q_0) = I_w(0) + K_w(0)$ imply

$$I_w(0) + K_w(0) \le \frac{2^{(\beta\wedge 1)-2} H_\beta^2}{c^{1+(\beta\wedge 1)}} \left(\mathbb{E}[\|(I_d - \Pi_w)X_0\|_2^2]\right)^{\beta\wedge 1}. \tag{B.26}$$

Dropping the nonnegative mutual information term, we conclude that

$$K_w(0) \le \frac{2^{(\beta\wedge 1)-2} H_\beta^2}{c^{1+(\beta\wedge 1)}} \left(\mathbb{E}[\|(I_d - \Pi_w)X_0\|_2^2]\right)^{\beta\wedge 1}. \tag{B.27}$$

All these quantities vanish when $d_{\text{eff}}(w) = d$.

B.3 Projected risk and the rate in the main theorem

In the definition of $\mathcal{R}_w$ in (5.1), insert $\Pi_w \nabla \log p_t$ between the fitted score and the low-frequency target. Since projection is contractive, the inequality $\|u + v\|^2 \le 2\|u\|^2 + 2\|v\|^2$ and Theorem B.1 imply

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_n}[\mathcal{R}_w(\widehat{s})] &\le 2\mathbb{E}_{\mathcal{D}_n}[\mathcal{R}(\widehat{s})] + 2\mathcal{E}_{\text{dep}} \\ &\le 2\mathsf{Est}_n(w) + 8(\mathcal{E}_{\text{dep}} + \mathcal{E}_{\text{cutoff}}) \\ &\le 2\mathsf{Est}_n(w) + 16\mathcal{I}(P_{\text{data}}^{(w)}, P_{\text{data}}^{(>w)}) + 16 K_w(0).\end{aligned} \tag{B.28}$$

Bound $16K_w(0)$ using (B.27), or bound the sum of the information terms using (B.26). Under Assumption 5.2, the projected-risk bound becomes

$$\mathbb{E}_{\mathcal{D}_n}[\mathcal{R}_w(\widehat{s})] \le 2\mathsf{Est}_n(w) + \frac{2^{(\beta\wedge 1)+2}H_\beta^2}{c^{1+(\beta\wedge 1)}} \begin{cases} \kappa^{\beta\wedge 1}(1+w)^{-2\iota(\beta\wedge 1)}, & \text{polynomial decay,} \\ \kappa^{\beta\wedge 1}e^{-2\rho w(\beta\wedge 1)}, & \text{exponential decay.} \end{cases}$$

This is the remainder bound in Corollary 5.3. It remains to simplify the statistical term at the prescribed times.

At the main theorem's times $t_0 = 1/n$ and $T = \log n$, the inequalities $1 - e^{-x} \le x$ and $1 - e^{-x} \ge x/2$ for $0 < x \le 1$ give $n \le \sigma_{t_0}^{-2} \le 2n$, and (B.1) gives $J \le 2\log n$. To separate the fixed parameters from these time choices, put

$$K_0 = 5 + \beta + \beta^{-1} + C + C^{-1} + c^{-1} + 2H_\beta.$$

Then $\mathsf{K}_\beta \le K_0 \log(edn)$. For the logarithmic factor, the inequalities $1 + \sigma_{t_0}^{-2} \le 3n$ and $\log\log(edn) \le \log(edn)$ imply

$$\Lambda_n \le \{4 + \log(3K_0)\}\log(edn).$$

Substituting these estimates into (B.5), we obtain the explicit statistical bound

$$\begin{aligned} 2\mathsf{Est}_n(w) \le\ & 2^{21}(d_{\mathrm{eff}}(w) + \lceil\beta\rceil + 2)^{68(d_{\mathrm{eff}}(w)+\lceil\beta\rceil+1)} \\ & \times K_0^{68(d_{\mathrm{eff}}(w)+\lceil\beta\rceil+1)(1+\beta^{-1})}\{4 + \log(3K_0)\}^{d_{\mathrm{eff}}(w)/2+3} \\ & \times L^2[\log(edn)]^{68(d_{\mathrm{eff}}(w)+\lceil\beta\rceil+1)(1+\beta^{-1})+d_{\mathrm{eff}}(w)/2+4} \\ & \times \{(d^4/n)^{\gamma_L} + d^5/n\}. \end{aligned} \tag{B.29}$$

The prescribed output cap satisfies

$$S_t \le \frac{64\sqrt{3d}}{\sigma_t} K_0^2[\log(edn)]^{5/2},$$

because $\epsilon_n \ge e^{-2}/n$.

The depth range is nonempty for sufficiently large n. Indeed, let L_n be the least integer satisfying (A.28) at accuracy e^{-2}/n and cutoff $w_{\max}$. The logarithm in that expression is $\Lambda_n + 2$, so we have that

$$\begin{aligned} L_n \le\ & 1 + 64(d_{\mathrm{eff}}(w_{\max}) + \lceil\beta\rceil + 2)^{64(\lceil\beta\rceil+2)} \\ & \times K_0^{64(\lceil\beta\rceil+2)(1+\beta^{-1})}\{6 + \log(3K_0)\}^3[\log(edn)]^{64(\lceil\beta\rceil+2)(1+\beta^{-1})+3}. \end{aligned}$$

For fixed $d, w_{\max}, \beta, C, c, H_\beta$, the right side is at most $\sqrt{n}$ for all sufficiently large n. The exact relation between the finite-depth exponent and its limit is

$$n^{-\gamma_L} = n^{-2\beta/(2\beta+d_{\mathrm{eff}}(w))} \exp\left\{\frac{4\beta\log n}{(2\beta + d_{\mathrm{eff}}(w))(L+1)}\right\}.$$

The logarithm in the common depth bound is at least $\log n$, and its coefficient is at least 64. Hence $n \ge 2$ gives $L \ge 2\log n$, which bounds the exponential factor by e.

Since $0 < \gamma_L < 1$, each term in braces in (B.29) is at most $d^5 n^{-\gamma_L}$. Thus, it holds that

$$2\mathsf{Est}_n(w) = \widetilde{O}_{\beta,C,c,H_\beta,d_{\mathrm{eff}}(w)}(L^2 d^5 n^{-\gamma_L}), \tag{B.30}$$

with the logarithmic powers and constant dependence displayed in (B.29). Combining this with (B.28) and (B.27) gives the rate and remainder terms in Theorem 5.1. Using the joint remainder bound instead proves Corollary 5.3. The constants and logarithmic powers depend on $d_{\mathrm{eff}}(w)$, so this notation does not give a uniform rate over growing cutoffs.

The minimum-depth choice satisfies $L_n = \mathcal{O}_{d,w_{\max},\beta,C,c,H_\beta}((\log n)^{64(\lceil\beta\rceil+2)(1+\beta^{-1})+3})$. At this depth, the statistical term is $\widetilde{O}(n^{-2\beta/(2\beta+d_{\mathrm{eff}}(w))})$ for fixed parameters. The bound is independent of the available number of branches at the stated depth, width, norm budgets, and output cap.

C Proof of Proposition 5.4

We first bound the distribution estimation error for general training times. The following lemma controls the full data law as well as its low-frequency projection, with separate terms for score estimation, Gaussian initialization, and early stopping.

For general $0 < t_0 < T$, run the reverse SDE in Section 5.2 for duration $T - t_0$ and define

$$\widehat{P} = \mathrm{Law}(\widehat{X}^{\leftarrow}_{T-t_0} \mid \mathcal{D}_n).$$

The Gaussian initialization and Brownian motion are independent of the training data, and the score is evaluated only at forward times in $[t_0, T]$. At reverse time $T - t_0$, the exact reverse process initialized from P_T has distribution P_{t_0}. The proposition compares $\widehat{P}$ with P_{data}, so we must also control the error from early stopping. For this comparison, the proof uses an exact reverse process that runs from P_{T-t_0} to P_{data} in the same duration $T - t_0$. Its score is evaluated at forward times that are t_0 earlier than those used by the learned sampler. For the initialization and stopping errors, define

$$\begin{aligned} \mathcal{E}_{\mathrm{init}}(h) &:= \frac{1}{2}\left[e^{-h}\mathbb{E}\big[\|X_0\|_2^2\big] + d\{-e^{-h} - \log(1-e^{-h})\}\right], \qquad h > 0, \\ \mathcal{E}_{\mathrm{stop}} &:= \mathrm{KL}(P_{\mathrm{data}} \| P_{t_0}). \end{aligned} \tag{C.1}$$

Lemma C.1 (From score error to distribution error)**.** Suppose Assumption 4.1 holds. Let $0 < t_0 \le 1, T > t_0$, and suppose the fitted finite ReLU network satisfies $\|\widehat{s}(x,t)\| \le H/\sigma_t$ for a deterministic constant H. Use the full-score risk $\mathcal{R}$ from Appendix B.1 and J from (B.1). For every real $q > 2$ satisfying

$$q\, t_0 \sqrt{d}\, \frac{|C-1|}{\min\{1, C\}} \le \frac{1}{2}, \tag{C.2}$$

we have

$$\begin{aligned} &\mathbb{E}_{\mathcal{D}_n}\big[\mathrm{KL}((\Pi_w)_\sharp P_{\mathrm{data}} \| (\Pi_w)_\sharp \widehat{P})\big] \\ &\quad \le \mathcal{E}_{\mathrm{init}}(T - t_0) + 2\mathcal{E}_{\mathrm{stop}} \\ &\qquad + \frac{2H_\beta}{c}\big(\mathbb{E}_{\mathcal{D}_n}\big[\mathcal{R}(\widehat{s})\big]\big)^{1-2/q}\left[2\{H^2 + \sqrt{d(d+2)}\}J\right]^{2/q}. \end{aligned} \tag{C.3}$$

The inequality also holds without projection. Its short-time term satisfies

$$\begin{aligned} \mathcal{E}_{\mathrm{stop}} \le t_0^{\beta\wedge 1}\Bigg[\frac{8H_\beta^2 d^{\beta\wedge 1}}{c}\Bigg\{(2/C)^{\beta\wedge 1}\left(\frac{1/2 + |C-1|}{\min\{1,C\}}\right)^{2(\beta\wedge 1)} + \min\{1,C\}^{-(\beta\wedge 1)}\Bigg\} \\ + 4H_\beta d\left(\frac{|C-1|}{\min\{1,C\}}\right)^2\Bigg]. \end{aligned} \tag{C.4}$$

The projected bound follows from the bound for the full distribution by data processing. Steps 1–4 prove the lemma. Step 5 combines it with the full-score estimate in Theorem B.1, using the times in (C.11), to obtain (C.14) for the stopped sampler defined above. We proceed as follows.

- Step 1 applies Girsanov's theorem to compare the learned sampler with the exact reverse process ending at P_{data}.
- Step 2 bounds the error from shifting the score time by t_0 using $\mathrm{KL}(P_{\mathrm{data}} \| P_{t_0})$.
- Step 3 bounds the score error averaged under P_{t-t_0} by the score error averaged under P_t, which appears in $\mathcal{R}$.
- Step 4 bounds the initialization and early-stopping errors and obtains the distribution bound.
- Step 5 chooses t_0 and T so that these two errors decay faster than the statistical term, then applies the score estimation bound.

• **Step 1: apply Girsanov's theorem to the reverse processes.** Condition on the training data. Set $h = T - t_0$ and compare this sampler with the exact reverse of the forward process on $[0, h]$. The exact reverse process starts at P_h, has law P_{h-u} at time u, and ends at P_{data}. Its drift is $x/2 + \nabla \log p_{h-u}(x)$, whereas the fitted drift uses time $T - u$. The reverse drift formula applies away from forward time zero (Haussmann & Pardoux, 1986, Theorem 2.1 and equation (2.3), pp. 1189–1190).

Write $v = h - u$. The difference of scores splits as

$$\widehat{s}(x, v + t_0) - \nabla \log p_v(x) = \widehat{s}(x, v + t_0) - \nabla \log p_{v+t_0}(x) + \nabla \log p_{v+t_0}(x) - \nabla \log p_v(x).$$

The first difference is the estimation error evaluated under P_v. The second compares the true scores at two times; Steps 2–3 bound these terms separately.

For $\delta > 0$, both drifts are globally Lipschitz on $[0, h-\delta]$. For the true drift, the conditional Gaussian formula implies

$$\nabla^2 \log p_v(x) = -\sigma_v^{-2} I_d + \alpha_v^2 \sigma_v^{-4} \operatorname{Cov}(X_0 \mid X_v = x),$$
$$\operatorname{Cov}(X_0 \mid X_v = x) \preceq \frac{H_\beta}{c} \frac{\sigma_v^2}{\alpha_v^2 + C\sigma_v^2} I_d.$$

The covariance bound follows by bounding the conditional density by H_β/c times the density of $\mathsf{N}(\alpha_v x/(\alpha_v^2 + C\sigma_v^2), \{\sigma_v^2/(\alpha_v^2 + C\sigma_v^2)\} I_d)$. It implies $\|\nabla^2 \log p_v\|_{\mathrm{op}} \le (1 + H_\beta/c)/\sigma_\delta^2$. The fitted finite ReLU network is globally Lipschitz, so both processes are nonexplosive on this interval.

Apply Girsanov after stopping the drift difference on exit from a ball (Oksendal, 2013, Theorem 8.6.6, pp. 167–168). The stopped difference is bounded. As the ball grows, the stopped learned process converges to the learned process by nonexplosion; the path-entropy bound passes to the limit by lower semicontinuity of entropy. The entropy chain rule adds $\mathrm{KL}(P_h \| \mathsf{N}(0, I_d))$ for initialization. The factor $1/2$ in that formula combines with the inequality $\|a + b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ to bound the endpoint entropy by

$$\begin{aligned} \mathrm{KL}(P_\delta \| \operatorname{Law}(\widehat{X}^{\leftarrow}_{h-\delta})) \le\ & \mathrm{KL}(P_h \| \mathsf{N}(0, I_d)) \\ & + \int_\delta^h \mathbb{E}_{P_v}\left[\|\widehat{s}(X, v + t_0) - \nabla \log p_{v+t_0}(X)\|^2\right] dv \\ & + \int_\delta^h \mathbb{E}_{P_v}\left[\|\nabla \log p_{v+t_0} - \nabla \log p_v\|^2\right] dv. \end{aligned} \tag{C.5}$$

• **Step 2: bound the error from shifting the score time by t_0.** The last integral in (C.5) compares $\nabla \log p_v$ with $\nabla \log p_{v+t_0}$. Both densities solve the same OU equation as functions of v. Differentiating their relative entropy and integrating by parts, we obtain

$$\frac{d}{dv} \mathrm{KL}(P_v \| P_{v+t_0}) = -\frac{1}{2} \mathbb{E}_{P_v}\left[\|\nabla \log p_v - \nabla \log p_{v+t_0}\|^2\right].$$

For $v \ge \delta > 0$, the Gaussian bounds justify integration by parts using compact cutoffs and then letting their supports grow. Integrating this identity expresses the time-shift error as

$$\int_\delta^h \mathbb{E}_{P_v}\left[\|\nabla \log p_v - \nabla \log p_{v+t_0}\|^2\right] dv = 2\{\mathrm{KL}(P_\delta \| P_{\delta+t_0}) - \mathrm{KL}(P_h \| P_T)\}. \tag{C.6}$$

As $\delta \downarrow 0$, the first entropy converges to $\mathcal{E}_{\mathrm{stop}}$ by dominated convergence, using the Gaussian bounds. The integral is therefore at most $2\mathcal{E}_{\mathrm{stop}}$ in this limit.

• **Step 3: change the expectation from P_{t-t_0} to P_t.** Set $t = v + t_0$ in the first integral in (C.5). Its score error is evaluated under P_{t-t_0}, whereas $\mathcal{R}$ integrates the error under P_t. The change of density is

$$\begin{aligned} & \mathbb{E}_{X \sim P_{t-t_0}}\left[\|\widehat{s}(X, t) - \nabla \log p_t(X)\|^2\right] \\ & \quad = \mathbb{E}_{X \sim P_t}\left[\frac{p_{t-t_0}(X)}{p_t(X)} \|\widehat{s}(X, t) - \nabla \log p_t(X)\|^2\right]. \end{aligned}$$

We bound the $L^q(P_t)$ norm of this density ratio, then apply Hölder's inequality and a fourth-moment bound on the score error. Completing the square in the Gaussian convolution, we obtain

$$c \le \frac{p_t(x)}{\phi_{d,(\alpha_t^2 + C\sigma_t^2)/C}(x)} \le H_\beta.$$

For this calculation only, put $z = (\alpha_{t-t_0}^2/C + \sigma_{t-t_0}^2)/(\alpha_t^2/C + \sigma_t^2)$. Then $|z - 1| \le t_0 |C - 1| / \min\{1, C\}$. Direct Gaussian integration bounds the norm of the density ratio by

$$\left\| \frac{p_{t-t_0}}{p_t} \right\|_{L^q(P_t)} \le H_\beta c^{-1+1/q} z^{-d(q-1)/(2q)} [q - (q-1)z]^{-d/(2q)}. \tag{C.7}$$

Write $z = 1 + u$. By condition (C.2), $|u| \le 1/(2q\sqrt{d})$. Since the derivative satisfies

$$\left| \frac{d}{du}\{-(q-1)\log(1+u) - \log(1-(q-1)u)\} \right| \le 2q(q-1)|u|,$$

the logarithm of the Gaussian factor in (C.7) is at most $d(q-1)u^2/2 \le 1/(8q) < \log 2$. Thus, using $c \le 1$, we obtain

$$\sup_{t_0 \le t \le T} \|p_{t-t_0}/p_t\|_{L^q(P_t)} \le 2H_\beta/c. \tag{C.8}$$

We next control the higher moment needed to use this ratio. The conditional Gaussian score identity is $\nabla \log p_t(X_t) = -\mathbb{E}[Z \mid X_t]/\sigma_t$. By conditional Jensen's inequality, $\mathbb{E}\big[\|\nabla \log p_t(X_t)\|^4\big] \le d(d+2)/\sigma_t^4$. By Minkowski's inequality and the output cap, the square root of the fourth moment satisfies

$$\big(\mathbb{E}_{\mathcal{D}_n, P_t}\big[\|\widehat{s} - \nabla \log p_t\|^4\big]\big)^{1/2} \le \frac{2\{H^2 + \sqrt{d(d+2)}\}}{\sigma_t^2}.$$

Interpolating between the first and second moments of the *squared* error and applying Hölder's inequality with (C.8), we find

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_n, P_{t-t_0}}\big[\|\widehat{s} - \nabla \log p_t\|^2\big] \\
&\le \frac{2H_\beta}{c}\big(\mathbb{E}_{\mathcal{D}_n, P_t}\big[\|\widehat{s} - \nabla \log p_t\|^2\big]\big)^{1-2/q}\left[\frac{2\{H^2 + \sqrt{d(d+2)}\}}{\sigma_t^2}\right]^{2/q}.
\end{aligned}$$

The interpolation exponents satisfy $1/(q/(q-1)) = (1-2/q) + (2/q)/2$. Applying Hölder's inequality to the time integral, we obtain

$$\begin{aligned}
&\int_{t_0}^{T} \mathbb{E}_{\mathcal{D}_n, P_{t-t_0}}\big[\|\widehat{s} - \nabla \log p_t\|^2\big] dt \\
&\le \frac{2H_\beta}{c}\big(\mathbb{E}_{\mathcal{D}_n}\big[\mathcal{R}(\widehat{s})\big]\big)^{1-2/q}\Big[2\{H^2 + \sqrt{d(d+2)}\}J\Big]^{2/q}.
\end{aligned} \tag{C.9}$$

• Step 4: bound initialization and early stopping and conclude the lemma.

⋆ *Initialization.* By convexity of $\mathrm{KL}(\cdot\|\mathsf{N}(0, I_d))$, the Gaussian mixture P_h satisfies

$$\mathrm{KL}(P_h\|\mathsf{N}(0, I_d)) \le \mathcal{E}_{\mathrm{init}}(h).$$

Evaluating the relative entropy of each Gaussian mixture component yields the mean and variance terms in (C.1).

⋆ *Error from stopping at positive forward time.* Use $\mathrm{KL}(p\|q) \le \int (p-q)^2/q$. At the full cutoff, the Gaussian product identity in (A.3) expresses the density as

$$p_t(x) = \phi_{d,(\alpha_t^2 + C\sigma_t^2)/C}(x)\mathbb{E}\big[f\big(\frac{\alpha_t}{\alpha_t^2 + C\sigma_t^2}x + \frac{\sigma_t}{\sqrt{\alpha_t^2 + C\sigma_t^2}}G\big)\big].$$

Here $G \sim \mathsf{N}(0, I_d)$. The function f is globally Hölder continuous of order $\beta \wedge 1$ with constant H_β. We also have that

$$|1 - \frac{\alpha_t}{\alpha_t^2 + C\sigma_t^2}| \le t\frac{1/2 + |C-1|}{\min\{1, C\}}, \qquad \frac{\sigma_t^2}{\alpha_t^2 + C\sigma_t^2} \le t/\min\{1, C\}.$$

Subtract the two densities by first changing their bounded factors and then their Gaussian factors. Applying the inequality $(a+b)^2 \le 2a^2 + 2b^2$ to these two differences, we obtain

$$\begin{aligned}
\int \frac{(p_0 - p_t)^2}{p_t} \le \frac{2}{c}\int \frac{\phi_{d,C^{-1}}^2}{\phi_{d,(\alpha_t^2+C\sigma_t^2)/C}}\left| f(x) - \mathbb{E}\left[f\left(\frac{\alpha_t x}{\alpha_t^2 + C\sigma_t^2} + \frac{\sigma_t G}{\sqrt{\alpha_t^2 + C\sigma_t^2}}\right)\right]\right|^2 dx \\
+ 2H_\beta \int \frac{(\phi_{d,C^{-1}} - \phi_{d,(\alpha_t^2+C\sigma_t^2)/C})^2}{\phi_{d,(\alpha_t^2+C\sigma_t^2)/C}}\, dx.
\end{aligned} \tag{C.10}$$

If $t|C-1|/\min\{1, C\} \le 1/(2\sqrt{d})$, the Gaussian weight in the first integral has mass at most 2 and, after normalization, covariance at most $(2/C)I_d$. Indeed, with $z = 1/(\alpha_t^2 + C\sigma_t^2)$, its mass is $[z(2-z)]^{-d/2}$ and its normalized covariance is $[C(2-z)]^{-1}I_d$. The second Gaussian integral is this mass minus one, at most $2d(t|C-1|/\min\{1, C\})^2$.

Hölder continuity and Jensen's inequality bound the squared difference by

$$2H_\beta^2\{|1-\frac{\alpha_t}{\alpha_t^2+C\sigma_t^2}|^{2(\beta\wedge1)}\|x\|^{2(\beta\wedge1)}+(\sigma_t^2/(\alpha_t^2+C\sigma_t^2))^{\beta\wedge1}d^{\beta\wedge1}\}.$$

Insert the Gaussian moment bounds in (C.10) and use $t^{2(\beta\wedge1)}\le t^{\beta\wedge1}$, $t^2\le t^{\beta\wedge1}$ for $t\le1$. This proves (C.4). Its condition follows from (C.2) since $q>2$.

⋆ *Clean endpoint and projection.* Let $\delta\downarrow0$ in (C.5). By continuity of the paths, the laws of the true and learned endpoints converge weakly to P_{data} and $\widehat{P}$, respectively. Joint lower semicontinuity of relative entropy, Fatou's lemma over the training samples, (C.6), and (C.9) prove (C.3) before projection. The projected bound follows by data processing under Π_w.

• **Step 5: choose the times and combine the error bounds.** The initialization bound decays exponentially in $T-t_0$, while (C.4) bounds the early-stopping error by a constant times $t_0^{\beta\wedge1}$. We choose $T-t_0$ proportional to $\log n$ and t_0 proportional to n^{-2}, and verify below that both errors decay faster than $n^{-\gamma_L}$. Specifically, for $n\ge2$, set

$$t_0=\frac{n^{-2}}{8\sqrt{d}(1+C+C^{-1})},\qquad T=t_0+\log(2en). \tag{C.11}$$

Use the common architecture in (B.2) at these times, with its minimum admissible depth and width, and require $L^2\le n$. For the fixed cutoff w, use the accuracy and norm budgets in (B.3) and (B.4), together with the output bound and empirical minimizer specified in Theorem B.1. Evaluate K_β at the horizon in (C.11).

⋆ *Bound the fractional power of the score estimation error.* Take $q=4\log(en)$ and $H=64\sqrt{d}\,\mathsf{K}_\beta^2\sqrt{\log(ed/\epsilon_n)}$. Since $|C-1|/\min\{1,C\}\le1+C+C^{-1}$, we have that

$$qt_0\sqrt{d}\,\frac{|C-1|}{\min\{1,C\}}\le\frac{\log(en)}{2n^2}\le\frac{1}{2}.$$

For every $x\ge0$, it holds that

$$x^{1-1/(2\log(en))}\le\sqrt{e}\,(x+n^{-1}).$$

For $x\ge n^{-1}$ this follows by dividing by x; for $x<n^{-1}$, use monotonicity. Substituting this scalar bound in Lemma C.1, we obtain

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}_n}\left[\mathrm{KL}((\Pi_w)_\sharp P_{\mathrm{data}}\|(\Pi_w)_\sharp\widehat{P})\right]\\&\quad\le\mathcal{E}_{\mathrm{init}}(T-t_0)+2\mathcal{E}_{\mathrm{stop}}\\&\qquad+\frac{2\sqrt{e}\,H_\beta}{c}\left[2\{H^2+\sqrt{d(d+2)}\}J\right]^{1/(2\log(en))}\left\{\mathbb{E}_{\mathcal{D}_n}\left[\mathcal{R}(\widehat{s})\right]+n^{-1}\right\}.\end{aligned} \tag{C.12}$$

The factors in this display have the explicit bounds

$$\begin{aligned}\mathsf{K}_\beta&\le6+\beta+\beta^{-1}+C+C^{-1}+c^{-1}+2H_\beta+\log(en),\\H^2&\le4096d\mathsf{K}_\beta^4\log(e^3dn),\\J&\le\log\{16e^2\sqrt{d}(1+C+C^{-1})n^3\}.\end{aligned}$$

The last inequality uses $e^{t_0}-1\ge t_0$ and $T\le1+\log(2en)$. Consequently, for all sufficiently large n with the problem parameters fixed, we have that

$$2\{H^2+\sqrt{d(d+2)}\}J\le(en)^{2\log2-1}, \tag{C.13}$$

and the multiplier in (C.12) is at most $4H_\beta/c$.

⋆ *Bound the endpoint errors and check the depth.* Integrating $\|x\|^2p_{\mathrm{data}}(x)$ using the density upper bound, we obtain $\mathbb{E}\left[\|X_0\|^2\right]\le H_\beta d/C$. For $b=(2en)^{-1}$, expanding $-\log(1-b)$ in a series shows that $-b-\log(1-b)\le b^2/[2(1-b)]\le b^2$. Therefore, we obtain

$$\mathcal{E}_{\mathrm{init}}(T-t_0)\le\frac{H_\beta d}{4Cen}+\frac{d}{8e^2n^2}.$$

The stopping term is bounded by (C.4), with the t_0 in (C.11). The depth prescribed in (A.28) is admissible for the balanced accuracy because the following bounds hold:

$$1+\sigma_{t_0}^{-2}\le1+16\sqrt{d}(1+C+C^{-1})n^2,\qquad\epsilon_n\ge e^{-2}/n.$$

Indeed, $1 - e^{-t_0} \geq t_0/2$, and $0 < \gamma_L < 1$ in (B.3). These inequalities and the displayed bound for K_β show directly from (A.28) that its minimum integer depth satisfies $L^2 \leq n$ for sufficiently large n. Thus the same architecture and budgets satisfy all conditions of Theorem B.1.

⋆ *Apply the full-score estimation bound.* Under (C.13), we substitute (B.6) into (C.12) to obtain

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_n}\left[\mathrm{KL}((\Pi_w)_\sharp P_{\mathrm{data}} \| (\Pi_w)_\sharp \widehat{P})\right] \\
&\quad \leq \frac{H_\beta d}{4Cen} + \frac{d}{8e^2 n^2} + 2\mathcal{E}_{\mathrm{stop}} + \frac{4H_\beta}{c}\{\mathsf{Est}_n(w) + n^{-1}\} \\
&\qquad + \frac{32H_\beta}{c}\{I_w(t_0) - I_w(T) + K_w(t_0) - K_w(T)\}.
\end{aligned}$$

Here (B.5) specifies $\mathsf{Est}_n(w)$ with numerical constants. The information sum is at most $I_w(0) + K_w(0)$. By (B.27), the information remainder satisfies

$$\frac{32H_\beta}{c}\{I_w(0) + K_w(0)\} \leq \frac{32H_\beta}{c} I_w(0) + \frac{2^{(\beta\wedge 1)+3} H_\beta^3}{c^{2+(\beta\wedge 1)}} \left(\mathbb{E}\left[\|(I_d - \Pi_w)X_0\|_2^2\right]\right)^{\beta\wedge 1}.$$

⋆ *Conclude the stated rate.* Since $d_{\mathrm{eff}}(w) \geq 1$, we have that

$$0 < \gamma_L < \frac{2\beta}{2\beta + d_{\mathrm{eff}}(w)} < \min\{1, 2(\beta \wedge 1)\}.$$

The explicit endpoint bounds consequently decay faster than $n^{-\gamma_L}$, with the class parameters fixed. Combining the explicit estimation bound in (B.5), the bounds above for J and K_β, and the prescribed minimum depth, we conclude that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_n}\left[\mathrm{KL}((\Pi_w)_\sharp P_{\mathrm{data}} \| (\Pi_w)_\sharp \widehat{P})\right] \\
&\quad \leq \widetilde{\mathcal{O}}_{\beta,C,c,H_\beta,d_{\mathrm{eff}}(w)}\left(L^2\left\{(d^4/n)^{\gamma_L} + d^5/n\right\}\right) \\
&\qquad + \frac{32H_\beta}{c}\mathcal{I}(P_{\mathrm{data}}^{(w)}, P_{\mathrm{data}}^{(>w)}) + \frac{2^{(\beta\wedge 1)+3} H_\beta^3}{c^{2+(\beta\wedge 1)}} \left(\mathbb{E}\left[\|(I_d - \Pi_w)X_0\|_2^2\right]\right)^{\beta\wedge 1}.
\end{aligned} \tag{C.14}$$

Only powers of $\log(edn)$ are suppressed. As $d^{4\gamma_L} n^{-\gamma_L} + d^5/n \leq 2d^5 n^{-\gamma_L}$, this proves the statistical term in Proposition 5.4. The remainder constants are explicit in (C.14). By Theorem B.1, the bound is uniform in the available number of branches when the depth, width, norm bounds, and output bound are fixed. At full cutoff both exact residuals vanish. Moreover, it holds that

$$n^{-\gamma_L} = n^{-2\beta/(2\beta + d_{\mathrm{eff}}(w))} \exp\left\{\frac{4\beta \log n}{(2\beta + d_{\mathrm{eff}}(w))(L+1)}\right\}.$$

The exponential factor is bounded at the prescribed minimum depth, so the full-cutoff rate is $\widetilde{\mathcal{O}}(n^{-2\beta/(2\beta+d)})$.

C.1 A LOWER BOUND FOR THE HÖLDER DENSITY CLASS

Let $\mathcal{P}_\beta$ be the class in Assumption 4.1, with $C > 0$, $\beta > 0$, $0 < c < 1$, and $H_\beta > 1$ fixed. We prove the lower bound in Section 5.2 on a subclass whose discarded frequencies are independent reference Gaussians. The information remainder vanishes on this subclass.

Theorem C.2. Let $r = \lceil \beta \rceil + 1$ and define

$$a_{w,\beta} := \frac{\min\{1/4, 1-c, (H_\beta - 1)/3\}}{3(2\pi e)^{d_{\mathrm{eff}}(w)/2} 2^{r d_{\mathrm{eff}}(w)} (r-1)! (18r\sqrt{d_{\mathrm{eff}}(w)})^{r-1} \max\{1, C^{\beta/2}\}}. \tag{C.15}$$

For every $n \geq 1$, it holds that

$$\inf_{\widehat{P}} \sup_{P \in \mathcal{P}_\beta} \mathbb{E}\left[\mathrm{KL}((\Pi_w)_\sharp P \| (\Pi_w)_\sharp \widehat{P})\right] \geq \frac{a_{w,\beta}^2}{2^{2\beta+5}} \left(\frac{(r!)^2}{(2r+1)!}\right)^{2d_{\mathrm{eff}}(w)} n^{-2\beta/(2\beta+d_{\mathrm{eff}}(w))}.$$

The infimum includes all randomized probability-valued estimators based on n full d-dimensional observations.

If $c = 1$ or $H_\beta = 1$, normalization forces $f \equiv 1$, so the class is a singleton for fixed C and has zero minimax risk.

Proof. The proof has three steps.

- Step 1 constructs a family of retained densities and verifies the Hölder bound.
- Step 2 bounds the distance between neighboring sample laws.
- Step 3 reduces distribution estimation to testing the perturbation signs.

Use $k = \lceil \beta \rceil - 1$ and $\tau = \beta - k$ from (E.1), so $r = k + 2$.

• **Step 1: construct a family in the Hölder class.** Let $\eta(u) = [u(1-u)]^r$ on $[0,1]$ and zero elsewhere, and set

$$\psi(z) = \left\{ \sum_{j=0}^{3} (-1)^j \binom{3}{j} \eta(8z_1 - 2j) \right\} \prod_{i=2}^{d_{\mathrm{eff}}(w)} \eta(z_i).$$

The first-coordinate copies have disjoint supports in $[0,1]$. Since their coefficients sum to zero, we have that

$$\int \psi = 0, \qquad \int |\psi| = \left(\frac{(r!)^2}{(2r+1)!} \right)^{d_{\mathrm{eff}}(w)}. \tag{C.16}$$

The second identity uses $\sum_{j=0}^{3} \binom{3}{j}/8 = 1$ and the beta integral $\int_0^1 [u(1-u)]^r du = (r!)^2/(2r+1)!$.

For an integer $N \geq 1$, let $h = N^{-1}$ and $\nu \in \{0, \dots, N-1\}^{d_{\mathrm{eff}}(w)}$. For signs $\theta_\nu \in \{-1, 1\}$, define

$$q_{\theta,h}(z) = \phi_{d_{\mathrm{eff}}(w),1}(z) + a_{w,\beta} h^\beta \sum_{\nu} \theta_\nu \psi((z - h\nu)/h).$$

Embed this density into $\mathbb{R}^d$ by setting

$$p_{\theta,h}(x) = \left(\frac{C}{2\pi} \right)^{d/2} e^{-C\|x\|_2^2/2} \frac{q_{\theta,h}(\sqrt{C}\, U_w^\top x)}{\phi_{d_{\mathrm{eff}}(w),1}(\sqrt{C}\, U_w^\top x)}.$$

Let $P_{\theta,h}$ denote the distribution with density $p_{\theta,h}$. For $X \sim P_{\theta,h}$, the high-frequency coordinates $V_w^\top X$ have distribution $\mathsf{N}(0, C^{-1} I_{d - d_{\mathrm{eff}}(w)})$ and are independent of the low-frequency coordinates $U_w^\top X$. In particular, $I_w(t) = K_w(t) = 0$ for every $t \geq 0$ on this family. Normalization follows from the zero integral in (C.16).

To check the Hölder bound, write $v = q_{\theta,h}/\phi_{d_{\mathrm{eff}}(w),1} - 1$. The derivatives of η through order $k+1$ vanish at its endpoints. Consequently, the grid perturbations and their derivatives join continuously, and at most one cell contributes at each point. By expanding η and differentiating the Gaussian reciprocal, we obtain the following bounds for $0 \leq j \leq k+1$:

$$\|\partial^j \psi\|_{\infty,\mathrm{op}} \leq 3\, 2^{r d_{\mathrm{eff}}(w)} (16 r \sqrt{d_{\mathrm{eff}}(w)})^j,$$
$$\sup_{z \in [0,1]^{d_{\mathrm{eff}}(w)}} \|\partial^j (\phi^{-1}_{d_{\mathrm{eff}}(w),1})(z)\|_{\mathrm{op}} \leq (2\pi e)^{d_{\mathrm{eff}}(w)/2} j! (1 + \sqrt{d_{\mathrm{eff}}(w)})^j.$$

For the first inequality, $\|\eta^{(j)}\|_\infty \leq 2^r (2r)^j$, the first-coordinate scaling contributes 8^j, and expansion in unit directions contributes $d_{\mathrm{eff}}(w)^{j/2}$. For the second, the derivative is a sum over partitions of j directions into singletons and pairs; there are at most $j!$ such partitions. Leibniz's rule bounds the derivatives of the perturbation as follows:

$$\|\partial^j v\|_{\infty,\mathrm{op}} \leq a_{w,\beta} Q_{w,\beta} h^{\beta - j}, \qquad Q_{w,\beta} := 3(2\pi e)^{d_{\mathrm{eff}}(w)/2} 2^{r d_{\mathrm{eff}}(w)} (r-1)! (18 r \sqrt{d_{\mathrm{eff}}(w)})^{r-1}.$$

Here $16 r \sqrt{d_{\mathrm{eff}}(w)} + 1 + \sqrt{d_{\mathrm{eff}}(w)} \leq 18 r \sqrt{d_{\mathrm{eff}}(w)}$, and $h \leq 1$. The derivative bounds of orders k and $k+1$ imply

$$\begin{aligned} \|\partial^k v(z) - \partial^k v(z')\|_{\mathrm{op}} &\leq a_{w,\beta} Q_{w,\beta} \min\{2h^\tau, h^{\tau-1} \|z - z'\|_2\} \\ &\leq 2 a_{w,\beta} Q_{w,\beta} \|z - z'\|_2^\tau. \end{aligned}$$

Composition with $\sqrt{C}\, U_w^\top$ multiplies each derivative bound and the Hölder seminorm by at most $\max\{1, C^{\beta/2}\}$. With the choice (C.15), the ambient density ratio satisfies

$$f \geq c, \qquad \frac{3}{4} \leq f \leq \frac{5}{4}, \qquad \|f\|_{\mathcal{H}^\beta} \leq H_\beta.$$

Hence all $P_{\theta,h}$ belong to one fixed class $\mathcal{P}_\beta$, independently of N and the signs.

• **Step 2: bound the distance between neighboring sample laws.** Let P_+ and P_- be two members of this family that differ in one sign, and let q_- be the density of $\sqrt{C}\,U_w^\top X$ for $X \sim P_-$. The common Gaussian density of the high-frequency coordinates cancels in the chi-squared divergence. Since $q_- \geq (3/4)\phi_{d_{\mathrm{eff}}(w),1}$, we have that

$$\begin{aligned}\chi^2(P_+\|P_-) &\leq \frac{16}{3} a_{w,\beta}^2 (2\pi e)^{d_{\mathrm{eff}}(w)/2} h^{2\beta+d_{\mathrm{eff}}(w)} \int \psi^2 \\ &\leq \frac{1}{3} h^{2\beta+d_{\mathrm{eff}}(w)}.\end{aligned}$$

The last inequality uses $|\psi| \leq 3$ on a set of volume at most one and $a_{w,\beta} \leq [12(2\pi e)^{d_{\mathrm{eff}}(w)/2}]^{-1}$. Choose $N = \lceil n^{1/(2\beta+d_{\mathrm{eff}}(w))} \rceil$. Applying the product formula for chi-squared divergence and then the inequality $\mathrm{TV}(P,Q) \leq \frac{1}{2}\sqrt{\chi^2(P\|Q)}$, we find

$$\mathrm{TV}(P_+^{\otimes n}, P_-^{\otimes n}) \leq \tfrac{1}{2}\sqrt{e^{nh^{2\beta+d_{\mathrm{eff}}(w)}/3} - 1} < \tfrac{1}{2}.$$

An independent random seed for the estimator does not change this bound.

• **Step 3: reduce estimation to testing the signs.** For an arbitrary probability-valued estimator, let $\widehat{Q} = (\sqrt{C}\,U_w^\top)_\sharp \widehat{P}$. The coordinate map is a measurable bijection on the retained subspace, so it holds that

$$\mathrm{KL}((\Pi_w)_\sharp P_{\theta,h} \| (\Pi_w)_\sharp \widehat{P}) = \mathrm{KL}(q_{\theta,h}(z)\,dz \| \widehat{Q}). \tag{C.17}$$

Put $b_\nu(z) = \operatorname{sgn}\psi((z - h\nu)/h)$, with $\operatorname{sgn}(0) = 0$. Decode $\widehat{\theta}_\nu$ as the sign of $\int b_\nu\, d(\widehat{Q} - \phi_{d_{\mathrm{eff}}(w),1}(z)\,dz)$, resolving a tie as 1. By (C.16), the corresponding integral for the true law is

$$a_{w,\beta} h^{\beta+d_{\mathrm{eff}}(w)} \left(\frac{(r!)^2}{(2r+1)!}\right)^{d_{\mathrm{eff}}(w)} \theta_\nu.$$

A wrong sign requires a discrepancy of at least this magnitude. Since the b_ν have disjoint nonzero supports and $|b_\nu| \leq 1$, the dual formula for total variation implies the following lower bound:

$$\mathrm{TV}(q_{\theta,h}(z)\,dz, \widehat{Q}) \geq \frac{a_{w,\beta}}{2} h^{\beta+d_{\mathrm{eff}}(w)} \left(\frac{(r!)^2}{(2r+1)!}\right)^{d_{\mathrm{eff}}(w)} \sum_\nu \mathbb{1}\{\widehat{\theta}_\nu \neq \theta_\nu\}.$$

Write $\mathbb{E}_\theta$ for expectation over the sample from $P_{\theta,h}^{\otimes n}$ and any independent randomness in the estimator. Average uniformly over all sign vectors. For each sign, the sum of the two testing errors is at least $1 - \mathrm{TV}(P_+^{\otimes n}, P_-^{\otimes n}) \geq 1/2$; its average error probability is therefore at least $1/4$. There are $N^{d_{\mathrm{eff}}(w)} = h^{-d_{\mathrm{eff}}(w)}$ signs, so we obtain

$$\frac{1}{2^{N^{d_{\mathrm{eff}}(w)}}} \sum_\theta \mathbb{E}_\theta\big[\mathrm{TV}(q_{\theta,h}(z)\,dz, \widehat{Q})\big] \geq \frac{a_{w,\beta}}{8} \left(\frac{(r!)^2}{(2r+1)!}\right)^{d_{\mathrm{eff}}(w)} h^\beta.$$

By Pinsker's inequality $\mathrm{KL} \geq 2\,\mathrm{TV}^2$ and Jensen's inequality, the average KL risk satisfies

$$\frac{1}{2^{N^{d_{\mathrm{eff}}(w)}}} \sum_\theta \mathbb{E}_\theta\big[\mathrm{KL}(q_{\theta,h}(z)\,dz \| \widehat{Q})\big] \geq \frac{a_{w,\beta}^2}{32} \left(\frac{(r!)^2}{(2r+1)!}\right)^{2d_{\mathrm{eff}}(w)} h^{2\beta}.$$

Finally, $N \leq 2n^{1/(2\beta+d_{\mathrm{eff}}(w))}$ and the supremum risk is at least the average over this family. Together with (C.17), these facts prove the claimed coefficient and rate. □

D Numerical details

D.1 Construction and properties of the synthetic distribution

We give the construction of the fixed synthetic distribution used in the simulation experiments and verify Assumptions 4.1 and 5.2. The ambient dimension is $d = 15$, and the distribution is held fixed as the training sample size, regularization coefficient, and evaluation cutoff vary. We use a fixed finite Hölder radius in the regularity condition; all distributional constants below are independent of these experimental choices.

Construction in a real Fourier basis. Let $U \in \mathbb{R}^{15\times 15}$ be the real orthogonal Fourier basis corresponding to the complex DFT matrix F in Section 3.2, with columns ordered as the DC component followed by the cosine and sine components at frequencies $1, \dots, 7$. For $\ell = 0, \dots, 14$ and $k = 1, \dots, 7$, these columns are

$$(u_0)_\ell = \frac{1}{\sqrt{15}}, \qquad (u_{k,c})_\ell = \sqrt{\frac{2}{15}} \cos\left(\frac{2\pi k\ell}{15}\right), \qquad (u_{k,s})_\ell = \sqrt{\frac{2}{15}} \sin\left(\frac{2\pi k\ell}{15}\right). \tag{D.1}$$

Thus $U^\top U = I_{15}$. Let $Z = (Z_1, \dots, Z_{15})^\top$ have independent coordinates with

$$Z_j \sim \frac{1}{2}\mathcal{N}(0,1) + \frac{1}{2}\mathcal{N}(0, \tau^2_{k(j)}), \qquad \tau_k^2 = (1+k)^{-(2s+1)}, \qquad s = \frac{1}{4}, \tag{D.2}$$

where $(k(1), \dots, k(15)) = (0, 1, 1, \dots, 7, 7)$, and define $X_0 = UZ$. The target distribution P_{data} is the law of $X_0 \in \mathbb{R}^{15}$; $Z = U^\top X_0$ denotes its Fourier coordinates. The mixture indicators are sampled independently for every coordinate and observation. In particular, this is a product of univariate mixtures, rather than a mixture with one indicator shared by all coordinates.

For a cutoff $w \in \{0, \dots, 7\}$, let U_w contain the first $m_w = 1 + 2w$ columns of U. Then $\Pi_w = U_w U_w^\top = F^* \Lambda_w F$. retains the DC component and both real coordinates at each frequency up to w. A cutoff is used to evaluate an already fitted model; it does not change the data-generating distribution.

Gaussian-relative density and Assumption 4.1. Write $z = U^\top x$, and let φ_{15} be the standard Gaussian density on $\mathbb{R}^{15}$. Since U is orthogonal,

$$\begin{aligned} p_{\text{data}}(x) &= \varphi_{15}(x) f(x), \\ f(x) &= \prod_{j=1}^{15} \left[\frac{1}{2} + \frac{1}{2\tau_{k(j)}} \exp\left\{-\frac{\tau_{k(j)}^{-2} - 1}{2} z_j^2\right\}\right]. \end{aligned} \tag{D.3}$$

The DC factor is identically one because $\tau_0 = 1$. Every other factor lies between $1/2$ and $(1 + \tau_k^{-1})/2$, giving

$$2^{-14} \le f(x) \le \prod_{k=1}^{7} \left(\frac{1 + \tau_k^{-1}}{2}\right)^2 \approx 28435.10531. \tag{D.4}$$

Each positive-order derivative of a nonconstant factor is a polynomial times a Gaussian function and is globally bounded. Finite products and the orthogonal change of variables preserve this property. Consequently, $f \in C^\infty(\mathbb{R}^{15})$, and every derivative of fixed order is bounded. For any fixed $\beta > 0$, the next bounded derivative controls the Hölder seminorm at distances at most one, while boundedness controls larger distances. Hence $\|f\|_{H^\beta} \le H_\beta$ for a finite constant H_β. This verifies the Gaussian-relative density, positive lower bound, and Hölder regularity in Assumption 4.1, with $C = 1$, $c = 2^{-14}$, and a fixed radius H_β. The spectral parameter $s = 1/4$ is distinct from the Hölder smoothness index β.

Non-Gaussianity, moments, and tails. The distribution is centered and has a strictly positive density on $\mathbb{R}^{15}$. For a coordinate at frequency k, let

$$v_k = \mathbb{E} Z_j^2 = \frac{1 + \tau_k^2}{2}. \tag{D.5}$$

For $k \ge 1$, its fourth cumulant is $\mathbb{E}Z_j^4 - 3(\mathbb{E}Z_j^2)^2 = \frac{3}{4}(1 - \tau_k^2)^2 > 0$, establishing non-Gaussianity. Moreover,

$$\mathbb{E} e^{aZ_j} = \frac{1}{2} e^{a^2/2} + \frac{1}{2} e^{\tau^2_{k(j)} a^2/2} \le e^{a^2/2}. \tag{D.6}$$

Independence and orthogonality imply $\mathbb{E} e^{u^\top X_0} \le e^{\|u\|_2^2/2}$ for every $u \in \mathbb{R}^{15}$. Thus X_0 is sub-Gaussian with covariance proxy I_{15}, and its actual covariance is $U \operatorname{diag}(v_{k(1)}, \dots, v_{k(15)}) U^\top$.

Spectral energy and Assumption 5.2(A). In the unitary complex DFT, the coefficients at frequencies k and $15 - k$ each have expected squared magnitude v_k: each is formed from the cosine

and sine coordinates with normalization $1/\sqrt{2}$. Therefore,

$$E_{\text{tail}}(w) := \mathbb{E}\|(I_{15} - \Pi_w)X_0\|_2^2 = 2\sum_{k=w+1}^{7} v_k$$
$$= (7-w) + \sum_{k=w+1}^{7} (1+k)^{-3/2}. \tag{D.7}$$

Evaluating this finite sum at the eight admissible cutoffs gives

$$\max_{0\le w\le 7}(1+w)^{1/2}E_{\text{tail}}(w) = 9.3195993603\ldots, \tag{D.8}$$

where the maximum is over integers and is attained at $w = 2$. Hence

$$E_{\text{tail}}(w) \le 9.32(1+w)^{-1/2}, \qquad w = 0,\ldots,7, \tag{D.9}$$

which verifies Assumption 5.2(A) with $s = 1/4$ and $S_{\text{pol}} = 9.32$. This is an upper envelope for the fixed 15-dimensional distribution. The actual coordinate energy is $v_k = \frac{1}{2} + \frac{1}{2}(1+k)^{-3/2}$, including the broad Gaussian component's contribution $1/2$; it is not an exact power law. No dimension-uniform spectral constant is asserted.

Analytic scores under the forward process. For the OU process in the main text, write $\alpha_t = e^{-t/2}$ and $\sigma_t^2 = 1 - e^{-t}$. Orthogonal Gaussian corruptions preserve independence of the Fourier coordinates. If φ_a denotes the centered univariate Gaussian density with variance a, the noisy coordinate at frequency k has density

$$q_{k,t}(z) = \frac{1}{2}\varphi_1(z) + \frac{1}{2}\varphi_{a_{k,t}}(z), \qquad a_{k,t} = \sigma_t^2 + \alpha_t^2\tau_k^2. \tag{D.10}$$

Its score is available analytically:

$$s_{k,t}(z) := \partial_z \log q_{k,t}(z) = -z\left[1 + r_{k,t}(z)(a_{k,t}^{-1} - 1)\right], \qquad r_{k,t}(z) = \frac{\varphi_{a_{k,t}}(z)}{\varphi_1(z) + \varphi_{a_{k,t}}(z)}. \tag{D.11}$$

The full score in data coordinates is obtained by multiplying this vector of coordinate scores by U. At each cutoff, the retained marginal score equals the corresponding block of the full score. It also equals the retained block of the score obtained by starting the OU process from $\Pi_w X_0$. This identity connects the analytic evaluation below to R_w in the main text.

Dependence and Gaussian-completion terms. For every $t \ge 0$, the retained and discarded Fourier blocks are independent, so $I_w(t) = 0$ and the dependence remainder $\mathcal{E}_{\text{dep}}$ is zero. The discarded marginal need not equal the reference Gaussian. With $C = 1$, define

$$K_w(t) = 2\sum_{k=w+1}^{7} \text{KL}(q_{k,t}\|\varphi_1). \tag{D.12}$$

OU entropy dissipation gives the exact Gaussian-completion term

$$\mathcal{E}_{\text{cutoff}}(w) = 2\{K_w(t_0) - K_w(T)\}$$
$$= \int_{t_0}^{T} \sum_{j=m_w+1}^{15} \mathbb{E}\left[\{s_{k(j),t}(Z_{t,j}) + Z_{t,j}\}^2\right] dt, \tag{D.13}$$

where $Z_t = U^\top X_t$. It is positive for $w < 7$ and vanishes at $w = 7$. Thus E_{tail} is the discarded second moment entering the spectral bound, whereas $\mathcal{E}_{\text{cutoff}}$ measures the discarded score's mismatch with its Gaussian completion.

D.2 Implementation details of the simulation experiments

The simulations use the distribution in Appendix D.1. Unless stated otherwise, the architecture and diffusion interval are fixed across all conditions: $M = W = 64$, $[t_0, T] = [10^{-4}, 8]$, and output cap $S = 128$. The learning curves use $n \in \{128, 256, 512, 1024, 2048, 4096\}$ at $\lambda = 0.01$. The regularization sweep compares $\lambda \in \{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$; the combined figure shows its $n = 2048$ slice. The reference coefficient 0.01 was fixed before test evaluation.

Training, validation, and test data. We use five independent data seeds, $17, 29, 43, 59, 71$, and two initialization restarts per (n, λ) and data seed. Each seed has an i.i.d. training pool of 8192 observations; the training set D_n is its first n observations. Training sets are therefore nested across sample sizes within a seed. Separate random streams generate 4096 validation observations and $N_{\text{test}} = 16384$ test observations per seed. These clean data sets are independent of one another. A seed's clean validation and test sets are shared across experimental conditions, and common corruption draws are used within each evaluation stream to make comparisons less noisy. Training, objective-audit, validation, and test corruption streams are distinct. Checkpoint selections are frozen before test data are accessed for the corresponding training family.

Score-network architecture and initialization. The network receives the full data vector and time, $[x^\top, t]^\top \in \mathbb{R}^{16}$. Each of its 64 branches has four affine layers of width 64, with ReLU activations after the first three layers. Their outputs are concatenated into a 4096-dimensional vector and passed through a shared 64-unit ReLU layer and a 30-unit ReLU layer. Subtracting the last 15 coordinates of the latter from the first 15 gives a raw score $h_\theta(x, t) \in \mathbb{R}^{15}$. A fixed two-affine-layer ReLU map applies coordinatewise clipping:

$$[g_\theta(x,t)]_j = \operatorname{clip}\left([h_\theta(x,t)]_j, -\frac{128}{\sqrt{15}}, \frac{128}{\sqrt{15}}\right). \tag{D.14}$$

Hence $\|g_\theta(x,t)\|_2 \le 128$. The shared output network has four affine layers when the two fixed clipping layers are included. There are $1{,}132{,}510$ trainable scalar parameters.

Initialization represents the clipped Gaussian score $-x$. One branch provides signed identity features; the other branch parameters are randomly initialized. The initial squared parameter norms are 4216 in the branches and 60 in the trainable shared layers. All 64 shared channels remain trainable. In particular, the 34 channels beyond the signed identity features have positive initial activations, while their outgoing head weights start at zero to preserve the initial score. Those outgoing weights can receive gradients immediately. The initialization uses neither the Fourier basis nor the analytic mixture score. The same architecture is trained on full observations for every cutoff; no cutoff-specific input projection is supplied to the network.

DSM objective and uniform quadratic penalty. For $X_t = \alpha_t X^{(i)} + \sigma_t \varepsilon$, with $\varepsilon \sim \mathsf{N}(0, I_{15})$, training minimizes

$$\begin{aligned} \widehat{J}_\lambda(\theta) &= \widehat{L}(g_\theta; D_n) + \lambda\|\theta\|_2^2, \\ \widehat{L}(g_\theta; D_n) &= \frac{1}{n}\sum_{i=1}^{n}\int_{t_0}^{T} \mathbb{E}_\varepsilon \left\| g_\theta(\alpha_t X^{(i)} + \sigma_t\varepsilon, t) + \frac{\varepsilon}{\sigma_t}\right\|_2^2 dt. \end{aligned} \tag{D.15}$$

Here θ contains every trainable weight and bias in both the branches and shared layers. The fixed clipping coefficients are excluded. The penalty has no factor $1/2$ and is not divided by sample size, parameter count, or an initialization norm. We implement it explicitly in the loss; Adam's built-in weight-decay parameter is zero. Thus the penalty contributes $2\lambda\theta$ to the gradient before gradient clipping and the Adam update.

To estimate the unweighted time integral efficiently, define

$$\mathcal{I} = \int_{t_0}^{T} \frac{dt}{\sigma_t^2} = \log\left(\frac{e^T - 1}{e^{t_0} - 1}\right) \approx 17.20995, \qquad q(t) = \frac{1}{\mathcal{I}\sigma_t^2}. \tag{D.16}$$

Each update samples $B = 512$ training indices with replacement, independent Gaussian corruptions, and stratified time draws from q. Its loss estimate is

$$\frac{\mathcal{I}}{B}\sum_{b=1}^{B} \|\sigma_{t_b} g_\theta(X_{t_b}, t_b) + \varepsilon_b\|_2^2 + \lambda\|\theta\|_2^2. \tag{D.17}$$

The importance distribution and factor $\mathcal{I}$ make this an unbiased estimate of Eq. (D.15). Times are obtained by stratifying $u \in (0,1)$ into 512 equal intervals, randomly permuting the draws, and applying $t = \log[1 + \exp\{\log(e^{t_0} - 1) + \mathcal{I}u\}]$.

Optimization and training budget. We use Adam with $(\beta_1, \beta_2) = (0.9, 0.999)$, numerical stabilizer 10^{-8}, and global gradient-norm clipping at 10. Training uses FP32, without mixed precision, TF32, or an exponential moving average of the parameters. Independent trajectories are executed concurrently on four NVIDIA A100 GPUs with 80 GB of memory each, using PyTorch 2.7.1 with CUDA 12.6. Each trajectory retains its own batch of 512 observations and optimizer state. The default learning-rate schedule is

Update range	Learning rate
1–2000	linear warmup from 3×10^{-5} to 3×10^{-4}
2001–24000	3×10^{-4}
24001–48000	10^{-4}
48001–96000	3×10^{-5}
96001–144000	10^{-5}
144001–192000	3×10^{-6}

Every trajectory receives at least 96000 updates. Audit-based extensions may increase the horizon to 144000 and then 192000 updates.

For positive λ, the last 12000 updates of each reached horizon form a refinement stage. At its start, we restore the best independently audited checkpoint and rebalance each branch's final affine layer against its corresponding columns in the shared aggregation matrix. If their squared norms are a_m and b_m, respectively, the rescaling $(A_{m,4}, b_{m,4}) \mapsto (A_{m,4}, b_{m,4})/c_m$ and $C_m \mapsto c_m C_m$, with $c_m = (a_m/b_m)^{1/4}$ for positive a_m, b_m, preserves the network function and minimizes the penalty along that scale direction. We reset Adam and use learning rate 10^{-5} throughout the refinement stage. These updates are included in the stated horizons and override the default schedule above. Zero-penalty trajectories omit this rebalancing stage.

Checkpoint audits and continuation decisions. The candidate pool includes initialization, checkpoints every 4000 updates, checkpoints immediately before and after rebalancing, and horizon endpoints. Each candidate's original penalized training objective is audited on the entire fixed D_n, using a corruption bank independent of training and monitoring. Audits use 64 Gauss–Legendre nodes in log time and initially four independent antithetic Gaussian pairs per observation and time node. The analytically known empirical DSM loss of the Gaussian score $-x$ is used as a control variate: we estimate the DSM difference with the same corruptions and add back the exact Gaussian-reference loss. This reduces Monte Carlo variance without changing the audited objective.

At each horizon, continuation is triggered by a material audited improvement, a best checkpoint within the last 8000 updates, or a material penalty decrease from that horizon's rebalancing step. For objective improvement Δ, we require

$$\Delta > \max\{3\ \mathrm{SE}_{\mathrm{paired}},\ 0.01 \max\{1, |J_{\mathrm{init}} - J_{\mathrm{best}}|\}\} \tag{D.18}$$

to call the improvement material. Here J_{init} and J_{best} are audited penalized objective values, and the paired standard error is computed conditionally on the fixed training set from independent antithetic replicates. The first comparison uses candidates available by update 72000 as its reference; later comparisons use the preceding horizon. When the paired uncertainty interval crosses the material-improvement threshold, the number of audit pairs is doubled, up to 16. A material discrepancy between restarts can also extend the less successful restart, subject to the same 192000-update cap. Selected objective estimates receive an additional 128-node quadrature check.

For each data seed and (n, λ), the primary checkpoint minimizes the independently audited penalized training objective over saved candidates from both restarts. Within a restart, exact ties favor the earlier checkpoint; ties across restarts favor the lower restart index. A separate secondary selection minimizes full validation score risk, evaluated with 32 log-time quadrature nodes and one antithetic pair on the independent validation set. The combined figure uses the primary selection throughout. The two restarts are candidates for selection, not additional replicates in the plotted average. Previously completed $\lambda = 0$ controls are reused only where the data, initialization, optimizer, audit rules, and frozen training code agree with the present zero-penalty protocol.

The trajectories supplying the 50 distinct primary fits in the combined figure all reached 192000 updates. Of these fits, 45 retain the conservative late-checkpoint diagnostic; four of those also retain a material-improvement diagnostic. A late-checkpoint flag records where the best saved candidate

occurs and does not by itself quantify the remaining optimization error. All fits are retained in the reported results.

Independent-test retained-score risk. For a frozen fitted network, write $s_{w,t}$ for the score of the retained noisy marginal. We evaluate

$$R_w(g_\theta) = \int_{t_0}^{T} \mathbb{E}_{X_t \sim P_t} \left\| U_w^\top g_\theta(X_t, t) - s_{w,t}(U_w^\top X_t) \right\|_2^2 dt$$
$$= \int_{t_0}^{T} \mathbb{E} \sum_{j=1}^{m_w} \left\{ [U^\top g_\theta(UZ_t, t)]_j - s_{k(j),t}(Z_{t,j}) \right\}^2 dt, \tag{D.19}$$

using Eq. (D.11). The expectation is over the full noisy data law, and the network still receives all 15 coordinates. This risk contains neither the parameter penalty nor an added cutoff remainder. It is not divided by m_w or by $T - t_0$.

Let $(t_q, \omega_q)_{q=1}^{Q}$ be the log-time Gauss–Legendre rule with weights transformed to integrate dt. For each clean test Fourier vector $Z^{(i)}$, node q, and independent pair r, draw $\varepsilon_{irq} \sim \mathsf{N}(0, I_{15})$ and set $Z_{irq}^{\pm} = \alpha_{t_q} Z^{(i)} \pm \sigma_{t_q} \varepsilon_{irq}$. With P antithetic pairs, the estimator is

$$\widehat{R}_w = \sum_{q=1}^{Q} \frac{\omega_q}{2PN_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \sum_{r=1}^{P} \sum_{a \in \{-,+\}} \sum_{j=1}^{m_w} \left\{ [U^\top g_\theta(UZ_{irq}^a, t_q)]_j - s_{k(j),t_q}(Z_{irq,j}^a) \right\}^2. \tag{D.20}$$

The default is $Q = 64$ and $P = 1$, with evaluation batch size 16384. The fixed evaluation rule increases P to 2 and then 4 if the estimated full-risk Monte Carlo standard error exceeds 2% of its estimate. The current combined figure uses $Q = 64$, $P = 1$ for every fitted model; additional 128-node checks are recorded for the displayed $n = 128, 2048, 4096$ conditions. Model evaluations use FP32, while analytic scores, quadrature, and risk accumulation use FP64. An antithetic pair is treated as one replicate when assessing uncertainty.

Aggregation and empirical frequency frontier. Each primary checkpoint and its test corruptions are reused across all eight cutoffs. The plotted value is the arithmetic mean of the five data-seed risk estimates. The combined figure omits error bars. Its sample-size curves use $w = 1, 3, 5, 7$, corresponding to $m_w = 3, 7, 11, 15$, while its full-range and enlarged cutoff panels display the same test measurements on different vertical ranges.

The horizontal reference $R_\star = 1$ was introduced after inspecting the results as a descriptive accuracy budget. On the mean-risk curves, the largest qualifying cutoffs are respectively *none*, *none*, $1, 3, 6, 7$ for $n = 128, 256, 512, 1024, 2048, 4096$. It does not enter training or checkpoint selection. Because Eq. (D.19) is a cumulative sum of nonnegative coordinate errors, increasing risk with w alone does not establish an impossibility threshold for each added frequency. These fixed-architecture, fixed-time-window experiments describe finite-sample behavior under the specified training procedure; they do not test the asymptotic exponent of Theorem 5.1.

D.3 CIFAR-10 TRAINING AND SAMPLING PROTOCOL

Data and preprocessing. We reserve 5,000 images from the official 50,000-image training split for validation and construct nested training subsets of sizes $n \in \{5{,}000, 10{,}000, 20{,}000, 40{,}000\}$ from the remaining 45,000 images. Each training subset contains exactly $n/10$ images from each class. Class labels are used only to construct the subsets; the models are unconditional. The subsets and validation indices remain fixed throughout training. Minibatch indices are sampled with replacement from the corresponding finite training subset. Images are mapped from 8-bit RGB to $[-1, 1]$, with independent random horizontal flips of probability $1/2$ during training. We do not use uniform dequantization. Validation images are evaluated without augmentation. The official 10,000-image test split supplies the common real-image reference for pFID; it is not used to select checkpoints.

Architecture and diffusion process. We use the continuous-VP CIFAR-10 DDPM++ configuration from the Score-SDE implementation (Song et al., 2020b), pinned to commit `cb1f359f4aadf0ff9a5e122fe8fffc9451fd6e44`. The architecture has 128 base channels, channel multipliers $(1, 2, 2, 2)$, four residual blocks per resolution, attention at 16×16 resolution, GroupNorm, Swish activations, and dropout probability 0.1. It contains 61,804,419 trainable parameters. All models are initialized from scratch and trained on complete noisy images, with diffusion time supplied as an additional input.

Writing $Y_0 \in [-1, 1]^{3\times32\times32}$ for a training image, the forward process is

$$\begin{aligned} Y_t &= \alpha_t Y_0 + \sigma_t \epsilon, \\ \alpha_t &= \exp\left[-\tfrac{1}{2}(0.1t + 9.95t^2)\right], \qquad \sigma_t^2 = 1 - \alpha_t^2, \end{aligned} \tag{D.21}$$

where $\epsilon \sim N(0, I)$ and training times are uniform on $[10^{-5}, 1]$. This is the VP schedule $\beta(t) = 0.1 + 19.9t$ used by the reference implementation.

Objective and optimization. Let $D = 3 \cdot 32 \cdot 32 = 3072$, and let $\epsilon_\theta(Y_t, t)$ denote the network's noise prediction. The native training objective averages squared error over images and all D image components:

$$\widehat{\mathcal{L}}_{\text{native}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{a,t,\epsilon}\left[\frac{1}{D}\left\|\epsilon_\theta(\alpha_t a(Y_i) + \sigma_t\epsilon, t) - \epsilon\right\|_2^2\right], \tag{D.22}$$

where a is the random horizontal-flip augmentation. Fresh times and Gaussian corruptions are sampled at every update, while the clean training dataset remains fixed. We optimize $\widehat{\mathcal{L}}_{\text{native}}(\theta) + \lambda\|\theta\|_2^2$, penalizing all trainable parameters, including biases. There is no factor $1/2$ or normalization by the number of parameters in the penalty. Optimizer weight decay is set to zero, so this penalty is applied exactly once. The sample-size comparison uses $\lambda = 0$; the displayed penalty comparison at $n = 20{,}000$ uses $\lambda \in \{0, 10^{-7}, 10^{-6}, 10^{-5}\}$.

We use Adam with learning rate 2×10^{-4}, coefficients $(0.9, 0.999)$, numerical constant 10^{-8}, a 5,000-update linear warmup, global batch size 128, and gradient-norm clipping at 1. Exponential moving averages (EMA) use the reference implementation's decay parameter 0.9999 and its initial update-count correction. The global batch size is unchanged when training is distributed across two GPUs. Each displayed condition uses one training repetition (run seed 711; initialization seed 10711). The transferred $n = 20{,}000$, $\lambda = 10^{-6}$ checkpoint uses the same architecture, training subset, loss normalization, and validation protocol; its stored validation DSM was independently reproduced.

Training budget, early stopping, and checkpoint selection. We use a common 200,000-update window for checkpoint selection. Within this window, we retain the EMA checkpoint with the smallest unpenalized validation DSM. Checkpoints beyond update 200,000 are excluded from the reported comparison. Validation is performed every 5,000 updates and at saved milestones. Each of the 5,000 validation images receives four fixed Gaussian corruptions, with one time sampled in each of four equal-width strata of $[10^{-5}, 1]$. The same draws, generated with seed 79001, are reused across checkpoints and conditions. Dropout is disabled during validation. The loss is averaged over corruptions within each image and then over images; it excludes the L_2 penalty.

The penalty runs additionally enable validation-based stopping after 50,000 updates: training stops after six consecutive regular checks more than 1% above the best validation loss, or after 100,000 updates without a cumulative relative improvement of at least 0.2%. None of the three displayed positive-penalty runs triggers this rule before 200,000 updates. The 10k unregularized trajectory was manually stopped after validation deterioration, at update 183,598; its last validation was at 180,000. Table 1 gives the selected checkpoints and their full-image FIDs. All three positive-penalty models select update 200,000; the comparison therefore concerns performance within the specified budget.

Sampling. Each selected EMA model generates 10,000 images with the same continuous-VP reverse sampler. We initialize with $N(0, I)$ at $t = 1$ and use 1,000 Euler–Maruyama predictor steps down to $t = 10^{-3}$, without a corrector or probability-flow substitution. We enable the reference sampler's final denoising step. Sampling uses batch size 128 and float32 arithmetic with TF32 disabled. Images are indexed as two blocks of 5,000, each processed in batches of at most 128. The random seed for each batch is 910001 plus its first absolute image index; this rule is shared across

Table 1: Validation-selected EMA checkpoints used in the CIFAR-10 comparison. Selection is restricted to updates at most 200,000. Validation DSM excludes the parameter penalty. FID uses 10,000 generated images and 10,000 official test images.

Training images n	λ	Selected update	Validation DSM	FID
5,000	0	20,000	0.03023881	15.10
10,000	0	40,000	0.02966227	10.62
20,000	0	90,000	0.02915971	8.03
40,000	0	200,000	0.02870772	6.93
20,000	10^{-7}	200,000	0.02916011	10.26
20,000	10^{-6}	200,000	0.03001353	20.48
20,000	10^{-5}	200,000	0.03128808	41.86

models. Parallel generation preserves these batch boundaries and seeds. The output is mapped back to $[0, 1]$, clipped, and converted to 8-bit RGB by truncation. These saved images are reused at every cutoff.

D.4 Fourier-projected evaluation and spectral energy

Fourier projection. All projections are computed at the original 32×32 resolution with an orthonormal two-dimensional discrete Fourier transform, separately in each RGB channel. For integer frequencies $k_x, k_y \in \{-16, \dots, 15\}$, define

$$\Pi_w x = \mathcal{F}^{-1}[\mathbf{1}\{\max(|k_x|, |k_y|) \le w\}\, \mathcal{F}x]\,. \tag{D.23}$$

The mask is square, includes both members of each conjugate pair, and is identical across channels. We evaluate pFID at every integer $w \in \{1, \dots, 16\}$, without spatial downsampling. Fourier projection is used only for evaluation; each model generates complete images before any projection is applied.

Clipping and quantization. Real images and saved generated images are first represented on the normalized 8-bit grid in $[0, 1]$. After projection, both image sets pass through the same elementwise operation

$$\mathcal{Q}(z) = \frac{1}{255}\,\lfloor 255\ \mathrm{clip}(z, 0, 1)\rfloor\,. \tag{D.24}$$

The floor corresponds to the implementation's truncating conversion from clipped nonnegative values to `uint8`. Thus the preprocessing implicit in Eq. (6.1) is

$$\mathrm{pFID}_w = \mathrm{FID}\big(\{\mathcal{Q}(\Pi_w x_i)\}_{i=1}^N, \{\mathcal{Q}(\Pi_w \widehat{x}_i)\}_{i=1}^N\big)\,, \qquad N = 10{,}000. \tag{D.25}$$

In particular, generated images are quantized once when saved and again after projection. At $w = 16$, the implementation returns the original 8-bit images directly, avoiding artificial one-level changes from an FFT round trip followed by truncation. Consequently, pFID_{16} equals ordinary FID for the same images and evaluation pipeline.

Inception features and FID. We use the frozen TensorFlow Hub `tensorflow/tfgan/eval/inception/1` model and its 2,048-dimensional `pool_3` activations, following the Score-SDE evaluation implementation. The extractor receives 8-bit RGB images, converts them to float32 in $[-1, 1]$, and performs resizing inside the reference SavedModel; there is no additional external resize. Feature batches contain 32 images. FID (Heusel et al., 2017) is computed with TensorFlow-GAN 2.0.0 from empirical feature means and sample covariances, using the same extractor and preprocessing for real and generated images at every cutoff. The reference is the same 10,000 official test images for all models, rather than the published 50,000-image reference statistics sometimes used for the full-data CIFAR-10 benchmark. Our full-cutoff values therefore describe this 10k-versus-10k protocol.

Empirical recovered range. Let $\mathcal{W} = \{1, \dots, 16\}$. For the illustrative tolerance $\tau = 6.5$, the circles in Figure 2(d) identify

$$\widehat{w}_\tau = \max\left\{ w \in \mathcal{W} : \max_{\substack{v \in \mathcal{W} \\ v \le w}} \mathrm{pFID}_v \le \tau \right\}, \tag{D.26}$$

when the set is nonempty. This requires every tested cutoff up to the marked cutoff to meet the tolerance and uses no interpolation. The threshold was chosen after the initial evaluation for visualization; it is not a confidence-calibrated recovery boundary. The initially coarser cutoff grid was subsequently completed to all integers using the same checkpoints, generated images, and metric implementation. For the reported $\lambda = 0$ models, the resulting cutoffs are $6, 7, 8, 11$ at $n = 5{,}000, 10{,}000, 20{,}000, 40{,}000$, respectively. At $n = 40{,}000$, $\mathrm{pFID}_{11} = 6.4304$ and $\mathrm{pFID}_{12} = 6.5545$, placing the last qualifying cutoff at $w = 11$.

Spectral tail energy. For Figure 2(a), we use all 60,000 original CIFAR-10 images, including the official training and test splits, in canonical $[0, 1]$ RGB coordinates. We compute

$$\widehat{E}_{\mathrm{tail}}(w) = \frac{1}{60{,}000} \sum_{i=1}^{60{,}000} \left\|(I - \Pi_w)x_i\right\|_2^2, \qquad w = 0, \ldots, 16. \tag{D.27}$$

The squared norm sums over all 3,072 image components, without division by image dimension or per-image standardization. This calculation uses float64 Fourier energies and Parseval's identity; the projected images are not clipped or quantized for this diagnostic. The mean tail energies at $w = 4, 8, 12$ are respectively 23.3428, 7.0440, and 1.7879, and the tail at $w = 16$ is exactly zero. This is the empirical counterpart of the spectral-tail quantity in Assumption 5.2. Its finite-resolution decay does not establish a population decay envelope or the paper's other distributional assumptions. Using all 60,000 images here is solely a descriptive data diagnostic and does not enlarge any model's training set.

Checks and interpretation. Numerical checks verify Parseval's identity, projection idempotence, nesting, full-cutoff identity, and agreement between direct residual energy and Fourier tail sums, with errors below 10^{-8}. Ordinary FID is also computed independently to check the full-cutoff identity. All displayed conditions use one training repetition, so the curves provide point estimates rather than across-training-seed error bars. FID has finite-sample variability, and the displayed threshold is descriptive. Finally, the native objective in Eq. (D.22) corresponds to σ_t^2-weighted score DSM through $s_\theta = -\epsilon_\theta/\sigma_t$; it is not the unweighted OU risk analyzed in the theory. Likewise, pFID measures discrepancy in Inception feature moments after projection and image preprocessing, not the projected KL divergence or exact score risk. These experiments study the empirical frequency-recovery perspective on real images without identifying these distinct quantities.

E Helper lemmas and background

This section contains the auxiliary results used in the proofs:

- Hölder regularity of f_w and bounds on $\sigma_t \nabla f_{w,t}$ (Appendix E.1);
- ReLU approximation of elementary operations (Appendix E.2);
- covering numbers for the parallel networks and shared output layers (Appendix E.3);
- entropy bounds for the two remainders (Appendix E.4);
- norm bounds implied by weight decay (Appendix E.5).

E.1 Hölder regularity and Gaussian projections

We use the following Hölder convention and prove the two bounds needed in the approximation argument:

- the Gaussian integral defining f_w preserves the Hölder norm bound on f;
- $\sigma_t \nabla f_{w,t}$ is uniformly bounded by H_β.

For a j times continuously differentiable map $F : \mathbb{R}^r \to \mathbb{R}^q$, write $\partial^j F(x)$ for its jth total derivative, with $\partial^0 F(x) = F(x)$ and

$$\partial^j F(x)[u_1, \ldots, u_j] := \frac{\partial^j}{\partial s_1 \cdots \partial s_j} F\left(x + \sum_{i=1}^{j} s_i u_i\right)\Bigg|_{s_1 = \cdots = s_j = 0}.$$

Its operator norm is the supremum of the output norm over unit directions; at $j = 0$ it is $\|F(x)\|_2$.

Definition E.1 (Hölder norm)**.** For $\beta > 0$, let $k = \lceil \beta \rceil - 1$ and $\tau = \beta - k \in (0, 1]$. For a k times continuously differentiable map F, define

$$\|F\|_{\mathcal{H}^\beta} := \max_{0 \le j \le k} \sup_x \|\partial^j F(x)\|_{\mathrm{op}} + \sup_{x \ne y} \frac{\|\partial^k F(x) - \partial^k F(y)\|_{\mathrm{op}}}{\|x - y\|_2^\tau}. \tag{E.1}$$

At integer β, this is the $C^{\beta-1,1}$ convention.

• **Step 1: bound the Hölder norm of f_w.** Since the derivatives of f are bounded and U_w is an isometry, we may differentiate the integral defining f_w with respect to u to obtain

$$\partial^j f_w(u)[u_1, \dots, u_j] = \mathbb{E}\Big[\partial^j f(U_w u + C^{-1/2} V_w G_{d - d_{\mathrm{eff}}(w)})[U_w u_1, \dots, U_w u_j]\Big].$$

The derivatives therefore satisfy

$$\sup_u \|\partial^j f_w(u)\|_{\mathrm{op}} \le \sup_x \|\partial^j f(x)\|_{\mathrm{op}}, \qquad 0 \le j \le k. \tag{E.2}$$

For the Hölder seminorm, use the same $G_{d-d_{\mathrm{eff}}(w)}$ in the difference at u and u'. It follows that

$$\sup_{u \ne u'} \frac{\|\partial^k f_w(u) - \partial^k f_w(u')\|_{\mathrm{op}}}{\|u - u'\|_2^\tau} \le \sup_{x \ne x'} \frac{\|\partial^k f(x) - \partial^k f(x')\|_{\mathrm{op}}}{\|x - x'\|_2^\tau}. \tag{E.3}$$

Adding these bounds proves $\|f_w\|_{\mathcal{H}^\beta} \le H_\beta$. The pointwise bounds $c \le f \le H_\beta$ are preserved by the integral defining f_w, so $c \le f_w \le H_\beta$.

• **Step 2: bound the normalized score numerator.** By (A.5), for each unit vector v, the normalized directional derivative satisfies

$$|v^\top \sigma_t \nabla f_{w,t}(u)| \le H_\beta \mathbb{E}\big[|v^\top G_{d_{\mathrm{eff}}(w)}|\big] = \sqrt{2/\pi}\, H_\beta.$$

Here $\alpha_t / \sqrt{\alpha_t^2 + C\sigma_t^2} \le 1$, and $v^\top G_{d_{\mathrm{eff}}(w)}$ is standard Gaussian. Taking the supremum over v proves

$$\|\sigma_t \nabla f_{w,t}(u)\|_2 \le H_\beta. \tag{E.4}$$

E.2 ReLU approximation of elementary operations

The following lemmas implement multiplication, Gaussian functions, and the scalar diffusion coefficients by ReLU networks. The bounds account for registers carried through the layers, biases, and the constant coordinates used in rescaling.

Set $c_0 = 2^{14}$. The estimates below hold with this constant for all dimensions, smoothness indices, approximation accuracies, and time intervals specified below. All logarithms are natural unless a base is shown. A signed scalar register consists of its positive and negative parts. Thus an identity operation on an arbitrary real register uses two ReLU neurons, whose weights must be included in the norm bound.

Lemma E.2 (Multiplication with bounded coefficients)**.** For $R \ge 1$ and $0 < \eta \le 1$, a ReLU network denoted by Mult satisfies

$$\sup_{|u|,|v| \le R} |\mathrm{Mult}(u, v) - uv| \le \eta, \qquad \mathrm{Mult}(u, 0) = \mathrm{Mult}(0, v) = 0.$$

Its width is at most eight, its depth is at most $8[1 + \log(R^2/\eta)]$, and every weight and bias has absolute value at most four. Copying q additional signed registers through the construction uses width at most $8 + 2q$, without changing its depth bound.

Proof. We use the square approximation and adapt the multiplication construction of Yarotsky (2017, Propositions 2–3) to input bound $M = R$. We check the explicit width, depth, and coefficient bounds below. Let S_n be the dyadic square interpolant in the proof of Proposition 2, so that $\|S_n - s^2\|_{L^\infty([0,1])} \le 2^{-2n-2}$. For $n = \lceil \frac{1}{2} \log_2(2R^2/\eta) \rceil$, the polarization formula is

$$\mathrm{Mult}(u, v) = R^2 \left\{ S_n\left(\frac{|u + v|}{2R}\right) - S_n\left(\frac{|u - v|}{2R}\right) \right\}.$$

Its error is at most $2R^2 2^{-2n-2} \le \eta$, and its two arguments coincide whenever either input is zero.

In each square circuit, three triangular-function neurons and one partial-sum neuron are followed by two nonnegative register updates. The two parallel copies therefore have width eight and use $2n$ layers, with coefficients bounded by four. Write $R^2 = 2^a b$, where $a = \lceil \log_2 R^2 \rceil$ and $1/2 < b \le 1$. Implement the final gain by multiplying the signed difference by b and doubling it a times. Initialization, absolute values, the signed difference, and the final output add at most four layers. Thus the depth satisfies

$$2n + a + 4 \le 8 + \frac{2 \log R^2 + \log(1/\eta)}{\log 2} \le 8[1 + \log(R^2/\eta)].$$

Copying each additional signed register uses two neurons with unit weights and does not add layers. □

Lemma E.3 (ReLU approximation of elementary computations)**.** Suppose a function is evaluated on a given input domain in $S \ge 1$ successive steps. At each step, at most $V \ge 1$ scalar values are computed or retained for later use. Each new value is an affine combination, a product of two values, their minimum or maximum, or a clipped value, formed from values available at the preceding step. Operations within a step are performed in parallel; initial and final affine maps count toward S. Assume all affine coefficients, biases, and fixed clipping endpoints have absolute value at most $A \ge 2$, and every exact input to a product lies in $[-A, A]$ throughout the input domain. Other values may be unbounded.

For $0 < \eta \le 1$, there is a ReLU network with uniform error at most η in each output coordinate, width at most $10(V + 1)$, and depth at most

$$16(S + 1)^2 \log\left(\frac{8(V + 1)A}{\eta}\right). \tag{E.5}$$

Every weight and bias has absolute value at most four. If scalar operations are evaluated one at a time, the width is at most $2V + 10$.

Proof. We replace each product by the network in Lemma E.2 and implement the other operations exactly. The proof has two steps:

- Choose the multiplication accuracy and bound the accumulated error.
- Count the layers and neurons used by these operations.

• **Step 1: control the accumulated error.** Use Lemma E.2 on $[-2A, 2A]^2$ with accuracy

$$\eta_* = \frac{\eta}{(S + 1)[16(V + 1)A]^{S+1}}. \tag{E.6}$$

On this square, multiplication is $4A$-Lipschitz in the maximum norm. Thus the error of the approximated product is at most $4A$ times the largest operand error plus η_*. After the first step, an affine combination uses at most V values and amplifies their maximum error by at most VA. Minimum, maximum, and clipping amplify this error by at most two. These bounds also apply when the values entering affine operations or clips are unbounded.

Starting from exact inputs, induction shows that the largest absolute error among the values after step $s \le S$ is at most

$$\eta_* \sum_{j=0}^{s-1} [16(V + 1)A]^j < \eta.$$

The exact inputs to each product lie in $[-A, A]$. Since their errors are less than $\eta \le 1$, their approximations lie in $[-2A, 2A]$. This verifies the multiplication domain at each induction step and proves the output error bound.

• **Step 2: bound the network depth and width.** An affine map with coefficients and biases bounded by A is implemented by dividing its coefficients and biases by a power of two at least A, representing each output by its positive and negative parts, and restoring the scale by repeated doubling. This takes at most $\lceil \log_2 A \rceil + 2$ layers, with coefficients bounded by four. Minimum and maximum are implemented exactly using $\max(a, b) = b + \mathrm{ReLU}(a - b)$ and $\min(a, b) = a - \mathrm{ReLU}(a - b)$; clipping is a composition of these two operations.

Write $\ell = \log(8(V+1)A/\eta) \geq \log 32$. By Lemma E.2, the parallel multiplication networks at one step require at most $8[1 + \log(4A^2/\eta_*)]$ layers. The affine operations, clips, and copies add at most 5ℓ layers. To substitute (E.6) into this count, we use

$$1 + \log(4A^2/\eta) \leq 2\ell, \qquad \log(S+1) \leq S\ell/5, \qquad \log(16(V+1)A) \leq 6\ell/5.$$

The total number of layers is therefore at most

$$S\big\{8[1 + \log(4A^2/\eta_*)] + 5\ell\big\} \leq 16S(S+2)\ell \leq 16(S+1)^2\ell.$$

Each new scalar output requires at most eight neurons, whether it is computed by multiplication, an affine operation, or clipping. Copying a scalar value uses two neurons for its positive and negative parts. Thus computing at most V new values while retaining at most V others uses at most $10V$ neurons. More precisely, p parallel products together with V copied values use at most $8p + 2V$ neurons. For sequential scalar operations, one operation and the copied values use at most $2V + 8$ neurons. Allowing two more neurons for a constant value proves the stated width bounds. Each operation uses only its specified arguments, and retained values are copied without adding connections to other values. □

Lemma E.4 (Gaussian density and distribution function)**.** Write $\phi(z) = (2\pi)^{-1/2}e^{-z^2/2}$ and $\Phi(z) = \int_{-\infty}^{z} \phi(u)\,du$. For $0 < \eta \leq 1/4$, there are scalar ReLU networks $\bar{\phi}, \bar{\Phi}$ with ranges in $[0,1]$, uniform error at most η on the whole real line, bounded coefficients as in Lemma E.3, width at most 34, and depth at most $c_0[\log(8/\eta)]^3$.

Proof. The proof has two steps:

- Control the Gaussian tails and polynomial approximation errors.
- Implement the polynomials using the multiplication network.

• **Step 1: bound the polynomial approximation errors.** Put $\ell = \log(8/\eta)$ and $b = \sqrt{2\ell}$. Clipping the input to $[-b, b]$ changes either Gaussian function by at most $\eta/8$. On this interval, set $x = z/b$ and approximate $\phi(z)$ and $\Phi(z)$ by the respective Taylor polynomials

$$\frac{1}{\sqrt{2\pi}} \sum_{j=0}^{q} \frac{(-b^2/2)^j x^{2j}}{j!}, \qquad \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \sum_{j=0}^{q} \frac{(-1)^j b^{2j+1} x^{2j+1}}{2^j j!(2j+1)},$$

where $q = \lceil 4eb^2 + 2\log(8b/\eta) \rceil \leq 25\ell$. Taylor's remainder bound $u^{q+1}/(q+1)!$ for e^{-u}, together with $(q+1)! \geq ((q+1)/e)^{q+1}$, bounds both polynomial errors by $\eta/4$. The error bound for the integrated polynomial includes the factor b. The sums of the absolute coefficients, and hence the partial sums in Horner's rule on $[-1,1]$, are at most $16\eta^{-2}$.

• **Step 2: implement the polynomials.** Apply Lemma E.3 to each Horner evaluation with error $\eta/4$. Each evaluation requires no more than $S = 2q + 9$ stages, including input and output clips, and twelve live registers. Thus the width is at most 34. Take $A = 32\eta^{-2}$ in (E.6). Since $S + 1 \leq 53\ell$ and $\log(16 \cdot 13A) \leq 3.35\ell$, the local tolerance satisfies $\log(1/\eta_*) \leq 180\ell^2$. The sharper multiplier count $2n + a + 4$ in Lemma E.2 is therefore at most $272\ell^2$. There are at most $q + 2 \leq 26\ell$ products; the intervening affine gains cost at most 4ℓ layers each, and the clips add at most 16. The total depth is consequently bounded by

$$26\ell(272\ell^2 + 4\ell) + 16 < 8192\ell^3 \leq c_0\ell^3.$$

Clipping both outputs to $[0,1]$ cannot increase the error. Together with the tail and polynomial errors, this proves the lemma. □

Lemma E.5 (Known scalar diffusion coefficients)**.** Let $K \geq 2 + T + C + C^{-1}$ and $0 < \rho \leq 1/4$. There are simultaneous scalar ReLU approximations to the functions

$$\alpha_t, \quad \sigma_t, \quad \sigma_t^{-1}, \quad \sigma_t^{-2}, \quad \frac{\alpha_t}{\alpha_t^2 + C\sigma_t^2}, \quad \frac{\sigma_t}{\sqrt{\alpha_t^2 + C\sigma_t^2}}, \quad \frac{\sqrt{\alpha_t^2 + C\sigma_t^2}}{\sigma_t}, \quad \alpha_t/\sqrt{\alpha_t^2 + C\sigma_t^2}$$

and $C/(\alpha_t^2 + C\sigma_t^2)$, each with absolute error at most ρ on $[t_0, T]$. These networks have width at most 58, coefficients bounded by four, and depth at most

$$c_0^8 K^4 \left[\log\left(\frac{eK(1+\sigma_{t_0}^{-2})}{\rho}\right)\right]^3. \tag{E.7}$$

The approximation of σ_t^{-1} is nonnegative and at most $(1+\rho)/\sigma_t$. The approximation of σ_t lies in $[0, 1]$.

Proof. The proof has three steps:

- Approximate the exponential coefficients by polynomials.
- Compute inverse square roots and bound the errors in the remaining coefficients.
- Bound the depth and width of the resulting networks.

• **Step 1: approximate the exponentials.** Put $A_0 = 16K^2(1+\sigma_{t_0}^{-2})$ and $\zeta = \rho/(100A_0)^{100}$. Taylor polynomials of degree $\lceil 4eT + 2\log(8/\zeta)\rceil$ approximate e^{-t} and $e^{-t/2}$ on $[0, T]$ to error ζ, by the remainder estimate in Lemma E.4. Clip them to $[0, 1-\sigma_{t_0}^2]$ and $[0, 1]$, respectively. Use these approximations in $\sigma_t^2 = 1 - e^{-t}$ and $\alpha_t^2 + C\sigma_t^2 = 1 + (C-1)\sigma_t^2$, and clip the approximation of the latter coefficient to $[K^{-1}, K]$.

• **Step 2: approximate the remaining coefficients.** For the inverse square roots, set $a = \sigma_{t_0}^2/(4K^2)$ and, for $x \in [a, 1]$, use the Newton iteration

$$z_0 = 1, \qquad z_{j+1} = \tfrac{1}{2} z_j(3 - x z_j^2).$$

The quantities $y_j = x z_j^2$ remain in $(0, 1]$ and increase by a factor at least $25/16$ while $y_j \le 1/2$. Thereafter their residuals satisfy $1 - y_{j+1} = (1-y_j)^2(4-y_j)/4 \le (1-y_j)^2$. It follows that $|z_n - x^{-1/2}| \le \zeta$ for

$$n = \left\lceil \frac{\log(1/a)}{\log(25/16)} \right\rceil + \left\lceil \log_2 \log_2 \left(\frac{2}{\zeta\sqrt{a}}\right) \right\rceil + 2. \tag{E.8}$$

Apply this iteration to the approximations of σ_t^2 and $(\alpha_t^2 + C\sigma_t^2)/K$; both lie in $[a, 1]$. Multiplying the second output by $K^{-1/2}$ approximates $(\alpha_t^2 + C\sigma_t^2)^{-1/2}$. The remaining outputs require fewer than forty affine or product operations on arguments bounded by A_0. Since $|(x^{-1/2})'| \le a^{-3/2}/2 \le A_0^2$ on $[a, 1]$, their combined error is at most $2A_0^2(400A_0)^{40}\zeta < \rho/2$.

• **Step 3: bound the depth and width.** To bound the implementation cost, write $H = \log(eK(1+\sigma_{t_0}^{-2})/\rho)$. The inequalities $K \ge 4$, $H \ge 4$, $\log(100A_0) \le 3H$, and $\log(1/a) \le 2H$ bound the exponential degree by $160KH$ and the Newton iteration count in (E.8) by $9H$. Evaluate the two polynomials in parallel by Horner's rule and the two inverse roots in parallel, using five stages per iteration. Including the output operations and clips, the number of stages satisfies

$$S + 1 \le 2(160KH) + 45H + 46 \le 340KH.$$

Every affine coefficient and exact intermediate value has absolute value at most $\exp(2KH)$: the Horner partial sums are bounded by e^K, the Newton intermediates by $4/a$, and the final arithmetic values by A_0. Twenty live registers and two parallel multiplier blocks use at most 56 neurons during the paired recursions; at most 24 registers and width 58 suffice for the final scalar operations.

Compile the circuit to additional error $\rho/2$ using (E.6). The logarithm in (E.5) is at most $3KH$, so the depth is bounded by $16(340KH)^2(3KH) < 2^{23}K^3H^3 \le c_0^8 K^4 H^3$. Finally, clip the approximations of σ_t and σ_t^{-1} to $[0, 1]$ and $[0, 2\sigma_{t_0}^{-1}]$, respectively. These clips do not increase the error, and $\sigma_t^{-1} \ge 1$ implies the stated upper bound $(1+\rho)/\sigma_t$. □

Lemma E.6 (Gaussian interval moments and a cell network)**.** Each cell contribution (d_z, u_z) in (A.16) admits one multidimensional ReLU branch. Use the cells and midpoint Taylor polynomials

of Step 4, with N cells, input radius R_ϵ, and clean radius R_{clean}. Use the accuracy δ from (A.11). The following quantities specify the internal approximation accuracies:

$$\begin{aligned} K &= 2 + d_{\text{eff}}(w) + k + T + C + C^{-1} + c^{-1} + 2H_\beta, \qquad Q = \binom{d_{\text{eff}}(w) + k}{k}, \\ A &= 2^{12} K^4 (Q+1)(N+1)(1 + R_{\text{clean}} + K R_\epsilon)(1 + \sigma_{t_0}^{-2}), \\ \nu &= \frac{\delta}{2^{14}(N+1)(d_{\text{eff}}(w)+1)^2 (Q+1) K}, \\ \rho &= \frac{\nu}{(8A)^{4k+16}}, \qquad \eta_\Phi = \frac{\nu}{(8A)^{4k+12}}. \end{aligned} \tag{E.9}$$

For each cell, the branch outputs $(\bar{d}_z, \bar{u}_z)$ satisfy, simultaneously for all $t \in [t_0, T]$ and $\|U_w^\top x\|_\infty \le R_\epsilon$,

$$|\bar{d}_z - d_z| \le \frac{\delta}{2N}, \qquad \|\bar{u}_z - u_z\|_2 \le \frac{\delta}{2N}. \tag{E.10}$$

Its width is at most $4d_{\text{eff}}(w) + 106$, including the direct signed input lift $(U_w^\top x, -U_w^\top x, t, -t)$, which uses $2d_{\text{eff}}(w) + 2$ neurons and has a sum of squared weight norms and squared bias norms of $2(d_{\text{eff}}(w) + 1)$. In particular, this width is at most $64(d + d_{\text{eff}}(w) + k + 3)$. All coefficients are bounded by four, and its depth is at most

$$c_0^{16}(Q+1)^2 K^{20} \Lambda^3, \qquad \Lambda = \log\left(\frac{ed(N+1)(1+\sigma_{t_0}^{-2})(K + R_\epsilon + R_{\text{clean}})}{\delta} \right). \tag{E.11}$$

Proof. The proof has the following steps.

- Apply the Gaussian interval-moment recurrence to each cell.
- Bound the errors when the Gaussian density and distribution function are approximated.
- Combine the interval moments to obtain the numerator and denominator for one cell.
- Count the intermediate values that must be stored at the same time.
- Bound the total depth after replacing each arithmetic operation by a ReLU network.

• **Step 1: apply the interval-moment recurrence.** Fix a cell and a coordinate, and suppress their indices in T_r and J_r. Let b_-, b_+ be the interval endpoints and z its midpoint. Apply the univariate recurrence in Kan & Robotti (2017, Section 2, p. 931) to the variable $u - z$, with mean $\alpha_t (U_w^\top x)_i / (\alpha_t^2 + C\sigma_t^2) - z$, variance $\sigma_t^2 / (\alpha_t^2 + C\sigma_t^2)$, and integration limits $b_- - z, b_+ - z$. The standardized endpoints are

$$A_\pm = \frac{\sqrt{\alpha_t^2 + C\sigma_t^2}}{\sigma_t} \left(b_\pm - \frac{\alpha_t (U_w^\top x)_i}{\alpha_t^2 + C\sigma_t^2} \right).$$

The recurrence starts from $T_0 = \Phi(A_+) - \Phi(A_-)$. Substituting these parameters and using the definition of J_r in (A.15), we obtain

$$\begin{aligned} J_r &= (b_- - z)^r \phi(A_-) - (b_+ - z)^r \phi(A_+) + \frac{r\sigma_t}{\sqrt{\alpha_t^2 + C\sigma_t^2}} T_{r-1}, \\ T_r &= \left(\frac{\alpha_t (U_w^\top x)_i}{\alpha_t^2 + C\sigma_t^2} - z \right) T_{r-1} + \frac{\sigma_t}{\sqrt{\alpha_t^2 + C\sigma_t^2}} J_{r-1}, \qquad r \ge 1. \end{aligned} \tag{E.12}$$

For $r = 0$, omit the last term in the first line to compute J_0. For $r \ge 2$, substituting the formula for J_{r-1} into the second line computes T_r from T_{r-1} and T_{r-2}. Only two consecutive T moments need to be stored; each J_r can then be computed from the first line. Since $|u - z| \le h'/2 \le 1$ on each cell, the moments satisfy

$$|T_r| \le 1, \qquad |J_r| \le \int_{\mathbb{R}} |y| \phi(y)\, dy \le 1.$$

Clipping a computed moment to $[-1, 1]$ therefore cannot increase its approximation error.

• **Step 2: approximate the scalar functions.** Take the polynomial–Newton construction from Lemma E.5 with accuracy ρ, including the approximation of $\sqrt{\alpha_t^2 + C\sigma_t^2}/\sigma_t$. At this point retain its affine and product operations as a finite arithmetic circuit; the entire cell circuit will be implemented by a ReLU network after the approximation errors have been bounded. Clip the approximation of $\sigma_t/\sqrt{\alpha_t^2 + C\sigma_t^2}$ from below at $\sigma_{t_0}/(2\sqrt{K})$. The exact coefficient is at least $\sigma_{t_0}/\sqrt{K}$, so this clipping cannot increase the error. Form the two standardized endpoints using the approximation of $\sqrt{\alpha_t^2 + C\sigma_t^2}/\sigma_t$. Their errors are at most $A^2\rho$. Both ϕ and Φ are one-Lipschitz, so this also bounds the errors in their exact values at the perturbed endpoints.

Next use the explicit polynomial constructions of Lemma E.4, with total Gaussian-function error at most η_Φ, to evaluate those four endpoint values. Each such routine clips its argument to magnitude at most $\sqrt{2\log(8/\eta_\Phi)}$ before polynomial evaluation; the lemma's Gaussian tail estimate includes the error from this clipping. Only the arguments of the Gaussian functions are clipped in this way: the coefficients $\alpha_t(U_w^\top x)_i/(\alpha_t^2 + C\sigma_t^2) - z$ and $\sigma_t/\sqrt{\alpha_t^2 + C\sigma_t^2}$, and the powers of $b_\pm - z$, remain unchanged in the moment recurrence.

Insert these approximate values into the recurrence and clip each computed T_r, J_r to $[-1, 1]$. The exact moments belong to this interval, so each clip cannot increase the error. Each of the four Gaussian values has error at most $\eta_\Phi + A^2\rho$, and the starting moments T_0, J_0 have at most twice this error. In the recurrence for T_r in terms of T_{r-1} and T_{r-2}, the sum of the absolute coefficients multiplying the preceding moment errors is at most $8A^3$; errors in these coefficients and the boundary values add at most $8A^3(\eta_\Phi + A^2\rho)$. These estimates use $r \le k \le A$, $|b_\pm - z| \le 1$, and the bound A on the absolute values of $\alpha_t(U_w^\top x)_i/(\alpha_t^2 + C\sigma_t^2) - z$ and $\sigma_t/\sqrt{\alpha_t^2 + C\sigma_t^2}$. Iterate this bound through at most k updates and then apply the formula for J_r, in which the coefficient multiplying the moment error is at most $2A^2$. Since $A \ge 2^{12}$, each T_r and J_r is approximated with error at most

$$(4A)^{3k+4}(\eta_\Phi + A^2\rho) \le 2\nu.$$

• **Step 3: assemble the cell outputs.** The coefficients $b_{z,r} = \partial^r f_w(z)/r! - c\,\mathbb{1}_{\{r=0\}}$ are fixed numbers of magnitude at most K. For each multi-index, multiply the $d_{\text{eff}}(w)$ interval moments in (A.16), replacing the ith T factor by J for numerator coordinate i. Sum the Q weighted terms and multiply the numerator by $\alpha_t/\sqrt{\alpha_t^2 + C\sigma_t^2}$. These operations define the approximate cell outputs before their ReLU implementation.

Products of numbers in $[-1, 1]$ amplify their maximum error by at most the number of factors. The preceding moment errors therefore contribute at most $8(d_{\text{eff}}(w)+1)QK\nu$ to each output coordinate; the approximation of $\alpha_t/\sqrt{\alpha_t^2 + C\sigma_t^2}$ contributes at most another $QK\rho$. The Euclidean error is at most $\sqrt{d_{\text{eff}}(w)}$ times the coordinate bound. Substituting the tolerances from (E.9), we find that

$$\sqrt{d_{\text{eff}}(w)}\{8(d_{\text{eff}}(w) + 1)QK\nu + QK\rho\} < \frac{\delta}{8N}.$$

Since $d_{\text{eff}}(w) \ge 1$, the same bound covers the scalar denominator output.

• **Step 4: bound the number of registers.** Evaluate monomials and numerator coordinates sequentially. For each of the Q multi-indices there are at most $d_{\text{eff}}(w)(d_{\text{eff}}(w) + 1)$ one-dimensional moment calls: $d_{\text{eff}}(w)$ for the denominator and $d_{\text{eff}}(w)$ for each numerator coordinate. A call uses at most $14(k+1)+10$ affine and product stages after its four Gaussian values are available. Each two-step recurrence update needs at most three products, three affine operations, and two clips; forming the final J_r adds at most five stages. Recompute these values when needed instead of storing all $d_{\text{eff}}(w)k$ moments. Besides the $d_{\text{eff}}(w)$ retained inputs, keep $d_{\text{eff}}(w)$ numerator accumulators, one denominator accumulator, and a fixed number of registers for scalar operations. More explicitly, reserve $2d_{\text{eff}}(w) + 11$ registers for the retained inputs, output accumulators, nine scalar coefficients, and time; four for means and standardized endpoints; four for Gaussian values; twelve for evaluating scalar functions; five for the moment recurrence; eight for products and temporaries; and one constant. This is at most $2d_{\text{eff}}(w) + 48$ registers. Sequential compilation uses width $4d_{\text{eff}}(w) + 106$. The paired scalar recursions need only $4d_{\text{eff}}(w) + 58$ neurons while carrying the retained inputs and accumulators. The exact signed input lift uses $2d_{\text{eff}}(w) + 2$ neurons and need not retain any ambient input coordinate. This proves the stated width bounds.

• **Step 5: implement the circuit and count its layers.** To bound its size, define

$$\begin{aligned} E &= 1 + K + \log\left(\frac{eA}{\nu}\right), \\ S &= \left\lceil c_0(Q+1)(d_{\text{eff}}(w)+1)^2(k+1)^2KE \right\rceil, \\ A_* &= \exp\{c_0(k+1)^2KE\}. \end{aligned} \tag{E.13}$$

Here S bounds the number of arithmetic stages, and A_* bounds every coefficient and intermediate magnitude. To verify these claims, first use $A \geq 2^{12}$ and $E \geq 8$ to obtain

$$\log(8/\eta_\Phi) \leq 16(k+1)E, \qquad \log\left(\frac{eK(1+\sigma_{t_0}^{-2})}{\rho}\right) \leq 23(k+1)E.$$

Thus each Gaussian polynomial has degree at most $400(k+1)E$. Four sequential Horner evaluations, followed by the recurrence, endpoint operations, and product accumulation, take at most $4096(k+1)E$ stages per moment call. The initial scalar-coefficient calculations require at most $7820K(k+1)E$ stages. With the final accumulations included, the total stage count satisfies

$$\begin{aligned} S_{\text{actual}} &\leq 7820K(k+1)E + 4096Qd_{\text{eff}}(w)(d_{\text{eff}}(w)+1)(k+1)E + 40Q(d_{\text{eff}}(w)+1) \\ &\leq 2^{11}(Q+1)(d_{\text{eff}}(w)+1)^2(k+1)^2KE \leq S. \end{aligned}$$

The second inequality uses $K \geq 4$, $d_{\text{eff}}(w) \geq 1$, $Q \geq 1$, and $E \geq 8$. To bound the intermediate values, first note that the scalar-coefficient calculations have absolute values at most $\exp\{46K(k+1)E\}$, and each Gaussian polynomial has sum of absolute coefficients at most $16\eta_\Phi^{-2}$. Each partial sum of cell terms is at most $2(Q+1)K$. Each moment computed before clipping has absolute value at most $4A^3$; clipping reduces this bound to one. The logarithms of these upper bounds are at most $64(k+1)KE \leq c_0(k+1)^2KE$, so all the quantities are bounded by A_*.

Now apply Lemma E.3 with final coordinate tolerance $\eta = \delta/[2^{14}(N+1)(d_{\text{eff}}(w)+1)]$ and the bounds S, A_* above. The additional vector error is at most $\sqrt{d_{\text{eff}}(w)}\eta < \delta/(8N)$. Adding it to the scalar-function error already allocated above proves (E.10). The resulting network has every weight and bias bounded by four.

It remains to express its depth in the parameters of the statement. We have $\log(Q+1) \leq (d_{\text{eff}}(w)+k+1)\log 2 \leq K\log 2$. Substituting $\log(1+R_{\text{clean}}+KR_\epsilon) \leq \log K + \log(K+R_\epsilon+R_{\text{clean}})$ into the definitions, we find that $E \leq 16K\Lambda$. The compiler logarithm is at most $2c_0(k+1)^2KE$, while $S+1$ is at most twice the unrounded expression in (E.13). Hence its depth is bounded by

$$\begin{aligned} 128c_0^3(Q+1)^2(d_{\text{eff}}(w)+1)^4(k+1)^6K^3E^3 &\leq 2^{19}c_0^3(Q+1)^2K^{16}\Lambda^3 \\ &\leq c_0^{16}(Q+1)^2K^{20}\Lambda^3. \end{aligned}$$

This proves (E.11). □

Lemma E.7 (Norm bounds under constant-coordinate rescaling)**.** Suppose $L \geq 2$ and a depth-L branch g, including any identity padding, has width at most $W-1$ and the sum of squared weight and bias norms in each layer is at most $E \geq 0$. For any $q > 0$, the function qg has a depth-L, width-W representation with sum of squared Frobenius norms and squared bias norms at most

$$L(E+1)q^{2/L}. \tag{E.14}$$

Proof. We adapt the positive-homogeneity rescaling argument used for Claim 1 in Neyshabur et al. (2015), adding one coordinate to include the biases. Append a constant-one neuron in the first hidden layer, replace subsequent hidden maps by $\left[\begin{smallmatrix} A_j & b_j \\ 0 & 1 \end{smallmatrix}\right]$ with zero bias, and use $[A_L \; b_L]$ in the output layer. The added coordinates contribute at most $L-1$ to the squared parameter norm. Multiplying the first matrix and bias, and each subsequent augmented matrix, by $q^{1/L}$ multiplies the output by q, by positive homogeneity. The resulting sum of squared norms is at most

$$q^{2/L}\{LE + (L-1)\} \leq L(E+1)q^{2/L}.$$

□

E.3 Covering parallel branches and shared output layers

We first cover the sum of the parallel branches and then the shared output layers. All uniform errors below are measured in the Euclidean norm of the output. Discarding small branches follows the argument in (Zhang et al., 2024c, Section 4.2, Lemma 4.1, equation (11)). That result concerns convolutional residual networks and suppresses logarithmic factors; the argument below retains explicit constants for the branch biases, affine aggregation, and shared layers in (3.5).

Lemma E.8 (Covering the affine aggregation of centered branches)**.** Let $L \geq 2$, $W \geq d$, $R \geq 1$, and let $B, B_C \geq 0$. Consider every encoder $g(y) = \sum_{m=1}^{M} C_m g_m(y) + b_C \in \mathbb{R}^W$, $y \in \mathbb{R}^{d+1}$, whose branches have exact depth L, widths and output dimensions at most W, sum of squared weight norms and squared bias norms at most B, and $\sum_m \|C_m\|_F^2 + \|b_C\|_2^2 \leq B_C$. The auxiliary B_C is used only in this lemma. Put

$$V_e = \sqrt{B_C}(B/L)^{L/2}, \quad O_e = \sqrt{B_C}(1+B/L)^{L/2}, \quad d_e = (L-1)(W^2+W)+(d+1+W)W.$$

For $\eta > 0$, set $K = \max\{1, \lceil (4RV_e/\eta)^{2/(L-1)} \rceil\}$. The union over every finite M has an external uniform η-cover on $\|y\| \leq R$ with logarithmic cardinality at most

$$\mathcal{E}_R(\eta) := W \log(1+8O_e/\eta) + K \log(1+4RKV_e/\eta) + Kd_e \log(1+64R(L+1)\sqrt{W}V_e/\eta). \quad \text{(E.15)}$$

An internal cover of radius η has logarithmic cardinality at most $\mathcal{E}_R(\eta/2)$.

Proof. The proof has the following steps.

- Separate the constant output and bound its magnitude.
- Rescale each centered branch so that its parameters are bounded.
- Construct a finite cover of the rescaled branches.
- Discard branches whose contribution is small and bound the resulting error.
- Cover the coefficients of the remaining branches and add the approximation errors.

• **Step 1: separate and bound the constant output.** We first separate the constant contribution of all biases. With aggregation blocks $[C_1, \ldots, C_M]$, the encoder can be written as

$$g(y) = g(0) + \sum_m C_m\{g_m(y) - g_m(0)\}.$$

To bound $g(0)$, stack the first-layer maps vertically and the subsequent maps in block-diagonal form, append one shared constant coordinate to each layer, and write $e_j = \sum_m (\|A_j^m\|_F^2 + \|b_j^m\|_2^2)$. Each augmented layer has squared Frobenius norm $1 + e_j$; the last readout $[C_1, \ldots, C_M, b_C]$ has norm at most $\sqrt{B_C}$. The augmented input at $y = 0$ has norm one. ReLU does not increase Euclidean norm, and the identity activation in the last affine layer has the same property. The norm of the constant output is bounded by the product of the layer norms:

$$\|g(0)\|_2 \leq \sqrt{B_C} \prod_{j=1}^{L} \sqrt{1+e_j} \leq \sqrt{B_C}(1+B/L)^{L/2} = O_e.$$

The second inequality follows from the arithmetic–geometric mean inequality. Apply the Euclidean-ball covering bound in (Vershynin, 2018, Corollary 4.2.13) with dimension W, ball radius O_e, and covering radius $\eta/4$. The constant outputs can therefore be covered by at most $(1 + 8O_e/\eta)^W$ centers.

• **Step 2: rescale the centered branches.** We normalize each centered branch and bound its biases without changing its values on the input ball. If $\|C_m\|_F = 0$, or if some $\|A_j^m\|_F = 0$, the centered contribution of this branch is zero and can be discarded. For every other branch suppress m temporarily and write its hidden outputs as $h_0(y) = y$ and $h_j(y) = \mathrm{ReLU}(A_j h_{j-1}(y) + b_j)$ for $1 \leq j < L$. Define $\bar{A}_j = A_j/\|A_j\|_F$ for $1 \leq j \leq L$ and set $z_0(y) = y$. For $1 \leq j < L$, define the normalized hidden states and thresholds by

$$z_j(y) = \frac{h_j(y) - h_j(0)}{\prod_{i=1}^{j} \|A_i\|_F}, \qquad \tau_j = \frac{A_j h_{j-1}(0) + b_j}{\prod_{i=1}^{j} \|A_i\|_F}.$$

By positive homogeneity of ReLU, the normalized hidden states satisfy

$$z_j(y) = \mathrm{ReLU}(\bar{A}_j z_{j-1}(y) + \tau_j) - \mathrm{ReLU}(\tau_j), \qquad 1 \le j < L.$$

Because $\|\bar{A}_j\|_F = 1$, its operator norm is at most one. Subtracting the same shifted ReLU at zero and using its one-Lipschitz property proves $\|z_j(y)\|_2 \le \|z_{j-1}(y)\|_2 \le R$ inductively.

Each coordinate of $\bar{A}_j z_{j-1}(y)$ lies in $[-R, R]$. On this interval, $\mathrm{ReLU}(s+\tau) - \mathrm{ReLU}(\tau)$ equals s for $\tau \ge R$ and zero for $\tau \le -R$. Clipping every threshold to $[-R, R]$ therefore leaves the recursion unchanged.

Keep C_m for this branch's readout block, and let $\bar{C} = C_m/\|C_m\|_F$ and $a = \|C_m\|_F \prod_{j=1}^{L} \|A_j\|_F$. The branch now has the form

$$C_m\{g_m(y) - g_m(0)\} = a\, v(y), \qquad v(y) = \bar{C}\bar{A}_L z_{L-1}(y).$$

The normalized branch v satisfies $v(0) = 0$, $\sup_{\|y\|_2 \le R} \|v(y)\|_2 \le R$, and is one-Lipschitz. Zero padding makes every hidden and branch-output dimension exactly W. Its parameters consist of $W(d+1)$ entries in $\bar{A}_1$, $(L-1)W^2$ in the remaining branch matrices, $(L-1)W$ thresholds, and W^2 in $\bar{C}$. Thus their total count is d_e.

• **Step 3: cover the rescaled branches.** We cover the normalized parameters and propagate their errors through the branch. Cover every matrix's Frobenius unit ball with radius $\zeta/[4R(L+1)]$ and every threshold cube with Euclidean radius $\zeta/[8(L-1)]$. Choose matrix centers in their unit balls and threshold centers in their cubes. For every branch represented by a cover center, the intermediate outputs therefore still have norm at most R.

The elementary shifted-ReLU estimate used to propagate these errors is

$$\|[\mathrm{ReLU}(u+\tau) - \mathrm{ReLU}(\tau)] - [\mathrm{ReLU}(\widetilde{u}+\widetilde{\tau}) - \mathrm{ReLU}(\widetilde{\tau})]\|_2 \le \|u - \widetilde{u}\|_2 + 2\|\tau - \widetilde{\tau}\|_2.$$

Applying this estimate successively through the hidden layers and readouts, we find that

$$\|v - \widetilde{v}\|_\infty \le R\sum_{j=1}^{L} \|\bar{A}_j - \widetilde{A}_j\|_F + R\|\bar{C} - \widetilde{C}\|_F + 2\sum_{j=1}^{L-1} \|\tau_j - \widetilde{\tau}_j\|_2 \le \zeta/2.$$

Apply the covering bounds in Vershynin (2018, Proposition 4.2.12 and Corollary 4.2.13) to each matrix ball of radius one and each threshold cube contained in the radius-$R\sqrt{W}$ Euclidean ball, using the respective covering radii above. The proposition permits centers inside the cube. The matrix and threshold cover factors are $1 + 8R(L+1)/\zeta$ and $1 + 16R(L-1)\sqrt{W}/\zeta$, respectively. Multiplying these factors with exponents equal to their parameter counts, we obtain

$$\log \mathcal{N}_{\mathrm{shape}}(\zeta) \le d_e \log\left(1 + \frac{16R(L+1)\sqrt{W}}{\zeta}\right).$$

• **Step 4: bound the error from discarding small branches.** We use the norm budgets to bound the error of keeping only K branches. Restore the index m, set $e_m = \sum_j (\|A_j^m\|_F^2 + \|b_j^m\|_2^2)$, and let $a_m = \|C_m\|_F \prod_j \|A_j^m\|_F$. Apply arithmetic–geometric mean inside each branch and then Hölder's inequality across branches to obtain

$$\begin{aligned}\sum_m a_m^{2/(L+1)} &\le \sum_m (e_m/L)^{L/(L+1)} \|C_m\|_F^{2/(L+1)} \\ &\le (B/L)^{L/(L+1)} B_C^{1/(L+1)} = V_e^{2/(L+1)}.\end{aligned}$$

If $V_e = 0$, there is no nonconstant contribution and Step 1 already proves the lemma. Assume $V_e > 0$ and order the amplitudes decreasingly. Apply the best-K-term approximation inequality (Foucart et al., 2010, equation (1.5)) to the amplitude vector, with source exponent $2/(L+1)$ and target exponent one. Since $L \ge 2$, the source exponent is strictly smaller than the target exponent, as required. Combining this inequality with the preceding norm bound, we obtain

$$\sum_{m>K} a_m \le K^{-(L-1)/2} \left(\sum_m a_m^{2/(L+1)}\right)^{(L+1)/2} \le V_e K^{-(L-1)/2}.$$

The bound also holds for an empty tail. The same amplitude budget implies $\sum_m a_m \le V_e$ by monotonicity of the ℓ^p quasinorms. Because each normalized branch has norm at most R, the chosen value of K ensures that

$$\left\| \sum_{m>K} a_m v_m \right\|_\infty \le R V_e K^{-(L-1)/2} \le \eta/4.$$

• **Step 5: cover the remaining coefficients and add the errors.** We combine the errors from covering the constant output, discarding small branches, and approximating the remaining normalized branches and amplitudes. Cover each of the at most K retained normalized branches with accuracy $\zeta = \eta/(4V_e)$. Because the amplitudes sum to at most V_e, their combined approximation error is at most $\eta/4$. The logarithm of the number of choices for these branches is at most

$$K d_e \log\left(1 + 64R(L+1)\sqrt{W}\, V_e/\eta\right).$$

Round each retained amplitude down to a multiple of $\eta/(4RK)$. Each amplitude belongs to $[0, V_e]$, so at most $(1 + 4RKV_e/\eta)^K$ choices are needed. The total amplitude error is at most $K\{\eta/(4RK)\}R = \eta/4$, because the normalized branches in the cover also have norm at most R.

The four errors come from the constant-output cover, the discarded branches, the normalized-branch covers, and amplitude rounding. Each is at most $\eta/4$; their sum is at most η. Multiplying the three finite-cover sizes proves the stated logarithmic bound. The centers need not belong to the original norm class. For an internal cover, retain one original function from each nonempty radius-$\eta/2$ external ball. These functions form a cover of radius η by the triangle inequality, with no increase in the number of centers. □

Lemma E.9 (Covering parallel branches with shared output layers)**.** Let $L \ge 2$, $W \ge d$, $R \ge 1$, and $B, B_{\text{out}} \ge 0$. Consider the union of (3.5), with $S_t \equiv \infty$, over all finite numbers of branches. Zero padding makes the shared bottleneck dimension exactly W. Define

$$\begin{gathered} V_e = \sqrt{B_{\text{out}}}(B/L)^{L/2}, \qquad O_e = \sqrt{B_{\text{out}}}(1 + B/L)^{L/2}, \\ A_h = (1 + B_{\text{out}}/L)^{L/2}, \qquad R_h = O_e + RV_e, \\ d_e = (L-1)(W^2 + W) + (d + 1 + W)W, \qquad P_h = L(W^2 + W), \\ K_\delta = \max\{1, \lceil (16RA_hV_e/\delta)^{2/(L-1)} \rceil\}. \end{gathered} \tag{E.16}$$

For every $\delta > 0$, an external uniform δ-cover on $\|y\| \le R$ has logarithmic cardinality at most

$$\begin{aligned} \mathcal{H}_R(\delta) := {} & W \log(1 + 32A_hO_e/\delta) + K_\delta \log(1 + 16RK_\delta A_hV_e/\delta) \\ & + K_\delta d_e \log(1 + 256R(L+1)\sqrt{W}A_hV_e/\delta) + \log L \\ & + P_h \log\left(1 + \frac{4\sqrt{LB_{\text{out}}}\sqrt{1 + R_h^2}A_h}{\delta}\right). \end{aligned} \tag{E.17}$$

The logarithmic cardinality of an internal δ-cover is at most $\mathcal{H}_R(\delta/2)$. The displayed bound is independent of the available branch count M.

Proof. The proof has the following steps.

- Apply the preceding lemma to the sum of the parallel branches.
- Bound how much the shared output layers amplify an input error.
- Construct a finite cover of the shared-layer parameters.
- Combine the two covers and add their logarithmic cardinalities.

• **Step 1: cover the sum of the parallel branches.** Write the full network as $\rho \circ h$, where $h = \sum_m C_m g_m + b_C$ is its first shared affine map. The map ρ includes the first shared ReLU and the remaining affine layers. The joint budget implies both $\sum_m \|C_m\|_F^2 + \|b_C\|^2 \le B_{\text{out}}$ and a shared-layer norm bound at most B_{out}. Enlarging to this product of two budget sets is legitimate when

bounding a covering number. Apply Lemma E.8, with its auxiliary $B_C = B_{\text{out}}$, to construct an internal encoder cover of radius $\delta/(2A_h)$, with logarithmic cardinality at most $\mathcal{E}_R(\delta/(4A_h))$. Every such encoder and internal center satisfies

$$\sup_{\|y\|\le R} \|h(y)\| \le \|h(0)\| + RV_e \le O_e + RV_e = R_h.$$

Thus the inputs to the shared output layers lie in the ball of radius R_h.

• **Step 2: bound the Lipschitz constant of the shared output layers.** Let ρ have $j \le L$ affine layers, and let e_i be the sum of squared weights and biases in shared layer i. The Lipschitz constant with respect to the input is at most $\prod_{i=1}^{j} \sqrt{e_i}$. To account for biases and parameter perturbations, augment each layer with its constant coordinate, whose Frobenius norm is $\sqrt{1+e_i}$. Inserting $L-j$ factors equal to one and applying the arithmetic–geometric mean inequality, we obtain

$$\text{Lip}(\rho) \le \prod_{i=1}^{j} \sqrt{1+e_i} \le (1 + B_{\text{out}}/L)^{L/2} = A_h.$$

• **Step 3: cover the parameters of the shared output layers.** Fix the number $j \le L$ of affine layers in ρ and pad its hidden widths to W. Its scalar parameter count is at most $P_h = L(W^2+W)$. Join two parameter vectors for ρ by a straight segment inside their Euclidean radius-$\sqrt{B_{\text{out}}}$ ball. At almost every point along the segment, the chain rule implies that the derivative of the network along the segment satisfies

$$\|\dot\rho(z)\| \le \sqrt{1+R_h^2} \sum_{i=1}^{j} \|\dot\theta_i\| \prod_{k\ne i} \sqrt{1+e_k} \le \sqrt{L}\sqrt{1+R_h^2} A_h \|\dot\theta\|.$$

ReLU is differentiable along these piecewise-polynomial paths except at finitely many breakpoints, and is absolutely continuous. Integrating along the segment proves the same Lipschitz bound for changes in the parameters. Cover the parameter ball with Euclidean covering radius $\delta/[2\sqrt{L}\sqrt{1+R_h^2}A_h]$, taking centers inside the ball. Apply the Euclidean-ball covering bound in Vershynin (2018, Corollary 4.2.13) with dimension at most P_h and ball radius $\sqrt{B_{\text{out}}}$. Taking the union over the L possible depths, we obtain

$$\log N_h(\delta/2) \le \log L + P_h \log\left(1 + \frac{4\sqrt{LB_{\text{out}}}\sqrt{1+R_h^2}A_h}{\delta}\right).$$

The first shared affine map has already been covered through its centered branch contributions in Step 1, so it is not counted again here.

• **Step 4: combine the two covers.** Choose h_* and ρ_* from the two covers to approximate h and ρ, respectively. Then the total approximation error satisfies

$$\|\rho(h(y)) - \rho_*(h_*(y))\| \le A_h\|h(y) - h_*(y)\| + \|\rho(h_*(y)) - \rho_*(h_*(y))\| \le \delta.$$

Both terms are at most $\delta/2$. Multiplying the two cover sizes, we obtain

$$\log N_{\text{comp}}(\delta) \le \mathcal{E}_R(\delta/(4A_h)) + \log N_h(\delta/2) \le \mathcal{H}_R(\delta).$$

Substituting (E.15) into this bound proves (E.17). For internal centers, retain one original function in each nonempty external ball of radius $\delta/2$. This doubles its radius without increasing cardinality, and also preserves any imposed output cap. □

E.4 Entropy identities for the remainders

The following identity bounds the two score residuals in Appendix B.2 by the corresponding relative entropies at time zero. Let Q_t have the auxiliary density q_t from Appendix A.

Proposition E.10. Under Assumption 4.1, for $0 < t_0 < T$, we have that

$$\mathcal{E}_{\text{dep}} + \mathcal{E}_{\text{cutoff}} = 2\{\text{KL}(P_{t_0}\|Q_{t_0}) - \text{KL}(P_T\|Q_T)\}. \tag{E.18}$$

Equivalently, it holds that

$$\mathcal{E}_{\text{dep}} + \mathcal{E}_{\text{cutoff}} = 2\{I_w(t_0) - I_w(T)\} + 2\{K_w(t_0) - K_w(T)\}. \tag{E.19}$$

In particular, we have that

$$\mathcal{E}_{\text{dep}} + \mathcal{E}_{\text{cutoff}} \le 2\{I_w(0) + K_w(0)\} = 2\,\text{KL}(P_0\|Q_0). \tag{E.20}$$

Proof. All terms vanish when $d_{\text{eff}}(w) = d$. Otherwise, the proof has two steps:

- Express $\mathrm{KL}(P_t \| Q_t)$ as $I_w(t) + K_w(t)$.
- Differentiate this relative entropy and integrate over time.

• **Step 1: separate the two entropy terms.** In the orthogonal coordinates $(U_w^\top x, V_w^\top x)$, the law Q_t is the product of the retained marginal of P_t and $\mathsf{N}(0, (\alpha_t^2/C + \sigma_t^2) I_{d-d_{\text{eff}}(w)})$. The relative-entropy decomposition for a product reference measure (Gray, 1990, Theorem 5.3.1 and Corollary 5.3.1), applied to these two coordinate blocks, therefore reads

$$\mathrm{KL}(P_t \| Q_t) = I_w(t) + K_w(t), \qquad t \geq 0.$$

• **Step 2: integrate relative-entropy dissipation.** The laws P_t and Q_t are OU evolutions of P_0 and Q_0; under Q_0, the discarded block is independent of the retained block and has law $\mathsf{N}(0, C^{-1} I_{d-d_{\text{eff}}(w)})$. The multivariate Gaussian-perturbation identity (Yoshida, 2017, Lemma 2, equations (9)–(10)) applies with noise variance $\sigma_t^2/\alpha_t^2 = e^t - 1$, followed by multiplication by α_t. Relative entropy is unchanged by this multiplication, whereas the integral of the squared score difference is multiplied by $\alpha_t^{-2} = e^t$. This is also the derivative of the noise variance, so the identity becomes

$$\frac{d}{dt} \mathrm{KL}(P_t \| Q_t) = -\frac{1}{2} \mathbb{E}_{P_t} \left[\| \nabla \log p_t(X_t) - \nabla \log q_t(X_t) \|_2^2 \right].$$

To check regularity, completing the square in the Gaussian convolution formulas shows that $p_t / \phi_{d,(\alpha_t^2 + C\sigma_t^2)/C}$ and $q_t / \phi_{d,(\alpha_t^2 + C\sigma_t^2)/C}$ lie in $[c, H_\beta]$. Their spatial derivatives are bounded on compact positive-time intervals. Thus the densities are smooth and positive, with Gaussian tails and finite Fisher information, as required by the cited identity.

By (A.4) and (B.18), the time integral of the squared score difference is $\mathcal{E}_{\text{dep}} + \mathcal{E}_{\text{cutoff}}$. Integrating the derivative proves (E.18), and the decomposition in Step 1 proves (E.19). Finally, the data-processing inequality (Gray, 1990, Corollary 5.2.2), applied to the common map $(x, z) \mapsto \alpha_{t_0} x + \sigma_{t_0} z$ with independent $z \sim \mathsf{N}(0, I_d)$, implies $\mathrm{KL}(P_{t_0} \| Q_{t_0}) \leq \mathrm{KL}(P_0 \| Q_0)$. Together with $\mathrm{KL}(P_T \| Q_T) \geq 0$ and Step 1, this proves (E.20). □

E.5 Weight decay and constrained norm budgets

The quadratic penalties in (4.3) control the sums of squared weights and biases in (3.5), as shown below.

Proposition E.11 (Norm bounds implied by weight decay)**.** Let $0 < t_0 < T$, $S_t \geq 0$, and $\lambda, \lambda_{\text{out}} > 0$. Any global minimizer of (4.3) belongs to $\mathcal{F}(M, L, W, B, B_{\text{out}}, S_t)$ with

$$B = \frac{d}{\lambda} \log \frac{e^T - 1}{e^{t_0} - 1}, \qquad B_{\text{out}} = \frac{d}{\lambda_{\text{out}}} \log \frac{e^T - 1}{e^{t_0} - 1}.$$

Proof. Setting all weights and biases to zero gives the zero network, which satisfies the output bound and has zero penalty. By the definition of the denoising loss in (3.3), its loss is

$$\widehat{\mathcal{L}}(0; \mathcal{D}_n) = d \int_{t_0}^{T} \sigma_t^{-2} \, \mathrm{d}t = d \log \frac{e^T - 1}{e^{t_0} - 1}.$$

The penalized objective at a global minimizer is no larger than its value at the zero network. Since the denoising loss is nonnegative, the parameters of this minimizer satisfy

$$\lambda \sum_{m=1}^{M} \sum_{j=1}^{L} \left(\|A_j^m\|_{\mathrm{F}}^2 + \|b_j^m\|_2^2 \right) + \lambda_{\text{out}} \sum_{j=1}^{L} \left(\|A_j^{\text{out}}\|_{\mathrm{F}}^2 + \|b_j^{\text{out}}\|_2^2 \right) \leq d \log \frac{e^T - 1}{e^{t_0} - 1}.$$

Both sums are nonnegative. Dividing by the respective penalty coefficients proves the two norm bounds. □

Increasing either penalty coefficient decreases the corresponding bound. The bounds are independent of the number of parallel networks.

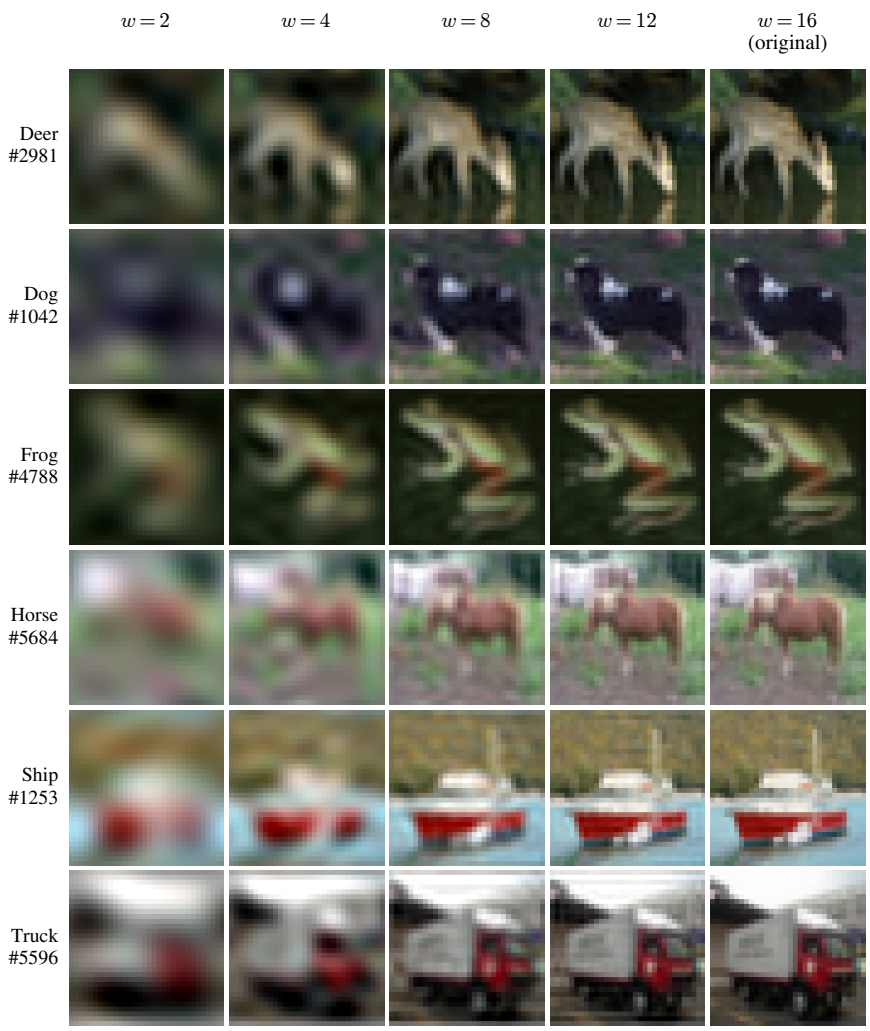


Figure 3: **Fourier resolutions of real CIFAR-10 images.** Each row shows the same official test image at cutoffs $w \in \{2, 4, 8, 12, 16\}$, displaying $\mathcal{Q}(\Pi_w x)$. At $w = 16$, the displayed image is the original. Six classes were selected uniformly without replacement, followed by one uniformly sampled test image per class, using seed 20260926. Row labels give classes and zero-based test indices.

F Visualizing Fourier projections on CIFAR-10

To illustrate the frequency resolutions considered in our experiments and the image sets compared by pFID, we visualize real and generated CIFAR-10 images at several Fourier cutoffs. All displayed images use $\mathcal{Q}(\Pi_w x)$ with the same channelwise square Fourier mask, clipping, and 8-bit quantization as the evaluation protocol in Appendix D.4. Projections are computed on the original 32×32 grid and enlarged using nearest-neighbor interpolation, with a common RGB range and no per-image contrast normalization. Figure 3 follows six fixed test images across cutoffs, illustrating how the retained visual content changes with w. Figure 4 compares real samples with unconditional samples from models trained on different dataset sizes at $\lambda = 0$. Within each row, the same images are reused across cutoffs. These examples provide visual context for the full-sample pFID measurements; the real and generated rows are distribution samples rather than paired reconstructions.

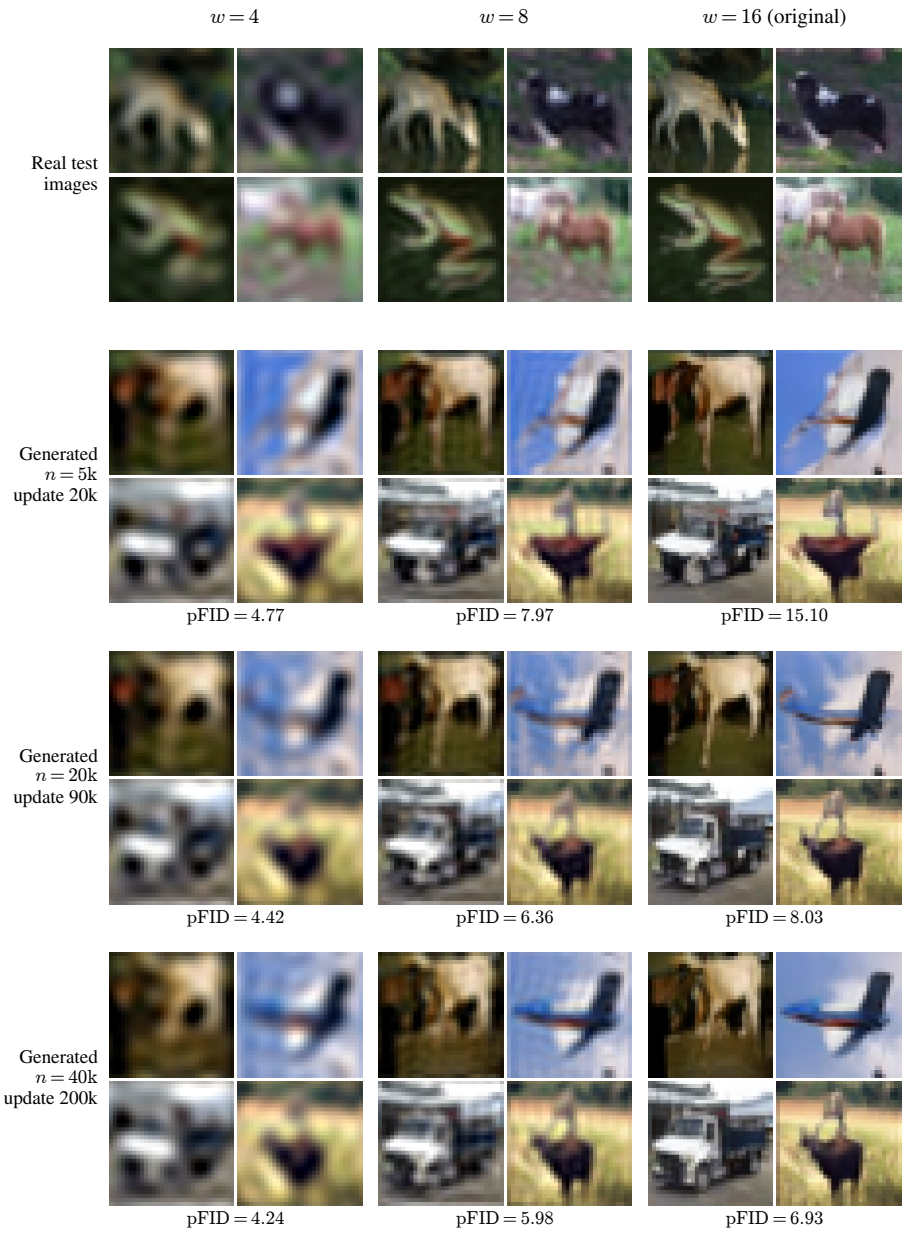


Figure 4: **Real and generated images at matched Fourier cutoffs.** Each cell shows four samples through $\mathcal{Q}(\Pi_w x)$, reused across columns. Models use $\lambda = 0$ and validation-selected EMA checkpoints within 200000 updates. Real images are the first four rows of Figure 3; generated indices 974, 4006, 4071, 5425 (zero-based, row-major) were sampled once with seed 20260927 and reused across models. pFID annotations use the full 10000-versus-10000 evaluation. Shared generated indices do not guarantee matching classes or objects across models.