CONDITIONAL DIFFUSION WITHOUT ERM ORACLES: FROM FINITE NEURAL TRAINING TO SAMPLING GUARANTEES

Anonymous authors
Paper under double-blind review

ABSTRACT

Conditional diffusion guarantees based on empirical risk minimization do not directly describe the accuracy achieved by finite neural-network training. We establish an end-to-end sampling guarantee for a finite deep ReLU denoiser trained in all layers by finitely many steps of full-batch gradient descent, without an empirical risk minimization oracle. Under explicit conditional tail and regularity assumptions, the resulting reverse sampler achieves conditional total-variation error $\widetilde{\mathcal{O}}(\varepsilon)$, averaged over the conditioning variable, for the conditional law at a fixed positive diffusion stopping time. The sufficient clean-sample size is $\widetilde{\mathcal{O}}(\varepsilon^{-[2+(D+1)/\bar{k}]})$, where $D = d_x + d_y$ combines the data and conditioning dimensions, $k \geq 1$ is the integer regularity order in the conditional-law assumptions, and $\bar{k} = \min\{k, (D+2)/2\}$ is the effective smoothness. Our analysis derives denoising regularity and approximation bounds from the conditional data law and transfers population learning guarantees to the actual finite network. The key step uses Duhamel-based trajectory comparisons to control how discrepancies between neural and kernel training accumulate over a finite horizon and affect predictions at unseen inputs. This yields an explicit tradeoff between training duration and sufficient network width, without a Polyak–Łojasiewicz condition or a positive lower bound on the empirical kernel spectrum. The bounds keep clipping, learning, and finite-width errors explicit. Recovering the original conditional law additionally requires smoothing control and a corresponding resource calibration.

1 INTRODUCTION

Diffusion models generate samples by learning to reverse a process that gradually adds noise to observations (Ho et al., 2020). In conditional generation, this reverse process must also respond to an observed covariate, so the learned denoiser depends jointly on the noisy observation, diffusion time, and conditioning variable. Understanding the resulting sampling accuracy therefore requires connecting three things: the available data, the network to represent the denoiser, and the number of training steps.

Existing theoretical guarantees address different parts of this connection. Sampling analyses establish distributional accuracy given a sufficiently accurate score estimate (Chen et al., 2023b; Benton et al., 2024). Statistical analyses of conditional diffusion additionally study score estimation, often through a neural estimator defined by empirical risk minimization (Yuan et al., 2023; Fu et al., 2024; Tang et al., 2025; Hu et al., 2025). These results explain what can be achieved by an accurate estimator, but an empirical risk minimization assumption does not specify how finitely many gradient steps produce that estimator. A guarantee for the trained generative model must account for this training procedure and its error at previously unseen inputs.

Recent work has begun to address finite optimization in diffusion and flow matching (Gaur et al., 2026; Han et al., 2024; He et al., 2026). These analyses study different architectures, sampling objectives, and optimization assumptions. Our focus is the joint problem of conditional sampling with a finite deep network whose hidden layers and readout are all trained, with approximation and learning guarantees derived from assumptions on the data law. This motivates the following question:

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.

How much data and training are sufficient for a finite deep network, trained in every layer, to generate accurate samples from a conditional distribution without an empirical risk minimization oracle?

Our answer. We establish an end-to-end guarantee for the specified family of centered deep ReLU denoisers, with independent Gaussian initialization and every hidden layer and readout trained by full-batch gradient descent. Under the conditional tail and regularity assumptions in Assumptions 3.1 and 3.2, Theorem 1 gives a sufficient clean-sample size

$$N = \widetilde{O}\Big(\varepsilon^{-[2+(D+1)/\bar{k}]}\Big), \qquad D = d_x + d_y, \qquad \bar{k} = \min\Big\{k, \frac{D+2}{2}\Big\},$$

where d_x and d_y are the data and conditioning dimensions, $k \geq 1$ is the integer conditional-law regularity order in Assumption 3.2, and $\bar{k}$ is the effective smoothness. For fixed depth and diffusion stopping time $t_0 > 0$, explicit finite widths and a deterministic training schedule yield, with high probability,

$$\mathbb{E}_Y[\mathrm{TV}(p_{t_0}(\cdot \mid Y), \widehat{p}_{t_0}(\cdot \mid Y))] \leq \widetilde{O}(\varepsilon).$$

Here p_{t_0} is the conditional law after forward diffusion for time t_0, and $\widehat{p}_{t_0}$ is the output law of the learned reverse sampler. The expectation uses a fresh conditioning variable with the trained predictor fixed. The theorem specifies sufficient network widths and training durations from supplied model-class bounds. Recovery of the original conditional law additionally requires smoothing control and a corresponding resource calibration (Corollary 1).

Technical approach. We derive denoising regularity from the conditional data law and obtain population learning bounds for the limiting-kernel predictor. To transfer these bounds to the trained network, we compare neural, finite-kernel, and limiting-kernel training trajectories. Duhamel's identity expresses their discrepancies as accumulated one-step errors arising from kernel drift, nonlinear remainders, and finite-width kernel approximation. Controlling these errors on the training sample and at fresh inputs yields an explicit tradeoff between training duration and sufficient network width, without requiring a Polyak–Łojasiewicz condition or a positive lower bound on the empirical Gram spectrum. Score-weighted training then links the resulting population denoising error to sampling accuracy. Appendix F.1 explains the Duhamel identity and these trajectory comparisons.

Contributions. Our main contributions are as follows.

1. **Conditional sampling guarantees from finite neural training.** We connect finitely many gradient steps on an all-trained deep ReLU network to averaged conditional total-variation accuracy. The result specifies sufficient sample size, training duration, network width, and reverse-sampling resources without assuming access to an empirical risk minimizer.
2. **Population control over a finite training horizon.** We establish Duhamel-based trajectory comparisons between neural, finite-kernel, and limiting-kernel training trajectories. The bounds account for unequal hidden widths and a trainable Gaussian readout, control predictions at unseen inputs, and quantify how sufficient width depends on the training horizon.
3. **Approximation and learning derived from the conditional law.** We construct bounded approximating functions for the denoising target and establish population learning bounds under the discrete time distribution used by the sampler. These bounds retain clipping, localization, approximation, and statistical errors, allowing them to be calibrated jointly with finite-width and neural-training errors.

2 Problem Setup: Multilayer Conditional Score Learning

Data and forward diffusion. We observe N i.i.d. clean pairs $(X_{0,i}, Y_i)$, where $X_0 \in \mathbb{R}^{d_x}$ is the data variable and $Y \in \mathbb{R}^{d_y}$ is the conditioning variable. Write $D = d_x + d_y$ and $\mu_y = \mathcal{L}(X_0 \mid Y = y)$. The forward process is

$$\mathrm{d}X_t = -X_t\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \qquad X_0 \mid Y = y \sim \mu_y,$$

where B is Brownian motion independent of (X_0, Y). For each fixed $t > 0$, conditional on (X_0, Y), the law of X_t is $N(\alpha_t X_0, \sigma_t^2 I_{d_x})$, where $\alpha_t = e^{-t}$ and $\sigma_t^2 = 1 - e^{-2t}$. Let $p_t(\cdot \mid y)$ denote the

Table 1: Comparison of diffusion guarantees. ✓ indicates a property established in the compared result; × indicates that it is outside that guarantee's scope. Assumptions and error criteria differ across rows.

Work	Conditional diffusion	Finite GD/SGD	Deep all-layer training	Derived approximation	Error guarantee
Yuan et al. (2023)	✓	×	×	✓[a]	Latent stopped TV
Fu et al. (2024)	✓	×	×	✓	Averaged TV
Tang et al. (2025)	✓	×	×	✓	Averaged TV; manifold W_1
Hu et al. (2025)	✓	×	×	✓	Averaged TV
Gupta et al. (2024)	×	×	×	×	Mixed TV/W_2
Gaur et al. (2026)	×	✓	✓[b]	×	Stopped TV
Han et al. (2024)	×	✓	×	✓	Weighted L^2 score error
This work	✓	✓	✓	✓	**Averaged stopped TV**

Finite GD/SGD means a guarantee for a finite optimizer iterate, rather than an assumed empirical-risk minimizer. Deep all-layer training requires such an analysis beyond the two-layer setting. Derived approximation means bounds obtained from distributional or target regularity, rather than assumed realizability. Stopped TV uses a positive diffusion-time target; mixed TV/W_2 means TV proximity to a law that is W_2-close to the data law. [a]Includes Yuan et al.'s Appendix F extension under conditional-score regularity. [b]Counts Gaur et al.'s full-parameter SGD update as all-layer training.

conditional law of X_t and, for $t > 0$, also its density. We learn its score through the denoising target $f^\star(x, y, t) = \mathbb{E}[X_0 \mid X_t = x, Y = y]$. Tweedie's identity gives

$$s^\star(x, t, y) := \nabla_x \log p_t(x \mid y) = \frac{\alpha_t f^\star(x, y, t) - x}{\sigma_t^2}. \tag{1}$$

Training times and frozen data. Fix a positive stopping time t_0 and a terminal time $T \geq 1$, with $t_0 < T$. Training and sampling use the same grid with M intervals, $t_0 < t_1 < \cdots < t_M = T$. Set $\Delta_j = t_{j+1} - t_j$ and $\mathsf{T}_M = T - t_0$. We require $\Delta_j \leq \kappa_{\text{grid}} \min\{1, t_j\}$, where $0 < \kappa_{\text{grid}} \leq 1$ and $\kappa_{\text{grid}} = \widetilde{O}(M^{-1})$. Since the reverse sampler evaluates scores at the right endpoints, we draw training times from

$$\begin{aligned} c(t) &= \frac{\alpha_t^2}{\sigma_t^4}, \qquad \mathcal{W} = \sum_{j=0}^{M-1} \Delta_j c(t_{j+1}), \\ \mathbb{P}(J = j) &= \frac{\Delta_j c(t_{j+1})}{\mathcal{W}}, \qquad U = (X_{t_{J+1}}, Y, t_{J+1}) \in \mathbb{R}^{D+1}. \end{aligned} \tag{2}$$

Here J is independent of the clean pair and the forward noise. The weight $c(t)$ makes the sampler's accumulated squared score error equal to $\mathcal{W}$ times the population denoising error (Lemma C.3). For each clean pair, draw one independent index J_i and one $\varepsilon_i \sim N(0, I_{d_x})$, independently of the clean data, and form

$$U_i = \big(\alpha_{t_{J_i+1}} X_{0,i} + \sigma_{t_{J_i+1}} \varepsilon_i, Y_i, t_{J_i+1}\big).$$

These augmented observations are held fixed throughout training.

Labels and network inputs. We clip the clean labels and map the augmented inputs to a sphere with input dimension $d_0 = D + 2$. For $q = 1, \ldots, d_x$, let $Z_{B_N,q} = \text{clip}(X_{0,q}, [-B_N, B_N])$ and define

$$\widetilde{Z}_{B_N,q} = B_N^{-1} Z_{B_N,q}, \qquad \psi_\rho(u) = \frac{(u/\rho, 1)}{\sqrt{1 + \|u\|_2^2/\rho^2}} \in \mathbb{S}^{d_0 - 1}. \tag{3}$$

For a confidence parameter $\delta \in (0, 1)$, the scales $B_N, \rho_N = \Theta(\log(N/\delta) + T)$ are calibrated in Section B. Write $X_i = \psi_{\rho_N}(U_i)$ for the network inputs and $P_X = \mathcal{L}(\psi_{\rho_N}(U))$ for their population law. The corresponding labels $\widetilde{Z}_{B_N,i,q}$ lie in $[-1, 1]$. Their unnormalized regression target is $f^\star_{B_N,q}(u) = \mathbb{E}[Z_{B_N,q} \mid U = u]$. Clipping introduces an exponentially small squared regression bias (Lemma C.4); the injective map ψ_{ρ_N} provides bounded network inputs while retaining the information in U.

Network architecture and initialization. We use fully connected ReLU networks with nonincreasing hidden widths (Nguyen & Mondelli, 2020; Nguyen et al., 2021). Let $L \geq 3$ be the depth, $H = L - 1$ the number of hidden layers, $d_H = m$ the final hidden width, and $d_L = 1$ the output dimension. The widths satisfy

$$d_1 \geq \cdots \geq d_H, \qquad 1 \leq \frac{d_\ell}{d_{\ell+1}} \leq \kappa_{\mathrm{pyr},N}, \quad 1 \leq \ell < H, \tag{4}$$

where $\kappa_{\mathrm{pyr},N} \geq 1$ controls the ratios of adjacent hidden widths. With $\bar{h}_0(v) = v$ and coordinatewise ReLU $\sigma(z) = \max\{z, 0\}$, define

$$\bar{h}_\ell(v) = \sqrt{\frac{2}{d_\ell}}\,\sigma(W_\ell \bar{h}_{\ell-1}(v)), \quad 1 \leq \ell \leq H, \qquad F_\theta(v) = W_L \bar{h}_H(v), \tag{5}$$

where $W_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$ and $\theta = (W_1, \ldots, W_L)$. Each output coordinate q uses a separate network. Every entry of every matrix, including the readout, is initialized independently from $N(0,1)$, independently across networks and of the training data. For each network, we retain its initialization function and use $F^c_\theta = F_\theta - F_{\theta_0}$, with θ_0 denoting that network's initial parameters. This makes the initial prediction zero without changing the derivatives with respect to the trained parameters.

Finite gradient-descent training. For each coordinate q, full-batch gradient descent minimizes the squared loss on the frozen inputs and normalized labels. Choose a step size $0 < \eta \leq 1/(4L)$ and a stopping scale $\lambda_N > 0$, and set $R_\star = \lceil (\eta\lambda_N)^{-1} \rceil$. The loss, updates, and final denoiser are

$$\begin{aligned} \widehat{\mathcal{L}}_q(\theta) &= \frac{1}{2N} \sum_{i=1}^{N} \big(F^c_\theta(X_i) - \widetilde{Z}_{B_N,i,q}\big)^2, \\ \theta_{q,r+1} &= \theta_{q,r} - \eta \nabla \widehat{\mathcal{L}}_q(\theta_{q,r}), \qquad 0 \leq r < R_\star, \\ \widehat{f}_q(u) &= B_N F^c_{\theta_{q,R_\star}}\big(\psi_{\rho_N}(u)\big). \end{aligned} \tag{6}$$

All hidden matrices and the readout are updated. Section B prescribes the scales, stopping rule, and sufficient widths using the supplied class bounds in Section G.1, retaining depth, width ratios dependence.

Conditional sampling and target law. Assemble $\widehat{f} = (\widehat{f}_1, \ldots, \widehat{f}_{d_x})^\top$ and define the learned score by

$$\widehat{s}(x,t,y) = \frac{\alpha_t \widehat{f}(x,y,t) - x}{\sigma_t^2}.$$

Given a condition y, the reverse sampler starts from $\gamma_0 = N(0, I_{d_x})$ and uses $\widehat{s}$ on the same time grid, with the update in (75) and $\bar{s} = \widehat{s}$. Its output law is $\widehat{p}_{t_0}(\cdot \mid y)$, and the target is the forward marginal $p_{t_0}(\cdot \mid y)$. For the error decomposition, $\widetilde{p}_{t_0}(\cdot \mid y)$ denotes the output of the same learned sampler initialized from $p_T(\cdot \mid y)$ instead of γ_0. We measure sampling accuracy by $\mathbb{E}_Y \mathrm{TV}(p_{t_0}(\cdot \mid Y), \widehat{p}_{t_0}(\cdot \mid Y))$, using a fresh condition while holding the trained predictor fixed. Comparison with the original law μ_Y additionally requires controlling the smoothing bias from time 0 to t_0. The training and sampling pseudocode is given in Section G.

3 Assumptions

Our analysis uses two assumptions on the conditional data law: exponential tail control and regularity with respect to the conditioning variable. An additional density regularity assumption is used when extending the guarantee from the positive-time law $p_{t_0}(\cdot \mid y)$ to the original law $p_0(\cdot \mid y)$.

Assumption 3.1 (Conditional exponential tails)**.** *There is a family $(\mu_y)_{y \in \mathbb{R}^{d_y}}$ of probability measures that agrees with $\mathcal{L}(X_0 \mid Y = y)$ for P_Y-almost every y. For some $\lambda_X, \lambda_Y > 0$ and finite M_X, M_Y,*

$$\mathbb{E} e^{\lambda_Y \|Y\|_2} \leq M_Y, \qquad \sup_{y \in \mathbb{R}^{d_y}} \int e^{\lambda_X \|z\|_2} \mu_y(\mathrm{d}z) \leq M_X. \tag{7}$$

Uniform conditional tail control also appears in earlier analyses of conditional diffusion models. Fu et al. (2024, Assumption 3.1) impose a Gaussian upper envelope on the conditional density for compactly supported guidance; their Assumption 4.4 treats unbounded, sub-Gaussian guidance using a bounded Hölder factor multiplying a Gaussian envelope. Hu et al. (2025, Assumptions 3.1–3.2) likewise impose Gaussian conditional tails on a compact guidance domain, while Tang et al. (2025, Section 3.1) work with compact data and covariate domains. Our exponential-moment condition allows both X_0 and Y to be unbounded and accommodates exponential rather than Gaussian tails.

Assumption 3.2 (Order-k relative conditional-law regularity)**.** *Fix an integer $k \geq 1$. The family chosen in Assumption 3.1 is C^k as a map $y \mapsto \mu_y$ in the weighted total-variation norm*

$$\|\nu\|_{\mathrm{TV},1} := \int (1 + \|z\|_2)\, |\nu|(\,\mathrm{d}z),$$

where $|\nu|$ denotes the total variation measure of the signed measure ν. For every multi-index $\alpha \in \mathbb{N}_0^{d_y}$ with $1 \leq |\alpha| \leq k$, there are a measurable function S_α and constants $C_\alpha < \infty$ and $a_\alpha, b_\alpha \geq 0$ such that, for every $y \in \mathbb{R}^{d_y}$,

$$\begin{aligned} \partial_y^\alpha \mu_y(\,\mathrm{d}z) &= S_\alpha(z, y)\mu_y(\,\mathrm{d}z), \\ |S_\alpha(z, y)| &\leq C_\alpha(1 + \|y\|_2)^{a_\alpha}(1 + \|z\|_2)^{b_\alpha}. \end{aligned} \tag{8}$$

The first identity is an equality of signed measures, and the second holds μ_y-almost everywhere. The constants are independent of y and z. Write $b_k = \max_{1 \leq |\alpha| \leq k} b_\alpha$. No density of Y is required.

Regularity of the conditional distribution is also a central input in existing diffusion theory. Fu et al. (2024); Hu et al. (2025) use Hölder regularity of the conditional density or its Gaussian-envelope factor, while Tang et al. (2025) allow different smoothness orders in the data and conditioning variables. Here, regularity in y is expressed directly through the conditional law: its derivatives are absolutely continuous with respect to μ_y, with polynomially bounded relative derivatives. Gaussian noising supplies spatial smoothness at positive times.

Assumption 3.3 (Additional regularity for original-law recovery)**.** *For P_Y-almost every y, the measure μ_y has a bounded density $p_0(\cdot \mid y)$ that is β-Hölder in x for some $\beta \in (0, 1]$. Both the density bound and the Hölder constant are uniform in y.*

Density smoothness also appears in the original-law guarantees of Fu et al. (2024); Tang et al. (2025); Hu et al. (2025). In our proof, Assumption 3.3 controls the total-variation bias between $p_{t_0}(\cdot \mid y)$ and $p_0(\cdot \mid y)$ as $t_0 \downarrow 0$, allowing the positive-time guarantee to be transferred to the original conditional law in Corollary 1. The guarantee at a fixed positive stopping time uses only Assumptions 3.1 and 3.2.

Section H verifies these assumptions for explicit conditional families. In particular, the mixture of fixed Gaussian components with logistic mixing weights in Proposition H.2 satisfies Assumption 3.2 for every finite k and Assumption 3.3 with $\beta = 1$. With suitably separated components, this family includes non-log-concave conditional distributions. Section A and Table 1 compare the analyzed settings and assumptions with existing results. Section 5 explains how the training,learning, and sampling bounds combine.

4 Main Results

We now establish a conditional sampling guarantee for the network obtained after a prescribed number of gradient-descent updates. The result specifies sufficient sample size, training duration, network width, and reverse-sampling steps, without assuming access to an empirical risk minimizer. Our primary guarantee concerns the conditional law $p_{t_0}(\cdot \mid Y)$ at a fixed diffusion stopping time $t_0 > 0$. We then extend the guarantee to the original conditional law; a sharper rate under higher conditional-law regularity is stated in the appendix.

Recall that $D = d_x + d_y$. Let $k \geq 1$ be the integer regularity order in Assumption 3.2, and define

$$\begin{aligned} s &= \frac{D+2}{2}, \qquad \gamma = \frac{D+1}{D+2}, \qquad \bar{k} = \min\{k, s\}, \\ r &= \frac{\bar{k}}{\bar{k} + \gamma s}, \qquad \ell = \frac{s}{\bar{k} + \gamma s}. \end{aligned} \tag{9}$$

The exponents r and ℓ determine the statistical rate and training horizon, respectively, and satisfy $r + \gamma\ell = 1$. Throughout, $\widetilde{O}$ suppresses log factors; dependence on t_0^{-1} is retained when t_0 varies.

Theorem 1 (Finite-training conditional diffusion guarantee)**.** *Suppose Assumptions 3.1 and 3.2 hold at integer order k. Fix depth $L \geq 3$ and diffusion stopping time $t_0 > 0$, and supply valid numerical class bounds as specified in Section G.1. Run Algorithm 1 with the centered ReLU networks, independent Gaussian initialization, and frozen score-weighted augmentation from Section 2, training all layers by full-batch gradient descent (6) with $0 < \eta \leq 1/(4L)$.*

For sufficiently small $\varepsilon > 0$ and $0 < \delta < \delta_0$, there is a deterministic sufficient clean-sample threshold

$$N_{\mathrm{cal}}(\varepsilon, \delta) = \widetilde{O}\Big(\varepsilon^{-2/r}\Big) = \widetilde{O}\Big(\varepsilon^{-\left[2+(D+1)/\bar{k}\right]}\Big). \tag{10}$$

For $N \geq N_{\mathrm{cal}}$, choose the clipping, localization, and stopping scales according to Section B, using the constrained stopping scale λ_N in (44). Train each coordinate network for

$$R_\star = \lceil (\eta\lambda_N)^{-1} \rceil$$

updates. Choose hidden widths satisfying (4) and the sufficient width condition (59); Lemma B.5 gives a finite choice satisfying that condition. Choose a terminal time $T > t_0$ of order $\log(1/\varepsilon)$ and an admissible grid with $M = \widetilde{\Theta}(\varepsilon^{-2})$ reverse steps.

Then, with probability at least $1 - C\delta$ over the clean training pairs, frozen augmentation, and initialization, the reverse sampler satisfies

$$\mathbb{E}_Y \mathrm{TV}(p_{t_0}(\cdot \mid Y), \widehat{p}_{t_0}(\cdot \mid Y)) \leq \widetilde{O}(\varepsilon). \tag{11}$$

Here Y is a fresh conditioning variable and the trained predictor is held fixed. At fixed depth, dimensions, adjacent-width ratios, and t_0, the prescribed training horizon and a sufficient choice of final hidden width obey

$$R_\star = \widetilde{O}\big(\eta^{-1} N^{\ell}\big), \qquad m = \widetilde{O}\big(N^{13\ell+3r}\big). \tag{12}$$

The theorem connects the output of finite neural training to conditional sampling accuracy. Its sample exponent decreases as the available conditional-law regularity increases, until the effective order reaches $\bar{k} = s$. The probability statement concerns the randomness used to construct and train the predictor; the TV error is then averaged over an independent condition drawn from P_Y. The width and training scales in (12) quantify sufficient resources for this guarantee.

If $k \geq \lceil (D+2)/2 \rceil$, then $\bar{k} = s$ and $N = \widetilde{O}(\varepsilon^{-[4-2/(D+2)]})$ at fixed t_0; Corollary B.7 gives the full statement and width bound. The preceding results keep the diffusion stopping time fixed. To recover the original conditional law, we let t_0 decrease with the target accuracy and use Assumption 3.3 to control the smoothing bias.

Corollary 1 (Recovery of the original conditional law)**.** *Suppose Assumptions 3.1 to 3.3 hold, and use the architecture and training procedure of Theorem 1, with the supplied class bounds specified in Section G.1. Put $b = \bar{k}/s$, $j = \lceil \bar{k} \rceil$, and $\chi = j \max\{4, b_j\}$, where $b_j = \max_{1 \leq |\alpha| \leq j} b_\alpha$ in Assumption 3.2. A sufficient deterministic clean-sample threshold is $N_{\mathrm{cal}}^{(0)}(\varepsilon, \delta) = \widetilde{O}(\varepsilon^{-p})$, where*

$$\begin{aligned} p &= \max\{p_{\mathrm{stat}}, p_{\mathrm{gate}}\}, \\ p_{\mathrm{stat}} &= 2\left(1 + \frac{\gamma}{b}\right)\left(1 + \frac{1}{\beta}\right) + \frac{4\gamma\chi}{\beta b}, \qquad p_{\mathrm{gate}} = \frac{2}{b}\left(1 + \frac{1 + 2\chi}{\beta}\right). \end{aligned} \tag{13}$$

For $N \geq N_{\mathrm{cal}}^{(0)}$, choose

$$t_0 = \widetilde{\Theta}(\varepsilon^{2/\beta}), \qquad T = \Theta(\log(1/\varepsilon)), \qquad M = \widetilde{\Theta}(\varepsilon^{-2}).$$

Use the constrained stopping scale (44) and train for $R_\star = \lceil (\eta\lambda_N)^{-1} \rceil$ updates. Choose $B_N \gtrsim \log(1/(\varepsilon t_0))$ and calibrate localization and width according to Theorem B.9, using this shrinking t_0 and the resulting training horizon. Then, with probability at least $1 - C\delta$ over the clean training pairs, frozen augmentation, and initialization,

$$\mathbb{E}_Y \mathrm{TV}(p_0(\cdot \mid Y), \widehat{p}_{t_0}(\cdot \mid Y)) \leq \widetilde{O}(\varepsilon),$$

where Y is fresh and the trained predictor is held fixed.

The choice of t_0 makes the smoothing bias $\widetilde{O}(t_0^{\beta/2}) = \widetilde{O}(\varepsilon)$. The two sample exponents account for statistical learning and the population-comparison floor, respectively; both retain the small-time growth of the target regularity, score-error conversion. Theorem B.9 gives the detailed proof.

5 Proof Overview: From Finite Training to Conditional Sampling

The proof links conditional sampling accuracy to the population error of the finite trained network. Two steps are central: controlling all-layer training at fresh inputs, and deriving approximation and statistical bounds for the discrete-time training distribution. All population norms below use fresh inputs with the trained predictor held fixed.

5.1 From sampling error to population prediction error

We first reduce conditional sampling error to population prediction errors. Write $P_U = \mathcal{L}(U)$ for the augmented training law and recall the denoising target $f^\star(u) = \mathbb{E}[X_0 \mid U = u]$. For $s_f(x,t,y) = (\alpha_t f(x,y,t) - x)/\sigma_t^2$, the score-weighted time distribution in (2) gives

$$\sum_{j=0}^{M-1} \Delta_j \mathbb{E}\,\|s_f - s^\star\|_{t_{j+1}}^2 = \mathcal{W}\,\|f - f^\star\|_{L^2(P_U)}^2, \tag{14}$$

where each score expectation is evaluated at $(X_{t_{j+1}}, t_{j+1}, Y)$. Thus the regression loss uses precisely the time weighting required by the sampling bound.

Let $g_{q,j}^K$ and $g_{q,j}^\infty$ be zero-start kernel GD on the same frozen inputs and normalized labels, using the initialization kernel and the limiting full NTK, respectively. We separate the neural-to-kernel discrepancy, finite-to-limiting-kernel discrepancy, and limiting-kernel prediction error as

$$\mathcal{E}_{\text{opt}} := B_N^2 \sum_{q=1}^{d_x} \left\| F_{\theta_{q,R_\star}^c}^c - g_{q,R_\star}^K \right\|_{L^2(P_X)}^2, \mathcal{E}_{\text{width}} \quad := B_N^2 \sum_{q=1}^{d_x} \left\| g_{q,R_\star}^K - g_{q,R_\star}^\infty \right\|_{L^2(P_X)}^2,$$

$$\mathcal{E}_{\text{ker}} := \sum_{q=1}^{d_x} \left\| B_N g_{q,R_\star}^\infty \circ \psi_{\rho_N} - f_{B_N,q}^\star \right\|_{L^2(P_U)}^2.$$

Here $B_N g_{q,R_\star}^\infty \circ \psi_{\rho_N}$ is the qth limiting-kernel denoising predictor after $R_\star$ GD updates, with ψ_{ρ_N} mapping the augmented input to the sphere and B_N undoing label normalization. Clipping contributes $\mathcal{E}_{\text{clip}} = e^{-cB_N}$. Combining the squared triangle inequality with the score-to-sampling bound yields

$$\boxed{\mathbb{E}_Y \mathrm{TV}(p_{t_0}, \widehat{p}_{t_0}) \le \mathcal{E}_{\text{term}} + \mathcal{E}_{\text{disc}} + C\sqrt{\mathcal{W}(\mathcal{E}_{\text{clip}} + \mathcal{E}_{\text{opt}} + \mathcal{E}_{\text{width}} + \mathcal{E}_{\text{ker}})}.} \tag{15}$$

The error terms on the right are defined and the bound (15) is derived in Section B, sub-sections B.1-B.3. Here both laws are conditional on the fresh variable Y; $\mathcal{E}_{\text{term}}$ and $\mathcal{E}_{\text{disc}}$ account for Gaussian initialization of the sampler and time discretization. $C > 0$ depends only on the fixed tail parameters in Assumption 3.1 and is independent of N, m, $R_\star$, M, T, and t_0. Short OU and exponential-tail arguments bound $\mathcal{E}_{\text{term}}$ and $\mathcal{E}_{\text{clip}}$, respectively (Lemmas C.1 and C.4). For $\mathcal{E}_{\text{disc}}$ and the score-to-sampling reduction, we apply Benton et al. (2024) conditionally and average over Y (Lemmas C.2 and C.3). The remaining terms require the population bounds developed below: Section 5.2 controls $\mathcal{E}_{\text{opt}}$ and $\mathcal{E}_{\text{width}}$, and Section 5.3 controls $\mathcal{E}_{\text{ker}}$.

5.2 Controlling finite neural training at fresh inputs

The error $\mathcal{E}_{\text{opt}}$ measures the population prediction discrepancy between the trained neural network and gradient descent with its fixed initialization kernel. The error $\mathcal{E}_{\text{width}}$ measures the corresponding discrepancy between this finite-width kernel predictor and the limiting full-NTK predictor; both errors are evaluated at fresh inputs. Movement and perturbation bounds control the changing network, while full-NTK concentration controls the finite-kernel discrepancy. Duhamel expansions track how these perturbations accumulate during training and affect predictions at fresh inputs.

Controlling the changing network. Training changes every hidden matrix, the ReLU activation patterns, and the Gaussian-initialized readout. Our perturbation bounds control these changes jointly for the unequal-width architecture, uniformly over inputs and activation changes within the prescribed movement ball. They therefore apply to the patterns selected during training. A simultaneous movement and residual argument shows that the iterates remain within distance $O(\sqrt{1+h})$ of

initialization and that the normalized training residuals remain bounded, where $h = \eta R_\star$. Substituting this movement radius into the local estimates gives the perturbation scale

$$q_\star = C_{L,\kappa_{\mathrm{pyr,N}}} \sqrt{\nu_m} (1+h)^{1/6} m^{-1/6},$$

where ν_m is the logarithmic factor in (219). The movement argument ensures that these local estimates remain valid at every update up to $R_\star$.

Accumulating perturbations on the sample. We compare the centered network with zero-start gradient descent under its initialization kernel, using the same frozen sample. Their predictions initially agree. Each neural update introduces perturbations through changes in the tangent kernel and the nonlinear remainder of the prediction update. The discrete Duhamel identity expresses the difference between the two training trajectories as the sum of these perturbations, each propagated through subsequent fixed-kernel updates. Under the prescribed step size, these fixed-kernel updates do not amplify empirical errors. The movement and residual bounds above therefore control the accumulated sample discrepancy by $O(q_\star h)$ over the effective training horizon h.

Transferring the comparison to fresh inputs. To control population prediction error, we also track how these sample discrepancies affect predictions at inputs unseen during training. At a fresh input, the difference between the two prediction updates includes an average of the training-sample discrepancies, weighted by similarities under the initialization kernel. Our kernel bounds control this averaged cross-kernel map independently of the sample size. Accumulating the update differences gives two contributions: direct changes in the kernel similarities and nonlinear remainders contribute $O(q_\star h)$, while the sample discrepancies accumulated earlier feed into later prediction updates and contribute at most $O(q_\star h^2)$. These two effects explain the $h + h^2$ dependence in the following population comparison. Section F.1 explains the identity and how it yields this trajectory comparison.

Proposition 5.1 (Finite-horizon population comparison)**.** *Under the architecture, step-size, and width calibration of Theorem 1 and Section B, conditional on the clean pairs and frozen augmentation, with probability at least $1-\delta$ over initialization, simultaneously for all $q \le d_x$,*

$$\sup_{0 \le j \le R_\star} \left\| F^c_{\theta_{q,j}} - g^K_{q,j} \right\|_{L^2(P_X)} \le C_{L,\kappa_{\mathrm{pyr,N}}} \sqrt{\nu_m} (1+h)^{1/6} (h+h^2) m^{-1/6}. \tag{16}$$

Let $\mathfrak{r}_N$ denote the deterministic squared population-risk envelope in (46), which satisfies $\mathfrak{r}_N = \widetilde{O}(N^{-r})$ under the fixed-t_0 calibration. The width calibration makes the displayed bound at most $C\mathfrak{r}_N^{1/2}/(B_N\sqrt{d_x})$, hence $\mathcal{E}_{\mathrm{opt}} \le C\mathfrak{r}_N$. This comparison uses no lower Gram-eigenvalue bound or PL inequality. The proof and the resulting bound on $\mathcal{E}_{\mathrm{opt}}$ are given in Section C.7.

Passing to the limiting kernel. We next compare the two kernel GD trajectories. Uniform concentration of the full initialization NTK, including the readout contribution, controls both their training updates and their cross-kernels at fresh inputs. A second Duhamel comparison then gives $\mathcal{E}_{\mathrm{width}} \le C\mathfrak{r}_N$ under the same width calibration, as proved in Section C.6. Comparing the trajectories transfers their predictions without requiring a limiting-kernel approximating function to belong to the finite-rank RKHS. The required arguments are developed in Section F.

5.3 Learning under the actual diffusion training distribution

The error $\mathcal{E}_{\mathrm{ker}}$ measures the squared population prediction error of the rescaled limiting full-NTK predictor relative to the clipped denoising target at fresh inputs. We bound $\mathcal{E}_{\mathrm{ker}}$ by combining approximation derived from the conditional law with capacity and early-stopping estimates under the actual input law P_X. Because training times are discrete, this law is supported on time slices of zero spherical surface measure. Approximation error averaged under uniform spherical measure alone therefore does not control the error on the training distribution. Throughout this subsection, the kernel is the limiting full NTK K_L^{full} of the specified ReLU network on sphericalized inputs, including contributions from every hidden layer and the readout.

From conditional-law regularity to bounded approximants. We differentiate the Gaussian observation kernel to handle derivatives in the noisy input and time, and use the relative conditional-law assumption to handle derivatives in the conditioning variable. This gives explicit bounds

on joint derivatives of the clipped denoising target up to order k on bounded input regions with $t \geq t_0$. We extend the normalized target to the sphere and use the identification $\mathcal{H}_L^{\text{full}} \simeq H^s(\mathbb{S}^{D+1})$, with $s = (D+2)/2$, to construct RKHS approximants. When $\bar{k} < s$, smoothing cancels lower-order smoothing bias and gives uniform approximation, with ranges bounded independently of the smoothing scale. When $\bar{k} = s$, the target extension itself is an admissible RKHS comparator. Uniform approximation controls error on the actual time slices, and the range bound controls statistical fluctuations. In both regimes, error outside the localization region is retained as $\tau_N = \mathbb{P}(\|U\|_2 > \rho_N)$.

Capacity for the population and empirical designs. The bounded RKHS approximants constructed in the preceding paragraph provide the comparators for the bias term $A_{\bar{k},N}(t_0)^2\lambda_N^b$ in (18) ($A_{\bar{k},N}(t_0)$ is defined in Equation (148)). To obtain the statistical term, we also need to control the kernel's effective dimension: the number of feature directions resolved at scale λ. The early-stopping argument requires this controlfor both the population law P_X and the empirical training law.Using harmonic projection and a uniform tail bound (Vakili et al., 2021b;a), projection onto degrees at most J has rank $O(J^{D+1})$, while the remaining squared feature norm is uniformly $O(J^{-1})$. Consequently, for every probability law P on the sphere,

$$\mathsf{df}_P(\lambda) \leq C\left(J^{D+1} + \frac{1}{\lambda J}\right) \lesssim \lambda^{-\gamma}, \qquad \gamma = \frac{D+1}{D+2}, \tag{17}$$

where $0 < \lambda \leq 1$ and $J \asymp \lambda^{-1/(D+2)}$. This supplies the factor $1/(N\lambda_N^\gamma)$ in the statistical term of (18), including for our discrete-time design.

Let $b = \bar{k}/s$, and $A_{\bar{k},N}(t_0)$ bound the target-extension norms, and put $\mathcal{L}_N = \log(CNd_x/\delta)$. The bounded-comparator early-stopping bound and regularized population–empirical operator comparison yield, with high probability over the clean pairs and augmentation,

$$\mathcal{E}_{\text{ker}} \leq Cd_xB_N^2\left[A^2\lambda_N^b + \frac{\mathcal{L}_N}{N\lambda_N^\gamma} + \tau_N\right], \tag{18}$$

provided $N\lambda_N \geq C\mathcal{L}_N$. The three terms quantify approximation and stopping bias, statistical fluctuation, and localization error, respectively. The capacity bound (17) is proved in Section E.4, and the learning-error bound (18) is derived in Section C.5.

5.4 Choosing the resources and completing the sampling bound

Fix t_0, depth, dimensions, and adjacent-width ratios. Under the logarithmic localization $\rho_N = \Theta(\log(N/\delta) + T)$ specified in Section B, A grows at most polylogarithmically with N. The constrained stopping rule balances the first two terms in (18) while retaining $\lambda_N \geq C\mathcal{L}_N/N$. After controlling clipping and localization, it gives $\mathcal{E}_{\text{ker}} = \widetilde{O}(N^{-r})$ and $\eta R_\star = \widetilde{O}(N^\ell)$. The width calibration makes both trajectory discrepancies no larger than this statistical scale, with sufficient width $m = \widetilde{O}(N^{13\ell+3r})$. For the remaining sampling terms, let $M_2 = \sup_y \int \|z\|_2^2 \mu_y(\mathrm{d}z) < \infty$. OU convergence and the conditional reverse-process bound of Benton et al. (2024) give

$$\mathcal{E}_{\text{term}} \leq Ce^{-T}\sqrt{d_x + M_2}, \qquad \mathcal{E}_{\text{disc}} \leq \widetilde{O}\left(\sqrt{\frac{d_x + M_2}{M}}\right) \tag{19}$$

on the admissible grid. These bounds and their conditional hypotheses are established in Lemmas C.1 and C.2. Since $\mathcal{W} \lesssim t_0^{-1}$, the learning contribution to (15) is $\widetilde{O}(N^{-r/2})$ at fixed t_0. Taking $N \geq N_{\text{cal}} = \widetilde{O}(\varepsilon^{-2/r})$, $T = \Theta(\log(1/\varepsilon))$, and $M = \widetilde{\Theta}(\varepsilon^{-2})$ proves Theorem 1 on the joint learning and initialization event. For original-law recovery, the smoothing-bias bound in Lemma C.13 supplies the additional error under Assumption 3.3. The resources are then recalibrated while retaining their dependence on the shrinking t_0. The complete assembly of these bounds into the theorem and corollary proofs is given in Section B.6. A detailed roadmap of the proof structure is provided in the appendix contents and overview.

6 Conclusion

We establish conditional diffusion guarantees for finitely trained deep ReLU networks without ERM oracles. Duhamel comparisons and approximation from the conditional law give explicit sample, training, width, and sampling requirements.

AI USE STATEMENT

Generative AI tools were used for writing assistance. The authors verified all mathematical statements, proofs, citations, and experimental results, and take full responsibility for the content of the submission.

REFERENCES

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 242–252. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/allen-zhu19a.html.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *Advances in neural information processing systems*, 32, 2019.

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.

Alberto Bietti and Francis Bach. Deep equals shallow for relu networks in kernel regimes. *arXiv preprint arXiv:2009.14397*, 2020.

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 Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=zyLVMgsZ0U_.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1eK3i09YQ.

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.

Khashayar Gatmiry, Jonathan Kelner, and Holden Lee. Learning mixtures of gaussians using diffusion models. In *Proceedings of the 38th Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 2403–2456, 2025.

Mudit Gaur, Prashant Trivedi, Sasidhar Kunapuli, Amrit Singh Bedi, and Vaneet Aggarwal. Improved sample complexity bounds for diffusion model training without empirical risk minimizer access. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=CFdNqqlqOv.

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.

Yinbin Han, Meisam Razaviyayn, and Renyuan Xu. Neural network-based score estimation in diffusion models: Optimization and generalization. In *International Conference on Learning Representations*, volume 2024, pp. 42520–42558, 2024.

Yihan He, Qishuo Yin, Yuan Cao, Jianqing Fan, and Han Liu. A theory on flow matching with neural networks. *arXiv preprint arXiv:2606.10089*, 2026.

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

Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17(none):1 – 6, 2012. doi: 10.1214/ECP.v17-2079. URL https://doi.org/10.1214/ECP.v17-2079.

Jerry Yao-Chieh Hu, Weimin Wu, Yi-Chen Lee, Yu-Chao Huang, Minshuo Chen, and Han Liu. On statistical rates of conditional diffusion transformers: Approximation, estimation and minimax optimality. In *International Conference on Learning Representations*, volume 2025, pp. 12591–12675, 2025.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HygegyrYwH.

Kedar Karhadkar, Michael Murray, and Guido Montúfar. Bounds for the smallest eigenvalue of the ntk for arbitrary spherical data of arbitrary dimension. *Advances in Neural Information Processing Systems*, 37:138197–138249, 2024.

Ilja Kuzborskij and Csaba Szepesvári. Learning lipschitz functions by gd-trained shallow overparameterized relu neural networks. *arXiv preprint arXiv:2212.13848*, 2022.

Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards non-asymptotic convergence for diffusion-based generative models. In *International Conference on Learning Representations*, volume 2024, pp. 51686–51697, 2024.

Quynh Nguyen, Marco Mondelli, and Guido F Montufar. Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep relu networks. In *International Conference on Machine Learning*, pp. 8119–8129. PMLR, 2021.

Quynh N Nguyen and Marco Mondelli. Global convergence of deep networks with one wide layer followed by pyramidal topology. *Advances in Neural Information Processing Systems*, 33: 11961–11972, 2020.

Atsushi Nitanda and Taiji Suzuki. Optimal rates for averaged stochastic gradient descent under neural tangent kernel regime. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PULSD5qI2N1.

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.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *The Journal of Machine Learning Research*, 15(1): 335–366, 2014.

Kulin Shah, Sitan Chen, and Adam Klivans. Learning mixtures of gaussians using the ddpm objective. *Advances in Neural Information Processing Systems*, 36:19636–19649, 2023.

Rong Tang, Lizhen Lin, and Yun Yang. Conditional diffusion models are minimax-optimal and manifold-adaptive for conditional distribution estimation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=NltQraRnbW.

Hans Triebel. *Theory of function spaces III*. Springer, 2006.

Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and trends® in machine learning*, 8(1-2):1–230, 2015.

Sattar Vakili, Michael Bromberg, Jezabel Garcia, Da-shan Shiu, and Alberto Bernacchia. Uniform generalization bounds for overparameterized neural networks. *arXiv preprint arXiv:2109.06099*, 2021a.

Sattar Vakili, Kia Khezeli, and Victor Picheny. On information gain and regret bounds in gaussian process bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 82–90. PMLR, 2021b.

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.

Hui Yuan, Kaixuan Huang, Chengzhuo Ni, Minshuo Chen, and Mengdi Wang. Reward-directed conditional diffusion: Provable distribution estimation and reward improvement. *Advances in Neural Information Processing Systems*, 36:60599–60635, 2023.

Junyu Zhou, Puyu Wang, Yunwen Lei, Yiming Ying, and Ding-Xuan Zhou. Optimal rates for generalization of gradient descent methods with deep neural networks. *arXiv preprint arXiv:2606.06764*, 2026.

APPENDIX CONTENTS AND OVERVIEW

The appendix begins with related work, then presents the proofs of the main guarantees, and ends with supplementary procedures, examples, and reference material. Section titles and page numbers below are linked.

	Section	Page
Related work		
A	Related work and detailed comparisons	14
Proofs of the main guarantees		
B	Main results and parameter calibration	17
C	Component-error bounds	26
D	Conditional-law regularity and regression identities	32
E	Kernel approximation and population learning	35
F	Finite neural training and kernel comparisons	50
F.1	A guide to the Duhamel comparison	50
Supplementary material		
G	Complete training and conditional sampling procedure	63
H	Examples and non-examples of conditional laws	65
I	Parameter and computation costs	68
J	Optional geometry and curvature certificates	68
K	Notation	73
L	Assumption, architecture, and proof-dependency ledger	74

Reading the proofs. The proof is organized around the sampling-error decomposition in Section B. It first converts conditional sampling error into population denoising error, together with terminal-initialization, discretization, and clipping errors. Inserting the initialization-kernel and limiting-kernel predictors then separates the learning contribution into three terms: the neural-to-kernel discrepancy $\mathcal{E}_{\mathrm{opt}}$, the finite-to-limiting-kernel discrepancy $\mathcal{E}_{\mathrm{width}}$, and the limiting-kernel prediction error $\mathcal{E}_{\mathrm{ker}}$. The component bounds are established in Section C, using the supporting arguments described below.

The statistical argument begins in Section D, which derives regularity of the clipped denoising target from the conditional-law assumptions. Section E uses this regularity to construct bounded RKHS approximants, bounds the effective dimension under the actual population and empirical input laws, and proves the early-stopping oracle. These ingredients give the bound on $\mathcal{E}_{\mathrm{ker}}$ in Section C.5. The bound retains approximation and stopping bias, statistical fluctuation, and localization error, both below saturation and when the target extension itself is an RKHS comparator.

The finite-training argument is developed in Section F. Local perturbation estimates bound changes in the tangent kernel and nonlinear prediction remainders. A movement argument ensures that these estimates remain valid throughout training. The Duhamel comparison then accumulates the perturbations on the sample and transfers their effects to fresh inputs. This gives the neural-to-kernel bound in Section C.7, which proves Proposition 5.1 and controls $\mathcal{E}_{\mathrm{opt}}$. Full-NTK concentration and a second trajectory comparison give the bound on $\mathcal{E}_{\mathrm{width}}$ in Section C.6.

The stopping and width calibrations make all three learning errors bounded by a constant multiple of the squared-risk scale $\mathfrak{r}_N$. Combining them with the terminal-initialization, discretization, and clipping bounds gives the stopped-law sampling guarantee. For original-law recovery, an additional smoothing-bias bound is required, and the resources are recalibrated to retain the dependence on the shrinking diffusion stopping time t_0. The complete assembly and proofs of Theorem 1 and Corollary 1, together with the higher-regularity corollary, are given in Section B.6.

Supplementary material. The complete procedure and its supplied class bounds, illustrative conditional families, and computation costs follow the proofs. The geometry results in Section J use

additional conditions and are not invoked by the main sampling guarantees. The notation table and dependency ledger provide reference summaries.

A Related Work and Detailed Comparisons

A.1 Diffusion guarantees and scope of comparisons

Unconditional diffusion. Sampling analyses connect score error to distributional accuracy (Chen et al., 2023b; Benton et al., 2024; Li et al., 2024). Statistical results cover nearly minimax distribution estimation under Besov regularity (Oko et al., 2023), recovery on low-dimensional linear subspaces (Chen et al., 2023a), and empirical Bayes score estimation (Wibisono et al., 2024). Gupta et al. (2024) obtain sample-complexity bounds polynomial in dimension and network size under sufficient score representability and empirical risk minimization. Results on learning Gaussian mixtures with score-based objectives (Shah et al., 2023; Gatmiry et al., 2025) address particular structured distribution models and are complementary to our conditional-law setting.

Conditional diffusion. Conditional diffusion theory covers reward-directed generation (Yuan et al., 2023), classifier-free conditional estimation (Fu et al., 2024), minimax estimation and manifold adaptation (Tang et al., 2025), and transformer score models (Hu et al., 2025). The compared results analyze ERM-defined neural estimators; our guarantee concerns the output of finitely many gradient steps that train every network layer. An ERM theorem does not itself impose lazy training.

Finite-training guarantees for diffusion. Optimization-aware analyses explicitly connect neural training to score or sampling accuracy. Gaur et al. (2026) analyze stochastic gradient descent for smooth multilayer score networks under a population Polyak–Łojasiewicz condition; their total-variation guarantee at positive diffusion stopping time retains an approximation remainder. Han et al. (2024) analyze gradient descent and an explicit early-stopping rule for two-layer ReLU score networks with a fixed readout. Their theorem controls a weighted squared score error and uses an empirical Gram eigenvalue lower bound. These works establish finite-optimizer guarantees for unconditional diffusion. Our result combines external conditioning with finite training of every layer of a deep ReLU denoiser and an averaged conditional total-variation guarantee.

Position of this work. Theorem 1 gives an averaged conditional stopped-TV guarantee for the finite neural iterate without an ERM oracle, PL condition, or assumed positive empirical Gram lower bound. Table 1 summarizes the compared regimes; Section A.2 and Table 2 detail their differing assumptions, metrics, and sampling targets.

A.2 Detailed assumptions and theorem comparisons

Table 2 expands the compact main comparison. The cited 2026 Han revision supersedes the assumed stopping-error condition in the 2024 conference analysis (Han et al., 2024). We compare the explicit rule in its Theorem 3.12 and score theorem 3.13. The latter controls a weighted squared score error, so it is not directly a conditional TV rate.

Comparison of conditional-tail assumptions. The uniform moment condition is a restriction on the whole conditional family, not only its marginal distribution. It is comparable to established uniform tail assumptions: Assumption 3.1 of Fu et al. (2024) requires common constants $C_1, C_2 > 0$ with $p_0(x \mid y) \le C_1 \exp(-C_2 \|x\|_2^2/2)$ for every $y \in [0,1]^{d_y}$. Their Assumption 4.4 gives an unbounded-conditioning analogue through a uniformly controlled Hölder multiplier of a fixed Gaussian envelope. Assumption 3.2 of Hu et al. (2025) uses the same Gaussian factorization on a bounded conditioning domain; their Assumption 3.1 also imposes conditional Gaussian tails. The bounded-domain setting of Tang et al. (2025) supplies a different, compact-support comparison. On any common conditioning domain, a uniform envelope $p_0(x \mid y) \le Ae^{-a\|x\|_2^2}$ implies

$$\sup_y \mathbb{E}[e^{\lambda\|X_0\|_2} \mid Y = y] \le A \int_{\mathbb{R}^{d_x}} e^{\lambda\|x\|_2 - a\|x\|_2^2}\, dx < \infty.$$

Thus our tail requirement permits slower exponential decay on that domain. This comparison concerns tails only: it does not order the complete assumption classes, their conditioning domains, or their

Table 2: Assumptions and approximation treatment in the compared results. Equation and theorem numbers in each row belong to the cited work.

Work / locator	Distributional or optimization inputs	Approximation and width interpretation
Yuan et al. (2023), (3.2), (3.8), Assumption 4.3, Lemma D.3	Sub-Gaussian latent tail envelope; realizable score class in the stated regime; reward and subspace hypotheses.	ERM over an encoder–decoder class. Latent-space TV enters the reward analysis; no finite-GD width calibration.
Fu et al. (2024), (2.6)–(2.7), Theorems 3.2, 3.4, 4.2	Conditional density Hölder regularity and Gaussian tail envelope; extra KL control in the weaker regime.	ReLU approximation is calibrated through network class size. Original-law TV is averaged over data and conditions; no optimizer theorem.
Tang et al. (2025), (8), Section 2.3, Theorems 1–2	Bounded domains, density/covariate regularity; additional geometric conditions for manifold adaptation.	Time-dependent ReLU class sizes control approximation. Density TV and manifold Wasserstein regimes differ; no optimizer theorem.
Hu et al. (2025), (2.1), Assumptions 3.1–3.2, Theorems 3.1–3.4	Hölder density with Gaussian envelope; stronger density factorization in the improved regime.	Transformer approximation and covering bounds calibrate the class and averaged TV. The output map is part of the class; no finite-iterate training analysis.
Gaur et al. (2026), Assumptions 1–4, Theorem 1	Finite second moment, population-loss PL, loss smoothness, and bounded gradient variance.	The TV theorem retains ϵ_{approx}. Width W and depth enter the per-time sample condition; no approximation-to-width calibration is supplied there.
Han et al. (2024), Assumptions 3.2, 3.5, 3.6, 3.8; Theorem 3.13	Bounded data, regular forward coefficient, Lipschitz posterior mean, and a lower eigenvalue bound on the empirical Gram matrix.	RKHS approximation is balanced using an explicit stopping rule; fixed readout. The final theorem states a sufficient width, given below.
He et al. (2026), Lemma 3.1, Assumptions 1–4, Theorems 3.2–3.5	Anti-concentration, path regularity, Gram positivity, and the prescribed conditional base coupling. Sampling additionally uses bounded support.	Both layers train. The structured velocity target is controlled in Theorem 3.3; Theorem 3.5 gives W_1. Width and step size depend on the Gram spectrum.
This work, Theorem 1	Uniform conditional exponential moments for a smooth law version and relative law derivatives. No assumed empirical Gram lower bound.	Population comparator error is calibrated with early stopping; all layers, including the Gaussian readout, train. Corollary B.4 gives sufficient $m = \widetilde{O}(N^{13\ell+3r})$.

He et al. condition their velocity field on the sampled base variable; the target law has no external covariate. Their learning analysis allows sub-Gaussian data, while the sampling theorem uses bounded support. Their Gram positivity is supported by an auxiliary lemma; width and step size retain its minimum eigenvalue.

smoothness requirements. In particular, an all-y bound excludes unbounded conditional location families such as $X_0 = Y + Z$ with independent standard Gaussian Y, Z, since the conditional exponential moment is at least $e^{\lambda|y|}$.

Sufficient width is not a common complexity scale. In their notation, Han et al.'s Theorem 3.13 states $m \gtrsim \mathrm{poly}(d)N^{26+4/(d+3)}$. He et al.'s Theorem 3.2 uses

$$m \gtrsim \frac{n^5\|\boldsymbol{U}^{(0)} - \boldsymbol{V}\|_2^2\, d\log(nd)}{\kappa^2\lambda_{\min}(\boldsymbol{H}^{(\infty)})\delta^3},$$

with further calibration in Theorem 3.3; their κ is an initialization scale. These sufficient bounds have different horizons, metrics, and retained factors and are not lower bounds or a ranking of practical sizes. Gaur et al.'s Theorem 1 counts samples at each training time, retaining a W^{2D} factor in its own depth notation. An uncontrolled approximation remainder cannot be read as a complete arbitrary-accuracy guarantee.

A.3 Sources of the training and population-learning arguments

Conditional-law and sampling interfaces. Relative-law assumptions yield regularity of the clipped posterior mean and localized RKHS comparators. Population transfer compares the neural iterate with initialization-kernel GD and then limiting full-NTK GD, controlling fresh-input predictions and every trained block, including the readout. The score-weighted right-endpoint design identifies the sampler's integrated score loss with regression risk under the frozen At fixed t_0, the sufficient clean-sample exponent is $2+(D+1)/\bar{k}$, where $\bar{k} = \min\{k, (D+2)/2\}$. This equals $2+(D+1)/k$ below the RKHS threshold and $4 - 2/(D+2)$ when $k \geq \lceil (D+2)/2 \rceil$. These are algorithmic upper bounds, not minimax assertions.

Deep-network trajectory comparison. Zhou et al. (2026), version 1, already establish the population comparison through neural, linearized, initialization-kernel, and limiting kernel iterates (equation (9), Propositions 1–3); their Theorems 5 and 10 cover GD and SGD under capacity and source assumptions. Their main architecture and initialization are specified in equations (1)–(2), with equal hidden widths, a paired last hidden layer, and a fixed signed readout. Section 5, particularly Lemma 32, additionally proves kernel concentration for nonsymmetric initialization with a fixed signed readout. Consequently, the generic comparison principle and nonsymmetric concentration are prior methodology. Our required extensions concern the jointly independent Gaussian, trainable-readout, unequal-width architecture: Theorem F.8 controls moving gates and readout, Theorem F.10 controls the full initialization kernel, and Proposition F.13 and Lemma F.11 give the finite-horizon population transfer. Source correspondence is recorded in Sections F.3 and F.4. We fix depth and allow its constants to grow; polynomial-in-depth resource control is not asserted.

Uniform tails and the effective dimension. The finite-dimensional projection method of Vakili et al. (2021b), Theorem 3, controls kernel information gain by a rank contribution and a uniform residual-kernel bound. For neural kernels, Vakili et al. (2021a), Theorem 2 and Appendix G (Lemma 4), use the spherical-harmonic addition theorem to bound that residual uniformly, despite the absence of a uniform bound on the individual harmonics. Specializing their activation order to ReLU gives the same $O(J^{-1})$ high-degree tail used here. We claim the needed trace consequence in Proposition E.19, rather than a new harmonic principle: rank–tail compression gives

$$\mathrm{tr}[T_P(T_P + \lambda I)^{-1}] \leq C\{(1+J)^{D+1} + \lambda^{-1}(1+J)^{-1}\}$$

for every probability law P, and choosing $J \asymp \lambda^{-1/(D+2)}$ yields the capacity exponent in the early-stopping theorem. Applying the same result to the empirical law bounds the realized noise complexity. The distinction matters because the discrete score-weighted time design may be singular relative to spherical surface measure.

Other optimization and sampling tools. The NTK and lazy-training literature supplies exact-kernel computation, finite-width optimization, and shallow population guarantees (Jacot et al., 2018; Arora et al., 2019; Du et al., 2019; Allen-Zhu et al., 2019; Ji & Telgarsky, 2020; Nitanda & Suzuki, 2021; Chizat et al., 2019). Pyramidal and deep-ReLU curvature results include Nguyen & Mondelli

(2020); Nguyen et al. (2021); Karhadkar et al. (2024); KMM supplies initialization curvature for the optional branch, not persistence. Population learning combines early stopping (Raskutti et al., 2014), heat smoothing, and endpoint analysis (Bietti & Bach, 2020); sampling uses Chen et al. (2023b); Benton et al. (2024); Li et al. (2024).

B Proofs of the Main Results and Parameter Calibration

This appendix assembles the component bounds into the proofs of Theorem 1 and Corollary 1 and the higher-regularity corollary. The component estimates are proved in Section C, with foundations in Sections D to F. The supplied numerical class bounds are listed in Section G.1. All expectations below are over fresh variables independent of the frozen training sample and initialization unless stated otherwise. Once a predictor has been trained, it is held fixed throughout the population and sampling comparisons.

B.1 Sampling laws and the total-variation reduction

Condition on the realized training sample and initialization, so that the learned score $\widehat{s}$ is fixed. All expectations below use a fresh condition Y and a fresh conditional forward OU path. We first identify the initialization, discretization, and score errors, and then introduce the deterministic budgets used in the sampling bound.

The three sampling laws. For a fixed condition y, the target $p_{t_0}(\cdot \mid y)$ is the endpoint of the exact reverse-OU process started from $p_T(\cdot \mid y)$ and driven by the true score $s^\star$. Let $\widetilde{p}_{t_0}(\cdot \mid y)$ be the endpoint of the discretized learned-score chain in (75) when started from $p_T(\cdot \mid y)$. The implementable law $\widehat{p}_{t_0}(\cdot \mid y)$ uses the same updates but starts from $\gamma_0 = N(0, I_{d_x})$. If $Q_{\widehat{s},y}$ denotes the composition of these discrete transition kernels, then

$$\widetilde{p}_{t_0}(\cdot \mid y) = p_T(\cdot \mid y) Q_{\widehat{s},y}, \qquad \widehat{p}_{t_0}(\cdot \mid y) = \gamma_0 Q_{\widehat{s},y}.$$

The triangle inequality and contraction under this common Markov kernel give

$$\begin{aligned} &\mathrm{TV}(p_{t_0}(\cdot \mid y), \widehat{p}_{t_0}(\cdot \mid y)) \\ &\quad \le \mathrm{TV}(p_{t_0}(\cdot \mid y), \widetilde{p}_{t_0}(\cdot \mid y)) + \mathrm{TV}(p_T(\cdot \mid y), \gamma_0)\,. \end{aligned} \tag{20}$$

Terminal-initialization error. The exact mismatch at the starting time of the reverse sampler is

$$\delta_{\mathrm{term}}(T) := \mathbb{E}_Y \mathrm{TV}(p_T(\cdot \mid Y), \gamma_0) = \frac{1}{2}\mathbb{E}_Y \int_{\mathbb{R}^{d_x}} |p_T(x \mid Y) - \varphi(x)|\ \mathrm{d}x, \tag{21}$$

where φ is the standard Gaussian density. It also bounds the propagated initialization discrepancy $\mathbb{E}_Y \mathrm{TV}(\widetilde{p}_{t_0}(\cdot \mid Y), \widehat{p}_{t_0}(\cdot \mid Y))$ by the same contraction. Put

$$M_2 := \operatorname*{ess\,sup}_y \mathbb{E}[\|X_0\|_2^2 \mid Y = y] < \infty,$$

where finiteness follows from Assumption 3.1 and Lemma D.1. Since $T \ge 1$, Lemma C.1 applies and yields

$$\delta_{\mathrm{term}}(T) \le C e^{-T} \sqrt{d_x + M_2}. \tag{22}$$

That lemma uses the Gaussian-mixture representation of the forward OU law, a relative-entropy bound, and Pinsker's inequality.

Discretization and score-estimation errors. The reverse sampler freezes the score at the state and time at the start of each reverse update, corresponding to the forward time t_{j+1}. The squared error from freezing the true score is

$$\mathcal{D}_{\mathrm{grid}} := \sum_{j=0}^{M-1} \int_{t_j}^{t_{j+1}} \mathbb{E}\left\| s^\star(X_t, t, Y) - s^\star(X_{t_{j+1}}, t_{j+1}, Y) \right\|_2^2\ \mathrm{d}t. \tag{23}$$

Here X_t and $X_{t_{j+1}}$ belong to the same forward OU path. The error from replacing the true score by the learned score at the evaluation points is

$$\mathcal{S}(\widehat{s}) := \sum_{j=0}^{M-1} \Delta_j \mathbb{E} \left\| \widehat{s}(X_{t_{j+1}}, t_{j+1}, Y) - s^\star(X_{t_{j+1}}, t_{j+1}, Y) \right\|_2^2 . \tag{24}$$

Both quantities are evaluated under the true forward process.

The path-space comparison used in the proof of Lemma C.2 separates the drift discrepancy into these two contributions. The squared triangle inequality, data processing, Pinsker's inequality, and averaging over Y give

$$\mathbb{E}_Y \mathrm{TV}(p_{t_0}(\cdot \mid Y), \widetilde{p}_{t_0}(\cdot \mid Y)) \le C\sqrt{\mathcal{S}(\widehat{s})} + C\sqrt{\mathcal{D}_{\mathrm{grid}}}. \tag{25}$$

For the admissible grid $\Delta_j \le \kappa_{\mathrm{grid}} \min\{1, t_j\}$, $0 < \kappa_{\mathrm{grid}} \le 1$, the discretization estimate underlying that lemma gives

$$\mathcal{D}_{\mathrm{grid}} \le C \left(\kappa_{\mathrm{grid}}^2 d_x M + \kappa_{\mathrm{grid}} d_x T + \kappa_{\mathrm{grid}} M_2 \right). \tag{26}$$

Specifically, this is the conditional version of Benton et al. (2024, Lemma 2, with the general-moment modification in Appendix C, Eq. (19)); the path-space step uses their Proposition 3 and Lemma 8, as detailed in Lemma C.2. The square root of $\mathcal{D}_{\mathrm{grid}}$ controls the sampling effect of discretization, including when the score is exact.

The error budgets in the sampling bound. Choose a sufficiently large constant C and define

$$\begin{aligned} \mathcal{E}_{\mathrm{term}} &:= C e^{-T} \sqrt{d_x + M_2}, \\ \mathcal{E}_{\mathrm{disc}} &:= C \sqrt{(\kappa_{\mathrm{grid}}^2 M + \kappa_{\mathrm{grid}} T)(d_x + M_2)}. \end{aligned} \tag{27}$$

Thus $\delta_{\mathrm{term}}(T) \le \mathcal{E}_{\mathrm{term}}$ and $C\sqrt{\mathcal{D}_{\mathrm{grid}}} \le \mathcal{E}_{\mathrm{disc}}$, after enlarging the budget constant if needed; the latter uses $T \ge 1$. The symbols $\mathcal{E}_{\mathrm{term}}$ and $\mathcal{E}_{\mathrm{disc}}$ denote these deterministic upper bounds on the initialization and discretization contributions. For the prescribed grid with $\kappa_{\mathrm{grid}} = \widetilde{O}(M^{-1})$ and the logarithmic terminal-time calibration,

$$\mathcal{E}_{\mathrm{disc}} = \widetilde{O}\left(\sqrt{\frac{d_x + M_2}{M}} \right).$$

Combining (20) with the preceding bounds yields

$$\begin{aligned} &\mathbb{E}_Y \mathrm{TV}(p_{t_0}(\cdot \mid Y), \widehat{p}_{t_0}(\cdot \mid Y)) \\ &\quad \le \mathcal{E}_{\mathrm{term}} + \mathcal{E}_{\mathrm{disc}} + C\sqrt{\mathcal{S}(\widehat{s})}. \end{aligned} \tag{28}$$

The next subsection converts $\mathcal{S}(\widehat{s})$ into clean-sample regression error and clipping bias using the prescribed score-weighted time law.

B.2 Score error and clean-sample regression

Condition on the realized training sample and initialization. The prescribed score-weighted right-endpoint law is

$$\mathbb{P}(J = j) = \frac{\Delta_j c(t_{j+1})}{\mathcal{W}}, \qquad \mathcal{W} = \sum_{j=0}^{M-1} \Delta_j c(t_{j+1}), \qquad c(t) = \frac{e^{-2t}}{(1 - e^{-2t})^2}.$$

For any square-integrable predictor f, Tweedie's identity gives $s_f - s^\star = (\alpha_t/\sigma_t^2)(f - f^\star)$. Consequently,

$$\sum_{j=0}^{M-1} \Delta_j \mathbb{E} \left\| s_f(X_{t_{j+1}}, t_{j+1}, Y) - s^\star(X_{t_{j+1}}, t_{j+1}, Y) \right\|_2^2 = \mathcal{W} \mathbb{E} \left\| f(U) - f^\star(U) \right\|_2^2 . \tag{29}$$

Both sides use fresh variables and the unclipped posterior mean. Since c is decreasing, the right-endpoint sum obeys

$$0 < \mathcal{W} \le \int_{t_0}^{T} c(t)\, dt = \frac{1}{2} \left\{ \frac{1}{1 - e^{-2t_0}} - \frac{1}{1 - e^{-2T}} \right\} \le \frac{1}{4t_0}. \tag{30}$$

The exact identity therefore needs no worst-time factor $(T - t_0)c(t_0)$. Clipping is handled by a separate inequality below.

For the learned predictor, define

$$\mathcal{E}_{\mathrm{reg}}(\widehat{f}) = \sum_{q=1}^{d_x} \mathbb{E}\left|\widehat{f}_q(U) - f^{\star}_{B_N,q}(U)\right|^2, \qquad \mathcal{E}_{\mathrm{clip}} = e^{-cB_N}. \tag{31}$$

The squared triangle inequality and Lemma C.4 give

$$\mathbb{E}\left\|\widehat{f}(U) - f^{\star}(U)\right\|_2^2 \le 2\mathcal{E}_{\mathrm{reg}}(\widehat{f}) + C\mathcal{E}_{\mathrm{clip}}. \tag{32}$$

Proposition B.1 (Architecture-independent learning-to-sampling reduction)**.** *Under Assumption 3.1, condition on any realized square-integrable $\widehat{f}$. Then*

$$\mathbb{E}_Y \mathrm{TV}(p_{t_0}(\cdot \mid Y), \widehat{p}_{t_0}(\cdot \mid Y)) \le Ce^{-T}\sqrt{d_x + M_2} + \widetilde{O}\left(\sqrt{\frac{d_x + M_2}{M}}\right) + C\sqrt{\mathcal{W}\big(\mathcal{E}_{\mathrm{reg}}(\widehat{f}) + \mathcal{E}_{\mathrm{clip}}\big)}. \tag{33}$$

Proof. Insert (29), (B.2), and (32) into (28). □

B.3 The finite-trajectory decomposition

For every output coordinate q, let $g^K_{q,R_\star}$ and $g^\infty_{q,R_\star}$ denote zero-start full-batch kernel-GD predictors at the prescribed stopping time. The first uses the finite initialization kernel and the second the limiting full NTK; both use the same frozen examples, normalized clipped labels, and step size as the nonlinear network. Define

$$\mathcal{E}_{\mathrm{opt}} = B_N^2 \sum_{q=1}^{d_x} \left\| F^c_{\theta_{q,R_\star}} - g^K_{q,R_\star} \right\|^2_{L^2(P_X)}, \tag{34}$$

$$\mathcal{E}_{\mathrm{width}} = B_N^2 \sum_{q=1}^{d_x} \left\| g^K_{q,R_\star} - g^\infty_{q,R_\star} \right\|^2_{L^2(P_X)}, \tag{35}$$

$$\mathcal{E}_{\mathrm{ker}} = \sum_{q=1}^{d_x} \left\| B_N g^\infty_{q,R_\star} \circ \psi_{\rho_N} - f^\star_{B_N,q} \right\|^2_{L^2(P_U)}. \tag{36}$$

Here $\mathcal{E}_{\mathrm{opt}}$ is a neural-to-kernel trajectory discrepancy, not the optimization suboptimality of an ERM. Finite-stopping bias remains inside the limiting-kernel oracle that bounds $\mathcal{E}_{\mathrm{ker}}$.

For each q, inserting the two kernel trajectories gives

$$\begin{aligned}
&\left\| B_N F^c_{\theta_{q,R_\star}} \circ \psi_{\rho_N} - f^\star_{B_N,q} \right\|_{L^2(P_U)} \\
&\quad \le B_N \left\| F^c_{\theta_{q,R_\star}} - g^K_{q,R_\star} \right\|_{L^2(P_X)} + B_N \left\| g^K_{q,R_\star} - g^\infty_{q,R_\star} \right\|_{L^2(P_X)} \\
&\qquad + \left\| B_N g^\infty_{q,R_\star} \circ \psi_{\rho_N} - f^\star_{B_N,q} \right\|_{L^2(P_U)}.
\end{aligned} \tag{37}$$

Using $(a+b+c)^2 \le 3(a^2+b^2+c^2)$ and summing over q yields

$$\mathcal{E}_{\mathrm{reg}}(\widehat{f}) \le 3\big(\mathcal{E}_{\mathrm{opt}} + \mathcal{E}_{\mathrm{width}} + \mathcal{E}_{\mathrm{ker}}\big). \tag{38}$$

Substitution into Proposition B.1 gives

$$\mathbb{E}_Y \mathrm{TV}(p_{t_0}(\cdot \mid Y), \widehat{p}_{t_0}(\cdot \mid Y)) \le \mathcal{E}_{\mathrm{term}} + \mathcal{E}_{\mathrm{disc}} + C\sqrt{\mathcal{W}\big(\mathcal{E}_{\mathrm{clip}} + \mathcal{E}_{\mathrm{opt}} + \mathcal{E}_{\mathrm{width}} + \mathcal{E}_{\mathrm{ker}}\big)}. \tag{39}$$

The limiting-kernel oracle is an upper bound on $\mathcal{E}_{\mathrm{ker}}$; it is not an exact additive decomposition into approximation and estimation errors.

B.4 Component bounds and statistical calibration

Let $k \geq 1$ be the integer regularity order in Assumption 3.2, and recall $\bar{k} = \min\{k, s\}$ from (9). For the general calibration, fix an auxiliary order $1 \leq u \leq \bar{k}$ and define

$$s = \frac{D+2}{2}, \qquad \gamma = \frac{D+1}{D+2}, \qquad b = \frac{u}{s}, \\ r_u = \frac{u}{u+\gamma s}, \qquad \ell_u = \frac{s}{u+\gamma s}. \tag{40}$$

These definitions include the saturated case $u = s$ and satisfy $r_u + \gamma \ell_u = 1$. The main results use $u = \bar{k}$, for which $r_u = r$ and $\ell_u = \ell$. Retaining the auxiliary order allows the original-law analysis below to also use lower available regularity.

Put $j = \lceil u \rceil \leq k$ and apply the integer-order extension construction at order j. This supplies the required $C^u_* \cap H^u$ envelope, where C^u_* denotes the Hölder–Zygmund space. Let $A = A_{u,N}(t_0) \geq 1$ be a deterministic envelope for the normalized extension norms, uniformly over coordinates. The positive-time extension construction gives

$$A \leq C \rho_N^{c_j} (1+T)^{c_j} (1 + t_0^{-1})^{\chi_j}, \qquad \chi_j = j \max\{4, b_j\}, \qquad A = \widetilde{O}(t_0^{-\chi_j}), \tag{41}$$

where $b_j = \max_{1 \leq |\alpha| \leq j} b_\alpha$ in Assumption 3.2. The global extension and its qualified heat comparators also have a common supremum bound depending only on the fixed derivative order, independently of A. Write $\tau_N = \mathbb{P}(\|U\|_2 > \rho_N)$ and $\mathcal{L}_N = \log(CNd_x/\delta)$, with fixed constants enlarged to include the depth and confidence allocations.

For numerical calibration, replace τ_N in every prescribed risk envelope by a known upper bound. For example, if $\rho_N > T$, put $a_N = (\rho_N - T)/3$ and use

$$\tau_N \leq \overline{\tau}_N := \min\left\{1,\ M_X e^{-\lambda_X a_N} + M_Y e^{-\lambda_Y a_N} + 2^{d_x/2} e^{1-a_N}\right\}. \tag{42}$$

Indeed, $\|U\|_2 \leq \|X_0\|_2 + \|Y\|_2 + \|\varepsilon\|_2 + T$; apply a union bound and Markov's inequality, using $\mathbb{E}e^{\|\varepsilon\|_2} \leq e\, 2^{d_x/2}$ for a standard Gaussian. Thus the schedule requires neither an unknown population tail probability nor an empirical estimate of the extension norm.

Proposition B.2 (Bias–variance bound under the actual design)**.** *For $0 < \eta \leq 1/(4L)$ and a deterministic $0 < \lambda \leq 1$ satisfying the regularized comparison condition, the limiting-kernel iterate at $R_\star = \lceil (\eta\lambda)^{-1} \rceil$ satisfies, with probability at least $1 - \delta$,*

$$\mathcal{E}_{\mathrm{ker}} \leq C d_x B_N^2 \left[A^2 \lambda^b + \frac{\mathcal{L}_N}{N\lambda^\gamma} + \tau_N \right]. \tag{43}$$

Constants in the comparison condition may be enlarged for rounding.

Proof. For $u < s$, take the qualified iterated-heat comparator $R_v^{(J)} F$ with fixed integer $J > u/2$ and $v = \lambda^{1/s}$. Its squared uniform approximation error on the localized region is at most $CA^2 v^u$, and its squared RKHS norm is at most $CA^2 v^{-(s-u)}$. Their sum, with the norm multiplied by λ, is at most $CA^2 \lambda^{u/s}$. Off the localized region, boundedness contributes $C\tau_N$. At saturation $u = s$, the extension itself has squared RKHS norm at most CA^2 and approximation error at most $C\tau_N$.

Apply Theorem E.24 to this deterministic bounded comparator in each coordinate and use Proposition E.19 for its variance. Normalized clipped labels have universal sub-Gaussian noise. Allocate δ/d_x per coordinate, multiply by B_N^2, and sum. Finally, $\lambda/2 \leq (\eta R_\star)^{-1} \leq \lambda$, so rounding only changes constants. This proof retains localization even in the saturated case. □

Proposition B.3 (Deterministic stopping with a comparison floor)**.** *For sufficiently large $N/\mathcal{L}_N$, prescribe*

$$\lambda_{\mathrm{bal}} = \left(\frac{\mathcal{L}_N}{NA^2} \right)^{1/(b+\gamma)}, \quad \lambda_N = \max\left\{ \lambda_{\mathrm{bal}}, C_0 \frac{\mathcal{L}_N}{N} \right\}, \\ R_\star = \left\lceil (\eta \lambda_N)^{-1} \right\rceil. \tag{44}$$

For sufficiently large fixed C_0, this schedule satisfies

$$N(\eta R_\star)^{-1} \geq CL \log\left(\frac{CL}{(\eta R_\star)^{-1} \delta / d_x} \right). \tag{45}$$

A deterministic squared-risk envelope is

$$\mathfrak{r}_N = Cd_x B_N^2 \left[A^{\frac{2\gamma}{b+\gamma}} \left(\frac{\mathcal{L}_N}{N} \right)^{\frac{b}{b+\gamma}} + A^2 \left(\frac{\mathcal{L}_N}{N} \right)^b + \tau_N \right]. \tag{46}$$

The definitions permit deterministic dependence of A on N, t_0. In an instantiated schedule, the risk envelope uses $\overline{\tau}_N$ from (42), or any other supplied deterministic upper bound in place of τ_N.

Proof. The floor implies $\lambda_N \geq C_0 \mathcal{L}_N / N$ and bounds the logarithm in (45) by $C' \mathcal{L}_N$, also at the rounded scale $\lambda_N/2$. Since $A \geq 1$, both terms defining λ_N are eventually at most one. Use $\max\{x, y\}^b \leq x^b + y^b$ to bound the bias in (43), and $\lambda_N^{-\gamma} \leq \lambda_{\mathrm{bal}}^{-\gamma}$ to bound its variance. The resulting balanced term and floor bias are exactly those in (46). □

For $1 \leq u < s$, the nonsaturated calibration is

$$A = A_{u,N}(t_0), \qquad b = u/s, \qquad R_{\mathrm{cmp,N}} = CA^2 \lambda_N^{-(s-u)/s}, \tag{47}$$

with $(\lambda_N, R_\star, \mathfrak{r}_N)$ from (44)–(46). At saturation $u = s$, the corresponding prescription is

$$\begin{aligned} A &= A_{s,N}(t_0), \qquad b = 1, \\ \lambda_N &= \max\left\{ \left(\frac{\mathcal{L}_N}{NA^2} \right)^{1/(1+\gamma)}, C_0 \frac{\mathcal{L}_N}{N} \right\}, \\ \mathfrak{r}_N &= Cd_x B_N^2 \left[A^{\frac{2\gamma}{1+\gamma}} \left(\frac{\mathcal{L}_N}{N} \right)^{\frac{1}{1+\gamma}} + A^2 \frac{\mathcal{L}_N}{N} + \tau_N \right]. \end{aligned} \tag{48}$$

These are stopping rules for the unpenalized neural loss. No regularizer is added to training. Take $B_N, \rho_N = \Theta(\log(N/\delta) + T)$ with sufficient constants so that clipping and localization are absorbed into the prescribed risk budget. With shrinking t_0, these constants may instead be chosen explicitly from (72) below.

At the balanced scale the norm factors are

$$\begin{aligned} \lambda_{\mathrm{bal}} &= \widetilde{\Theta}\left(A^{-2s/(u+\gamma s)} N^{-s/(u+\gamma s)} \right), \\ \mathfrak{r}_N &= \widetilde{O}\left(d_x B_N^2 A^{2\gamma s/(u+\gamma s)} N^{-u/(u+\gamma s)} \right), \\ R_{\mathrm{cmp,N}} &= \widetilde{O}\left(A^{2s(1+\gamma)/(u+\gamma s)} N^{(s-u)/(u+\gamma s)} \right) \qquad (u < s). \end{aligned} \tag{49}$$

The balanced formulas require, up to logarithms,

$$N^{u-1/2} \gtrsim A^{2s}. \tag{50}$$

Indeed, $b + \gamma - 1 = (u - 1/2)/s > 0$. At fixed t_0, A is at most polylogarithmic, so the balanced scale is feasible for all sufficiently large N, and the floor term decays faster. The floor in (44) remains essential when t_0 varies.

B.5 Explicit finite-width choices

Use the calibration at a fixed auxiliary order $1 \leq u \leq \bar{k}$ from (40). Fix $L \geq 3$, put $H = L - 1$, $d_0 = D + 2$, and let $m = d_H$. Under (4),

$$m \leq d_\ell \leq \kappa_{\mathrm{pyr},N}^{H-\ell} m, \qquad 1 \leq \ell \leq H. \tag{51}$$

For an adjacent-width bound $\kappa \geq 1$, let $\Gamma_{L,\kappa}$ be a sufficiently large finite envelope for the initialization, perturbation, and trajectory constants:

$$\Gamma_{L,\kappa} := \max\{1, C_{L,\kappa}^{\mathrm{init}}, C_{L,\kappa}^{\mathrm{loc}}, C_{L,\kappa}^{\mathrm{traj}}\}. \tag{52}$$

Here κ denotes an architecture bound, not a regularity order. The constants are those in the local proofs below, enlarged by a maximum when combining prerequisites. No numerical or polynomial-in-depth envelope is asserted. If $\kappa_{\mathrm{pyr},N}$ varies with N, its dependence is retained in $\Gamma_{L,\kappa_{\mathrm{pyr},N}}$.

Set $h = \eta R_\star$ and

$$\varepsilon_{\mathrm{coup},N} = \min\left\{1, \frac{c_{\mathrm{coup}}}{B_N\sqrt{d_x}}\mathfrak{r}_N^{1/2}\right\}, \qquad \Xi_N = h + h^2. \tag{53}$$

Define

$$\begin{aligned}\nu_m &= 1 + d_0 H \log(e\kappa_{\mathrm{pyr},N} m) + \log(64 H d_x/\delta),\\ q_\star &= C_{L,\kappa_{\mathrm{pyr},N}}\sqrt{\nu_m}(1+h)^{1/6} m^{-1/6}.\end{aligned} \tag{54}$$

The separate movement and trajectory scales are

$$\begin{aligned}W_{\mathrm{mov}} &= (1+h)^7,\\ W_{\mathrm{ker}} &= \Xi_N^6 \varepsilon_{\mathrm{coup},N}^{-6},\\ W_{\mathrm{rate}} &= (1+h)\Xi_N^6 \varepsilon_{\mathrm{coup},N}^{-6},\\ W_{\mathrm{stat}} &= 1 \vee W_{\mathrm{mov}} \vee W_{\mathrm{ker}} \vee W_{\mathrm{rate}}.\end{aligned} \tag{55}$$

For a positive width scale W, write

$$\mathsf{Width}(W): \qquad m \geq C\Gamma_{L,\kappa_{\mathrm{pyr},N}}\nu_m^3 W. \tag{56}$$

In particular, the statistical condition is

$$m \geq C\Gamma_{L,\kappa_{\mathrm{pyr},N}}\nu_m^3 W_{\mathrm{stat}}. \tag{57}$$

These are implicit logarithmic width inequalities, because ν_m contains $\log m$. For fixed choices of the other quantities they admit finite solutions.

The movement scale supplies the radius $4\sqrt{1+h}$ and the residual bound in Proposition F.13. The nonlinear population coupling has size $Cq_\star\Xi_N$. Requiring it to be at most $\varepsilon_{\mathrm{coup},N}$ gives the W_{rate} branch after taking sixth powers:

$$q_\star\Xi_N \lesssim \varepsilon_{\mathrm{coup},N} \quad \Longleftarrow \quad m \gtrsim \Gamma_{L,\kappa_{\mathrm{pyr},N}}\nu_m^3(1+h)\Xi_N^6\varepsilon_{\mathrm{coup},N}^{-6}. \tag{58}$$

Full-NTK concentration and direct comparison of the two zero-start kernel trajectories give population discrepancy at most $C\Xi_N\sqrt{\nu_m}m^{-1/6}$, which yields W_{ker}. Thus (57) controls movement, both population comparisons, and their initialization and local-perturbation prerequisites. In particular, a convenient sufficient condition is

$$\boxed{m \geq C_{L,\kappa_{\mathrm{pyr},N}}\nu_m^3(1+h)^{13}\varepsilon_{\mathrm{coup},N}^{-6}.} \tag{59}$$

Indeed, $\Xi_N = h(1+h)$ and $\varepsilon_{\mathrm{coup},N} \leq 1$, so the displayed expression dominates every branch of W_{stat}. The full derivations are Corollary C.9, Proposition C.10, and Theorem C.12. No inverse curvature factor enters these statistical conditions.

For clarity about the population normalization, the averaged cross-kernel operator $\mathcal{A}_K v = N^{-1}\sum_i K(\cdot, X_i)v_i$ has norm $O(1)$ from $\ell_{2,N}$ to $L^2(P_X)$; the unaveraged operator has norm $O(N)$. The Duhamel accumulation therefore gives $\Xi_N = h + h^2$. Section F.2 gives the deterministic derivation.

Corollary B.4 (Sufficient statistical-width exponent)**.** *Fix $1 \leq u \leq \bar{k}$. At fixed t_0, depth, dimensions, and adjacent-width ratios, the calibrated scales satisfy*

$$h = \widetilde{O}(N^{\ell_u}), \qquad \varepsilon_{\mathrm{coup},N}^{-6} = \widetilde{O}(N^{3r_u}), \qquad m = \widetilde{O}(N^{13\ell_u + 3r_u}) \tag{60}$$

as a sufficient width choice with sufficiently large constants and logarithmic factors. For the main results, $u = \bar{k}$ gives $r_u = r$ and $\ell_u = \ell$. In the saturated regime $u = s$, this gives $m = \widetilde{O}(N^{16/(1+\gamma)})$.

Proof. Use the exact risk and stopping calibrations in (47) and (48). Their positive envelope and clipping factors grow at most polylogarithmically for fixed t_0; choose the envelopes at least one. Substitution into (53), rather than inversion of a one-sided risk bound, gives the stated inverse-tolerance scale. Equation (59) then contributes

$$13\ell_u + 3r_u = 3 + (13 - 3\gamma)\ell_u.$$

Lemma B.5 resolves the remaining logarithmic dependence on m without changing this power. □

For $1 \le u < s$, the separate branch exponents are

$$\begin{aligned} E_{\mathrm{mov}}(u) &= 7\ell_u, \qquad E_{\mathrm{ker}}(u) = 12\ell_u + 3r_u, \\ E_{\mathrm{rate}}(u) &= E_{\mathrm{stat}}(u) = 13\ell_u + 3r_u. \end{aligned} \tag{61}$$

These sufficient width costs leave the calibrated clean-sample exponents unchanged.

The symbols A, a, b in the following lemma and its application are local to the logarithmic-width calculation.

Lemma B.5 (Resolving a logarithmic width condition)**.** *Let $A \ge 1$, $a \ge 1$, and $b \ge 1$ be deterministic. Put*

$$Q = a + b\log\big(eA(a+b)^3\big), \qquad m = \lceil 4096AQ^3 \rceil .$$

Then $m \ge A(a + b\log m)^3$.

Proof. Set $t = Q/(a+b) \ge 1$. Since $m \le 8192AQ^3$,

$$\begin{aligned} a + b\log m &\le a + b\log(8192AQ^3) \\ &= Q + b(\log 8192 - 1) + 3b\log t \\ &\le Q\big(\log 8192 + 3/e\big) < 16Q. \end{aligned}$$

For the last inequality use $b \le Q$, $b/Q \le 1/t$, and $\sup_{t\ge 1}(\log t)/t = 1/e$. Cubing yields

$$A(a + b\log m)^3 \le 4096AQ^3 \le m.$$

□

Apply this with

$$\begin{aligned} a &= 1 + d_0 H\log(e\kappa_{\mathrm{pyr},N}) + \log(64Hd_x/\delta), \\ b &= d_0 H, \qquad A = \max\{1, C_{L,\kappa_{\mathrm{pyr},N}}(1+h)^{13}\varepsilon_{\mathrm{coup},N}^{-6}\}. \end{aligned}$$

This yields an explicit finite choice once the existing proof constants are specified. The proof constants remain analytical envelopes, so this is not a numerically practical width prescription.

Failure-probability allocation. For the statistical event, allocate δ across the operator, misspecification, and noise events of kernel learning, and the rectangular initialization, perturbation, and finite/full-NTK concentration events, including the union over scalar heads. Clipping, population localization, and the sampler reduction are deterministic bounds conditional on the trained predictor. Any sample-localization event used by an optional certificate receives its own allocation. For the optional geometry event, allocate additional portions to Euclidean separation and KMM initialization. Replacing local failure levels by constant multiples of $\delta/(Ld_x)$ changes only logarithmic factors. Geometry conditioning events are measurable with respect to the frozen covariates, so output-head initializations may still be union-bounded conditionally.

B.6 Assembly of the main theorem and corollaries

The fixed-t_0 theorem is the primary learning guarantee: its sample and width exponents keep t_0 fixed while the requested accuracy improves. The original-law theorem below is a separate calibration in which t_0 shrinks. Both the extension envelope and the score-to-regression factor then grow, so the fixed-t_0 resource powers cannot be used for that different target without recalibration.

Proposition B.6 (Fresh-population risk of the finite trained predictor)**.** *Suppose Assumptions 3.1 and 3.2 hold. Fix $1 \le u \le \bar{k}$, use the corresponding calibration (47) or (48), and impose* $\mathsf{Width}(W_{\mathrm{stat}})$*. Then, with probability at least $1 - C\delta$, simultaneously for all output heads,*

$$\sum_{q=1}^{d_x} \mathbb{E}_{U\sim P_U}\left[\left|\widehat{f}_q(U) - f^{\star}_{B_N,q}(U)\right|^2 \,\middle|\, \textit{frozen training sample and initialization}\right] \le C\mathfrak{r}_N. \tag{62}$$

Proof. Under $\mathsf{Width}(W_{\mathrm{stat}})$, Proposition C.10 and Theorem C.12 give $\mathcal{E}_{\mathrm{opt}} \le d_x B_N^2 \varepsilon_{\mathrm{coup},N}^2$, and Corollary C.9 gives $\mathcal{E}_{\mathrm{width}} \le d_x B_N^2 \varepsilon_{\mathrm{coup},N}^2$. By the definition of $\varepsilon_{\mathrm{coup},N}$, both are at most a sufficiently small constant multiple of $\mathfrak{r}_N$. For $u < s$, Theorem C.7 and the localization choice give $\mathcal{E}_{\mathrm{ker}} \le C\mathfrak{r}_N$; for $u = s$, use Proposition C.8. Equation (38) proves (62). This argument invokes no separation event, KMM certificate, lower Gram eigenvalue, or curvature-persistence result. □

Define the stopped-law envelope

$$\mathsf{Env}_N(z) = \mathcal{E}_{\mathrm{term}} + \mathcal{E}_{\mathrm{disc}} + C\sqrt{\mathcal{W}(z + e^{-cB_N})}. \tag{63}$$

By Propositions B.1 and B.6,

$$\mathbb{E}_Y \mathrm{TV}(p_{t_0}(\cdot \mid Y), \widehat{p}_{t_0}(\cdot \mid Y)) \le \mathsf{Env}_N(\mathfrak{r}_N). \tag{64}$$

Proof of Theorem 1. Set $u = \bar{k}$ in (40), so $b = \bar{k}/s$, $r_u = r$, and $\ell_u = \ell$. Use Proposition B.6 and (64). Choose $T = \Theta(\log(1/\varepsilon))$ and an admissible grid with $M = \widetilde{\Theta}(\varepsilon^{-2})$ to control initialization and discretization. At fixed t_0, $\mathcal{W}$ is bounded and A, B_N grow at most polylogarithmically. Since

$$b + \gamma - 1 = \frac{\bar{k} - 1/2}{s} > 0,$$

the comparison floor is eventually below λ_{bal}, and (46) is $\widetilde{O}(N^{-r})$. The learned-score term is $\widetilde{O}(N^{-r/2})$, so the sufficient sample exponent is

$$\frac{2}{r} = 2 + \frac{D+1}{\bar{k}}.$$

The deterministic threshold can be enlarged to meet clipping, localization, and the comparison condition for every larger N. Their logarithmic requirements do not alter this exponent. The finite width exists by Lemma B.5, and Corollary B.4, with $u = \bar{k}$, gives its scale. A fixed union allocation over the statistical and initialization events gives probability at least $1 - C\delta$. This proof does not invoke the optional geometry branch. □

Corollary B.7 (Sample complexity under higher conditional-law regularity)**.** *Under the conditions of Theorem 1, suppose $k \ge \lceil s \rceil = \lceil (D+2)/2 \rceil$. Then $\bar{k} = s$, and the stopping calibration is (48). For fixed t_0, depth, dimensions, and adjacent-width ratios, the guarantee (11) holds with sufficient sample size and width*

$$N = \widetilde{O}\left(\varepsilon^{-[4-2/(D+2)]}\right), \qquad m = \widetilde{O}\left(N^{16/(1+\gamma)}\right).$$

This regime follows from the conditional-law regularity in Assumption 3.2: the proof derives the corresponding regularity of the positive-time denoising target and uses it to obtain the approximation bound. Further increases in k leave the displayed sample exponent unchanged.

Proof. Use the positive-time extension at integer order $j_s = \lceil s \rceil \le k$ and retain its H^s norm. The full-NTK RKHS equivalence and bounded extension give comparator error $C\tau_N$ and squared norm CA^2. Thus the rate-critical terms are $\lambda A^2 + (N\lambda^\gamma)^{-1}$. At fixed t_0, the constrained rule reaches their balance, with squared risk $\widetilde{O}(N^{-1/(1+\gamma)})$. The finite-network comparisons have that same budget. Taking a square root in the sampling reduction gives sample exponent $2(1+\gamma) = 4 - 2/(D+2)$. Substituting $r_s = \ell_s = 1/(1+\gamma)$ into (60) gives width exponent $16/(1+\gamma)$. □

Proposition B.8 (Small-time risk with the comparison floor)**.** *Fix $1 \le u \le \bar{k}$ and suppose $A = \widetilde{O}(t_0^{-\chi})$, with χ the exponent for the required extension order. For $u < s$, put*

$$\theta_u = \frac{2\gamma s \chi}{u + \gamma s}. \tag{65}$$

The constrained calibration obeys

$$\mathfrak{r}_N(t_0) = \widetilde{O}\left(t_0^{-\theta_u} N^{-r_u} + t_0^{-2\chi} N^{-u/s} + B_N^2 \tau_N\right). \tag{66}$$

The first term alone is sufficient only when (50) holds. At saturation $u = s$, take

$$j_s = \lceil s \rceil, \qquad \chi_{j_s} = j_s \max\{4, b_{j_s}\}, \qquad \theta_{\mathrm{sat}} = \frac{2\gamma\chi_{j_s}}{1+\gamma}. \tag{67}$$

Then replace the first term by $t_0^{-\theta_{\mathrm{sat}}} N^{-1/(1+\gamma)}$ and the second by $t_0^{-2\chi_{j_s}} N^{-1}$.

Proof. Substitute $A = \widetilde{O}(t_0^{-\chi})$ into both nonlocalization terms of (46); the logarithmic clipping factors are included in the notation. The same substitution with $b = 1$ proves the saturated assertion. No floor term is dropped. □

Theorem B.9 (Original-law TV with a feasible stopping schedule)**.** *Suppose the statistical hypotheses of Theorem 1 and Assumption 3.3 with $0 < \beta \le 1$ hold. Fix $1 \le u \le \bar{k}$, put $b = u/s$, and use the calibration at this order. Let $A = \widetilde{O}(t_0^{-\chi})$, using $j = \lceil u \rceil \le k$ and $\chi = \chi_j = j \max\{4, b_j\}$ as in (41). On the common event,*

$$\begin{aligned}\mathbb{E}_Y \mathrm{TV}(p_0(\cdot \mid Y), \widehat{p}_{t_0}(\cdot \mid Y)) \le \mathcal{E}_{\mathrm{term}} + \mathcal{E}_{\mathrm{disc}} + C\sqrt{\mathcal{W}(\mathfrak{r}_N + e^{-cB_N})} \\ + C t_0^{\beta/2} \left(1 + \log \frac{C}{t_0}\right)^{d_x+\beta}.\end{aligned} \tag{68}$$

A sufficient deterministic sample threshold for error $\widetilde{O}(\varepsilon)$, with probability at least $1 - C\delta$, is $\widetilde{O}(\varepsilon^{-p})$, where

$$\begin{aligned} p &= \max\{p_{\mathrm{stat}}, p_{\mathrm{gate}}\}, \\ p_{\mathrm{stat}} &= 2\left(1 + \frac{\gamma}{b}\right)\left(1 + \frac{1}{\beta}\right) + \frac{4\gamma\chi}{\beta b}, \qquad p_{\mathrm{gate}} = \frac{2}{b}\left(1 + \frac{1+2\chi}{\beta}\right). \end{aligned} \tag{69}$$

Use the constrained stopping rule and choose

$$t_0 = \widetilde{\Theta}(\varepsilon^{2/\beta}), \quad T = \Theta(\log(1/\varepsilon)), \quad M = \widetilde{\Theta}(\varepsilon^{-2}), \quad B_N \gtrsim \log \frac{1}{\varepsilon t_0}, \tag{70}$$

with clipping/localization errors at most $c\varepsilon^2/\mathcal{W}$. Recalibrate width using this actual stopping horizon and regression budget. At saturation $u = s$, the explicit exponent is

$$p = \max\left\{2(1+\gamma)\left(1 + \frac{1}{\beta}\right) + \frac{4\gamma\chi_{\lceil s \rceil}}{\beta},\ 2\left(1 + \frac{1+2\chi_{\lceil s \rceil}}{\beta}\right)\right\}. \tag{71}$$

Proof. The triangle inequality and Lemma C.13 give (68). Under uniform conditional exponential tails and density Hölder regularity, that smoothing lemma has uniform bound

$$C t_0^{\beta/2} \left(1 + \log(C/t_0)\right)^{d_x+\beta}.$$

Choose the terminal time and grid as above. Since $\mathcal{W} \le 1/(4t_0)$, it suffices to use normalized regression budget $\tau = c\varepsilon^2 t_0/(d_x B_N^2)$. Each of the first two terms inside the brackets in (46) is at most $C\tau$ provided

$$N \gtrsim \mathcal{L}_N \max\left\{A^{2\gamma/b}\tau^{-(1+\gamma/b)},\ A^{2/b}\tau^{-1/b}\right\}. \tag{72}$$

Equivalently, a bias-admissible scale $\lambda \asymp (\tau/A^2)^{1/b}$ needs both $N \gtrsim \mathcal{L}_N \tau^{-1}\lambda^{-\gamma}$ for variance and $N \gtrsim \mathcal{L}_N \lambda^{-1}$ for population comparison.

Substitute $A = \widetilde{O}(t_0^{-\chi})$ and $t_0 = \widetilde{\Theta}(\varepsilon^{2/\beta})$ to obtain the two powers in (69). Thus the comparison floor is enforced even when the unconstrained balance would be too small. Sufficient logarithmic clipping and localization levels absorb their tails. The population couplings must each be at most $c\varepsilon\sqrt{t_0}/(B_N\sqrt{d_x})$; the same finite-width formula with this horizon and tolerance supplies that budget. Allocate fixed fractions of δ to the learning and initialization/coupling events. The sampler bound and smoothing bound then give the asserted high-probability TV guarantee. A logarithmic reduction of t_0 makes the smoothing term $O(\varepsilon)$ if this is desired without a tilde. □

Proof of Corollary 1. Apply Theorem B.9 with $u = \bar{k}$. Then $b = \bar{k}/s$, $j = \lceil \bar{k} \rceil$, and $\chi = j \max\{4, b_j\}$, as stated in the corollary. The shrinking t_0 changes the extension norm, constrained stopping rule, coupling tolerance, and sufficient width. All are chosen again from deterministic bounds before training. □

Remark B.10 (Small-time smoothness and numerical disclosure). The choice $u = \bar{k}$ gives the best fixed-t_0 sample exponent available from this analysis. Original-law recovery also depends on the derivative-envelope power $\chi_{\lceil u \rceil}$, so the largest available order need not minimize the original-law sample exponent. The general formulation of Theorem B.9 therefore permits any $1 \le u \le \bar{k}$. A lower order may improve the small-time calibration; its extension envelope, stopping horizon, and width must all be calibrated using that same order. The main corollary uses $u = k$.

For $D = 2$, $k = \beta = 1$, and $\chi_1 = 4$, one has $\bar{k} = u = 1$, $s = 2$, $b = 1/2$, and $\gamma = 3/4$, giving

$$\theta_1 = 24/5, \qquad p_{\mathrm{stat}} = 34, \qquad p_{\mathrm{gate}} = 40, \qquad p = 40.$$

Using 34 alone would fail to impose the population-comparison condition. For this same $k = 1$ case, the fixed-t_0 theorem has clean-sample exponent $2 + (D+1)/\bar{k} = 5$, whereas original-law recovery has the sufficient exponent 40 above. The different exponents describe different targets, and the associated stopping horizon and width must also be recomputed when t_0 shrinks. These large powers are conservative sufficient bounds, not necessity claims. The logistic fixed-component mixture has $b_k = 0$, hence $\chi_k = 4k$; it does not improve this general derivative envelope.

C Bounds for the Component Errors

This appendix states the component bounds used by the sampling decomposition. Section B gives their deterministic calibration and assembles the main results; Sections D to F provide the regularity, kernel, concentration, and perturbation inputs. Population norms and expectations throughout use a fresh draw while the realized training sample and initialization are held fixed.

C.1 Terminal initialization

Lemma C.1 (Terminal OU law is close to Gaussian initialization)**.** *Let $\gamma_0 = N(0, I_{d_x})$. Under Assumption 3.1, for $e^{-2T} \le 1/2$,*

$$\mathbb{E}_Y \mathrm{TV}(p_T(\cdot \mid Y), \gamma_0) \le C e^{-T} \sqrt{d_x + M_2}. \tag{73}$$

If Q_y is any Markov kernel, then

$$\mathrm{TV}(p_T(\cdot \mid y) Q_y, \gamma_0 Q_y) \le \mathrm{TV}(p_T(\cdot \mid y), \gamma_0). \tag{74}$$

Proof. Conditional on $X_0 = z$, $X_T \sim N(\alpha_T z, \sigma_T^2 I)$. Convexity of relative entropy gives

$$\begin{aligned} \mathrm{KL}(p_T(\cdot \mid y) \| \gamma_0) &\le \frac{1}{2}\left[\alpha_T^2 \mathbb{E}(\|X_0\|_2^2 \mid Y = y) + d_x(\sigma_T^2 - 1 - \log \sigma_T^2)\right] \\ &\le C e^{-2T}\big(\mathbb{E}(\|X_0\|_2^2 \mid Y = y) + d_x\big). \end{aligned}$$

Pinsker's inequality, averaging, and Jensen's inequality prove (73). Equation (74) is total-variation contraction under a common Markov kernel. □

Thus $\mathcal{E}_{\mathrm{term}} = C e^{-T} \sqrt{d_x + M_2}$ controls the terminal component.

C.2 Reverse stability and discretization

For a fixed score estimator $\bar{s}$ and condition y, the implementable reverse chain is

$$\begin{aligned} &\widehat{X}_M \sim \gamma_0 = N(0, I_{d_x}), \\ &\widehat{X}_j = e^{\Delta_j} \widehat{X}_{j+1} + 2(e^{\Delta_j} - 1)\bar{s}(\widehat{X}_{j+1}, t_{j+1}, y) + \sqrt{e^{2\Delta_j} - 1}\, Z_j, \qquad j = M-1, \dots, 0, \end{aligned} \tag{75}$$

with independent standard Gaussian Z_j. The output law is $\widehat{p}_{t_0}(\cdot \mid y)$. Initializing the same kernels from $p_T(\cdot \mid y)$ instead defines $\widetilde{p}_{t_0}(\cdot \mid y)$. Indeed, solving $dV_u = (V_u + 2a)du + \sqrt{2}\, dB_u$ over an interval of length Δ, with a frozen at its initial state, gives the deterministic coefficient $2(e^{\Delta} - 1)$ and noise covariance $2\int_0^{\Delta} e^{2(\Delta - u)} du\, I = (e^{2\Delta} - 1) I$.

Lemma C.2 (Conditional reverse-process stability). *Assume $T \geq 1$, $0 < \kappa_{\mathrm{grid}} \leq 1$, and $\Delta_j \leq \kappa_{\mathrm{grid}} \min\{1, t_j\}$. For a fixed measurable, finite-valued score estimator $\bar{s}$, set*

$$\mathcal{S}(\bar{s}) = \sum_{j=0}^{M-1} \Delta_j \mathbb{E} \left\| \bar{s}(X_{t_{j+1}}, t_{j+1}, Y) - s^\star(X_{t_{j+1}}, t_{j+1}, Y) \right\|_2^2 .$$

Then

$$\mathbb{E}_Y \mathrm{TV}(p_{t_0}(\cdot \mid Y), \widetilde{p}_{t_0}(\cdot \mid Y)) \leq C\sqrt{\mathcal{S}(\bar{s})} + C\sqrt{\kappa_{\mathrm{grid}}^2 d_x M + \kappa_{\mathrm{grid}} d_x T + \kappa_{\mathrm{grid}} M_2}, \tag{76}$$

$$\mathbb{E}_Y \mathrm{TV}(p_{t_0}(\cdot \mid Y), \widehat{p}_{t_0}(\cdot \mid Y)) \leq C\sqrt{\mathcal{S}(\bar{s}) + \mathcal{E}_{\mathrm{disc}}} + Ce^{-T}\sqrt{d_x + M_2},$$
$$\mathcal{E}_{\mathrm{disc}} := C\sqrt{(\kappa_{\mathrm{grid}}^2 M + \kappa_{\mathrm{grid}} T)(d_x + M_2)}. \tag{77}$$

All score expectations use forward marginals and a fresh Y, conditionally on the realized training data and initialization.

Proof. If $\mathcal{S}(\bar{s}) = \infty$ the claim is immediate. Otherwise fix y outside the null set where either the conditional score sum or the conditional second moment is infinite. Apply Benton et al. (2024, Appendix C, Theorem 2), using its Lemma 8 and Proposition 3 for the path-space bound with exact initialization. The version is arXiv:2308.03686v3, 6 March 2024. Define the reverse clock by

$$\tau_k = T - t_{M-k}, \quad 0 \leq k \leq M, \qquad \gamma_k^{\mathrm{B}} = \tau_{k+1} - \tau_k = \Delta_{M-k-1}. \tag{78}$$

The source's grid is τ_k, step count is M, early-stopping parameter is our t_0, dimension is d_x, and data law is $p_0(\cdot \mid y)$. For $k = M - j - 1$, $T - \tau_k = t_{j+1}$ and $T - \tau_{k+1} = t_j$, so

$$\gamma_k^{\mathrm{B}} = \Delta_j \leq \kappa_{\mathrm{grid}} \min\{1, T - \tau_{k+1}\}.$$

Under the true reverse path law Q_y, the state V_{τ_k} has law $p_{T-\tau_k}(\cdot \mid y) = p_{t_{j+1}}(\cdot \mid y)$. Thus the source's discrete score-error term, written under Q_y as in its Proposition 3 and the proof of Lemma 8, is exactly the conditional version $\mathcal{S}_y(\bar{s})$ of the displayed sum. Its general-moment path-space estimate is

$$\mathrm{KL}(Q_y \| P_y^{p_T}) \leq C\{\mathcal{S}_y(\bar{s}) + \kappa_{\mathrm{grid}}^2 d_x M + \kappa_{\mathrm{grid}} d_x T + \kappa_{\mathrm{grid}} \mathbb{E}[\|X_0\|_2^2 \mid Y = y]\}.$$

This is an imported discretization theorem, including its localized Girsanov argument; we do not impose an unproved Novikov condition. The source's Eq. (2) and Appendix F fix the $\sqrt{2}$ diffusion normalization, and its Proposition 3 fixes the reverse-law expectation; these conventions avoid the inconsistent coefficient and marginal subscripts in its displayed Eqs. (4) and (7).

Data processing and Pinsker give the conditional endpoint TV bound. Average over Y, use Jensen's inequality and the uniform moment bound from Lemma D.1. This proves (76). The initialization difference is bounded by Lemma C.1 and Markov-kernel contraction. Finally $T \geq 1$ bounds the three discretization terms by $(\kappa_{\mathrm{grid}}^2 M + \kappa_{\mathrm{grid}} T)(d_x + M_2)$. □

C.3 Regression-to-score interface

The main algorithm uses the score-weighted law

$$c(t) = \frac{\alpha_t^2}{\sigma_t^4}, \qquad \mathcal{W} = \sum_{j=0}^{M-1} \Delta_j c(t_{j+1}), \qquad \pi_j = \frac{\Delta_j c(t_{j+1})}{\mathcal{W}}, \qquad U = (X_{t_{J+1}}, Y, t_{J+1}). \tag{79}$$

Here J is independent of the clean pair and the Gaussian augmentation. Because the final coordinate identifies J, changing positive grid probabilities changes the design law but does not change the conditional target at a specified time slice.

Lemma C.3 (Exact weighted regression-to-score transfer). *For any square-integrable predictor f, define $s_f(x, t, y) = (\alpha_t f(x, y, t) - x)/\sigma_t^2$. Then, under (79),*

$$\sum_{j=0}^{M-1} \Delta_j \mathbb{E} \left\| s_f(X_{t_{j+1}}, t_{j+1}, Y) - s^\star(X_{t_{j+1}}, t_{j+1}, Y) \right\|_2^2 = \mathcal{W}\, \mathbb{E} \left\| f(U) - f^\star(U) \right\|_2^2, \tag{80}$$

and

$$0 < \mathcal{W} \leq \frac{1}{2} \left[\frac{1}{1 - e^{-2t_0}} - \frac{1}{1 - e^{-2T}} \right] \leq \frac{1}{4t_0}. \tag{81}$$

Proof. By Lemma D.2, $\|s_f - s^\star\|_2^2 = c(t)\,\|f - f^\star\|_2^2$. Sum with weights Δ_j and substitute the probabilities (79). Since $c'(t) < 0$, the right-endpoint sum satisfies

$$\mathcal{W} \le \int_{t_0}^{T} c(t)dt = \frac{1}{2}\left[\frac{1}{1-e^{-2t_0}} - \frac{1}{1-e^{-2T}}\right] \le \frac{1}{2(e^{2t_0}-1)} \le \frac{1}{4t_0}.$$

□

Interval-length variant. If one instead chooses $\pi_j^{\mathrm{u}} = \Delta_j/(T - t_0)$ at the same right endpoints, the exact identity concerns score risk and gives $\mathcal{S}(s_f) = (T - t_0)\mathbb{E}_{\mathrm{u}}\,\|s_f - s^\star\|_2^2 \le (T - t_0)c(t_0)\mathbb{E}_{\mathrm{u}}\,\|f - f^\star\|_2^2$. This is a distinct training law; the main algorithm and calibrations use (79).

C.4 Clipping, iid augmentation, and localization

Lemma C.4 (Clipping yields centered sub-Gaussian noise)**.** *Let* $\xi_{B_N,q} = Z_{B_N,q} - f^\star_{B_N,q}(U)$. *Uniformly over any independent choice of positive time probabilities,*

$$\mathbb{E}[\xi_{B_N,q} \mid U] = 0, \qquad \mathbb{E}[e^{v\xi_{B_N,q}/B_N} \mid U] \le \exp(v^2/2), \tag{82}$$

and

$$\sum_{q=1}^{d_x} \mathbb{E}|f^\star_{B_N,q}(U) - f^\star_q(U)|^2 \le Ce^{-cB_N}. \tag{83}$$

Consequently,

$$\mathcal{S}(s_f) \le 2\mathcal{W}\left\{\mathbb{E}\left\|f(U) - f^\star_{B_N}(U)\right\|_2^2 + Ce^{-cB_N}\right\}. \tag{84}$$

Proof. Conditional centering is the definition of $f^\star_{B_N,q}$. Hoeffding's lemma applies to $Z_{B_N,q}/B_N \in [-1, 1]$. Conditional Jensen and the tower property yield

$$\mathbb{E}\left\|f^\star_{B_N}(U) - f^\star(U)\right\|_2^2 \le \mathbb{E}\,\|Z_{B_N} - X_0\|_2^2 \le \mathbb{E}[\|X_0\|_2^2\,\mathbf{1}_{\{\|X_0\|_2 > B_N\}}].$$

The marginal law of X_0 is unchanged by independent time sampling. The exponential moment in Assumption 3.1 and $r^2\mathbf{1}_{\{r>B_N\}} \le C_{\lambda_X}e^{\lambda_X r}e^{-\lambda_X B_N/2}$ give (83). The squared-triangle inequality with (80) proves (84). The exact identity has the true target $f^\star$; replacing it by the clipped target requires this explicit bias term. □

The tuples $(X_{0,i}, Y_i, J_i, \varepsilon_i)$ are iid, and the pairs $(U_i, Z_{B_N,i})$ are measurable images of these tuples. Drawing each augmentation once and freezing it throughout training therefore preserves an iid regression sample. The injective sphericalization preserves the conditional expectation and the noise bound (82). No observation at every time slice and no lower bound on $\min_j \pi_j$ is needed for these claims or for the distribution-free oracle below.

Lemma C.5 (Covariate localization)**.** *There exist* $c, C > 0$*, uniformly over the independent time probabilities, such that, for* $\rho \ge C(1 + T)$*,*

$$\mathbb{P}(\|U\|_2 > \rho) \le Ce^{-c(\rho - T)}. \tag{85}$$

Thus

$$\rho_N = C(1 + T) + C\log\frac{CN}{\zeta_{\mathrm{loc}}} \tag{86}$$

implies $\mathbb{P}(\max_{i\le N}\|U_i\|_2 \le \rho_N) \ge 1 - \zeta_{\mathrm{loc}}$.

Proof. For every possible $t_{J+1} \le T$,

$$\|U\|_2 \le \|X_0\|_2 + \|\varepsilon\|_2 + \|Y\|_2 + T.$$

Apply a union bound with the exponential tails of X_0, Y and the Gaussian norm tail. Another union bound over the observations gives the sample statement. The population learning theorem uses the full design law; it does not replace it by a conditional truncated law. □

The choices in Section B absorb both (83) and $B_N^2\mathbb{P}(\|U\|_2 > \rho_N)$ into the stated risk envelope. The population localization remainder is retained until that deterministic choice is made.

C.5 LIMITING-KERNEL LEARNING

Throughout this subsection use $0 < \eta \le 1/(4L)$ for the full NTK with diagonal $K(x,x) = L$.

Use the kernel-GD notation from Section E.5: $K = K_L^{\text{full}}$, $\mathcal{H} = \mathcal{H}_L^{\text{full}}$, $T = S^*S$, $T_N = S_N^* S_N$, and $g_r = q_r(T_N) S_N^* Y_N$, with $\lambda_r = (\eta r)^{-1}$. The following consequence states exactly which oracle is used for the bounded heat comparators.

Corollary C.6 (Bounded-comparator early-stopping oracle)**.** *Assume the observation and step-size hypotheses of Theorem E.24, and let $\mathcal{G} \subset \mathcal{H}$ be a finite deterministic family with $\sup_{h\in\mathcal{G}} \|h\|_\infty \le M_{\mathcal{G}} < \infty$. Under its regularized-sample condition, with probability at least $1-\delta$,*

$$\|Sg_r - f\|^2_{L^2(P_X)} \le C \min_{h\in\mathcal{G}} \left\{ \|Sh - f\|^2_{L^2(P_X)} + \lambda_r \|h\|^2_{\mathcal{H}} \right\} + \frac{C}{N}\left[(1+M_{\mathcal{G}})^2 \log\frac{3|\mathcal{G}|}{\delta} + \sigma^2 \left(C_{L,d_0} \lambda_r^{-\gamma} + \log\frac{3}{\delta} \right) \right], \qquad \gamma = \frac{D+1}{D+2}. \tag{87}$$

Proof. This is Theorem E.24 with $M = M_{\mathcal{G}}$. Its proof separately allocates the operator, misspecification, and conditional-noise failure probabilities; the empirical effective dimension is bounded deterministically by Proposition E.19. Thus no empirical Gram lower bound or unbounded-comparator absorption is used here. □

The approximation and finite-stopping terms remain distinct inside (87). The comparator construction in Section E.3 satisfies its uniform range hypothesis, including in the saturated regime.

Theorem C.7 (General-smoothness full-NTK learning rate)**.** *Suppose Assumptions 3.1 and 3.2 hold at integer order $k \ge 1$. Let $s = (D+2)/2$, $\bar{k} = \min\{k, s\}$, and fix $1 \le u \le \bar{k}$ with $u < s$. Use the extension at integer order $j = \lceil u \rceil \le k$ in the comparator construction of Theorem E.17. Put $A = A_{u,N}(t_0) \ge 1$, $\gamma = (D+1)/(D+2)$, and $\mathcal{L}_N = \log(CNd_x/\delta)$. The actual score-weighted design obeys*

$$\mathsf{df}_{P_X}(\lambda) \le C_{L,d_0} \lambda^{-\gamma}, \qquad 0 < \lambda \le 1. \tag{88}$$

Use the constrained stopping scale in (44), with $b = u/s$, and take $R_\star = \lceil (\eta\lambda_N)^{-1} \rceil$. Assume N is large enough that $\lambda_N \le 1$. With probability at least $1-\delta$,

$$\sum_{q=1}^{d_x} \left\| B_N g^{\infty}_{q,R_\star} \circ \psi_{\rho_N} - f^\star_{B_N,q} \right\|^2_{L^2(P_U)} \le C d_x B_N^2 \left[A^{\frac{2\gamma s}{u+\gamma s}} \left(\frac{\mathcal{L}_N}{N} \right)^{\frac{u}{u+\gamma s}} + A^2 \left(\frac{\mathcal{L}_N}{N} \right)^{u/s} + \mathbb{P}(\|U\|_2 > \rho_N) \right]. \tag{89}$$

For fixed $t_0 > 0$ and the stated logarithmic choices of T, B_N, ρ_N, this is $\widetilde{O}(N^{-r_u}) = \widetilde{O}(N^{-2u/(2u+D+1)})$, where r_u is defined in (40).

Proof. Equation (88) is Proposition E.19; it applies directly to P_X. For any deterministic $0 < \lambda \le 1$, choose a single comparator per output coordinate from Theorem E.17, with

$$R(\lambda) = CA^2 \lambda^{-(s-u)/s}, \qquad \mathcal{G}_q = \{h_{q,R(\lambda)}\}. \tag{90}$$

Its squared approximation error after normalization is at most $CA^2\lambda^{u/s} + C\mathbb{P}(\|U\|_2 > \rho_N)$, and its stopping bias satisfies

$$\lambda \left\| h_{q,R(\lambda)} \right\|^2_{\mathcal{H}} \le CA^2 \lambda^{u/s}.$$

The bounded linear extension and the finite sum of heat contractions in Lemma E.15 give a uniform range bound $\left\| h_{q,R(\lambda)} \right\|_\infty \le M_u$, independent of N, λ, t_0. The normalized labels have conditional sub-Gaussian proxy one by Lemma C.4. Apply Corollary C.6 with failure probability δ/d_x and sum over outputs. The resulting rate-critical bound is

$$C d_x B_N^2 \left\{ A^2 \lambda^{u/s} + \frac{\mathcal{L}_N}{N\lambda^\gamma} + \mathbb{P}(\|U\|_2 > \rho_N) \right\}.$$

For $R_\star = \lceil (\eta\lambda_N)^{-1} \rceil$, the actual filter scale lies between $\lambda_N/2$ and λ_N. The fixed constant in the sample floor is enlarged to retain the regularized-sample condition at that actual scale. Substituting

$$\lambda_N = \max\left\{ \left(\frac{\mathcal{L}_N}{NA^2} \right)^{s/(u+\gamma s)}, C_0 \frac{\mathcal{L}_N}{N} \right\}$$

and using

$$(\max\{a_1, a_2\})^{u/s} \le a_1^{u/s} + a_2^{u/s}, \qquad a_1, a_2 \ge 0,$$

gives (89). At fixed t_0, A is polylogarithmic and the floor term decays faster. The tail is absorbed only after its displayed localization calibration. □

For the main theorem, apply the preceding result with $u = \bar{k}$ when $\bar{k} < s$. When $k \ge \lceil s \rceil$, we have $\bar{k} = s$; the following proposition supplies the corresponding saturated bound.

Proposition C.8 (RKHS-membership saturation with localization)**.** *Suppose the bounded global extension $F_{N,q}$ lies in $H^s(\mathbb{S}^p) \simeq \mathcal{H}_L^{\text{full}}$ and $\|F_{N,q}\|_{\mathcal{H}} \le A = A_{s,N}(t_0)$, for every output coordinate. Use*

$$\lambda_N = \max\left\{ \left(\frac{\mathcal{L}_N}{NA^2} \right)^{1/(1+\gamma)}, C_0 \frac{\mathcal{L}_N}{N} \right\}, \qquad \gamma = \frac{D+1}{D+2}, \tag{91}$$

with N large enough that $\lambda_N \le 1$ and $R_\star = \lceil (\eta\lambda_N)^{-1} \rceil$. With probability at least $1-\delta$,

$$\sum_{q=1}^{d_x} \left\| B_N g^{\infty}_{q,R_\star} \circ \psi_{\rho_N} - f^{\star}_{B_N,q} \right\|^2_{L^2(P_U)} \le C d_x B_N^2 \left[A^{2\gamma/(1+\gamma)} \left(\frac{\mathcal{L}_N}{N} \right)^{1/(1+\gamma)} + A^2 \frac{\mathcal{L}_N}{N} + \mathbb{P}(\|U\|_2 > \rho_N) \right]. \tag{92}$$

At fixed t_0, the squared-risk exponent is $1/(1+\gamma)$ and the learned-score contribution to TV is $\widetilde{O}(N^{-1/[2(1+\gamma)]})$.

Proof. Use the singleton comparator family $\{F_{N,q}\}$ in Corollary C.6. The extension is uniformly bounded and agrees with $B_N^{-1} f^{\star}_{B_N,q} \circ \psi_{\rho_N}^{-1}$ on the localized region. Therefore its squared population approximation error is at most $C\mathbb{P}(\|U\|_2 > \rho_N)$; it need not vanish globally. The remaining terms are $CA^2\lambda + C\mathcal{L}_N/(N\lambda^\gamma)$. Substitute (91), account for the constant factor between λ_N and the actual filter scale, sum over coordinates, and restore the factor B_N^2. The floor and tail terms remain explicit until the fixed-t_0 or original-law calibration is made. Finally use Lemma C.3 for the TV statement. □

C.6 Finite-width kernel approximation

Corollary C.9 (Finite-to-infinite full-NTK coupling)**.** *Under Theorem F.10, suppose $\epsilon_K \le 1$, $\eta \le 1/(4L)$, and let $h = \eta R$. For zero-start trajectories with the finite initialization kernel and K_L^{full} on the same sample and labels $|Z_i| \le 1$,*

$$\sup_{r \le R} \left\| g_r^K - g_r^\infty \right\|^2_{L^2(P_X)} \le C_{L,\kappa_{\text{pyr,N}}} \nu_m (h + h^2)^2 m^{-1/3}. \tag{93}$$

Consequently, in addition to the initialization/concentration hypotheses,

$$m \ge C_{L,\kappa_{\text{pyr,N}}} \nu_m^3 (h + h^2)^6 \varepsilon_{\text{coup},N}^{-6} \tag{94}$$

implies

$$\sup_{r \le R} \left\| g_r^K - g_r^\infty \right\|_{L^2(P_X)} \le \varepsilon_{\text{coup},N}. \tag{95}$$

Both conclusions also hold in the supremum norm on the sphere.

Proof. Apply Lemma F.11 with $K_1 = K_{\theta_0}$, $K_2 = K_L^{\text{full}}$, and diagonal bound $M = 2L$. The choice $\eta \le 1/(4L)$ ensures contraction for both empirical kernel updates. Insert (242) to obtain

$$\sup_{r \le R} \left\| g_r^K - g_r^\infty \right\|_{L^\infty} \le C_{L,\kappa_{\text{pyr,N}}}(h + h^2)\sqrt{\nu_m}\, m^{-1/6}.$$

Squaring and integrating proves (93); taking sixth powers of the tolerance condition proves (94). □

The direct trajectory comparison is necessary because a pointwise kernel gap does not place the full-NTK comparator in the finite-rank RKHS.

C.7 Finite nonlinear training

Proposition C.10 (Horizon-calibrated nonlinear-to-kernel coupling)**.** *Assume the hypotheses and events of Propositions F.13 and F.14, and put*

$$h = \eta R_\star, \qquad \Xi_N = \eta R_\star + (\eta R_\star)^2 = h + h^2. \tag{96}$$

With $q_\star$ from (256), simultaneously for all heads,

$$\sup_{r \le R_\star} \left\| F_{\theta_r}^c - g_r^K \right\|_{L^2(P_X)} \le C_{L,\kappa_{\text{pyr,N}}} q_\star \Xi_N. \tag{97}$$

This holds for any input law P_X supported on the sphere. In particular, if

$$q_\star \Xi_N \le c_{L,\kappa_{\text{pyr,N}}} \varepsilon_{\text{coup},N}, \tag{98}$$

then

$$\sup_{r \le R_\star} \left\| F_{\theta_r}^c - g_r^K \right\|_{L^2(P_X)} \le \varepsilon_{\text{coup},N}. \tag{99}$$

In addition to the movement and initialization hypotheses, a sufficient width for this tolerance is

$$m \ge C_{L,\kappa_{\text{pyr,N}}} \nu_m^3 (1 + \eta R_\star) \Xi_N^6 \varepsilon_{\text{coup},N}^{-6}. \tag{100}$$

Proof. By Propositions F.7 and F.13, take $B_0 \le 2$ and $\kappa_{0,P} \le C_{L,\kappa_{\text{pyr,N}}}$ in Proposition F.3. These are the envelope and residual constants for the averaged population operator, with no factor of $\sqrt{N}$. The drift and remainder bounds (258)–(261) give

$$\frac{\eta \delta_K}{N} + \tau_X \le C\eta q_\star, \qquad \delta_{k,P} \le C q_\star, \qquad \tau_P \le C\eta q_\star.$$

Thus the sample discrepancy after r steps is at most $C\eta r q_\star$. The exact population recursion (211) sums this sample discrepancy once more and contributes at most $C(\eta R_\star)^2 q_\star$. The cross-kernel drift and population Taylor remainders together contribute at most $C\eta R_\star q_\star$. Equivalently, substitution into (209) gives

$$\sup_{r \le R_\star} \left\| F_{\theta_r}^c - g_r^K \right\|_{L^2(P_X)} \le C q_\star \{\eta R_\star + (\eta R_\star)^2\} = C q_\star \Xi_N.$$

This proves (97) and (99). Substitute the definition of $q_\star$ into (98) and take sixth powers to obtain (100). The argument also works in the supremum norm. Contraction and finite-horizon summation replace any use of a lower Gram eigenvalue or an assumed PL inequality. The discrepancy measures predictions along the actual neural trajectory; it is not an assumed ERM optimization gap. □

Corollary C.11 (A sufficient width for both population comparisons)**.** *Let $0 < \varepsilon_{\text{coup},N} \le 1$, $h = \eta R_\star$, and $\eta \le 1/(4L)$. For a sufficiently large fixed-depth/width-ratio constant, the condition*

$$m \ge C_{L,\kappa_{\text{pyr,N}}} \nu_m^3 (1 + h)^{13} \varepsilon_{\text{coup},N}^{-6} \tag{101}$$

implies movement and makes both (95) and (99) hold through $R_\star$ on the common initialization and local events. The width condition is implicit only through $\log m$ and admits finite solutions for any fixed architecture ratio and horizon.

Proof. Movement requires $m \geq C\nu_m^3(1+h)^7$. The two comparison requirements are respectively $C\nu_m^3\Xi_N^6\varepsilon_{\mathrm{coup},N}^{-6}$ and $C\nu_m^3(1+h)\Xi_N^6\varepsilon_{\mathrm{coup},N}^{-6}$. Since $\Xi_N = h + h^2 \leq (1+h)^2$ and $\varepsilon_{\mathrm{coup},N} \leq 1$, (101) dominates all three, as well as $m \geq C\nu_m^3$ and the initialization/local-width conditions after increasing the constant. The sharper maximum of the separate conditions is the envelope used in Section B. □

Theorem C.12 (Finite-horizon controlled-pyramidal trajectory theorem)**.** *Under the architecture and step-size conditions of Theorem 1, if* $\mathsf{Width}(W_{\mathrm{stat}})$ *holds, then with probability at least* $1-\delta$*, simultaneously for all heads, the movement bound* (252)*, local scales* (257)*–*(261)*, and both population comparisons* (95) *and* (99) *hold through* $R_\star$*.*

Proof. The statistical width dominates (227), (241), (233), (251), (94), and (100). Use a fixed subdivision of δ among initialization, local perturbation, and concentration, including the union over the scalar heads in ν_m. Replacing each event failure parameter by that fixed fraction changes only the constant inside the logarithm and the fixed width constant. Intersect the events of Proposition F.7 and Theorems F.8 and F.10, then apply Propositions C.10, F.13 and F.14 and Corollary C.9 deterministically on that intersection. The union bound gives probability at least $1-\delta$. No empirical minimum-eigenvalue condition enters this branch. □

For the tolerance $\varepsilon_{\mathrm{coup},N} = \min\{1, c\mathfrak{r}_N^{1/2}/(B_N\sqrt{d_x})\}$, squaring each head's two discrepancies, multiplying by B_N^2, and summing over d_x heads gives

$$\mathcal{E}_{\mathrm{opt}} \leq C\mathfrak{r}_N, \qquad \mathcal{E}_{\mathrm{width}} \leq C\mathfrak{r}_N. \tag{102}$$

Combined with the limiting-kernel oracle and weighted sampling reduction, these bounds control the finite trained predictor at the same risk scale.

C.8 Original-law smoothing

Lemma C.13 (OU smoothing bias for the original law)**.** *Under Assumptions 3.1 and 3.3, for sufficiently small* $t \in (0,1)$*,*

$$\mathbb{E}_Y \mathrm{TV}(p_0(\cdot \mid Y), p_t(\cdot \mid Y)) \leq Ct^{\beta/2}\left(1 + \log\frac{C}{t}\right)^{d_x+\beta} + Cd_x t. \tag{103}$$

Proof. Fix y, write $p = p_0(\cdot \mid y)$, $d = d_x$, $\alpha = e^{-t}$, $\sigma = \sqrt{1-e^{-2t}}$, and let $W \sim N(0, I_d)$. The density of $\alpha X_0 + \sigma W$ is $p_t(x) = \alpha^{-d}\mathbb{E}[p((x-\sigma W)/\alpha)]$. For $Q_R = [-R,R]^d$, add and subtract the unscaled expectation. The Hölder condition contributes at most $C[t^\beta R^{d+\beta} + t^{\beta/2}R^d]$ on Q_R. Conditional exponential tails and the Gaussian tail give $\mathrm{Tail}(R,t,y) \leq Ce^{-cR} + Ce^{-cR^2/t}$, while $|1-\alpha^d| \leq Cdt$. Set $R = C_0(1+\log(C/t))$, divide the L^1 bound by two, and average over Y. □

This is the only component using Assumption 3.3. Section B combines it with the stopped-law bound; it does not arise by setting $t_0 = 0$ in Lemma C.3.

D Conditional-Law Regularity and Regression Identities

We derive the moment, regression, and positive-time target-regularity properties used in the sampling and learning bounds. The proofs of kernel learning and finite-network comparison follow in Sections E and F; verifiable conditional families and non-examples are collected in Section H.

D.1 Derived Properties from the Primitive Assumptions

Lemma D.1 (Uniform finite moments)**.** *Under Assumption 3.1, for every finite* $p \geq 1$*,*

$$\operatorname*{ess\,sup}_y \mathbb{E}[\|X_0\|_2^p \mid Y = y] < \infty, \qquad \mathbb{E}\,\|Y\|_2^p < \infty. \tag{104}$$

In particular $M_2 < \infty$*.*

Proof. For every $\lambda > 0$ and $p \geq 1$, $z^p e^{-\lambda z}$ is bounded on $[0, \infty)$. Hence $z^p \leq C_{p,\lambda} e^{\lambda z}$. Apply this to $\|X_0\|_2$ conditionally on Y, and to $\|Y\|_2$ unconditionally. □

Lemma D.2 (Conditional clean-sample prediction identity)**.** *For every $t > 0$ and every (x, y) with $p_t(x \mid y) > 0$,*

$$s^\star(x, t, y) = \frac{\alpha_t f^\star(x, y, t) - x}{\sigma_t^2}. \tag{105}$$

Proof. Condition on $Y = y$. Differentiating

$$p_t(x \mid y) = \int \varphi_{\sigma_t}(x - \alpha_t z) \mu_y(\mathrm{d}z)$$

under the integral gives

$$\nabla_x p_t(x \mid y) = -\sigma_t^{-2} \int (x - \alpha_t z) \varphi_{\sigma_t}(x - \alpha_t z) \mu_y(\mathrm{d}z).$$

Divide by $p_t(x \mid y)$ and recognize the posterior mean. □

Lemma D.3 (Regression projection identity)**.** *For every square-integrable measurable f,*

$$\mathbb{E}\,\|f(U) - X_0\|_2^2 - \mathbb{E}\,\|f^\star(U) - X_0\|_2^2 = \mathbb{E}\,\|f(U) - f^\star(U)\|_2^2. \tag{106}$$

Proof. Expand $f(U) - X_0 = (f(U) - f^\star(U)) + (f^\star(U) - X_0)$; the cross term has conditional expectation zero given U. □

D.2 Deriving Higher-Order Target Regularity from the Conditional Law

We derive the localized target envelope from Assumptions 3.1 and 3.2. The conditional-law version is defined on all conditioning values used in the extension; no conditioning density is divided out.

D.2.1 The conditioning domain and the tail quantifier

Lemma D.4 (Full-support equivalence of the moment bounds)**.** *Let $y \mapsto \mu_y$ be a weakly continuous conditional-law version on $\mathbb{R}^{d_y}$, let $S = \mathrm{supp}(P_Y)$, and suppose*

$$\operatorname*{ess\,sup}_{y \sim P_Y} \int e^{\lambda \|z\|} \mu_y(\mathrm{d}z) \leq M < \infty.$$

Then $\int e^{\lambda \|z\|} \mu_y(\mathrm{d}z) \leq M$ for every $y \in S$. If $S = \mathbb{R}^{d_y}$, this is equivalent to the all-y moment bound in Assumption 3.1. In particular the equivalence applies under Assumption 3.2 whenever Y has full topological support, including a nondegenerate Gaussian law.

Proof. For each positive integer m, the function $\varphi_m(z) = \min\{e^{\lambda \|z\|}, m\}$ is bounded and continuous, so $g_m(y) = \int \varphi_m(z) \mu_y(\mathrm{d}z)$ is continuous. The essential bound provides a P_Y-full-measure set on which $g_m \leq M$. This set is dense in S: every neighborhood of a point of S has positive P_Y measure. Continuity gives $g_m \leq M$ throughout S. Monotone convergence as $m \to \infty$ proves the assertion. The reverse implication from a pointwise bound to an essential bound is immediate. Weighted-TV continuity implies weak continuity because bounded test functions are continuous linear functionals on the weighted measure space. □

For compact or gapped support this argument does not control the chosen version outside S. For example, $Y \sim \mathrm{Unif}[-1, 1]$ and the smooth version $\mu_y = N(y, 1)$ satisfy the essential moment condition, whereas that particular version has no all-y uniform exponential moment. The all-y condition supplies the full-neighborhood control needed below; a different extension from the conditioning support would require its own norm estimates.

D.2.2 WEIGHTED VARIATION AND POSTERIOR MOMENTS

For y, y' in a radius-R ball, the first-order part of Assumption 3.2 implies

$$\|\mu_y - \mu_{y'}\|_{\mathrm{TV},1} := \sup_{|g(z)|\le 1+\|z\|_2} \left| \int g \, \mathrm{d}(\mu_y - \mu_{y'}) \right| \le L_\mu(R) \, \|y - y'\|_2 \,, \tag{107}$$

where $L_\mu(R)$ grows polynomially. Indeed, along $y_s = y + s(y' - y)$,

$$\int g \, \mathrm{d}(\mu_{y'} - \mu_y) = \int_0^1 (y' - y)^\top \mathbb{E}_{\mu_{y_s}}[g(Z) S_e(Z, y_s)] \, \mathrm{d}s,$$

and conditional exponential tails uniformly control every polynomial moment.

Set $\alpha_t = e^{-t}$ and $v_t = \sigma_t^2 = 1 - e^{-2t}$, and define the Gaussian kernel and posterior by

$$\begin{aligned} K_t(x, z) &= (2\pi v_t)^{-d_x/2} \exp\left(-\frac{\|x - \alpha_t z\|^2}{2v_t} \right), \\ \Pi_{x,y,t}(dz) &= \frac{K_t(x, z)\mu_y(dz)}{\mathcal{D}(x, y, t)}, \qquad \mathcal{D}(x, y, t) = \int K_t(x, z)\mu_y(dz). \end{aligned} \tag{108}$$

Here $\mathcal{D}(x, y, t) > 0$ for every x, y and $t > 0$. For $B \ge 1$, a coordinate $q \le d_x$, and $h_{B,q}(z) = B^{-1} \operatorname{clip}(z_q, [-B, B])$, write

$$g_{B,q}(x, y, t) = \int h_{B,q}(z)\Pi_{x,y,t}(dz) = B^{-1} f^\star_{B,q}(x, y, t). \tag{109}$$

In particular, $|g_{B,q}| \le 1$ for every positive time.

Lemma D.5 (Posterior moments throughout the conditioning domain)**.** *Under Assumptions 3.1 and 3.2, let $t \ge t_0 > 0$, $R \ge 1$, and $\|x\| \le R$. For every real $q \ge 0$,*

$$\sup_y \int (1 + \|z\|)^q \Pi_{x,y,t}(dz) \le C_q (1 + R)^{2q} (1 + t_0^{-1})^q. \tag{110}$$

The bound is independent of any upper time horizon.

Proof. The assertion for $q = 0$ is immediate. Put $r_0 = \lambda_X^{-1} \log(2M_X)$. By Markov's inequality and (7), $\mu_y(\|z\| \le r_0) \ge 1/2$ for every y. Because $\alpha_t \le 1$ and $v_t \ge v_{t_0}$,

$$\mathcal{D}(x, y, t) \ge \tfrac{1}{2}(2\pi v_t)^{-d_x/2} \exp\left(-\frac{(R + r_0)^2}{2v_{t_0}} \right).$$

On the other hand, the numerator defining $\Pi_{x,y,t}(\|z\| > u)$ is at most $(2\pi v_t)^{-d_x/2} M_X e^{-\lambda_X u}$. Hence, with $a = \log(2M_X) + (R + r_0)^2/(2v_{t_0})$,

$$\Pi_{x,y,t}(\|z\| > u) \le \min\{1, e^{a - \lambda_X u}\}.$$

Split the tail integral for $(1 + \|z\|)^q$ at $u = a/\lambda_X$. The part below that point is at most $(1 + a/\lambda_X)^q$; after translation, the remaining integral is bounded by $C_q(1 + a)^q \int_0^\infty (1 + w)^{\max\{q-1,0\}} e^{-\lambda_X w} \, dw$. Thus the moment is at most $C_q(1 + a)^q$. Finally, $v_{t_0}^{-1} \le C(1 + t_0^{-1})$, proving the claim. □

D.2.3 GAUSSIAN DERIVATIVES AND QUOTIENT INDUCTION

Lemma D.6 (Derivatives of the unnormalized integrals)**.** *Let $Q_0 = 1 + t_0^{-1}$ and $M_k = \max\{4, b_k\}$. For a multi-index $\nu = (a, \beta, j)$ in (x, y, t) with $1 \le |\nu| = |a| + |\beta| + j \le k$, put $N_{B,q}(x, y, t) = \int h_{B,q}(z) K_t(x, z)\mu_y(dz)$. Uniformly for $\|x\| + \|y\| \le R$, $t \ge t_0 > 0$, and $B \ge 1$,*

$$\frac{|\partial^\nu \mathcal{D}| + |\partial^\nu N_{B,q}|}{\mathcal{D}} \le C_k (1 + R)^{c_k} Q_0^{M_k|\nu|}. \tag{111}$$

Proof. We first record an elementary Gaussian derivative bound. For $|a| + j \le k$,

$$\frac{|\partial_x^a \partial_t^j K_t(x,z)|}{K_t(x,z)} \le C_k(1+R)^{c_k} Q_0^{|a|+2j}(1+\|z\|)^{|a|+2j}, \qquad \|x\| \le R, \quad t \ge t_0. \tag{112}$$

For completeness, put $w_t = v_t^{-1}$. Then $\alpha_t' = -\alpha_t$, $w_t' = -2\alpha_t^2 w_t^2$, and

$$\partial_{x_i} \log K_t = -w_t(x_i - \alpha_t z_i), \qquad \partial_t \log K_t = -d_x \alpha_t^2 w_t - \alpha_t w_t (x - \alpha_t z)^\top z + \alpha_t^2 w_t^2 \|x - \alpha_t z\|^2 .$$

Starting with $K_t/K_t = 1$, the identities $\partial_i(K_t P)/K_t = \partial_i P + P \partial_i \log K_t$ show inductively that the derivative ratio is a polynomial in x, z, α_t, w_t. Its degrees in z and x, and its largest exponent of w_t, are each at most $|a| + 2j$. A time derivative of a coefficient raises the exponent of w_t by at most one; multiplication by $\partial_t \log K_t$ raises it and the spatial degrees by at most two. An x derivative raises the corresponding bounds by at most one. Since $0 < \alpha_t \le 1$ and $1 \le w_t \le C Q_0$, this proves (112).

Set $S_0 = 1$, $b_0 = a_0 = 0$. Differentiation of the integrals gives

$$\partial_x^a \partial_y^\beta \partial_t^j \mathcal{D} = \int (\partial_x^a \partial_t^j K_t)(x,z) S_\beta(z,y) \mu_y(dz),$$

and the same formula for $N_{B,q}$ with the additional factor $h_{B,q}$. These identities are justified locally on every finite positive-time rectangle: all Gaussian derivatives appearing as test functions are bounded in z there, and depend continuously on (x,t) in the bounded test-function norm. The assumed C^k differentiation of μ_y in weighted total variation therefore permits the y derivatives; the Gaussian derivative bounds permit the remaining derivatives under the integral. Equivalently, one may first take the y derivatives and then differentiate their finite signed-measure integrals. The polynomial relative-score bounds and the uniform exponential moment supply the needed dominating functions on compact sets.

Divide the displayed integral by $\mathcal{D}$, use $|h_{B,q}| \le 1$, and apply (112) and Lemma D.5 with moment order $|a| + 2j + b_\beta$. The resulting power of Q_0 is $2|a| + 4j + b_\beta$. If $\beta \ne 0$, then $b_\beta \le b_k \le M_k|\beta|$; if $\beta = 0$, its contribution is zero. Consequently this power is at most $M_k(|a| + j + |\beta|)$. All powers of $1 + R$ can be bounded by one fixed power depending on k and the finitely many relative-score parameters. This proves (111). □

Proposition D.7 (Higher-order target envelope)**.** *Under Assumptions 3.1 and 3.2, $g_{B,q}$ is C^k on $\mathbb{R}^{d_x+d_y} \times (0,\infty)$ and, for every $R, B \ge 1$ and $T \ge t_0 > 0$,*

$$\max_{|\nu| \le k} \sup_{\substack{\|x\|+\|y\| \le R \\ t \in [t_0, T]}} |\partial^\nu g_{B,q}(x,y,t)| \le C_k(1+R)^{c_k}(1 + t_0^{-1})^{\chi_k}, \qquad \chi_k = k \max\{4, b_k\}. \tag{113}$$

Proof. The differentiability follows from the preceding integral identities and $\mathcal{D} > 0$. We prove the sharper order-by-order statement with exponent $M_k|\nu|$ by induction. At order zero, $|g_{B,q}| \le 1$. Differentiating $\mathcal{D} g_{B,q} = N_{B,q}$ gives, for $|\nu| \ge 1$,

$$\partial^\nu g_{B,q} = \frac{\partial^\nu N_{B,q}}{\mathcal{D}} - \sum_{0 < \omega \le \nu} \binom{\nu}{\omega} \frac{\partial^\omega \mathcal{D}}{\mathcal{D}} \, \partial^{\nu-\omega} g_{B,q}. \tag{114}$$

Each product in the sum has power at most $M_k|\omega| + M_k|\nu - \omega| = M_k|\nu|$ by Lemma D.6 and the induction hypothesis. Its polynomial factor in R also has a bounded exponent: only finitely many derivatives through order k occur. Increasing C_k, c_k once after the induction proves the claimed bound. No derivatives of the relative scores S_β are assumed or used. □

E Kernel Approximation and Population Learning

These results bound the limiting-kernel prediction error. We establish the full-NTK recursion and spectrum, construct bounded approximants from the conditional-law regularity, control capacity under arbitrary input laws, and prove the early-stopping oracle used in Sections B and C.

E.1 Kernel Preliminaries and Exact Full-NTK Recursion

E.1.1 Shallow kernel diagonal

Lemma E.1 (Shallow diagonal and baseline capacity)**.** *For*

$$K_{\mathrm{gate}}(x,x') = (x^\top x')\mathbb{P}_{w\sim N(0,I)}(w^\top x \ge 0, w^\top x' \ge 0), \tag{115}$$

one has $K_{\mathrm{gate}}(x,x) = 1/2$ on the unit sphere. Consequently, its population operator T satisfies

$$\mathrm{tr}(T) = 1/2, \qquad \mathsf{df}(\lambda) = \mathrm{tr}[T(T+\lambda I)^{-1}] \le \frac{1}{2\lambda}. \tag{116}$$

Proof. By symmetry, $\mathbb{P}(w^\top x \ge 0) = 1/2$. For a positive trace-class integral operator, $\mathrm{tr}(T) = \int K(x,x)\,\mathrm{d}P_X(x) = 1/2$. If $\{\mu_j\}$ are its nonnegative eigenvalues, then $\sum_j \mu_j/(\mu_j+\lambda) \le \lambda^{-1}\sum_j \mu_j$. □

E.1.2 Finite-width full-NTK blocks

For $\ell \le H = L-1$, define

$$D_\ell(\theta,x) = \mathrm{diag}\big(\mathbf{1}\{W_\ell \bar{h}_{\ell-1}(x) > 0\}\big) \tag{117}$$

and the normalized backward row vector

$$b_\ell(\theta,x) = W_L \prod_{r=H}^{\ell+1}\left(\sqrt{\frac{2}{d_r}} D_r(\theta,x) W_r\right)\sqrt{\frac{2}{d_\ell}} D_\ell(\theta,x), \tag{118}$$

where the empty product is the identity. Then

$$\nabla_{W_\ell} F_\theta(x) = b_\ell(\theta,x)^\top \bar{h}_{\ell-1}(\theta,x)^\top, \qquad \nabla_{W_L} F_\theta(x) = \bar{h}_H(\theta,x)^\top. \tag{119}$$

Thus

$$\begin{aligned} K_\theta(x,x') = &\left\langle \bar{h}_H(x), \bar{h}_H(x')\right\rangle \\ &+ \sum_{\ell=1}^{H} \langle b_\ell(x), b_\ell(x')\rangle \left\langle \bar{h}_{\ell-1}(x), \bar{h}_{\ell-1}(x')\right\rangle. \end{aligned} \tag{120}$$

The first term is the trainable-readout block. Under iid initialization none of the hidden-layer blocks cancels.

Lemma E.2 (Infinite-width covariance recursion)**.** *Fix finitely many inputs on $\mathbb{S}^{d_0-1}$. As every hidden width diverges with bounded adjacent ratios,*

$$\left\langle \bar{h}_r(x), \bar{h}_r(x')\right\rangle \longrightarrow C_r(\langle x,x'\rangle) \tag{121}$$

in probability, jointly over the fixed input set, where $C_0(u) = u$ and $C_r = \mathfrak{a}_1 \circ C_{r-1}$.

Proof. At $r = 0$ the claim is immediate. Conditional on layer $r-1$, the pairs of preactivations are iid centered Gaussian. The conditional law of large numbers and

$$2\mathbb{E}[\sigma(G)\sigma(G')] = \mathfrak{a}_1(\rho) \tag{122}$$

for jointly standard Gaussian (G,G') of correlation ρ yield the covariance recursion. The diagonal recursion gives unit limiting norms, and induction completes the proof. Bounded aspect ratios are not needed for a finite input set; they enter the uniform finite-width bounds later. □

Proposition E.3 (Exact full-NTK recursion)**.** *For the product-Frobenius metric and normalization in (5), the infinite-width initialization NTK is*

$$K_L^{\mathrm{full}}(x,x') = \Theta_H(\langle x,x'\rangle), \qquad H = L-1, \tag{123}$$

where $\Theta_0 = C_0$ and

$$\Theta_r(u) = C_r(u) + \Theta_{r-1}(u)\mathfrak{a}_0(C_{r-1}(u)). \tag{124}$$

Proof. For a linear network, the tangent kernel is $C_0(u) = u$. Suppose a depth-$(r-1)$ network has covariance C_{r-1} and NTK Θ_{r-1}. Adding a ReLU layer produces covariance C_r by Lemma E.2. Its new readout block contributes $C_r(u)$. Every inherited parameter derivative is multiplied by the product of the two ReLU derivatives, whose Gaussian average is

$$2\mathbb{E}[\mathbf{1}\{G > 0\}\mathbf{1}\{G' > 0\}] = \mathfrak{a}_0(C_{r-1}(u)). \tag{125}$$

Therefore inherited tangent blocks contribute $\Theta_{r-1}(u)\mathfrak{a}_0(C_{r-1}(u))$. Adding the readout block proves (124). The unweighted sum follows from the product-Frobenius metric. This factorization is an infinite-width statement. At finite width the ReLU gate and the Gaussian quadratic form involving the same row are dependent. Lemma F.9 and Theorem F.10 bound the resulting mean correction by $C\,\|A\|_{\mathrm{op}} = O(m^{-1})$ under the present normalization, justifying the limiting gate factor above. □

Remark E.4 (Full-NTK diagonal). Since $C_r(1) = \mathfrak{a}_0(1) = 1$,

$$\Theta_r(1) = r + 1, \qquad K_L^{\mathrm{full}}(x, x) = L. \tag{126}$$

Hence $\mathsf{df}(\lambda) \le L/\lambda$ for every population input law.

E.2 Endpoint Analysis, Harmonic Positivity, and Full-NTK Spectrum

Put $c_0 = \sqrt{2}/\pi$ and $a_0 = 2\sqrt{2}/(3\pi)$. As $t \downarrow 0$,

$$\mathfrak{a}_0(1-t) = 1 - c_0 t^{1/2} + O(t^{3/2}), \qquad \mathfrak{a}_1(1-t) = 1 - t + a_0 t^{3/2} + O(t^{5/2}), \tag{127}$$

$$\mathfrak{a}_0(-1+t) = c_0 t^{1/2} + O(t^{3/2}), \qquad \mathfrak{a}_1(-1+t) = a_0 t^{3/2} + O(t^{5/2}). \tag{128}$$

These follow from the Taylor expansion of $\arccos(1-t)$ and $\mathfrak{a}_1' = \mathfrak{a}_0$.

Lemma E.5 (Positive-endpoint expansion)**.** *For every fixed $r \ge 0$,*

$$C_r(1-t) = 1 - t + O(t^{3/2}), \tag{129}$$

and

$$\Theta_r(1-t) = (r+1) - A_r t^{1/2} + O(t), \qquad A_r = c_0 \frac{r(r+1)}{2}. \tag{130}$$

Proof. The covariance expansion follows inductively from (127). Since $\Theta_0(1-t) = 1 - t$, $A_0 = 0$. If $\Theta_{r-1}(1-t) = r - A_{r-1}t^{1/2} + O(t)$, then

$$\mathfrak{a}_0(C_{r-1}(1-t)) = 1 - c_0 t^{1/2} + O(t),$$

and the recursion gives $A_r = A_{r-1} + r c_0$, hence the formula. □

For the negative endpoint define $q_1 = 0$, $q_r = \mathfrak{a}_1(q_{r-1})$ for $r \ge 2$, and $p_r = \mathfrak{a}_0(q_r)$. Then $p_1 = 1/2$ and $p_r \in (1/2, 1)$ for $r \ge 2$.

Lemma E.6 (Negative-endpoint expansion)**.** *For $r = 1$,*

$$\Theta_1(-1+t) = -c_0 t^{1/2} + O(t^{3/2}). \tag{131}$$

For every $r \ge 2$, there is S_r such that

$$\Theta_r(-1+t) = S_r - \beta_r t^{1/2} + O(t), \qquad \beta_r = c_0 \prod_{j=1}^{r-1} p_j. \tag{132}$$

Moreover $0 < \beta_r \le c_0/2$ and $A_r > \beta_r$.

Proof. The $r = 1$ formula follows by inserting (128) into the recursion. Since $C_1(-1+t) = O(t^{3/2})$, analyticity of $\mathfrak{a}_0, \mathfrak{a}_1$ at the interior points reached after the first layer gives, inductively,

$$C_{r-1}(-1+t) = q_{r-1} + O(t^{3/2}), \qquad \mathfrak{a}_0(C_{r-1}(-1+t)) = p_{r-1} + O(t^{3/2}).$$

Thus $\beta_r = \beta_{r-1} p_{r-1}$. The inequalities follow from $p_1 = 1/2$, later $p_j < 1$, and $A_r = c_0 r(r+1)/2$. □

Lemma E.7 (Convergent endpoint expansions and differentiated remainders)**.** *For each fixed $r \geq 0$, each of the functions $C_r(1-z^2)$, $C_r(-1+z^2)$, $\Theta_r(1-z^2)$, and $\Theta_r(-1+z^2)$, initially defined for small $z \geq 0$, extends to a real-analytic function of z near zero. Consequently, if G is any of C_r, Θ_r, then for either endpoint, every integer $M \geq 0$, and every derivative order $j \geq 0$,*

$$\frac{\mathrm{d}^j}{\mathrm{d}t^j}\left[G(\pm(1-t)) - \sum_{n=0}^{M} a_{n,\pm} t^{n/2}\right] = O\left(t^{(M+1)/2-j}\right), \qquad t \downarrow 0, \tag{133}$$

where the coefficients are those of the convergent series in $z = \sqrt{t}$. The constants may depend on r, M, j.

Proof. The identity $\arccos(1-z^2) = 2\arcsin(z/\sqrt{2})$ for small $z \geq 0$ shows that $\mathfrak{a}_0(1-z^2)$ has an analytic extension in z. The same holds for $\mathfrak{a}_1(1-z^2)$, either from its explicit formula or by integrating $\mathfrak{a}_1' = \mathfrak{a}_0$; the square root in that formula is $z\sqrt{2-z^2}$. The identities $\mathfrak{a}_0(-u) = 1 - \mathfrak{a}_0(u)$ and $\mathfrak{a}_1(-u) = \mathfrak{a}_1(u) - u$ give the corresponding assertion at -1.

At the positive endpoint, inductively write $C_{r-1}(1-z^2) = 1 - z^2 R_{r-1}(z)$ with R_{r-1} analytic and $R_{r-1}(0) = 1$; the latter equality follows from (129). The analytic square root of R_{r-1} with value one at zero gives the analytic variable $v = z\sqrt{R_{r-1}(z)}$. Composing the preceding endpoint expansions with v proves analyticity of both $\mathfrak{a}_i(C_{r-1}(1-z^2))$, $i = 0, 1$, and preserves the displayed factorization for C_r. The recursion (124) then proves analyticity of Θ_r. At the negative endpoint the first composition has just been handled; after it, $C_1(-1) = 0$ and every subsequent endpoint value q_r lies in $[0, 1)$, where both $\mathfrak{a}_i$ are ordinarily analytic. Composition and multiplication again complete the induction.

Finally, truncating a convergent power series in z after degree M leaves $z^{M+1}Q(z)$ with Q analytic. Applying $\mathrm{d}/\mathrm{d}t = (2z)^{-1}\,\mathrm{d}/\mathrm{d}z$ j times gives $z^{M+1-2j}Q_j(z)$ with Q_j analytic. This proves the remainder bound; thus no derivative is inferred from a bare $O(t)$ estimate. □

Lemma E.8 (Positive coefficients after two hidden layers)**.** *The Maclaurin series of C_2 has a strictly positive coefficient at every degree. Consequently every C_r and Θ_r with $r \geq 2$ has strictly positive coefficients at every degree.*

Proof. Using $\mathfrak{a}_0(u) = 1/2 + \pi^{-1}\arcsin u$ and $\mathfrak{a}_1' = \mathfrak{a}_0$,

$$\mathfrak{a}_1(u) = \frac{1}{\pi} + \frac{1}{2}u + \sum_{k\geq 0} c_k u^{2k+2}, \qquad c_k > 0. \tag{134}$$

Thus C_1 has positive coefficients in degrees $0, 1$ and every positive even degree. Its square has a positive coefficient at every degree: odd degree $n \geq 3$ contains the product of degrees 1 and $n-1$. Since (134) has a positive quadratic coefficient, $C_2 = \mathfrak{a}_1(C_1)$ has all coefficients positive. Composition with $\mathfrak{a}_1$ preserves this property, and Θ_r contains C_r plus a product with nonnegative coefficients. □

Lemma E.9 (No missing harmonic modes)**.** *For $H \geq 2$, every spherical-harmonic eigenvalue of $\Theta_H(\langle x, x'\rangle)$ is strictly positive.*

Proof. Write $\Theta_H(u) = \sum_{n\geq 0} a_n u^n$ with $a_n > 0$. The monomial kernel $(x^\top x')^n$ is the inner product of symmetric tensor features and has a strictly positive restriction to the degree-n harmonic subspace. Therefore the degree-n eigenvalue of the nonnegative sum is positive. □

Theorem E.10 (Two-sided full-NTK spectrum and Sobolev identification)**.** *For fixed $L \geq 3$,*

$$c_{L,d_0}(1+\ell)^{-d_0} \leq \mu_{L,\ell}^{\mathrm{full}} \leq C_{L,d_0}(1+\ell)^{-d_0}, \qquad \ell \geq 0. \tag{135}$$

Consequently,

$$c\,\|h\|^2_{H^{d_0/2}(\mathbb{S}^{d_0-1})} \leq \|h\|^2_{\mathcal{H}_L^{\mathrm{full}}} \leq C\,\|h\|^2_{H^{d_0/2}(\mathbb{S}^{d_0-1})}. \tag{136}$$

Proof. The normalization matches Bietti & Bach (2020, Eq. (7)) exactly: $C_r = \kappa^{r+1}$ and $\Theta_r = \kappa_{\mathrm{NTK}}^{r+1}$ in their notation, so the kernel here is the unnormalized full NTK with diagonal $H+1=L$. We verify the hypotheses of their full endpoint theorem (Bietti & Bach, 2020, Theorem 7), which also supplies their Corollary 3 for this recursion.

First, the arc-cosine functions are analytic on $(-1,1)$ and every covariance iterate maps this interval into $(-1,1)$, so Θ_H is C^∞ there. By Lemma E.7 with $M=3$,

$$\Theta_H(\pm(1-t)) = P_\pm(t) + c_{\pm,1}t^{1/2} + c_{\pm,2}t^{3/2} + R_\pm(t), \qquad R_\pm^{(j)}(t) = O(t^{2-j}),$$

for every fixed $j \geq 0$, where $P_\pm$ has degree at most one. Thus the finite noninteger exponents are $\nu_1 = 1/2$ and $\nu_2 = 3/2$ (a coefficient may be zero). Taking $\epsilon = 1/4$ meets $0 < \epsilon < \nu_2 - \nu_1$ and gives the required remainder $O(t^{\nu_1+1+\epsilon-j})$ for every derivative. Derivatives with respect to the kernel argument introduce the required factor $(-1)^j$ at the positive endpoint. These checks establish the full theorem's interior smoothness, finite-expansion, and differentiated-remainder hypotheses.

By Lemmas E.5 and E.6, $c_{+,1} = -A_H$ and $c_{-,1} = -\beta_H$. The even and odd asymptotic constants are proportional to $-A_H - \beta_H$ and $-A_H + \beta_H$, respectively; both are nonzero because $H \geq 2$. The resulting eigenvalue order is $\ell^{-d_0-2\nu_1+1} = \ell^{-d_0}$. Positive definiteness ensures that the two nonzero asymptotic constants are positive, and Lemma E.9 handles the finitely many remaining modes.

The degree-ℓ Laplace–Beltrami eigenvalue is $\ell(\ell+d_0-2)$, and $1+\ell(\ell+d_0-2) \asymp (1+\ell)^2$. The RKHS weights are $(\mu_{L,\ell}^{\mathrm{full}})^{-1} \asymp (1+\ell)^{d_0}$, which are the $H^{d_0/2}$ weights, including the constant mode. □

Lemma E.11 (Shallow spectrum on the even subspace)**.** *For $L=2$, the full kernel satisfies*

$$\mu_{2,2j}^{\mathrm{full}} \asymp (1+2j)^{-d_0} \quad (j \geq 0), \qquad \mu_{2,\ell}^{\mathrm{full}} = 0 \quad (\ell \geq 3 \text{ odd}),$$

and its degree-one eigenvalue is positive. In particular, $H_{\mathrm{even}}^{d_0/2}(\mathbb{S}^{d_0-1})$ embeds continuously in $\mathcal{H}_2^{\mathrm{full}}$, with equivalent norms on this even subspace.

Proof. The recursion gives

$$\Theta_1(u) = \mathfrak{a}_1(u) + u\mathfrak{a}_0(u) = u + \frac{\sqrt{1-u^2} + 2u\arcsin u}{\pi}.$$

The second term is even; (134) and the positive arcsine coefficients show that its Maclaurin coefficient at every even degree is strictly positive. Monomial-kernel positivity then gives all even harmonic eigenvalues positive, whereas the linear term has only degree one. The endpoint proof above applies with $A_1 = \beta_1 = c_0$: its even leading constant is nonzero, giving the stated decay. The harmonic description of the RKHS proves the norm claim. □

Remark E.12 (Global spectrum and the choice of depth). Two hidden layers make the full spherical RKHS equal to the unrestricted Sobolev space in Theorem E.10. This global distinction does not establish a shallow approximation obstruction on the actual localized upper cap: Lemma E.16 constructs an even extension agreeing with the target there, with controlled Sobolev norm. The main theorem studies the stated deep, all-layers-trained architecture; it does not assert that depth is necessary for representing its denoising target. The separate optional KMM curvature result also has its own $L \geq 3$ hypothesis.

E.3 Uniform Iterated-Heat Construction of Source-Free Comparators

Let $p = d_0 - 1 = D+1$, $s = d_0/2 = (D+2)/2$, and let $P_t = e^{t\Delta_{\mathbb{S}^p}}$ be the spherical heat semigroup. We use uniform approximation on a localized positive-time domain so that the comparator bound holds for the atomic-time training design. The target is initially defined only at positive physical times. Throughout, k denotes an available integer regularity order, while u denotes an auxiliary real smoothness order.

E.3.1 LINEAR EXTENSION FROM THE POSITIVE-TIME DOMAIN

For $\rho \geq \max\{1, T+1\}$ define

$$G_\rho^+ = \{v = (x, y, t) : \|v\| \leq \rho,\ t \in [t_0, T]\}, \qquad \mathcal{C}_\rho^+ = \psi_\rho(G_\rho^+) \subset \mathbb{S}^p. \tag{137}$$

Choose fixed smooth cutoffs $\theta : \mathbb{R} \to [0,1]$ and $\chi : \mathbb{R}^D \to [0,1]$ such that $\theta(r) = 1$ for $r \leq 1/4$, $\theta(r) = 0$ for $r \geq 1/2$, $\chi(z) = 1$ for $\|z\| \leq 1$, and $\chi(z) = 0$ for $\|z\| \geq 2$. Their derivatives of every fixed order are bounded. For a fixed integer $k \geq 1$, let $a_1, \ldots, a_{k+1}$ be the unique real solution of

$$\sum_{i=1}^{k+1} a_i (-i)^r = 1, \qquad r = 0, \ldots, k. \tag{138}$$

The coefficient matrix is a Vandermonde matrix at distinct points, so it is invertible. In particular, $C_k^{\mathrm{ref}} = \sum_i |a_i|$ is a finite constant depending only on k.

Lemma E.13 (Linear positive-time extension)**.** *Let $T \geq t_0 > 0$ and $\rho \geq \max\{1, T+1\}$. Suppose $g(z,t)$ is C^k for $z \in \mathbb{R}^D$ and $t > 0$, is bounded by one, and its derivatives through order k are bounded by A on $\{\|z\| \leq 2\rho,\ t_0 \leq t \leq T + k + 1\}$. There is a linear construction $g \mapsto H$ such that $H \in C^k(\mathbb{R}^p)$,*

$$H(z,t) = g(z,t) \quad (\|z\| \leq \rho,\ t \in [t_0, T]), \tag{139}$$

$$\|H\|_\infty \leq C_k, \qquad \max_{|\nu| \leq k} \|\partial^\nu H\|_\infty \leq C_k A, \tag{140}$$

$$\operatorname{supp}(H) \subset \{\|z\| \leq 2\rho,\ t_0 - 1 \leq t \leq T + 1\} \subset \{\|(z,t)\| \leq \sqrt{5}\rho\}. \tag{141}$$

The operator constants have no dependence on t_0, $T - t_0$, T, or ρ.

Proof. On the half-space $t \geq t_0$, set

$$H(z,t) = \chi(z/\rho)\theta(t - T)g(z,t).$$

For $t_0 - 1 < t < t_0$, set

$$H(z,t) = \chi(z/\rho)\theta(t_0 - t)\sum_{i=1}^{k+1} a_i g\big(z, t_0 + i(t_0 - t)\big), \tag{142}$$

and put $H = 0$ for $t \leq t_0 - 1$. The cutoffs make the expressions zero in neighborhoods of the outer gluing boundaries. Near $t = t_0$, both time cutoffs are one: $t_0 \leq T$ and θ is one on $(-\infty, 1/4]$. The left derivative of order r in time at t_0 equals

$$\sum_i a_i (-i)^r \partial_t^r g(z, t_0) = \partial_t^r g(z, t_0).$$

The same equality holds after any spatial derivatives with total order at most k. Thus all derivatives through order k match across $t = t_0$.

Every argument of g used in the reflected portion has positive time between t_0 and $t_0 + k + 1 \leq T + k + 1$; in the upper portion the cutoff restricts nonzero terms to $t \leq T + 1$. Each differentiation of a reflected term introduces only powers of the fixed integer index i. Derivatives of the spatial cutoff introduce factors $\rho^{-1} \leq 1$. The product rule therefore proves the C^k bound with constant C_k. The L^∞ bound follows directly from $|g| \leq 1$ and $\sum_i |a_i| < \infty$, independently of the derivative bound A. Matching and support are immediate; on the support, $|t| \leq T + 1 \leq \rho$ and $\|z\| \leq 2\rho$. All operations are linear in g, with range controlled by the fixed reflection coefficients. □

Lemma E.14 (Higher-order cap extension)**.** *Define*

$$\psi_\rho(v) = \frac{(v/\rho, 1)}{\sqrt{1 + \|v\|^2/\rho^2}},$$

and let $g = g_{B,q}$ from (109). Under Assumptions 3.1 and 3.2 at integer order k, for $\rho \geq \max\{1, T+1\}$ there is $F_{B,q,\rho} \in C^k(\mathbb{S}^p)$ such that

$$F_{B,q,\rho}(\psi_\rho(v)) = g_{B,q}(v) \quad (\|v\| \leq \rho,\ t \in [t_0, T]), \tag{143}$$

$$\|F_{B,q,\rho}\|_\infty \le C_k, \tag{144}$$

$$\|F_{B,q,\rho}\|_{C^k(\mathbb{S}^p)} + \|F_{B,q,\rho}\|_{H^k(\mathbb{S}^p)} \le C_k \rho^{c_k}(1+t_0^{-1})^{\chi_k}. \tag{145}$$

Consequently, for every real $0 < u \le k$,

$$\begin{aligned}\|F_{B,q,\rho}\|_{C^u_*(\mathbb{S}^p)} + \|F_{B,q,\rho}\|_{H^u(\mathbb{S}^p)} &\le A_{u,k}(\rho, t_0),\\ A_{u,k}(\rho,t_0) &= C_{k,u}\rho^{c_k}(1+t_0^{-1})^{\chi_k}.\end{aligned} \tag{146}$$

Here C^u_* *denotes the Hölder–Zygmund space. This envelope corresponds to the extension constructed at integer order* k.

Proof. On $\|z\| \le 2\rho$, one has $\|x\| + \|y\| \le 2\sqrt{2}\rho$. Proposition D.7, with upper horizon $T+k+1$, therefore gives the hypothesis of Lemma E.13 with $A \le C_k\rho^{c_k}(1+t_0^{-1})^{\chi_k}$. Let H be its extension. For $\omega \in \mathbb{S}^p$ with $\omega_{p+1} > 0$, define

$$F_{B,q,\rho}(\omega) = H\left(\rho\frac{\omega_{1:p}}{\omega_{p+1}}\right),$$

and set it to zero for $\omega_{p+1} \le 0$. By (141), its support on the upper hemisphere has $\omega_{p+1} \ge 1/\sqrt{6}$. It thus vanishes on an open band about the equator, making the zero extension C^k on the entire sphere.

On a slightly larger fixed cap, for example $\omega_{p+1} > 1/(2\sqrt{6})$, every chart derivative through order k of $\rho\omega_{1:p}/\omega_{p+1}$ is bounded by $C_k\rho$. The finite-order chain rule on a fixed finite atlas gives $\|F\|_{C^k} \le C_k(1+\rho)^k\|H\|_{C^k}$. This proves the C^k bound, while the range bound is unchanged by composition. Equation (143) follows from (139).

For clarity, the integer Sobolev bound requires no loss of regularity. For a smooth function, the spectral norm obeys

$$\|F\|^2_{H^k} = \langle F, (I-\Delta_{\mathbb{S}^p})^k F\rangle = \sum_{i=0}^k \binom{k}{i}\langle F, (-\Delta_{\mathbb{S}^p})^i F\rangle.$$

If $i = 2r$, the summand is $\|\Delta^r F\|^2_{L^2}$; if $i = 2r+1$, it is $\|\nabla\Delta^r F\|^2_{L^2}$. Each uses at most $i \le k$ derivatives, so is bounded by $C_k\|F\|^2_{C^k}$ on the fixed sphere. Smooth approximation extends this estimate to C^k functions.

For $u \le k$, monotonicity of the spectral Sobolev weights gives $\|F\|_{H^u} \le \|F\|_{H^k}$. Finally, in a fixed atlas the finite-difference definition of C^u_* implies

$$\|F\|_{C^u_*} \le C_{k,u}\|F\|_{C^k}.$$

Indeed, a finite difference of order $k+1$ is bounded by $C\|F\|_{C^k}|h|^k$ for $|h| \le 1$, and the zeroth-order term handles larger increments. This includes the integer endpoint in the Zygmund sense and proves (146). □

For the training calibration, fix a real order $0 < u \le k$ and put $j = \lceil u\rceil \le k$. Apply Lemma E.14 at integer order j, using the corresponding bounds from Assumption 3.2. Write $F_{N,q} = F_{B_N,q,\rho_N}$ for this construction, with $\rho_N \ge \max\{1, T+1\}$. The dependence of $F_{N,q}$ on the selected order u is suppressed in the notation. The resulting target match is

$$F_{N,q}(\psi_{\rho_N}(v)) = B_N^{-1} f^\star_{B_N,q}(v), \qquad \|v\| \le \rho_N, \quad t \in [t_0, T]. \tag{147}$$

The normalized extension satisfies

$$\begin{aligned}\|F_{N,q}\|_{C^u_*} + \|F_{N,q}\|_{H^u} &\le A_{u,N}(t_0), \qquad \|F_{N,q}\|_\infty \le C_j,\\ A_{u,N}(t_0) &= C_{j,u}\rho_N^{c_j}(1+t_0^{-1})^{\chi_j} \ge 1, \qquad \chi_j = j\max\{4, b_j\}.\end{aligned} \tag{148}$$

Here $b_j = \max_{1\le|\alpha|\le j} b_\alpha$. At logarithmic localization the displayed polynomial dependence on ρ_N is explicit; the extension introduces no further power of t_0^{-1}. All targets evaluated in the reflection proof have time at least t_0. For the statistical calibration, $1 \le u \le \bar{k}$; the main results use $u = \bar{k}$.

E.3.2 A QUALIFIED HEAT SMOOTHER

For an integer $J \geq 1$, set

$$R_t^{(J)} = I - (I - P_t)^J. \tag{149}$$

Lemma E.15 (Qualified heat smoothing). *Let $0 < u < s$, choose a fixed integer $J > u/2$, and let $F \in C_*^u(\mathbb{S}^p) \cap H^u(\mathbb{S}^p)$. For $0 < t \leq 1$,*

$$\left\| R_t^{(J)} F - F \right\|_\infty^2 \leq C_{p,u,J} t^u \|F\|_{C_*^u}^2, \tag{150}$$

$$\left\| R_t^{(J)} F \right\|_{H^s}^2 \leq C_{p,u,J} t^{-(s-u)} \|F\|_{H^u}^2, \tag{151}$$

$$\left\| R_t^{(J)} F \right\|_\infty \leq (2^J - 1) \|F\|_\infty. \tag{152}$$

Proof. Since $F - R_t^{(J)} F = (I - P_t)^J F$, the heat-semigroup characterization of Hölder–Zygmund spaces on a compact smooth manifold gives

$$\left\| (I - P_t)^J F \right\|_\infty \leq C t^{u/2} \|F\|_{C_*^u}, \qquad J > u/2;$$

see Triebel (2006). This proves (150).

For the Sobolev estimate, expand in spherical harmonics. If $\lambda_\ell = \ell(\ell + p - 1)$ and $r_J(z) = 1 - (1 - e^{-z})^J$, then $0 \leq r_J(z) \leq \min\{1, Je^{-z}\}$. Consequently,

$$\begin{aligned} \left\| R_t^{(J)} F \right\|_{H^s}^2 &\leq \left[\sup_{z \geq 0} (1+z)^{s-u} r_J(tz)^2 \right] \|F\|_{H^u}^2 \\ &\leq C_{p,u,J} t^{-(s-u)} \|F\|_{H^u}^2. \end{aligned}$$

Finally,

$$R_t^{(J)} = \sum_{i=1}^{J} (-1)^{i+1} \binom{J}{i} P_{it},$$

and every P_{it} is an L^∞ contraction. Summing the absolute binomial coefficients proves (152). □

E.3.3 LOCALIZED COMPARATORS AND THE SHALLOW PARITY RESTRICTION

Lemma E.16 (Even extension on the localized upper cap). *Fix $0 < u \leq \min\{k, s\}$ and let $F = F_{N,q}$ be the normalized extension constructed at $j = \lceil u \rceil$. Let $\mathcal{A}F(\omega) = F(-\omega)$ denote antipodal pullback, and set $EF = F + \mathcal{A}F$. Then EF is even and agrees with F throughout the upper hemisphere. For every $a \geq 0$ for which $F \in H^a$,*

$$\|EF\|_{H^a} \leq 2 \|F\|_{H^a}, \qquad \|EF\|_\infty \leq \|F\|_\infty. \tag{153}$$

For the saturated construction $u = s$, with $j_s = \lceil s \rceil \leq k$, one has $EF \in \mathcal{H}_2^{\mathrm{full}}$ and $\|EF\|_{\mathcal{H}_2^{\mathrm{full}}} \leq C A_{s,N}(t_0)$. For $u < s$ and a fixed integer $J > u/2$, the even functions $h_t = R_t^{(J)} EF$ satisfy, for $0 < t \leq 1$,

$$\|h_t\|_{\mathcal{H}_2^{\mathrm{full}}}^2 \leq C t^{-(s-u)} A_{u,N}(t_0)^2, \tag{154}$$

$$\sup_{\omega \in \mathcal{C}_{\rho_N}^+} |h_t(\omega) - F(\omega)|^2 \leq C t^u A_{u,N}(t_0)^2, \qquad \|h_t\|_\infty \leq C_{j,J}. \tag{155}$$

Here $\mathcal{H}_2^{\mathrm{full}}$ is the full NTK RKHS of one hidden ReLU layer, and its even-subspace Sobolev identification is given in Lemma E.11.

Proof. The support of F lies in $\{\omega_{p+1} \geq 1/\sqrt{6}\}$, so $\mathcal{A}F$ vanishes on the upper hemisphere and the two supports are disjoint. This proves matching and the range bound. On degree-ℓ spherical harmonics, $\mathcal{A}$ multiplies coefficients by $(-1)^\ell$. It therefore is an isometry in every spectral Sobolev norm and commutes with the spherical Laplacian and its heat semigroup. The triangle inequality

proves the H^a estimate; EF has only even harmonic coefficients. On those coefficients the shallow eigenvalues are comparable to $(1+\ell)^{-d_0}$, hence

$$\|EF\|_{\mathcal{H}_2^{\mathrm{full}}} \le C \|EF\|_{H^s}$$

whenever the latter norm is finite. For the construction at $u = s$, this norm is controlled by (148) with $j = j_s$.

Antipodal pullback also preserves the heat-semigroup characterization of C_*^u, so

$$\|EF\|_{C_*^u} \le C_u \|F\|_{C_*^u}.$$

Since the heat semigroup commutes with $\mathcal{A}$, h_t remains even. Apply Lemma E.15 to EF, then use the shallow Sobolev identification on the even subspace and (148). The uniform approximation to EF is an approximation to F on $\mathcal{C}^+_{\rho_N}$, proving the claims. □

Choosing t in (154) by the same norm budget as below gives the same localized approximation exponents, with changed constants. Off the localization event, the bounded range gives the same tail remainder. Thus the shallow missing odd modes alone do not obstruct either the saturated or the lower-regularity comparator construction. This observation concerns approximation; the main theorem is stated for the selected $L \ge 3$ architecture and does not assert a separately calibrated shallow training theorem.

E.3.4 General source-free approximation profile

Theorem E.17 (General-smoothness comparator)**.** *Suppose Assumptions 3.1 and 3.2 hold at integer order k. Fix $0 < u \le k$ with $u < s$ and a fixed integer $J > u/2$. Use the normalized extension constructed at $j = \lceil u \rceil \le k$, and put*

$$a_u = \frac{u}{s-u} = \frac{2u}{D+2-2u}. \tag{156}$$

For every squared norm budget $R \ge C A_{u,N}(t_0)^2$, there is $h_{q,R} \in \mathcal{H}_L^{\mathrm{full}} \simeq H^s(\mathbb{S}^p)$ such that

$$\|h_{q,R}\|^2_{\mathcal{H}_L^{\mathrm{full}}} \le R, \qquad \|h_{q,R}\|_\infty \le C_{j,J}, \tag{157}$$

and

$$\begin{aligned} \left\| B_N h_{q,R} \circ \psi_{\rho_N} - f^\star_{B_N,q} \right\|^2_{L^2(P_U)} &\le C B_N^2 A_{u,N}(t_0)^{\frac{2s}{s-u}} R^{-a_u} + C_{j,J} B_N^2 \tau_N, \\ \tau_N &= \mathbb{P}(\|U\|_2 > \rho_N). \end{aligned} \tag{158}$$

Proof. Take $h_{q,R} = R_t^{(J)} F_{N,q}$ and choose

$$t = \left(\frac{C A_{u,N}(t_0)^2}{R} \right)^{1/(s-u)} \le 1.$$

By Theorem E.10 and Lemma E.15, the RKHS norm is at most R. On the localized positive-time cap,

$$\begin{aligned} \|h_{q,R} - F_{N,q}\|^2_\infty &\le C A_{u,N}(t_0)^2 t^u \\ &\le C A_{u,N}(t_0)^{2s/(s-u)} R^{-u/(s-u)}. \end{aligned}$$

This uniform estimate can be integrated against the actual design P_X, despite its singularity relative to uniform spherical measure. Off the localization event, (152), the uniform range bound in (148), and $|B_N^{-1} f^\star_{B_N,q}| \le 1$ give the $C_{j,J}\tau_N$ remainder. The same heat contraction bound proves the asserted uniform envelope for $h_{q,R}$. Multiply by B_N^2. □

Proposition E.18 (Saturated comparator and its localization remainder)**.** *Suppose $k \ge \lceil s \rceil$ in Assumption 3.2. Use the normalized extension with $u = s$ and $j_s = \lceil s \rceil$. Enlarge the fixed constant in $A_{s,N}(t_0)$ to include the RKHS norm equivalence. Then $F_{N,q} \in \mathcal{H}_L^{\mathrm{full}}$, $\|F_{N,q}\|_{\mathcal{H}_L^{\mathrm{full}}} \le A_{s,N}(t_0)$, $\|F_{N,q}\|_\infty \le C_{j_s}$, and*

$$\sum_{q=1}^{d_x} \left\| B_N F_{N,q} \circ \psi_{\rho_N} - f^\star_{B_N,q} \right\|^2_{L^2(P_U)} \le C_{j_s} d_x B_N^2 \tau_N, \qquad \tau_N = \mathbb{P}(\|U\|_2 > \rho_N). \tag{159}$$

Proof. Equations (148) and (136) give the norm bound and RKHS membership. By (147), the error is zero on $\{\|U\|_2 \le \rho_N\}$. On its complement, $|F_{N,q}| \le C_{j_s}$ and $|B_N^{-1} f^\star_{B_N,q}| \le 1$ bound the squared error by $C_{j_s} B_N^2$, after enlarging the constant. Integrate and sum over coordinates. □

Thus RKHS membership removes the localized heat-approximation term, while finite stopping contributes regularization bias and the population error retains its off-localization remainder. By Lemma C.5, $\tau_N \le Ce^{-c(\rho_N - T)}$. For a prescribed budget $\varepsilon_{\mathrm{loc}} > 0$, taking

$$\rho_N \ge \max\left\{1, T+1,\, T + c^{-1}\log\frac{C_{j_s} C d_x B_N^2}{\varepsilon_{\mathrm{loc}}}\right\}$$

absorbs the term in (159) into that budget. The calibrated learning bound in Proposition C.8 uses this comparator and performs that absorption explicitly.

E.4 Full-NTK Capacity for Arbitrary Training Distributions

The projection-and-tail method is established in information-gain analysis (Vakili et al., 2021b); for neural kernels, the uniform harmonic tail via the addition theorem appears in Vakili et al. (2021a, Theorem 2 and Appendix G). We record its resolvent-trace consequence for arbitrary population and empirical designs, in the form required by the early-stopping oracle.

Let $d_0 = D + 2$, and let σ be normalized surface measure on $\mathbb{S}^{d_0-1}$. The full NTK in Theorem E.10 is rotationally invariant and has the expansion

$$K(x,x') = \sum_{\ell=0}^{\infty} \mu_\ell \sum_{a=1}^{M_\ell} Y_{\ell,a}(x) Y_{\ell,a}(x'), \qquad 0 \le \mu_\ell \le C_{L,d_0}(1+\ell)^{-d_0}, \tag{160}$$

where the harmonics are orthonormal in $L^2(\sigma)$ and

$$M_\ell \le C_{d_0}(1+\ell)^{d_0-2}. \tag{161}$$

The multiplicity estimate is the standard spherical-harmonic dimension bound. Lower eigenvalue bounds are needed for the RKHS identification, but not for the capacity argument here.

Define the feature map and its covariance by

$$\Phi(x) = (\sqrt{\mu_\ell} Y_{\ell,a}(x))_{\ell,a}, \qquad T_P = \mathbb{E}_{X\sim P}[\Phi(X)\otimes\Phi(X)].$$

The addition theorem gives $\sum_a Y_{\ell,a}(x)^2 = M_\ell$ for every x. Consequently the series for $\|\Phi(x)\|^2$ has a summable, uniform tail. The feature map is well defined in ℓ_2, the kernel expansion converges uniformly by Cauchy–Schwarz, and T_P is positive and trace class for every probability law P on the sphere. Its nonzero eigenvalues agree with those of the kernel integral operator, or equivalently with those of the covariance operator in the RKHS. Thus either realization defines the same effective dimension

$$\mathsf{df}_P(\lambda) = \mathrm{tr}[T_P(T_P+\lambda I)^{-1}].$$

Proposition E.19 (Uniform capacity bound over all designs)**.** *There is a constant C_{L,d_0} depending only on the constants in (160) and the dimension such that, for every probability law P on $\mathbb{S}^{d_0-1}$ and $0 < \lambda \le 1$,*

$$\mathsf{df}_P(\lambda) \le C_{L,d_0}\lambda^{-\gamma}, \qquad \gamma = \frac{d_0-1}{d_0} = \frac{D+1}{D+2}. \tag{162}$$

In particular, for every deterministic collection $x_1,\dots,x_N$, its empirical covariance satisfies

$$\widehat{\mathsf{df}}(\lambda) := \mathrm{tr}[T_N(T_N+\lambda I)^{-1}] \le C_{L,d_0}\lambda^{-\gamma}. \tag{163}$$

No density, separation, or randomness assumption is needed for this last inequality.

Proof. For an integer $J \ge 1$, let Π_J project onto the harmonic feature coordinates of degrees at most J. Summing multiplicities gives $\mathrm{rank}(\Pi_J) \le C_{d_0}(1+J)^{d_0-1}$. Tonelli's theorem and the addition theorem give, for every P,

$$\mathrm{tr}[(I-\Pi_J)T_P] = \mathbb{E}_{X\sim P}\,\|(I-\Pi_J)\Phi(X)\|^2 \tag{164}$$

$$= \sum_{\ell > J} \mu_\ell M_\ell \le C_{L,d_0} \sum_{\ell > J} (1+\ell)^{-2} \le \frac{C_{L,d_0}}{1+J}. \tag{165}$$

Put $A = T_P(T_P + \lambda I)^{-1}$. Functional calculus gives $0 \preceq A \preceq I$ and $A \preceq \lambda^{-1} T_P$. Since A is trace class, splitting its trace over the two orthogonal coordinate subspaces is legitimate and yields

$$\begin{aligned} \operatorname{tr}(A) &= \operatorname{tr}(\Pi_J A \Pi_J) + \operatorname{tr}((I - \Pi_J) A (I - \Pi_J)) \\ &\le \operatorname{rank}(\Pi_J) + \lambda^{-1} \operatorname{tr}[(I - \Pi_J) T_P] \\ &\le C_{L,d_0} \left((1+J)^{d_0 - 1} + \frac{1}{\lambda (1+J)} \right). \end{aligned}$$

No commutation of Π_J with T_P is used: the inequalities are applied after compression. Taking $J = \lceil \lambda^{-1/d_0} \rceil$ proves (162). Apply that result to $P = N^{-1} \sum_i \delta_{x_i}$ to obtain (163). □

Application to the training law. The law $P_X = \mathcal{L}(\psi_{\rho_N}(U))$ may be singular relative to uniform spherical measure: a discrete choice of the training time puts its support on finitely many time slices. The proposition applies to this law without a density-domination argument, and the constant does not involve the number of time points or their probabilities. It therefore applies in particular to the score-weighted right-endpoint training law. It also applies directly to the realized empirical distribution, so (163) bounds the noise term in Theorem E.24 without a separate empirical-effective-dimension concentration event. The regularized operator comparison and its sample-size condition remain necessary for the population prediction bound.

E.5 Kernel Filters, Concentration, and Early-Stopping Guarantees

Let $K = K_L^{\mathrm{full}}$, let $\mathcal{H} = \mathcal{H}_L^{\mathrm{full}}$, and assume $K(x,x) \le \kappa_K^2$. Write $S : \mathcal{H} \to L^2(P_X)$ for inclusion, $T = S^*S$, $S_N h = N^{-1/2}(h(X_i))_{i=1}^N$, and $T_N = S_N^* S_N$. Kernel GD from zero is

$$g_r = q_r(T_N) S_N^* Y_N, \qquad q_r(v) = \eta \sum_{j=0}^{r-1} (1 - \eta v)^j, \qquad Y_N = N^{-1/2}(Y_i)_{i=1}^N. \tag{166}$$

Put $\lambda_r = (\eta r)^{-1}$ and

$$\mathsf{df}_{P_X}(\lambda) = \operatorname{tr}\left[T(T + \lambda I)^{-1} \right]. \tag{167}$$

The covariates are independent with their common actual training law P_X. No absolute continuity of this law is assumed. The sample size, step size, and stopping time are prescribed before observing the sample.

E.5.1 Regularized operator comparison with an explicit sample gate

The comparison argument holds for any separable real RKHS $\mathcal{H}$ with measurable feature map $x \mapsto K_x$ and $K(x,x) \le \kappa_K^2$, where $\kappa_K > 0$. Let $X_1, \dots, X_N$ be independent with common law P_X. Write

$$T = \mathbb{E}[K_X \otimes K_X], \qquad T_N = \frac{1}{N} \sum_{i=1}^N K_{X_i} \otimes K_{X_i}, \qquad L_\lambda = \frac{\kappa_K^2}{\lambda}.$$

Both operators are trace class. Throughout this subsection, $0 < \lambda \le \kappa_K^2$, so $L_\lambda \ge 1$.

Lemma E.20 (Regularized comparison)**.** *Let $0 < \delta_o < 1$. If*

$$N\lambda \ge 64 \kappa_K^2 \log\left(\frac{16(1 + \kappa_K^2/\lambda)}{\delta_o} \right), \tag{168}$$

then, with probability at least $1 - \delta_o$,

$$\left\| (T + \lambda I)^{-1/2} (T_N - T) (T + \lambda I)^{-1/2} \right\|_{\mathrm{op}} \le \frac{1}{2}. \tag{169}$$

On this event,

$$\tfrac{1}{2}(T + \lambda I) \preceq T_N + \lambda I \preceq \tfrac{3}{2}(T + \lambda I), \qquad T \preceq 2(T_N + \lambda I). \tag{170}$$

In particular, for the inclusion $S : \mathcal{H} \to L^2(P_X)$ *and empirical evaluation* $S_N h = N^{-1/2}(h(X_i))_{i=1}^N$,

$$\|Sh\|^2_{L^2(P_X)} \le 2\big(\|S_N h\|_2^2 + \lambda \|h\|^2_{\mathcal{H}}\big) \quad \textit{for every } h \in \mathcal{H}. \tag{171}$$

Proof. Put $v_i = (T+\lambda I)^{-1/2} K_{X_i}$ and $C_\lambda = \mathbb{E}[v_i \otimes v_i] = T(T+\lambda I)^{-1}$. Then

$$0 \preceq C_\lambda \preceq I, \qquad \operatorname{tr} C_\lambda = \mathsf{df}_{P_X}(\lambda) \le L_\lambda, \qquad \|v_i\|^2 \le L_\lambda.$$

The independent centered self-adjoint operators $Z_i = v_i \otimes v_i - C_\lambda$ obey

$$\|Z_i\|_{\mathrm{op}} \le L_\lambda, \qquad \mathbb{E}[Z_i^2] = \mathbb{E}[\|v_i\|^2 v_i \otimes v_i] - C_\lambda^2 \preceq L_\lambda C_\lambda. \tag{172}$$

The first inequality follows from $-C_\lambda \preceq Z_i \preceq v_i \otimes v_i$ and $L_\lambda \ge 1$.

For completeness, the dimension factor in matrix Bernstein can be kept finite without asserting that the RKHS itself is finite dimensional. Let P_n be increasing finite-rank orthogonal projections converging strongly to the identity. Compress Z_i to $P_n\mathcal{H}$, and append a zero coordinate. A variance upper bound for their sum, in either order of matrix multiplication, is

$$V_n = N L_\lambda \begin{pmatrix} P_n C_\lambda P_n & 0 \\ 0 & 1 \end{pmatrix}.$$

Indeed $(P_n Z_i P_n)^2 \preceq P_n Z_i^2 P_n$. The added coordinate ensures $\|V_n\|_{\mathrm{op}} = N L_\lambda$, while the intrinsic dimension of the two variance blocks is

$$\frac{2 \operatorname{tr} V_n}{\|V_n\|_{\mathrm{op}}} = 2\big(\operatorname{tr}(P_n C_\lambda P_n) + 1\big) \le 2(\mathsf{df}_{P_X}(\lambda) + 1).$$

The intrinsic matrix Bernstein inequality of Tropp (2015, Theorem 7.3.1), applied to these variance upper bounds, therefore gives

$$\mathbb{P}\left(\left\| \sum_{i=1}^N P_n Z_i P_n \right\|_{\mathrm{op}} > \frac{N}{2} \right) \le 8(\mathsf{df}_{P_X}(\lambda) + 1) \exp\left(-\frac{3N}{28 L_\lambda} \right) \tag{173}$$

provided $N/2 \ge \sqrt{N L_\lambda} + L_\lambda/3$; the sufficient condition $N \ge 16 L_\lambda$ implies this restriction. Each Z_i is compact, since it is a rank-one operator minus a trace-class operator. Consequently the compressed sums converge in operator norm to $\sum_i Z_i$. Passing to the limit in the tail bound is legitimate (or first use any threshold smaller than $N/2$ and then increase it). This proves the same bound for the infinite-dimensional sum.

Condition (168) implies $N \ge 16 L_\lambda$ and makes the right-hand side of (173) at most δ_o, using $\mathsf{df}_{P_X}(\lambda) \le L_\lambda$. Divide the sum by N to obtain (169). Congruence by $(T+\lambda I)^{1/2}$ yields the first two Loewner inequalities in (170). The remaining statements follow by taking quadratic forms. □

Lemma E.21 (Optional empirical-to-population effective-dimension bound)**.** *Define* $\ell_\lambda(x) = \langle K_x, (T+\lambda I)^{-1} K_x \rangle$. *For* $0 < \delta_\ell < 1$, *outside an event of probability* δ_ℓ, *its empirical mean satisfies*

$$\frac{1}{N} \sum_i \ell_\lambda(X_i) \le 2\mathsf{df}_{P_X}(\lambda) + \frac{L_\lambda \log(1/\delta_\ell)}{N}. \tag{174}$$

On the intersection with (169),

$$\widehat{\mathsf{df}}(\lambda) \le 4\mathsf{df}_{P_X}(\lambda) + \frac{2 L_\lambda \log(1/\delta_\ell)}{N}. \tag{175}$$

The joint failure probability is at most $\delta_o + \delta_\ell$.

Proof. The leverage is bounded between zero and L_λ, its expectation is $\mathsf{df}_{P_X}(\lambda)$, and its variance is at most $L_\lambda \mathsf{df}_{P_X}(\lambda)$. Scalar Bernstein gives, with $u = \log(1/\delta_\ell)$,

$$\frac{1}{N} \sum_i \ell_\lambda(X_i) \le \mathsf{df}_{P_X}(\lambda) + \sqrt{\frac{2 L_\lambda \mathsf{df}_{P_X}(\lambda) u}{N}} + \frac{L_\lambda u}{3N}.$$

Use $\sqrt{2ab} \le a + b/2$ to get (174). The lower comparison in (170) implies $(T_N + \lambda I)^{-1} \preceq 2(T + \lambda I)^{-1}$. Taking the trace after congruence by $T_N^{1/2}$ gives

$$\widehat{\mathsf{df}}(\lambda) \le 2\,\mathrm{tr}[T_N(T + \lambda I)^{-1}] = \frac{2}{N}\sum_i \ell_\lambda(X_i),$$

which proves the assertion. This scalar event is additional to the operator-comparison event; it does not follow from that event alone. For the spherical kernel, Proposition E.19 instead provides (163) deterministically. □

E.5.2 Filters and deterministic parts of the oracle

For $r \ge 1$, $0 < \eta\kappa_K^2 \le 1$, set

$$q_r(v) = \eta\sum_{j=0}^{r-1}(1 - \eta v)^j, \qquad b_r(v) = 1 - vq_r(v), \qquad \lambda = (\eta r)^{-1}.$$

Here $q_r(0) = \eta r$. Since $\|T_N\|_{\mathrm{op}} \le \kappa_K^2$, all the following scalar bounds apply to T_N.

Lemma E.22 (Filter bounds). *For $0 \le v \le \kappa_K^2$,*

$$0 \le vq_r(v) \le 1, \qquad 0 \le b_r(v) \le \frac{\lambda}{v + \lambda}, \qquad 0 \le q_r(v) \le \frac{2}{v + \lambda}. \tag{176}$$

In particular,

$$(v + \lambda)b_r(v)^2 \le \lambda, \qquad v(v + \lambda)q_r(v)^2 \le \frac{4v}{v + \lambda} \le 4. \tag{177}$$

Proof. For $v > 0$, summing the geometric series gives $vq_r(v) = 1 - (1 - \eta v)^r$; continuity covers $v = 0$. For $z \in [0, 1]$, $(1-z)^r \le (1+rz)^{-1}$, so $b_r(v) \le (1+\eta rv)^{-1}$. Also $q_r(v) \le \min\{\eta r, v^{-1}\} \le 2/(v + \lambda)$, with the evident interpretation at zero. The two energy bounds follow by squaring these inequalities. □

Suppose the responses have the form $Y_i = f(X_i) + \xi_i$, where $|f| \le 1$. Let $Y_N = N^{-1/2}(Y_i)_i$ and $\xi_N = N^{-1/2}(\xi_i)_i$. Zero-start kernel gradient descent satisfies $g_r = q_r(T_N)S_N^*Y_N$. For a deterministic comparator $h \in \mathcal{H}$, write $\Delta = f - Sh$, $\Delta_N = N^{-1/2}(\Delta(X_i))_i$, and $d_h = q_r(T_N)S_N^*\Delta_N$. Exact algebra yields

$$Sg_r - f = -Sb_r(T_N)h + (Sd_h - \Delta) + Sq_r(T_N)S_N^*\xi_N. \tag{178}$$

On the operator event, the bias term obeys

$$\|Sb_r(T_N)h\|_{L^2(P_X)}^2 \le 2\langle h, b_r(T_N)(T_N + \lambda I)b_r(T_N)h\rangle \le 2\lambda\,\|h\|_{\mathcal{H}}^2. \tag{179}$$

For the misspecification term, spectral calculus gives

$$\|S_N d_h\|_2^2 + \lambda\,\|d_h\|_{\mathcal{H}}^2 = \langle \Delta_N, S_N q_r(T_N)(T_N + \lambda I)q_r(T_N)S_N^*\Delta_N\rangle \le 4\,\|\Delta_N\|_2^2. \tag{180}$$

To justify the last inequality, the nonzero eigenvalues of the matrix between the two vectors are $v(v + \lambda)q_r(v)^2$, with v ranging over the nonzero eigenvalues of T_N; use (177). Therefore

$$\|Sd_h - \Delta\|_{L^2(P_X)}^2 \le 16\,\|\Delta_N\|_2^2 + 2\,\|\Delta\|_{L^2(P_X)}^2. \tag{181}$$

This step uses no minimum empirical Gram eigenvalue.

E.5.3 Noise control and the bounded-comparator oracle

Assume that, conditional on $X_1, \dots, X_N$, the variables $\xi_1, \dots, \xi_N$ are independent and satisfy

$$\mathbb{E}[\exp(u\xi_i) \mid X_1, \dots, X_N] \le \exp(\sigma^2 u^2/2), \qquad u \in \mathbb{R}. \tag{182}$$

This condition includes conditional centering. Independence implies the corresponding joint conditional moment-generating-function bound for every linear combination of the ξ_i.

Lemma E.23 (Conditional quadratic-form bound for kernel GD)**.** *For every $u > 0$, with conditional probability at least $1 - e^{-u}$, on the operator-comparison event,*

$$\|Sq_r(T_N)S_N^*\xi_N\|_{L^2(P_X)}^2 \le \frac{24\sigma^2}{N}\big(\widehat{\mathsf{df}}(\lambda) + u\big). \tag{183}$$

More precisely, the conditional exceptional event has probability at most e^{-u} for every fixed covariate realization; the displayed conclusion follows on its intersection with the operator event.

Proof. Conditional on the covariates, let

$$A_r = S_N q_r(T_N) T q_r(T_N) S_N^* \succeq 0.$$

The squared norm in (183) equals $N^{-1}\xi^\top A_r \xi$, for the unnormalized vector $\xi = (\xi_1, \dots, \xi_N)^\top$. On the operator event,

$$A_r \preceq 2S_N q_r(T_N)(T_N + \lambda I) q_r(T_N) S_N^*.$$

The eigenvalues of the right-hand side are bounded by $8v/(v+\lambda)$, by (177). It follows that

$$\|A_r\|_{\mathrm{op}} \le 8, \qquad \operatorname{tr} A_r \le 8\widehat{\mathsf{df}}(\lambda), \qquad \operatorname{tr}(A_r^2) \le 64\widehat{\mathsf{df}}(\lambda). \tag{184}$$

Apply the sub-Gaussian quadratic-form inequality of Hsu et al. (2012) (Theorem 1 in the arXiv version), conditionally, to $A_r^{1/2}\xi$. It gives

$$\xi^\top A_r \xi \le \sigma^2 \left[\operatorname{tr} A_r + 2\sqrt{\operatorname{tr}(A_r^2)u} + 2\|A_r\|_{\mathrm{op}} u\right]$$

outside a conditional exceptional event of probability e^{-u}. Writing $a = \widehat{\mathsf{df}}(\lambda)$, the bracket is at most $8a + 16\sqrt{au} + 16u \le 16a + 24u \le 24(a+u)$. Divide by N. □

Theorem E.24 (Bounded-comparator early-stopping oracle)**.** *Let $K = K_L^{\mathrm{full}}$ satisfy Theorem E.10. Assume the observation model (182) and $|f| \le 1$. Let $\mathcal{G} \subset \mathcal{H}$ be a deterministic finite family, of cardinality $J_{\mathcal{G}} \ge 1$, such that $\|h\|_\infty \le M$ for every $h \in \mathcal{G}$. Fix r and η independently of the sample, with $0 < \eta\kappa_K^2 \le 1$ and $0 < \lambda = (\eta r)^{-1} \le \min\{1, \kappa_K^2\}$. If (168) holds with $\delta_o = \delta/3$, then with probability at least $1 - \delta$,*

$$\begin{aligned}\|Sg_r - f\|_{L^2(P_X)}^2 \le C \min_{h\in\mathcal{G}} \Big\{ \|Sh - f\|_{L^2(P_X)}^2 + \lambda \|h\|_{\mathcal{H}}^2 \Big\} \\ + \frac{C}{N}\left[(1+M)^2 \log\frac{3J_{\mathcal{G}}}{\delta} + \sigma^2\left(C_{L,d_0}\lambda^{-\gamma} + \log\frac{3}{\delta}\right)\right],\end{aligned} \tag{185}$$

where C is universal and $\gamma = (D+1)/(D+2)$. The empirical measure may be supported on finitely many time slices.

Proof. Allocate failure probability $\delta/3$ to each of: operator comparison, simultaneous misspecification concentration, and the conditional noise event. The first allocation is justified by Lemma E.20.

For fixed h, let $a_h = \|f - Sh\|_{L^2(P_X)}^2$ and $B = 1 + M$. The independent variables $W_i = (f(X_i) - h(X_i))^2$ satisfy $0 \le W_i \le B^2$, $\mathbb{E}W_i = a_h$, and $\operatorname{Var}(W_i) \le B^2 a_h$. Scalar Bernstein, followed by a union bound over the deterministic family, gives simultaneously for every $h \in \mathcal{G}$, with probability at least $1 - \delta/3$,

$$\|\Delta_N\|_2^2 \le a_h + \sqrt{\frac{2B^2 a_h u_m}{N}} + \frac{B^2 u_m}{3N} \le 2a_h + \frac{B^2 u_m}{N}, \qquad u_m = \log\frac{3J_{\mathcal{G}}}{\delta}. \tag{186}$$

Combining with (181) yields

$$\|Sd_h - \Delta\|_{L^2(P_X)}^2 \le 34 a_h + \frac{16B^2 u_m}{N}.$$

Take $u_n = \log(3/\delta)$ in Lemma E.23; conditional failure integrates to at most $\delta/3$. Proposition E.19 then replaces $\widehat{\mathsf{df}}(\lambda)$ by $C_{L,d_0}\lambda^{-\gamma}$ without another random event.

On the intersection of these three events, use $\|a+b+c\|^2 \le 3(\|a\|^2+\|b\|^2+\|c\|^2)$ in (178), together with (179) and (183). For every $h \in \mathcal{G}$, this gives

$$\|Sg_r - f\|^2_{L^2(P_X)} \le 102a_h + 6\lambda\|h\|^2_{\mathcal{H}} + \frac{48B^2u_m}{N} + \frac{72\sigma^2}{N}(C_{L,d_0}\lambda^{-\gamma} + u_n).$$

Taking the minimum proves (185), for example with $C = 128$. A union bound completes the failure-probability calculation. The comparator family and stopping time are deterministic; a data-dependent selection among additional stopping times requires its own simultaneous event or a separate argument. □

Remark E.25 (Relation to the population-effective-dimension formulation). For a general bounded kernel, the same proof gives a noise contribution $C\sigma^2(\widehat{\mathsf{df}}(\lambda)+\log(1/\delta))/N$ on the operator event. To express it using $\mathsf{df}_{P_X}(\lambda)$, allocate $\delta/4$ each to operator comparison, leverage concentration, misspecification concentration, and conditional noise. Lemma E.21 then adds $C\sigma^2 L_\lambda \log(4/\delta)/N^2$. Under (168), $N \ge L_\lambda$, so this is bounded by the already present $C\sigma^2\log(4/\delta)/N$ term. This establishes the population-effective-dimension oracle with an explicit allocation of its additional event. The spherical oracle (185) needs only the three events in its proof.

Remark E.26 (Arbitrary RKHS comparators and the absorption condition). If a deterministic finite family is specified without a uniform sup bound, one still has $|f(x)-h(x)|^2 \le 2+2\kappa_K^2\|h\|^2_{\mathcal{H}}$ by the reproducing property. Applying the preceding Bernstein argument to each comparator therefore introduces

$$\frac{C\log(3J_{\mathcal{G}}/\delta)}{N}(1+\kappa_K^2\|h\|^2_{\mathcal{H}}).$$

Its comparator-norm part is absorbed into $C\lambda\|h\|^2_{\mathcal{H}}$ under the additional condition

$$N\lambda \ge \kappa_K^2\log(3J_{\mathcal{G}}/\delta). \tag{187}$$

Thus a generic finite-family oracle must either retain this term or state the absorption condition. The heat comparators used here have a common sup bound depending only on the chosen smoothness order: the global extension is bounded, and $\|[I-(I-P_t)^J]F\|_\infty \le (2^J-1)\|F\|_\infty$. They therefore satisfy Theorem E.24 directly. The localized target extension need agree with the regression target only on the localization region: its remaining population approximation error, including the saturated case, stays inside $\|Sh-f\|^2_{L^2(P_X)}$ and must be carried until the tail calibration absorbs it.

E.5.4 Rate balance with the target-norm constant retained

Lemma E.27 (Three-term balance). *Suppose, up to logarithms,*

$$\mathcal{R}(R,\lambda) \le C_0R^{-a} + \lambda R + N^{-1}\lambda^{-\gamma}, \qquad a>0, \quad 0<\gamma\le 1. \tag{188}$$

Let $q = a+\gamma(a+1)$. The choices

$$R_N \asymp N^{1/q}C_0^{(\gamma+1)/q}, \qquad \lambda_N \asymp N^{-(a+1)/q}C_0^{-1/q} \tag{189}$$

produce, whenever the oracle is valid at these choices,

$$\mathcal{R}(R_N,\lambda_N) \le \widetilde{O}\left(C_0^{\gamma/q}N^{-a/q}\right). \tag{190}$$

Proof. Balancing C_0R^{-a} and λR gives $R \asymp (C_0/\lambda)^{1/(a+1)}$. Their common value is $C_0^{1/(a+1)}\lambda^{a/(a+1)}$. Balance this with $N^{-1}\lambda^{-\gamma}$ and substitute back. □

Remark E.28 (Capacity and the feasible stopping range). Proposition E.19 establishes $\gamma = (D+1)/(D+2)$ for the actual training law and its empirical measure. For $1 \le u \le \bar{k}$ with $u < s$, put

$$a = a_u = \frac{u}{s-u}, \qquad C_0 \asymp A_{u,N}(t_0)^{2s/(s-u)}.$$

The unconstrained balance then gives squared-risk exponent $r_u = u/(u+\gamma s)$ and stopping exponent $\ell_u = s/(u+\gamma s)$. When $u = \bar{k} < s$, these are the exponents r, ℓ in the main theorem.

The sample condition (168) must still hold. At fixed t_0, it holds eventually for the prescribed calibration because $u \ge 1$ implies $\ell_u < 1$. When t_0 varies, the target envelope can grow polynomially, and the admissible lower bound on λ can be active. The constrained calibration in Section B retains this lower bound; the unconstrained balance alone is not used to assert a rate in that regime.

F FINITE NEURAL TRAINING AND KERNEL COMPARISONS

These results control the neural-to-kernel and finite-to-limiting-kernel trajectory discrepancies at fresh inputs. We give the deterministic Duhamel interface, prove local perturbation and initialization bounds, and close the movement argument for the prescribed training horizon. The separate geometry and curvature certificates are collected in Section J.

F.1 A GUIDE TO THE DUHAMEL COMPARISON

A *training trajectory* is the sequence of predictor functions generated by gradient descent. Here we compare the neural trajectory $F^c_{\theta_r}$, the finite initialization-kernel trajectory g^K_r, and the limiting-kernel trajectory g^∞_r. A *Duhamel-based trajectory comparison* bounds the difference between two such sequences by expressing how discrepancies introduced at individual updates accumulate over training. We use this argument first to compare neural and finite-kernel training, and then to compare finite-kernel and limiting-kernel training.

The Duhamel identity. Duhamel's principle, also called variation of constants, expresses a perturbed evolution in terms of a reference evolution and its accumulated perturbations. Its discrete form is elementary: if

$$d_{r+1} = Gd_r + \xi_r$$

for a fixed linear operator G, then repeated substitution gives the exact identity

$$d_R = G^R d_0 + \sum_{r=0}^{R-1} G^{R-1-r}\xi_r.$$

Here d_r is the trajectory discrepancy, ξ_r is the *one-step perturbation*, and G^{R-1-r} propagates that perturbation through the remaining updates. The identity becomes a comparison bound once the propagators and perturbations are controlled. In particular, if G is a contraction and $d_0 = 0$, then

$$\|d_R\| \le \sum_{r=0}^{R-1} \|\xi_r\|.$$

This bound requires no strict contraction or positive spectral gap.

Training residuals and trajectory discrepancies. Fix one normalized output coordinate, the frozen training inputs $\mathbf{X} = (X_1, \dots, X_N)$, and their normalized labels Z. For the neural-to-kernel comparison, define

$$\varrho_r = F^c_{\theta_r}(\mathbf{X}) - Z, \qquad d_r = F^c_{\theta_r}(\mathbf{X}) - g^K_r(\mathbf{X}).$$

The *training residual* ϱ_r measures the neural predictor's error against the training labels. The *trajectory discrepancy* d_r measures its difference from the kernel predictor. The perturbation ξ_r is a third quantity: it is the term introduced when the two update equations are subtracted. We distinguish these quantities explicitly rather than using "Duhamel residual" for them.

Neural-to-kernel trajectory comparison. Write $K^X_r = K_X(\theta_r)$ for the tangent Gram matrix at update r and let

$$G_0 = I - \frac{\eta}{N} K^X_0.$$

Subtracting the finite-kernel update from the neural update gives

$$d_{r+1} = G_0 d_r + \xi_r, \qquad \xi_r = -\frac{\eta}{N}(K^X_r - K^X_0)\varrho_r + e^X_r,$$

as stated in (210). The first part of ξ_r comes from *kernel drift*: the network's tangent kernel changes as its parameters move. The second part, e^X_r, is the nonlinear Taylor remainder defined in (200). Thus the Duhamel identity tracks how these two effects separate neural training from training with the fixed initialization kernel.

Fix the prescribed horizon $R = R_\star$ and write $h = \eta R$. Centering and zero-start kernel GD give $d_0 = 0$. The step-size condition makes G_0 a positive-semidefinite contraction. The movement argument bounds the training residuals, and the perturbation estimates then give

$$\|\xi_r\|_{\ell_{2,N}} \le C\eta q_\star, \qquad 0 \le r < R,$$

where $\|v\|_{\ell_{2,N}} = N^{-1/2}\|v\|_2$ and $q_\star$ is the finite-width perturbation scale in (256); see Propositions F.13 and F.14. Applying the identity yields

$$\|d_r\|_{\ell_{2,N}} \le Cq_\star\eta r, \qquad \sup_{r\le R}\|d_r\|_{\ell_{2,N}} \le Cq_\star h.$$

Transferring the comparison to fresh inputs. To control population prediction error, define the function $d_r^P = F_{\theta_r}^c - g_r^K$. Subtracting the prediction updates at a fresh input gives

$$d_{r+1}^P = d_r^P - \eta\mathcal{A}_{K_{\theta_0}}d_r + \zeta_r, \qquad \zeta_r = -\eta\big(\mathcal{A}_{K_{\theta_r}} - \mathcal{A}_{K_{\theta_0}}\big)\varrho_r + e_r^P,$$

where e_r^P is the fresh-input Taylor remainder in (201). The averaged cross-kernel operator is

$$(\mathcal{A}_K v)(x) = \frac{1}{N}\sum_{i=1}^N K(x, X_i)v_i,$$

as in (198). Its norm from $\ell_{2,N}$ to $L^2(P_X)$ is $O(1)$ at fixed kernel bounds, where P_X is the fresh-input distribution.

Centering gives $d_0^P = 0$, and the population perturbation estimates give $\|\zeta_r\|_{L^2(P_X)} \le C\eta q_\star$. There are consequently two contributions to the accumulated population discrepancy. Summing the direct perturbations ζ_r contributes $O(q_\star h)$. Transferring the sample discrepancies through $\mathcal{A}_{K_{\theta_0}}$ contributes

$$O\left(\eta\sum_{r<R} q_\star\eta r\right) = O(q_\star h^2).$$

This explains the bound

$$\sup_{r\le R}\left\|F_{\theta_r}^c - g_r^K\right\|_{L^2(P_X)} \le Cq_\star(h + h^2)$$

proved in Proposition F.3. After squaring, summing over coordinates, and restoring the label scale B_N, this controls $\mathcal{E}_{\text{opt}}$ in (15).

Finite-to-limiting-kernel trajectory comparison. The second comparison applies the same identity to g_r^K and g_r^∞, trained from zero on the same inputs and labels. Here both kernels are fixed during training. The one-step perturbation is the difference between their Gram matrices acting on a kernel training residual; there is no nonlinear Taylor remainder. Uniform kernel approximation bounds this perturbation, and the same sample-to-population argument gives

$$\sup_{r\le R}\left\|g_r^K - g_r^\infty\right\|_{L^2(P_X)} \le C\epsilon_K(h + h^2)$$

at fixed kernel diagonal bounds, where ϵ_K is the uniform difference between the initialization and limiting kernels. Lemma F.11 gives the precise statement, including the dependence on the kernel bounds. This comparison controls $\mathcal{E}_{\text{width}}$.

Role in the full guarantee. The two Duhamel-based trajectory comparisons transfer the limiting-kernel learning guarantee to the actual finite neural network. The statistical error $\mathcal{E}_{\text{ker}}$ of the limiting-kernel predictor is bounded separately in Section E. The Duhamel identity is classical; the analysis here establishes the perturbation and movement bounds needed for the specified all-trained network and transfers the resulting comparisons to fresh inputs. These finite-horizon bounds require neither a Polyak–Łojasiewicz condition nor an inverse minimum Gram eigenvalue. Section A.3 discusses the prior comparison methodology and the extensions used here.

F.2 Exact All-Trained Pyramidal Architecture and Deterministic Interface

Remark F.1 (Proof sequence). The statistical argument first establishes uniform initialization controls, then movement and activation-change estimates, and finally a finite-horizon population comparison. The all-trained multilayer architecture requires unequal rectangular widths, simultaneous movement of every hidden matrix, a moving Gaussian readout, and both empirical and population Taylor remainders. The separate curvature branch combines frozen-design separation and the KMM initialization certificate with the same perturbation estimates. The shallow analyses of Kuzborskij & Szepesvári (2022) provide related context, but no shallow rate or width claim is imported.

F.2.1 KMM NORMALIZATION AND CENTERED TRAINING

Let $H = L - 1$ and $d_L = 1$. The original KMM-normalized representation is

$$h_0(x) = x, \qquad h_\ell(x) = \sigma(W_\ell h_{\ell-1}(x)), \quad \ell = 1, \dots, H, \qquad F_\theta(x) = \left(\prod_{\ell=1}^{H} \sqrt{\frac{2}{d_\ell}} \right) W_L h_H(x). \tag{191}$$

Positive homogeneity of ReLU makes this identical to the normalized recursion (5). Every entry of every W_ℓ is initialized independently from $N(0, 1)$, and the parameter space carries the product-Frobenius metric

$$\|\theta\|_F^2 = \sum_{\ell=1}^{L} \|W_\ell\|_F^2 . \tag{192}$$

The algorithm trains $F_\theta^c = F_\theta - F_{\theta_0}$. Since the subtracted term is fixed,

$$F_{\theta_0}^c \equiv 0, \qquad \nabla_\theta F_\theta^c(x) = \nabla_\theta F_\theta(x), \tag{193}$$

so the KMM tangent geometry is unchanged.

F.2.2 JACOBIANS AND KERNEL TRAJECTORIES

Let $P = \sum_{\ell=1}^{L} d_\ell d_{\ell-1}$ and define

$$j(\theta, x) = \nabla_\theta F_\theta(x) \in \mathbb{R}^P, \qquad J_X(\theta) = \begin{bmatrix} j(\theta, X_1)^\top \\ \vdots \\ j(\theta, X_N)^\top \end{bmatrix}, \tag{194}$$

$$K_X(\theta) = J_X(\theta) J_X(\theta)^\top, \qquad k_\theta(x, X) = j(\theta, x)^\top J_X(\theta)^\top. \tag{195}$$

For one coordinate head, finite initialization-kernel GD is

$$\begin{aligned} g_0^K(x) &= 0, \\ g_{r+1}^K(x) &= g_r^K(x) - \frac{\eta}{N} k_{\theta_0}(x, X)\big(g_r^K(X) - Z\big). \end{aligned} \tag{196}$$

The infinite full-NTK trajectory g_r^∞ is defined by replacing K_{θ_0} with K_L^{full}.

We use the normalized empirical norm

$$\|v\|_{\ell_{2,N}} = N^{-1/2} \|v\|_2 \tag{197}$$

and define the *averaged* cross-kernel operator by

$$(\mathcal{A}_{K_\theta} v)(x) = \frac{1}{N} k_\theta(x, X) v = \frac{1}{N} \sum_{i=1}^{N} K_\theta(x, X_i) v_i. \tag{198}$$

If $\sup_{x,i} |K_\theta(x, X_i)| \le M$, Cauchy–Schwarz gives

$$\|\mathcal{A}_{K_\theta}\|_{\ell_{2,N} \to L^2(P_X)} \le M, \qquad \|k_\theta(\cdot, X)\|_{\ell_{2,N} \to L^2(P_X)} \le NM. \tag{199}$$

These bounds also hold for the supremum norm at the output. The factor N in the unaveraged operator cannot generally be replaced by $\sqrt{N}$: for $K \equiv 1$ and $v = (1, \dots, 1)$, the input norm is one and $k(x, X)v = N$. All population comparisons below use (198) consistently.

F.2.3 TAYLOR REMAINDERS

For the GD step $\Delta\theta_r = \theta_{r+1} - \theta_r$, define

$$e_r^X = F_{\theta_r + \Delta\theta_r}^c(X) - F_{\theta_r}^c(X) - J_X(\theta_r)\Delta\theta_r, \tag{200}$$

$$e_r^P(x) = F_{\theta_r + \Delta\theta_r}^c(x) - F_{\theta_r}^c(x) - j(\theta_r, x)^\top \Delta\theta_r. \tag{201}$$

Remark F.2 (Why centering is algorithmic rather than statistical). Without centering, the zero-start kernel comparison would have the random shifted target $Z - F_{\theta_0}(X)$ and the population oracle would need a separate bound on the initialization function. The centered model avoids that term while preserving iid Gaussian initialization, a trainable readout, the product metric, and the KMM kernel.

F.2.4 HORIZON-BASED DETERMINISTIC COUPLING

Proposition F.3 (Finite-horizon neural-to-kernel coupling)**.** *Fix a horizon R. Suppose, uniformly over $r \le R$,*

$$\|J_X(\theta_r) - J_X(\theta_0)\|_{\mathrm{op}} \le \varepsilon_J, \tag{202}$$

$$\left\|\mathcal{A}_{K_{\theta_r}} - \mathcal{A}_{K_{\theta_0}}\right\|_{\ell_{2,N}\to L^2(P_X)} \le \delta_{k,P}, \tag{203}$$

$$N^{-1/2}\left\|e_r^X\right\|_2 \le \tau_X \left\|F^c_{\theta_r}(X) - Z\right\|_{\ell_{2,N}}, \tag{204}$$

$$\left\|e_r^P\right\|_{L^2(P_X)} \le \tau_P \left\|F^c_{\theta_r}(X) - Z\right\|_{\ell_{2,N}}. \tag{205}$$

Let $J_0 = \|J_X(\theta_0)\|_{\mathrm{op}}$ and

$$\delta_K = 2J_0\varepsilon_J + \varepsilon_J^2. \tag{206}$$

Assume

$$\left\|\mathcal{A}_{K_{\theta_0}}\right\|_{\ell_{2,N}\to L^2(P_X)} \le \kappa_{0,P}, \qquad \sup_{r\le R}\left\|F^c_{\theta_r}(X) - Z\right\|_{\ell_{2,N}} \le B_0, \tag{207}$$

and that $I - (\eta/N)K_X(\theta_0)$ is a positive-semidefinite contraction. Then, for $d_r = F^c_{\theta_r}(X) - g_r^K(X)$,

$$\sup_{r\le R}\|d_r\|_{\ell_{2,N}} \le B_0 R\left(\frac{\eta\delta_K}{N} + \tau_X\right), \tag{208}$$

and

$$\sup_{r\le R}\left\|F^c_{\theta_r} - g_r^K\right\|_{L^2(P_X)} \le B_0\left[\eta R^2\kappa_{0,P}\left(\frac{\eta\delta_K}{N} + \tau_X\right) + \eta R\delta_{k,P} + R\tau_P\right]. \tag{209}$$

No lower tangent-Gram eigenvalue is used in these two conclusions.

Proof. Let $\varrho_r = F^c_{\theta_r}(X) - Z$. Subtracting the finite-kernel sample recursion (196) from the nonlinear sample recursion gives

$$d_{r+1} = \left(I - \frac{\eta}{N}K_X(\theta_0)\right)d_r - \frac{\eta}{N}\big(K_X(\theta_r) - K_X(\theta_0)\big)\varrho_r + e_r^X. \tag{210}$$

Every propagator in its Duhamel expansion is a contraction. Thus one sums the R forcing terms once, rather than combining a residual geometric sum with a second resolvent sum, and obtains (208).

For the population discrepancy $d_r^P = F^c_{\theta_r} - g_r^K$, subtraction at a fresh test input gives the exact recursion

$$d_{r+1}^P = d_r^P - \eta\mathcal{A}_{K_{\theta_0}}d_r - \eta(\mathcal{A}_{K_{\theta_r}} - \mathcal{A}_{K_{\theta_0}})\varrho_r + e_r^P. \tag{211}$$

The empirical argument above gives $\|d_r\|_{\ell_{2,N}} \le B_0 r(\eta\delta_K/N + \tau_X)$. Take population norms in (211), insert the averaged-operator envelope (207), and sum over $r < R$. Since $\sum_{r<R} r \le R^2/2$, the resulting bound is at most (209). This argument uses the classical finite-horizon Duhamel expansion and contraction of a positive-semidefinite kernel update; it uses no inverse minimum eigenvalue or PL inequality. □

Corollary F.4 (Calibrated horizon tolerances)**.** *For a target $\varepsilon_{\mathrm{coup},N}$, (209) is at most $\varepsilon_{\mathrm{coup},N}$ whenever*

$$\frac{\eta\delta_K}{N} + \tau_X \le \frac{\varepsilon_{\mathrm{coup},N}}{4B_0\eta R^2\kappa_{0,P}}, \tag{212}$$

$$\delta_{k,P} \le \frac{\varepsilon_{\mathrm{coup},N}}{4B_0\eta R}, \tag{213}$$

$$\tau_P \le \frac{\varepsilon_{\mathrm{coup},N}}{4B_0 R}. \tag{214}$$

A sufficient Jacobian tolerance for any desired value of δ_K is

$$\varepsilon_J \le \min\left\{J_0, \frac{\delta_K}{3J_0}\right\}. \tag{215}$$

F.3 Local Perturbation for Controlled Pyramids with a Trainable Readout

The following arguments prove the local perturbation estimates for the independently initialized, all-trained architecture used in the theorem. They retain the gate-pattern entropy and all logarithmic factors explicitly.

F.3.1 Source correspondence and attribution

The relevant locations in Zhou et al. Zhou et al. (2026), version 1, are Lemmas 15 (initialization), 17 (forward perturbation), 18 (gate counts), 19–20 (products and backward perturbation), 22 (near-linearity), and 16 (uniform kernel concentration). That work's principal setup uses a paired last hidden layer and a fixed signed readout. In particular, Lemma 20's proof invokes a cancellation from that initialization. Our independent Gaussian, all-trained network requires its own argument. The following proofs address that architecture directly, rather than assert that these source lemmas already cover it. We credit that work for the related proof strategy; the bounds used below follow from the internal proofs here.

F.3.2 Notation and elementary probability tools

Let $H = L - 1 \geq 1$, $d_H = m$, and suppose

$$m \leq d_\ell \leq \kappa_{\mathrm{pyr}}^{H-1} m \quad (1 \leq \ell \leq H), \qquad \kappa_{\mathrm{pyr}} \geq 1.$$

Write $d_{\max} = \max_{1\leq \ell \leq H} d_\ell$. The controlled-pyramid condition in the manuscript implies these inequalities. Set

$$\alpha_\ell = \sqrt{2/d_\ell}, \quad h_0(x) = x, \quad h_\ell(x) = \alpha_\ell \sigma(W_\ell h_{\ell-1}(x)), \quad F_\theta(x) = a^\top h_H(x), \qquad x \in \mathbb{S}^{d_0-1}. \tag{216}$$

Here and in Section F.4, h_ℓ denotes the normalized $\bar{h}_\ell$ of (5), and $a = W_L^\top$. All entries of $(W_{1,0}, \ldots, W_{H,0}, a_0)$ are independent standard Gaussian variables. Set $\sigma'(0) = 0$ and let j_θ be the resulting chain-rule Jacobian, including the readout. All parameter gradients and GD steps use the product-Frobenius metric. For a parameter perturbation write

$$\|\Delta\theta\|_F^2 = \sum_{\ell=1}^{L} \|\Delta W_\ell\|_F^2, \qquad \|\Delta\theta\|_{\mathrm{op},\infty} = \max_{1\leq \ell \leq L} \|\Delta W_\ell\|_{\mathrm{op}}. \tag{217}$$

The readout row's operator norm is its Euclidean norm, and every product-Frobenius ball is contained in the operator ball of the same radius. For a deterministic radius $\mathcal{R} \geq 1$, put

$$\mathcal{B}_{\mathcal{R}} = \{\theta : \max(\max_{\ell \leq H} \|W_\ell - W_{\ell,0}\|_{\mathrm{op}}, \|a - a_0\|_2) \leq \mathcal{R}\} = \mathcal{B}_{\mathcal{R}}(\theta_0). \tag{218}$$

Throughout this and the full-NTK concentration appendix $C, c > 0$ may change from line to line and depend only on $(L, \kappa_{\mathrm{pyr}})$. Define

$$\nu_m = 1 + d_0 H \log(e\kappa_{\mathrm{pyr}} m) + \log(64 H d_x/\delta), \qquad 0 < \delta < 1. \tag{219}$$

For a calibrated architecture, κ_{pyr} means the selected $\kappa_{\mathrm{pyr},N}$. All displayed logarithms are retained in ν_m; no polynomial dependence on depth is asserted. Enlarging C absorbs fixed multiples of this entropy. Probability bounds are made simultaneous over d_x heads by the failure allocation already included in ν_m. Two prescribed radii, used below, are handled by splitting δ once more and enlarging C.

We record the elementary inequalities used in the proofs. If independent centered variables satisfy $\mathbb{E}|Y_i|^p \leq p! B^p$ for $p \geq 2$, expanding $\mathbb{E}e^{tY_i}$ and summing its geometric remainder gives $\log \mathbb{E}e^{tY_i} \leq Ct^2B^2$ for $|t| \leq c/B$. Multiplication of moment generating functions and optimization in t therefore give

$$\Pr\left(\left|n^{-1}\sum_i Y_i\right| > CB\{\sqrt{u/n} + u/n\}\right) \leq 2e^{-u}. \tag{220}$$

For Bernoulli variables of mean at most p, the same expansion, retaining the variance factor p, gives $\sum_i Y_i \leq C(np + u)$ with probability at least $1 - e^{-u}$, where here Y_i denotes the uncentered indicator.

For $g \sim N(0, I)$ and a deterministic real matrix A, replace A by $S = (A + A^\top)/2$ in $g^\top Ag$. Diagonalization and $\mathbb{E}e^{t\lambda(g_1^2-1)} = e^{-t\lambda}(1-2t\lambda)^{-1/2}$ show

$$|g^\top Ag - \operatorname{tr} A| \le C\{\|S\|_F \sqrt{u} + \|S\|_{\mathrm{op}} u\} \tag{221}$$

outside a set of probability $2e^{-u}$. Indeed, the logarithm of the product of these generating functions is at most $Ct^2 \|S\|_F^2$ when $|t| \|S\|_{\mathrm{op}} \le c$; optimizing t gives the stated bound. Applying this argument to $\left\|g^\top B\right\|_2^2$ also yields

$$\left\|g^\top B\right\|_2 \le C\{\|B\|_F + \|B\|_{\mathrm{op}} \sqrt{u}\} \tag{222}$$

with probability at least $1 - 2e^{-u}$. Finally, $1/4$-nets on the two unit spheres, of cardinalities at most 9^p and 9^q, and the scalar Gaussian tail give

$$\|G\|_{\mathrm{op}} \le C(\sqrt{p} + \sqrt{q} + \sqrt{u}) \tag{223}$$

for a $p \times q$ standard Gaussian matrix, outside probability $2e^{-u}$. The net controls the full operator norm by a fixed multiple of its maximum bilinear form on net pairs. These arguments apply conditionally whenever the matrix or subspace being tested is fixed before the Gaussian variable is drawn.

Lemma F.5 (Initialization geometry and margin counts)**.** *If $m \ge C\nu_m$, with probability at least $1 - \delta$ all normalized hidden operator norms and the normalized readout norm are bounded by C, and uniformly in $x \in \mathbb{S}^{d_0-1}$,*

$$c \le \|h_{\ell,0}(x)\|_2 \le C \qquad (0 \le \ell \le H). \tag{224}$$

For any prescribed τ with $cm^{-1/3} \le \tau \le C$, the same type of event, with its failure probability allocated separately, satisfies

$$\sup_x \#\{i : |(W_{\ell,0} h_{\ell-1,0}(x))_i| \le \tau\} \le C(d_\ell \tau + \nu_m). \tag{225}$$

A fixed number of choices of τ per layer can be included simultaneously.

Proof. The operator bounds follow from (223); $m \ge C\nu_m$ also controls d_0 and the Gaussian failure parameter. For fixed preceding weights, the possible initialization gate patterns through layer $\ell - 1$ partition the input into at most $\exp(Cd_0 H \log(ed_{\max}))$ cells, including lower-dimensional faces. To see this, on each existing cell every new preactivation is linear in x. An arrangement of n linear hyperplanes in dimension at most d_0 has at most $(Cn)^{2d_0}$ faces: select a linearly independent subset of vanishing constraints, then count regions of the remaining hyperplanes in the resulting subspace using the recurrence that a new hyperplane splits at most the number of regions on that hyperplane. Multiplication over the H layers gives the claimed bound. On each cell, $h_{\ell-1,0}(x)$ belongs to a fixed linear subspace of dimension at most d_0. These subspaces depend only on the preceding weights.

Condition on those weights. For a fixed unit vector v in one such subspace, $2d_\ell^{-1} \sum_i \sigma((W_{\ell,0}v)_i)^2$ has mean one and exponentially small constant-deviation probability by (220). The Gaussian squares have the required factorial moment bounds. Use a constant-resolution net of the unit sphere of each subspace. On the operator event, the map $v \mapsto \|\alpha_\ell \sigma(W_{\ell,0}v)\|_2$ is C-Lipschitz. Choose the net resolution sufficiently small, depending only on $(L, \kappa_{\mathrm{pyr}})$. A union bound then places this norm between $1/2$ and 2 for every unit vector in the union of subspaces. Conditional failure is at most $\exp(C\nu_m - cm)$. Induction and positive homogeneity give (224), with bounds 2^{-H} and 2^H sufficient for fixed depth.

For the margin estimate, first consider unit v in these same subspaces. A fixed vector has at most $C(d_\ell\tau + u)$ rows with $|(W_{\ell,0}v)_i| \le 2\tau$, except with probability e^{-u}: use the Gaussian small-ball probability $C\tau$ and the Bernoulli bound. Cover each unit subspace sphere at resolution $\tau/(C\sqrt{m})$. On the operator event, each coordinate changes by at most τ on moving to the net. The logarithm of the total net size is at most

$$Cd_0 H \log(ed_{\max}) + Cd_0 \log(C\sqrt{m}/\tau) \le C\nu_m.$$

A union bound with $u = C\nu_m$ proves the count uniformly over unit vectors. Divide $h_{\ell-1,0}(x)$ by its norm and use its lower bound in (224); changing τ by a fixed factor changes only C. For τ bounded below by a fixed constant the estimate is trivial after enlarging C. Conditional union bounds over layers complete the proof without asserting independence of successive layers' events. □

Lemma F.6 (Sparse restrictions of frozen backward rows)**.** *If $m \geq C\nu_m$, there is an event of probability at least $1-\delta$ on which the following holds simultaneously for every layer, initialization input gate pattern, and subset $S \subseteq [d_\ell]$ of size $s \geq 1$. Define*

$$B_{\ell,0}(x) = \left[\prod_{r=H}^{\ell+1} \alpha_r D_{r,0}(x) W_{r,0}\right] \alpha_\ell I, \qquad \Pi_S = \mathrm{diag}(\mathbf{1}_S).$$

Then

$$\left\|a_0^\top B_{\ell,0}(x)\Pi_S\right\|_2 \leq C\sqrt{\frac{s\log(ed_\ell/s)+\nu_m}{m}}. \tag{226}$$

Products are ordered from the output toward the input; an empty product is an identity matrix.

Proof. Condition on all hidden weights and their operator event. The pattern count in Lemma F.5 is at most $e^{C\nu_m}$. For every pattern, dense operator multiplication gives $\|B_{\ell,0}\|_{\mathrm{op}} \leq C/\sqrt{m}$ and $\|B_{\ell,0}\Pi_S\|_F \leq C\sqrt{s/m}$. The readout a_0 is independent of these matrices. Apply (222), union bound over at most $\binom{d_\ell}{s} \leq (ed_\ell/s)^s$ supports, and set $u = C\nu_m + 2s\log(ed_\ell/s)$. Union bounds over s, layers, and heads are absorbed in ν_m. This proves (226) even for a support selected after seeing the readout or the training trajectory, because the event holds for every support. A diagonal with entries of absolute value at most one and support S obeys the same bound. □

F.3.3 Initialization controls used by the trajectory argument

Proposition F.7 (Uniform pyramidal initialization controls)**.** *Suppose*

$$m \geq C_{L,\kappa_{\mathrm{pyr}}}\nu_m. \tag{227}$$

Then, simultaneously for all heads, with probability at least $1-\delta$,

$$\sup_{x\in\mathbb{S}^{d_0-1}} \left\|\bar{h}_{\ell,0}(x)\right\|_2 \leq C_{L,\kappa_{\mathrm{pyr}}}, \qquad \ell = 0,\dots,H, \tag{228}$$

$$\sup_{x\in\mathbb{S}^{d_0-1}} \left\|b_{\ell,0}(x)\right\|_2 \leq C_{L,\kappa_{\mathrm{pyr}}}, \qquad \ell = 1,\dots,H, \tag{229}$$

$$\sup_{x\in\mathbb{S}^{d_0-1}} \left\|j(\theta_0,x)\right\|_2 \leq J_\star := C_{L,\kappa_{\mathrm{pyr}}}\sqrt{L}, \tag{230}$$

$$\left\|J_X(\theta_0)\right\|_{\mathrm{op}} \leq J_\star\sqrt{N}. \tag{231}$$

For the averaged population cross-kernel operator $\mathsf{A}_{K_0}v = N^{-1}k_{\theta_0}(\cdot,X)v$, these bounds imply

$$\|\mathsf{A}_{K_0}\|_{\ell_{2,N}\to L^2(P_X)} \leq \kappa_{0,P} := J_\star^2, \qquad \|k_{\theta_0}(\cdot,X)\|_{\ell_{2,N}\to L^2(P_X)} \leq N\kappa_{0,P}. \tag{232}$$

The centered initial residual satisfies $\|Z_q\|_2 \leq \sqrt{N}$.

Proof. The forward bound is Lemma F.5. For backward rows, gate matrices are contractions, and normalized hidden operators and the normalized Gaussian readout are bounded on that lemma's event. Dense multiplication in (118), with the final factor α_ℓ, therefore bounds every initial backward row by $C_{L,\kappa_{\mathrm{pyr}}}$. The rank-one gradient-block formula (119) gives the pointwise Jacobian bound after summing squared Frobenius norms over the L blocks. Bounding the matrix operator norm by its Frobenius norm gives (231). For any v, the pointwise Cauchy–Schwarz estimate is

$$\left|N^{-1}j(\theta_0,x)^\top J_X(\theta_0)^\top v\right| \leq J_\star^2 N^{-1/2}\left\|v\right\|_2 = J_\star^2\left\|v\right\|_{\ell_{2,N}}.$$

Integrating proves (232). Finally, $F_{\theta_0}^c = 0$ and the normalized labels have absolute value at most one. □

F.3.4 Gate, forward, backward, Jacobian, and Taylor perturbations

Theorem F.8 (Local perturbation for iid Gaussian controlled pyramids)**.** *For a prescribed $\mathcal{R} \geq 1$, suppose*

$$m \geq C\nu_m^3\mathcal{R}^2. \tag{233}$$

Define

$$q(\mathcal{R}, m) = C\sqrt{\nu_m}\,\mathcal{R}^{1/3} m^{-1/6}. \tag{234}$$

With probability at least $1-\delta$*, uniformly over inputs and* $\theta, \widetilde{\theta} \in \mathcal{B}_{\mathcal{R}}$*,*

$$\#\operatorname{supp}(D_\ell(\theta, x) - D_{\ell,0}(x)) \le C(d_\ell \mathcal{R})^{2/3}, \tag{235}$$

$$\|h_\ell(\theta, x) - h_{\ell,0}(x)\|_2 \le C\mathcal{R}/\sqrt{m}, \tag{236}$$

$$\|b_\ell(\theta, x) - b_{\ell,0}(x)\|_2 \le q(\mathcal{R}, m), \tag{237}$$

$$\|j_\theta(x) - j_{\theta_0}(x)\|_2 \le q(\mathcal{R}, m), \tag{238}$$

$$|F_{\widetilde{\theta}}(x) - F_\theta(x) - \langle j_\theta(x), \widetilde{\theta} - \theta\rangle| \le Cq(\mathcal{R}, m) \left\|\widetilde{\theta} - \theta\right\|_{\mathrm{op},\infty}. \tag{239}$$

The backward row is $b_\ell = a^\top [\prod_{r=H}^{\ell+1} \alpha_r D_r W_r] \alpha_\ell D_\ell$*. Write* $q = q(\mathcal{R}, m)$ *inside the proof. All constants include fixed-depth factors. The support-count bound also gives* $d_\ell^{-1} \|D_\ell(\theta, x) - D_{\ell,0}(x)\|_0 \le C\mathcal{R}^{2/3} m^{-1/3}$*. The same pointwise bounds imply bounds in every population* $L^2(P_X)$ *norm.*

Proof. The source correspondence above credits the related perturbation strategy of Zhou et al. (2026); we prove the stated architecture-specific bounds using the independent Gaussian readout and the uniform events just established.

All normalized hidden operators in the ball are bounded by C since $\mathcal{R}/\sqrt{m} \le c$. ReLU is 1-Lipschitz, so, with $e_\ell = \sup_x \|h_\ell(\theta, x) - h_{\ell,0}(x)\|_2$,

$$e_\ell \le Ce_{\ell-1} + C\mathcal{R}/\sqrt{m}, \qquad e_0 = 0.$$

This proves (236) directly, without a sparse-product argument. It also gives

$$\sup_x \|W_\ell h_{\ell-1}(\theta, x) - W_{\ell,0} h_{\ell-1,0}(x)\|_2 \le C\mathcal{R}.$$

A changed gate either had initial absolute preactivation at most τ or requires a preactivation change of magnitude exceeding τ. Lemma F.5 therefore bounds its count by

$$C(d_\ell \tau + \nu_m + \mathcal{R}^2/\tau^2).$$

Take $\tau = (\mathcal{R}^2/d_\ell)^{1/3}$, with fixed constants adjusted for comparable widths. The conditions of the margin lemma hold, or the count bound is trivial when this threshold is a fixed constant or larger. Since $m \ge C\nu_m^3 \mathcal{R}^2$ and $\mathcal{R} \ge 1$, the additive ν_m is absorbed by $(d_\ell \mathcal{R})^{2/3}$. This proves (235).

For backward drift, first split $a = a_0 + (a - a_0)$. The second part has norm at most $C\mathcal{R}/\sqrt{m}$ by dense operator bounds and the final α_ℓ factor. For the part with a_0, telescope the product from its output end so that every factor to the left of the changed factor is at initialization and every factor to its right may be perturbed. For a layer factor use

$$D_r W_r - D_{r,0} W_{r,0} = (D_r - D_{r,0}) W_r + D_{r,0}(W_r - W_{r,0}).$$

A gate term has frozen prefix $a_0^\top B_{r,0}$ restricted to at most $C(d_r \mathcal{R})^{2/3}$ coordinates. Its remaining product has bounded operator norm: when present, the unnormalized W_r is followed by a product with an end factor of order $m^{-1/2}$. Lemma F.6 thus bounds each gate term by

$$C\sqrt{\frac{(d_r\mathcal{R})^{2/3}\log(ed_r) + \nu_m}{m}} \le C\sqrt{\nu_m}\,\mathcal{R}^{1/3} m^{-1/6}.$$

A weight term has a bounded frozen row, one weight perturbation of norm $\mathcal{R}$, and a remaining product of norm $C/\sqrt{m}$, hence is $O(\mathcal{R}/\sqrt{m})$. There are at most $2H+1$ terms. This proves (237). The independence used was only that of a_0 and the frozen hidden matrices in Lemma F.6; no independence of trained gates or trained weights was invoked.

The hidden gradient block is $b_\ell^\top h_{\ell-1}^\top$, interpreted as the matrix with entries $b_{\ell,i} h_{\ell-1,j}$, and the readout gradient is h_H. Initial forward and backward norms are bounded by C. Subtract the two rank-one blocks, use the forward and backward bounds, and sum squared Frobenius norms over the L blocks. This gives (238). More specifically, each hidden block difference has nuclear norm at most Cq,

since it is the sum of two rank-one matrices with the stated factor bounds; the readout difference has Euclidean norm at most Cq.

The segment between θ and $\widetilde{\theta}$ stays in the convex ball. Along it the network output is absolutely continuous and piecewise polynomial. Its derivative agrees almost everywhere with the chain-rule Jacobian paired with the direction. If a preactivation is identically zero on a segment piece, its ReLU output has zero derivative on that piece, consistent with the convention at zero. Integrate the difference of Jacobians along the segment. Their block nuclear norms relative to one another are bounded by Cq, by subtracting their initialization values. The inequality $|\operatorname{tr}(A^\top B)| \le \|A\|_* \|B\|_{\mathrm{op}}$, summed over layers, proves (239). It also proves the weaker but useful bound $Cq \left\|\widetilde{\theta} - \theta\right\|_F$. □

F.4 Finite-Width Concentration of the Standard Full NTK

Use the notation and elementary Gaussian inequalities of Section F.3, including the entropy ν_m in (219). All constants depend only on the fixed depth and controlled width ratio. The exact initialization retains every tangent block, including the trainable-readout covariance block.

This proof uses forward derivative matrices and their traces. It avoids any assumption that a backward vector is independent of the weights that created its gates.

Lemma F.9 (A gated Gaussian quadratic form)**.** *Let $w \sim N(0, I_n)$, let u, v be nonzero deterministic vectors, and put $\chi(w) = \mathbf{1}\{w^\top u > 0, w^\top v > 0\}$ and $p = \mathbb{E}\chi(w)$. For any deterministic real $n \times n$ matrix A,*

$$\left|\mathbb{E}[\chi(w) w^\top A w] - p \operatorname{tr} A\right| \le C \|A\|_{\mathrm{op}}. \tag{240}$$

If $n \le Cm$ and $\|A\|_{\mathrm{op}} \le C/m$, the centered variable $\chi(w) w^\top A w$ has factorial moment bounds with a constant depending only on these C's.

Proof. Let P project onto $\operatorname{span}\{u, v\}$, of rank at most two. The indicator depends only on Pw. The vector $(I - P)w$ is independent of Pw and centered. The mixed quadratic terms consequently have zero expectation even after multiplication by χ. The perpendicular quadratic term has expectation $p \operatorname{tr}((I - P)A)$. Subtracting $p \operatorname{tr} A$ leaves

$$\mathbb{E}[\chi(w)(Pw)^\top A (Pw)] - p \operatorname{tr}(PA),$$

whose magnitude is at most $\|A\|_{\mathrm{op}} (\mathbb{E}\|Pw\|_2^2 + \operatorname{rank} P) \le 4\|A\|_{\mathrm{op}}$. For the moment claim, replace A by its symmetric part. Its absolute eigenvalues sum to at most $n\|A\|_{\mathrm{op}} \le C$. The absolute quadratic form is bounded by $\sum_i |\lambda_i| g_i^2$. Jensen's inequality with weights proportional to $|\lambda_i|$, followed by the Gaussian square moment formula, bounds its pth moment by $p! C^p$. Multiplication by an indicator and centering preserve this bound after changing C. □

Theorem F.10 (Uniform full-NTK concentration with explicit logarithms)**.** *Suppose*

$$m \ge C_{L, \kappa_{\mathrm{pyr}}} \nu_m^3. \tag{241}$$

With probability at least $1 - \delta$, simultaneously over the heads,

$$\sup_{x, y \in \mathbb{S}^{d_0 - 1}} |K_{\theta_0}(x, y) - K_L^{\mathrm{full}}(x, y)| \le C\sqrt{\nu_m}\, m^{-1/6} =: \epsilon_K. \tag{242}$$

Here $K_{\theta_0} = \langle j_{\theta_0}(x), j_{\theta_0}(y)\rangle$ includes every hidden block and the readout block. If $\epsilon_K \le 1$, then

$$\sup_x K_{\theta_0}(x, x) \le 2L. \tag{243}$$

Proof. The related sphere-uniform concentration strategy appears in Lemma 16 of Zhou et al. (2026), version 1. The derivation below proves the independent-Gaussian, all-trained case directly, including its readout block.

Step 1: exact forward derivative recursion. For fixed inputs x, y, let $\mathcal{J}_\ell(x)$ be the derivative of $h_{\ell,0}(x)$ with respect to all hidden parameters through layer ℓ, and set

$$A_\ell(x, y) = \mathcal{J}_\ell(x)\mathcal{J}_\ell(y)^\top, \quad c_\ell(x, y) = \langle h_{\ell,0}(x), h_{\ell,0}(y)\rangle, \quad t_\ell(x, y) = \operatorname{tr} A_\ell(x, y).$$

Initialize $A_0 = 0$, $t_0 = 0$, and $c_0 = \langle x, y\rangle$; the subscript zero here denotes network depth, not diffusion stopping time. New-layer and preceding-layer parameter coordinates are disjoint. Direct differentiation therefore gives the exact matrix identity

$$A_\ell = \frac{2}{d_\ell} D_{\ell,0}(x)\{c_{\ell-1} I + W_{\ell,0} A_{\ell-1} W_{\ell,0}^\top\} D_{\ell,0}(y). \tag{244}$$

On the dense operator and forward-norm events,

$$\|\mathcal{J}_\ell(x)\|_{\mathrm{op}} \le C\,\|\mathcal{J}_{\ell-1}(x)\|_{\mathrm{op}} + C/\sqrt{m}.$$

The new-parameter block has norm at most $\alpha_\ell \|h_{\ell-1,0}(x)\|_2$. The old-parameter block is the previous derivative multiplied on the left by $\alpha_\ell D_{\ell,0} W_{\ell,0}$, proving this inequality. Thus, uniformly in the inputs,

$$\|\mathcal{J}_\ell(x)\|_{\mathrm{op}} \le C/\sqrt{m}, \quad \|A_\ell(x,y)\|_{\mathrm{op}} \le C/m, \quad \|A_\ell(x,y)\|_F \le C/\sqrt{m}. \tag{245}$$

In particular, $|t_\ell| \le C$.

Step 2: covariance recursion on a deterministic input net. Take a c/m-net $\mathcal{V}$ of the input sphere, with c a sufficiently small architecture-dependent constant. It has $\log|\mathcal{V}| \le C d_0 \log(em) \le C\nu_m$. Condition on all weights before layer ℓ. The covariance c_ℓ is the average of d_ℓ independent variables $2\sigma(w^\top h_{\ell-1,0}(x))\sigma(w^\top h_{\ell-1,0}(y))$. Their factorial moment bounds follow from $2|ab| \le a^2 + b^2$ and the bounded previous forward norms. Apply (220) with failure parameter $C\nu_m$, then union bound over net pairs and layers. Every covariance differs from its conditional mean by at most

$$e_m = C\{\sqrt{\nu_m/m} + \nu_m/m\}.$$

For a centered Gaussian pair with variances a, b and covariance c, its doubled ReLU covariance is

$$V(a,b,c) = \sqrt{ab}\,\phi(c/\sqrt{ab}), \quad \phi(z) = \frac{\sqrt{1-z^2} + (\pi - \arccos z)z}{\pi}.$$

This identity follows by integrating in polar coordinates in the span of the two Gaussian linear forms. Since $0 \le \phi'(z) \le 1$, V is Lipschitz on the bounded variance range bounded away from zero in Lemma F.5. Induction gives, on all net pairs,

$$|c_\ell(x,y) - C_\ell(\langle x, y\rangle)| \le C e_m, \qquad C_0(z) = z, \quad C_\ell(z) = \phi(C_{\ell-1}(z)). \tag{246}$$

The diagonal case is included, so each random variance differs from one by at most Ce_m.

Step 3: the hidden-derivative trace recursion. Taking the trace in (244), and conditioning on preceding layers, gives an average of independent row variables

$$t_\ell = \frac{2}{d_\ell} \sum_{i=1}^{d_\ell} \chi(w_i)\{c_{\ell-1} + w_i^\top A_{\ell-1} w_i\}.$$

Lemma F.9, (245), and (220) imply simultaneously over the net

$$t_\ell = \psi(\widehat{\rho}_{\ell-1})(c_{\ell-1} + t_{\ell-1}) + r_\ell, \qquad |r_\ell| \le C(e_m + m^{-1}), \tag{247}$$

where $\widehat{\rho}_{\ell-1} = c_{\ell-1}(x,y)/\sqrt{c_{\ell-1}(x,x)c_{\ell-1}(y,y)}$ and $\psi(z) = 1 - \arccos(z)/\pi = 2\mathbb{E}\chi$. This calculation includes the dependence of the gates on their Gaussian row, through the rank-two correction in (240). It does not discard that dependence. The bound $|\psi(z) - \psi(z')| \le C\sqrt{|z - z'|}$ follows by integrating $(1-z^2)^{-1/2}$, including at the endpoints. Equation (246) therefore makes the gate expectation error at most $C\sqrt{e_m}$. Induct on (247) to obtain

$$|t_\ell - T_\ell(\langle x, y\rangle)| \le C\sqrt{e_m}, \quad T_0 = 0, \quad T_\ell = \psi(C_{\ell-1})(C_{\ell-1} + T_{\ell-1}).$$

All conditional estimates above are made on the event of good preceding layers. Bounding each next-layer conditional failure and adding these probabilities, together with failures of the dense events, justifies their simultaneous use; independence of the good events is unnecessary.

Step 4: the Gaussian readout and the full limiting kernel. The exact full kernel is

$$K_{\theta_0}(x,y) = c_H(x,y) + a_0^\top A_H(x,y) a_0. \tag{248}$$

Conditional on the hidden weights, apply (221) to the independent a_0 and use (245). Union bounding over net pairs gives an additional error at most Ce_m. Thus the net kernel converges to $C_H + T_H$ with error $C\sqrt{e_m}$. Writing $\Theta_0(z) = z$ and

$$\Theta_\ell(z) = C_\ell(z) + \psi(C_{\ell-1}(z))\Theta_{\ell-1}(z)$$

shows $C_H + T_H = \Theta_H = K_L^{\text{full}}$. This is the full recursion; c_H in (248) is exactly the trainable-readout covariance block. On the diagonal $\Theta_H(1) = H + 1 = L$.

Step 5: extend from the net to the sphere. Use Theorem F.8 at radius one. For unit x, y with $\|x - y\|_2 \le c/m$, define

$$W_1' = W_{1,0} + W_{1,0}(y - x)x^\top,$$

leaving other weights unchanged. Its perturbation has norm at most $C\sqrt{m}\,\|x - y\|_2 \le 1$. The hidden forward passes at (θ', x) and (θ_0, y) agree exactly. All gradient blocks agree except that the first block has input factor x instead of y; its difference is at most $C\,\|x - y\|_2$. Consequently

$$\|j_{\theta_0}(x) - j_{\theta_0}(y)\|_2 \le q(1, m) + C\,\|x - y\|_2 .$$

The uniform initialization Jacobian norm is at most C by the dense forward/backward bounds. Hence moving either input of the finite kernel to the net costs at most $C\sqrt{\nu_m}\,m^{-1/6}$. The limiting recursion is $1/2$-Hölder in its initial inner product, by the Lipschitz property of ϕ and the $1/2$-Hölder property of ψ; its net-extension cost is at most $Cm^{-1/2}$. Finally $\sqrt{e_m} \le C(\nu_m/m)^{1/4} \le C\sqrt{\nu_m}\,m^{-1/6}$ under $m \ge C\nu_m^3$. Combining all three errors proves (242). The diagonal assertion follows from $L + 1 \le 2L$. □

F.4.1 Finite-to-infinite kernel trajectories

Lemma F.11 (Two-kernel gradient-descent comparison)**.** *Let K_1, K_2 be positive semidefinite kernels satisfying $\sup_x K_j(x, x) \le \kappa_K^2$, and let $g_r^{(j)}$ be zero-start full-batch kernel-GD trajectories on the same sample and labels with $|Z_i| \le 1$. If $0 < \eta\kappa_K^2 \le 1$ and $\epsilon_K = \sup_{x,x'} |K_1(x, x') - K_2(x, x')|$, then, writing $h = \eta R$, for every integer $R \ge 1$,*

$$\left\| g_R^{(1)} - g_R^{(2)} \right\|_{L^\infty} \le C\big(h + \kappa_K^2 h^2\big)\epsilon_K, \tag{249}$$

$$\left\| g_R^{(1)} - g_R^{(2)} \right\|_{L^2(P_X)}^2 \le C\big(h^2 + \kappa_K^4 h^4\big)\epsilon_K^2. \tag{250}$$

Proof. Let K_j^X be the Gram matrices, $G_r^{(j)} = g_r^{(j)}(X)$, and $e_r^{(j)} = G_r^{(j)} - Z$. Since $\left\|K_j^X\right\|_{\text{op}} \le N\kappa_K^2$, the matrices $I - (\eta/N)K_j^X$ are positive-semidefinite contractions and $\left\|e_r^{(j)}\right\|_2 \le \|Z\|_2 \le \sqrt{N}$. For $D_r = G_r^{(1)} - G_r^{(2)}$,

$$D_{r+1} = \left(I - \frac{\eta}{N}K_1^X\right) D_r + \frac{\eta}{N}(K_2^X - K_1^X)e_r^{(2)},$$

so $\|D_r\|_2 \le \eta r\sqrt{N}\epsilon_K$. At a test point, subtract the prediction recursions, use kernel Cauchy–Schwarz $\|k_{1,x}\|_2 \le \kappa_K^2\sqrt{N}$ and the uniform-gap bound $\|k_{1,x} - k_{2,x}\|_2 \le \sqrt{N}\epsilon_K$, then sum from 0 to $R - 1$. This proves (249); squaring proves (250). □

Remark F.12 (Why the finite-to-infinite branch is retained). A uniform pointwise gap between two kernels does not imply Loewner domination of their integral operators, nor does it embed the infinite-kernel RKHS into the finite-kernel RKHS. In particular, K_{θ_0} has finite rank, whereas the source-free comparators in Theorem E.17 are generic elements of the infinite full-NTK RKHS. The explicit trajectory comparison above is therefore retained rather than replacing the population oracle by one for K_{θ_0}.

F.5 Movement Bootstrap and Atomic Scales

The local theorem in Section F.3 applies on a prescribed operator ball. We show that centered GD remains in such a ball and record the empirical and population scales used in Section C. The proof makes the finite-trajectory displacement calculation explicit. All constants below depend only on fixed depth and the controlled width ratio; the logarithmic factor ν_m is retained.

F.5.1 SELF-CONSISTENT MOVEMENT

Proposition F.13 (Centered-GD movement bootstrap)**.** *Let $R \geq 1$ be an integer, $0 < \eta \leq 1/(4L)$, and $h = \eta R$. Train the centered empirical loss on any fixed sample with $|Z_i| \leq 1$. With ν_m as in (256), suppose*

$$m \geq C_{L,\kappa_{\mathrm{pyr,N}}} \nu_m^3 (1+h)^7. \tag{251}$$

On the intersection of the uniform initialization, full-NTK-concentration, and local-perturbation events, every scalar head satisfies

$$\|\theta_r - \theta_0\|_F^2 \leq 2\eta r, \qquad \left\|F_{\theta_r}^c(X) - Z\right\|_{\ell_{2,N}} \leq 2\|Z\|_{\ell_{2,N}} \leq 2, \qquad 0 \leq r \leq R. \tag{252}$$

No lower Gram-eigenvalue condition is used.

Proof. Set $\mathcal{R} = 4\sqrt{1+h}$ and let $Q = C_{L,\kappa_{\mathrm{pyr,N}}} \sqrt{\nu_m} \mathcal{R}^{1/3} m^{-1/6}$, including the Jacobian and Taylor constants in Q. The width condition implies (233) on this ball and makes both Qh and $Q\mathcal{R}$ as small as needed below. It also ensures that the initialization kernel error is at most one. By (243), the initial diagonal is at most $2L$. Increasing the width constant makes the Jacobian norm throughout the ball at most $J_b = \sqrt{3L}$, so $\eta J_b^2 \leq 3/4$.

Write $\varrho_r = F_{\theta_r}^c(X) - Z$, $e_r = \|\varrho_r\|_{\ell_{2,N}}$, $d_r = \theta_r - \theta_0$, and $g_r = \nabla\widehat{\mathcal{L}}(\theta_r)$. Inside the ball, $\|g_r\|_F \leq J_b e_r$. The sample Taylor expansion is

$$\varrho_{r+1} = \left(I - \frac{\eta}{N} K_X(\theta_r)\right)\varrho_r + e_r^X, \tag{253}$$

with

$$\left\|e_r^X\right\|_{\ell_{2,N}} \leq Q\eta J_b e_r. \tag{254}$$

The kernel has largest eigenvalue at most NJ_b^2, hence its update matrix is a positive-semidefinite contraction. Induction therefore gives $e_r \leq (1+\eta J_b Q)^r e_0 \leq e^{J_b Q h} e_0 \leq 2e_0 \leq 2$.

For displacement, Taylor's formula between θ_r and θ_0 yields

$$J_X(\theta_r) d_r = F_{\theta_r}^c(X) + v_r, \qquad \|v_r\|_{\ell_{2,N}} \leq Q\|d_r\|_F.$$

Consequently, by Cauchy–Schwarz and $\|Z\|_{\ell_{2,N}} \leq 1$,

$$\langle g_r, d_r\rangle = \frac{1}{N}\varrho_r^\top J_X(\theta_r) d_r \geq e_r^2 - e_r - Q\|d_r\|_F e_r.$$

Expanding the next displacement gives

$$\begin{aligned}\|d_{r+1}\|_F^2 - \|d_r\|_F^2 &= -2\eta\langle g_r, d_r\rangle + \eta^2\|g_r\|_F^2 \\ &\leq -\eta(2-\eta J_b^2)e_r^2 + 2\eta e_r + 2\eta Q\|d_r\|_F e_r \\ &\leq \eta + 4\eta Q\mathcal{R} \leq 2\eta,\end{aligned} \tag{255}$$

using $-e_r^2 + 2e_r \leq 1$ and $Q\mathcal{R} \leq 1/4$.

To justify both Taylor applications without circularity, use a simultaneous induction. If $\|d_r\|_F \leq \sqrt{2h}$ and $e_r \leq 2$, then $\eta\|g_r\|_F \leq 2\eta J_b \leq 1$; the entire next-step segment stays in the larger ball because $\sqrt{2h} + 1 < 4\sqrt{1+h}$. The residual and displacement inequalities can therefore be applied to that step. They establish the same bounds at $r+1$ and close the induction, since the product operator norm is at most the product Frobenius norm. This proof uses upper Jacobian and Taylor bounds only. □

F.5.2 ATOMIC EMPIRICAL AND POPULATION SCALES

At $R = R_\star$, set

$$\begin{aligned}R_{\mathrm{op},\star} &= 4\sqrt{1+\eta R_\star}, \\ \nu_m &= 1 + d_0 H\log(e\kappa_{\mathrm{pyr,N}} m) + \log(64 H d_x/\delta), \\ q_\star &= C_{L,\kappa_{\mathrm{pyr,N}}}\sqrt{\nu_m}(1+\eta R_\star)^{1/6} m^{-1/6}.\end{aligned} \tag{256}$$

The radius includes every next-step segment, not just the iterate endpoints.

Proposition F.14 (Jacobian, averaged cross-kernel, and Taylor scales)**.** *On the movement and local-perturbation events, uniformly for $r \le R_\star$,*

$$\|J_X(\theta_r) - J_X(\theta_0)\|_{\mathrm{op}} \le C\sqrt{N}\, q_\star, \tag{257}$$

$$\|K_X(\theta_r) - K_X(\theta_0)\|_{\mathrm{op}} \le C_{L,\kappa_{\mathrm{pyr,N}}} N q_\star, \tag{258}$$

$$\left\|\mathcal{A}_{K_{\theta_r}} - \mathcal{A}_{K_{\theta_0}}\right\|_{\ell_{2,N}\to L^2(P_X)} \le C_{L,\kappa_{\mathrm{pyr,N}}} q_\star, \tag{259}$$

$$N^{-1/2}\left\|e_r^X\right\|_2 \le C_{L,\kappa_{\mathrm{pyr,N}}} \eta q_\star \left\|\varrho_r\right\|_{\ell_{2,N}}, \tag{260}$$

$$\left\|e_r^P\right\|_{L^2(P_X)} \le C_{L,\kappa_{\mathrm{pyr,N}}} \eta q_\star \left\|\varrho_r\right\|_{\ell_{2,N}}. \tag{261}$$

Here $\mathcal{A}$ is the averaged operator in (198); the unaveraged cross-kernel drift is bounded by $C_{L,\kappa_{\mathrm{pyr,N}}} N q_\star$ in the same norms.

Proof. The pointwise Jacobian drift in (238), evaluated at $R_{\mathrm{op},\star}$, is at most $Cq_\star$. Bounding the empirical Jacobian operator norm by its Frobenius norm proves (257). The initialization and perturbed Jacobian norms are bounded uniformly by constants, and therefore

$$\sup_{x,y} |K_{\theta_r}(x,y) - K_{\theta_0}(x,y)| \le C_{L,\kappa_{\mathrm{pyr,N}}} q_\star.$$

This proves the Gram bound and, by (199), the averaged cross-kernel bound. The same argument bounds the initialization averaged operator by $C_{L,\kappa_{\mathrm{pyr,N}}}$. Finally, $\|\Delta\theta_r\|_F \le C_{L,\kappa_{\mathrm{pyr,N}}} \eta \left\|\varrho_r\right\|_{\ell_{2,N}}$. Apply (239) on the next-step segment guaranteed by the movement proof. Its uniform pointwise remainder bound implies both the empirical and population Taylor scales. In fact the population conclusions also hold in the supremum norm on the input sphere. □

G Complete Training and Conditional Sampling Procedure

G.1 Information supplied to the theoretical schedule

The deterministic calibration is parameterized by a specified conditional-law class. To instantiate it numerically, the learner is supplied with positive lower bounds on admissible λ_X, λ_Y and finite upper bounds on M_X, M_Y in Assumption 3.1; the integer regularity order k and upper bounds on $C_\alpha, a_\alpha, b_\alpha$ for $1 \le |\alpha| \le k$ in Assumption 3.2; and the dimensions, depth, and adjacent-width bound. The effective regularity is $\bar{k} = \min\{k, (D+2)/2\}$. We use the same symbols for conservative supplied values. For original-law recovery, the supplied information also includes β and uniform upper bounds on the density and its Hölder seminorm in Assumption 3.3. The stopped-law theorem does not use this last information.

These bounds determine valid moment, clipping, localization, and extension envelopes. In particular, A in Section B is a supplied or derived upper bound on the relevant extension norm, not that unknown norm itself. All unspecified constants in the displayed sufficient conditions must be replaced by valid upper bounds from the proofs, including the architecture envelope $\Gamma_{L,\kappa_{\mathrm{pyr},N}}$ in (52). The proofs establish finite envelopes at fixed architecture; they do not provide a numerically optimized constant table. Thus the result gives a theoretical schedule once valid class and proof bounds are fixed, rather than a tuning rule determined by N, ε, δ alone. Estimating the class bounds, adapting to unknown smoothness, and choosing stopping times from the data would require additional analysis.

G.2 Training and sampling procedure

Algorithm 1 combines the learning and generation stages of the method. Training produces a denoiser and its associated score estimator. Conditional generation then uses this fixed estimator and fresh Gaussian noise; it can be repeated without retraining.

Inputs and parameter calibration. The inputs are N i.i.d. clean pairs $\{(X_{0,i}, Y_i)\}_{i=1}^N$, a generation condition $y \in \mathbb{R}^{d_y}$, a target accuracy $\varepsilon > 0$, a confidence parameter $\delta \in (0,1)$, and the numerical class bounds specified in Section G.1. Fix a depth $L \ge 3$ and a diffusion stopping time $t_0 > 0$. Section B prescribes the terminal time $T \ge 1$, grid size M, clipping threshold B_N, input scale ρ_N, and sufficient layer widths. The grid $t_0 < \cdots < t_M = T$ satisfies $\Delta_j = t_{j+1} - t_j \le \kappa_{\mathrm{grid}} \min\{1, t_j\}$, with $0 < \kappa_{\mathrm{grid}} \le 1$ and $\kappa_{\mathrm{grid}} = \widetilde{O}(M^{-1})$. The widths satisfy (4) and (59); Lemma B.5 gives an explicit sufficient choice.

For completeness, the deterministic training schedule uses

$$s = \frac{D+2}{2}, \qquad \bar{k} = \min\{k, s\}, \qquad b = \frac{\bar{k}}{s}, \qquad \gamma = \frac{D+1}{D+2}, \qquad D = d_x + d_y.$$

Let $A = A_{\bar{k},N}(t_0) \ge 1$ be the deterministic upper bound on the required normalized extension norms, obtained from the supplied class bounds using integer derivative order $j = \lceil \bar{k} \rceil \le k$. This is the $u = \bar{k}$ specialization of the general calibration in Section B, with the extension envelope in (148). Put $\mathcal{L}_N = \log(CNd_x/\delta)$. Choose $0 < \eta \le 1/(4L)$ and set

$$\lambda_N = \max\left\{ \left(\frac{\mathcal{L}_N}{NA^2} \right)^{1/(b+\gamma)}, \frac{C_0 \mathcal{L}_N}{N} \right\}, \qquad R_\star = \left\lceil (\eta \lambda_N)^{-1} \right\rceil.$$

The constants and sample-size conditions are those of Proposition B.3 and Theorem 1. These choices are fixed before training. In particular, the comparison floor $C_0\mathcal{L}_N/N$ is retained, and A is a supplied or derived bound rather than an unknown population norm.

For original-law recovery, choose t_0 according to Corollary 1 and evaluate the extension envelope and resource bounds at that shrinking time. The main corollary also uses $u = \bar{k}$. To invoke Theorem B.9 at a lower order $1 \le u \le \bar{k}$, instead use $b = u/s$ and $A = A_{u,N}(t_0)$, constructed at $j = \lceil u \rceil$, and recalibrate the sufficient sample size, stopping horizon, and width at that same order.

Network and preprocessing definitions. Each coordinate uses the scalar network in (5), with $d_0 = D + 2$, $H = L - 1$ hidden layers, and $d_L = 1$. All hidden matrices and the readout are trained.

Throughout the algorithm, use

$$\alpha_t = e^{-t}, \qquad \sigma_t^2 = 1 - e^{-2t}, \qquad \psi_{\rho_N}(v) = \frac{(v/\rho_N, 1)}{\sqrt{1 + \|v\|_2^2/\rho_N^2}}.$$

Define the score-weighted time probabilities by

$$\mathcal{W} = \sum_{j=0}^{M-1} \Delta_j \frac{\alpha_{t_{j+1}}^2}{\sigma_{t_{j+1}}^4}, \qquad \pi_j = \frac{\Delta_j \alpha_{t_{j+1}}^2}{\mathcal{W}\sigma_{t_{j+1}}^4}.$$

Thus the network input contains the noisy observation, condition, and right-endpoint diffusion time.

Algorithm 1 Finite neural training and conditional generation

Require: Clean pairs $\{(X_{0,i}, Y_i)\}_{i=1}^N$, condition y, and calibrated grid, $B_N, \rho_N, \eta, R_\star$, and layer widths.
Ensure: Denoiser $\widehat{f}$, score estimator $\widehat{s}$, and one generated sample $\widehat{X}_0$ at forward time t_0.

Phase I: Training
for $i = 1, \dots, N$ **do**
  Draw J_i with $\mathbb{P}(J_i = j) = \pi_j$ and $\varepsilon_i \sim N(0, I_{d_x})$, independently of each other and of all other draws and clean data.
  Set $U_i \leftarrow (\alpha_{t_{J_i+1}} X_{0,i} + \sigma_{t_{J_i+1}} \varepsilon_i, Y_i, t_{J_i+1})$.
  Set $X_i \leftarrow \psi_{\rho_N}(U_i)$ and $\widetilde{Z}_{B_N,i,q} \leftarrow B_N^{-1} \operatorname{clip}(X_{0,i,q}, [-B_N, B_N])$ for every q.
end for
Freeze this dataset for all coordinates and all training steps.
for $q = 1, \dots, d_x$ **do**
  Initialize a separate network $\theta_{q,0}$ with all matrix entries independently distributed as $N(0,1)$.
  Retain a fixed copy of $F_{\theta_{q,0}}$ and cache $c_{i,q} \leftarrow F_{\theta_{q,0}}(X_i)$ for all i.
  for $r = 0, \dots, R_\star - 1$ **do**
    Compute $e_{i,q,r} \leftarrow F_{\theta_{q,r}}(X_i) - c_{i,q} - \widetilde{Z}_{B_N,i,q}$ for all i.
    Update all hidden matrices and the readout simultaneously:

$$\theta_{q,r+1} \leftarrow \theta_{q,r} - \frac{\eta}{N} \sum_{i=1}^N e_{i,q,r} \, \nabla_\theta F_\theta(X_i)|_{\theta=\theta_{q,r}} .$$

  end for
  Define $\widehat{f}_q(v) \leftarrow B_N[F_{\theta_{q,R_\star}}(\psi_{\rho_N}(v)) - F_{\theta_{q,0}}(\psi_{\rho_N}(v))]$.
end for
Assemble $\widehat{f}(v) \leftarrow (\widehat{f}_1(v), \dots, \widehat{f}_{d_x}(v))^\top$.
Define $\widehat{s}(x,t,y) \leftarrow (\alpha_t \widehat{f}(x,y,t) - x)/\sigma_t^2$.
Phase II: Conditional generation
Draw $\widehat{X}_M \sim N(0, I_{d_x})$ independently of training.
for $j = M-1, \dots, 0$ **do**
  Draw a fresh independent $Z_j \sim N(0, I_{d_x})$ and set $a_j \leftarrow \widehat{s}(\widehat{X}_{j+1}, t_{j+1}, y)$.
  Set $\widehat{X}_j \leftarrow e^{\Delta_j} \widehat{X}_{j+1} + 2(e^{\Delta_j} - 1) a_j + \sqrt{e^{2\Delta_j} - 1}\, Z_j$.
end for
return $\widehat{f}$, $\widehat{s}$, and $\widehat{X}_0$.

Frozen data and centered training. Each clean pair supplies one time index and one Gaussian augmentation, both retained throughout optimization. The coordinate networks share this dataset but have independent hidden layers and readouts; initialization is also independent of the training data. The cached values $c_{i,q}$ implement the centered loss in (6). The initialization networks remain fixed and are also evaluated at fresh inputs when computing $\widehat{f}$. Clipping applies to the training labels; no additional clipping of the learned denoiser or generated states is introduced. The stopping scale λ_N determines the training horizon, while the empirical objective is the squared loss in (6).

Sampling convention and reuse. Phase II evaluates the score at the current reverse state and the larger forward-time endpoint t_{j+1}, matching (75). The returned $\widehat{X}_0$ is the state at forward time t_0, despite its algorithmic index 0. Conditional on the trained predictor, its law is $\widehat{p}_{t_0}(\cdot \mid y)$. Further samples are obtained by repeating Phase II with fresh Gaussian draws and the desired condition.

The learned networks, initialization copies, input map, and training grid remain fixed. The time probabilities π_j are used to construct the training data; generation traverses every grid interval in reverse order. Lemmas C.2 and C.3 connect this procedure to the sampling guarantee.

H Examples and Non-Examples of Conditional Laws

In each family below the distribution of Y must have an exponential moment as in Assumption 3.1; a nondegenerate Gaussian law is one choice. All conditional-law and relative-derivative statements hold on the entire chosen domain $\mathbb{R}^{d_y}$.

Globally controlled Gaussian location–scale families. Let $\mu_y = N(m(y), \Sigma(y))$, where m, Σ are C^k, $\sup_y \|m(y)\| \le M$, and $cI \preceq \Sigma(y) \preceq CI$ for fixed $c, C > 0$. Assume derivatives of m, Σ through order k grow at most polynomially in y. Differentiating a Gaussian density gives relative derivatives that are polynomials in z with polynomially bounded coefficients in y; locally uniform Gaussian domination gives C^k weighted-TV differentiability. Moreover, for $Z \sim N(0, I)$ and any fixed $\lambda > 0$,

$$\sup_y \int e^{\lambda\|z\|} \mu_y(\mathrm{d}z) \le e^{\lambda M} \mathbb{E} e^{\lambda\sqrt{C}\|Z\|} < \infty.$$

Thus Assumptions 3.1 and 3.2 hold. Uniform ellipticity also bounds the Gaussian densities and their spatial gradients, giving Assumption 3.3 with $\beta = 1$. A concrete nonlinear example is $m(y) = a\tanh(v^\top y)$ and $\Sigma(y) = (\sigma_0^2 + \sigma_1^2 \ell(w^\top y))I$, where $\sigma_0 > 0$ and $\ell(u) = (1 + e^{-u})^{-1}$. Control only on the support of Y does not automatically verify the all-domain condition for an arbitrary chosen version.

Finite conditional mixtures. For varying components, the product rule requires polynomial envelopes for $|\partial^\alpha w_j|/w_j$, component relative derivatives, and weighted-TV differentiability through the required order. Bounds on the weights and their ordinary derivatives alone do not control these ratios. Uniform conditional exponential tails must be verified separately. The following fixed-component result gives a direct sufficient condition, followed by an explicit smooth Gaussian example. Structured mixture-learning results (Shah et al., 2023; Gatmiry et al., 2025) address complementary settings.

Lemma H.1 (Relative derivatives of a fixed-component mixture). *Let $\nu_1, \dots, \nu_J$ be fixed probability measures on $\mathbb{R}^{d_x}$, and let $w_j : \mathbb{R}^{d_y} \to (0, 1)$ be C^k functions with $\sum_j w_j = 1$. Suppose the component measures have the moments needed for the weighted-TV norm under consideration. If, for each $1 \le |\alpha| \le k$,*

$$|\partial^\alpha w_j(y)/w_j(y)| \le C_\alpha(1 + \|y\|)^{a_\alpha} \quad \text{for every } j, \tag{262}$$

then $\mu_y = \sum_j w_j(y)\nu_j$ is C^k in weighted TV and admits

$$\partial^\alpha \mu_y = S_\alpha(\cdot, y)\mu_y, \qquad |S_\alpha(z, y)| \le C_\alpha(1 + \|y\|)^{a_\alpha}. \tag{263}$$

If $\max_j \int e^{\lambda\|z\|}\nu_j(dz) < \infty$, then the conditional exponential-tail bound is uniform in y.

Proof. Use $\nu = \sum_j \nu_j$ as a common reference measure, and write $f_j = d\nu_j/d\nu$, $f_y = \sum_j w_j(y) f_j$. Since the sum is finite, $y \mapsto \mu_y$ is a finite linear combination of fixed finite-norm vectors in the weighted measure space. Differentiability there follows from that of the scalar weights. At points where $f_y > 0$, put $\pi_j(z, y) = w_j(y) f_j(z)/f_y(z)$. Then

$$S_\alpha(z, y) = \sum_j \pi_j(z, y) \frac{\partial^\alpha w_j(y)}{w_j(y)}.$$

The responsibilities are nonnegative and sum to one, proving the bound. The value of S_α on a μ_y-null set is immaterial. Finally, $\int e^{\lambda\|z\|}\mu_y(dz)$ is a convex combination of the corresponding component moments. □

Proposition H.2 (Smooth conditional mixture with bounded relative derivatives). *Let $d_x, d_y \ge 1$, $a \in \mathbb{R}^{d_x}$, $v \in \mathbb{R}^{d_y}$, $\sigma > 0$, and*

$$\begin{aligned} Y \sim N(0, I_{d_y}), \qquad w(y) = \frac{1}{1 + e^{-v^\top y}}, \\ \mu_y = w(y)N(a, \sigma^2 I_{d_x}) + (1 - w(y))N(-a, \sigma^2 I_{d_x}). \end{aligned} \tag{264}$$

This family satisfies Assumptions 3.1 and 3.2 at every finite order k. One may take $a_\alpha = b_\alpha = 0$ for all relative derivatives. Consequently $b_k = 0$, and the conservative envelope (113) gives $\chi_k = 4k$. Assumption 3.3 holds with $\beta = 1$.

For $d_x = d_y = 1$, $v \neq 0$, and $a > \sigma$, the conditional density at $y = 0$ is bimodal and not log-concave; failure of log-concavity also holds for a neighborhood of conditioning values with positive probability.

Proof. Write $\phi_+(z)$ and $\phi_-(z)$ for the two fixed Gaussian densities, $f(z \mid y) = w\phi_+ + (1-w)\phi_-$, and $\pi_+(z,y) = w\phi_+/f$. Let $\ell(u) = (1+e^{-u})^{-1}$. Its derivatives satisfy

$$\ell^{(n)}(u) = \ell(u)(1-\ell(u))P_{n-1}(\ell(u)), \qquad n \geq 1, \tag{265}$$

where $P_0 = 1$ and

$$P_n(t) = (1-2t)P_{n-1}(t) + t(1-t)P'_{n-1}(t).$$

This follows by differentiation and induction. Each P_n is bounded on $[0,1]$. Therefore both $|\ell^{(n)}|/\ell$ and $|\ell^{(n)}|/(1-\ell)$ are bounded at each fixed finite order. For a multi-index α with $n = |\alpha|$, let $v^\alpha = \prod_i v_i^{\alpha_i}$. Direct differentiation gives

$$\begin{aligned} S_\alpha(z,y) &= v^\alpha \ell^{(n)}(v^\top y)\frac{\phi_+(z) - \phi_-(z)}{f(z \mid y)} \\ &= v^\alpha P_{n-1}(w(y))\big(\pi_+(z,y) - w(y)\big). \end{aligned} \tag{266}$$

The second line uses $(\phi_+ - \phi_-)/f = (\pi_+ - w)/(w(1-w))$. Since $|\pi_+ - w| \leq 1$, the bound is uniform in both z and y. In particular $S_{e_i} = v_i(\pi_+ - w)$. The preceding lemma verifies weighted-TV differentiability; fixed Gaussian components admit all polynomially weighted moments and all exponential moments of their Euclidean norm.

For completeness, if $Z \sim N(0, I_{d_x})$, then for every fixed $\lambda > 0$,

$$\sup_y \mathbb{E}[e^{\lambda\|X_0\|} \mid Y = y] \leq e^{\lambda\|a\|}\mathbb{E}e^{\lambda\sigma\|Z\|} < \infty.$$

The same argument gives an exponential moment for $\|Y\|$. Together with the relative-derivative calculation, these verify both primitive assumptions. The two component densities and their spatial gradients have bounded suprema, independent of y. Their convex combinations are therefore uniformly bounded and uniformly Lipschitz in z, giving the optional assumption with $\beta = 1$.

In one dimension at $y = 0$, $w = 1/2$ and f is even. Its derivatives at zero satisfy

$$f'(0 \mid 0) = 0, \qquad f''(0 \mid 0) = \left(\frac{a^2}{\sigma^4} - \frac{1}{\sigma^2}\right)\phi_\sigma(a) > 0.$$

Thus zero is a strict local minimum; continuity, symmetry, and Gaussian decay give a maximum on each side. There are exactly two modes: for $z > 0$, $f'(z) = 0$ is equivalent to $z = a\tanh(az/\sigma^2)$. The right side minus z is strictly concave on $(0,\infty)$, has positive derivative at zero, and tends to $-\infty$, hence has exactly one positive root. Symmetry gives the negative root. Moreover $(\log f)''(0 \mid 0) > 0$, contradicting log-concavity. Continuity of this strict inequality in y gives its persistence on a neighborhood of zero, which has positive Gaussian probability. □

Fixed-component mixtures with exponential tails. The fixed-component lemma also covers non-Gaussian components. In one dimension take $\nu_\pm = \mathrm{Laplace}(\pm a, b)$ with $b > 0$, and use the logistic weight from Proposition H.2. For $0 < \lambda < b^{-1}$,

$$\int e^{\lambda|z|}\nu_\pm(\,\mathrm{d}z) \leq \frac{e^{\lambda|a|}}{1-b\lambda} < \infty.$$

The weight calculation is unchanged, so $b_k = 0$ and $\chi_k = 4k$ at every finite order. Each Laplace density is bounded and globally Lipschitz; its convex mixtures satisfy Assumption 3.3 with $\beta = 1$. Alternatively fixed atomic components satisfy the first two assumptions but need not satisfy the optional density condition. Thus a density for $X_0 \mid Y = y$ is not required for the stopped-law theorem.

Conditional exponential families. Let $\mu_y(\,\mathrm{d}z) = \exp\{\eta(y)^\top T(z) - A(\eta(y))\}\nu(\,\mathrm{d}z)$. The first relative derivative is

$$S_{e_j}(z, y) = \partial_{y_j}\eta(y)^\top \big(T(z) - \mathbb{E}_{\mu_y} T(Z)\big). \tag{267}$$

A concrete sufficient class takes a probability measure ν with an exponential moment, bounded sufficient statistics T, and a globally bounded C^k parameter map η whose derivatives grow polynomially in y. The density relative to ν is then bounded above and below by constants, uniformly in y, so the exponential moment is uniform. Differentiation under the integral is valid on compact parameter sets. Repeated quotient and product rules give relative derivatives polynomial in the derivatives of η and in bounded centered sufficient statistics; hence the weighted-TV condition holds with $b_k = 0$. Families with unbounded sufficient statistics require their own uniform moment and parameter-domain checks.

A noisy conditioning channel with fixed compact data support. Suppose X_0 has a probability law ν supported in $\|z\| \le R_0$, h is bounded on that support, and $Y = h(X_0) + \xi$ for independent $\xi \sim N(0, \tau^2 I)$, $\tau > 0$. The Bayes version is defined for every y and is proportional to

$$\exp\Big(\tau^{-2} y^\top h(z) - \tfrac{1}{2}\tau^{-2}\,\|h(z)\|^2\Big)\,\nu(\,\mathrm{d}z).$$

It has the same fixed compact support for every y, so its exponential moment is at most $e^{\lambda R_0}$. Derivatives of the unnormalized likelihood in y insert products of the bounded coordinates of $h(z)/\tau^2$. Quotient induction gives bounded relative derivatives at every finite order, and differentiation in weighted TV is justified locally by the compact support. Thus $a_\alpha = b_\alpha = 0$ is possible. Finally Y has an exponential moment because $h(X_0)$ is bounded and ξ is Gaussian. An arbitrary noisy conditioning channel need not have a uniform conditional moment; the fixed-support condition here supplies it explicitly.

An excluded unbounded Gaussian location model. Let Y, Z be independent standard Gaussian scalars and $X_0 = Y + Z$. Although X_0 is Gaussian and the relative derivatives of $N(y, 1)$ have polynomial growth, for every $\lambda > 0$ Jensen's inequality gives

$$\mathbb{E}[e^{\lambda|X_0|} \mid Y = y] \ge e^{\lambda|\mathbb{E}[X_0|Y=y]|} = e^{\lambda|y|}.$$

The P_Y-essential supremum is already infinite. Hence marginal light tails alone do not imply the conditional tail assumption, even for a smooth Gaussian joint model. This exclusion reflects uniform conditional control, not a distinction between pointwise and essential bounds for full-support Y.

Moving conditional support. The relative identity can fail when a conditional-law derivative has a singular boundary or transport component not dominated by μ_y. A nonconstant point-mass family $\mu_y = \delta_{g(y)}$ generally fails even TV continuity, and so is excluded. Constant atomic laws and smooth mixtures on a fixed common support can satisfy the assumptions. A fixed common manifold carrying smooth positive conditional densities relative to one reference measure may likewise be covered; common dominating conditional geometry, rather than ambient Lebesgue density, is what the relative-law condition requires.

I Parameter and Computation Costs

For fixed depth $L \geq 3$, input dimension, and adjacent-width ratios, a dense scalar pyramid has

$$P = \sum_{h=1}^{H} d_h d_{h-1} + d_H = \Theta(m^2)$$

trainable parameters. There are d_x independent scalar networks, with no shared backbone. Centering additionally stores the frozen initial network and changes constants, while the forward/backward dense cost remains $O(d_x N m^2)$ per full-batch update.

Use the main-theorem exponents r, ℓ from (9), evaluated at effective order $\bar{k} = \min\{k, s\}$, and put $E = 13\ell + 3r$. Table 3 translates the sufficient statistical width into costs at fixed step size and fixed $t_0 > 0$.

Table 3: Sufficient statistical resource scales at fixed t_0. All entries suppress logarithms and fixed architecture constants. The parameter and training-cost entries retain the d_x independent output heads. These are upper bounds from the proof, not necessary costs or empirical predictions.

Resource	General $\bar{k} = \min\{k, s\}$	Saturation $\bar{k} = s$
Last hidden width m	$N^{13\ell+3r}$	$N^{16/(1+\gamma)}$
GD updates $R_\star$	$\eta^{-1} N^{\ell}$	$\eta^{-1} N^{1/(1+\gamma)}$
Total parameters	$d_x N^{2E}$	$d_x N^{32/(1+\gamma)}$
Dense training operations	$d_x \eta^{-1} N^{1+\ell+2E}$	$d_x \eta^{-1} N^{1+33/(1+\gamma)}$

Substituting the sufficient sample size $N = \widetilde{O}(\varepsilon^{-2/r})$ gives

$$m = \widetilde{O}\left(\varepsilon^{-(6+26s/\bar{k})}\right), \qquad \eta R_\star = \widetilde{O}\left(\varepsilon^{-2s/\bar{k}}\right).$$

The capacity exponent γ enters the sample-size bound but cancels from these sufficient width and effective-horizon exponents in accuracy. In particular, when $\bar{k} = s$, the width exponent is 32 and the effective-horizon exponent is 2. For original-law recovery, the shrinking t_0 requires the resource recalibration in Theorem B.9.

Storing and evaluating the fixed initialization functions for centering changes storage and inference costs by a constant factor. The inference cost per generated vector is $O(d_x M P)$. These conservative sufficient scales establish finite computation; they are not necessary resource requirements or claims of practical efficiency at the worst-case calibration.

J Optional Geometry and Curvature Certificates

This section collects the additional frozen-design separation, initialization-curvature, and persistence guarantees. They use the additional density and width conditions stated here and are not needed for the population-learning or sampling results in Theorem 1 and Corollary 1.

J.1 Optional Frozen-Design Geometry and KMM Initialization

This optional branch is separate from the population-risk and sampling theorems. Its extra density assumption is stated here and is not required by the main statistical argument.

Assumption J.1 (Optional local density of the condition)**.** *The condition Y has a Lebesgue density q_Y such that $\operatorname{ess\,sup}_{\|y\|_2 \leq R} q_Y(y) \leq M_q(R) < \infty$ for every finite R. No polynomial growth of $M_q(R)$ is assumed.*

Lemma J.2 (Localized small-ball inequality for the condition)**.** *Under Assumption J.1, for independent Y, Y' and $0 < r \leq 1$,*

$$\mathbb{P}(\|Y - Y'\|_2 \leq r, \|Y\|_2 \leq R, \|Y'\|_2 \leq R) \leq C_Y(R) r^{d_y}, \qquad C_Y(R) = v_{d_y} M_q(R), \tag{268}$$

where v_d is the volume of the unit ball in $\mathbb{R}^d$.

Proof. Integrate first over y' with both $\|y\|' \le R$ and $\|y - y'\| \le r$. Bounding $q_Y(y')$ by $M_q(R)$ gives $v_{d_y} M_q(R) r^{d_y}$. The remaining integral of $q_Y(y)$ is at most one. For an unconditional model ($d_y = 0$), take $C_Y(R) = 1$ and omit this assumption. □

For two independent indices drawn from the main training law, define

$$\omega_M(r) = \sum_{j,\ell=0}^{M-1} \pi_j \pi_\ell \mathbf{1}_{\{|t_{j+1} - t_{\ell+1}| \le r\}}, \qquad \pi_j = \frac{\Delta_j c(t_{j+1})}{\mathcal{W}}, \qquad b_M := \max_j \frac{\pi_j}{\Delta_j} = \frac{c(t_1)}{\mathcal{W}}. \tag{269}$$

Write $C_X(t_0) = v_{d_x}(4\pi\sigma_{t_0}^2)^{-d_x/2}$.

Lemma J.3 (Pairwise small-ball bound). *Let $U = (X_{t_{J+1}}, Y, t_{J+1})$ and let U' be an independent copy. Under Assumption J.1, for $0 < r \le 1$,*

$$\mathbb{P}(\|U - U'\|_2 \le r, \|Y\|_2 \le R, \|Y'\|_2 \le R) \le C_X(t_0) C_Y(R) r^D \omega_M(r). \tag{270}$$

Proof. Condition on both clean pairs and indices. Unless $|t_{J+1} - t_{J'+1}| \le r$, the event is impossible. Otherwise the spatial difference is Gaussian, with covariance $(\sigma^2_{t_{J+1}} + \sigma^2_{t_{J'+1}})I \succeq 2\sigma^2_{t_0} I$; its probability in any radius-r ball is at most $C_X(t_0) r^{d_x}$. The event also requires $\|Y - Y'\|_2 \le r$. Use Lemma J.2, the independence of the indices from the clean pairs, and average over the indices. □

Lemma J.4 (Collision mass for the score-weighted grid). *Let $\Delta_{\max} = \max_j \Delta_j \le \kappa_{\text{grid}}$. Then*

$$\omega_M(r) \le \min\{1, 2b_M(r + \Delta_{\max})\} \le \min\{1, 2b_M(r + \kappa_{\text{grid}})\}. \tag{271}$$

Proof. For fixed j, the indices $I_j = \{\ell : |t_{\ell+1} - t_{j+1}| \le r\}$ are consecutive. The corresponding disjoint intervals $(t_\ell, t_{\ell+1}]$ have union of length at most $2r + 2\Delta_{\max}$. Hence

$$\sum_{\ell \in I_j} \pi_\ell \le b_M \sum_{\ell \in I_j} \Delta_\ell \le 2b_M(r + \Delta_{\max}).$$

Average against π_j and also use $\omega_M(r) \le 1$. □

Proposition J.5 (A-priori minimum separation of the frozen design). *Let $\delta_U = \min_{i \ne j} \|U_i - U_j\|_2$. Then*

$$\mathbb{P}(\delta_U \le r) \le \frac{N(N-1)}{2} C_X(t_0) C_Y(R) r^D \omega_M(r) + \mathbb{P}(\max_{i \le N} \|Y_i\|_2 > R). \tag{272}$$

Set

$$\mathsf{A}_{\text{sep,N}} = N^2 C_X(t_0) C_Y(\rho_N) \tag{273}$$

and choose ρ_N so that the final probability is at most ζ_{loc}. With probability at least $1 - \zeta_{\text{sep}} - \zeta_{\text{loc}}$,

$$\delta_U \ge \underline{\delta}_U := c_D \left[\left(\frac{\zeta_{\text{sep}}}{\mathsf{A}_{\text{sep,N}} b_M \kappa_{\text{grid}}} \right)^{1/D} \wedge \left(\frac{\zeta_{\text{sep}}}{\mathsf{A}_{\text{sep,N}} b_M} \right)^{1/(D+1)} \wedge 1 \right]. \tag{274}$$

Every quantity on the right is deterministic before the sample is drawn. One may replace κ_{grid} by $\Delta_{\max}$ for a sharper floor.

Proof. Apply Lemma J.3 to every unordered pair. On the localization event set $R = \rho_N$. By Lemma J.4, the resulting bound is at most $\mathsf{A}_{\text{sep,N}} b_M (r^{D+1} + \kappa_{\text{grid}} r^D) + \zeta_{\text{loc}}$. A sufficiently small c_D in (274) makes each monomial at most $\zeta_{\text{sep}}/2$. □

The interval-length variant has $b_M = 1/(T - t_0)$; this value cannot be used for the score-weighted main algorithm. Likewise, $\kappa_{\text{grid}} = \widetilde{O}(M^{-1})$ cannot be replaced by M^{-1} inside an exact deterministic floor without retaining its logarithmic factors.

Lemma J.6 (Upper-hemisphere separation). *If $\|u\|_2, \|v\|_2 \le \rho$, then*

$$\|\psi_\rho(u) - \psi_\rho(v)\|_2 \ge \frac{\|u - v\|_2}{2\sqrt{2}\rho}, \qquad \|\psi_\rho(u) + \psi_\rho(v)\|_2 \ge \sqrt{2}. \tag{275}$$

Hence, on $\max_i \|U_i\|_2 \le \rho_N$,

$$\delta_{\text{sph}} := \min_{i \ne j} \min\{\|X_i - X_j\|_2, \|X_i + X_j\|_2\} \ge \min\left\{\sqrt{2}, \frac{\delta_U}{2\sqrt{2}\rho_N}\right\}. \tag{276}$$

Proof. The last coordinate of every point in the localized cap is at least $1/\sqrt{2}$, and $\psi_\rho^{-1}(x) = \rho x_{1:d_0-1}/x_{d_0}$ is $2\sqrt{2}\rho$-Lipschitz there. This proves the first inequality. The last coordinates of $\psi_\rho(u)$ and $\psi_\rho(v)$ sum to at least $\sqrt{2}$, proving the second. □

J.1.1 From a certified tangent block to the full Gram matrix

Lemma J.7 (Tangent-block domination)**.** *Partition the initialization Jacobian by parameter blocks,*

$$J_X(\theta_0) = \begin{bmatrix} J^{(1)} & \cdots & J^{(L)} \end{bmatrix}. \tag{277}$$

Then

$$K_X(\theta_0) = \sum_{\ell=1}^{L} J^{(\ell)} J^{(\ell)\top} \succeq J^{(\ell_0)} J^{(\ell_0)\top} \qquad \text{for every } \ell_0 \in [L]. \tag{278}$$

Proof. Every summand is positive semidefinite, so deleting all but one can only decrease the matrix in Loewner order. □

Proposition J.8 (Optional KMM initialization certificate: exact import)**.** *Suppose $d_0 \geq 3$, $L \geq 3$, $d_L = 1$, and $d_\ell \geq d_{\ell+1}$ for $1 \leq \ell < L$. Allocate $\zeta_{\text{init}} = c\delta < 1/3$ with a fixed sufficiently small $c > 0$. Let $\delta_\star \in (0, 1/2]$ be the deterministic spherical separation floor and let Λ_L, λ_0 be as in (281). In addition to the statistical width, impose the separate source-theorem widths (283). Then, with probability at least $1 - \zeta_{\text{sep}} - \zeta_{\text{loc}} - \zeta_{\text{init}}$,*

$$K_X^{(q)}(\theta_{q,0}) \succeq \lambda_0 I_N \quad \text{simultaneously for } q \in [d_x]. \tag{279}$$

Proof. We apply Karhadkar et al. (2024, Theorem 8), in the final NeurIPS 2024 paper, Section 4, to each fixed realized design on the separation and localization event. The source uses $n = N$, d_0 as here, $\epsilon = \zeta_{\text{init}}/d_x < 1/3$, and separation parameter $\delta = \delta_\star$. Its quantity called λ is

$$\left(1 + \frac{d_0 \log(1/\delta_\star)}{\log d_0}\right)^{-3} \delta_\star^4 = \Lambda_L,$$

up to any explicitly fixed constant absorbed in λ_0. Its hypotheses are $d_H \gtrsim 2^L \log(NL/\epsilon)$ and $d_1 \gtrsim (N/\Lambda_L)\log(N/\Lambda_L)\log(N/\epsilon)$. The fixed failure-allocation constants are absorbed by the constants in the separately imposed endpoint conditions.

For the normalization map, its Eq. (7) defines the network by multiplying the unnormalized depth-L ReLU output by $\prod_{\ell=1}^{L-1}\sqrt{2/d_\ell}$. ReLU's positive homogeneity makes this the same function of the same Gaussian matrices as (5), as shown in (191). The source uses the product-Frobenius metric and includes all matrices, including the readout, in its parameter space. Thus both Jacobians and full Gram matrices coincide; centering subtracts a fixed function and changes neither. Theorem 8 therefore yields the full-matrix lower bound with $\lambda_0 = c_{\text{KMM}}\Lambda_L$. No assertion about an unspecified individual tangent block is needed. Condition on the design, apply the theorem to each independent initialization, union-bound over the heads, and add the separation/localization failure probabilities. □

The certificate is an optional external-theorem consequence with its own endpoint widths and extra density assumption. The statistical and sampling results do not invoke it. The elementary block-domination lemma above remains valid, but is not used to replace any source-theorem hypothesis.

J.2 Separate optional geometry width

The following conditions supplement the statistical width calibration in Section B. For this optional branch, assume the locally bounded density of Y and put $\zeta_{\text{sep}} = c\delta/(Ld_x)$. With the actual weighted time law, define

$$\begin{aligned} \mathsf{A}_{\text{sep,N}} &= N^2 C_X(t_0) C_Y(\rho_N), \qquad b_M = \max_j \frac{\mathbb{P}(J=j)}{\Delta_j} = \frac{c(t_1)}{\mathcal{W}}, \\ \underline{\delta}_U &= c\left[\left(\frac{\zeta_{\text{sep}}}{\mathsf{A}_{\text{sep,N}} b_M \kappa_{\text{grid}}}\right)^{1/D} \wedge \left(\frac{\zeta_{\text{sep}}}{\mathsf{A}_{\text{sep,N}} b_M}\right)^{1/(D+1)} \wedge 1\right]. \end{aligned} \tag{280}$$

The exact grid factor is retained: replacing it by M^{-1} would discard its logarithmic dependence. Sphericalization gives

$$\begin{aligned} \underline{\delta}_{\mathrm{sph}} &= \min\left\{\sqrt{2}, \frac{\delta_U}{2\sqrt{2}\rho_N}\right\}, \qquad \delta_\star = \min\{1/2, \underline{\delta}_{\mathrm{sph}}\}, \\ \Lambda_L &= \left(1 + \frac{d_0 \log(1/\delta_\star)}{\log d_0}\right)^{-3} \delta_\star^4, \quad \lambda_0 = c_{\mathrm{KMM}} \Lambda_L. \end{aligned} \tag{281}$$

These deterministic quantities precede the sample. Set

$$W_{\mathrm{curv}} = (1 + \eta R_\star) N^6 \lambda_0^{-6}, \qquad W_{\mathrm{geom}} = W_{\mathrm{stat}} \vee W_{\mathrm{curv}}. \tag{282}$$

The optional certificate requires $\mathsf{Width}(W_{\mathrm{geom}})$ and the source's endpoint-width conditions under the normalization and hypothesis mapping in Proposition J.8:

$$\begin{aligned} d_H &\geq C 2^L \log\left(\frac{NLd_x}{\delta}\right), \\ d_1 &\geq C \frac{N}{\Lambda_L} \log\left(\frac{N}{\Lambda_L}\right) \log\left(\frac{Nd_x}{\delta}\right). \end{aligned} \tag{283}$$

Here $d_0 \geq 3$, $L \geq 3$, and the widths are pyramidal, as required by the identified import. These endpoint conditions are separate from the statistical width. Under all these conditions, Propositions J.8 and J.10 give $K_X(\theta_r) \succeq (\lambda_0/2) I_N$ through $R_\star$ because $\|K_X(\theta_r) - K_X(\theta_0)\|_{\mathrm{op}} \leq CNq_\star$. No such event is needed in the main theorem.

At fixed t_0, on grids with $M = \widetilde{\Theta}(N^r)$, the weighted normalizer tends to the positive integral of c as the mesh vanishes; thus b_M remains bounded. If in addition $M_q(\rho_N) = \mathrm{polylog}(N)$, then

$$\delta_\star \geq \widetilde{\Omega}(N^{-e_{\mathrm{sep}}}), \qquad e_{\mathrm{sep}} = \max\left\{\frac{2-r}{D}, \frac{2}{D+1}\right\}, \tag{284}$$

so a sufficient curvature-persistence width has exponent

$$E_{\mathrm{curv}} = 6 + \ell + 24 e_{\mathrm{sep}}, \qquad E_{\mathrm{geom}} = \max\{13\ell + 3r, E_{\mathrm{curv}}\}. \tag{285}$$

The endpoint conditions (283) remain explicit. Under mere local density boundedness, no polynomial envelope for $M_q(\rho_N)$ follows; the deterministic conditions above are then the statement of record.

J.3 Deterministic curvature persistence

Fix a horizon R and assume that (202) holds uniformly for $r \leq R$. Use the Jacobian and kernel definitions of Section F.2, with J_0, ε_J, and $\delta_K = 2J_0\varepsilon_J + \varepsilon_J^2$ as in Proposition F.3.

On the deterministic-separation event, Proposition J.8 gives

$$K_X(\theta_0) \succeq \lambda_0 I_N, \qquad \lambda_0 = c_L \Lambda_L. \tag{286}$$

This is an initialization statement only; the following perturbation argument proves persistence. Population comparisons are established separately in Section F.

Proposition J.9 (Curvature persistence is a separate perturbation consequence)**.** *On the initialization event* (286)*, if*

$$\delta_K \leq \lambda_0/2, \tag{287}$$

then

$$K_X(\theta_r) \succeq \frac{\lambda_0}{2} I_N, \qquad r \leq R. \tag{288}$$

Proof. The Jacobian perturbation gives $\|K_X(\theta_r) - K_X(\theta_0)\|_{\mathrm{op}} \leq 2J_0\varepsilon_J + \varepsilon_J^2 = \delta_K$. Apply Weyl's inequality to (286). □

J.4 Optional Curvature Persistence

The following deterministic implication is independent of the source of an initialization eigenvalue certificate. It is separate from the statistical population-risk analysis.

Proposition J.10 (Persistence conditional on initialization curvature)**.** *Suppose the movement and local-perturbation hypotheses in Propositions F.13 and F.14 hold. On any event where $K_X(\theta_0) \succeq \lambda_0 I_N$ for a fixed $\lambda_0 > 0$, if*

$$Nq_\star \le c_{L,\kappa_{\mathrm{pyr,N}}}\lambda_0, \tag{289}$$

then

$$K_X(\theta_r) \succeq \frac{\lambda_0}{2} I_N, \qquad 0 \le r \le R_\star. \tag{290}$$

In addition to the stated movement and initialization hypotheses, a sufficient width is

$$m \ge C_{L,\kappa_{\mathrm{pyr,N}}}\nu_m^3(1+\eta R_\star)N^6\lambda_0^{-6}. \tag{291}$$

Proof. The Jacobian bounds (257) and (231) give $\|K_X(\theta_r) - K_X(\theta_0)\|_{\mathrm{op}} \le CNq_\star$. Choose the constant in (289) so that this is at most $\lambda_0/2$. Weyl's inequality then proves (290). Substitute $q_\star = C\sqrt{\nu_m}(1+\eta R_\star)^{1/6}m^{-1/6}$ and take sixth powers to obtain (291). This step propagates initialization curvature; it does not create it. □

Corollary J.11 (Optional finite-horizon geometry certificate)**.** *Impose Assumption J.1, the deterministic separation calibration, the exact KMM endpoint widths (283), and the additional persistence width (291), as well as the main movement/perturbation conditions. Allocate total failure probability at most δ across these events and output heads. Then, with probability at least $1-\delta$,*

$$K_X^{(q)}(\theta_{q,r}) \succeq \frac{\lambda_0}{2} I_N, \qquad q \in [d_x], \quad 0 \le r \le R_\star. \tag{292}$$

For each scalar empirical loss $\widehat{\mathcal{L}}_q = (2N)^{-1}\|e_q\|_2^2$, this implies the pathwise inequality $\left\|\nabla\widehat{\mathcal{L}}_q\right\|_2^2 \ge (\lambda_0/N)\widehat{\mathcal{L}}_q$ wherever the stated Jacobian is used. This optional conclusion is not invoked in the statistical branch.

Proof. The exact normalization map in Proposition J.8 supplies the initialization event under the separately stated source-theorem widths. Apply Proposition J.10 on its intersection with the movement/perturbation event and union-bound the allocated failures. Finally, $\left\|\nabla\widehat{\mathcal{L}}_q\right\|_2^2 = N^{-2}e_q^\top K_X^{(q)}e_q \ge \lambda_0\|e_q\|_2^2/(2N^2) = (\lambda_0/N)\widehat{\mathcal{L}}_q$. □

K NOTATION

Symbol	Meaning
$X_0, Y;\ d_x, d_y$	Clean observation and condition; their dimensions.
$D = d_x + d_y$	Joint data–condition dimension.
$d_0 = D + 2$	Input dimension after time augmentation and sphericalization.
$L,\ H = L - 1$	Network depth including readout; number of hidden layers.
$m = d_H,\ d_1, \dots, d_H$	Last hidden width and hidden layer widths.
$\kappa_{\mathrm{pyr,N}}$	Adjacent hidden-width ratio envelope.
k	Primitive integer derivative order in the conditional-law regularity assumption.
$s = (D+2)/2$	Full-NTK Sobolev order.
$\bar{k} = \min\{k, s\}$	Effective regularity used in the main results.
$\gamma = (D+1)/(D+2)$	Distribution-free effective-dimension exponent.
u	Auxiliary real regularity order in the appendix. Statistical calibrations use $1 \le u \le \bar{k}$; the main results take $u = \bar{k}$.
$j = \lceil u \rceil,\ j_s = \lceil s \rceil$	Integer derivative orders used to construct the order-u and saturated target extensions.
$b = u/s$	Bias exponent; in the main results, $b = \bar{k}/s$.
r, ℓ	Main-result exponents $\bar{k}/(\bar{k}+\gamma s)$ and $s/(\bar{k}+\gamma s)$; $r + \gamma\ell = 1$.
r_u, ℓ_u	Auxiliary exponents $u/(u+\gamma s)$ and $s/(u+\gamma s)$; $r_u + \gamma\ell_u = 1$. They equal r, ℓ at $u = \bar{k}$.
$a_u = u/(s-u)$	Nonsaturated comparator approximation exponent, defined for $0 < u < s$.
$A_{u,N}(t_0)$	Norm envelope for the localized normalized target extension constructed at derivative order $j = \lceil u \rceil$.
χ_j	Small-time target-envelope power at integer derivative order j; the order-u calibration uses $\chi = \chi_{\lceil u \rceil}$.
θ_u	Balanced nonsaturated risk power $2\gamma s\chi_{\lceil u \rceil}/(u+\gamma s)$.
θ_{sat}	Saturated risk power $2\gamma\chi_{\lceil s \rceil}/(1+\gamma)$.
β	Density smoothness used for the original-law comparison.

Symbol	Meaning (continued)
N, δ	Number of iid clean pairs; local failure-probability parameter.
N_{cal}	Deterministic sufficient clean-sample threshold.
t_0, T, M	Diffusion stopping time, terminal noising time, and reverse-step count.
$\Delta_j, \kappa_{\mathrm{grid}}$	Reverse-grid step lengths and admissibility factor.
$c(t), \mathcal{W}$	$e^{-2t}/(1-e^{-2t})^2$ and $\sum_j \Delta_j c(t_{j+1})$; time weights and score conversion.
B_N, ρ_N	Label clipping threshold and input localization/lift scale.
$\mathfrak{r}_N$	Calibrated squared population-risk scale.
$\eta, \lambda_N, R_\star$	GD step size, constrained effective regularization, and training updates.
$\lambda_{\mathrm{bal}}, \mathcal{L}_N$	Balanced scale and confidence logarithm in (44).
$h = \eta R_\star, \Xi_N$	Effective training horizon and $h + h^2$ accumulation factor.
$\varepsilon_{\mathrm{coup},N}$	Normalized scalar-head coupling tolerance; distinct from TV accuracy ε.
$\nu_m, q_\star$	Logarithmic entropy factor and finite-width perturbation scale.
P_U, P_X	Fresh augmented design law and its sphericalized pushforward.
g^K, g^∞	Zero-start GD predictors for the finite and limiting full NTKs.
$p_{\mathrm{stat}}, p_{\mathrm{gate}}$	Original-law statistical and population-comparison sample exponents.
$W_{\mathrm{stat}}, W_{\mathrm{geom}}$	Statistical and separate optional geometry width scales.

The auxiliary order u is distinct from the random augmented input U. In regularity statements derived from the conditional-law assumption, $0 < u \le k$; abstract approximation lemmas state their own admissible ranges. The subscript r on a training iterate is an integer update index; the standalone exponent r in (9) is a scalar. Training updates $R_\star$ and sampling steps M count different operations.

L ASSUMPTION, ARCHITECTURE, AND PROOF-DEPENDENCY LEDGER

Constant dependence. Architecture constants denoted $C_{L,\kappa_{\mathrm{pyr}}}$ depend only on fixed depth and the displayed adjacent-width envelope. Separate statistical and sampling constants retain their stated diffusion, regularity, and tail dependence. No polynomial-in-depth bound is asserted, and no constant silently depends on m, $R_\star$, or the realized separation. If the width-ratio envelope varies with N, that dependence is retained explicitly. The statistical coupling width is independent of the KMM curvature parameter; the additional geometry width contains the explicit factor λ_0^{-6}. The sole headline result is Theorem 1, for the all-trained controlled pyramid.

Table 4: Status and role of assumptions and design choices.

Condition or component	Type	Role in the theorem
Uniform conditional exponential tails on the chosen law version	Data distribution	Supplies posterior bounds throughout the regularity domain, as well as conditional moments, localization, clipping bias, and terminal-law control.
Locally bounded density of Y	Data distribution	Supplies frozen-design small-ball control for the optional curvature certificate. Local boundedness alone does not give a polynomial envelope for $M_q(\rho_N)$.
Order-k relative conditional-law regularity	Data distribution	Derives a localized joint C^k envelope for the clipped posterior mean. The main theorem uses $\bar{k} = \min\{k, s\}$; the appendix also allows statistical calibrations with $1 \le u \le \bar{k}$.
Controlled adjacent width ratios $\kappa_{\mathrm{pyr,N}}$	Architecture	Controls rectangular operator bounds. The finite constant envelope $\Gamma_{L,\kappa}$ in (52) retains fixed-depth and width-ratio dependence.
Gaussian initialization and centering	Algorithm	Independent Gaussian hidden weights and readout admit the local proofs; centering gives zero initial prediction while all parameter blocks move.
Finite stopping rule and step size	Algorithm	Supplied class bounds (Section G.1) determine the calibration. The deterministic maximum of the balanced scale and comparison floor controls population learning; the width calibration keeps the trajectory in the proved local region.
Optional density Hölder regularity	Data distribution	Needed only to compare the learned stopped law with p_0.
Actual-design capacity exponent	Derived	The harmonic addition theorem gives $\mathsf{df}_P(\lambda) \lesssim \lambda^{-\gamma}$ for every law, including each empirical measure; no design density is required.
ERM access, source membership, favorable Gram matrix	Not assumed	Population learning uses uniform comparators and early stopping; nonlinear coupling controls the finite iterate directly. Additional curvature is proved only under W_{geom}.

Attribution and depth scope. Section A positions the comparison strategy and harmonic-tail argument; exact source correspondences accompany the local proofs. The global full-sphere Sobolev identification uses $L \ge 3$. Lemma E.16 shows that the shallow parity gap alone does not obstruct localized comparators. The perturbation and movement arguments are formulated for $H \ge 1$; the optional KMM curvature certificate uses its separately stated deep-network hypotheses. No shallow representation or trajectory impossibility is asserted.

Table 5: Proof-dependency map and local/external split.

Module	Status	Exact consequence used
Higher-order target regularity	Proved in Section D.2	Polynomial localized C^k envelope with explicit small-t_0 power, obtained from Gaussian derivatives, posterior moments, and quotient induction.
Frozen-design geometry	Proved in Section J.1	Correct admissible-grid collision bound, a-priori Euclidean separation floor, and spherical directional separation.
Initialization curvature	Architecture-matched full-NTK theorem imported	The specified KMM theorem certifies the full initialization Gram matrix under its normalization, directional separation, and endpoint widths; this is optional.
Full-NTK recursion and spectrum	Proved in Sections E.1 and E.2	Exact all-trained recursion and $\mathcal{H}_L^{\mathrm{full}} \simeq H^{(D+2)/2}$ with no missing harmonic parity for $L \geq 3$.
Source-free comparator	Proved in Section E.3	A bounded linear positive-time extension followed by qualified heat approximation works under the actual design; its localization remainder persists at saturation.
Kernel early stopping	Proved in Section E.5	General-smoothness population risk without an inverse empirical-Gram penalty; RKHS saturation handled separately.
Gaussian initialization and sparse rows	Proved in Section F.3	Uniform margins, gate-pattern entropy, and all-support restrictions of the independent Gaussian readout; no unspecified random-product input.
Rectangular and moving-readout perturbation	Proved in Lemma F.5 and Theorem F.8	Direct Jacobian and Taylor bounds $C\sqrt{\nu_m}\mathcal{R}^{1/3}m^{-1/6}$ for the actual all-trained architecture.
Movement and nonlinear/kernel coupling	Proved in Section F.5 and Proposition C.10	Explicit centered-GD movement bootstrap and population coupling with $\Xi_N = h + h^2$, using averaged cross-kernel operators and no inverse curvature.
Curvature persistence	Proved in Proposition J.10 after imported KMM initialization	Bare perturbation condition $Nq_\star \lesssim \lambda_0$, yielding the separate λ_0^{-6} geometry width.
Controlled-pyramid full-NTK concentration	Proved in Section F.4	Conditional covariance and gated-quadratic trace recursions, Gaussian readout concentration, and net extension give $C\sqrt{\nu_m}m^{-1/6}$; direct two-kernel comparison completes transfer.
Capacity calculation	Proved for every design in Section E.4	Uniform harmonic feature-tail bounds and a projection trace split yield $\gamma = (D+1)/(D+2)$, including for atomic time and empirical laws.
Score and sampler conversion	Proved/invoked in Section C	Right-endpoint sampling matches the explicit integrator; score-weighted times give conversion factor $\mathcal{W} \leq 1/(4t_0)$. Original-law calibration retains the comparison floor.