CONTINUUM-EQUIVALENT NEURAL OPERATORS CAN LEARN DIFFERENT INTERACTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural operators should recover the same physical interaction as input resolution increases, yet vanishing weight decay can erase interactions that remain fully expressible in the continuum. Guarantees of continuum expressivity and consistent representation cannot predict this selection: a fixed interaction may require diverging coefficients, and standard parameter penalties charge those coefficients rather than the finite interaction they encode. This mismatch makes architecture and regularization decisive for whether learned continuum models preserve the underlying physics. We establish a realization-cost principle that characterizes consistent recovery through target approximations with vanishing penalized parameter cost. Counterterm-coupled regularization implements this principle in finite Wick coordinates, preserving the intended interaction objective and restoring consistency. We demonstrate the mechanism using cubic responses of two-dimensional Gaussian free fields with a fixed continuum target and smoothing observation, where continuum-equivalent architectures erase or recover the same target under identical vanishing weight decay. For standard ReLU channel networks, the principle yields the sharp threshold $\lambda_\varepsilon C_\varepsilon^{3/L} \to 0$, linking penalty strength λ_ε, field variance C_ε at input cutoff ε, and depth L. Exact population analysis, matched architectural controls, and independently verified polynomial-network training demonstrate the predicted separation and its correction at moderate resolutions.

1 INTRODUCTION

A physical operator should not change when its input field is represented at a finer resolution. This requirement motivates function-space models and cross-resolution analysis in operator learning (Kovachki et al., 2023; Gao et al., 2025). For cubic responses of singular stochastic fields, refinement can make the coefficients representing a fixed interaction diverge. Wick-based neural operators subtract lower-degree counterterms when representing stochastic partial differential equations (Shi et al., 2026). We study these responses through a fixed smoothing observation and ask whether regularization penalizes the finite interaction or the cancellation needed to represent it.

The continuum learning limit is determined by realization cost (the parameter penalty needed to approximate the physical interaction), not by continuum expressivity alone. We characterize consistency through vanishing penalized recovery cost, derive contraction and null-extension laws, and prove the sharp $C_\varepsilon^{3/L}$ cost law for standard ReLU channel networks. Here ε is the input cutoff, C_ε its field variance, and L the network depth. Prediction-equivalent shallow and deep architectures therefore erase or recover the same cubic under the same vanishing penalty.

The comparison uses a fixed realizable continuum target, population risk, and global minimizers. The architectures are prediction-class equivalent but cost-inequivalent: they have identical attainable predictor limits and different costs of realizing them. Figure 1 links this cost difference to the learned limits; Section 6 gives the exact result.

Coordinate dependence of regularization is established context. Martini et al. (2019) show how penalizing lower-degree coefficients makes polynomial regression depend on the coding of its predictors. Lassas & Siltanen (2004) demonstrate that refinement can destroy an intended edge-preserving Bayesian inductive bias, while Lassas et al. (2009) construct discretization-invariant priors by preserving information from a continuum model. In neural networks, Parhi & Nowak (2026),

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.

revisiting Parhi & Nowak (2022), identify the function-space regularization represented by specific weight-decay objectives.

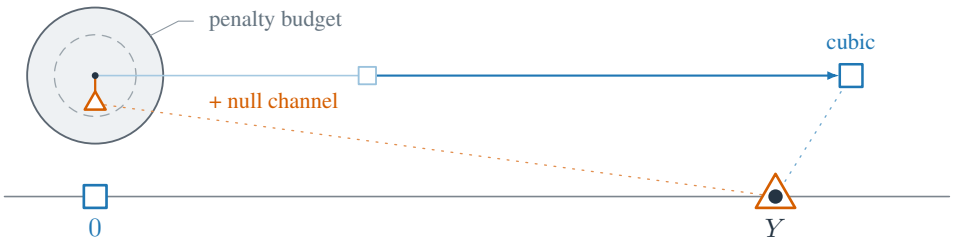


Figure 1: **Same target, different learned interactions.** Upper markers are representations of the same target Y; dotted guides indicate their common continuum limit. The arrow denotes refinement. Under $\lambda_\varepsilon = C_\varepsilon^{-1}$, the penalty-budget radius grows as $\sqrt{C_\varepsilon}$, while the cubic representation recedes as C_ε and the null-augmented witness stays bounded. Lower markers show the learned limits: 0 for the cubic and Y for the augmented model. Geometry is schematic; the lower line fixes the renormalized linear coefficient to zero.

The fixed-architecture law follows from one coupling: Wick renormalization subtracts a linear counterterm from the raw cubic. Squared-weight decay charges this cancellation at a scale proportional to penalty strength times squared field variance. This charge vanishes, remains finite, or forces a zero Laplacian in the channel variables. In one channel, that last constraint removes every nonzero cubic interaction.

Counterterm-coupled regularization penalizes the finite linear interaction and retains the hidden-weight cost, adding no learned parameters. It preserves the intended regularized objective even at positive strength. Faster decay and an unpenalized counterterm channel also give asymptotic consistency, but express different regularization choices.

For variance C, the Wick polynomials are $x^{3}_{:C} = x^3 - 3Cx$ and $x^{5}_{:C} = x^5 - 10Cx^3 + 15C^2x$. They give the architectural identity

$$-x^3 + \frac{x^5}{5C} = x^{3}_{:C} + \frac{1}{5C}x^{5}_{:C}. \tag{C}$$

The added degree-five Wick component (fifth Gaussian chaos) vanishes under refinement, making the extension continuum-null; its contractions provide a bounded-weight representation of the unit cubic. For quintic coefficient c, the bound $|C_\varepsilon c| \le 1$ and schedule $\lambda_\varepsilon = C_\varepsilon^{-1}$ give the low-cost interval $[-5, 5]$ (Proposition 6.2).

Cubic responses are the local interactions in Φ^4_2-type stochastic quantization (Da Prato & Debussche, 2003). The main theorems fix channel dimension d and a width $m \ge \binom{d+2}{3}$, with every neuron parameter trainable. Three claims are distinct: objective invariance preserves a specified regularized learning problem under a coordinate change; continuum consistency recovers its target as the cutoff and penalty vanish; finite-data observable accuracy is measured for a trained predictor. Sections 3–5 develop the selection law and its correction. Section 6 establishes the architectural separation, then Section 6.5 gives the general realization-cost and ReLU-depth laws, proved in Appendix D. Appendices E–F supply the numerical evidence.

2 Related Work

Representation-cost and compositional-space frameworks identify the function-level preferences of neural penalties (Parhi & Nowak, 2022; 2026; Ongie & Parhi, 2026). Continuum-invariant priors preserve an intended bias under refinement (Lassas & Siltanen, 2004; Lassas et al., 2009), while cross-resolution operator analysis addresses representation mismatch (Gao et al., 2025). DLR-Net and Wick–Hermite operator models encode singular-field structure (Gong et al., 2023; Shi et al., 2026). We characterize the learning limit when physical realization costs diverge as weight decay vanishes. Appendix G gives the detailed comparison.

3 A PHYSICAL LEARNING PROBLEM WITH A FIXED CONTINUUM TARGET

The rough input, its cutoff representation, and the finite interaction coefficients play different roles. Fixed output smoothing makes prediction error a Hilbert-space squared loss while retaining the singularity inside the nonlinear computation.

3.1 FIELDS, CUTOFFS, AND OBSERVATIONS

Let $\mathbb{T}^2 = (\mathbb{R}/2\pi\mathbb{Z})^2$, equipped with normalized Lebesgue measure μ, and let Δ_x be its spatial Laplacian. On a common probability space Ω, the input $X = (X_1, \dots, X_d)$ consists of independent centered real Gaussian free fields with covariance $(1 - \Delta_x)^{-1}$. Let $\mathbb{E}$ denote expectation and δ_{ij} the indicator of $i = j$. For smooth real test functions f, g,

$$\mathbb{E}[X_i(f)X_j(g)] = \delta_{ij}\langle (1 - \Delta_x)^{-1} f, g\rangle_{L^2(\mu)}. \tag{1}$$

The positive mass removes the zero-mode singularity. Each channel is a random distribution rather than a pointwise-defined function.

For $0 < \varepsilon \le 1$, use the heat regularization

$$X_\varepsilon = e^{\varepsilon^2\Delta_x/2}X, \qquad C_\varepsilon = \mathbb{E}[X_{\varepsilon,1}(x)^2], \qquad t_\varepsilon = 3C_\varepsilon. \tag{2}$$

Stationarity makes C_ε independent of x. It diverges logarithmically: $C_\varepsilon \asymp \log(1/\varepsilon)$ as $\varepsilon \downarrow 0$. We observe outputs through the fixed operator

$$S = e^{\Delta_x}, \qquad \mathsf{H} = L^2(\mathbb{T}^2, \mu). \tag{3}$$

The heat time in S is fixed; only the input cutoff changes. All operations and the H-norm are evaluated in the continuum. Thus ε is a physical regularization scale, not a numerical quadrature parameter. Appendix A.1–A.2 constructs the fields and proves the needed moment bounds directly.

3.2 TRAINABLE CUBIC INTERACTIONS AND THEIR COUNTERTERMS

Consider a one-hidden-layer cubic operator with a linear skip channel,

$$N_{\varepsilon,\theta,b} = S\left[\sum_{j=1}^{m} a_j(v_j^\top X_\varepsilon)^3 + b^\top X_\varepsilon\right], \qquad \theta = ((a_j, v_j))_{j=1}^m \in (\mathbb{R}\times\mathbb{R}^d)^m, \quad b \in \mathbb{R}^d. \tag{4}$$

Every a_j, v_j, and b is trainable. Vector norms are Euclidean; $\|\cdot\|_{\mathrm{op}}$ denotes operator norm. The tensor below records the represented cubic; its neuron factorization determines the cost of realizing it. Let $\mathcal{A} = \mathrm{Sym}^3(\mathbb{R}^d)$ be the symmetric third-order tensors, with the Frobenius inner product summed over ordered index triples. With $(v^{\otimes 3})_{ijk} = v_i v_j v_k$, set

$$A_\theta = \sum_{j=1}^{m} a_j v_j^{\otimes 3}, \qquad p_A(z) = \sum_{i,j,k=1}^{d} A_{ijk} z_i z_j z_k, \qquad D = \dim\mathcal{A} = \binom{d+2}{3}. \tag{5}$$

We fix $m \ge D$, a sufficient width to represent every cubic tensor because rank-one symmetric cubes span $\mathcal{A}$.

The contraction map $L : \mathcal{A} \to \mathbb{R}^d$ records the lower-order term created by Wick renormalization:

$$(LA)_i = \sum_{j=1}^{d} A_{ijj}, \qquad LA_\theta = \sum_{j=1}^{m} a_j\|v_j\|^2 v_j. \tag{6}$$

Define the Wick cubic at cutoff ε by

$$:\!p_A(X_\varepsilon)\!: = \sum_{i,j,k} A_{ijk}\big[X_{\varepsilon,i}X_{\varepsilon,j}X_{\varepsilon,k} - C_\varepsilon(\delta_{ij}X_{\varepsilon,k} + \delta_{ik}X_{\varepsilon,j} + \delta_{jk}X_{\varepsilon,i})\big]. \tag{7}$$

The contraction identity is exact:

$$p_A(X_\varepsilon) = :\!p_A(X_\varepsilon)\!: + t_\varepsilon(LA)^\top X_\varepsilon. \tag{8}$$

Consequently, the finite coefficient of the linear interaction is

$$\beta = b + t_\varepsilon LA_\theta, \qquad N_{\varepsilon,\theta,b} = F_\varepsilon(A_\theta, \beta), \qquad F_\varepsilon(A, \beta) = S[:\!p_A(X_\varepsilon)\!: + \beta^\top X_\varepsilon]. \tag{9}$$

We call (A, β) the *renormalized coefficients*; b is the *bare* linear coefficient. A fixed interaction (A, β) is represented by $b_\varepsilon = \beta - t_\varepsilon LA$. Its divergence is therefore a representation cost, not an increase in the interaction being learned.

3.3 AN IDENTIFIABLE LIMITING REGRESSION PROBLEM

The covariance calculation establishes the continuum target, identifies its coefficients, and supplies the curvature for the selection theorem.

Proposition 3.1 (Continuum features and risk)**.** *There is a linear map $F : \mathcal{A} \times \mathbb{R}^d \to L^2(\Omega; \mathsf{H})$ such that $F_\varepsilon \to F$ in operator norm. There are constants $\alpha, \gamma > 0$, independent of A, β, for which*

$$\mathbb{E}\langle F(A,\beta), F(B,\eta)\rangle_{\mathsf{H}} = \alpha\langle A, B\rangle_{\mathrm{F}} + \gamma\beta^\top\eta. \tag{10}$$

Fix any $A_* \in \mathcal{A}$ and $\beta_* \in \mathbb{R}^d$, and define the same target at every cutoff by

$$Y = F(A_*, \beta_*), \qquad \mathcal{R}_\varepsilon(A,\beta) = \frac{1}{2}\mathbb{E}\|F_\varepsilon(A,\beta) - Y\|_{\mathsf{H}}^2. \tag{11}$$

Then $\mathcal{R}_\varepsilon$ converges locally uniformly to

$$\mathcal{R}(A,\beta) = \frac{\alpha}{2}\|A - A_*\|_{\mathrm{F}}^2 + \frac{\gamma}{2}\|\beta - \beta_*\|^2. \tag{12}$$

Moreover, for all sufficiently small ε, there are constants $c_0 > 0$ and $K_* < \infty$ such that $\mathcal{R}_\varepsilon(A,\beta) \geq c_0(\|A\|_{\mathrm{F}}^2 + \|\beta\|^2) - K_*$.

The cubic and linear features are orthogonal third and first Gaussian chaos components. Spatial correlations determine α, γ; channel independence and isotropy give the tensor inner product in (10). Appendix A.2 proves the proposition. The target is fixed and realizable in the continuum; consistency means $\mathcal{R}_\varepsilon \to 0$.

4 WHAT VANISHING WEIGHT DECAY LEARNS

We now minimize prediction error plus ordinary squared-parameter regularization:

$$J_\varepsilon^{\mathrm{bare}}(\theta, b) = \frac{1}{2}\mathbb{E}\|N_{\varepsilon,\theta,b} - Y\|_{\mathsf{H}}^2 + \frac{\lambda_\varepsilon}{2}\left[\mathcal{W}(\theta) + \|b\|^2\right], \qquad \mathcal{W}(\theta) = \sum_{j=1}^m (a_j^2 + \|v_j\|^2), \tag{13}$$

where $\lambda_\varepsilon > 0$ and $\lambda_\varepsilon \to 0$. "Weight decay" means this explicit objective. An objective gap is the objective value minus its infimum; a near-minimizing sequence has gaps tending to zero.

4.1 THE PENALTY ON THE REPRESENTED INTERACTION

Feature learning must be included in the analysis. Replacing $\mathcal{W}(\theta)$ by a quadratic penalty on A_θ would change the problem. Instead, define the exact least cost of representing a tensor:

$$\Psi_m(A) = \min_{\theta : A_\theta = A} \mathcal{W}(\theta). \tag{14}$$

For $m \geq D$, this minimum exists for every A. The function Ψ_m is nonnegative, lower semicontinuous, and bounded on bounded subsets of $\mathcal{A}$. Indeed, a fixed rank-one tensor basis supplies a bounded-cost representation of every bounded tensor set, and bounded parameter costs yield compact minimizing sequences. Appendix A.3 gives the details.

Using (9), minimization of (13) is exactly equivalent, at the level of represented predictors, to minimization of

$$\widetilde{J}_\varepsilon^{\mathrm{bare}}(A,\beta) = \mathcal{R}_\varepsilon(A,\beta) + \frac{\lambda_\varepsilon}{2}\Psi_m(A) + \frac{\lambda_\varepsilon}{2}\|\beta - t_\varepsilon LA\|^2. \tag{15}$$

On bounded coefficient sets the hidden-weight cost vanishes, since $\lambda_\varepsilon\Psi_m \to 0$ uniformly. The last term need not vanish. Its effective strength is

$$\mu_\varepsilon = \lambda_\varepsilon t_\varepsilon^2 = 9\lambda_\varepsilon C_\varepsilon^2. \tag{16}$$

This quantity measures the regularization cost of encoding a fixed nonzero contraction. It is the single scale governing the limiting bias.

4.2 AN EXACT LIMITING PREDICTOR

The surviving penalty sees only tensor contractions; let L^* be the adjoint of L and define the trace and harmonic components of A_* by

$$\kappa_d = \frac{d+2}{3}, \qquad A_{\mathrm{T}} = \kappa_d^{-1} L^* L A_*, \qquad A_{\mathrm{H}} = A_* - A_{\mathrm{T}}. \tag{17}$$

These are the orthogonal projections onto the range $\operatorname{ran} L^*$ and kernel $\ker L$. With Δ_z the Laplacian in the channel variable $z \in \mathbb{R}^d$,

$$\Delta_z p_A(z) = 6(LA)^\top z, \qquad p_{A_{\mathrm{T}}}(z) = \kappa_d^{-1} |z|^2 (LA_*)^\top z. \tag{18}$$

Harmonicity means $\Delta_z p_A = 0$.

Theorem 4.1 (Cutoff-induced structural bias)**.** *Fix $d \geq 1$ and $m \geq \binom{d+2}{3}$. Suppose $\varepsilon \downarrow 0$, $\lambda_\varepsilon > 0$, $\lambda_\varepsilon \to 0$, and $\mu_\varepsilon \to \tau \in [0, \infty]$. The limiting optimization problem for* (15) *is*

$$\mathcal{J}_\tau(A, \beta) = \begin{cases} \mathcal{R}(A, \beta) + \dfrac{\tau}{2} \|LA\|^2, & \tau < \infty, \\ \mathcal{R}(A, \beta), & \tau = \infty \text{ and } LA = 0, \\ +\infty, & \tau = \infty \text{ and } LA \neq 0. \end{cases} \tag{19}$$

Minimum values converge to $\min \mathcal{J}_\tau$. Every sequence of global minimizers of (13)*, and every sequence whose objective gaps tend to zero, has represented coefficients satisfying*

$$\boxed{\widehat{A}_\varepsilon \longrightarrow A_\tau := A_{\mathrm{H}} + \frac{\alpha}{\alpha + \kappa_d \tau} A_{\mathrm{T}}, \qquad \widehat{\beta}_\varepsilon \longrightarrow \beta_*.} \tag{20}$$

At $\tau = \infty$ the fraction is defined as zero. The learned outputs converge in $L^2(\Omega; \mathsf{H})$ to $F(A_\tau, \beta_*)$, and their prediction errors obey

$$\lim_{\varepsilon \downarrow 0} \mathcal{R}_\varepsilon(\widehat{A}_\varepsilon, \widehat{\beta}_\varepsilon) = \frac{\alpha}{2} \left(\frac{\kappa_d \tau}{\alpha + \kappa_d \tau} \right)^2 \|A_{\mathrm{T}}\|_{\mathrm{F}}^2, \tag{21}$$

where the ratio is one at $\tau = \infty$.

The three regimes have different statistical meanings. When $\tau = 0$, regularization disappears in the coordinates of the interaction. When $0 < \tau < \infty$, a nonzero contraction penalty survives despite $\lambda_\varepsilon \to 0$. When $\tau = \infty$, learning enforces $LA = 0$ and projects the target cubic onto the harmonic subspace. Only the trace component is suppressed; the linear coefficient and harmonic cubic component are recovered.

Appendix B.1 proves the compactness and limiting-minimizer argument for finite and infinite τ, including vanishing objective gaps. The identity $LL^* = \kappa_d I$ and trace projection $\kappa_d^{-1} L^* L$ give the predictor (20) and risk (21).

4.3 SHARP CONSISTENCY AND EXACT EXAMPLES

Corollary 4.2 (Necessary and sufficient decay rate)**.** *Under the fixed-model assumptions of Theorem 4.1, retain $\lambda_\varepsilon \to 0$ but do not assume that μ_ε converges. For every nonharmonic target, $LA_* \neq 0$,*

$$\mathcal{R}_\varepsilon(\widehat{A}_\varepsilon, \widehat{\beta}_\varepsilon) \longrightarrow 0 \quad \Longleftrightarrow \quad \lambda_\varepsilon C_\varepsilon^2 \longrightarrow 0. \tag{25}$$

Harmonic targets, $LA_* = 0$, are recovered for every vanishing penalty schedule. Consequently, consistency for all *targets holds exactly when $\lambda_\varepsilon \log^2(1/\varepsilon) \to 0$.*

For one channel, L is the identity and $\ker L = \{0\}$. Taking $\lambda_\varepsilon = C_\varepsilon^{-1} \to 0$ gives $\mu_\varepsilon = 9C_\varepsilon \to \infty$. A target with nonzero cubic coefficient a_* is then learned as the purely linear operator $F(0, \beta_*)$, with limiting prediction error $\alpha a_*^2/2$.

For several channels, the interaction $p(z) = |z|^2 z_1$ is purely trace, since its tensor is $L^* e_1$, where e_1 is the first coordinate vector. The same regime eliminates this interaction in every channel dimension. In contrast, for $d \geq 2$, $p(z) = z_1^3 - 3z_1 z_2^2$ is harmonic and is retained.

5 COUNTERTERM-COUPLED REGULARIZATION

Counterterm-coupled regularization penalizes finite Wick coefficients while retaining hidden-weight learning. It preserves the chosen objective and restores consistency as regularization vanishes; at positive strength, the intended statistical bias remains.

Definition 5.1 (Counterterm-coupled regularization)**.** For the operator (4), define

$$\boxed{\mathcal{P}_\varepsilon^{\mathrm{ren}}(\theta, b) = \mathcal{W}(\theta) + \|b + t_\varepsilon L A_\theta\|^2 = \sum_j (a_j^2 + \|v_j\|^2) + \|b + t_\varepsilon \sum_j a_j \|v_j\|^2 v_j\|^2.} \tag{26}$$

The corrected objective is

$$J_\varepsilon^{\mathrm{ren}}(\theta, b) = \frac{1}{2}\mathbb{E}\|N_{\varepsilon,\theta,b} - Y\|_{\mathsf{H}}^2 + \frac{\lambda_\varepsilon}{2}\mathcal{P}_\varepsilon^{\mathrm{ren}}(\theta, b). \tag{27}$$

The bijection $(\theta, \beta) \mapsto (\theta, b = \beta - t_\varepsilon L A_\theta)$ preserves the hypothesis class and parameter count. In Wick coordinates the model is

$$\overline{N}_{\varepsilon,\theta,\beta} = S\left[\sum_j a_j \{(v_j^\top X_\varepsilon)^3 - 3C_\varepsilon \|v_j\|^2 (v_j^\top X_\varepsilon)\} + \beta^\top X_\varepsilon\right], \tag{28}$$

and the penalty is exactly $\mathcal{W}(\theta) + \|\beta\|^2$. Thus counterterm coupling can equivalently be expressed as ordinary regularization of this Wick-parameterized network.

Theorem 5.2 (Consistency without a cutoff-dependent decay condition)**.** *Fix d, m, A_*, β_* as above. For every $\lambda_\varepsilon > 0$ with $\lambda_\varepsilon \to 0$, every sequence of global minimizers of* (27)*, or of parameters with objective gaps $\eta_\varepsilon \to 0$, satisfies*

$$\mathcal{R}_\varepsilon(\widehat{A}_\varepsilon, \widehat{\beta}_\varepsilon) \longrightarrow 0, \qquad (\widehat{A}_\varepsilon, \widehat{\beta}_\varepsilon) \longrightarrow (A_*, \beta_*). \tag{29}$$

No condition on $\lambda_\varepsilon C_\varepsilon^2$ is needed. More precisely, for any fixed hidden representation θ_* of A_*,

$$\mathcal{R}_\varepsilon(\widehat{A}_\varepsilon, \widehat{\beta}_\varepsilon) \le \mathcal{R}_\varepsilon(A_*, \beta_*) + \frac{\lambda_\varepsilon}{2}\left[\mathcal{W}(\theta_*) + \|\beta_*\|^2\right] + \eta_\varepsilon. \tag{30}$$

Appendix B.2 proves (30) by comparison with the fixed renormalized target and also gives the fixed-positive-penalty limit and the consistent unpenalized-counterterm alternative.

The same phenomenon with sampled fields. Let n_ε independent copies of (X, Y) replace population expectations by sample averages in either objective. For fixed d, m, if $n_\varepsilon \to \infty$, the conclusions of Theorems 4.1 and 5.2 hold in probability for measurable empirical global minimizers, and for empirical objective gaps tending to zero in probability. No additional rate relation between n_ε and ε is necessary. Appendix B.3 proves the result through cutoff-uniform fourth moments and finite-dimensional empirical Gram matrices. Thus increasing the sample size reproduces the population bias of the bare penalty; it does not eliminate it.

6 A CONTINUUM-NULL CHANNEL CHANGES THE LEARNED INTERACTION

The preceding theorem fixes the architecture; we now change its realization cost while holding its continuum predictor set fixed. A cutoff-indexed family is the relevant comparison because physical refinement changes both the input field and the coefficients required to realize its nonlinear interactions. Keep the scalar spatial model, the observation $S = e^{\Delta_x}$, the fixed target $Y = F(1, 0) = S{:}X^3{:}$, and $\lambda_\varepsilon = C_\varepsilon^{-1}$. Write $F_{p,\varepsilon} = S{:}X_\varepsilon^p{:}$ for $p = 1, 3, 5$. Add one independently trainable raw quintic neuron:

$$N_\varepsilon^{(5)} = S[u_3(v_3 X_\varepsilon)^3 + u_5(v_5 X_\varepsilon)^5 + bX_\varepsilon], \quad a = u_3 v_3^3, \quad c = u_5 v_5^5, \quad |C_\varepsilon c| \le 1. \tag{31}$$

Ordinary weight decay penalizes $u_3^2 + v_3^2 + u_5^2 + v_5^2 + b^2$. The original scalar architecture omits the quintic neuron. All displayed weights are trainable.

For a scalar degree-$p \ge 1$ neuron the exact least squared-weight cost is

$$\Psi_p(s) = \min_{uv^p = s}(u^2 + v^2) = (p+1)p^{-p/(p+1)}|s|^{2/(p+1)}. \tag{32}$$

For $s \neq 0$, minimize $s^2/t^p + t$ over $t = v^2 > 0$; strict convexity gives $t = (ps^2)^{1/(p+1)}$. At $s = 0$, both weights vanish. Thus the exact reduced ordinary penalty is $\Psi_3(a) + \Psi_5(c) + b^2$, preserving hidden-weight learning. It is not quadratic ridge on the represented coefficients.

6.1 THE SEARCHED CANCELLATION AND ITS PHYSICAL CAPACITY

We seek a bounded-weight realization of the target whose fifth-chaos component vanishes.

With $C = C_\varepsilon$, the coefficients of $F_{p,\varepsilon}$ in the network output are

$$r_5 = c, \qquad r_3 = a + 10Cc, \qquad r_1 = b + 3Ca + 15C^2c. \tag{33}$$

The polynomial match in (C) yields $a = -1$, $c = (5C)^{-1}$, and $b = 0$. Appendix C.1 gives its neural realization, bounded weight cost, and vanishing fifth-chaos bound.

Let $\mathcal{H}^{(p)}_{\mathrm{cont}}$ be the set of $L^2(\Omega; \mathsf{H})$ predictor limits of the degree-p family as $\varepsilon \downarrow 0$. Orthogonality and the positive limiting Gram bound r_3, r_1 on bounded-risk sequences; the vanishing fifth component makes every limit cubic-plus-linear. Setting $c = 0$ realizes every such limit, so $\mathcal{H}^{(3)}_{\mathrm{cont}} = \mathcal{H}^{(5)}_{\mathrm{cont}} = \{F(a, \beta) : a, \beta \in \mathbb{R}\}$. This equality has no raw-weight budget: bounded original weights force a zero cubic limit, whereas the extension reaches Y at bounded cost. The raw branch itself need not vanish; its contractions implement the cancellation identity.

6.2 GLOBAL OUTCOMES AND MATCHED CONTROLS

Chaos orthogonality gives the exact spatial risk for the global-optimum and matched-control comparisons.

Let $A_0 = \mathbb{E}\|Y\|^2$, $A_{3,\varepsilon} = \mathbb{E}\|F_{3,\varepsilon}\|^2$, $B_{1,\varepsilon} = \mathbb{E}\|F_{1,\varepsilon}\|^2$, $D_{5,\varepsilon} = \mathbb{E}\|F_{5,\varepsilon}\|^2$, $K_{3,\varepsilon} = \mathbb{E}\langle F_{3,\varepsilon}, Y\rangle$, and $E_{3,\varepsilon} = A_{3,\varepsilon} - 2K_{3,\varepsilon} + A_0$. The exact spatial population risk is

$$\mathcal{R}_\varepsilon = \tfrac{1}{2}[A_{3,\varepsilon} r_3^2 - 2K_{3,\varepsilon} r_3 + A_0 + B_{1,\varepsilon} r_1^2 + D_{5,\varepsilon} r_5^2]. \tag{38}$$

Theorem 6.1 (Architectural sensitivity under a continuum-null extension)**.** *For the scalar cutoff-aware family* (31)*, with $|C_\varepsilon c| \le 1$, fixed target $Y = S{:}X^{\otimes 3}$, and ordinary squared-weight decay on all trainable parameters, every sequence of global minimizers is consistent whenever $\lambda_\varepsilon > 0$ and $\lambda_\varepsilon \to 0$. For the fixed schedule $\lambda_\varepsilon = C_\varepsilon^{-1}$,*

$$(r_3, r_1, C_\varepsilon c, a) \longrightarrow (1, 0, 1/5, -1), \qquad \mathcal{R}_\varepsilon \longrightarrow 0. \tag{39}$$

The original cubic architecture instead has $(r_3, r_1) \to (0, 0)$ and risk tending to $A_0/2$. The same limits hold for objective gaps tending to zero.

The parameter-matched placebo adds an independent Wick-quintic neuron $u_5 v_5^{\otimes 5} F_{5,\varepsilon}$ to the original cubic and has the same population-optimal predictor set. The corrected reduced penalty $\Psi_3(r_3) + \Psi_5(r_5) + r_1^2$ gives identical cubic and quintic optimal-predictor sets at every cutoff and consistency for every vanishing penalty schedule. Appendix C.2 gives the control objectives, transported map, and proof.

Let Π_I denote Euclidean projection onto a closed interval I.

Proposition 6.2 (The null channel selects a low-cost interval)**.** *In the scalar architecture of Section 6, replace $|C_\varepsilon c| \le 1$ by $|C_\varepsilon c| \le M$, where $0 \le M < \infty$ is fixed. Take the fixed target $Y_{a_*} = a_* F_{3,0}$, with $F_{3,0} = S{:}X^{\otimes 3}$, denote its risk by $\mathcal{R}^{a_*}_\varepsilon$, and set $\lambda_\varepsilon = C_\varepsilon^{-1}$. The continuum predictor class is unchanged. Every sequence of global or vanishing-objective-gap minimizers satisfies*

$$(r_3, r_1, C_\varepsilon c, a) \longrightarrow (s_M, 0, s_M/5, -s_M), \qquad s_M = \Pi_{[-5M, 5M]} a_*,$$

and

$$\mathcal{R}^{a_*}_\varepsilon \longrightarrow \frac{A_0}{2}(a_* - s_M)^2, \qquad A_0 = \mathbb{E}\|F_{3,0}\|^2.$$

The constraint selects the vanishing-cost part of the common prediction class. It controls class equality rather than causing target recovery: Appendix C.3 proves the same erasure–recovery separation for unconstrained cubic and quintic networks, both with the original input and with known-variance input standardization. Under $\lambda_\varepsilon = C_\varepsilon^{-1}$, their unconstrained quintic minimizers satisfy $C_\varepsilon c \to a_*/5$. By contrast, common bounded-cost recovery sequences would force both architectures to recover the same realizable target.

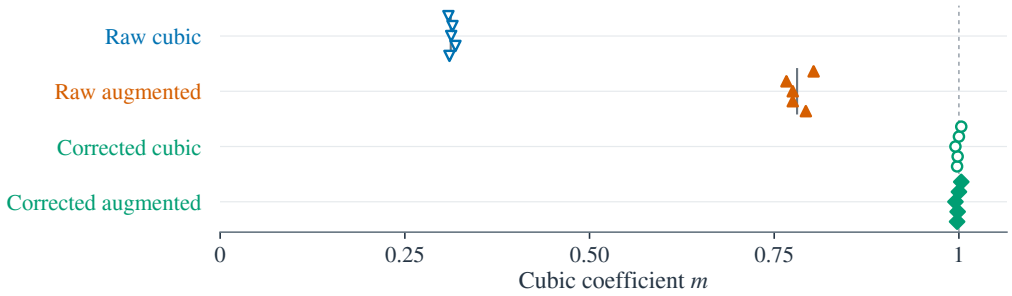


Figure 2: **Architectural separation in trained networks.** All 20 accepted Adam–Newton endpoints at 129^2, with five paired seeds per model. The cubic coefficient is normalized by its unregularized population optimum: $m = r_3/r_{0,129}$. Vertical offsets separate seeds; short ticks mark regularized population optima and the dashed line marks $m = 1$. Every endpoint passes the empirical-gap and population-agreement criteria (Appendix E).

6.3 Executed finite-cutoff certificates

At the physical cutoff defined by $C_\varepsilon = 10^4$, for which $\varepsilon < e^{-1248.625}$, every original/placebo global optimum has $\rho \geq 0.9271$ and every raw-augmented optimum has $\rho \leq 0.006301$, where $\rho = 2\mathcal{R}/A_0$. These outward-rounded spatial-moment bounds certify separation at every global minimizer at that finite cutoff (Appendix C.4).

6.4 Moderate-grid neural training

At 129^2, all twenty registered Adam–Newton runs meet the independently evaluated empirical-gap tolerance and population-coefficient criterion: the maximum normalized gap is 4.03×10^{-9} and the maximum normalized coefficient deviation is 0.0224614. Figure 2 shows the raw models' separation and corrected models' agreement. Appendix E gives the methods and results across resolutions (Figure E.1).

6.5 Consistency is a realization-cost property

The preceding separation is a realization-cost effect. For a real Hilbert space $\mathcal{H}$, target $Y \neq 0$, prediction $F_\varepsilon(\theta)$, and penalty $P_\varepsilon(\theta) \geq 0$, write $J_\varepsilon = \frac{1}{2}\|F_\varepsilon - Y\|_{\mathcal{H}}^2 + \frac{\lambda_\varepsilon}{2} P_\varepsilon$, where $\lambda_\varepsilon > 0$ tends to zero. A recovery sequence satisfies $F_\varepsilon(\theta_\varepsilon) \to Y$ in $\mathcal{H}$; consistency requires this convergence for near-minimizers. Constants in the bounds below are positive and independent of ε.

Theorem 6.3 (Vanishing-cost characterization of consistent learning)**.** *Assume a zero prediction with zero penalty, and output-scaling maps $T_{\varepsilon,t}$ satisfying $F_\varepsilon(T_{\varepsilon,t}\theta) = tF_\varepsilon(\theta)$ and $P_\varepsilon(T_{\varepsilon,t}\theta) \leq t^a P_\varepsilon(\theta)$ for $0 < t < 1$ and a fixed $a > 0$. Then every near-minimizing sequence is consistent if and only if some such sequence is consistent, if and only if there is a recovery sequence with $F_\varepsilon(\theta_\varepsilon^*) \to Y$ and $\lambda_\varepsilon P_\varepsilon(\theta_\varepsilon^*) \to 0$. Moreover, suppose a cost scale $s_\varepsilon \to \infty$, a recovery sequence has $P_\varepsilon \leq K s_\varepsilon$, and, on bounded-risk sets,*

$$|\langle F_\varepsilon(\theta), Y\rangle| \leq B s_\varepsilon^{-\nu}(1 + P_\varepsilon(\theta)^\nu), \qquad \nu > 0.$$

Then near-minimizers are consistent exactly when $\lambda_\varepsilon s_\varepsilon \to 0$; if $\lambda_\varepsilon s_\varepsilon \to \infty$, their predictions converge to zero.

Appendix D proves the theorem and its null-extension specialization: a diverging skip cost enforces cancellation by the null channels; a finite surviving cost penalizes squared distance to their attainable counterterms. This yields the contraction law and the interval in Proposition 6.2.

For a scalar-input channel network with ReLU activation $z_+ = \max\{z, 0\}$ and L affine layers, let f_θ denote its output, with trainable weights W_j and biases b_j. Set $P(\theta) = \sum_{j=1}^{L}(\|W_j\|_{\mathrm{F}}^2 + \|b_j\|^2)$ and observe $S f_\theta(X_\varepsilon)$ through $S = e^{\Delta_x}$.

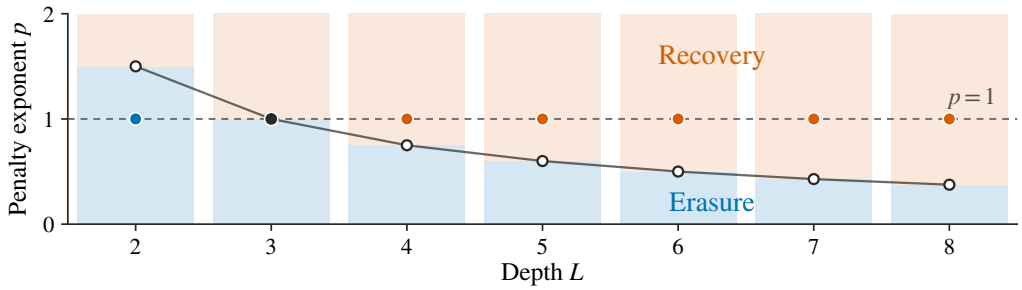


Figure 3: **Depth changes the consistency threshold.** For $\lambda_\varepsilon = \kappa C_\varepsilon^{-p}$, $\kappa, p > 0$, Theorem 6.4 gives recovery above $p = 3/L$ and erasure below it; the boundary is inconsistent. Depth is integer, and connecting segments are guides. The dashed $p = 1$ schedule erases at $L = 2$, is critical at $L = 3$, and recovers for $L \geq 4$.

Theorem 6.4 (Depth controls the consistency threshold)**.** *For every fixed $L \geq 2$, the unions over finite hidden widths have the same finite-cutoff prediction class. Every sequence recovering $Y = S{\cdot}X^3$ has cost at least $c_L C_\varepsilon^{3/L}$ eventually, uniformly over width. A width-four recovery sequence has cost at most $K_L C_\varepsilon^{3/L}$. Consequently, for $\lambda_\varepsilon \to 0$, all near-minimizers are consistent exactly when*

$$\lambda_\varepsilon C_\varepsilon^{3/L} \longrightarrow 0.$$

If this product diverges, their predictions converge to zero. At $\lambda_\varepsilon = \kappa / C_\varepsilon$, $\kappa > 0$, depth two erases the cubic, depth three is inconsistent, and every fixed depth $L \geq 4$ is consistent.

The width-independent bias bound supplies the lower cost; a four-channel piecewise-affine recovery function supplies the upper cost, including its infinite higher-chaos remainder. Output scaling gives the exponent $a = 2/L$ in Theorem 6.3. Figure 3 displays the depth-dependent threshold.

Standard channel-network instance. Here $Z_0 = \mathbb{E}\|Y_{129}\|_{\mathsf{H}}^2$, where Y_{129} is the target's conditional mean on the reference 129^2 input grid (Appendix E.1). Penalty continuation reaches a low-cost deep solution from every archived initialization at both tested grids. All ten primary deep endpoints have certified objective advantages of at least $0.297958 Z_0$ over the unrestricted-width shallow optimum, same-function shallow-to-deep realization-cost ratios of at least 23, and certified numerical precision; their certified population-gradient upper bounds are at most 8.09×10^{-6}, so first-order stationarity is not certified at the registered 10^{-6} gate (Appendix F).

7 Discussion and Conclusion

Equal continuum prediction classes can support different learned interactions because the same physical cancellation has architecture-dependent parameter cost. The limiting law isolates a surviving contraction penalty; full Wick-coordinate regularization preserves the intended interaction objective. The erasure–recovery separation holds with the class-preserving quintic bound, without that bound, and after input-variance normalization. At 129^2, Adam–Newton training meets the independently evaluated empirical-optimum and population-agreement criteria in all 20 runs.

The general realization-cost theorem makes the polynomial counterexample an instance of a broader selection principle. For standard ReLU channel networks, recovery cost scales sharply as $C_\varepsilon^{3/L}$, so prediction-equivalent shallow and deep channel architectures can erase or recover the same cubic under the same vanishing penalty. Together with the completed moderate-grid training result, this identifies realization cost, not continuum expressivity alone, as the quantity controlling the learned interaction.

REPRODUCIBILITY STATEMENT

The release supplies the model definitions, proofs, executable configurations, immutable registrations, checkpoints, and numerical certificates underlying the paper. The polynomial study includes population-optimum enclosures and independently evaluated empirical optimization gaps. The standard-channel study includes the unrestricted-width shallow certificate, exact same-function cost bounds, and separate objective-precision and population-gradient certificates. A claim-to-artifact index identifies the canonical inputs, terminal outcomes, reproduction commands, and hashes. Historical studies and superseded protocol generations remain intact in the archive. Rebuilding the publication source verifies its references and reproduces the canonical PDF.

AI USE STATEMENT

Large language models assisted with rewriting parts of the manuscript and debugging code. The author reviewed and verified all changes and takes full responsibility for the content of this work.

REFERENCES

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017. URL https://jmlr.org/papers/v18/14-546.html.

Giuseppe Da Prato and Arnaud Debussche. Strong solutions to the stochastic quantization equations. *The Annals of Probability*, 31(4):1900–1916, 2003. doi: 10.1214/aop/1068646370.

Wenhan Gao, Ruichen Xu, Yuefan Deng, and Yi Liu. Discretization-invariance? on the discretization mismatch errors in neural operators. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=J9FgrqOOni.

Shiqi Gong, Peiyan Hu, Qi Meng, Yue Wang, Rongchan Zhu, Bingguang Chen, Zhiming Ma, Hao Ni, and Tie-Yan Liu. Deep latent regularity network for modeling stochastic partial differential equations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):7740–7747, 2023. doi: 10.1609/aaai.v37i6.25938.

Svante Janson. *Gaussian Hilbert Spaces*, volume 129 of *Cambridge Tracts in Mathematics*. Cambridge University Press, 1997.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24(89):1–97, 2023. URL https://jmlr.org/papers/v24/21-1524.html.

Matti Lassas and Samuli Siltanen. Can one use total variation prior for edge-preserving Bayesian inversion? *Inverse Problems*, 20(5):1537–1563, 2004. doi: 10.1088/0266-5611/20/5/013.

Matti Lassas, Eero Saksman, and Samuli Siltanen. Discretization-invariant Bayesian inversion and Besov space priors. *Inverse Problems and Imaging*, 3(1):87–122, 2009. doi: 10.3934/ipi.2009.3.87.

Johannes W. R. Martini, Francisco Rosales, Ngoc-Thuy Ha, Johannes Heise, Valentin Wimmer, and Thomas Kneib. Lost in translation: On the problem of data coding in penalized whole genome regression with interactions. *G3: Genes|Genomes|Genetics*, 9(4):1117–1129, 2019. doi: 10.1534/g3.118.200961.

NeuralOperator contributors. ChannelMLP: pointwise channel mixing. NeuralOperator source, commit 00b7d86f8d74ff0af55da53eb585fe26df9c71f0, 2026. URL https://neuraloperator.github.io/dev/modules/generated/neuralop.layers.channel_mlp.ChannelMLP.html.

NIST. Digital library of mathematical functions, section 7.9: Continued fractions, 2026. URL https://dlmf.nist.gov/7.9.

Greg Ongie and Rahul Parhi. Representation costs in data science: Foundations and the quasi-Banach spaces of deep neural networks, 2026. URL https://arxiv.org/abs/2606.14954v4. arXiv:2606.14954v4.

Rahul Parhi and Robert D. Nowak. What kinds of functions do deep neural networks learn? insights from variational spline theory. *SIAM Journal on Mathematics of Data Science*, 4(2):464–489, 2022. doi: 10.1137/21M1418642.

Rahul Parhi and Robert D. Nowak. Compositional function spaces for deep learning. *SIAM Review*, 68(1):127–149, 2026. doi: 10.1137/25M1802948. SIGEST revision of Parhi and Nowak (2022).

Tim G. J. Rudner, Sanyam Kapoor, Shikai Qiu, and Andrew Gordon Wilson. Function-space regularization in neural networks: A probabilistic perspective. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29275–29290. PMLR, 2023. URL https://proceedings.mlr.press/v202/rudner23a.html.

Dai Shi, Luke Thompson, Andi Han, Peiyan Hu, Junbin Gao, and José Miguel Hernández-Lobato. Wiener chaos expansion based neural operator for singular stochastic partial differential equations. arXiv:2603.08219v1, 2026. URL https://arxiv.org/abs/2603.08219v1.

Appendices A–C establish the Gaussian learning model, its limiting objectives, and the polynomial architectural separation. Appendix D proves the general realization-cost and ReLU-depth laws. Appendices E–F give the population calculations, training protocols, and numerical certificates for the two reported demonstrations. Appendix G positions these results against the closest related work.

A Gaussian features and neural realization

A.1 Construction and covariance estimates

Let W_i be independent centered Gaussian families on H with $\mathbb{E}[W_i(f)W_i(g)] = \langle f, g\rangle_{\mathsf{H}}$ (isonormal processes). Set $X_i(f) = W_i((1-\Delta_x)^{-1/2}f)$. This realizes (1), and $\mathbb{E}\|X_i\|^2_{H^{-r}} = \sum_{k\in\mathbb{Z}^2}(1+|k|^2)^{-1-r} < \infty$ for every $r > 0$, where H^{-r} is the Sobolev space of order $-r$. Exponential Fourier decay makes every positive heat cutoff almost surely smooth. All cutoffs use this common probability space. With p_s the heat kernel relative to μ, their cross-covariance at displacement $r \in \mathbb{T}^2$ is

$$G_{\varepsilon,\delta}(r) = \sum_{k\in\mathbb{Z}^2} \frac{e^{-(\varepsilon^2+\delta^2)|k|^2/2}}{1+|k|^2} e^{ik\cdot r} = \int_0^\infty e^{-u} p_{u+(\varepsilon^2+\delta^2)/2}(r)\, du. \tag{A.1}$$

A zero cutoff denotes a covariance kernel; $G = G_{0,0}$ is defined off the diagonal. Write $\rho(r) = \operatorname{dist}_{\mathbb{T}^2}(r,0)$ and $\log_+ a = \max\{\log a, 0\}$. Here K denotes a finite constant independent of the cutoffs and may change between bounds.

Lemma A.1 (Logarithmic covariance control)**.** *For $0 \le \varepsilon, \delta \le 1$ and $r \neq 0$,*

$$0 < G_{\varepsilon,\delta}(r) \le K[1 + \log_+(1/\rho(r))]. \tag{A.2}$$

As $(\varepsilon,\delta) \to (0,0)$, $G_{\varepsilon,\delta}(r) \to G(r)$ off the diagonal, and

$$C_\varepsilon = 2\pi \log(1/\varepsilon) + O(1). \tag{A.3}$$

Proof. The periodic Gaussian formula and its nearest-lattice-point estimate give

$$p_s(r) = \frac{\pi}{s} \sum_{n\in\mathbb{Z}^2} e^{-|r+2\pi n|^2/(4s)} > 0, \qquad p_s(r) \le K[1 + s^{-1}e^{-\rho(r)^2/(Ks)}] \quad (0 < s \le 2).$$

For $s \ge 1$, p_s is uniformly bounded. Put $a = (\varepsilon^2+\delta^2)/2 \le 1$ in (A.1). The part $u \ge 1$ is bounded; the singular part is at most

$$K\int_a^{1+a} v^{-1}e^{-\rho(r)^2/(Kv)}\, dv \le K\int_0^2 v^{-1}e^{-\rho(r)^2/(Kv)}\, dv \le K[1+\log_+(1/\rho(r))].$$

Split at $v = \rho(r)^2$ to obtain the last inequality. Positivity follows from the kernel formula. At fixed $r \neq 0$, the same estimate uniformly bounds $p_{u+a}(r)$ near $u = a = 0$, so dominated convergence proves the covariance limit. At the origin, $p_v(0) = \pi/v + O(1)$ uniformly for $0 < v \le 2$; hence

$$C_\varepsilon = \pi\int_0^1 \frac{e^{-u}}{u+\varepsilon^2}\, du + O(1) = \pi \log\frac{1+\varepsilon^2}{\varepsilon^2} + O(1).$$

Replacing e^{-u} by 1 costs a bounded integral. Every finite power of the logarithmic envelope is integrable on $\mathbb{T}^2$. □

A.2 Wick covariance, feature limits, and uniform moments

For general degree p, the Wick polynomial is $x^{p:}_C = C^{p/2}H_p(x/\sqrt{C})$, where $H_p(z) = (-1)^p e^{z^2/2}(d/dz)^p e^{-z^2/2}$ is the probabilists' Hermite polynomial. Let $h = p_2$, the smooth, bounded, positive kernel of S^*S. Define

$$\alpha_{\varepsilon,\delta} = 6\int_{\mathbb{T}^2} hG^3_{\varepsilon,\delta}\, d\mu, \qquad \gamma_{\varepsilon,\delta} = \int_{\mathbb{T}^2} hG_{\varepsilon,\delta}\, d\mu, \qquad (\alpha,\gamma) = (\alpha_{0,0},\gamma_{0,0}). \tag{A.4}$$

Lemma A.2 (Cross-cutoff feature isometry)**.** *For positive cutoffs and every scalar order $p \geq 1$, put $F_{p,\varepsilon} = S[:X^p_{\varepsilon,1}:]$. Then*

$$\mathbb{E}\langle F_{p,\varepsilon}, F_{p,\delta}\rangle_{\mathsf{H}} = p! \int_{\mathbb{T}^2} h(r) G_{\varepsilon,\delta}(r)^p \, d\mu(r). \tag{35}$$

Distinct orders are orthogonal. For $A, B \in \mathcal{A}$ and $\beta, \eta \in \mathbb{R}^d$,

$$\mathbb{E}\langle F_\varepsilon(A,\beta), F_\delta(B,\eta)\rangle_{\mathsf{H}} = \alpha_{\varepsilon,\delta}\langle A, B\rangle_{\mathrm{F}} + \gamma_{\varepsilon,\delta}\beta^\top\eta. \tag{A.5}$$

The constants converge to $\alpha, \gamma > 0$ as both cutoffs vanish.

Proof. Wick subtraction removes internal pairings, leaving $p!$ bijections between equal-order factors and none between different orders (Janson, 1997). For tensors, channel independence and symmetry give the pointwise cubic covariance $6G_{\varepsilon,\delta}(x-y)^3\langle A, B\rangle_{\mathrm{F}}$; the linear covariance is $G_{\varepsilon,\delta}(x-y)\beta^\top\eta$. Integrating against $h(x-y)$ proves both identities by stationarity and $\mu(\mathbb{T}^2) = 1$; smooth Gaussian moments justify Fubini. Lemma A.1 gives dominated convergence for every fixed order, and positivity of h, G gives positive limiting constants. □

Proof of Proposition 3.1. The isometry gives

$$\begin{aligned}\mathbb{E}\|F_\varepsilon(A,\beta) - F_\delta(A,\beta)\|^2_{\mathsf{H}} &= (\alpha_{\varepsilon,\varepsilon} + \alpha_{\delta,\delta} - 2\alpha_{\varepsilon,\delta})\|A\|^2_{\mathrm{F}} \\ &\quad + (\gamma_{\varepsilon,\varepsilon} + \gamma_{\delta,\delta} - 2\gamma_{\varepsilon,\delta})\|\beta\|^2 \longrightarrow 0.\end{aligned}$$

Thus the limit defines a linear F. Applying this to a finite orthonormal coefficient basis proves operator-norm convergence; passing to the limit proves (10). The same Cauchy argument gives convergence and uniform second moments for each fixed scalar order. Letting $\delta \downarrow 0$ in the cross-cutoff identity gives

$$\begin{aligned}\mathcal{R}_\varepsilon(A,\beta) = \frac{\alpha_{\varepsilon,\varepsilon}}{2}\|A\|^2_{\mathrm{F}} - \alpha_{\varepsilon,0}\langle A, A_*\rangle_{\mathrm{F}} + \frac{\alpha}{2}\|A_*\|^2_{\mathrm{F}} \\ + \frac{\gamma_{\varepsilon,\varepsilon}}{2}\|\beta\|^2 - \gamma_{\varepsilon,0}\beta^\top\beta_* + \frac{\gamma}{2}\|\beta_*\|^2.\end{aligned} \tag{A.6}$$

Coefficient convergence proves local uniform risk convergence. Eventually $\alpha_{\varepsilon,\varepsilon} \geq \alpha/2$, $\gamma_{\varepsilon,\varepsilon} \geq \gamma/2$, and the cross-cutoff coefficients are bounded. Young's inequality on the linear terms therefore gives the stated uniform coercivity. □

Write $V = \mathcal{A} \times \mathbb{R}^d$ with $\|(A,\beta)\|^2_V = \|A\|^2_{\mathrm{F}} + \|\beta\|^2$.

Lemma A.3 (Cutoff-uniform moments)**.** *For a finite K, every $c \in V$, and $0 < \varepsilon \leq 1$,*

$$\mathbb{E}\|F_\varepsilon(c)\|^4_{\mathsf{H}} \leq K\|c\|^4_V, \qquad \mathbb{E}\|F(c)\|^4_{\mathsf{H}} \leq K\|c\|^4_V. \tag{A.7}$$

Proof. Scalar hypercontractivity on chaos q gives $\|Z\|_{L^4} \leq 3^{q/2}\|Z\|_{L^2}$ (Janson, 1997). For finite orthogonal projections of an H-valued variable in that chaos,

$$\mathbb{E}\Big(\sum_r Z_r^2\Big)^2 \leq \sum_{r,s}(\mathbb{E}Z_r^4)^{1/2}(\mathbb{E}Z_s^4)^{1/2} \leq 3^{2q}\Big(\sum_r \mathbb{E}Z_r^2\Big)^2.$$

Monotone convergence extends this to H. Apply it to the cubic and linear components, of orders three and one: Lemma A.1 uniformly bounds their second moments. The $L^4(\Omega;\mathsf{H})$ triangle inequality bounds their sum. Fatou's lemma along an almost surely convergent subsequence of the L^2 feature convergence proves the limit bound. □

A.3 Neural realization and the exact penalty

Lemma A.4 (Representability and least parameter cost)**.** *For fixed $m \geq D$, every $A \in \mathcal{A}$ is represented by A_θ; $\Psi_m(A)$ is attained and lower semicontinuous, $\Psi_m(0) = 0$, and*

$$0 \leq \Psi_m(A) \leq K_m(1 + \|A\|^2_{\mathrm{F}}). \tag{A.8}$$

Proof. Orthogonality of B to every $v^{\otimes 3}$ implies the polynomial $p_B(v)$ vanishes identically, hence $B = 0$. Choose a rank-one basis $u_j^{\otimes 3}$ and its linear coordinates $c_j(A)$. The representation $a_j = c_j(A)$, $v_j = u_j$ for $j \le D$, with other units zero, proves the bound. All-zero parameters represent zero at zero cost. The representation constraint is closed, and bounded cost bounds all parameters in fixed finite dimension; every minimizing sequence therefore has a convergent subsequence attaining its infimum. If $A_n \to A$ with finite $\liminf \Psi_m(A_n)$, choose minimizing representations along a subsequence attaining that lower limit. Compactness gives a limit representing A, and continuity of $\mathcal{W}$ yields $\Psi_m(A) \le \liminf \Psi_m(A_n)$. An infinite lower limit needs no argument. □

Lemma A.5 (Exact reduction and minimizer existence)**.** *For every $\varepsilon, \lambda_\varepsilon > 0$, the infimum of $J_\varepsilon^{\mathrm{bare}}$ equals the minimum of* (15)*; the corresponding equality holds for $J_\varepsilon^{\mathrm{ren}}$ and* (B.4)*. Both neural minima are attained. A neural objective gap at most η induces a reduced gap at most η.*

Proof. Set $A = A_\theta$, $\beta = b + t_\varepsilon LA$. Equation (9) identifies the risks, and $\Psi_m(A) \le \mathcal{W}(\theta)$ gives the reduced inequality. Conversely, a least-cost representation and $b = \beta - t_\varepsilon LA$ give equality. Thus infima agree and gaps transfer. The bare objective is continuous and coercive in (θ, b) because its penalty bounds both. The corrected objective is continuous and coercive in (θ, β) for the same reason; this coordinate change is bijective. Finite-dimensional compactness gives both neural minima, hence the reduced minima. □

The reduction preserves the actual neural penalty and requires no convexity of the representation cost.

B Learning limits and sampling

B.1 The structural learning limit

Lemma B.1 (Convergence of reduced optimization problems)**.** *Suppose $\lambda_\varepsilon > 0$, $\lambda_\varepsilon \to 0$, and $\mu_\varepsilon = \lambda_\varepsilon t_\varepsilon^2 \to \tau \in [0, \infty]$. For $c_\varepsilon \to c$ in V,*

$$\liminf_{\varepsilon \downarrow 0} \widetilde{J}_\varepsilon^{\mathrm{bare}}(c_\varepsilon) \ge \mathcal{J}_\tau(c). \tag{B.1}$$

For every c with finite $\mathcal{J}_\tau(c)$, the constant recovery sequence satisfies

$$\widetilde{J}_\varepsilon^{\mathrm{bare}}(c) \longrightarrow \mathcal{J}_\tau(c). \tag{B.2}$$

Every fixed objective sublevel lies in one bounded coefficient set for all sufficiently small cutoffs.

Proof. Nonnegative penalties and uniform risk coercivity give the sublevel bound. For finite τ, $\lambda_\varepsilon \Psi_m \to 0$ uniformly on bounded sets, and

$$\frac{\lambda_\varepsilon}{2}\|\beta - t_\varepsilon LA\|^2 = \frac{\mu_\varepsilon}{2}\|LA\|^2 - \lambda_\varepsilon t_\varepsilon \beta^\top LA + \frac{\lambda_\varepsilon}{2}\|\beta\|^2. \tag{22}$$

Since $\lambda_\varepsilon t_\varepsilon = \mu_\varepsilon / t_\varepsilon \to 0$, the objectives converge locally uniformly to $\mathcal{J}_\tau$, proving both bounds. For $\tau = \infty$, bounded objective values imply

$$\|LA - \beta / t_\varepsilon\|^2 \le O(1)/\mu_\varepsilon \longrightarrow 0. \tag{23}$$

For $c_\varepsilon \to (A, \beta)$ with $LA \neq 0$, the left side tends to $\|LA\|^2 > 0$; its product with $\mu_\varepsilon/2$ diverges. If $LA = 0$, nonnegativity and risk convergence give the lower bound. At fixed (A, β) with $LA = 0$, the penalty is $\lambda_\varepsilon[\Psi_m(A) + \|\beta\|^2]/2 \to 0$, giving recovery. □

Lemma B.2 (Harmonic and trace components)**.** *The contraction adjoint and its composition satisfy*

$$(L^* v)_{ijk} = \frac{v_i \delta_{jk} + v_j \delta_{ik} + v_k \delta_{ij}}{3}, \qquad LL^* = \kappa_d I_d. \tag{24}$$

The orthogonal projections onto $\operatorname{ran} L^*$ *and* $\ker L$ *are $P_{\mathrm{T}} = \kappa_d^{-1} L^* L$ and $P_{\mathrm{H}} = I - P_{\mathrm{T}}$, respectively, and*

$$p_{L^* v}(z) = |z|^2 v^\top z, \qquad \Delta_z p_A(z) = 6(LA)^\top z. \tag{B.3}$$

Proof. Symmetry of A identifies $\sum_{i,j} A_{ijj} v_i$ with its Frobenius product against the tensor in (24). Summing $(L^* v)_{ijj}$ over j gives $(d+2)v_i/3$. Thus P_{T} is self-adjoint and $P_{\mathrm{T}}^2 = \kappa_d^{-2} L^*(LL^*)L = P_{\mathrm{T}}$, with the asserted range and kernel. Substitution and differentiation give (B.3). □

Proof of Theorem 4.1. Lemma A.5 transfers neural gaps to reduced gaps. Write $v_\varepsilon = \min \widetilde{J}_\varepsilon^{\mathrm{bare}}$ and $\widetilde{J}_\varepsilon^{\mathrm{bare}}(\widehat{c}_\varepsilon) \le v_\varepsilon + \eta_\varepsilon$, where $\eta_\varepsilon \to 0$. Comparison with zero gives $v_\varepsilon \le \mathbb{E}\|Y\|_{\mathsf{H}}^2/2$, so Lemma B.1 bounds $\widehat{c}_\varepsilon$. At any cluster point $\widehat{c}$, for every finite-objective c,

$$\mathcal{J}_\tau(\widehat{c}) \le \liminf \widetilde{J}_\varepsilon^{\mathrm{bare}}(\widehat{c}_\varepsilon) \le \limsup \widetilde{J}_\varepsilon^{\mathrm{bare}}(\widehat{c}_\varepsilon) \le \lim \widetilde{J}_\varepsilon^{\mathrm{bare}}(c) = \mathcal{J}_\tau(c).$$

Hence every cluster point minimizes the limit. Applying the lower bound to exact minimizers and recovery at a limiting minimizer gives $v_\varepsilon \to \min \mathcal{J}_\tau$. For finite τ, strict convexity gives the normal equations $(\alpha I + \tau L^* L)A = \alpha A_*$ and $\beta = \beta_*$. Lemma B.2 yields (20). For infinite τ, orthogonal projection onto $\ker L \times \mathbb{R}^d$ gives the unique minimizer $(A_{\mathrm{H}}, \beta_*)$. Uniqueness forces the entire near-minimizer sequence to converge. The predictions satisfy

$$\|F_\varepsilon(\widehat{c}_\varepsilon) - F(A_\tau, \beta_*)\|_{L^2(\Omega;\mathsf{H})} \le \|F_\varepsilon - F\|_{\mathrm{op}} \|\widehat{c}_\varepsilon\|_V + \|F\|_{\mathrm{op}} \|\widehat{c}_\varepsilon - (A_\tau, \beta_*)\|_V \longrightarrow 0.$$

Local uniform risk convergence then gives $\mathcal{R}_\varepsilon(\widehat{c}_\varepsilon) \to \alpha \|A_\tau - A_*\|_{\mathrm{F}}^2/2$; substitution of (20) is exactly (21). □

Proof of Corollary 4.2. If $\mu_\varepsilon \to 0$, apply Theorem 4.1 with $\tau = 0$. Otherwise some subsequence has $\mu_\varepsilon \ge a > 0$ and a further subsequence converges in $[a, \infty]$. If $LA_* \ne 0$, then $A_{\mathrm{T}} \ne 0$, so (21) is strictly positive there, excluding consistency. If $LA_* = 0$, every cutoff subsequence has a further subsequence on which μ_ε converges in $[0, \infty]$ and the limiting risk is zero. A subsequence of risks bounded below by a positive constant is therefore impossible. Finally $\mu_\varepsilon = 9\lambda_\varepsilon C_\varepsilon^2$ and $C_\varepsilon \asymp \log(1/\varepsilon)$ give the equivalent decay conditions. □

B.2 Corrected objectives and consistent alternatives

Proof of Theorem 5.2. For a fixed representation θ_* of A_*, choose $b_{\varepsilon,*} = \beta_* - t_\varepsilon L A_*$. Its corrected penalty is $\mathcal{W}(\theta_*) + \|\beta_*\|^2$. Target comparison gives

$$\mathcal{R}_\varepsilon(\widehat{A}_\varepsilon, \widehat{\beta}_\varepsilon) \le \mathcal{R}_\varepsilon(A_*, \beta_*) + \frac{\lambda_\varepsilon}{2}[\mathcal{W}(\theta_*) + \|\beta_*\|^2] + \eta_\varepsilon \longrightarrow 0.$$

Uniform coercivity bounds the coefficients; local uniform risk convergence forces every cluster point to the unique zero (A_*, β_*) of the limiting risk. □

The same coordinate transformation preserves the chosen regularizer at positive strength:

$$\widetilde{J}_\varepsilon^{\mathrm{ren}}(A, \beta) = \mathcal{R}_\varepsilon(A, \beta) + \frac{\lambda_\varepsilon}{2}[\Psi_m(A) + \|\beta\|^2]. \tag{B.4}$$

Proposition B.3 (A prescribed nonzero regularizer survives refinement)**.** *If $\lambda_\varepsilon > 0$ and $\lambda_\varepsilon \to \lambda_0 \in [0, \infty)$, the corrected reduced objectives converge variationally to*

$$\mathcal{J}_{\lambda_0}^{\mathrm{ren}}(A, \beta) = \mathcal{R}(A, \beta) + \frac{\lambda_0}{2}[\Psi_m(A) + \|\beta\|^2]. \tag{B.5}$$

Minimum values converge; global and vanishing-gap minimizing coefficients are bounded, and every cluster point minimizes this objective. At $\lambda_0 = 0$ its unique minimizer is (A_*, β_*); positive strength permits nonuniqueness.

Proof. Uniform risk coercivity and comparison with zero give compactness. For $c_\varepsilon \to c$, local uniform risk convergence and lower semicontinuity of Ψ_m give the limiting lower bound when $\lambda_0 > 0$; nonnegativity suffices when $\lambda_0 = 0$. At fixed c, finiteness of $\Psi_m(A)$ gives convergence to the displayed objective. The lower-bound and recovery argument in Theorem 4.1 proves convergence of minima and the cluster-point assertion. Positivity of α, γ gives uniqueness at zero strength. □

Proposition B.4 (Leaving the counterterm channel unpenalized)**.** *Global and vanishing-gap minimizers of*

$$J_\varepsilon^{\mathrm{free}}(\theta, b) = \mathcal{R}_\varepsilon(A_\theta, b + t_\varepsilon L A_\theta) + \frac{\lambda_\varepsilon}{2}\mathcal{W}(\theta)$$

are consistent for every target whenever $\lambda_\varepsilon > 0$ tends to zero.

Proof. The target representation gives

$$\mathcal{R}_\varepsilon(\widehat{A}_\varepsilon, \widehat{\beta}_\varepsilon) \le \mathcal{R}_\varepsilon(A_*, \beta_*) + \frac{\lambda_\varepsilon}{2}\mathcal{W}(\theta_*) + \eta_\varepsilon \longrightarrow 0.$$

At fixed cutoff, the positive-definite quadratic risk bounds represented coefficients on objective sublevels, while the penalty bounds θ. Thus $b = \beta - t_\varepsilon L A_\theta$ is bounded; continuity gives a minimum. Uniform coercivity and identifiability of the limiting risk prove coefficient consistency. □

B.3 Joint sampling and cutoff limits

Let $\varepsilon_k \downarrow 0$, $\lambda_k > 0$, and $n_k \to \infty$. At stage k, observe n_k independent pairs $(X_{\varepsilon_k}^{(r)}, Y^{(r)})$ from copies of the common field and its fixed target $Y = F(A_*, \beta_*)$. Different stages need not be independent. Replace the population risk by

$$\widehat{\mathcal{R}}_k(c) = \frac{1}{2n_k}\sum_{r=1}^{n_k} \|F_{\varepsilon_k}^{(r)}(c) - Y^{(r)}\|_{\mathsf{H}}^2, \tag{B.6}$$

leaving the penalty unchanged. Minimizers are measurable choices of neural parameters; their nonnegative gaps may be random.

Theorem B.5 (Empirical minimizers inherit the structural limit)**.** *Fix d, m, A_*, β_* as in Theorem 4.1, and let $\lambda_k \to 0$. For the bare objective, if $\lambda_k t_{\varepsilon_k}^2 \to \tau \in [0, \infty]$, empirical global minimizers and empirical near-minimizers whose gaps tend to zero in probability satisfy*

$$\widehat{A}_k \xrightarrow{\mathbb{P}} A_\tau, \qquad \widehat{\beta}_k \xrightarrow{\mathbb{P}} \beta_*.$$

Their conditional test risk on an independent field converges in probability to (21)*. For the corrected objective, the corresponding coefficients converge in probability to (A_*, β_*) and the conditional test risk tends to zero, for every vanishing penalty schedule. No further rate relation between n_k and ε_k is required.*

Choose an orthonormal basis $e_1, \dots, e_q$ of V, $q = D + d$, with its first D vectors in $\mathcal{A}$, and set $\Phi_{\varepsilon,\ell} = F_\varepsilon(e_\ell)$. Then

$$\widehat{\mathcal{R}}_k(c) = \tfrac{1}{2} c^\top \widehat{\Sigma}_k c - \widehat{r}_k^\top c + \tfrac{1}{2}\widehat{s}_k, \tag{B.7}$$

where

$$(\widehat{\Sigma}_k)_{\ell j} = \frac{1}{n_k}\sum_r \langle \Phi_{\varepsilon_k,\ell}^{(r)}, \Phi_{\varepsilon_k,j}^{(r)} \rangle_{\mathsf{H}}, \qquad (\widehat{r}_k)_\ell = \frac{1}{n_k}\sum_r \langle \Phi_{\varepsilon_k,\ell}^{(r)}, Y^{(r)} \rangle_{\mathsf{H}},$$
$$\widehat{s}_k = \frac{1}{n_k}\sum_r \|Y^{(r)}\|_{\mathsf{H}}^2. \tag{B.8}$$

Their population values, in these coordinates, are

$$\Sigma_\varepsilon = \mathrm{diag}(\alpha_{\varepsilon,\varepsilon} I_D, \gamma_{\varepsilon,\varepsilon} I_d), \quad r_\varepsilon = (\alpha_{\varepsilon,0} A_*, \gamma_{\varepsilon,0}\beta_*), \quad s = \mathbb{E}\|Y\|_{\mathsf{H}}^2. \tag{B.9}$$

Lemma B.6 (A cutoff-uniform quadratic law of large numbers)**.** *For K independent of k,*

$$\mathbb{E}\|\widehat{\Sigma}_k - \Sigma_{\varepsilon_k}\|_{\mathrm{F}}^2 + \mathbb{E}\|\widehat{r}_k - r_{\varepsilon_k}\|^2 + \mathbb{E}|\widehat{s}_k - s|^2 \le K/n_k. \tag{B.10}$$

Consequently $\widehat{\mathcal{R}}_k \to \mathcal{R}$ locally uniformly in probability, and its leading matrix has a fixed positive lower eigenvalue bound with probability tending to one.

Proof. Lemma A.3 and Cauchy–Schwarz give

$$\mathbb{E}|\langle \Phi_{\varepsilon,\ell}, \Phi_{\varepsilon,j}\rangle_{\mathsf{H}}|^2 \le (\mathbb{E}\|\Phi_{\varepsilon,\ell}\|_{\mathsf{H}}^4)^{1/2}(\mathbb{E}\|\Phi_{\varepsilon,j}\|_{\mathsf{H}}^4)^{1/2} \le K.$$

The same estimate bounds target cross-products and $\|Y\|_{\mathsf{H}}^2$ in L^2. Within-stage independence divides each variance by n_k; summing the fixed number of entries proves (B.10). The population coefficients converge to $\Sigma = \mathrm{diag}(\alpha I_D, \gamma I_d)$, $r = (\alpha A_*, \gamma\beta_*)$, and s. Thus empirical coefficients converge in probability, giving local uniform quadratic convergence. Eigenvalue continuity and $\min(\alpha, \gamma) > 0$ give the final assertion. □

Proof of Theorem B.5. The empirical risk depends only on represented coefficients, so Lemma A.5 transfers gaps unchanged. Coercivity in (θ, b) for the bare objective and (θ, β) for the corrected objective ensures empirical minima exist. From any stage subsequence, extract a further subsequence on which the quadratic coefficients and gaps converge almost surely, using Lemma B.6 and the gap assumption. Fix such an outcome. Eventually the leading matrix has a uniform positive lower eigenvalue, and the other coefficients are bounded; hence $\widehat{\mathcal{R}}_k(c) \ge a\|c\|_V^2 - B$. Comparison with zero bounds the minimum by $\widehat{s}_k/2$, so represented near-minimizers are bounded. The risks converge uniformly on bounded sets. For the bare penalty, the finite- and infinite-strength calculations of Lemma B.1 apply unchanged, including constant recovery sequences. Compactness and uniqueness in Theorem 4.1 give $\widehat{c}_k \to (A_\tau, \beta_*)$. For the corrected penalty, Proposition B.3 at zero strength gives $\widehat{c}_k \to (A_*, \beta_*)$. Every subsequence thus has a further subsequence with the asserted almost sure convergence, proving convergence in probability for the measurable choices. Independence of the test field identifies conditional test risk as $\mathcal{R}_{\varepsilon_k}(\widehat{c}_k)$; local uniform population risk convergence gives its stated limit. The cutoff-uniform bound (B.10) uses only $n_k \to \infty$, so no cutoff/sample-rate coupling enters the argument. □

C Null extensions and finite-cutoff certificates

C.1 Class equality and a neural witness

The ansatz $a = A$, $b = 0$, $c = q/C$ solves $r_3 = 1$, $r_1 = 0$ uniquely at $A = -1$, $q = 1/5$. It gives (C), with the actual neural realization and squared-weight cost

$$(u_3, v_3, u_5, v_5, b) = (-1, 1, (5C)^{-1}, 1, 0), \qquad P = 3 + (25C^2)^{-1}, \qquad |Cc| = 1/5 < 1.$$

By (35) and Lemma A.1, the features $F_{p,\varepsilon}$ converge in $L^2(\Omega; \mathsf{H})$ for $p = 1, 3, 5$. Write $F_{p,0}$ for these limits. Consequently, for every fixed capacity $M < \infty$,

$$\sup_{|C_\varepsilon c| \le M} \mathbb{E}\|cF_{5,\varepsilon}\|^2 \le \frac{M^2}{C_\varepsilon^2} \sup_\delta D_{5,\delta} \longrightarrow 0. \tag{36}$$

For a cutoff-indexed family $\mathcal{H}_\varepsilon$, define

$$\mathcal{H}_{\mathrm{cont}} = \left\{ f \in L^2(\Omega; \mathsf{H}) : \exists f_\varepsilon \in \mathcal{H}_\varepsilon,\ \|f_\varepsilon - f\|_{L^2(\Omega;\mathsf{H})} \xrightarrow[\varepsilon \downarrow 0]{} 0 \right\}. \tag{37}$$

Orthogonality bounds r_3, r_1 on bounded-risk sequences, since $A_{3,\varepsilon} \to A_0 > 0$ and $B_{1,\varepsilon} \to B_{1,0} > 0$. Coefficient compactness and (36) put every continuum limit in $\{F(r, s) : r, s \in \mathbb{R}\}$. Conversely, $c = 0$, $a = r$, $b = s - 3C_\varepsilon r$ realizes every such limit in both architectures, including $M = 0$. Thus their complete continuum classes agree, without a weight budget.

C.2 Global outcomes, exact controls, and the low-cost interval

Proof of Theorem 6.1 and the control equivalences. At fixed cutoff, the ordinary objective is continuous and coercive in the actual weights on the closed constraint set, so it attains its minimum. Independent minimizing factorizations (32) give exactly $\Psi_3(a) + \Psi_5(c) + b^2$. The witness bounds $\mathcal{R}_\varepsilon \le J_\varepsilon \le U_\varepsilon + \eta_\varepsilon$, where

$$U_\varepsilon = \frac{E_{3,\varepsilon}}{2} + \frac{D_{5,\varepsilon}}{50C_\varepsilon^2} + \frac{\lambda_\varepsilon}{2}\left(3 + \frac{1}{25C_\varepsilon^2}\right) \to 0, \tag{C.1}$$

for every objective gap $\eta_\varepsilon \to 0$. Orthogonality and positive limiting feature norms give $r_3 \to 1$, $r_1 \to 0$. When $\lambda_\varepsilon = 1/C_\varepsilon$, $b^2/(2C_\varepsilon) \le U_\varepsilon + \eta_\varepsilon$ gives $b/C_\varepsilon \to 0$, and

$$Cc = \frac{r_3}{5} - \frac{r_1 - b}{15C}, \qquad a = -r_3 + \frac{2(r_1 - b)}{3C} \tag{C.2}$$

gives the remaining limits. For the original model, comparison with zero bounds $b = O(\sqrt{C})$ and $r_1 = O(1)$, hence $a = (r_1 - b)/(3C) \to 0$. Its third-chaos risk tends to $A_0/2$; the upper bound $\mathcal{R}_\varepsilon \le A_0/2 + \eta_\varepsilon$ forces $r_1 \to 0$.

Let $J_\varepsilon^{(3)}(a, b)$ denote the original cubic objective after minimizing its neuron factorization at fixed (a, b). The Wick-quintic placebo has reduced objective $J_\varepsilon^{(3)}(a, b) + [D_{5,\varepsilon}c^2 + \lambda_\varepsilon \Psi_5(c)]/2$. The added term vanishes uniquely at $c = 0$, proving equality of its entire optimal-predictor set with the original set. For full transport, use

$$P_\varepsilon^{\mathrm{ren}} = \Psi_3(r_3) + \Psi_5(r_5) + r_1^2, \tag{40}$$

the least cost of independent Wick neurons, with inverse map

$$c = r_5, \qquad a = r_3 - 10Cr_5, \qquad b = r_1 - 3Cr_3 + 15C^2 r_5. \tag{C.3}$$

The corrected objective separates as

$$\tfrac{1}{2}[A_{3,\varepsilon} r_3^2 - 2K_{3,\varepsilon} r_3 + A_0 + \lambda_\varepsilon \Psi_3(r_3)] + \tfrac{1}{2}(B_{1,\varepsilon} + \lambda_\varepsilon) r_1^2 + \tfrac{1}{2}[D_{5,\varepsilon} r_5^2 + \lambda_\varepsilon \Psi_5(r_5)].$$

Its positive quadratic part is coercive on $|Cr_5| \le 1$; continuity and attained neuron factorizations give existence. Both final terms vanish uniquely at zero. The remaining cubic objective is identical in both corrected architectures, including any nonuniqueness. The target witness has objective at most $E_{3,\varepsilon}/2 + \lambda_\varepsilon \to 0$, proving consistency also for vanishing gaps. □

Proof of Proposition 6.2. Write $r = r_3$, $s = r_1$, $q = Cc$. The exact reduced objective is

$$J_\varepsilon(r, s, q) = \tfrac{1}{2}[A_{3,\varepsilon} r^2 - 2a_* K_{3,\varepsilon} r + a_*^2 A_0 + B_{1,\varepsilon} s^2 + D_{5,\varepsilon} q^2/C^2] + \frac{\Psi_3(r - 10q) + \Psi_5(q/C) + [s - 3C(r - 5q)]^2}{2C}, \qquad |q| \le M.$$

Bounded objective bounds r, s, q and b^2/C, so $r - 5q = (s - b)/(3C) \to 0$. Every limiting cubic therefore lies in $[-5M, 5M]$. Conversely, for every $|r| \le 5M$ and $s \in \mathbb{R}$, choose

$$(u_3, v_3, u_5, v_5, b) = (-r, 1, r/(5C), 1, s), \qquad (r_3, r_1, r_5) = (r, s, r/(5C)),$$
$$\frac{P}{2C} = \frac{r^2 + 2 + r^2/(25C^2) + s^2}{2C} \to 0.$$

This recovery construction includes both endpoints and $M = 0$. Nonnegativity of the penalty, the moment limits, and the uniform fifth-chaos bound give the matching lower and recovery limits after profiling over q. Let ι_I be zero on the interval I and $+\infty$ outside it. The limit is

$$J_0^M(r, s) = \frac{A_0}{2}(r - a_*)^2 + \frac{B_{1,0}}{2} s^2 + \iota_{[-5M, 5M]}(r).$$

Outside the interval, the bare-linear penalty diverges. Bounded sublevels are sequentially compact in (r, s, q). Comparing any minimizing subsequence with the recovery sequence for the unique minimizer $(s_M, 0)$ proves minimum-value convergence and convergence of every vanishing-gap minimizer. Then $q \to s_M/5$, $a = r - 10q \to -s_M$, and the exact risk converges to $A_0(a_* - s_M)^2/2$. Class equality follows from Appendix C.1. □

C.3 Capacity and deterministic input normalization

Proposition C.1 (Unconstrained and standardized erasure–recovery). *For $a_* \in \mathbb{R}$ and $\delta \in \{0, \frac{1}{2}\}$, set $Y_{a_*} = a_* F_{3,0}$, $Z_{\varepsilon,\delta} = C_\varepsilon^{-\delta} X_\varepsilon$, and*

$$N_\varepsilon^{(3,\delta)} = S[u_3 (v_3 Z_{\varepsilon,\delta})^3 + b Z_{\varepsilon,\delta}],$$
$$N_\varepsilon^{(5,\delta)} = N_\varepsilon^{(3,\delta)} + S[u_5 (v_5 Z_{\varepsilon,\delta})^5].$$

Minimize $\frac{1}{2}\mathbb{E}\|N - Y_{a_*}\|_{\mathsf{H}}^2 + P/(2C_\varepsilon)$ *over each architecture's parameters, with* P *their squared sum. Every global or vanishing-objective-gap minimizing sequence satisfies*

$$N_\varepsilon^{(3,\delta)} \to 0, \qquad N_\varepsilon^{(5,\delta)} \to Y_{a_*} \quad \text{in } L^2(\Omega;\mathsf{H}).$$

For the augmented model, the physical and Wick coefficients are

$$a = C^{-3\delta}u_3v_3^3, \quad c = C^{-5\delta}u_5v_5^5, \quad \ell = C^{-\delta}b,$$
$$r_3 = a + 10Cc, \qquad r_1 = \ell + 3Ca + 15C^2c, \qquad r_5 = c,$$
$$(r_3, r_1, Cc, a) \longrightarrow (a_*, 0, a_*/5, -a_*).$$

For any $M > |a_*|/5$*, all unconstrained augmented global optima eventually satisfy* $|Cc| < M$*. At every sufficiently fine cutoff their optimal-predictor set equals that obtained under* $|Cc| \le M$*.*

Proof. Continuity and coercivity in actual weights give existence. Expanding $Z_{\varepsilon,\delta}$ gives the displayed maps and $b^2 = C^{2\delta}\ell^2$. The physical witness $(a, c, \ell) = (-a_*, a_*/(5C), 0)$ has output $a_*F_{3,\varepsilon} + a_*F_{5,\varepsilon}/(5C)$ and least cost

$$P_{\mathrm{wit}}^\delta = \Psi_3(-a_*C^{3\delta}) + \Psi_5((a_*/5)C^{5\delta-1}),$$
$$U_\varepsilon^\delta = \frac{a_*^2E_{3,\varepsilon}}{2} + \frac{a_*^2D_{5,\varepsilon}}{50C^2} + \frac{P_{\mathrm{wit}}^\delta}{2C} \to 0.$$

Indeed (32) gives bounded cost for $\delta = 0$ and $O(C^{3/4}) + O(C^{1/2})$ for $\delta = 1/2$. Witness comparison proves augmented recovery; orthogonality gives $r_3 \to a_*$ and $r_1 \to 0$.

Zero-predictor comparison gives $J_\varepsilon \le a_*^2A_0/2 + \eta_\varepsilon$, hence $b = O(\sqrt{C})$ in either architecture. In the cubic model, bounded risk bounds a, r_1, and $a = (r_1 - C^{-\delta}b)/(3C) \to 0$. Moreover,

$$\mathcal{R}_\varepsilon = \tfrac{1}{2}\mathbb{E}\|aF_{3,\varepsilon} - a_*F_{3,0}\|^2 + \tfrac{1}{2}B_{1,\varepsilon}r_1^2 \le a_*^2A_0/2 + \eta_\varepsilon.$$

The first term tends to $a_*^2A_0/2$, so $r_1 \to 0$ and the output vanishes. For augmented near-minimizers, (C.2) with b replaced by $C^{-\delta}b$ recovers $Cc \to a_*/5$ and $a \to -a_*$. Thus their fifth-chaos output vanishes without a capacity constraint. If eventual strict feasibility failed, a sequence of violating global optima would contradict this coefficient limit. All unconstrained optima are therefore feasible; the two minimum values and entire optimal-predictor sets agree. □

For $a_* = 1$, $M = 1$ is eventually inactive in both conventions. $\delta = 1/2$ uses the known variance, rather than sample-dependent normalization. Removing the bound does enlarge the complete continuum class: for $\kappa \ne 0$, choose $c = \kappa$, $a = -10C\kappa$, $\ell = 15C^2\kappa$, $b = C^\delta\ell$. These realizable coefficients give $\kappa F_{5,\varepsilon} \to \kappa F_{5,0} \ne 0$, orthogonal to the cubic-plus-linear class by (35).

Remark C.2 (Common recovery sequences and actual realization cost). If both architectures admit actual-parameter recovery sequences $\theta_{\varepsilon,*}$ with vanishing risk and $\lambda_\varepsilon P_\varepsilon(\theta_{\varepsilon,*}) \to 0$, their vanishing-gap minimizers satisfy

$$0 \le \mathcal{R}_\varepsilon(\theta_\varepsilon) \le \mathcal{R}_\varepsilon(\theta_{\varepsilon,*}) + \tfrac{\lambda_\varepsilon}{2}P_\varepsilon(\theta_{\varepsilon,*}) + \eta_\varepsilon \longrightarrow 0.$$

Uniformly bounded recovery costs suffice when $\lambda_\varepsilon \to 0$. Erasure–recovery separation therefore requires different penalized realization costs.

C.4 A FINITE-CUTOFF CERTIFICATE

Write $A = A_{3,\varepsilon}$, $B = B_{1,\varepsilon}$, $K = K_{3,\varepsilon}$, $C = C_\varepsilon$, and $\lambda = \lambda_\varepsilon$. Profiling the original objective over its bare linear weight gives

$$b(a) = -\frac{3CBa}{B+\lambda}, \qquad j(a) = \frac{A_0}{2} + \frac{A+h}{2}a^2 - Ka + \frac{\lambda}{2}\Psi_3(a), \quad h = \frac{9\lambda C^2B}{B+\lambda}. \tag{C.4}$$

Since $K > 0$, every minimizer is nonnegative. At a positive minimizer, $j'(a) = 0$ and $\Psi_3'(a) > 0$ imply $a \le K/(A+h)$; this also holds at zero. Denote the original cubic risk by $\mathcal{R}_\varepsilon^{(3)}$. Then $K^2 \le A_0A$ and the third-chaos reverse triangle inequality give

$$\mathcal{R}_\varepsilon^{(3)} \ge \frac{A_0}{2}\left(\frac{h}{A+h}\right)^2. \tag{C.5}$$

This applies to every global minimizer, without convexity or uniqueness.

To bound the moments, put $s = \varepsilon^2$, $v_k = (1+|k|^2)^{-1}$, $h_k = e^{-2|k|^2}$, $T(\boldsymbol{k}) = \sum_j |k_j|^2$, and $w_p(\boldsymbol{k}) = h_{\sum_j k_j} \prod_j v_{k_j}$. Fourier expansion of (35) yields

$$M_p(s) = p! \sum_{\boldsymbol{k} \in (\mathbb{Z}^2)^p} w_p(\boldsymbol{k}) e^{-sT(\boldsymbol{k})}, \qquad (A_0, A, K, B, D_{5,\varepsilon}) = (M_3(0), M_3(s), M_3(s/2), M_1(s), M_5(s)). \tag{C.6}$$

Positive weights give

$$E_{3,\varepsilon} = 6 \sum_{\boldsymbol{k}} w_3(\boldsymbol{k}) (1 - e^{-sT(\boldsymbol{k})/2})^2 \le A_0 - A \le A_0. \tag{C.7}$$

Square shells contain $8n$ points at $|k|_\infty = n$. Integral comparison therefore gives, for $r > 1$,

$$S_r := \sum_k (1+|k|^2)^{-r} \le 1 + 8 \sum_{n \ge 1} n^{1-2r} \le 9 + \frac{4}{r-1}. \tag{C.8}$$

Also $H = \sum_k h_k < (1 + 2/6)^2 < 2$, since $n^2 \ge n$ and $e^2 > 7$. Write v^{*p} for p-fold lattice convolution and $\|\cdot\|_r$ for the $\ell^r(\mathbb{Z}^2)$ norm. Young's inequality gives $M_p(s) \le p! H \|v^{*p}\|_\infty \le 2p! S_{p/(p-1)}^{p-1}$ for $p = 3, 5$. Together with the zero tuples,

$$6 \le A_0, A \le 3468, \qquad B \ge 1, \qquad D_{5,\varepsilon} \le 93\,750\,000. \tag{C.9}$$

Positive truncation and monotone convergence justify the Fourier sums. For $v_k^{(s)} = e^{-s|k|^2} v_k$, the inequality $1 - e^{-x} \le x^{1/6}$ gives $\|v - v^{(s)}\|_{3/2} \le s^{1/6} 25^{2/3}$ and $\|v^{(s)}\|_{3/2} \le \|v\|_{3/2} \le 17^{2/3}$. Telescoping the triple convolution then yields

$$E_{3,\varepsilon} \le A_0 - A \le 36\, 25^{2/3} 17^{4/3} s^{1/6} < 14256\, \varepsilon^{1/3}, \tag{C.10}$$

where $25^2 < 9^3$ and $17^4 < 44^3$ verify the rounded constant.

The variance is continuous, strictly decreasing, and divergent at zero. For $s \le 1$, the same shells and integral comparison give

$$\begin{aligned} C_\varepsilon &\le 1 + 8 \sum_{n \ge 1} \frac{e^{-sn^2}}{n} \le 9 + 4 \int_s^\infty \frac{e^{-u}}{u}\, du \\ &\le 9 + 4 \log(1/s) + 4/e < 11 + 8 \log(1/\varepsilon). \end{aligned} \tag{C.11}$$

Every $C > 11$ thus specifies a unique cutoff below one. Using $e^{-x} \le 6!/x^6$ gives the rational envelope

$$E_{3,\varepsilon} \le E_{3,U}(C) := \min \left\{ 3468, 14256\, 6! \left(\frac{24}{C - 11} \right)^6 \right\}. \tag{C.12}$$

For $\lambda = 1/C$, $h \ge 9C^2/(C+1)$. Define $Q(C) = [9C^2/(9C^2 + 3468(C+1))]^2$. The sharp baseline and witness bounds are

$$\begin{aligned} \mathcal{R}_\varepsilon^{(3)} &\ge L_{\mathrm{sharp}}(C) := 3Q(C), \\ U(C) &= \frac{E_{3,U}(C)}{2} + \frac{93\,750\,000}{50C^2} + \frac{1}{2C} \left(3 + \frac{1}{25C^2} \right). \end{aligned} \tag{C.13}$$

For $\rho = 2\mathcal{R}/A_0$, exact control equivalence and $A_0 \ge 6$ imply

$$\rho_{\mathrm{original}}, \rho_{\mathrm{placebo}} \ge Q(C), \qquad \rho_{\mathrm{raw}} \le \min\{1, U(C)/3\}. \tag{C.14}$$

The substitution and outward rounding are checkable by integer cross-multiplication:

$$\begin{aligned} Q(10^4) &= \left(\frac{900000000}{934683468} \right)^2 > \frac{9271}{10^4}, \\ \frac{U(10^4)}{3} &= \frac{14256\, 6!}{6} \left(\frac{24}{9989} \right)^6 + \frac{1}{160} + \frac{1}{20000} + \frac{1}{150 \cdot 10^{12}} < \frac{6301}{10^6}. \\ C = 10^4 \quad &\Longrightarrow \quad \varepsilon < \exp[-(10000 - 11)/8] = e^{-1248.625}. \end{aligned} \tag{C.15}$$

At $C = 10^4$, both original and placebo satisfy $\rho \ge 0.9271$, and the raw augmentation satisfies $\rho \le 0.006301$; the variance bound gives $\varepsilon < e^{-1248.625}$. These are bounds for every global minimizer, not fitted losses. The original and placebo have identical optimal-risk sets. The release contains the exact arithmetic records.

D General realization-cost and depth laws

D.1 Vanishing penalized recovery cost

Proof of Theorem 6.3. Write $f_\varepsilon = F_\varepsilon(\theta_\varepsilon)$ and let $\eta_\varepsilon \to 0$ bound the objective gap. The zero realization gives

$$J_\varepsilon(\theta_\varepsilon) \le \tfrac{1}{2}\|Y\|^2 + \eta_\varepsilon,$$

so the risk and $\lambda_\varepsilon P_\varepsilon(\theta_\varepsilon)$ are bounded. A recovery sequence with vanishing penalized cost has $J_\varepsilon(\theta_\varepsilon^*) \to 0$. Comparison makes every near-minimizer consistent; in particular some near-minimizing sequence is consistent, since the nonnegative finite infimum always admits arbitrarily close approximants.

Conversely, suppose $f_\varepsilon \to Y$. Comparison with its admissible scaled realization gives, for fixed $0 < t < 1$,

$$\frac{\lambda_\varepsilon}{2}(1-t^a)P_\varepsilon(\theta_\varepsilon) \le \tfrac{1}{2}\|tf_\varepsilon - Y\|^2 - \tfrac{1}{2}\|f_\varepsilon - Y\|^2 + \eta_\varepsilon.$$

Thus

$$\limsup_{\varepsilon\downarrow 0} \lambda_\varepsilon P_\varepsilon(\theta_\varepsilon) \le \frac{(1-t)^2}{1-t^a}\|Y\|^2 \longrightarrow 0 \quad (t \uparrow 1).$$

The consistent near-minimizer is itself the required recovery sequence.

Under the additional cost and correlation hypotheses, $\lambda_\varepsilon s_\varepsilon \to 0$ gives sufficiency by the stipulated recovery sequence. Consistency implies $\langle f_\varepsilon, Y\rangle \to \|Y\|^2 > 0$; the correlation barrier forces $P_\varepsilon(\theta_\varepsilon) \ge c s_\varepsilon$ eventually. The proved necessity of vanishing penalized cost then gives $\lambda_\varepsilon s_\varepsilon \to 0$, including schedules without a limiting product.

If $\lambda_\varepsilon s_\varepsilon \to \infty$, zero comparison gives $P_\varepsilon = O(\lambda_\varepsilon^{-1})$. The barrier makes

$$|\langle f_\varepsilon, Y\rangle| \le B s_\varepsilon^{-\nu} + B'(\lambda_\varepsilon s_\varepsilon)^{-\nu} \to 0.$$

Expanding the same comparison and dropping the nonnegative penalty gives $\frac{1}{2}\|f_\varepsilon\|^2 \le \langle f_\varepsilon, Y\rangle + \eta_\varepsilon \to 0$. □

For a degree-p polynomial neuron, scaling both factors by $t^{1/(p+1)}$ scales its output by t and its cost by $t^{2/(p+1)}$. Scaling the linear skip by t completes the contraction. A common positive exponent works for finitely many degrees, including $a = 1/3$ for the cubic–quintic architecture. The output-scaling assumption permits comparison within the same admissible architecture.

D.2 Explicit contraction and null-extension limits

Proposition D.1 (Reduced null-extension selection)**.** *Let $z = (x, \beta)$ range over a fixed finite-dimensional space and Q be a nonempty compact convex set. Suppose the exact reduced neural infimum and its near-minimizers transfer to*

$$J_\varepsilon(z,q) = \tfrac{1}{2}\|A_\varepsilon z + E_\varepsilon q - Y\|^2 + \tfrac{\lambda_\varepsilon}{2}\left[h_\varepsilon(z,q) + \|\beta - t_\varepsilon(Lx - Bq)\|^2\right], \quad q \in Q.$$

Here the displayed maps are linear, $A_\varepsilon \to A$ and $E_\varepsilon \to 0$ in operator norm, A is injective, $t_\varepsilon \to \infty$, $\lambda_\varepsilon > 0$ tends to zero, and $h_\varepsilon \ge 0$ satisfies $\lambda_\varepsilon h_\varepsilon \to 0$ uniformly on bounded sets. If $\lambda_\varepsilon t_\varepsilon^2 \to \tau < \infty$, the limiting reduced objective is

$$J_\tau(z) = \tfrac{1}{2}\|Az - Y\|^2 + \tfrac{\tau}{2}\operatorname{dist}(Lx, BQ)^2.$$

If $\tau = \infty$ and $\{z : Lx \in BQ\}$ is nonempty, it is the same risk restricted to $Lx \in BQ$. Infimum values converge, and every near-minimizing z converges to the unique limiting minimizer.

Proof. Injectivity of A gives a uniform lower bound on $\|A_\varepsilon z\|$ for small ε. Compactness of Q and $E_\varepsilon \to 0$ then bound z on bounded objective sublevels. Along a subsequence, $z_\varepsilon \to z$ and $q_\varepsilon \to q \in Q$. For finite τ,

$$\lambda_\varepsilon\|\beta_\varepsilon - t_\varepsilon(Lx_\varepsilon - Bq_\varepsilon)\|^2 \longrightarrow \tau\|Lx - Bq\|^2.$$

Nonnegativity of the hidden cost proves the lower bound. A fixed pair (z, q) has vanishing penalized hidden cost and gives the matching recovery bound. Minimizing over compact Q gives the displayed distance.

For infinite τ, bounded objective requires $Lx = Bq$ in the limit. For any feasible pair, the fixed sequence has skip penalty $\lambda_\varepsilon \|\beta\|^2 \to 0$ and vanishing hidden charge. Feasibility provides a bounded comparison sequence. These lower and upper bounds prove convergence of infima and near-minimizers. The risk is strictly convex in z, and the added distance penalty or constraint is convex, giving uniqueness. □

With $Q = \{0\}$, $E_\varepsilon = 0$, $z = (A, \beta)$ and the reduced hidden cost Ψ_m, this gives Theorem 4.1's contraction penalty and harmonic constraint. For the scalar quintic extension set $z = (r, s)$, $q = C_\varepsilon c$, $Q = [-M, M]$,

$$A_\varepsilon z = rF_{3,\varepsilon} + sF_{1,\varepsilon}, \quad E_\varepsilon q = (q/C_\varepsilon)F_{5,\varepsilon}, \quad Lx = r, \quad Bq = 5q, \quad t_\varepsilon = 3C_\varepsilon,$$
$$h_\varepsilon = \Psi_3(r - 10q) + \Psi_5(q/C_\varepsilon).$$

The fifth-chaos moment bound gives $E_\varepsilon \to 0$. At $\lambda_\varepsilon = C_\varepsilon^{-1}$ the limiting constraint is $r \in [-5M, 5M]$, exactly Proposition 6.2. The bounded-cost witness for Theorem 6.1, and the unconstrained and standardized recovery sequences in Appendix C.3, satisfy Theorem 6.3. Equality of the corrected finite-cutoff optimal predictor sets uses the separable orthogonal control objective in Appendix C.2.

D.3 The sharp ReLU depth law

Proof of Theorem 6.4. Prediction classes. Every finite scalar-input affine/ReLU network is continuous piecewise affine with finitely many knots, hence has a representation

$$f(x) = c + sx + \sum_j d_j (x - t_j)_+.$$

The identity $x = x_+ - (-x)_+$ represents its linear part in one hidden layer. Conversely, identity layers preserve the nonnegative outputs of any ReLU hidden layer and embed every shallow network into greater depth. Unions over finite widths therefore have equal prediction classes at each cutoff, and so do their images under S.

Width-independent barrier. Write $C = C_\varepsilon$ and remove all biases to obtain f_{θ^0}. Positive homogeneity in one scalar input makes this function linear on each half-line; its odd part is linear. Its third Hermite coefficient is therefore zero. The ReLU Lipschitz property gives

$$\sup_x |f_\theta(x) - f_{\theta^0}(x)| \le \sum_{\ell=1}^{L} \|b_\ell\| \prod_{j>\ell} \|W_j\|_{\mathrm{op}} \le T_L(P), \qquad T_L(P) = \sum_{k=1}^{L} (P/k)^{k/2}.$$

The final bound uses $\|W\|_{\mathrm{op}} \le \|W\|_{\mathrm{F}}$ and arithmetic–geometric mean on each product's k squared parameter norms. For probabilists' Hermite polynomials H_n and $Z \sim \gamma = N(0, 1)$, $\mathbb{E}H_3(Z)^2 = 6$, so

$$a_3 = \frac{\mathbb{E}[f_\theta(\sqrt{C}Z)H_3(Z)]}{6C^{3/2}}, \qquad |a_3| \le \frac{T_L(P)}{\sqrt{6}C^{3/2}} \le K_L \frac{1 + P^{L/2}}{C^{3/2}}.$$

Gaussian-chaos orthogonality gives $\langle Sf_\theta(X_\varepsilon), Y\rangle = a_3 \langle F_{3,\varepsilon}, Y\rangle$. The last factor tends to $\|Y\|^2 > 0$. Recovery forces $a_3 \to 1$, hence $P \ge c_L C^{3/L}$ eventually. No width enters the bound.

A four-channel recovery function. Let ϕ, Φ be the standard Gaussian density and CDF, and set

$$p = 2[1 - \Phi(1)], \quad a = \phi(1)/3, \quad g(z) = (z - 1)_+ - (-z - 1)_+, \quad \psi(z) = \frac{g(z) - pz}{a}.$$

Gaussian integration by parts gives $\mathbb{E}[g(Z)Z] = p$ and $\mathbb{E}[g(Z)H_3(Z)] = 2\phi(1)$. Thus $\mathbb{E}[\psi H_1] = 0$ and $\mathbb{E}[\psi H_3]/3! = 1$. Oddness and linear growth imply the $L^2(\gamma)$ expansion

$$\psi = H_3 + \sum_{\substack{n \ge 5 \\ n \text{ odd}}} a_n H_n, \qquad \sum_n n! a_n^2 = \|\psi\|^2_{L^2(\gamma)} < \infty.$$

For $f_\varepsilon(x) = C_\varepsilon^{3/2}\psi(x/\sqrt{C_\varepsilon})$, its third-chaos observation is exactly $F_{3,\varepsilon}$. Write $R_\varepsilon = Sf_\varepsilon(X_\varepsilon) - F_{3,\varepsilon}$ for the remaining observed component. The heat-kernel representation in Appendix A gives $0 < G_\varepsilon := G_{\varepsilon,\varepsilon} \le C_\varepsilon$. With $h = p_2$ and $\rho_\varepsilon = G_\varepsilon/C_\varepsilon$, it obeys

$$\begin{aligned}\|R_\varepsilon\|^2_{L^2(\Omega;\mathsf{H})} &= C_\varepsilon^3 \sum_{\substack{n\ge 5\\ n\ \text{odd}}} n! a_n^2 \int h\rho_\varepsilon^n \, d\mu \\ &\le C_\varepsilon^{-2}\|\psi\|^2_{L^2(\gamma)} \int h G_\varepsilon^5 \, d\mu \longrightarrow 0.\end{aligned}$$

Nonnegative terms justify the sum–integral interchange; Lemma A.1 uniformly bounds the last integral. Together with $F_{3,\varepsilon} \to Y$, this proves recovery including the infinite higher-chaos tail.

The width-four realization has first weight rows $(1,-1,1,-1)/\sqrt{C}$, first biases $(-1,-1,0,0)$, intermediate identity matrices, and final row $(1,-1,-p,p)/a$. Multiply every weight and the first bias by $s = C^{3/(2L)}$, leaving all other biases zero. Its output is f_ε and its exact cost is

$$P = C^{3/L}\left[4/C + 2 + 4(L-2) + \|(1,-1,-p,p)/a\|^2\right] \le K_L C^{3/L}$$

for sufficiently small cutoff.

Selection. Scaling W_j by $t^{1/L}$ and b_j by $t^{j/L}$ scales each layer-j output by $t^{j/L}$, so it scales the final output by t and the penalty by at most $t^{2/L}$. Apply Theorem 6.3 with $s_\varepsilon = C_\varepsilon^{3/L}$ and $\nu = L/2$. This proves the sharp threshold, erasure when its product diverges, and the three stated depth regimes at $\lambda_\varepsilon = \kappa/C_\varepsilon$. □

E Polynomial population and training verification

Numerically resolved optima show selection on all four grids. All twenty Adam–Newton endpoints at 129^2 pass the independent empirical-gap and population-coefficient criteria.

E.1 Fixed target, covariances, and spatial moments

The Fourier coefficients on $\mathbb{T}^2$ satisfy $X_{-k} = \overline{X_k}$ and $\mathbb{E}|X_k|^2 = (1+|k|^2)^{-1}$. For $N \in \{33, 65, 129, 257\}$, write $\sum_N$ for the sum over $\|k\|_\infty \le K_N$, and set

$$K_N = (N-1)/2, \quad \varepsilon_N = 8/N, \quad U_N = P_{K_N}X, \quad X_N = e^{\varepsilon_N^2\Delta_x/2}U_N, \quad S = e^{\Delta_x}.$$

Here P_K retains Fourier modes $\|k\|_\infty \le K$. The input covariance, cross-target covariance, and sharp variance are distinct:

$$\begin{aligned} g_N(x) &= \sum_N \frac{e^{-\varepsilon_N^2|k|^2}e^{ik\cdot x}}{1+|k|^2}, \qquad C_N = g_N(0), \\ g_{N,0}(x) &= \sum_N \frac{e^{-\varepsilon_N^2|k|^2/2}e^{ik\cdot x}}{1+|k|^2}, \quad C_N^{\mathrm{sharp}} = \sum_N \frac{1}{1+|k|^2}. \end{aligned}$$

For the fixed target $Y = S{:}X^3{:}$, the labels are $Y_N = S[U_N^3 - 3C_N^{\mathrm{sharp}}U_N]$. To prove population equivalence, truncate X at $L > K_N$ and write $P_L X = U_N + V_{N,L}$ with independent Gaussian summands. Conditional Gaussian moments give $\mathbb{E}[{:}(U_N+V_{N,L})^3{:} \mid U_N] = {:}U_N^3{:}_{C_N^{\mathrm{sharp}}}$. The smoothed Wick cubic converges as $L \to \infty$ in $L^2(\Omega;\mathsf{H})$; conditional expectation is an L^2 contraction and commutes with S. Nonzero retained heat multipliers give $\sigma(U_N) = \sigma(X_N)$. Hence, for every square-integrable predictor measurable from X_N,

$$Y_N = \mathbb{E}[Y \mid X_N], \qquad \mathbb{E}\|f - Y\|^2 = \mathbb{E}\|f - Y_N\|^2 + \mathbb{E}\|Y - Y_N\|^2.$$

Thus the population minimizers agree. Finite-sample estimators need not agree, and Y_N differs from ${:}X_N^3{:}_{C_N}$.

Let h be the kernel of $S^*S = e^{2\Delta_x}$ and $F_{p,N} = S{:}X_N^p{:}_{C_N}$. Gaussian pairings give, in feature order $(F_{3,N}, F_{1,N}, F_{5,N})$,

$$A_N = 6\int hg_N^3, \quad B_N = \int hg_N, \quad D_N = 120\int hg_N^5, \quad T_N = 6\int hg_{N,0}^3,$$
$$G_N = \operatorname{diag}(A_N, B_N, D_N), \qquad c_N = (T_N, 0, 0), \qquad Z = \mathbb{E}\|Y_{129}\|^2 = 221.91452147549396.$$

All integrals use normalized measure. Float64 FFT convolution, padded to the smallest power of two at least $10K_N + 1$, resolves the fifth-degree spectrum; coefficients are weighted by $e^{-2|k|^2}$. Doubling padding checks evaluation, and doubling support at fixed ε_N checks truncation. The fixed Z uses the sharp cubic covariance. Define

$$r_{0,N} = T_N/A_N, \qquad m_N = r_3/r_{0,N}, \qquad E_N(r) = \frac{A_N(r_3 - r_{0,N})^2 + B_N r_1^2 + D_N r_5^2}{T_N^2/A_N}.$$

Here E_N is excess above the best represented unregularized predictor; the zero predictor has $E_N = 1$.

E.2 True neural objectives and global population searches

Put $a = u_3 v_3^3$ and $c = u_5 v_5^5$. The raw network is $S[aX_N^3 + cX_N^5 + bX_N]$; its corrected counterpart uses Wick neurons and linear weight β. Cubic models omit both quintic factors. The vector θ collects all trainable factors and the linear weight. Its penalty is $P = \|\theta\|^2$, and both augmented models enforce $|C_N c| \le 1$. Their represented coefficients and exact reduced penalties are

$$\begin{array}{lll} \text{raw:} & r = (a + 10C_N c,\ b + 3C_N a + 15C_N^2 c,\ c), & \Psi_3(a) + \Psi_5(c) + b^2, \\ \text{corrected:} & r = (a, \beta, c), & \Psi_3(a) + \Psi_5(c) + \beta^2, \end{array}$$

where Ψ_p is the exact neuron cost (32). The normalized objective is $\mathbb{E}\|f - Y_N\|^2/(2Z) + \bar\lambda_N P/2$, with $\bar\lambda_N = 10^{-3} C_{129}/C_N$ and $\lambda_N = Z\bar\lambda_N$. For raw augmentation, set $r = r_3$, $q = C_N r_5$ and profile the linear coordinate:

$$H_N = \frac{9C_N^2 \lambda_N B_N}{B_N + \lambda_N}, \qquad r_1 = \frac{3C_N\lambda_N}{B_N + \lambda_N}(r - 5q),$$
$$\frac{A_N}{2}(r - r_{0,N})^2 + \frac{D_N q^2}{2C_N^2} + \frac{H_N}{2}(r - 5q)^2 + \frac{\lambda_N}{2}\big[\Psi_3(r - 10q) + \Psi_5(q/C_N)\big].$$

The last line is the unnormalized objective up to a constant. Comparison with zero confines minimizers to $[0, 2r_{0,N}] \times [-1, 1]$. Each subdivision cell has a lower bound from the exact minimum of its convex quadratic part plus lower bounds on the neural costs; sign-preserving cells also use the concave-power secants. Candidates include zero, $q = 0$, $r - 10q = 0$, and both constraint edges. Every cell whose lower bound can attain the incumbent remains active, retaining joint coefficient intervals.

For raw cubic, $r = z^2$ reduces positive stationary points to $(A_N + H_N)z^3 - T_N z + \lambda_N 3^{-3/4} = 0$; all positive roots and zero are compared. Corrected cubic sets $H_N = 0$. Independent two-variable separable searches check the Wick-quintic placebo and corrected augmented equivalences; CPU high-precision checks verify roots and objective values. The numerical gates are padding discrepancy $\le 10^{-8}$, support change in $m_N \le 10^{-3}$, objective gap divided by $Z \le 10^{-7}$, coefficient uncertainty $\le 10^{-3}$ or a retained minimizing interval, and control agreement $\le 10^{-8}$. Propagation includes measured support changes, 10^{-12} relative moment and 10^{-10} absolute profile guards; separations must exceed ten times their numerical uncertainty.

E.3 Registered optimization and independent verification

At $N = 129$, seeds $101, 202, 303, 404, 505$ each use 8192 training, 1024 validation, and 2048 independent test fields, paired across all four models. For seed s, $\mathbb{E}_s$ averages the training split. Write $\Phi = (F_{3,N}, F_{1,N}, F_{5,N})$. The empirical statistics are $G_{s,ij} = \mathbb{E}_s\langle\Phi_i, \Phi_j\rangle$, $c_{s,i} = \mathbb{E}_s\langle\Phi_i, Y_N\rangle$, and $y_s = \mathbb{E}_s\|Y_N\|^2$, retaining every off-diagonal Gram entry. The exact objective is

$$J_s(\theta) = \frac{r(\theta)^\top G_s r(\theta) - 2c_s^\top r(\theta) + y_s}{2Z} + \frac{10^{-3}}{2}\|\theta\|^2.$$

Each seed initializes the same predictor in every model: cubic coefficient uniform on $[-0.05, 0.05]$, hidden factors uniform on $[0.8, 1.2]$, and zero represented linear and quintic coefficients. Raw skip weights include counterterms; quintic output factors start at zero. Optimizers read only training statistics and starting weights; endpoints are hash-committed before the evaluator opens held-out data or optimum records.

Adam–Newton procedure. CUDA float64 Adam uses paired shuffled indices, batch size 128, betas $(0.9, 0.999)$, epsilon 10^{-8}, zero built-in weight decay, and gradient clipping at 10. Two cosine phases use $\eta_t = \eta_0[0.01 + 0.99(1 + \cos(\pi t/n))/2]$, $0 \le t < n$, with $(n, \eta_0) = (20000, 10^{-3})$ then $(40000, 3 \cdot 10^{-4})$; moments and counter restart for the second phase. Deterministic full-batch refinement then uses the exact gradient and symmetric Hessian, with $R = \partial r/\partial\theta$, $C = C_{129}$, $\bar\lambda = 10^{-3}$, and $q = Cu_5v_5^5$. For matrices, $\|\cdot\|_2$ is the operator norm and $\lambda_{\min}$ the smallest eigenvalue:

$$
\begin{aligned}
&g = R^\top(G_s r - c_s)/Z + \bar\lambda\theta, \qquad H = \big[R^\top G_s R + \textstyle\sum_j (G_s r - c_s)_j \nabla^2 r_j\big]/Z + \bar\lambda I,\\
&d_i = \sqrt{\max(|H_{ii}|, \bar\lambda, 10^{-12})}, \quad D = \operatorname{diag}(d_i), \quad H_s = D^{-1}HD^{-1}, \quad g_s = D^{-1}g,\\
&B = H_s + \max\{0, 10^{-8}\max(1, \|H_s\|_2) - \lambda_{\min}(H_s)\}I,\\
&z = \operatorname*{arg\,min}_{-1-q \le \nabla q^\top D^{-1} z \le 1-q} \{g_s^\top z + \tfrac{1}{2} z^\top B z\}, \quad p = D^{-1}z,\\
&p \leftarrow p/\max\{1, \max_i |p_i|/(1 + |\theta_i|)\}.
\end{aligned}
$$

Cubic models omit the constraint. After every Adam update and Newton trial, infeasible quintic factors are both scaled by $|q|^{-1/6}$, rounded toward zero if needed. Trials $\alpha = 2^{-j}$, $j = 0, \dots, 39$, retract $\theta + \alpha p$ to θ_{trial}. Set $d = \theta_{\text{trial}} - \theta$ and $\Delta r = r(\theta + d) - r(\theta)$, and accept the first with $g^\top d < 0$ and

$$
\Delta J = \frac{(G_s r - c_s)^\top \Delta r + \frac{1}{2}\Delta r^\top G_s \Delta r}{Z} + \bar\lambda(\theta^\top d + \tfrac{1}{2} d^\top d) \le 10^{-4} g^\top d.
$$

The uncapped projected-gradient direction p_I uses $D = B = I$ above. Forty rejections trigger this fallback with the same cap and search. At iteration start, $\|p_I\|_\infty \le 10^{-10}$ stops as stationary; other stops are 2000 iterations, both searches failing, numerical failure, or registered resource ceilings. Every terminal refinement is primary.

The release preserves the optimization-limited 20,000-update study and final Adam–Newton endpoints, with both registrations, all checkpoints, and every historical decision.

For seed s and architecture a, let $m_{s,a}$ be the endpoint's normalized cubic coefficient and m_a^{pop} its population optimum. The evaluator encloses each empirical minimum $J_{s,a}^*$ using full Grams and exact neural costs. With $I(v) = [v^-, v^+]$ an enclosure, require $g_{s,a}^+ = J_s(\theta)^+ - (J_{s,a}^*)^- \le 10^{-4}$ and $I(m_{s,a} - m_a^{\text{pop}}) \subset [-0.05, 0.05]$. Paired raw coefficient differences must each exceed 0.10, with mean minus twice sample SD at least 0.25, and every raw augmented test error must be lower. Both corrected means must be at least 0.90; population RMS disagreement pooled over the five corrected pairs and normalized by $\sqrt{T_N^2/A_N}$ must be at most 0.05.

E.4 Final accepted results

Every endpoint stops as stationary, and all registered criteria pass (Table E.1). Figure E.1 places the same endpoints as Figure 2 on the population curves across resolutions.

Model	Runs	Range of m	$\max g^+$	$\max \lvert m - m^{\mathrm{pop}} \rvert^+$
Raw cubic	5	$[0.308973, 0.318495]$	1.81×10^{-9}	0.0063443
Raw augmented	5	$[0.766546, 0.803346]$	3.46×10^{-9}	0.0224614
Corrected cubic	5	$[0.995309, 1.003202]$	4.03×10^{-9}	0.0042240
Corrected augmented	5	$[0.995309, 1.003202]$	4.03×10^{-9}	0.0043242

Table E.1: Final polynomial endpoints, all five seeds per model. Coefficient ranges include numerical enclosures. Maxima bound the normalized empirical gap and absolute population-coefficient deviation; all bounds round outward.

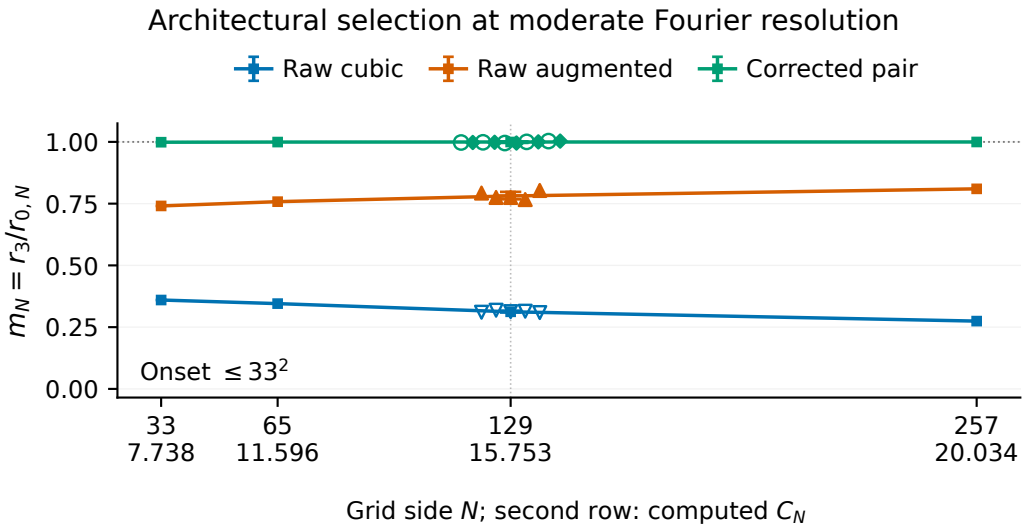


Figure E.1: **Architectural selection at moderate Fourier resolution.** Lines are the numerically resolved population optima; markers at 129^2 are the 20 Adam–Newton endpoints, with mean and one sample standard deviation across five paired seeds. Every empirical objective gap is below 4.03×10^{-9}, and every normalized cubic coefficient is within 0.0224614 of its population value. The registered tolerances are 10^{-4} and 0.05. The release retains the frozen 20,000-update endpoints and all stage-specific diagnostics.

F Standard-channel comparison and certification

F.1 Model and population objective

The standard component is NeuralOperator's `ChannelMLP` (NeuralOperator contributors, 2026): scalar input/output, width eight, ReLU, zero dropout, and $L \in \{2, 4\}$ pointwise affine layers. All weights and biases train. The pinned component is commit `00b7d86f8d74ff0af55da53eb585fe26df9c71f0`. Grids $N \in \{65, 257\}$ and seeds $101, 202, 303, 404, 505$ give twenty primary runs; the five depth-four 257^2 runs also have matched controls. Each equal-depth seed shares its initialization across grids. The physical target $Y = S{\cdot}X^3{\cdot}$, $S = e^{\Delta_x}$, Fourier supports, and moments C_N, A_N, T_N are those of Appendix E.1. The fixed constants are $Z_0 = Z = 221.91452147549396$ and $C_{129} = 15.75327195212808$, with $\bar{\lambda}_N = \lambda_N / Z_0 = 10^{-3} C_{129}/C_N$. For normalized Hermite polynomials $h_q = H_q/\sqrt{q!}$ and $Z \sim \gamma = N(0,1)$, define

$$f_\theta(\sqrt{C_N}z) = \sum_{q \geq 0} b_q h_q(z), \qquad v_{q,N} = \int h(g_N/C_N)^q \, d\mu,$$

$$r_3 = \frac{b_3}{\sqrt{6} C_N^{3/2}}, \qquad r_{0,N} = T_N/A_N, \qquad V_N = T_N^2/A_N.$$

Chaos orthogonality gives the excess objective J_N^{ex}; truncation at Hermite degree Q gives $\mathcal{J}_{Q,N}$, normalized by Z_0. The indicator $\mathbf{1}_{q=3}$ is one for $q=3$ and zero otherwise:

$$J_N^{\mathrm{ex}} = \frac{A_N}{2}(r_3 - r_{0,N})^2 + \frac{1}{2}\sum_{q\neq 3} v_{q,N} b_q^2 + \frac{\lambda_N}{2}P(\theta), \tag{F.1}$$

$$\mathcal{J}_{Q,N} = \frac{1}{2Z_0}\sum_{q=0}^{Q} v_{q,N}(b_q - b_q^*)^2 + \frac{\bar{\lambda}_N}{2}P(\theta), \qquad b_q^* = \mathbf{1}_{q=3}\sqrt{6}C_N^{3/2} r_{0,N}. \tag{F.2}$$

Here P is the sum of squared actual parameters; norms of random spatial outputs are in $L^2(\Omega;\mathsf{H})$. The fixed-target objective minus J_N^{ex} is $(\|Y\|^2 - V_N)/2$; the conditional-label objective minus it is $(\|Y_N\|^2 - V_N)/2$. Their difference is $\|Y - Y_N\|^2/2$. Thus the pointwise-feature projection $y_0 = r_{0,N}F_{3,N}$ and conditional mean Y_N have distinct roles.

F.2 Shallow comparison, exact functions, and numerical certificates

Unrestricted-width shallow optimum and same-function cost. Balancing a shallow output weight with its hidden affine pair gives the signed-measure formulation (Bach, 2017). Let $g_\omega = S(wX_N + b)_+$, with $\omega = (w,b)$ on the unit circle $\mathbb{S}^1$. Write μ for a finite signed measure on that circle, $\|\mu\|_{\mathrm{TV}}$ for its total variation, and β for the output bias. For a dual variable $v \in L^2(\Omega;\mathsf{H})$,

$$\inf_{\mu,\beta}\frac{1}{2}\left\|\int g_\omega\, d\mu + \beta - y_0\right\|^2 + \lambda_N\|\mu\|_{\mathrm{TV}} + \frac{\lambda_N}{2}\beta^2, \qquad D(v) = \langle y_0, v\rangle - \frac{1}{2}\|v\|^2 - \frac{\langle v,1\rangle^2}{2\lambda_N}.$$

Norm continuity of the atom map on the compact circle permits finite-atomic approximation without increasing total variation, so this infimum equals the finite-width union's infimum. Bounded variation and bias give a relaxed optimizer; strict convexity gives its unique prediction. Quadratic conjugacy makes $D(v)$ a lower bound whenever $\sup_\omega |\langle v, g_\omega\rangle| \le \lambda_N$. For $v = y_0$, its mean is zero, and Gaussian hinge integration, with $t = -b/(|w|\sqrt{C_N})$, gives

$$|\langle y_0, g_\omega\rangle| = \frac{T_N|w|\sqrt{C_N}|t|\phi(t)}{6C_N^{3/2}} \le \frac{T_N}{6C_N^{3/2}\sqrt{2\pi}}.$$

The equality applies for $w \neq 0$; at $w = 0$ the correlation is zero. The bound follows from $|w|\sqrt{C_N}|t| = |b| \le 1$. Outward enclosures give $0.298203 < 0.301477 \le \lambda_{65}$ and $0.159474 < 0.174497 \le \lambda_{257}$. Hence zero attains the dual value $V_N/2$ and is the unique optimal shallow prediction.

Every saved network has the exact all-input representation $f(x) = c + sx + \sum_j d_j(x - t_j)_+$: induction through the affine/ReLU layers splits each affine cell at its exact rational preactivation roots; continuity and slope jumps then give this hinge identity. Let $P_{\mathrm{shallow}}(f)$ be the infimum of squared-parameter cost over shallow realizations of f. It satisfies

$$P_{\mathrm{shallow}}(f) \ge 2\sum_j |d_j|\sqrt{1+t_j^2}. \tag{F.3}$$

Indeed, a neuron at $t = -b/w$ contributes jump $a|w|$, while $a^2 + w^2 + b^2 \ge 2|aw|\sqrt{1+t^2}$. Summing at shared knots uses the triangle inequality; affine and output-bias costs are nonnegative. Disjoint certified knot brackets bound net jumps from below; omitted knots can only add variation. Divide this lower bound by the enclosed actual deep penalty's upper endpoint.

Activation cells and outward integration. Exact dyadic parameters and rational root comparisons establish the cells on the whole real line. Isolated coincident roots have Gaussian measure zero. An identically zero preactivation on an interval is first-order irrelevant only when an exact downstream path test, opening zero gates and closing negative gates, finds no nonzero path to the output; otherwise its derivative remains unresolved. On every activation cell, including boundaries where output slope jumps cancel, differentiate slopes and intercepts in the actual parameters. Gaussian integration gives b_q and $u_{iq} = \partial_{\theta_i} b_q$; continuity cancels moving-boundary terms for b_q. Let $\sigma = \sqrt{C_N}$, choose the

registered binary64 reference σ_0, and set $F_0(z) = f_\theta(\sigma_0 z)$. With standard Gaussian density/CDF ϕ, Φ and $\bar\Phi = 1 - \Phi$, the reference-scale coefficients of $F_0(z) = c + sz + \sum_j \delta_j (z - \kappa_j)_+$ are

$$b_0 = c + \sum_j \delta_j [\phi(\kappa_j) - \kappa_j \bar\Phi(\kappa_j)], \quad b_1 = s + \sum_j \delta_j \bar\Phi(\kappa_j), \quad b_q = \sum_j \frac{\delta_j \phi(\kappa_j) h_{q-2}(\kappa_j)}{\sqrt{q(q-1)}} \ (q \ge 2).$$

Physical-input uncertainty is enclosed below.

Spatial moments. On the retained Fourier support, let $p_k = e^{-\varepsilon_N^2 |k|^2}/[C_N(1+|k|^2)]$ and $\mathcal{M}_N(t) = \sum_k p_k e^{tk_1}$. Then $v_{q,N} = \sum_k e^{-2|k|^2} (p^{*q})_k$, where p^{*q} is the q-fold convolution. Let $\widetilde v_{q,N}$ use circular convolution on an $M \times M$ grid and retain heat modes $\|k\|_\infty \le H = 12$. An alias entering these modes requires coordinate displacement at least $M - H$. Chernoff's inequality for either sign and a union bound over two coordinates, using square symmetry, give

$$\delta_a = \min_{j=-8,\dots,8} 4 e^{-t_j (M-H)} \mathcal{M}_N(t_j)^q, \qquad t_j = 2^j/[(N-1)/2],$$

$$\widetilde v_{q,N} - \delta_a \le v_{q,N} \le \widetilde v_{q,N} + e^{-2(H+1)^2}.$$

The last term bounds omitted heat-weighted probability mass. Arithmetic uncertainty enters outward; the first $M \in \{2048, 4096, 8192\}$ enclosing all degrees through 512 with alias error at most 10^{-14} is used.

Full objective and population-gradient enclosures. Outward arithmetic includes conversion, accumulation, and elementary-function remainders (NIST, 2026). Write $f = f_\theta$, let $\mathrm{Lip}(f)$ be its global Lipschitz constant, and use $\|\cdot\|_2$ for the $L^2(\gamma)$ norm. Set $L_z = \sigma_0 \,\mathrm{Lip}(f)$ and $U_i^0(z) = \partial_{\theta_i} f(\sigma_0 z)$, with $A_i, B_i \ge 0$ satisfying $|U_i^0(z)| \le A_i |z| + B_i$. Let F_{clip} extend F_0 constantly outside $[-B, B]$, $B = 16$. The omitted output, input-derivative, and parameter-derivative energies obey

$$\|F_0 - F_{\mathrm{clip}}\|_2^2 \le 4 L_z^2 \phi(B)/B^3, \qquad \mathbb{E}[(F_0'(Z))^2 \mathbf{1}_{|Z|>B}] \le 2 L_z^2 \bar\Phi(B),$$

$$\mathbb{E}[U_i^0(Z)^2 \mathbf{1}_{|Z|>B}] \le 2\{A_i^2 [B\phi(B) + \bar\Phi(B)] + 2 A_i B_i \phi(B) + B_i^2 \bar\Phi(B)\},$$

These bounds follow by Gaussian tail integration and are restored in the full-function enclosures. Physical-scale uncertainty is included through $\|f(\sigma Z) - f(\sigma_0 Z)\|_2 \le \mathrm{Lip}(f)|\sigma - \sigma_0|$. For possibly discontinuous $U_i^x = \partial_{\theta_i} f$, put $u_{iq}(\sigma) = \mathbb{E}[U_i^x(\sigma Z) h_q(Z)]$. Differentiating the Gaussian density gives

$$\frac{d}{d\sigma} \mathbb{E}[U_i^x(\sigma Z) h_q(Z)] = \frac{q\, u_{iq}(\sigma) + \sqrt{(q+1)(q+2)}\, u_{i,q+2}(\sigma)}{\sigma}.$$

Cauchy–Schwarz and the affine envelope enclose this perturbation over the physical variance interval; the same density argument bounds the derivative square-norm perturbation.

For the exact physical output $F_\theta(z) = f_\theta(\sqrt{C_N} z)$, set

$$U_i = \partial_{\theta_i} F_\theta, \quad V_Q = \mathbb{E} F_\theta^2 - \sum_{q=0}^{Q} b_q^2, \quad D_Q = \mathbb{E}(F_\theta')^2 - \sum_{q=1}^{Q} q b_q^2, \quad V_{i,Q} = \mathbb{E} U_i^2 - \sum_{q=0}^{Q} u_{iq}^2.$$

The Gaussian Sobolev identity follows from $h_q' = \sqrt{q}\, h_{q-1}$ and Parseval (Janson, 1997): for continuous piecewise-affine F_θ, $\mathbb{E}(F_\theta')^2 = \sum_{q \ge 1} q b_q^2$. Thus $\sum_{q>Q} b_q^2 \le \min\{V_Q, D_Q/(Q+1)\}$. For odd Q, positivity of h, $|g_N/C_N| \le 1$, and nonnegative observed-chaos norms imply $0 \le v_{q,N} \le v_{Q+1,N}$ for $q > Q$: the even power $Q+1$ bounds $|(g_N/C_N)^q|$ even where $g_N < 0$. Consequently,

$$\sum_{q>Q} v_{q,N} b_q^2 \le R_Q := v_{Q+1,N} \min\{V_Q, D_Q/(Q+1)\}. \tag{F.4}$$

The full objective is enclosed by the finite-degree objective plus $[0, R_Q/(2Z_0)]$. Parameter derivatives may jump, but ordinary $L^2(\gamma)$ Parseval gives $V_{i,Q} = \sum_{q>Q} u_{iq}^2$. Weighted Cauchy–Schwarz therefore proves

$$\left| Z_0^{-1} \sum_{q>Q} v_{q,N} b_q u_{iq} \right| \le Z_0^{-1} \sqrt{\Big(\sum_{q>Q} v_{q,N} b_q^2\Big)\Big(\sum_{q>Q} v_{q,N} u_{iq}^2\Big)} \le \frac{\sqrt{R_Q v_{Q+1,N} V_{i,Q}}}{Z_0}.$$

No Sobolev regularity of U_i is needed. Residual intervals are intersected with $[0,\infty)$; a negative upper endpoint invalidates the calculation. If $G_{i,Q}$ encloses the finite-degree gradient, including covariance, physical-scale, Gaussian clipping, and arithmetic terms, and upper selects an interval's upper endpoint, population stationarity is certified precisely when the exact derivative test resolves and

$$\max_i \left\{ \sup_{g\in G_{i,Q}} |g| + \mathrm{upper}\left(\frac{\sqrt{R_Q v_{Q+1,N} V_{i,Q}}}{Z_0} \right) \right\} \le 10^{-6}. \tag{F.5}$$

F.3 Registered optimization procedure

The registered procedure trains the same width-eight ChannelMLP at both depths and resolutions, forms the nonlinear response before introducing the final weight penalty, and refines the resulting parameters under the unchanged population objective. All primary runs use their own archived initialization; the matched control applies the final penalty from the first update.

Hidden weights are Kaiming-normal, output weights have variance $1/8$, first biases are uniform on $[-1,1]$, and remaining biases are zero. Each run uses $80{,}000$ full-batch binary64 Adam updates, betas $(0.9, 0.999)$, epsilon 10^{-8}, per-run Euclidean gradient cap ten, persistent optimizer state, and explicit penalty $\bar\lambda_N w(t) P/2$ without built-in decay. A $1{,}024$-node Gaussian rule evaluates degrees 0–31. For update index t, learning rate η_t, and $c(u) = (1-\cos(\pi u))/2$, the completed schedule is

$$w(t) = \begin{cases} 0 & t < 20000, \\ c((t-20000)/20000) & 20000 \le t < 40000, \\ 1 & 40000 \le t < 80000, \end{cases}$$

$$\eta_t = \begin{cases} 10^{-3} & t < 40000, \\ 10^{-3}[1 - 0.99c((t-40000)/40000)] & 40000 \le t < 80000. \end{cases}$$

Controls use $w \equiv 1$ and the same learning rates. From the terminal Adam checkpoint, the analytic Gauss–Newton refinement uses $(Q, \text{outer budget}) = (63,500), (127,500), (255,2000)$. Write $b = (b_0,\dots,b_Q)^\top$, $b^* = (b_0^*,\dots,b_Q^*)^\top$, and $\mathcal{J} = \mathcal{J}_{Q,N}$. With $\mathsf{R}_{qi} = u_{iq}$, $W = \mathrm{diag}(v_{q,N})/Z_0$, and damping $\mu > 0$, its direction and acceptance are

$$g = \mathsf{R}^\top W(b-b^*) + \bar\lambda_N \theta, \quad H = \mathsf{R}^\top W \mathsf{R} + \bar\lambda_N I, \quad (D^{-1}HD^{-1} + \mu I) z = -D^{-1} g, \quad p = D^{-1} z,$$

$$D_{ii} = \sqrt{\max\{H_{ii}, \bar\lambda_N, 10^{-12}\}}, \quad \mathrm{pred} = -g^\top p - p^\top H p/2 > 0, \quad \frac{\mathcal{J}(\theta) - \mathcal{J}(\theta+p)}{\mathrm{pred}} \ge 0.1.$$

Rescale p to relative coordinate size at most one. Each degree starts at $\mu = 10^{-3}$ and stops at $\|g\|_\infty \le 10^{-9}$, its outer budget, or twenty unsuccessful trials; its executable registration fixes the damping updates. These are the retained shallow endpoints.

Final deep and control endpoints start from their own $80{,}000$-update checkpoints with function-preserving positive scales $s_j = e^{x_j}$:

$$(W_1', b_1') = (s_1 W_1, s_1 b_1), \qquad (W_4', b_4') = (W_4/s_3, b_4),$$
$$(W_j', b_j') = ((s_j/s_{j-1}) W_j, s_j b_j) \quad (j = 2, 3).$$

ReLU homogeneity preserves the function; the resulting convex sum-of-exponentials penalty is balanced by damped Newton to $\|\nabla_x P\|_\infty \le 10^{-12}$ before each BFGS stage. Actual-parameter inverse BFGS uses degrees/accepted-step budgets $(255, 4000), (511, 8000)$, a relative step cap 0.5, and Armijo constant 10^{-4}. Three consecutive dual-direction line-search stalls terminate a stage. Degree 255 advances at finite-gradient bound 2×10^{-7}. Degree 511 certifies on crossing 5×10^{-7}, meeting 2×10^{-7}, and eligible multiples of 250 accepted steps; a full bound 2.5×10^{-7} allows early stopping. Every terminal state receives the degree-511 certificate.

F.4 Final endpoints and registered outcome

All ten primary deep endpoints meet the registered recovery, objective-advantage, same-function-cost, and numerical-precision criteria at both grids. Their objective advantages are at least $0.297958 Z_0$, and every exact trained function has a shallow realization-cost lower bound at least 23 times its actual

deep penalty. The certified population-gradient upper bounds are at most 8.09×10^{-6}; first-order stationarity is not certified at the registered 10^{-6} gate, and the preregistered joint outcome remains UNRESOLVED.

In the initial registered study, the default optimizer reached the near-zero stationary point in four of five deep initializations at each grid; seed 101 reached a nonzero lower-cost solution. The release preserves that study and all subsequent registered endpoints.

Table F.1 aggregates every run in each group. Figure F.1 links cubic recovery (left) to the cost of realizing the same functions (right). Here $m = r_3/r_{0,N}$, objective advantages and enclosure widths are divided by Z_0, noncubic energy is divided by V_N, and the gradient norm is $\|\nabla_\theta(J_N^{\mathrm{ex}}/Z_0)\|_\infty$ in actual parameters. The registered gates require $m \geq 0.75$, cubic advantage ≥ 0.20, objective advantage $\geq 0.02Z_0$, noncubic fraction ≤ 0.10, cost ratio ≥ 2, objective width $\leq 10^{-5}Z_0$, cubic width $\leq 10^{-4}$, identity checks, and effects exceeding ten times their propagated uncertainties. All ten shallow endpoints have certified stationarity and normalized global-optimum gap $\leq 10^{-4}$. Controls do not determine the joint primary outcome.

Table F.1: Final standard-channel endpoints, five runs per group. Ranges enclose all coefficients; advantages and cost ratios use minima, and errors, widths, and gradient bounds use maxima.

Quantity	Primary 65^2	Primary 257^2	Control 257^2
Cubic coefficient m	$[0.869158, 0.875513]$	$[0.908311, 0.911279]$	$[-2.11 \times 10^{-49}, 0.910669]$
Objective advantage, minimum	0.297958	0.373047	-6.50×10^{-10}
Same-function cost ratio, minimum	23.443	35.340	0.000
Noncubic energy fraction, maximum	0.046936	0.035097	0.032022
Objective width, maximum	1.18×10^{-7}	1.41×10^{-7}	6.81×10^{-8}
Population-gradient bound, maximum	8.09×10^{-6}	7.30×10^{-6}	3.27×10^{-6}
Every criterion except stationarity	5/5	5/5	1/5
Certified population stationarity	0/5	0/5	4/5

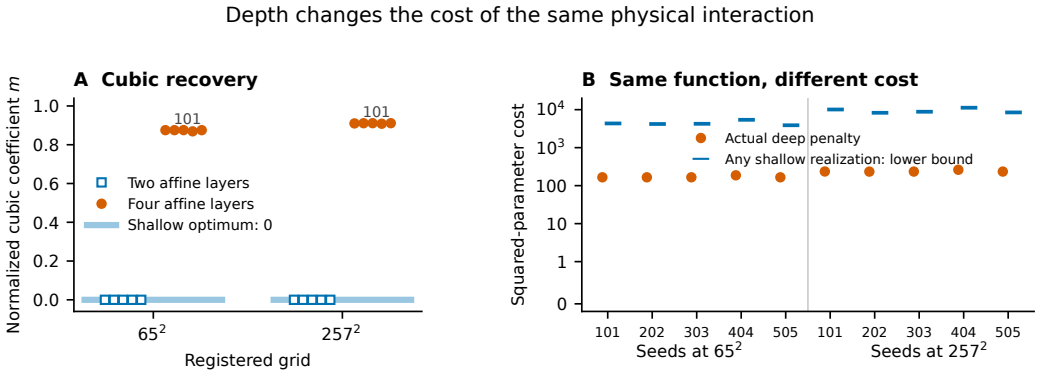


Figure F.1: **Realization-cost separation in standard channel networks.** All primary deep runs reach the registered recovery and objective-advantage criteria at both grids. The right panel compares each network's actual penalty with a certified lower bound for any shallow representation of exactly the same function. Numerical precision is certified for every primary endpoint; the largest certified population-gradient upper bound is 8.09×10^{-6}, above the registered 10^{-6} certification gate.

G CLOSEST RELATED WORK

Martini et al. (2019) characterize coding dependence of penalized polynomial interactions and identify translation-invariant formulations. Lassas & Siltanen (2004) establish that an intended prior effect can change under refinement; Lassas et al. (2009) develop discretization-invariant constructions. Parhi & Nowak (2022; 2026) connect neural parameter penalties with compositional function spaces, while Ongie & Parhi (2026) develop a general representation-cost framework. Rudner et al. (2023) express preferences over predictive functions through function-space empirical Bayes. Kovachki et al. (2023) formulate operator learning between function spaces; Gao et al. (2025) analyze cross-resolution representation mismatch. DLR-Net (Gong et al., 2023) and the Wick–Hermite operator construction (Shi et al., 2026) incorporate singular-field structure. The present results concern a different object: the learning limit when physical realization costs diverge while parameter regularization vanishes. Theorems 6.3–6.4 characterize that selection through penalized recovery cost and establish its sharp depth dependence for standard pointwise ReLU channel networks.