CHECKABLE NTK POSITIVITY AND FINITE-WIDTH GRADIENT DESCENT FOR SCALAR- AND VECTOR-VALUED PINNS WITH STRONG-FORM, WEAK-FORM, AND NONLOCAL LINEAR CONSTRAINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

We give checkable positive-definiteness criteria for the limiting neural tangent kernel (NTK) and high-probability finite-width gradient-descent guarantees for scalar- or vector-valued physics-informed neural networks (PINNs) with linear constraints. The constraints may be strong-form differential rows of any fixed finite order, including coupled systems, with any linear initial or boundary conditions; weak-form residual and boundary functionals; or finite-measure nonlocal observations such as integral, nonlocal-diffusion, and Dirac-data rows, all in any dimension. For each class, positive definiteness of the limiting NTK is equivalent to a rank condition on a finite coefficient, functional, or moment matrix of the fixed design: a certificate computed before training that detects structural zero modes. The model hypotheses are those of an ordinary two-layer PINN, a smooth nonpolynomial activation with bounded symmetric initialization, and are met by standard choices such as $\tanh$ with uniform initialization. Given a certificate, explicit width and step-size conditions ensure, with high probability, that the empirical NTK retains at least half the limiting gap and that full-batch gradient descent decreases the training loss geometrically at every iteration. Controlled experiments check the certificates and the finite-width mechanism.

1 INTRODUCTION

Physics-informed neural networks incorporate sampled differential equations and boundary data into a training loss (Dissanayake & Phan-Thien, 1994; Lagaris et al., 1998; 2000; Raissi et al., 2019). The solution error comprises approximation, discretization or generalization, and optimization components. Universal approximation supports expressivity (Hornik et al., 1989), and PINN error and partial differential equation (PDE) stability analyses address approximation and discretization under problem-dependent assumptions (De Ryck et al., 2024; Shin et al., 2020; Mishra & Molinaro, 2023), but the nonconvex training problem remains difficult (Wang et al., 2021; 2022).

When the width is large relative to the number of sampled constraints, NTK theory reduces this optimization to a spectral question: a strictly positive smallest tangent-kernel eigenvalue makes the residual dynamics of gradient descent (GD) contractive, provided the finite-width Gram matrix remains close to its initialization (Jacot et al., 2018; Lee et al., 2019; Du et al., 2019b). In a PINN this Gram matrix encodes sampled constraints rather than input examples: it is indexed by point–operator pairs for strong-form losses and by bounded linear functionals for weak-form or nonlocal losses; positivity must hold jointly across PDE, initial, and boundary constraints; and one physical site may carry several scalar equations. Postulating a spectral gap without verifying it therefore leaves a central premise of the convergence argument unresolved.

We replace the unverified qualitative nondegeneracy premise with exact structural criteria for finite collections of strong-form local differential, weak-form functional, and nonlocal linear constraints. The resulting gap is positive for each fixed nonredundant design, but its magnitude remains problem-

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.

dependent. Under the hypotheses stated below, our argument follows the chain

$$\text{nonredundancy} \iff \text{positive-definiteness certificate}$$
$$\implies \text{positive-definite finite-width kernel during GD} \implies \text{geometric loss decay.}$$

The exact verification in the first step is a *positive-definiteness certificate* (Definition 2.7). The final implication uses only such a certificate and local row regularity of the one-neuron prediction maps, so one finite-width GD argument covers strong-form differential rows of arbitrary fixed finite order together with admissible weak-form and nonlocal rows. Nonredundancy means independence of the co-located coefficient vectors, of the combined residual and boundary functionals, or of the loss-scaled signed-measure tuples, for the three classes respectively. Thus, for each fixed finite design in our setting, the limiting kernel is singular exactly when the sampled constraints have a structural dependency; compatible targets for such singular designs are treated by exact active-subspace compression in Section E.2. Our contributions are:

- **Checkable positive-definiteness certificates.** For strong-form rows of any fixed finite order, scalar or coupled, and for weak-form and finite-measure nonlocal rows, all in any dimension, we derive exact limiting-NTK nullspace and rank formulas through sitewise coefficient vectors, joint functional dependencies, or signed-measure dependencies. They cover repeated sites and reduce positive definiteness to the rank of a finite coefficient, functional, moment, or Dirac-atom matrix of the fixed design.
- **Finite-width GD for broad linear observations.** For an ordinary two-layer PINN, a smooth nonpolynomial activation with bounded symmetric initialization, a certificate plus explicit width and step-size conditions preserve at least half of the limiting spectral gap and give geometric empirical-loss decay under full-batch GD at every iteration, for differential rows of any fixed finite order and for admissible weak and nonlocal rows.
- **Experiments consistent with the theorems.** In six matched two-row constraint families and a 770-run study over eleven strong-form, weak-form, and nonlocal PDE losses, every certificate is fixed before training, measured ranks match the predicted nullspaces, at every tabulated width above 16 every run keeps at least half the reference gap at every checkpoint, and a separate per-step audit finds the loss nonincreasing at every scored iterate.

The appendix contains full proofs, the constructive certificate procedures with worked examples, a finite-witness lower bound on the smallest eigenvalue, the compatible-singular extension, and fallback criteria outside the biased analytic setting.

Related work Wang et al. (2022) derive the limiting PINN tangent kernel and identify loss imbalance. The closest finite-width, full-batch GD analysis of strong-form PINNs is that of Gao et al. (2023), for biased scalar cubic-ReLU networks with pointwise second-order constraints; our argument depends only on the induced linear observations, not their differential order. Operator-oriented analyses establish transformed-kernel limits while assuming a positive limiting Gram (Gan et al., 2025) or study admissible operators and random-feature training (Zhao & Luo, 2025); neither gives the joint exact nullspace used here. For positivity of the differential NTK (DNTK), Jin & Wu (2026a) prove the cross-location feature independence used in our pointwise argument, under distinct physical points and scalar operators acting componentwise; we identify the full nullspace for repeated sites, coupled outputs, joint residual–boundary systems, and dependent rows. Jin & Wu (2026b) bound both extreme eigenvalues for a bias-free two-layer rectified-power-unit (RePU) Poisson model with separated sites in dimension at least three; our finite-witness estimate applies more generally. Related optimization results treat projected GD (Nießen & Müller, 2026), stochastic wide-network training (Jin & Wu, 2027), or limits without a uniform spectral gap (Jiang et al., 2026). For weak-form PINNs, Kouma & Bomisso (2026) give a conditional, frequency-localized DNTK analysis; variational and nonlocal PINNs motivate our functional rows (Kharazmi et al., 2021; Pang et al., 2020), whose kernel treatment draws on collocation, operator observations, and measure embeddings (Schaback & Wendland, 2006; Pförtner et al., 2022; Sriperumbudur et al., 2011). Section C.1 compares assumptions and scope in detail.

2 SETTING

Let $\Omega \subset \mathbb{R}^{d_0}$ be the physical domain and write $[M] = \{1, \dots, M\}$ for any positive integer M. Let N denote the number of scalar constraint rows.

Definition 2.1 (Two-layer network)**.** For input dimension d_0, output dimension d_2, and width d_1, let

$$\begin{aligned} \mathbf{u}(\mathbf{x};\theta) &= \frac{1}{\sqrt{d_1}} \sum_{m=1}^{d_1} \sigma(\mathbf{w}_m^\top \mathbf{x} + \gamma b_m)\mathbf{a}_m, \\ \boldsymbol{\xi}_m &= (\mathbf{w}_m, b_m), \qquad \theta_m = (\boldsymbol{\xi}_m, \mathbf{a}_m), \\ \theta &= (\theta_1, \dots, \theta_{d_1}) \in \mathbb{R}^{d_1 d_\theta}, \end{aligned} \tag{1}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is the scalar activation, $\mathbf{w}_m \in \mathbb{R}^{d_0}$, $b_m \in \mathbb{R}$, $\mathbf{a}_m \in \mathbb{R}^{d_2}$, $d_\theta = d_0 + d_2 + 1$, and $\gamma \geq 0$. With the ridge functions $\varphi_{\boldsymbol{\xi},q} = \mathbf{e}_q \sigma(\mathbf{w}^\top \cdot + \gamma b)$ of a neuron $\boldsymbol{\xi} = (\mathbf{w}, b)$ in output $q \in [d_2]$, $\mathbf{u}(\cdot;\theta) = d_1^{-1/2} \sum_{m=1}^{d_1} \sum_{q=1}^{d_2} a_{m,q} \varphi_{\boldsymbol{\xi}_m,q}$.

Definition 2.2 (Sampled linear constraints)**.** After decomposing every vector-valued PDE, initial condition, and boundary condition into scalar rows, each constraint $i \in [N]$ is a linear observation of $\mathbf{u}(\cdot;\theta)$ with a loss scale $s_i > 0$, of one of three types.

(i) *Strong-form pointwise.* For $\mathbf{x}_i \in \overline{\Omega}$,

$$\begin{aligned} P_i(\theta) &= s_i(\mathcal{A}_i \mathbf{u})(\mathbf{x}_i;\theta), \qquad (\mathcal{A}_i \mathbf{u})(\mathbf{x}) = \sum_{q=1}^{d_2} (\mathcal{A}_{iq} u_q)(\mathbf{x}), \\ (\mathcal{A}_{iq} v)(\mathbf{x}) &= \sum_{\alpha \in I_{iq}} a_{iq\alpha}(\mathbf{x})\, \partial^\alpha v(\mathbf{x}), \end{aligned} \tag{2}$$

where every $I_{iq} \subset \mathbb{N}_0^{d_0}$ is finite and the coefficients are finite at $\mathbf{x}_i$; Dirichlet, Neumann, Robin, mixed, and variable-coefficient conditions all have this form.

(ii) *Weak-form functional.* $P_i(\theta) = \lambda_i(\mathbf{u}(\cdot;\theta))$ with $\lambda_i(f) = s_i a_i(f, v_i)$, where $\mathcal{X}$ is a trial space containing the candidate solutions, $\mathcal{V}_i$ is normed, $a_i : \mathcal{X} \times \mathcal{V}_i \to \mathbb{R}$ is a continuous bilinear map, and $v_i \in \mathcal{V}_i$ is a fixed test function.

(iii) *Finite-measure nonlocal.* $P_i(\theta) = \lambda_i(\mathbf{u}(\cdot;\theta))$ with $\lambda_i(f) = s_i \sum_{q=1}^{d_2} \int_D f_q \, d\nu_{iq}$ for a compact $D \subseteq \overline{\Omega}$ and finite signed regular Borel measures ν_{iq} on D.

The scale s_i carries the loss normalization, for example $s_i = N_r^{-1/2}$ for a group of N_r residual rows, and all three types enter the same loss and tangent Gram matrix.

Definition 2.3 (Loss and tangent Gram matrix)**.** Let $\mathbf{v}_P(\theta) = (P_i(\theta))_{i=1}^N$. If y_i denotes the unscaled target for row i, set $[\mathbf{v}_T]_i = s_i y_i$, so predictions and targets use the same row scaling. Define

$$\begin{aligned} \mathbf{r}(\theta) &= \mathbf{v}_T - \mathbf{v}_P(\theta), \quad \mathcal{L}(\theta) = \|\mathbf{r}(\theta)\|_2^2, \\ [\mathbf{J}(\theta)]_{i,:} &= \nabla_\theta P_i(\theta)^\top, \qquad i \in [N], \\ \mathbf{K}(\theta) &= \mathbf{J}(\theta)\mathbf{J}(\theta)^\top. \end{aligned} \tag{3}$$

Here $\mathbf{J}(\theta) \in \mathbb{R}^{N \times d_1 d_\theta}$. We write $\mathbf{K}_{(k)} = \mathbf{K}(\theta_{(k)})$.

Definition 2.4 (Gradient descent)**.** For a step size $\eta > 0$,

$$\theta_{(k+1)} = \theta_{(k)} - \eta \nabla_\theta \mathcal{L}(\theta_{(k)}).$$

We abbreviate $\mathbf{r}_{(k)} = \mathbf{r}(\theta_{(k)})$.

Assumption 2.5 (Initialization)**.** The vectors $\theta_{m,(0)}$ are i.i.d. from a width-independent, compactly supported law μ. The constraint design and target are fixed before initialization or sampled independently of it; in the latter case all probability statements are conditional on their realization. Let μ_h denote the $\boldsymbol{\xi}$-marginal of μ.

Assumption 2.6 (Classwise observation and feature conditions)**.** The fixed finite family of linear rows satisfies the clauses for one of the following classes. A subscript "reg" marks sufficient regularity for population-NTK existence and GD; "pd" marks the additional feature richness used for the exact positive-definiteness criterion.

($\mathsf{P}_{\mathrm{reg}}$) The rows are of type (i) in Definition 2.2, with largest order $r_{\max} < \infty$, and $\sigma \in C_{\mathrm{loc}}^{r_{\max}+1,1}(\mathbb{R})$.

(P_{pd}) In addition, σ is real analytic and nonpolynomial, $\gamma > 0$, and $\mathrm{int}(\mathrm{supp}\, \mu_h) \neq \varnothing$.

($\mathsf{W}_{\mathrm{reg}}$) The rows are of type (ii) in Definition 2.2, the trial space $\mathcal{X}$ is a real Banach space of $\mathbb{R}^{d_2}$-valued functions embedded continuously in $L^1_{\mathrm{loc}}(\Omega;\mathbb{R}^{d_2})$, so that $\mathcal{X}$-valued derivatives and limits may be computed pointwise, and every loss-scaled row λ_i belongs to $\mathcal{X}^*$. The one-neuron map $\mathcal{G}(\boldsymbol{\xi},\mathbf{a}) = \sum_{q=1}^{d_2} a_q \varphi_{\boldsymbol{\xi},q}$ is twice continuously Fréchet differentiable into $\mathcal{X}$ on an open neighborhood of every fixed parameter tube generated by $\operatorname{supp}\mu$.

(W_{pd}) In addition, $\operatorname{int}(\operatorname{supp}\mu_h) \neq \varnothing$, every $\boldsymbol{\xi} \mapsto \varphi_{\boldsymbol{\xi},q}$ is real analytic into $\mathcal{X}$, and

$$\overline{\operatorname{span}\{\varphi_{\boldsymbol{\xi},q} : \boldsymbol{\xi} \in \mathbb{R}^{d_0+1},\ q \in [d_2]\}}^{\mathcal{X}} = \mathcal{X}.$$

($\mathsf{N}_{\mathrm{reg}}$) The rows are of type (iii) in Definition 2.2, and $\sigma \in C^2_{\mathrm{loc}}(\mathbb{R})$.

(N_{pd}) In addition, σ is real analytic and nonpolynomial, $\gamma > 0$, and $\operatorname{int}(\operatorname{supp}\mu_h) \neq \varnothing$.

Each "pd" label denotes its displayed condition together with the corresponding "reg" clause; the bias-free case $\gamma = 0$ admitted by Definition 2.1 is used only in the appendix, chiefly for the counterexamples of Section D.1. Row independence is not assumed here: it is the design-side certificate below. Conditional output-weight centering is imposed only for GD.

These clauses model the deterministic empirical loss obtained after fixing a finite test or collocation family and any quadrature, including standard variational PINN (VPINN) and hp-VPINN formulations and forward nonlocal PINNs with fixed operator parameters. The analyticity and density clauses are sufficient conditions for the exact NTK characterization, not conditions normally checked in implementations. Table 1 summarizes the three classes. The only restriction is linearity of the rows, not their order, dimension, or form: any fixed finite-order linear operator, scalar or coupled, in any dimension, with any linear initial or boundary rows, is covered, as are fixed-test variational and finite-measure nonlocal observations, with all rows certified jointly.

The abstract GD theorem needs only the applicable "reg" clause and any positive-definiteness certificate; the "pd" clauses make that certificate equivalent to the checkable independence criterion in the table.

For every fixed linear row considered below, linearity and Definition 2.1 give $P_i(\theta) = d_1^{-1/2}\sum_{m=1}^{d_1} h_i(\theta_m)$ with $h_i(\boldsymbol{\xi},\mathbf{a}) = \mathbf{a}^\top \mathbf{f}_i(\boldsymbol{\xi})$, where $\mathbf{f}_i : \mathbb{R}^{d_0+1} \to \mathbb{R}^{d_2}$ applies row i to the ridge functions of one neuron: $[\mathbf{f}_i(\boldsymbol{\xi})]_q = s_i(\mathcal{A}_{iq}\sigma(\mathbf{w}^\top \cdot + \gamma b))(\mathbf{x}_i)$ for rows of type (i) and $[\mathbf{f}_i(\boldsymbol{\xi})]_q = \lambda_i(\varphi_{\boldsymbol{\xi},q})$ for types (ii) and (iii). Define the single-neuron Jacobian $\boldsymbol{\Psi}(z) \in \mathbb{R}^{N\times d_\theta}$

Table 1: The three observation classes, each a fixed finite family of linear rows under Assumption 2.5. In every class the "reg" clause gives existence of $\mathbf{K}^\infty_{(0)}$ and $\mathbf{K}_{(0)} \to \mathbf{K}^\infty_{(0)}$; the "pd" clause identifies $\ker \mathbf{K}^\infty_{(0)}$ with the listed relations, so row independence certifies $\mathbf{K}^\infty_{(0)} \succ 0$ through the output block; and that certificate, conditional centering, and Equation (9) give geometric decay of the empirical training loss, not of the PDE solution error (Corollary 4.3).

Class	Rows and representative PDEs	$\ker \mathbf{K}^\infty_{(0)}$ under "pd"	Result; certificate
Strong-form pointwise	Fixed finite-order linear differential rows, scalar or coupled: Poisson and variable-coefficient elliptic, transport, heat, wave, biharmonic, Stokes, linear elasticity, and linear initial or boundary rows	Sitewise relations among the loss-scaled coefficient vectors	Proposition 3.1; Section D.3.1
Weak-form functional	Fixed bounded residual and boundary functionals on one trial space: fixed-test variational elliptic or parabolic systems, tested traces, and the discontinuous-coefficient weak Poisson problem of Corollary C.19	Joint relations among the functionals in $\mathcal{X}^*$	Proposition 3.2; Section D.3.2
Finite-measure nonlocal	Finite signed-measure tuples on compact D: averages and moments, bounded-kernel integral and nonlocal-diffusion rows, Fredholm or Volterra rows, Dirac data, and atomic quadrature	Joint relations among the loss-scaled measure tuples	Proposition 3.3; Section D.3.3

by $[\mathbf{\Psi}(z)]_{ip} = \partial h_i(z)/\partial z_p$ and the output feature matrix $\mathbf{\Phi}_{\text{out}}(\boldsymbol{\xi}) \in \mathbb{R}^{N \times d_2}$ by $[\mathbf{\Phi}_{\text{out}}(\boldsymbol{\xi})]_{i,:} = \mathbf{f}_i(\boldsymbol{\xi})^\top$. Whenever these derivatives are square-integrable, define the population full and output-weight NTKs

$$\mathbf{K}^{\infty}_{(0)} = \mathbb{E}_{z\sim\mu}\left[\mathbf{\Psi}(z)\mathbf{\Psi}(z)^\top\right], \qquad \mathbf{K}^{\infty}_{\text{out}} = \mathbb{E}_{\boldsymbol{\xi}\sim\mu_h}\left[\mathbf{\Phi}_{\text{out}}(\boldsymbol{\xi})\mathbf{\Phi}_{\text{out}}(\boldsymbol{\xi})^\top\right].$$

Writing $\mathbf{\Psi} = (\mathbf{\Psi}_{\text{hid}}, \mathbf{\Phi}_{\text{out}})$ according to the hidden affine and output parameters gives

$$\mathbf{K}^{\infty}_{(0)} = \mathbf{K}^{\infty}_{\text{out}} + \mathbf{K}^{\infty}_{\text{hid}}, \qquad \mathbf{K}^{\infty}_{\text{hid}} = \mathbb{E}_{z\sim\mu}\left[\mathbf{\Psi}_{\text{hid}}(z)\mathbf{\Psi}_{\text{hid}}(z)^\top\right] \succeq \mathbf{0}, \tag{4}$$

and since both summands are positive semidefinite,

$$\mathbf{K}^{\infty}_{(0)} \succeq \mathbf{K}^{\infty}_{\text{out}}, \qquad \mathbf{K}^{\infty}_{(0)} \succeq \mathbf{K}^{\infty}_{\text{hid}}, \qquad \lambda_{\min}(\mathbf{K}^{\infty}_{(0)}) \geq \lambda_{\min}(\mathbf{K}^{\infty}_{\text{out}}) + \lambda_{\min}(\mathbf{K}^{\infty}_{\text{hid}}). \tag{5}$$

Definition 2.7 (Positive-definiteness certificate)**.** For a fixed loss-scaled row design, a *positive-definiteness certificate* is an exact proof that $\mathbf{K}^{\infty}_{(0)} \succ 0$. By Equation (5), it suffices to prove either $\mathbf{K}^{\infty}_{\text{out}} \succ 0$ or $\mathbf{K}^{\infty}_{\text{hid}} \succ 0$; the other block need not be checked.

The one-block condition is sufficient but not necessary, since $\ker \mathbf{K}^{\infty}_{(0)} = \ker \mathbf{K}^{\infty}_{\text{out}} \cap \ker \mathbf{K}^{\infty}_{\text{hid}}$. Our constructive certificates (Section D.3) use the output block because its features expose the row relations directly.

3 Positive definiteness from constraint independence

Here $\mathbf{K}_{(0)} \to \mathbf{K}^{\infty}_{(0)}$ means almost-sure convergence along nested widths and convergence in probability under fresh resampling. Each class follows the same pattern: the "reg" clause gives existence and convergence, the "pd" clause identifies $\ker \mathbf{K}^{\infty}_{(0)}$ with the linear relations among the rows in their natural representation, and a finite rank computation then checks that no relation exists. The proofs establish the same nullspace for the output block en route, so a trivial nullspace proves $\mathbf{K}^{\infty}_{\text{out}} \succ 0$ and hence certifies $\mathbf{K}^{\infty}_{(0)} \succ 0$.

3.1 Pointwise differential constraints

Let $\mathcal{S} = \{\mathbf{x}_i : i \in [N]\}$ be the set of sampled physical locations. For $\mathbf{x} \in \mathcal{S}$, let $I_{\mathbf{x}} = \{i : \mathbf{x}_i = \mathbf{x}\}$. Define the common coordinate set $\mathcal{C}_{\mathbf{x}} = \{(q, \alpha) : q \in [d_2], \alpha \in \bigcup_{i\in I_{\mathbf{x}}} I_{iq}\}$ and, extending each row's coefficient list by zeros in the derivative slots it does not use, the loss-scaled vectors

$$\mathbf{b}_i(\mathbf{x}) = \left(s_i a_{iq\alpha}(\mathbf{x})\right)_{(q,\alpha)\in\mathcal{C}_{\mathbf{x}}}, \qquad i \in I_{\mathbf{x}}.$$

Proposition 3.1 (Pointwise NTK: existence, nullspace, and rank)**.** *Under Assumption 2.5 and clause* P_{reg} *of Assumption 2.6,* $\mathbf{K}^{\infty}_{(0)}$ *exists and* $\mathbf{K}_{(0)} \to \mathbf{K}^{\infty}_{(0)}$*. If clause* P_{pd} *also holds, then*

$$\begin{aligned} \ker \mathbf{K}^{\infty}_{(0)} &= \left\{ \mathbf{c} \in \mathbb{R}^N : \sum_{i\in I_{\mathbf{x}}} c_i \mathbf{b}_i(\mathbf{x}) = \mathbf{0} \ \forall \mathbf{x} \in \mathcal{S} \right\}, \\ \operatorname{rank} \mathbf{K}^{\infty}_{(0)} &= \sum_{\mathbf{x}\in\mathcal{S}} \dim \operatorname{span}\{\mathbf{b}_i(\mathbf{x}) : i \in I_{\mathbf{x}}\}. \end{aligned} \tag{6}$$

Consequently, $\mathbf{K}^{\infty}_{(0)} \succ 0$ *if and only if every local family* $\{\mathbf{b}_i(\mathbf{x}) : i \in I_{\mathbf{x}}\}$ *is linearly independent.*

The output index q in $\mathbf{b}_i(\mathbf{x})$ already accounts for coupled vector-valued systems: several scalar equations at one point preserve positive definiteness exactly when their local coefficient vectors are independent. The check is a finite computation at each site: assemble $\mathbf{A}_{\mathbf{x}} = [\mathbf{b}_i(\mathbf{x})^\top]_{i\in I_{\mathbf{x}}}$ and verify $\operatorname{rank} \mathbf{A}_{\mathbf{x}} = |I_{\mathbf{x}}|$. The conclusion is qualitative for each fixed design; Proposition C.9 bounds $\lambda_{\min}(\mathbf{K}^{\infty}_{(0)})$ from below using finitely many parameter values and positive-measure sets on which the output-weight feature matrix has controlled variation.

3.2 Weak functional constraints

For rows of type (ii) in Definition 2.2, with $\mathcal{X}$ as in clause W_{reg}, define $\mathcal{R}_\lambda : \mathbb{R}^N \to \mathcal{X}^*$, $\mathcal{R}_\lambda \mathbf{c} = \sum_{i=1}^N c_i \lambda_i$.

Proposition 3.2 (Weak-form NTK: existence, nullspace, and rank)**.** *Under Assumption 2.5 and clause* $\mathsf{W}_{\mathrm{reg}}$ *of Assumption 2.6,* $\mathbf{K}_{(0)}^{\infty}$ *exists and* $\mathbf{K}_{(0)} \to \mathbf{K}_{(0)}^{\infty}$*. If clause* W_{pd} *also holds, then*

$$\ker \mathbf{K}_{(0)}^{\infty} = \ker \mathcal{R}_{\lambda}, \qquad \operatorname{rank} \mathbf{K}_{(0)}^{\infty} = \operatorname{rank} \mathcal{R}_{\lambda} = \dim \operatorname{span}\{\lambda_1, \dots, \lambda_N\}.$$

Thus $\mathbf{K}_{(0)}^{\infty} \succ 0$ *if and only if* $\lambda_1, \dots, \lambda_N$ *are linearly independent in* $\mathcal{X}^*$*.*

If $\lambda_i(f) = s_i a(f, v_i)$ for a common continuous coercive bilinear form a on $\mathcal{X}$, independence of $v_1, \dots, v_N$ suffices. Separate residual and boundary independence does not: the combined family must be independent (Theorem C.17 and Corollary C.18). Independence in $\mathcal{X}^*$ is decided by finitely many numbers. Under W_{pd} each $F_{iq}(\boldsymbol{\xi}) = \lambda_i(\varphi_{\boldsymbol{\xi},q})$ is real analytic, so $\mathbf{K}_{(0)}^{\infty} \succ 0$ exactly when finitely many derivatives of the F_{iq} at one fixed center have row rank N (Proposition C.13); and since, for real analytic nonpolynomial σ and $\gamma > 0$, differentiating $\sigma(\mathbf{w}^{\top}\mathbf{x} + \gamma b)$ in $\mathbf{w}$ and b at the origin yields the monomials $p_{\alpha}(\mathbf{x}) = \mathbf{x}^{\alpha}$ up to nonzero constants, derivatives may be replaced by moments: $\mathbf{K}_{(0)}^{\infty} \succ 0$ exactly when, for some finite K, the matrix $[s_i a_i(\mathbf{e}_q p_{\alpha}, v_i)]_{i \in [N],\, (q,\alpha):|\alpha| \le K}$ has row rank N (Corollary C.14). For a weak loss evaluated by quadrature, the rows are combinations $\lambda_i = \sum_j B_{ij} \mathcal{Q}_j$ of point evaluations $\mathcal{Q}_j(f) = \partial^{\alpha_j} f_{q_j}(\mathbf{x}_j)$; once the retained $\mathcal{Q}_j$ are shown independent in $\mathcal{X}^*$, positivity is equivalent to $\operatorname{rank} \mathbf{B} = N$.

3.3 Nonlocal integral constraints

For rows of type (iii) in Definition 2.2, set $\mathcal{X} = C(D; \mathbb{R}^{d_2})$, $\bar{\boldsymbol{\nu}}_i = s_i(\nu_{i1}, \dots, \nu_{id_2})$, and $\mathcal{R}_{\nu}\mathbf{c} = \sum_i c_i \bar{\boldsymbol{\nu}}_i$ for $\mathbf{c} \in \mathbb{R}^N$, with equality between measure tuples understood componentwise.

Proposition 3.3 (Nonlocal NTK: existence, nullspace, and rank)**.** *Under Assumption 2.5 and clause* $\mathsf{N}_{\mathrm{reg}}$ *of Assumption 2.6,* $\mathbf{K}_{(0)}^{\infty}$ *exists and* $\mathbf{K}_{(0)} \to \mathbf{K}_{(0)}^{\infty}$*. If clause* N_{pd} *also holds, then*

$$\ker \mathbf{K}_{(0)}^{\infty} = \ker \mathcal{R}_{\nu}, \qquad \operatorname{rank} \mathbf{K}_{(0)}^{\infty} = \operatorname{rank} \mathcal{R}_{\nu} = \dim \operatorname{span}\left\{\bar{\boldsymbol{\nu}}_1, \dots, \bar{\boldsymbol{\nu}}_N\right\}. \tag{7}$$

Consequently, $\mathbf{K}_{(0)}^{\infty} \succ 0$ *if and only if* $\bar{\boldsymbol{\nu}}_1, \dots, \bar{\boldsymbol{\nu}}_N$ *are linearly independent.*

Proof of Propositions 3.2 and 3.3. Lemma B.3 under $\mathsf{W}_{\mathrm{reg}}$, respectively Proposition B.6 under $\mathsf{N}_{\mathrm{reg}}$, verifies the abstract row bounds, so Theorem C.7 gives existence and convergence. The nullspace follows from Theorem C.17, respectively from Theorem C.21 applied to the combined residual and boundary tuples. □

The same moment test applies here with entries $s_i \int_D p_{\alpha}\, d\nu_{iq}$ in place of $s_i a_i(\mathbf{e}_q p_{\alpha}, v_i)$ (Corollary C.14); for measures with finitely many atoms, such as Dirac data or quadrature rows, independence of the tuples is the rank of a finite coefficient matrix over the distinct atoms. Section D.3 gives the exact procedures with worked examples.

4 Finite-width gradient descent across operator classes

The abstract transfer uses one local row-regularity condition.

Definition 4.1 (Local row regularity)**.** Fix $R > 0$. The rows satisfy (G_R) if there are width-independent constants $A_{0,i}, B_i, L_i < \infty$ such that, for every $z_0 \in \operatorname{supp} \mu$ and all $z, z' \in \overline{\mathbb{B}}_R(z_0)$,

$$|h_i(z_0)| \le A_{0,i}, \qquad \|\nabla h_i(z)\|_2 \le B_i, \qquad \|\nabla h_i(z) - \nabla h_i(z')\|_2 \le L_i \|z - z'\|_2. \tag{8}$$

This restates the row bounds in Definition B.1 and is a proof condition, not an extra PDE assumption: the "reg" clauses of Assumption 2.6 imply it classwise. Set $A_{0,*}^2 = \sum_i A_{0,i}^2$, $B_*^2 = \sum_i B_i^2$, $L_*^2 = \sum_i L_i^2$, $R_K = \operatorname{ess\,sup}_{z \sim \mu} \|\boldsymbol{\Psi}(z)\|_2^2 \le B_*^2$, and, for $0 < \delta < 1$ and $\mathbf{h}(z) = (h_i(z))_{i=1}^N$,

$$H_0^2 = \mathbb{E}_{z \sim \mu} \|\mathbf{h}(z)\|_2^2 \le A_{0,*}^2, \qquad \overline{r}_{\delta} = \|\mathbf{v}_T\|_2 + H_0 + A_{0,*}\sqrt{2 \log \tfrac{2}{\delta}}.$$

Under the conditional-centering hypothesis below, Lemma E.1 shows that $\overline{r}_{\delta}$ bounds the initial residual with probability at least $1 - \delta/2$, independently of width.

Theorem 4.2 (Finite-width GD for fixed linear observations)**.** *Under Assumption 2.5, suppose* (G_R) *of Definition 4.1 holds and* $\mathbb{E}[\mathbf{a} \mid \boldsymbol{\xi}] = \mathbf{0}$*. Let a positive-definiteness certificate give*

$$0 < \lambda_0 \le \lambda_{\min}(\mathbf{K}^{\infty}_{(0)}), \qquad \Lambda := B_*^2, \qquad c_{Ch} := \frac{1}{4} + \frac{3}{4}\log\frac{3}{4}.$$

A lower bound on either NTK block in Equation (5) *may serve as* λ_0*. For* $0 < \delta < 1$*, suppose*

$$d_1 \ge \max\left\{ \frac{R_K}{c_{Ch}\lambda_0}\log\frac{2N}{\delta}, (320 + 128\sqrt{6})\frac{L_*^2\Lambda\bar{r}_\delta^2}{\lambda_0^3}, \frac{16B_*^2\bar{r}_\delta^2}{R^2\lambda_0^2} \right\}, \qquad 0 < \eta\Lambda \le \frac{1}{2}. \tag{9}$$

Then, with probability at least $1-\delta$*, simultaneously for every* $k \in \mathbb{N}_0$*,*

$$\begin{aligned} \max_{m\in[d_1]} \|\theta_{m,(k)} - \theta_{m,(0)}\|_2 &\le R, \\ \mathbf{K}_{(k)} &\succeq \frac{\lambda_0}{2}\mathbf{I}_N, \\ \mathcal{L}(\theta_{(k)}) &\le (1-\eta\lambda_0)^k \mathcal{L}(\theta_{(0)}). \end{aligned} \tag{10}$$

With a fixed number of loss groups, $N_g^{-1/2}$ row scaling for a group of N_g rows, and uniform unscaled row and target bounds, all constants above except λ_0 are group-size independent. Example C.10 evaluates every constant for an exact continuum mean observation and obtains $d_1 \ge 1{,}356$ at confidence 0.95.

Corollary 4.3 (Classwise GD transfer)**.** *Under Assumption 2.5, suppose* $\mathbb{E}[\mathbf{a} \mid \boldsymbol{\xi}] = \mathbf{0}$ *and one of the following certificate conditions from Assumption 2.6 holds:*

(i) P_{pd} *holds and every co-located coefficient family is linearly independent;*
(ii) W_{pd} *holds and the combined functional family is linearly independent in* $\mathcal{X}^*$*; or*
(iii) N_{pd} *holds and the combined measure tuples are linearly independent.*

Then (G_R) *holds and* $\lambda_0 := \lambda_{\min}(\mathbf{K}^{\infty}_{(0)}) > 0$*. Consequently, for every* $0 < \delta < 1$*, Equation* (9) *implies Equation* (10) *with probability at least* $1-\delta$*.*

Proof. In each branch the cited rank condition proves $\mathbf{K}^{\infty}_{\mathrm{out}} \succ 0$, hence $\mathbf{K}^{\infty}_{(0)} \succ 0$ by Equation (5); (G_R) follows from Proposition B.7, Lemma B.3, or Proposition B.6, and Theorem 4.2 applies. □

Scope of the transfer Lemma B.3 transfers C^2 (or $C^{1,1}$) regularity of the complete one-neuron map into the trial space to (G_R) for arbitrary bounded rows, with explicit constants; Corollary B.5 and Proposition B.6 instantiate it for weak rows in H^k on a bounded Lipschitz domain in any dimension under $\sigma \in C^{k+2}_{\mathrm{loc}}$, with pointwise Dirichlet rows admissible once $k > d_0/2$ and pointwise rows of derivative order r once $k > d_0/2 + r$, and for finite signed measures on compact D under $\sigma \in C^2_{\mathrm{loc}}$; Lemma C.15 verifies the density clause of W_{pd} on the same spaces. Joint independence remains a separate check; Corollaries C.19 and C.20 verify everything for weak Poisson problems with discontinuous coefficients in one and two dimensions and, under conditional centering, give end-to-end weak-form GD instances. For singular designs whose targets satisfy the exact row relations, Section E.2 proves exact active-subspace compression and the corresponding GD guarantee.

Proof roadmap Matrix Chernoff gives $\mathbf{K}_{(0)} \succeq 3\lambda_0\mathbf{I}_N/4$ with high probability. Inside the radius-R tube, singular-value perturbation controls the gap through $(L_*/\sqrt{d_1})\|\theta - \theta_{(0)}\|_2$, the loss induction gives $\max_m \|\theta_{m,(k)} - \theta_{m,(0)}\|_2 \le 4B_*\|\mathbf{r}_{(0)}\|_2/(\sqrt{d_1}\lambda_0)$ and closes localization, and an integral Taylor remainder gives one-step contraction; the three width terms in Equation (9) enforce exactly these three steps. Full proofs are in Sections B and H.

5 Numerical checks

Three campaigns check the certificate route and the finite-width mechanism; designs, tolerances, full tallies, and provenance are in Section K. All use Equation (1) with $\sigma = \tanh$, $\gamma = 1$, i.i.d. $\mathrm{Unif}[-1,1]$ parameters, both layers trained, and $\eta = 0.2/\lambda_{\max}(\mathbf{K}_{(0)})$. The broad study trains

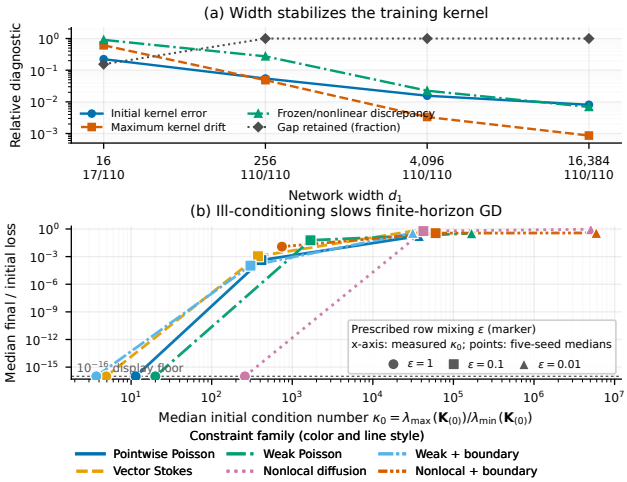


Figure 1: (a) Pooled medians over the 110 runs (eleven PDE designs, ten seeds) at each width of the initial relative kernel error against the randomized quasi-Monte Carlo (QMC) population reference $\widehat{\mathbf{K}}^{\infty}$, the maximum relative kernel drift, and the endpoint discrepancy from the frozen-kernel trajectory (Table 9); the gray curve is the fraction of runs with $\lambda_{\min}(\mathbf{K}_{(k)}) \geq \lambda_{\min}(\widehat{\mathbf{K}}^{\infty})/2$ at every saved checkpoint. (b) Five-seed median final-to-initial loss ratio at $d_1 = 16{,}384$ for the contrast target of each two-row family at row mixing $\varepsilon \in \{1, 0.1, 0.01\}$ (markers), against the measured initial condition number $\kappa_0 = \lambda_{\max}(\mathbf{K}_{(0)})/\lambda_{\min}(\mathbf{K}_{(0)})$; five medians at $\varepsilon = 1$ lie below the 10^{-16} display floor.

the eleven fixed linear PDE designs of Table 4 at seven widths and ten seeds for 3,000 full-batch steps (770 runs). The control study takes six two-row constraint families—co-located pointwise Poisson and vector Stokes rows, weak Poisson and nonlocal diffusion rows, and weak or nonlocal boundary–residual pairs—and gives each five row designs: the independent normalized pair (ℓ_1, ℓ_2), three mixings that replace ℓ_2 by $(\ell_1 + \varepsilon\ell_2)/\sqrt{1+\varepsilon^2}$ with $\varepsilon \in \{1, 0.1, 0.01\}$, and the exact duplicate $\varepsilon = 0$. At $d_1 \in \{16, 256, 4{,}096, 16{,}384\}$ with five seeds this gives 600 kernels; each of the 480 nonsingular kernels is trained on a manufactured and a contrast compatible target and each of the 120 duplicates on the manufactured target alone, giving 1,080 runs. A three-row audit uses designs with an exact row relation at $d_1 \in \{32, 128\}$ with three seeds. For each of the eleven designs the applicable certificate matrix of Section D.3 is built, rank-checked, and stored before any network is initialized; all eleven pass.

Width stabilizes the kernel and GD decreases the loss Endpoint training loss decreases in all 770 runs (Table 5). Pooled over the 110 runs at each displayed width, the initial relative kernel error, the maximum kernel drift, and the endpoint discrepancy from the frozen-kernel trajectory all decrease with width (Figure 1a), and the fraction of runs holding $\lambda_{\min}(\mathbf{K}_{(k)})$ at or above half the numerical population-reference gap at every saved checkpoint rises from $17/110$ at $d_1 = 16$ to $110/110$ at each larger displayed width. A per-step audit over thirty two-row runs finds the loss nonincreasing, up to a relative tolerance of 10^{-12}, at all 65,860 iterations that retain double-precision significance, with the largest per-step ratio matching the frozen-kernel factor $(1 - 2\eta\lambda_{\min}(\mathbf{K}_{(0)}))^2$ to a relative discrepancy of at most 1.2×10^{-5} (Section K.8). These finite-horizon diagnostics use a numerical reference, not a certified λ_0, and do not verify the constants in Equation (9).

Near-dependence shrinks the gap without breaking rank At the registered 10^{-12} relative tolerance all 480 independent two-row Gram matrices are classified nonsingular and all 120 duplicates singular (Table 7), an implementation check. For every $\varepsilon > 0$ the mixed design keeps full rank while $\lambda_{\min}(\mathbf{K}_{(0)})/\lambda_{\max}(\mathbf{K}_{(0)})$ shrinks, and at $d_1 = 16{,}384$ the contrast target's final-to-initial loss ratio increases across $\varepsilon \in \{1, 0.1, 0.01\}$ in all six families (Figure 1b, Table 8). The certified gap fixes only the guaranteed factor $1 - \eta\lambda_0$, not the observed finite-horizon loss.

Singular designs with compatible targets In the three-row audit, which exercises Section E.2, a compatible target and its null-displaced pair share one parameter path, so their losses differ by exactly 0.0625; that offset holds at all 19 saved checkpoints of 18 paired settings to within 4.44×10^{-16}. As Corollary E.4 predicts, the incompatible target's loss converges to the floor 0.0625 rather than to zero, and the shared active loss falls to roundoff in median (Figure 2).

6 Scope and conclusion

For each fixed finite design, structural nonredundancy determines whether the limiting gap is positive, and this premise is checkable through a finite coefficient, functional, or moment/Dirac matrix rather than assumed. The network and initialization hypotheses are easy to satisfy and easy to verify: each is a property of the chosen activation, architecture, and initialization law, read off from those choices without any computation, not a condition on the data or on the trajectory. An ordinary two-layer PINN meets all of them, with a smooth nonpolynomial activation such as $\tanh$, sigmoid, softplus, or GELU, a bias, all parameters trained, and i.i.d. bounded symmetric initialization such as the uniform law of Section 5; excluded are ReLU-type activations, unbounded initialization laws, and output-weight laws that vary with the width instead of the explicit $d_1^{-1/2}$ scaling of Equation (1). Necessary conditions are in Section E.4. Section D.1 shows that the classwise clauses of Assumptions 2.5 and 2.6 are not decorative: six constructions, each deleting or weakening one clause, admit independent rows whose limiting NTK is singular or fails to exist.

From training loss to solution error PINN error analyses built on stability of the underlying PDE bound the error between the trained network and the exact solution by a stability constant times the training residual plus a quadrature term (Mishra & Molinaro, 2023; De Ryck et al., 2024). Composed with Theorem 4.2, the geometric decay in Equation (10) drives the training term below any tolerance; the quadrature term remains. The composition presumes a stability estimate that we do not establish.

Full-batch gradient descent Full-batch updates reflect the problem, not a limitation of the analysis. A PINN loss over a fixed design (Definition 2.2 and assumption 2.5) is a deterministic finite sum without observation noise; on a certified design with a compatible target it is an interpolation problem with minimum value zero, and Theorem 4.2 shows that plain GD reaches any tolerance. Stochasticity therefore has no statistical role, and the remaining discretization error is the quadrature term of the design, which no optimizer changes; row subsampling is a computational device.

Depth At depth H (Definition F.19) one ingredient holds without any hypothesis: the tangent Gram matrix is a sum $\mathbf{K} = \sum_{l=1}^{H+1} \mathbf{K}^{[l]}$ of positive semidefinite layer blocks, so $\lambda_{\min}(\mathbf{K}) \geq \max_l \lambda_{\min}(\mathbf{K}^{[l]})$ and certifying one layer certifies the kernel (Lemma F.20). If, moreover, the rows satisfy a depth analogue (G_τ^H) of Definition 4.1 with a gap λ_0, the width and step-size conditions and the spectral and loss conclusions of Theorem 4.2 carry over termwise, with a layerwise localization in place of the per-neuron one (Proposition F.22). Whether (G_τ^H) holds for PINN rows at depth is open; we state its two missing ingredients as conjectures: that the analogues of Λ and L_*^2 are bounded by $C_0^{(1+r_{\max})H}$, the factor $1 + r_{\max}$ arising because a row of order $r_{\max}$ threads its derivatives through every layer, which order-zero observations (Du et al., 2019a; Allen-Zhu et al., 2019; Zou et al., 2020) avoid (Conjecture F.23), and that one layer block has the sitewise nullspace of Proposition 3.1 at every depth (Conjecture F.25). What the two-layer proof uses and depth lacks is the per-neuron sum $P_i(\theta) = d_1^{-1/2} \sum_m h_i(\theta_m)$ of independent terms.

Nonlinear constraints A nonlinear row, such as the convective term $u\,\partial_x u$ at a point, is a fixed nonlinear function of linear observations of the network. For such rows the situation is negative rather than open. Section F shows that the tangent Gram matrix at initialization converges in distribution to a random matrix that is deterministic for affine rows, whereas under a certificate for the base observations and centered isotropic output weights the diagonal entry of a genuinely nonlinear polynomial row has a nondegenerate random limit: its standard deviation stays of order one in the width instead of decaying like $d_1^{-1/2}$, so no deterministic gap exists to transfer. No choice of width exponent repairs this, since any deterministic limit of such a kernel block is zero. Whether GD converges in this regime is open.

Reproducibility statement

All assumptions are stated in Assumptions 2.5 and 2.6 and Definition 4.1, and the appendix gives complete proofs of every claim. The experimental designs, tolerances, provenance, and full tallies are in Section K. The code, configurations, tracked hash-validated aggregates, stored certificates, and locked environment are at `https://github.com/pinn-ntk-certificates/iclr2027`; its README gives the exact commands that run the tests and the smoke audit, reproduce Figure 1 and the three-row audit, and rerun the broad campaign, whose original per-run histories are not included.

AI use statement

In this work, we used generative AI tools to implement the numerical methods, to formulate auxiliary mathematical claims and assist in writing their proofs (Corollaries B.5 and C.20, Lemma C.15, Proposition E.5, and Example E.6), to provide feedback on the design and interpretation of the numerical experiments, and to audit existing proofs and code for errors. We have not used generative AI tools to develop the theoretical framework, to formulate or prove the main theorems, or to propose or refine hypotheses; generating synthetic data sets, translation, data cleaning, and qualitative analysis are not applicable to this work. Additionally, we used generative AI tools to draft and edit parts of the paper for readability, to search for and summarize related literature and draft parts of the related-work discussion, to write the repository documentation, and to check the consistency of reported numbers against tracked artifacts. We have reviewed all AI-assisted work: every AI-drafted proof was checked line by line by the authors, all AI-written code is covered by the unit tests and smoke runs in the released repository and was checked against the model, loss, and optimizer stated in the paper, and every number in the paper was traced to a tracked artifact. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

Robert A. Adams and John J. F. Fournier. *Sobolev Spaces*, volume 140 of *Pure and Applied Mathematics*. Academic Press, Amsterdam, second edition, 2003.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 242–252, 2019.

Andrea Bonfanti, Giuseppe Bruno, and Cristina Cipriani. The challenges of the nonlinear regime for physics-informed neural networks. In *Advances in Neural Information Processing Systems*, volume 37, pp. 41852–41881, 2024. doi: 10.52202/079017-1324.

Andrea Bonito, Ronald DeVore, Guergana Petrova, and Jonathan W. Siegel. Convergence and error control of consistent PINNs for elliptic PDEs. *IMA Journal of Numerical Analysis*, 46(1):90–148, 2026. doi: 10.1093/imanum/draf008.

Luís Carvalho, João L. Costa, José Mourão, and Gonçalo Oliveira. The positivity of the neural tangent kernel. *SIAM Journal on Mathematics of Data Science*, 7(2):495–515, 2025. doi: 10.1137/24M1659534.

Yuqian Cheng, Zhuo Chen, and Qian Lin. Consistency of physics-informed neural networks for second-order elliptic equations. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5389.

Tim De Ryck, Ameya D. Jagtap, and Siddhartha Mishra. Error estimates for physics-informed neural networks approximating the Navier–Stokes equations. *IMA Journal of Numerical Analysis*, 44(1): 83–119, 2024. doi: 10.1093/imanum/drac085.

Muditha W. M. G. Dissanayake and Nhan Phan-Thien. Neural-network-based approximations for solving partial differential equations. *Communications in Numerical Methods in Engineering*, 10 (3):195–201, 1994. doi: 10.1002/cnm.1640100303.

Nathan Doumèche, Gérard Biau, and Claire Boyer. On the convergence of PINNs. *Bernoulli*, 31(3): 2127–2151, 2025. doi: 10.3150/24-BEJ1799.

Simon S. Du, Jason D. Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1675–1685, 2019a.

Simon Shaolei Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019b.

Weiye Gan, Yicheng Li, Qian Lin, and Zuoqiang Shi. Neural tangent kernel of neural networks with loss informed by differential operators. arXiv preprint arXiv:2503.11029v1, 2025.

Yihang Gao, Yiqi Gu, and Michael K. Ng. Gradient descent finds the global optima of two-layer physics-informed neural networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10676–10707, 2023.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, pp. 8571–8580, 2018.

Deqing Jiang, Justin Sirignano, and Samuel N. Cohen. Global convergence of deep Galerkin and PINNs methods for solving partial differential equations. *SIAM Journal on Financial Mathematics*, 17(2):620–645, 2026. doi: 10.1137/24M1701502.

Bangti Jin and Longjun Wu. The differential neural tangent kernel and its positivity. arXiv preprint arXiv:2607.10200v1, 2026a.

Bangti Jin and Longjun Wu. On the extreme eigenvalues of the Gram matrix in physics-informed neural networks for the Poisson equation. arXiv preprint arXiv:2608.22426, 2026b.

Bangti Jin and Longjun Wu. Convergence of stochastic gradient methods for wide two-layer physics-informed neural networks for the Poisson equation. *Neural Networks*, 205:109335, 2027. doi: 10.1016/j.neunet.2026.109335. Preprint version: arXiv:2508.21571v2.

Ehsan Kharazmi, Zhongqiang Zhang, and George Em Karniadakis. hp-VPINNs: Variational physics-informed neural networks with domain decomposition. *Computer Methods in Applied Mechanics and Engineering*, 374:113547, 2021. doi: 10.1016/j.cma.2020.113547.

Ali Ouattara Kouma and Gossrin Jean-Marc Bomisso. Double screening in the training dynamics of variational physics-informed neural networks for heterogeneous coupled parabolic systems. arXiv preprint arXiv:2607.21352, 2026.

Aditi S. Krishnapriyan, Amir Gholami, Shandian Zhe, Robert M. Kirby, and Michael W. Mahoney. Characterizing possible failure modes in physics-informed neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 26548–26560, 2021.

Isaac E. Lagaris, Aristidis Likas, and Dimitrios I. Fotiadis. Artificial neural networks for solving ordinary and partial differential equations. *IEEE Transactions on Neural Networks*, 9(5):987–1000, 1998. doi: 10.1109/72.712178.

Isaac E. Lagaris, Aristidis Likas, and Dimitrios G. Papageorgiou. Neural-network methods for boundary value problems with irregular boundaries. *IEEE Transactions on Neural Networks*, 11 (5):1041–1049, 2000. doi: 10.1109/72.870037.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993. doi: 10.1016/S0893-6080(05)80131-5.

Siddhartha Mishra and Roberto Molinaro. Estimates on the generalization error of physics-informed neural networks for approximating PDEs. *IMA Journal of Numerical Analysis*, 43(1):1–43, 2023. doi: 10.1093/imanum/drab093.

J. Moser, C. Albert, and S. Ranftl. Resolving positive semi-definiteness in physics-informed kernels for scientific machine learning. arXiv preprint arXiv:2608.20985, 2026.

Jonas Nießen and Johannes Müller. Non-asymptotic analysis of projected gradient descent for physics-informed neural networks. In *Scientific Machine Learning: Emerging Topics*, pp. 65–92. Springer Nature Switzerland, 2026. doi: 10.1007/978-3-032-11527-0_4. Preprint version: arXiv:2505.07311v1.

Guofei Pang, Lu Lu, and George Em Karniadakis. fPINNs: Fractional physics-informed neural networks. *SIAM Journal on Scientific Computing*, 41(4):A2603–A2626, 2019. doi: 10.1137/18M1229845.

Guofei Pang, Marta D'Elia, Michael Parks, and George Em Karniadakis. nPINNs: Nonlocal physics-informed neural networks for a parametrized nonlocal universal Laplacian operator: Algorithms and applications. *Journal of Computational Physics*, 422:109760, 2020. doi: 10.1016/j.jcp.2020.109760.

Marvin Pförtner, Ingo Steinwart, Philipp Hennig, and Jonathan Wenger. Physics-informed Gaussian process regression generalizes linear PDE solvers. arXiv preprint arXiv:2212.12474, 2022.

Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. doi: 10.1016/j.jcp.2018.10.045.

Pratik Rathore, Weimu Lei, Zachary Frangella, Lu Lu, and Madeleine Udell. Challenges in training PINNs: A loss landscape perspective. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42159–42191, 2024.

Robert Schaback and Holger Wendland. Kernel techniques: From machine learning to meshless methods. *Acta Numerica*, 15:543–639, 2006. doi: 10.1017/S0962492906270016.

Yeonjong Shin, Jérôme Darbon, and George Em Karniadakis. On the convergence of physics-informed neural networks for linear second-order elliptic and parabolic-type PDEs. *Communications in Computational Physics*, 28(5):2042–2074, 2020. doi: 10.4208/cicp.OA-2020-0193.

Justin Sirignano, Konstantinos Spiliopoulos, and Samuel N. Cohen. Global convergence of DGM and PINN algorithms for solving nonlinear PDEs. arXiv preprint arXiv:2607.24726, 2026.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(70):2389–2410, 2011.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. doi: 10.1007/s10208-011-9099-z.

Sifan Wang, Yujun Teng, and Paris Perdikaris. Understanding and mitigating gradient pathologies in physics-informed neural networks. *SIAM Journal on Scientific Computing*, 43(5):A3055–A3081, 2021. doi: 10.1137/20M1318043.

Sifan Wang, Xinling Yu, and Paris Perdikaris. When and why PINNs fail to train: A neural tangent kernel perspective. *Journal of Computational Physics*, 449:110768, 2022. doi: 10.1016/j.jcp.2021.110768.

Xianliang Xu, Wang Kong, Jiaheng Mao, Zhongyi Huang, and Ye Li. Fast convergence of natural gradient descent for over-parameterized physics-informed neural networks. In *International Conference on Learning Representations*, 2026.

Xianliang Xu, Ting Du, Wang Kong, Bin Shan, Zhongyi Huang, and Ye Li. Convergence of implicit gradient descent for training two-layer physics-informed neural networks. *East Asian Journal on Applied Mathematics*, 17(1):59–94, 2027. doi: 10.4208/eajam.2025-011.150925.

Yaohua Zang, Gang Bao, Xiaojing Ye, and Haomin Zhou. Weak adversarial networks for high-dimensional partial differential equations. *Journal of Computational Physics*, 411:109409, 2020. doi: 10.1016/j.jcp.2020.109409.

Wei Zhao and Tao Luo. Convergence guarantees for gradient-based training of neural PDE solvers: From linear to nonlinear PDEs. arXiv preprint arXiv:2505.14002, 2025.

Zijian Zhou and Zhenya Yan. Is the neural tangent kernel of PINNs deep learning general partial differential equations always convergent? *Physica D: Nonlinear Phenomena*, 457:133987, 2024. doi: 10.1016/j.physd.2023.133987.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep ReLU networks. *Machine Learning*, 109:467–492, 2020.

APPENDIX CONTENTS

A Notation **15**

B Proofs for the NTK analysis of PINNs with linear constraints **16**

B.1 Abstract representation as a sum of per-neuron contributions 16

B.2 Abstract NTK analysis 22

C Extended positivity results **24**

C.1 Comparison with related theorem scopes 24

C.2 Deterministic limit, finite-width transfer, and local stability 29

C.3 Independence lemma, finite-witness bound, and an explicit example 30

C.4 General Gram and direct-sum criteria 32

C.5 Finite-dimensional rank conditions for analytic functionals 33

C.6 Weak residuals and boundary constraints 35

C.7 Nonlocal integral constraints and boundary measures 39

D Alternative positivity checks and singular examples **41**

D.1 Sharpness of the classwise hypotheses 41

D.2 Taylor-coefficient conditions for linear independence 49

D.3 Constructive positive-definiteness certificates 51

E Auxiliary optimization results **54**

E.1 Initialization error 54

E.2 Structural dependencies and compatible singular kernels 54

E.3 Fixed-kernel limiting loss 55

E.4 Necessary conditions: a loss floor and a width lower bound 55

F Nonlinear constraints and depth: obstructions to the present proof architecture **57**

F.1 Composite rows and an exact tangent Gram identity 57

F.2 Nondeterminism of the infinite-width kernel 59

F.3 No width exponent repairs the polynomial case 61

F.4 Toward depth: the assumptions, the implication, and what remains conjectural 64

G Proofs and auxiliary results for NTK nullspaces and positive definiteness **71**

G.1 Compatible training error for a singular NTK 75

G.2 Poisson proof by linear independence of parameter-derivative functions 77

H Proof of gradient descent convergence **77**

H.1 Proof skeleton 78

H.2 Taylor decomposition of consecutive predictions 79

H.3 Positive definiteness along gradient descent 80

H.4 Consequences of strong induction and a small learning step 80

I A singular NTK with a compatible target for `tanh` **83**

J Vector-valued pointwise constraints **85**

K Experimental protocols and finite-width diagnostics **85**

K.1 Matched architecture and experimental matrix 85

K.2 PDE benchmarks and manufactured data . . . 86
K.3 Compatible two-row controls . . . 89
K.4 Constraint independence and numerical rank . . . 90
K.5 Positive-definite and compatible singular GD . . . 90
K.6 Conditioning and the GD rate . . . 92
K.7 Width, NTK stability, and scope . . . 93
K.8 Per-iterate monotonicity . . . 94
K.9 Reproduction . . . 95

A NOTATION

This appendix consolidates the notation used in the main text and proofs. For a positive integer M, $[M] = \{1, \dots, M\}$. Input, weight, output, prediction, and residual vectors, as well as Gram matrices, Jacobians, and auxiliary coefficient or evaluation matrices, use boldface; the full and per-neuron parameter vectors θ and θ_m are exceptions. Superscript $\top$ denotes transpose.

Network and parameters The input dimension is d_0, the width is d_1, and the output dimension is d_2. For neuron $m \in [d_1]$, the hidden weight is $\mathbf{w}_m \in \mathbb{R}^{d_0}$, the hidden bias is $b_m \in \mathbb{R}$, the output weight is $\mathbf{a}_m \in \mathbb{R}^{d_2}$, and

$$\boldsymbol{\xi}_m = (\mathbf{w}_m, b_m), \qquad \theta_m = (\boldsymbol{\xi}_m, \mathbf{a}_m), \qquad \theta = (\theta_1, \dots, \theta_{d_1}).$$

Thus each unit has $d_\theta = d_0 + d_2 + 1$ trainable coordinates. We parenthesize iteration indices: $\theta_{(k)}$ is the complete iterate after k gradient-descent steps and $\theta_{m,(k)}$ is its mth block. When declared in $\mathbb{R}^{d_\theta}$, z, z', and z_0 are generic per-neuron parameter vectors and remain unbold; in activation formulas, z is instead a scalar dummy argument. In one-dimensional single-output examples, we suppress boldface for the weights w_m and a_m. We denote the initialization law of θ_m by μ and its hidden-parameter marginal by μ_h. The network is

$$\mathbf{u}(\mathbf{x}; \theta) = d_1^{-1/2} \sum_{m=1}^{d_1} \sigma(\mathbf{w}_m^\top \mathbf{x} + \gamma b_m) \mathbf{a}_m.$$

Here $\sigma : \mathbb{R} \to \mathbb{R}$ is the scalar activation and $\gamma \geq 0$ is the fixed bias scale of Definition 2.1.

Constraints and multi-indices The formulation has N scalar constraint rows. Their prediction and target vectors are $\mathbf{v}_P(\theta), \mathbf{v}_T \in \mathbb{R}^N$, and $\mathbf{r}(\theta) = \mathbf{v}_T - \mathbf{v}_P(\theta)$ is the residual vector; $\mathbf{r}_{(k)} = \mathbf{r}(\theta_{(k)})$. For a multi-index $\alpha \in \mathbb{N}_0^{d_0}$, $|\alpha| = \sum_j \alpha_j$, $\alpha! = \prod_j \alpha_j!$, $\mathbf{x}^\alpha = \prod_j x_j^{\alpha_j}$, and $\partial^\alpha = \prod_j \partial_{x_j}^{\alpha_j}$. For $\boldsymbol{\rho} \in \mathbb{N}_0^{d_0+1}$, the analogous parameter-space quantities are $|\boldsymbol{\rho}|$, $\boldsymbol{\rho}!$, and $\partial_{\boldsymbol{\xi}}^{\boldsymbol{\rho}}$. Likewise, $\partial_{\mathbf{w}}^{\alpha} = \prod_j \partial_{w_j}^{\alpha_j}$. The vector $\mathbf{e}_q \in \mathbb{R}^{d_2}$ is the qth standard basis vector, and $p_\alpha(\mathbf{x}) = \mathbf{x}^\alpha$. For a finite index set I, $\mathbb{R}^I$ denotes the associated real coordinate space. For finite-measure rows, ν_{iq} and ω_{jq} denote the finite signed regular Borel measures of residual row i and boundary row j in output q, with loss scales $s_i, t_j > 0$, and $\bar{\boldsymbol{\nu}}_i = s_i(\nu_{i1}, \dots, \nu_{id_2})$ is the loss-scaled measure tuple.

Jacobians, kernels, and matrix order The prediction Jacobian $\mathbf{J}(\theta) \in \mathbb{R}^{N \times d_1 d_\theta}$ has row $[\mathbf{J}(\theta)]_{i,:} = \nabla_\theta P_i(\theta)^\top$. Its tangent Gram matrix is $\mathbf{K}(\theta) = \mathbf{J}(\theta)\mathbf{J}(\theta)^\top$. We abbreviate $\mathbf{J}_{(k)} = \mathbf{J}(\theta_{(k)})$ and $\mathbf{K}_{(k)} = \mathbf{K}(\theta_{(k)})$; $\mathbf{K}_{(0)}^\infty$ is the deterministic infinite-width limit at initialization, and $\mathbf{K}_{\text{out}}^\infty$ is its output-weight contribution. For symmetric matrices, $\mathbf{A} \succeq \mathbf{B}$ (equivalently, $\mathbf{B} \preceq \mathbf{A}$) means that $\mathbf{A} - \mathbf{B}$ is positive semidefinite, and $\mathbf{A} \succ 0$ means that $\mathbf{A}$ is positive definite. We use $\ker$, rank, span, Range, and $\lambda_{\min}, \lambda_{\max}$ for nullspace, rank, linear span, column space, and the extreme eigenvalues. When $\mathbf{K}$ is positive definite, $\kappa(\mathbf{K}) = \lambda_{\max}(\mathbf{K})/\lambda_{\min}(\mathbf{K})$. We use $\mathbf{I}_N$ for the identity of size N. The smallest singular value is $\sigma_{\min}$, and $\operatorname{spec}(\mathbf{A})$ denotes the eigenvalue multiset of a square matrix $\mathbf{A}$. For a subspace $V \subseteq \mathbb{R}^N$, Π_V is the orthogonal projection onto V and $V^\perp$ is its orthogonal complement.

Per-neuron maps and kernel blocks Every loss-scaled linear row i, with scale $s_i > 0$ (Definition 2.2), has the per-neuron representation $P_i(\theta) = d_1^{-1/2} \sum_m h_i(\theta_m)$ with $h_i(\boldsymbol{\xi}, \mathbf{a}) = \mathbf{a}^\top \mathbf{f}_i(\boldsymbol{\xi})$, where $\mathbf{f}_i(\boldsymbol{\xi}) \in \mathbb{R}^{d_2}$ applies row i to the ridge functions $\varphi_{\boldsymbol{\xi},q}$ of one neuron (Definition 2.1 and the paragraph following Table 1). The single-neuron Jacobian $\boldsymbol{\Psi}(z) \in \mathbb{R}^{N \times d_\theta}$ has entries $\partial h_i(z)/\partial z_p$ and splits as $(\boldsymbol{\Psi}_{\mathrm{hid}}, \boldsymbol{\Phi}_{\mathrm{out}})$ into its hidden-parameter and output-weight columns, with $[\boldsymbol{\Phi}_{\mathrm{out}}(\boldsymbol{\xi})]_{i,:} = \mathbf{f}_i(\boldsymbol{\xi})^\top$ (the appendix also writes $\mathbf{F} = \boldsymbol{\Phi}_{\mathrm{out}}$); accordingly $\mathbf{K}^\infty_{(0)} = \mathbf{K}^\infty_{\mathrm{out}} + \mathbf{K}^\infty_{\mathrm{hid}}$ (Equation (4)). The condition (G_R) of Definition 4.1 bounds h_i, its gradient, and the gradient's Lipschitz modulus on closed balls of radius R around $\operatorname{supp}\mu$; $\mathcal{Z}_R = \operatorname{supp}\mu + \overline{\mathbb{B}}_R(\mathbf{0})$ is the corresponding compact per-neuron tube and $\mathcal{T}_R(\theta_{(0)})$ the product tube of all neurons around the initialization. For Banach-space-valued maps, $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm of a Fréchet derivative with respect to the Euclidean norm on the parameter space.

Norms and related operations The notation $\|\mathbf{v}\|_2$ is the Euclidean norm of a vector, whereas $\|\mathbf{A}\|_2$ and $\|\mathbf{A}\|_F$ are the spectral and Frobenius matrix norms. For functions, we write $\|f\|_{\mathcal{X}}$; $\|\nu\|_{\mathrm{TV}}$ denotes the total variation of a signed measure. Inner products are $\langle \cdot, \cdot \rangle$, with a subscript identifying the space when needed. Finally, $(t)_+ = \max\{t, 0\}$, and $\overline{\mathbb{B}}_R(z_0)$ is the closed Euclidean ball of radius R centered at z_0.

B Proofs for the NTK analysis of PINNs with linear constraints

This section proves the three items of Theorem C.7, which is stated in the next appendix where it is applied: existence of the deterministic limit, finite-width concentration, and local stability of the tangent Gram matrix. Section B.1 isolates the per-neuron regularity condition the proofs use and verifies it classwise; Theorem B.9 below then proves the three items under that condition.

B.1 Abstract representation as a sum of per-neuron contributions

We isolate the exact per-neuron sum in Theorem C.7 and the regularity conditions on each scalar prediction component used by the kernel and optimization arguments. We then verify them classwise: for bounded functional rows in Lemma B.3, with Sobolev instances in Corollaries B.4 and B.5, for finite-measure rows in Proposition B.6, and for finite-order strong-form linear differential constraints in Proposition B.7.

Definition B.1 (Abstract row-regularity condition)**.** Under Assumption 2.5, fix a target $\mathbf{v}_T$ and a finite collection of prediction components. Assemble them in $\mathbf{v}_P(\theta) \in \mathbb{R}^N$, and consider the loss $\mathcal{L}(\theta) = \|\mathbf{v}_P(\theta) - \mathbf{v}_T\|_2^2$. Write $\theta = (\theta_1, \ldots, \theta_{d_1})$, choose $R > 0$, and define the product neighborhood

$$\mathcal{T}_R(\theta_{(0)}) := \left\{ \theta : \max_{1 \le m \le d_1} \|\theta_m - \theta_{m,(0)}\|_2 \le R \right\}. \tag{11}$$

For each scalar prediction component i, one loss-scaled linear row of Definition 2.2, require

$$P_i(\theta) = \frac{1}{\sqrt{d_1}} \sum_{m=1}^{d_1} h_i(\theta_m), \qquad h_i(\boldsymbol{\xi}, \mathbf{a}) = \mathbf{a}^\top \mathbf{f}_i(\boldsymbol{\xi}). \tag{12}$$

The fixed data encoded by $\mathbf{f}_i$ may define any fixed scalar linear observation satisfying the representation in Equation (12) and the bounds in Equation (13) below. For a strong-form local differential row, they include its sampled evaluation point (its collocation point), its evaluated operator coefficients, and its loss scale s_i. Require deterministic constants $A_{0,i}, B_i, L_i < \infty$, uniform in d_1, such that, for every $z_0 \in \operatorname{supp}\mu$ and all $z, z' \in \overline{\mathbb{B}}_R(z_0)$,

$$\begin{aligned} |h_i(z_0)| &\le A_{0,i}, & \|\nabla h_i(z)\|_2 &\le B_i, \\ \|\nabla h_i(z) - \nabla h_i(z')\|_2 &\le L_i \|z - z'\|_2. \end{aligned} \tag{13}$$

If $\theta \in \mathcal{T}_R(\theta_{(0)})$ and every $\theta_{m,(0)} \in \operatorname{supp}\mu$, which holds almost surely under Assumption 2.5, then $\theta_m \in \overline{\mathbb{B}}_R(\theta_{m,(0)})$ for every m, so the bounds in Equation (13) apply to every neuron of θ with center $z_0 = \theta_{m,(0)}$. Define

$$A_{0,*}^2 = \sum_i A_{0,i}^2, \qquad B_*^2 = \sum_i B_i^2, \qquad L_*^2 = \sum_i L_i^2. \tag{14}$$

For later use, recall $d_\theta = d_0 + d_2 + 1$ and set $\mathbf{\Psi}(z) \in \mathbb{R}^{N \times d_\theta}$ by $[\mathbf{\Psi}(z)]_{ip} = \partial h_i(z)/\partial z_p$, $p \in [d_\theta]$. Differentiating Equation (12) in θ_m gives $\nabla_{\theta_m} P_i = d_1^{-1/2} \nabla h_i(\theta_m)$, so $\mathbf{J}(\theta) = d_1^{-1/2}[\mathbf{\Psi}(\theta_1) \cdots \mathbf{\Psi}(\theta_{d_1})]$. The gradient-Lipschitz bound yields the quadratic Taylor remainder used in the optimization argument through the integral of the gradient difference; this derivation does not require an everywhere-defined Hessian. The condition imposes neither analyticity nor nonpolynomiality.

For the width-independent bound on the initial training error, we additionally impose $\mathbb{E}_{z\sim\mu} h_i(z) = 0$ for every prediction component i. In the network parameterization of the main text, the single structural requirement $\mathbb{E}[\mathbf{a} \mid \boldsymbol{\xi}] = \mathbf{0}$ implies all these identities; independence between hidden and output weights is unnecessary.

Remark B.2 (Generality of Definition B.1). Finite-order strong-form differential constraints satisfy the condition under Assumption 2.5 and order-matched activation smoothness. For a tested weak, integro-differential, or fractional row, one must first verify Equations (12) and (13) for that row; linearity and bounded coefficients alone do not imply the required bounds.

Lemma B.3 (Automatic regularity for bounded functional rows)**.** *Under Assumption 2.5, so that* $\operatorname{supp}\mu$ *is compact, let* $\mathcal{X}$ *be a real Banach space of* $\mathbb{R}^{d_2}$*-valued functions embedded continuously in* L^1_{loc} *of its domain, so that an* $\mathcal{X}$*-limit that also converges almost everywhere coincides almost everywhere with its pointwise limit (this identifies Fréchet derivatives with classical parameter derivatives in the corollaries below; the lemma itself does not use it), let* $\lambda_i \in \mathcal{X}^*$*,* $i \in [N]$*, be fixed bounded linear functionals, and set*

$$\mathcal{G}(z) = \mathbf{a}\,\sigma(\mathbf{w}^\top \cdot + \gamma b) \in \mathcal{X}, \qquad z = (\mathbf{w}, b, \mathbf{a}) \in \mathbb{R}^{d_\theta}.$$

Fix $0 < R < \infty$ *and define the compact parameter tube*

$$\mathcal{Z}_R = \operatorname{supp}\mu + \overline{\mathbb{B}}_R(\mathbf{0}) = \{z : \operatorname{dist}(z, \operatorname{supp}\mu) \le R\}.$$

Suppose that $\mathcal{G}$ *is twice continuously Fréchet differentiable as an* $\mathcal{X}$*-valued map on an open neighborhood of* $\mathcal{Z}_R$*. Define*

$$M_0 = \sup_{z \in \operatorname{supp}\mu} \|\mathcal{G}(z)\|_{\mathcal{X}},$$
$$M_1 = \sup_{z \in \mathcal{Z}_R} \|D\mathcal{G}(z)\|_{\mathrm{op}},$$
$$M_2 = \sup_{z \in \mathcal{Z}_R} \|D^2\mathcal{G}(z)\|_{\mathrm{op}}.$$

Here $\|\cdot\|_{\mathrm{op}}$ *is taken with respect to the Euclidean norm on* $\mathbb{R}^{d_\theta}$*; for* $D^2\mathcal{G}(z)$ *it is the bilinear norm* $\sup_{\|\dot z\|_2 = \|\widehat z\|_2 = 1} \|D^2\mathcal{G}(z)[\dot z, \widehat z]\|_{\mathcal{X}}$*. These constants are finite. For the rows* $P_i(\theta) = \lambda_i(\mathbf{u}(\cdot;\theta))$ *of types (ii) and (iii) in Definition 2.2, since* $\mathbf{u}(\cdot;\theta) = d_1^{-1/2} \sum_m \mathcal{G}(\theta_m)$ *and* λ_i *is linear, the scalar maps* $h_i = \lambda_i \circ \mathcal{G}$ *give the representation in Equation* (12)*, with*

$$[\mathbf{f}_i(\boldsymbol{\xi})]_q = \lambda_i\big(\mathbf{e}_q \sigma(\mathbf{w}^\top \cdot + \gamma b)\big), \qquad \boldsymbol{\xi} = (\mathbf{w}, b),$$

and satisfy Equation (13) *with the explicit valid choices*

$$A_{0,i} = \|\lambda_i\|_{\mathcal{X}^*} M_0, \qquad B_i = \|\lambda_i\|_{\mathcal{X}^*} M_1, \qquad L_i = \|\lambda_i\|_{\mathcal{X}^*} M_2. \tag{15}$$

The same conclusion holds if $\mathcal{G}$ *is continuously Fréchet differentiable on an open neighborhood of* $\mathcal{Z}_R$ *and there is a finite constant* $M_{1,1}$ *such that*

$$\|D\mathcal{G}(z) - D\mathcal{G}(z')\|_{\mathrm{op}} \le M_{1,1}\|z - z'\|_2 \qquad (z, z' \in \mathcal{Z}_R).$$

In this $C^{1,1}$ *case, take* $L_i = \|\lambda_i\|_{\mathcal{X}^*} M_{1,1}$ *in place of the third choice in Equation* (15)*. In particular, Banach-valued real analyticity on an open neighborhood of the tube implies the displayed* C^2 *hypothesis and hence all three bounds.*

As a specialization, let $\mathcal{V}_i$ *be normed, let* $a_i : \mathcal{X} \times \mathcal{V}_i \to \mathbb{R}$ *be a continuous bilinear form, fix* $v_i \in \mathcal{V}_i$*, and define the loss-scaled weak row*

$$\lambda_i(f) = s_i a_i(f, v_i), \qquad s_i > 0.$$

If $\|a_i\|_{\mathrm{bil}}$ denotes its bilinear operator norm, then

$$\|\lambda_i\|_{\mathcal{X}^*} \le s_i \|a_i\|_{\mathrm{bil}} \|v_i\|_{\mathcal{V}_i}. \tag{16}$$

Consequently, substituting the right-hand side of Equation (16) *for $\|\lambda_i\|_{\mathcal{X}^*}$ in Equation* (15) *gives valid $A_{0,i}, B_i, L_i$ for every such weak row. Any bounded boundary functional on the same trial space is covered by the first part of the lemma.*

Proof. Compactness of $\operatorname{supp}\mu$ and finite dimensionality of the per-neuron parameter space make $\mathcal{Z}_R$ compact. Continuity of $\mathcal{G}, D\mathcal{G}, D^2\mathcal{G}$ therefore makes M_0, M_1, M_2 finite. Linearity of λ_i and the coordinate decomposition

$$\mathcal{G}(\boldsymbol{\xi}, \mathbf{a}) = \sum_{q=1}^{d_2} (\mathbf{a})_q \mathbf{e}_q \sigma(\mathbf{w}^\top \cdot + \gamma b)$$

give the claimed per-neuron representation. The Banach-space chain rule gives

$$Dh_i(z) = \lambda_i \circ D\mathcal{G}(z).$$

Boundedness of λ_i gives $|h_i(z_0)| \le \|\lambda_i\|_{\mathcal{X}^*} M_0$ on $\operatorname{supp}\mu$, and since $\nabla h_i(z)$ is the Riesz representative of $Dh_i(z)$, $\|\nabla h_i(z)\|_2 = \|Dh_i(z)\|_{\mathrm{op}} \le \|\lambda_i\|_{\mathcal{X}^*} M_1$ on $\mathcal{Z}_R$.

Fix $z_0 \in \operatorname{supp}\mu$ and $z, z' \in \overline{\mathbb{B}}_R(z_0)$. Their line segment remains in that ball and hence in $\mathcal{Z}_R$. The Banach-space fundamental theorem of calculus, applied to $D\mathcal{G}$, yields

$$\|D\mathcal{G}(z) - D\mathcal{G}(z')\|_{\mathrm{op}} \le M_2 \|z - z'\|_2.$$

Composition with λ_i proves the Lipschitz-gradient estimate and therefore Equation (15). The $C^{1,1}$ variant is the same argument with its defining Lipschitz bound. Finally, continuity and bilinearity give

$$|\lambda_i(f)| \le s_i \|a_i\|_{\mathrm{bil}} \|f\|_{\mathcal{X}} \|v_i\|_{\mathcal{V}_i},$$

which proves Equation (16) and the weak-row specialization. □

Corollary B.4 (Automatic regularity for standard weak H^1 rows)**.** *Let $\Omega \subset \mathbb{R}^{d_0}$ be a bounded open set and take $\mathcal{X} = H^1(\Omega; \mathbb{R}^{d_2})$. Suppose the per-neuron initialization law is width independent and compactly supported, fix $0 < R < \infty$, and assume $\sigma \in C^3_{\mathrm{loc}}(\mathbb{R})$. Then*

$$z = (\mathbf{w}, b, \mathbf{a}) \longmapsto \mathbf{a}\, \sigma(\mathbf{w}^\top \cdot + \gamma b)$$

is a C^2 map into $\mathcal{X}$ on an open neighborhood of the compact tube $\mathcal{Z}_R$ of Lemma B.3. Consequently, every fixed finite family of weak rows $\lambda_i(f) = s_i a_i(f, v_i)$, where $\mathcal{V}_i$ is normed, $a_i : \mathcal{X} \times \mathcal{V}_i \to \mathbb{R}$ is continuous and bilinear, and $v_i \in \mathcal{V}_i$ is fixed, together with any bounded boundary functionals on $H^1(\Omega; \mathbb{R}^{d_2})$, satisfies Equations (12) *and* (13)*, with the constants in Equation* (15)*.*

Here a boundary functional is covered only if it is already a bounded element of $H^1(\Omega; \mathbb{R}^{d_2})^*$, for example because an appropriate tested trace theorem applies. Pointwise Dirichlet evaluation is not automatically bounded on H^1 when $d_0 \ge 2$, and pointwise Neumann evaluation is not generally bounded on H^1; those rows require a smoother trial space or a bounded tested trace. This corollary verifies (G_R), equivalently the row bounds in Definition B.1. It does not establish Assumption 2.5, centering, Banach-valued analyticity or ridge density, joint functional independence, or a positive-definiteness certificate.

Proof. Write

$$t_z(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + \gamma b, \qquad \ell_{\dot z}(\mathbf{x}) = \dot{\mathbf{w}}^\top \mathbf{x} + \gamma \dot b$$

for a parameter direction $\dot z = (\dot{\mathbf{w}}, \dot b, \dot{\mathbf{a}})$, and use $\ell_{\widehat z}$ analogously for $\widehat z = (\widehat{\mathbf{w}}, \widehat b, \widehat{\mathbf{a}})$. The candidate first and second Fréchet derivatives of $\mathcal{G}(z) = \mathbf{a}\sigma(t_z)$ are

$$D\mathcal{G}(z)[\dot z] = \dot{\mathbf{a}}\, \sigma(t_z) + \mathbf{a}\, \sigma'(t_z) \ell_{\dot z}, \tag{17}$$

$$D^2\mathcal{G}(z)[\dot z, \widehat z] = \dot{\mathbf{a}}\, \sigma'(t_z) \ell_{\widehat z} + \widehat{\mathbf{a}}\, \sigma'(t_z) \ell_{\dot z} + \mathbf{a}\, \sigma''(t_z) \ell_{\dot z} \ell_{\widehat z}. \tag{18}$$

Their spatial weak derivatives are represented by their classical derivatives. In particular,

$$\begin{aligned}\nabla_{\mathbf{x}} D\mathcal{G}(z)[\dot{z}] &= \dot{\mathbf{a}}\mathbf{w}^\top\sigma'(t_z) + \mathbf{a}\dot{\mathbf{w}}^\top\sigma'(t_z) + \mathbf{a}\mathbf{w}^\top\sigma''(t_z)\ell_{\dot{z}},\\ \nabla_{\mathbf{x}} D^2\mathcal{G}(z)[\dot{z},\widehat{z}] &= \dot{\mathbf{a}}\big[\mathbf{w}^\top\sigma''(t_z)\ell_{\widehat{z}} + \widehat{\mathbf{w}}^\top\sigma'(t_z)\big]\\ &\quad + \widehat{\mathbf{a}}\big[\mathbf{w}^\top\sigma''(t_z)\ell_{\dot{z}} + \dot{\mathbf{w}}^\top\sigma'(t_z)\big]\\ &\quad + \mathbf{a}\big[\mathbf{w}^\top\sigma'''(t_z)\ell_{\dot{z}}\ell_{\widehat{z}} + \dot{\mathbf{w}}^\top\sigma''(t_z)\ell_{\widehat{z}} + \widehat{\mathbf{w}}^\top\sigma''(t_z)\ell_{\dot{z}}\big].\end{aligned}$$

Choose $\varepsilon > 0$ and work on the enlarged compact tube $\mathcal{Z}_{R+\varepsilon} = \operatorname{supp}\mu + \overline{\mathbb{B}}_{R+\varepsilon}(\mathbf{0})$, whose interior is an open neighborhood of $\mathcal{Z}_R$, as Lemma B.3 requires. Boundedness of Ω and compactness of this tube uniformly bound $\mathbf{x}$, all parameters, $t_z(\mathbf{x})$, and the affine functions $\ell_{\dot{z}}(\mathbf{x})$ for unit parameter directions. Hence $\sigma^{(j)}$, $0 \le j \le 3$, is bounded and uniformly continuous on one compact interval containing every relevant affine argument. The displayed formulas and their spatial derivatives are consequently uniformly bounded in $L^\infty(\Omega)$, in operator norm over unit directions.

Taylor's theorem in the scalar affine argument, with its remainder uniform on that compact interval, shows from the same formulas that the difference quotients defining Equations (17) and (18) converge in $W^{1,\infty}(\Omega;\mathbb{R}^{d_2})$. Uniform continuity of $\sigma^{(j)}$, $j \le 3$, also makes the two displayed derivative maps continuous in their corresponding operator norms. Thus $z \mapsto \mathcal{G}(z)$ is C^2 into $W^{1,\infty}$ on an open neighborhood of $\mathcal{Z}_R$. Since $|\Omega| < \infty$, the embedding

$$W^{1,\infty}(\Omega;\mathbb{R}^{d_2}) \hookrightarrow H^1(\Omega;\mathbb{R}^{d_2})$$

is continuous, so the same formulas are its first and second Fréchet derivatives as an H^1-valued map. Apply Lemma B.3 and Equation (16). □

The same argument applies to Sobolev spaces of any order on domains of any dimension, and a smoother trial space makes pointwise boundary rows admissible in every dimension.

Corollary B.5 (Automatic regularity for weak H^k rows in any dimension)**.** *Let $\Omega \subset \mathbb{R}^{d_0}$ be a bounded open set, $k \in \mathbb{N}_0$, and $\mathcal{X} = H^k(\Omega;\mathbb{R}^{d_2})$. Suppose the per-neuron initialization law is width independent and compactly supported, fix $0 < R < \infty$, and assume $\sigma \in C^{k+2}_{\mathrm{loc}}(\mathbb{R})$. Then $z = (\mathbf{w}, b, \mathbf{a}) \mapsto \mathbf{a}\,\sigma(\mathbf{w}^\top\cdot + \gamma b)$ is a C^2 map into $\mathcal{X}$ on an open neighborhood of the compact tube $\mathcal{Z}_R$, and every fixed finite family of bounded linear functionals on $\mathcal{X}$ satisfies Equations* (12) *and* (13)*, hence* (G_R)*, with the constants of Lemma B.3. If moreover Ω is a Lipschitz domain and $k > d_0/2 + r$ for an integer $r \ge 0$, every pointwise evaluation $f \mapsto \partial^\alpha f_q(\mathbf{x})$ with $|\alpha| \le r$ and $\mathbf{x} \in \overline{\Omega}$ is bounded on $\mathcal{X}$, so pointwise Dirichlet rows ($r = 0$) and pointwise Neumann or higher-order boundary rows ($r \ge 1$) are admissible weak rows in every dimension.*

Proof. Write $t_z(\mathbf{x}) = \mathbf{w}^\top\mathbf{x} + \gamma b$. For a multi-index β with $|\beta| \le k$, $\partial^\beta_{\mathbf{x}}[a_q\sigma(t_z(\mathbf{x}))] = a_q\,\mathbf{w}^\beta\,\sigma^{(|\beta|)}(t_z(\mathbf{x}))$, which as a function of z is a polynomial in $(\mathbf{w},\mathbf{a})$ times $\sigma^{(|\beta|)}(t_z(\mathbf{x}))$. On the enlarged compact tube $\mathcal{Z}_{R+\varepsilon}$, whose interior is an open neighborhood of $\mathcal{Z}_R$, the arguments $t_z(\mathbf{x})$, $\mathbf{x} \in \overline{\Omega}$, range over a compact interval on which $\sigma^{(|\beta|)}$, $\sigma^{(|\beta|+1)}$, and $\sigma^{(|\beta|+2)}$ are bounded and uniformly continuous, because $\sigma \in C^{k+2}_{\mathrm{loc}}$. Hence each $z \mapsto \partial^\beta_{\mathbf{x}}[\mathbf{a}\,\sigma(t_z(\cdot))]$ is C^2 into $C(\overline{\Omega};\mathbb{R}^{d_2})$, with the pointwise formulas as its derivatives, and since $C(\overline{\Omega}) \hookrightarrow L^2(\Omega)$ continuously on a bounded domain, the map is C^2 into $H^k(\Omega;\mathbb{R}^{d_2})$; this is the proof of Corollary B.4 with k spatial derivatives in place of one. Lemma B.3 then gives (G_R) for any finite family of bounded functionals. For the last claim, the Sobolev embedding $H^k(\Omega) \hookrightarrow C^r(\overline{\Omega})$ holds on bounded Lipschitz domains whenever $k > d_0/2 + r$ (Adams & Fournier, 2003, Theorem 4.12), so pointwise derivative evaluations of order at most r are bounded functionals on $\mathcal{X}$. □

Proposition B.6 (Automatic regularity of finite-measure rows)**.** *Let $D \subset \mathbb{R}^{d_0}$ be compact. For $i \in [N]$ and $q \in [d_2]$, let $\bar{\nu}_{iq} = s_i\nu_{iq}$ be the loss-scaled finite signed regular Borel measures of a type (iii) row in Definition 2.2, so that $\bar{\boldsymbol{\nu}}_i = (\bar{\nu}_{iq})_{q\in[d_2]}$, and define*

$$P_i(\theta) = \sum_{q=1}^{d_2}\int_D u_q(\mathbf{y};\theta)\,d\bar{\nu}_{iq}(\mathbf{y}).$$

Suppose the per-neuron initialization law μ is width independent and compactly supported (Assumption 2.5), fix $0 < R < \infty$, and assume $\sigma \in C^2_{\mathrm{loc}}(\mathbb{R})$. Then the induced per-neuron maps

satisfy Equations (12) *and* (13)*; in particular, clause* $\mathsf{N}_{\mathrm{reg}}$ *of Assumption 2.6 implies* (G_R) *of Definition 4.1.*

More precisely, write

$$\widetilde{\mathbf{y}} = (\mathbf{y}, \gamma), \qquad \tau_i = \left(\sum_{q=1}^{d_2} \|\bar{\nu}_{iq}\|_{\mathrm{TV}}^2 \right)^{1/2}, \qquad \mathcal{Z}_R = \operatorname{supp} \mu + \overline{\mathbb{B}}_R(\mathbf{0}),$$

and set

$$C_D = \max_{\mathbf{y} \in D} \|\widetilde{\mathbf{y}}\|_2, \qquad A_R = \max_{(\boldsymbol{\xi}, \mathbf{a}) \in \mathcal{Z}_R} \|\mathbf{a}\|_2,$$
$$T_{D,R} = \max_{\substack{(\boldsymbol{\xi}, \mathbf{a}) \in \mathcal{Z}_R \\ \mathbf{y} \in D}} |\boldsymbol{\xi}^\top \widetilde{\mathbf{y}}|, \qquad S_{j,R} = \max_{|t| \le T_{D,R}} |\sigma^{(j)}(t)|, \quad j = 0, 1, 2.$$

The following are valid choices:

$$\begin{aligned} A_{0,i} &= A_R S_{0,R} \tau_i, \\ B_i &= \tau_i \left(S_{0,R} + A_R S_{1,R} C_D \right), \\ L_i &= \tau_i \left(A_R S_{2,R} C_D^2 + S_{1,R} C_D \right). \end{aligned} \tag{19}$$

If $\mathbb{E}[\mathbf{a} \mid \boldsymbol{\xi}] = \mathbf{0}$*, then also* $\mathbb{E}[h_i(\boldsymbol{\xi}, \mathbf{a})] = 0$ *for every row.*

Proof. Define $\mathbf{f}_i : \mathbb{R}^{d_0+1} \to \mathbb{R}^{d_2}$ by

$$[\mathbf{f}_i(\boldsymbol{\xi})]_q = \int_D \sigma(\boldsymbol{\xi}^\top \widetilde{\mathbf{y}}) \, d\bar{\nu}_{iq}(\mathbf{y}).$$

Then

$$P_i(\theta) = \frac{1}{\sqrt{d_1}} \sum_{m=1}^{d_1} h_i(\theta_m), \qquad h_i(\boldsymbol{\xi}, \mathbf{a}) = \mathbf{a}^\top \mathbf{f}_i(\boldsymbol{\xi}).$$

Compactness of D, local smoothness of σ, and finite total variation justify differentiation under each integral. For $\mathbf{v}, \mathbf{v}' \in \mathbb{R}^{d_0+1}$,

$$D[\mathbf{f}_i]_q(\boldsymbol{\xi})[\mathbf{v}] = \int_D \sigma'(\boldsymbol{\xi}^\top \widetilde{\mathbf{y}}) \mathbf{v}^\top \widetilde{\mathbf{y}} \, d\bar{\nu}_{iq}(\mathbf{y}),$$
$$D^2[\mathbf{f}_i]_q(\boldsymbol{\xi})[\mathbf{v}, \mathbf{v}'] = \int_D \sigma''(\boldsymbol{\xi}^\top \widetilde{\mathbf{y}}) (\mathbf{v}^\top \widetilde{\mathbf{y}}) (\mathbf{v}'^\top \widetilde{\mathbf{y}}) \, d\bar{\nu}_{iq}(\mathbf{y}).$$

Consequently, throughout $\mathcal{Z}_R$,

$$\|\mathbf{f}_i\|_2 \le S_{0,R} \tau_i, \qquad \|D\mathbf{f}_i\|_{\mathrm{op}} \le S_{1,R} C_D \tau_i, \qquad \|D^2 \mathbf{f}_i\|_{\mathrm{op}} \le S_{2,R} C_D^2 \tau_i.$$

The gradient and Hessian of h_i have the block forms

$$\nabla h_i = \begin{pmatrix} (D\mathbf{f}_i)^\top \mathbf{a} \\ \mathbf{f}_i \end{pmatrix}, \qquad \nabla^2 h_i = \begin{pmatrix} \sum_q a_q D^2[\mathbf{f}_i]_q & (D\mathbf{f}_i)^\top \\ D\mathbf{f}_i & 0 \end{pmatrix}.$$

Bounding $\|D\mathbf{f}_i\|_{\mathrm{op}}$ by its Frobenius norm, using $\|\nabla h_i\|_2 \le \|(D\mathbf{f}_i)^\top \mathbf{a}\|_2 + \|\mathbf{f}_i\|_2$, $\|\sum_q a_q D^2[\mathbf{f}_i]_q\|_{\mathrm{op}} \le \sum_q |a_q| \, S_{2,R} C_D^2 \|\bar{\nu}_{iq}\|_{\mathrm{TV}} \le A_R S_{2,R} C_D^2 \tau_i$ (Cauchy–Schwarz over q), and $\left\| \left(\begin{smallmatrix} A & B^\top \\ B & 0 \end{smallmatrix} \right) \right\|_{\mathrm{op}} \le \|A\|_{\mathrm{op}} + \|B\|_{\mathrm{op}}$ gives Equation (19). The mean-value estimate on each convex radius-R ball then gives the required Lipschitz-gradient bound. Finally, $\mathbb{E}[h_i \mid \boldsymbol{\xi}] = \mathbb{E}[\mathbf{a} \mid \boldsymbol{\xi}]^\top \mathbf{f}_i(\boldsymbol{\xi}) = 0$, proving the centering assertion. □

We next check the componentwise conditions for the finite-order strong-form differential operators in Definition 2.2. The general results then apply directly to any other linear row for which the same requirements are verified.

Proposition B.7 (Row regularity of finite-order strong-form linear constraints)**.** *For the network in Definition 2.1, suppose Assumption 2.5 holds. Let $r_{\max}$ be the largest differential order among the finite collection of strong-form differential constraints in Definition 2.2. If $\sigma \in C^{r_{\max}+2}_{\mathrm{loc}}(\mathbb{R})$, or more generally $\sigma \in C^{r_{\max}+1,1}_{\mathrm{loc}}(\mathbb{R})$, then, for every fixed $0 < R < \infty$, this family satisfies Equations* (12) *and* (13) *with radius R and width-independent constants. In particular, Equation* (8) *is then automatic for such constraints. If σ is real analytic, each h_i, hence every entry of $\mathbf{\Psi}$, is real analytic in $z = (\boldsymbol{\xi}, \mathbf{a})$.*

Proof. For row i, set $t_m(\mathbf{x}_i) = \mathbf{w}_m^\top \mathbf{x}_i + \gamma b_m$. Applying each mixed derivative in Equation (2) to one neuron gives

$$P_i(\theta) = \frac{1}{\sqrt{d_1}} \sum_{m=1}^{d_1} h_i(\theta_m), \qquad h_i(\boldsymbol{\xi}_m, \mathbf{a}_m) = s_i \sum_{q=1}^{d_2} \sum_{\alpha \in I_{iq}} a_{iq\alpha}(\mathbf{x}_i)(\mathbf{a}_m)_q \sigma^{(|\alpha|)}(t_m(\mathbf{x}_i)) \mathbf{w}_m^\alpha. \tag{20}$$

where $a_{iq\alpha}(\mathbf{x}_i)$ are the operator coefficients of Equation (2) and $(\mathbf{a}_m)_q$ the output weights. Thus, the representation in Equation (12) holds with

$$[\mathbf{f}_i(\boldsymbol{\xi}_m)]_q = s_i \sum_{\alpha \in I_{iq}} a_{iq\alpha}(\mathbf{x}_i) \sigma^{(|\alpha|)}(t_m(\mathbf{x}_i)) \mathbf{w}_m^\alpha.$$

Each h_i is a finite sum of products of hidden-weight monomials, output-weight coordinates, and derivatives of σ of order at most $r_{\max}$. Hence $\sigma \in C^{r_{\max}+2}_{\mathrm{loc}}$ gives a continuously differentiable h_i with locally Lipschitz gradient; the weaker condition $\sigma \in C^{r_{\max}+1,1}_{\mathrm{loc}}$ yields the same conclusion directly. The tube $\mathcal{Z}_R = \operatorname{supp}\mu + \overline{\mathbb{B}}_R(\mathbf{0})$ is compact and width independent, so h_i and ∇h_i are uniformly bounded there, and, each ball $\overline{\mathbb{B}}_R(z_0)$ being convex, the mean-value inequality turns the uniform local Lipschitz constant of ∇h_i on $\mathcal{Z}_R$ into a Lipschitz constant on the ball. This proves Equation (13); explicit constants follow. Write $\widetilde{\mathbf{x}}_i = (\mathbf{x}_i, \gamma)$, $X_i = \|\widetilde{\mathbf{x}}_i\|_2$, $W_R = \max_{\mathcal{Z}_R} \max\{1, \|\mathbf{w}\|_2\}$, $A_R = \max_{\mathcal{Z}_R} \|\mathbf{a}\|_2$, $T_{\mathcal{S},R} = \max_i \max_{\mathcal{Z}_R} |\boldsymbol{\xi}^\top \widetilde{\mathbf{x}}_i|$, and $S_j = \max_{|t| \le T_{\mathcal{S},R}} |\sigma^{(j)}(t)|$ for $j \le r_{\max}+2$; in the $C^{r_{\max}+1,1}_{\mathrm{loc}}$ case let $S_{r_{\max}+2}$ be the Lipschitz constant of $\sigma^{(r_{\max}+1)}$ on $[-T_{\mathcal{S},R}, T_{\mathcal{S},R}]$. For $g_\alpha(\boldsymbol{\xi}) = \mathbf{w}^\alpha \sigma^{(|\alpha|)}(\boldsymbol{\xi}^\top \widetilde{\mathbf{x}}_i)$, the bounds $\|\nabla_{\mathbf{w}} \mathbf{w}^\alpha\|_2 \le |\alpha| W_R^{|\alpha|-1}$ and $\|\nabla^2_{\mathbf{w}} \mathbf{w}^\alpha\|_{\mathrm{op}} \le |\alpha|(|\alpha|-1) W_R^{|\alpha|-2}$ on $\mathcal{Z}_R$ and the product rule give, on $\mathcal{Z}_R$,

$$\begin{aligned} |g_\alpha| &\le G_0(\alpha) := W_R^{|\alpha|} S_{|\alpha|}, \\ \|\nabla g_\alpha\|_2 &\le G_1(\alpha) := W_R^{|\alpha|} \big(|\alpha| S_{|\alpha|} + X_i S_{|\alpha|+1}\big), \\ \mathrm{Lip}(\nabla g_\alpha) &\le G_2(\alpha) := W_R^{|\alpha|} \big(|\alpha|(|\alpha|-1) S_{|\alpha|} + 2|\alpha| X_i S_{|\alpha|+1} + X_i^2 S_{|\alpha|+2}\big). \end{aligned}$$

With $F_{k,iq} = s_i \sum_{\alpha \in I_{iq}} |a_{iq\alpha}(\mathbf{x}_i)|\, G_k(\alpha)$ and $\mathbf{F}_{k,i} = (F_{k,iq})_{q \in [d_2]}$, the block computation in the proof of Proposition B.6 yields the valid choices

$$A_{0,i} = A_R \|\mathbf{F}_{0,i}\|_2, \qquad B_i = \|\mathbf{F}_{0,i}\|_2 + A_R \|\mathbf{F}_{1,i}\|_2, \qquad L_i = \|\mathbf{F}_{1,i}\|_2 + A_R \|\mathbf{F}_{2,i}\|_2,$$

all proportional to s_i and independent of d_1. Finiteness of the sampled family makes the aggregate constants in Equation (14) finite. If σ is real analytic, Equation (20) is real analytic in the parameters of one neuron, and so is every entry of $\mathbf{\Psi}$. □

Remark B.8 (Weak and nonlocal rows). Let $\mathcal{X}$ be a real Banach space of $\mathbb{R}^{d_2}$-valued functions embedded continuously in L^1_{loc} as in clause $\mathsf{W}_{\mathrm{reg}}$, and let a loss-scaled bounded linear functional $\lambda_i \in \mathcal{X}^*$ induce a tested scalar constraint. Suppose the complete one-neuron map $z \mapsto \mathbf{a}\, \sigma(\mathbf{w}^\top \cdot + \gamma b)$ is C^2, or more generally $C^{1,1}$, as an $\mathcal{X}$-valued map on an open neighborhood of the compact parameter tube $\mathcal{Z}_R$. Then Lemma B.3 verifies Equations (12) and (13), with the explicit constants in Equation (15) (with $L_i = \|\lambda_i\|_{\mathcal{X}^*} M_{1,1}$ in the $C^{1,1}$ case). In particular, a continuous weak bilinear form with a fixed test function satisfies the dual-norm estimate in Equation (16); mixed weak-residual and bounded boundary rows are covered when they act on one common trial space. Finite-measure rows are the case $\mathcal{X} = C(D; \mathbb{R}^{d_2})$ with $\|\lambda_i\|_{\mathcal{X}^*} \le \sum_q \|\bar{\nu}_{iq}\|_{\mathrm{TV}}$; Proposition B.6 gives the sharper constants Equation (19) for them directly under $\sigma \in C^2_{\mathrm{loc}}$. Fixed-node quadrature rules without derivative evaluations are atomic finite-measure rows and fall under that proposition; rules containing derivative evaluations are a special case here whenever every evaluation used by the rule is bounded on the chosen trial space. Singular integral, fractional, and continuum weak constraints require this regularity check; coefficient boundedness alone does not suffice.

B.2 ABSTRACT NTK ANALYSIS

Theorem B.9 (Abstract NTK analysis)**.** *If the row-regularity condition in Definition B.1 holds, the deterministic-limit, finite-width concentration, and local stability statements in Theorem C.7 hold, with bounded continuous parameter-derivative functions. For strong-form differential constraints of any fixed order, these functions are real analytic in the per-neuron parameters when σ is, by Proposition B.7. Weak or nonlocal cases require separate verification of this property.*

We prove the three statements in Theorem C.7 individually under this condition.

Proof of item 1 of Theorem C.7. Order the coordinates of $z = \theta_m$, and define the parameter-derivative functions

$$\psi_{i,p}(z) = \partial_{z_p} h_i(z), \qquad p \in [d_\theta].$$

Take one infinite i.i.d. sequence of per-neuron parameter vectors and construct the width-d_1 network from its first d_1 vectors. Embedding these models in a common probability space gives almost-sure convergence as $d_1 \to \infty$; fresh independent resampling at each size gives convergence in probability. The prediction components and constraint design remain fixed, or one samples them independently of the parameter sequence; in the latter case, the following limit is conditional on the realized design. The sum over neurons in Equation (12) gives, exactly,

$$[\mathbf{K}_{(0)}]_{ij} = \frac{1}{d_1} \sum_{m=1}^{d_1} \sum_{p=1}^{d_\theta} \psi_{i,p}(\theta_{m,(0)}) \psi_{j,p}(\theta_{m,(0)}).$$

For fixed i, j, p, these terms are i.i.d. by Assumption 2.5 and bounded by Equation (13). The strong law therefore gives the almost-sure deterministic limit

$$[\mathbf{K}^\infty_{(0)}]_{ij} = \sum_{p=1}^{d_\theta} \mathbb{E}_{z\sim\mu}[\psi_{i,p}(z)\psi_{j,p}(z)].$$

Equivalently, $\mathbf{K}^\infty_{(0)} = \mathbb{E}_{z\sim\mu}[\mathbf{\Psi}(z)\mathbf{\Psi}(z)^\top]$, the population NTK of the main text. Among the d_θ coordinates of z, d_2 correspond to output weights and $d_0 + 1$ to hidden affine parameters. Each trainable coordinate $p \in [d_\theta]$ contributes the positive-semidefinite Gram matrix $[\langle \Phi_i^{(p)}, \Phi_j^{(p)} \rangle_{L^2(\mu)}]_{i,j\in[N]}$ of the functions $\Phi_i^{(p)} := \psi_{i,p} \in L^2(\mu)$, so $\mathbf{K}^\infty_{(0)}$ is the positive-semidefinite sum of these d_θ matrices. These functions are bounded by B_i and Lipschitz on every ball $\overline{\mathbb{B}}_R(z_0)$, $z_0 \in \operatorname{supp}\mu$, by Definition B.1, hence continuous on $\operatorname{supp}\mu$; in particular $\psi_{i,p} \in L^2(\mu)$. For the strong-form differential rows of Proposition B.7 with real analytic σ, they are analytic in the parameters of a single neuron. □

Proposition B.10 (Finite-width positive definiteness at initialization)**.** *Suppose the condition in Definition B.1 holds and the fixed design has a positive-definiteness certificate as in Definition 2.7, so $\mathbf{K}^\infty_{(0)} \succ 0$, and let $0 < \lambda_0 \le \lambda_{\min}(\mathbf{K}^\infty_{(0)})$ be a certified lower bound, and fix $0 < \delta < 1$. Set*

$$R_K := \operatorname*{ess\,sup}_{z\sim\mu} \|\mathbf{\Psi}(z)\|_2^2 \le B_*^2 \quad \textit{(by Equation (13))}, \qquad c_{Ch} := \frac{1}{4} + \frac{3}{4}\log\frac{3}{4}.$$

If

$$d_1 \ge \frac{R_K}{c_{Ch}\lambda_0} \log\left(\frac{N}{\delta}\right),$$

then $\lambda_{\min}(\mathbf{K}_{(0)}) \ge 3\lambda_0/4$ with probability at least $1 - \delta$.

Proof of Proposition B.10 (item 2 of Theorem C.7). Let $\mathbf{J}_{(0),m} \in \mathbb{R}^{N\times d_\theta}$ denote the initialization-Jacobian block for neuron m, and define

$$\mathbf{\Psi}_m = \sqrt{d_1}\,\mathbf{J}_{(0),m} = \mathbf{\Psi}(\theta_{m,(0)}), \qquad \mathbf{G}_m = \mathbf{\Psi}_m \mathbf{\Psi}_m^\top.$$

The per-neuron parameter vectors are i.i.d., each $\mathbf{G}_m$ is positive semidefinite, and

$$\mathbf{K}_{(0)} = \frac{1}{d_1} \sum_{m=1}^{d_1} \mathbf{G}_m, \qquad \mathbb{E}\mathbf{G}_m = \mathbf{K}^\infty_{(0)}.$$

Because $\lambda_{\max}(\mathbf{G}_m) = \|\mathbf{\Psi}_m\|_2^2$ and $\theta_{m,(0)} \in \operatorname{supp}\mu$ almost surely, the bounds in Equation (13) give the almost-sure estimate

$$\lambda_{\max}(\mathbf{G}_m) \le R_K = \operatorname*{ess\,sup}_{z\sim\mu} \|\mathbf{\Psi}(z)\|_2^2 \le \operatorname*{ess\,sup}_{z\sim\mu} \|\mathbf{\Psi}(z)\|_F^2 \le B_*^2.$$

Apply the lower-tail matrix Chernoff inequality (Tropp, 2012, Corollary 5.2) to the independent summands $\mathbf{X}_m = \mathbf{G}_m/d_1$, which satisfy $\lambda_{\max}(\mathbf{X}_m) \le R_K/d_1$ almost surely and $\mu_{\min} := \lambda_{\min}(\sum_m \mathbb{E}\mathbf{X}_m) = \lambda_{\min}(\mathbf{K}_{(0)}^{\infty}) \ge \lambda_0$. Since $\{\lambda_{\min}(\mathbf{K}_{(0)}) \le (1-\varepsilon)\lambda_0\} \subseteq \{\lambda_{\min}(\mathbf{K}_{(0)}) \le (1-\varepsilon)\mu_{\min}\}$ and the base $e^{-\varepsilon}/(1-\varepsilon)^{1-\varepsilon}$ lies in $(0,1]$, replacing $\mu_{\min}$ by λ_0 in the exponent only weakens the bound, and we obtain, for $0 \le \varepsilon < 1$,

$$\Pr\{\lambda_{\min}(\mathbf{K}_{(0)}) \le (1-\varepsilon)\lambda_0\} \le N\left[\frac{e^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}}\right]^{d_1\lambda_0/R_K}.$$

Setting $\varepsilon = 1/4$ gives the exponent $c_{\mathrm{Ch}} > 0$ (so $c_{\mathrm{Ch}}^{-1} \approx 29.207$) and yields

$$\Pr\left\{\lambda_{\min}(\mathbf{K}_{(0)}) \le \frac{3}{4}\lambda_0\right\} \le N\exp\left(-c_{\mathrm{Ch}}\frac{d_1\lambda_0}{R_K}\right).$$

The stated width makes the right-hand side at most δ, proving the claim. □

Remark B.11 (Chernoff bound in limiting-kernel geometry). The same argument can retain the population geometry. Define the largest eigenvalue of a one-neuron Gram term in limiting-NTK coordinates:

$$\tau_K := \operatorname*{ess\,sup}_{z\sim\mu} \lambda_{\max}\left((\mathbf{K}_{(0)}^{\infty})^{-1/2}\mathbf{\Psi}(z)\mathbf{\Psi}(z)^{\top}(\mathbf{K}_{(0)}^{\infty})^{-1/2}\right).$$

The resulting whitened matrices have expectation $\mathbf{I}_N$ and are bounded above by $\tau_K\mathbf{I}_N$. Matrix Chernoff therefore yields

$$\Pr\left\{\mathbf{K}_{(0)} \not\succeq \frac{3}{4}\mathbf{K}_{(0)}^{\infty}\right\} \le N\exp(-c_{\mathrm{Ch}}d_1/\tau_K).$$

Indeed, the inequality first gives the good event for $(\mathbf{K}_{(0)}^{\infty})^{-1/2}\mathbf{K}_{(0)}(\mathbf{K}_{(0)}^{\infty})^{-1/2}$, and congruence by $(\mathbf{K}_{(0)}^{\infty})^{1/2}$ gives the displayed matrix-order inequality. Thus $d_1 \ge c_{\mathrm{Ch}}^{-1}\tau_K\log(N/\delta)$ is an alternative sufficient width for the event $\mathbf{K}_{(0)} \succeq \frac{3}{4}\mathbf{K}_{(0)}^{\infty}$. Since $\tau_K \le R_K/\lambda_{\min}(\mathbf{K}_{(0)}^{\infty})$, it is never larger than the width in Proposition B.10 with $\lambda_0 = \lambda_{\min}(\mathbf{K}_{(0)}^{\infty})$, but it requires the whitened quantity τ_K rather than only R_K and a certified λ_0.

Remark B.12 (Comparison with entrywise concentration). An entrywise Hoeffding alternative with B_K an almost-sure upper bound on the range of each scalar summand requires $d_1 = \Omega(B_K^2N^2\lambda_0^{-2}\log(N/\delta))$. It takes a union over the N^2 entries and then uses the Frobenius norm. Matrix Chernoff exploits positivity and controls the required lower eigenvalue directly. Its sufficient width is $d_1 = \Omega(R_K\lambda_0^{-1}\log(N/\delta))$. With a fixed number of loss groups and each row divided by the square root of its group's sample count, $B_K = O(1/N)$ and $R_K = O(1)$, so the two widths are $\Omega(\lambda_0^{-2}\log(N/\delta))$ and $\Omega(\lambda_0^{-1}\log(N/\delta))$; without normalization $B_K = O(1)$ and $R_K = O(N)$ generically, and they are $\Omega(N^2\lambda_0^{-2}\log(N/\delta))$ and $\Omega(N\lambda_0^{-1}\log(N/\delta))$. In both regimes the Chernoff width is never larger, since, with $B_K = \max_i B_i^2$, $\lambda_0 \le \lambda_{\min}(\mathbf{K}_{(0)}^{\infty}) \le N^{-1}\sum_i B_i^2 \le B_K$ and $R_K \le B_*^2 \le NB_K$.

Proposition B.13 (Positive definiteness during training)**.** *Suppose the condition in Definition B.1 holds, let $\theta_{(0)} \in (\operatorname{supp}\mu)^{d_1}$, which holds almost surely under Assumption 2.5, and let $\theta \in \mathcal{T}_R(\theta_{(0)})$. Then $\lambda_{\max}(\mathbf{K}(\theta)) \le B_*^2$ and*

$$\lambda_{\min}(\mathbf{K}(\theta)) \ge \left(\sqrt{\lambda_{\min}(\mathbf{K}_{(0)})} - \frac{L_*}{\sqrt{d_1}}\|\theta - \theta_{(0)}\|_2\right)_+^2; \tag{21}$$

for an iterate $\theta = \theta_{(k)}$ this bounds $\lambda_{\min}(\mathbf{K}_{(k)})$. In particular, if $0 < \lambda_{\mathrm{thr}} < \lambda_{\min}(\mathbf{K}_{(0)})$ and

$$\frac{L_*}{\sqrt{d_1}}\|\theta_{(k)} - \theta_{(0)}\|_2 \le \sqrt{\lambda_{\min}(\mathbf{K}_{(0)})} - \sqrt{\lambda_{\mathrm{thr}}}, \tag{22}$$

then $\lambda_{\min}(\mathbf{K}_{(k)}) \ge \lambda_{\mathrm{thr}}$.

Proof of Proposition B.13 (item 3 of Theorem C.7). For any real matrix $\mathbf{A}$ with N rows, write $\sigma_{\min}(\mathbf{A}) := \sqrt{\lambda_{\min}(\mathbf{A}\mathbf{A}^\top)}$; this convention includes the necessary zero singular values when $\mathbf{A}$ has fewer than N columns. Recall that $\mathbf{\Psi}(z) \in \mathbb{R}^{N\times d_\theta}$ satisfies $[\mathbf{\Psi}(z)]_{ip} = \partial h_i(z)/\partial z_p$. Summing the row bounds in Equation (13) over $i \in [N]$ and using Equation (14) together with $\|\cdot\|_2 \le \|\cdot\|_F$ gives, for every $z_0 \in \operatorname{supp}\mu$ and all $z, z' \in \overline{\mathbb{B}}_R(z_0)$,

$$\|\mathbf{\Psi}(z)\|_2 \le B_*, \qquad \|\mathbf{\Psi}(z) - \mathbf{\Psi}(z')\|_F^2 = \sum_{i=1}^{N} \|\nabla h_i(z) - \nabla h_i(z')\|_2^2 \le L_*^2 \|z - z'\|_2^2.$$

Fix $\theta \in \mathcal{T}_R(\theta_{(0)})$. By Equation (11), $\theta_m \in \overline{\mathbb{B}}_R(\theta_{m,(0)})$ for every m, and $\theta_{m,(0)} \in \operatorname{supp}\mu$ by hypothesis, so the displayed bounds apply to each pair $(z, z') = (\theta_m, \theta_{m,(0)})$ with $z_0 = \theta_{m,(0)}$. Write the full prediction Jacobian as a horizontal concatenation over neurons:

$$\mathbf{J}(\theta) = \frac{1}{\sqrt{d_1}}[\mathbf{\Psi}(\theta_1) \ \cdots \ \mathbf{\Psi}(\theta_{d_1})].$$

Then

$$\begin{aligned}\|\mathbf{J}(\theta) - \mathbf{J}(\theta_{(0)})\|_2^2 \le \|\mathbf{J}(\theta) - \mathbf{J}(\theta_{(0)})\|_F^2 &= \frac{1}{d_1}\sum_{m=1}^{d_1} \|\mathbf{\Psi}(\theta_m) - \mathbf{\Psi}(\theta_{m,(0)})\|_F^2 \\ &\le \frac{L_*^2}{d_1}\sum_{m=1}^{d_1} \|\theta_m - \theta_{m,(0)}\|_2^2 = \frac{L_*^2}{d_1}\|\theta - \theta_{(0)}\|_2^2, \end{aligned} \tag{23}$$

so that $\|\mathbf{J}(\theta) - \mathbf{J}(\theta_{(0)})\|_2 \le L_*\|\theta - \theta_{(0)}\|_2/\sqrt{d_1}$. Also, because $\mathbf{K}(\theta) = \mathbf{J}(\theta)\mathbf{J}(\theta)^\top$, for every $\theta \in \mathcal{T}_R(\theta_{(0)})$,

$$\lambda_{\max}(\mathbf{K}(\theta)) \le \frac{1}{d_1}\sum_{m=1}^{d_1} \|\mathbf{\Psi}(\theta_m)\|_2^2 \le B_*^2.$$

Weyl's inequality for singular values, $|\sigma_N(\mathbf{A}) - \sigma_N(\mathbf{B})| \le \|\mathbf{A} - \mathbf{B}\|_2$, and Equation (23) yield

$$\begin{aligned}\sigma_{\min}(\mathbf{J}(\theta)) &\ge \sigma_{\min}(\mathbf{J}(\theta_{(0)})) - \|\mathbf{J}(\theta) - \mathbf{J}(\theta_{(0)})\|_2 \\ &\ge \sqrt{\lambda_{\min}(\mathbf{K}_{(0)})} - \frac{L_*}{\sqrt{d_1}}\|\theta - \theta_{(0)}\|_2.\end{aligned}$$

Taking the positive part and using $\lambda_{\min}(\mathbf{J}\mathbf{J}^\top) = \sigma_{\min}(\mathbf{J})^2$ proves Equation (21). Under the condition in Equation (22), the right-hand side of Equation (21) at $\theta = \theta_{(k)}$ is at least λ_{thr}, which establishes the final assertion. □

C Extended positivity results

C.1 Comparison with related theorem scopes

We compare results whose principal theorems concern an operator-transformed NTK limit, definiteness or spectrum, or optimization of a linear PINN in a wide-network or kernel regime. This includes the closest ordinary-NTK positivity and weak-form conditioning baselines. Table 2 records each result's stated setting and conclusion, followed by a constraint family treated here but outside that theorem's hypotheses. The comparison concerns theorem scope rather than paper-level generality; an outside-scope example is not a counterexample to the cited result. Complementary work on method history, solution consistency, generic kernel background, empirical failure modes, and nonlinear-PDE optimization is summarized after the table. For pairwise-distinct sites, Proposition 3.1 reduces to the familiar condition that every constraint have a nonzero evaluated coefficient vector $\mathbf{b}_i(\mathbf{x}_i)$, which under P_{pd} is then also necessary; the proof of Proposition 3.1 is in Section G.

Table 2: Comparison of direct theorem-level results. PD and PSD mean positive definite and positive semidefinite; GF, GD, IGD, and NGD denote gradient flow, gradient descent, implicit GD, and natural GD; SGD denotes stochastic GD. The final column identifies additional constraint structures treated in this paper.

Work and result	Setting and assumptions	Main conclusion	Additional scope treated here
Panel A: kernel limits, positivity, rank, and spectrum			
Wang et al. (2022), limiting PINN NTK	Fully connected PINNs with smooth activations and pointwise residual and boundary rows; the infinite-width statements assume bounded parameters and loss along GF and an invertible deterministic limiting matrix for exponential residual decay.	Derives the joint block DNTK, its deterministic initialization limit, and freezing along infinite-width GF. Exponential residual decay is obtained under an invertibility assumption; exact rank criteria and finite-width discrete-GD guarantees are outside its scope.	Two equal rows give a singular joint Gram even when the ordinary input NTK is PD. Proposition 3.1 and Corollary E.4 classify both equal and unequal targets.
Carvalho et al. (2025), ordinary-NTK PD	Sequential infinite-width fully connected network with nondegenerate Gaussian initialization and distinct raw inputs. With bias, the activation is continuous, almost everywhere differentiable, and nonpolynomial, with the integrability that keeps the Gaussian kernel expectations finite; without bias, sites must also be nonzero and pairwise nonproportional.	Proves strict PD for ordinary value-observation NTKs under its distinct-input conditions. Differential, weak, and nonlocal observation rows are outside its scope.	Our result treats operator-transformed and co-located rows through their coefficient rank and uses compact initialization to ensure the required moments are finite on compact input and parameter sets.
Gao et al. (2023), population PD and finite-width GD	Biased scalar two-layer cubic-ReLU network; Gaussian hidden and Rademacher output initialization; both layers trained. The joint heat-residual, Dirichlet, and initial sites are pairwise distinct, and the boundary weight is positive.	Establishes positivity of both population Gram blocks for the specified heat-equation design and gives GF and GD decay after a finite-width transfer.	A fourth-order strong row is identically zero for cubic ReLU away from its kink, whereas Proposition 3.1 under P_{pd} covers every fixed finite order; see Example C.1.
Gan et al. (2025), operator-informed NTK	Bias-free scalar deep network, sequential widths, one fixed linear differential operator on a compact convex set, continuous coefficients, and a C^{k+2} activation for an operator of order k, with polynomially bounded derivatives. The training theorem assumes a positive limiting Gram and uniform gradient continuity.	Proves uniform convergence in probability of the transformed kernel. Its training statement assumes limiting-Gram positivity and uniform gradient continuity; it does not characterize rank for arbitrary joint constraint rows.	Duplicate compatible rows have zero full gap but a positive one-dimensional compression. Joint residual–boundary families, weak rows, and nonlocal rows also lie outside the single-operator theorem.
Jin & Wu (2026a), qualitative DNTK PD	Shallow or deep fully connected networks with standard-Gaussian parameters and positive bias scale; pairwise-distinct physical sites; one nonzero evaluated scalar differential row per site; smooth nonpolynomial or RePU activation with order-matched regularity.	Theorem 2 establishes infinite-width covariance and DNTK positivity under its distinct-site assumptions. Exact nullspaces for arbitrary repeated-site and coupled-output rows are outside the stated setting.	Co-located value/derivative rows can be independent, and coupled-output rows can be dependent. Our local map gives both the exact PD criterion and the nullspace; see Examples C.2 and C.3.
Jin & Wu (2026b), extreme DNTK eigenvalues	Bias-free scalar two-layer RePU, Poisson residual plus Dirichlet rows, and the unweighted infinite DNTK. The lower bound needs $d_0 \geq 3$, angular δ-separation of all nonzero sites, and a RePU power restricted to a range set by that paper's parameters; upper bounds allow $d_0 \geq 2$.	Gives explicit lower and upper extreme-eigenvalue bounds by spherical harmonics under the stated separation assumptions. Loss weights require congruence scaling; finite-width optimization is outside its scope.	Repeated sites make the separation parameter vanish. Proposition 3.1 still distinguishes an independent value/derivative pair from a duplicate pair, and Corollary E.4 handles the latter on its active subspace.
Moser et al. (2026), transformed-kernel PSD	A scalar PSD base kernel with the required classical mixed derivatives and a finite matrix of pointwise linear differential operators whose coefficients are finite at the evaluation sites.	Uses reproducing-kernel Hilbert space (RKHS) representers to establish PSD. Strict PD, exact rank, an NTK limit, and optimization are outside the theorem; sharp divergence-form interfaces are explicitly deferred to weak formulations.	Duplicate operators give an exact structural nullspace, and a discontinuous diffusion interface gives a bounded weak row; see Examples C.4 and C.5.
Zhou & Yan (2024), DNTK scaling and freezing	Two-layer pointwise PINNs for continuous, possibly nonlinear differential operators with width scaling $d_1^{-\varrho}$ as in Definition F.1 (denoted N^{-s} in that paper); smooth bounded activation derivatives. The freezing theorem also assumes uniform parameter, operator-derivative, and time-integrated residual bounds.	Identifies homogeneity-dependent scaling thresholds for initialization convergence and proves finite-time kernel freezing for $\varrho > 1/4$ under the displayed trajectory bounds. It neither proves PD/rank nor supplies an optimizer rate.	Two co-located linear rows can be identical or independent while satisfying the same width scaling. Joint independence of their coefficient rows determines their rank, which kernel convergence alone does not.
Panel B: optimization and weak-form comparators			

Continued on the next page

Table 2 continued

Work and result	Setting and assumptions	Main conclusion	Additional scope treated here
Jiang et al. (2026), population training without a uniform gap	Scalar second-order linear PDE with the Dirichlet condition imposed by a lift; population L^2 residual, optionally with finitely many value data; shallow network, C_b^4 nonconstant activation, full-support initialization, and clipped continuous-time GF.	Sends width to infinity at each fixed time and then proves asymptotic residual convergence without a uniform spectral gap; solution convergence additionally assumes a bounded inverse. It gives neither a rate nor a uniform-in-time finite-width discrete-GD transfer.	A finite co-located residual–boundary pair may be singular or PD according to its evaluated coefficients. Our formula classifies both and gives a finite-width GD result, without claiming population solution consistency.
Nießen & Müller (2026), projected GD	Scalar homogeneous-Dirichlet Poisson problem on a bounded $C^{2,\beta}$ domain; one hidden layer; fixed paired output weights; only hidden weights trained; paired truncated-Gaussian initialization; per-neuron projection; and globally bounded activation derivatives through order four.	Theorem 6 gives an $O(T^{-1/2})$ time-averaged empirical-loss bound, together with width and approximation terms, under the stated projected-training setup. The result concerns time-averaged loss rather than last-iterate geometric decay.	Coupled multi-output pointwise residuals with joint boundary rows, trained in both layers by ordinary unprojected GD, satisfy the hypotheses of Proposition 3.1 and Corollary 4.3; see Example J.1.
Jin & Wu (2027), stochastic wide-PINN training	Biased scalar two-layer Poisson PINN at fixed residual and Dirichlet sites; both layers trained; Gaussian/Rademacher initialization; unbiased with-replacement SGD or modeled stochastic GF. The two population parameter blocks are each assumed PD, and the step size is small.	Gives expected loss decay on a no-exit event at sufficient width under the assumed positive population blocks and stated sampling model. It concerns stochastic rather than full-batch training and does not treat compatible singular systems.	Equal duplicate rows make both block Grams singular but leave a positive active compression for compatible targets. Our full-batch result applies after exact compression; see Example C.4.
Xu et al. (2027), implicit GD	Bias-free scalar two-layer PINN for the heat equation; fixed Rademacher output weights, with only Gaussian-initialized hidden weights trained; and pairwise nonparallel training inputs. The analytic nonpolynomial activation has bounded third derivative and globally Lipschitz derivatives through order three. In addition, for every $n \geq 2$, it assumes a common asymptotic profile ϕ with $\sigma^{(n)}(x)/\phi(x) \to c_n \neq 0$, where $\phi(x) \to 0$ and $x\phi(bx)/\phi(x) \to 0$ for every $b > 1$, in that paper's limiting regime. The limiting Gram is PD, and each implicit step is solved exactly.	Proves high-probability linear IGD loss decay for any fixed $\eta > 0$ at sufficient width; the sufficient-width condition depends on η. The guarantee concerns a special pointwise scalar system and exact implicit updates, not ordinary GD or a rank classification.	Co-located independent operators violate the nonparallel-site condition, and compatible duplicate rows have zero full gap. Our local-rank and active-subspace results cover both cases for explicit GD.
Xu et al. (2026), natural GD	Overparameterized scalar two-layer PINNs for the heat-equation model of Xu et al. (2027); fixed Rademacher output weights, with only Gaussian-initialized hidden weights trained. The smooth-activation analysis assumes a PD limiting Gram and regular derivatives, and NGD requires an invertible empirical Gram/Jacobian system along training.	Establishes improved GD step-size dependence and fast NGD convergence, including a quadratic local rate for smooth activations. It does not derive an exact rank/nullspace criterion or cover weak, nonlocal, or compatible singular systems.	A duplicated compatible observation makes the natural-gradient Gram singular but has a positive compressed kernel. Corollary E.4 gives explicit-GD convergence without inverting the full Gram.
Zhao & Luo (2025), admissible-operator and random-feature training	Two-layer tanh models for PINNs and Deep Ritz. The linear result requires an "admissible" pointwise operator that preserves feature independence; the nonlinear results freeze hidden features and use coercivity plus a Łojasiewicz condition for GF or IGD.	Proves convergence for its admissible linear class and convergence to critical points in the nonlinear random-feature setting. The linear admissibility hypothesis supplies the needed nondegeneracy rather than an exact rank formula for arbitrary joint rows.	Repeated-site mixed operators and deliberately dependent rows need not meet admissibility. Our local map says exactly when they are independent and gives the attainable loss for compatible or incompatible dependencies.
Kouma & Bomisso (2026), variational-NTK conditioning	Coupled parabolic VPINN with boundary/initial data imposed by a lift; compact space–time tests; deep vector network. The analysis assumes a frozen, translation-invariant DNTK with Hermitian-PD Fourier transform, frequency localization, an invertible symbol, and large convection.	Theorem 4.1 gives a conditional Schur-complement asymptotic for one frequency under the frozen, translation-invariant DNTK hypotheses. It does not give an exact full-matrix rank characterization or a finite-width optimizer theorem.	Two identical weak tests create a cross-test null vector at every width, which a per-frequency component-symbol calculation does not classify. Our weak nullspace and compatible active-subspace results do.
Panel C: present work			

Continued on the next page

Table 2 continued

Work and result	Setting and assumptions	Main conclusion	Additional scope treated here
This work	Biased two-layer, possibly multi-output networks; a compactly supported per-neuron law whose hidden marginal support has nonempty interior; analytic nonpolynomial activation for the exact pointwise rank/nullspace criterion. Weak rows additionally require Banach-valued analyticity and dense ridge-feature span; nonlocal rows use finite signed-measure tuples.	Gives exact limiting-NTK rank and nullspace formulas, a finite-witness lower bound, and high-probability finite-width explicit-GD convergence with stated width and step size. A fixed active-subspace version covers compatible exact dependencies.	Treats all six constructions below: excessive differential order for a fixed RePU is shown directly to annihilate the row, outside $\mathsf{P}_{\mathrm{reg}}$; dependent or independent same-site rows, coupled outputs, compatible duplicates, sharp-interface weak rows, and activations whose Gaussian moments diverge but whose compact-initialization moments remain finite are classified by the rank formulas.

The table does not compare results whose principal contribution is nonlinear-PDE optimization (Bonfanti et al., 2024; Sirignano et al., 2026), because for nonlinear constraints our contribution is negative rather than comparable: under conditional centering, second-moment isotropy of the output weights, and a certificate for the base observations, Section F.2 shows that a single genuinely nonlinear polynomial row has a nondegenerate random limiting kernel entry, so the deterministic limiting kernel on which every linear-row comparison in this table rests no longer exists, and Section F.3 shows that no width exponent yields a nonzero deterministic limit. The cited analyses reach nonlinear problems by routes that do not pass through such a kernel. The table likewise omits solution consistency and error control (Shin et al., 2020; Cheng et al., 2025; Doumèche et al., 2025; Bonito et al., 2026), and empirical failure or landscape analyses (Krishnapriyan et al., 2021; Rathore et al., 2024). Weak and nonlocal method papers (Kharazmi et al., 2021; Pang et al., 2020) motivate the functional rows but do not make the theorem-level NTK claims compared above. Functional kernel perspectives from collocation, operator observations, and measure embeddings (Schaback & Wendland, 2006; Pförtner et al., 2022; Sriperumbudur et al., 2011) supply broader background rather than direct PINN-optimization comparisons.

Within the omitted consistency category, Cheng et al. (2025) give conditions for consistency of physics-informed kernel GF and wide PINNs, finite-time NTK convergence, and nonasymptotic loss bounds while assuming a PD base NTK. They do not derive finite-data DNTK rank or a discrete-GD finite-width transfer.

The intermediate independence property of Jin & Wu (2026a, Definition 3, Property (LI)) permits distinct multi-index derivative labels at one site. Their main positivity theorem, however, assumes pairwise-distinct physical points and does not provide an arbitrary same-site coefficient-rank criterion or exact nullspace formula (Jin & Wu, 2026a, Theorem 2).

Example C.1 (Cubic ReLU degenerates for a fourth-order strong residual). Let $\rho_3(z) = (z_+)^3$ and consider the one-dimensional equation $u^{(4)}(x) = 1$. For the two-layer trial function

$$u_\theta(x) = \frac{1}{\sqrt{d_1}} \sum_{m=1}^{d_1} a_m \rho_3(w_m x + b_m),$$

every fixed finite collocation set avoids all activation nondifferentiabilities $w_m x_i + b_m = 0$ almost surely under continuous initialization. On that event, each summand either vanishes locally or agrees there with a cubic polynomial, and hence

$$\partial_x^4 u_\theta(x_i) = 0, \qquad \nabla_\theta \partial_x^4 u_\theta(x_i) = \mathbf{0}.$$

Every such residual row and column of the finite-width and expected infinite-width tangent kernels is therefore zero, whereas the target 1 leaves a strictly positive minimum achievable loss. Moreover, ρ_3 is not classically C^3 at its nondifferentiability point, so even third-derivative constraints require additional care. Thus the fixed cubic-ReLU second-order result of Gao et al. (2023) does not extend to arbitrary higher-order strong PDEs without changing the activation or formulation. This example does not contradict Jin & Wu (2026a): their assumptions exclude ρ_3 for fourth derivatives, and their Remark 1 prescribes a matching rectified-power unit instead.

Example C.2 (A shared PDE–boundary point can be singular or benign). Consider the closed interval $[0, 1]$ and collocate at the boundary point $x_\star = 0$ both the PDE operator $\mathcal{A}_\star = \partial_x + x\mathcal{I}$, where $\mathcal{I}$

is the identity operator, and the Neumann boundary operator $\mathcal{B} = \partial_x$. For positive scales s_Ω, s_∂, define $P_\Omega = s_\Omega(\mathcal{A}_\star u_\theta)(x_\star)$ and $P_\partial = s_\partial(\mathcal{B}u_\theta)(x_\star)$. Although $\mathcal{A}_\star \neq \mathcal{B}$ as written operators and both evaluated rows are nonzero,

$$\mathcal{A}_\star u_\theta(0) = \mathcal{B}u_\theta(0) = u'_\theta(0), \qquad s_\partial \nabla_\theta P_\Omega - s_\Omega \nabla_\theta P_\partial = \mathbf{0}.$$

The joint 2×2 Gram block has determinant zero at every width. This singularity shows why pairwise-distinct-point positivity does not automatically apply at a repeated physical location under only a nontriviality check on each constraint. Coincidence is not itself an obstruction: replacing $\mathcal{A}_\star$ by $\partial_x + \mathcal{I}$ gives the independent local coefficient rows $(1, 1)$ and $(0, 1)$ in coordinates (u, u'). Under the hypotheses of Proposition 3.1, this pair yields a PD limiting NTK although the sites coincide. Linear independence of the local coefficient vectors, rather than coincident sampling alone, distinguishes the two cases.

Example C.3 (Coupled vector rows require joint rank). For $\mathbf{u}_\theta = (u_{1,\theta}, u_{2,\theta})^\top$, impose at the same physical point

$$P_1 = u_{1,\theta}(x_\star), \qquad P_2 = u_{2,\theta}(x_\star), \qquad P_3 = u_{1,\theta}(x_\star) + u_{2,\theta}(x_\star).$$

Then $\nabla_\theta P_3 = \nabla_\theta P_1 + \nabla_\theta P_2$, so the joint NTK is singular at every width with null direction $(-1, -1, 1)^\top$. Thus scalar positivity for each output coordinate does not by itself imply positivity of an arbitrary coupled, scalarized multi-output constraint family. Multiple components alone are not the problem: for example, $u_1 + u_2$ and $u_1 - u_2$ have independent coefficient rows. Jin & Wu (2026a) allow vector outputs with operators acting componentwise; the additional issue here is coupled constraints at a repeated site. The local vectors $\mathbf{b}_i(\mathbf{x})$ record both the component and derivative indices, and their joint linear independence is the exact criterion; see also Example J.1.

Example C.4 (Compatible duplicate rows need only an active-subspace gap). Let ℓ be any nonzero pointwise, weak, or nonlocal constraint covered by Propositions 3.1 to 3.3, and duplicate its scaled prediction:

$$P_1(\theta) = P_2(\theta) = \ell(u_\theta).$$

If $g(\theta) = \|\nabla_\theta P_1(\theta)\|_2^2$, then the full tangent kernel is

$$\mathbf{K}(\theta) = g(\theta) \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \qquad \ker \mathbf{K}(\theta) = \operatorname{span}\{(1, -1)^\top\} \text{ whenever } g(\theta) > 0.$$

Thus every theorem that assumes a strictly positive full Gram matrix excludes this system. Equal targets $(c, c)^\top$ are nevertheless compatible: the error remains in $\mathcal{U}_{\text{dup}} = \operatorname{span}\{(1, 1)^\top\}$, and the compressed kernel on $\mathcal{U}_{\text{dup}}$ is positive whenever $g(\theta) > 0$. Under the applicable regularity and width hypotheses, Corollary E.4 therefore gives geometric finite-width GD convergence. For unequal targets $(c_1, c_2)^\top$, the unavoidable minimum of the objective $\|\mathbf{v}_P - \mathbf{v}_T\|_2^2$ is $(c_1 - c_2)^2/2 = \|\Pi_{\mathcal{N}_{\text{str}}} \mathbf{v}_T\|_2^2$, the irreducible term in Equation (52). This single construction applies, for example, to duplicate collocation rows and to duplicate weak test functions.

Example C.5 (A sharp interface is strong-form singular but weak-form regular). On $(-1, 1)$, let

$$\kappa(x) = \begin{cases} 1, & x < 0, \\ 2, & x \geq 0. \end{cases}$$

The divergence-form operator $-(\kappa u')'$ cannot be expanded at $x = 0$ as a strong-form finite sum of classical derivatives with pointwise coefficients unless an interface interpretation is added. Take the network trial space to be $\mathcal{X} = H^1(-1, 1)$ and the test space to be $H_0^1(-1, 1)$. For $u \in \mathcal{X}$ and $v \in H_0^1(-1, 1)$, the continuum weak row

$$\ell_v(u) = \int_{-1}^{1} \kappa(x) u'(x) v'(x)\, dx$$

is bounded on $\mathcal{X}$. Dirichlet values are imposed by the two separate, bounded trace rows $u \mapsto u(-1)$ and $u \mapsto u(1)$. The distinction between trial and test spaces is essential: an unconstrained tanh ridge $x \mapsto \tanh(wx + \gamma b)$ generally does not belong to $H_0^1(-1, 1)$. Corollary C.19 verifies, rather than assumes, the Banach-valued analyticity, density, and joint independence required for this precise formulation. This example lies outside the strong-form differential-operator theorem of Moser et al. (2026), which explicitly sends sharp interfaces to a weak formulation.

Example C.6 (Initialization law and moment existence). Consider the analytic nonpolynomial activation $\sigma(z) = \exp(z^4)$ and a nondegenerate Gaussian preactivation Z. Then

$$\mathbb{E}[\sigma(Z)^2] = \mathbb{E}[\exp(2Z^4)] = \infty.$$

Consequently, the ordinary infinite-width covariance and NTK expectations are not finite for this activation under Gaussian initialization. This illustrates why population-kernel statements require moment existence in addition to activation regularity and nonpolynomiality. Our compact per-neuron initialization ensures the relevant moments are finite on compact input and parameter sets; in particular, one nonzero value row has a finite positive limiting Gram entry under our assumptions.

Five of these six cases illustrate boundaries among theorem settings rather than counterexamples to the cited results, and Example C.6 shows only that moment existence is part of any Gaussian-initialization statement. Together they show why activation regularity must match differential order, why co-located and coupled rows require a joint rank test, why full-Gram positivity is unnecessary for compatible training, why sharp interfaces call for a weak formulation, and why population-kernel statements need moment existence in addition to activation regularity.

C.2 Deterministic limit, finite-width transfer, and local stability

Theorem C.7 (NTK analysis for finite linear constraints)**.** *Use the network and loss in Definitions 2.1 and 2.3, and suppose the row-regularity condition in Definition B.1 holds. Then:*

1. *At initialization, the deterministic infinite-width NTK exists and has the Gram representation*

$$[\mathbf{K}^{\infty}_{(0)}]_{ij} = \mathbb{E}_{z\sim\mu}\left[\langle \nabla h_i(z), \nabla h_j(z)\rangle\right] = \sum_{p=1}^{d_\theta} \mathbb{E}_{z\sim\mu}\left[\partial_{z_p} h_i(z)\, \partial_{z_p} h_j(z)\right].$$

   Thus it is the sum of $d_\theta = d_0 + d_2 + 1$ positive-semidefinite contributions, and bounded continuous parameter-derivative functions in an $L^2(\mu)$ space define each contribution. If every width-d_1 network uses the first d_1 vectors of one infinite i.i.d. sequence, then $\mathbf{K}_{(0)} \to \mathbf{K}^{\infty}_{(0)}$ almost surely; for separately resampled widths, convergence is in probability. For finite-order pointwise differential constraints, these parameter-derivative functions are analytic when σ is real analytic.

2. *Suppose the fixed design has a positive-definiteness certificate in the sense of Definition 2.7, so $\mathbf{K}^{\infty}_{(0)} \succ 0$, and let $0 < \lambda_0 \le \lambda_{\min}(\mathbf{K}^{\infty}_{(0)})$ be a certified lower bound. Define*

$$c_{Ch} = \frac{1}{4} + \frac{3}{4}\log\frac{3}{4}.$$

   For $0 < \delta < 1$, if

$$d_1 \ge \frac{R_K}{c_{Ch}\lambda_0}\log\left(\frac{N}{\delta}\right),$$

   then $\lambda_{\min}(\mathbf{K}_{(0)}) \ge 3\lambda_0/4$ with probability at least $1-\delta$. For the pointwise class under clause P_{pd} of Assumption 2.6, sitewise coefficient-vector independence gives the required positive-definiteness certificate through Lemma C.8, Proposition 3.1, and Section D.3.1.

3. *For every $\theta \in \mathcal{T}_R(\theta_{(0)})$, direct perturbation of the prediction Jacobian gives*

$$\lambda_{\min}(\mathbf{K}(\theta)) \ge \left(\sqrt{\lambda_{\min}(\mathbf{K}_{(0)})} - \frac{L_*}{\sqrt{d_1}}\|\theta - \theta_{(0)}\|_2\right)_+^2,$$

   and $\lambda_{\max}(\mathbf{K}(\theta)) \le B_*^2$.

The abstract proof appears in Theorem B.9, and Proposition B.7 verifies its per-neuron regularity hypotheses for the stated finite-order strong-form differential constraints. Matrix Chernoff yields the second conclusion, and singular-value perturbation of the Jacobian yields the third.

C.3 INDEPENDENCE LEMMA, FINITE-WITNESS BOUND, AND AN EXPLICIT EXAMPLE

The local rank characterization in Proposition 3.1 uses the following specialization of Jin & Wu (2026a, Theorem 4).

Lemma C.8 (Independence of activation-derivative polynomial functions)**.** *Let $\sigma \in C^\infty(\mathbb{R})$ be nonpolynomial, and let $\mathbf{x}_1, \dots, \mathbf{x}_M \in \mathbb{R}^{d_0}$ be pairwise distinct. For each $r \in [M]$, let $J_r \subset \mathbb{N}_0^{d_0}$ be finite and nonempty and let $(\kappa_{r,\alpha})_{\alpha \in J_r}$ be nonzero coefficients. Define*

$$Q_r(\mathbf{w}, v) = \sum_{\alpha \in J_r} \kappa_{r,\alpha} \sigma^{(|\alpha|)}(\mathbf{w}^\top \mathbf{x}_r + v)\mathbf{w}^\alpha. \tag{24}$$

Then $Q_1, \dots, Q_M$ are linearly independent as functions on $\mathbb{R}^{d_0+1}$.

Proof. Every summand in Equation (24) has differential degree equal to the degree of its $\mathbf{w}$-factor. Thus, after combining like terms, every Q_r is built from products of activation derivatives and monomials in $\mathbf{w}$, as required by Theorem 4 of Jin & Wu (2026a), which states that if such a combination with nonzero coefficients at pairwise distinct sites vanishes identically on $\mathbb{R}^{d_0+1}$, then σ is a polynomial. A nonzero dependence, after omitting indices with zero outer coefficients, would satisfy those hypotheses and would force σ to be a polynomial, a contradiction. □

The exact rank results are qualitative. The next proposition gives a generic finite-witness bound from finitely many values of the output feature matrix $\mathbf{F} = \mathbf{\Phi}_{\text{out}}$ defined below and its variation on sets of positive initialization probability. Unlike the Poisson-specific bounds of Jin & Wu (2026b), it is not a mesh-uniform or separation-asymptotic estimate.

Proposition C.9 (Finite-witness lower bound for the smallest eigenvalue)**.** *Use the two-layer network in Definition 2.1, with its output weights trainable, and let $\lambda_1, \dots, \lambda_N$ be any fixed linear observations on a common function space containing the coordinate ridge features. Set $P_i(\theta) = \lambda_i(\mathbf{u}(\cdot;\theta))$ and write $\mathbf{F}(\boldsymbol{\xi}) = \mathbf{\Phi}_{\text{out}}(\boldsymbol{\xi}) \in \mathbb{R}^{N \times d_2}$ for the output feature matrix of the main text,*

$$F_{iq}(\boldsymbol{\xi}) = \lambda_i(\mathbf{e}_q \sigma_{\mathbf{w},b}), \qquad \boldsymbol{\xi} = (\mathbf{w}, b), \qquad \sigma_{\mathbf{w},b}(\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + \gamma b).$$

Assume $F_{iq} \in L^2(\mu_h)$ for every i, q and that the full limiting NTK over all trainable parameters exists. The output-weight contribution is then $\mathbf{K}^\infty_{out} = \mathbb{E}_{\boldsymbol{\xi} \sim \mu_h}[\mathbf{F}(\boldsymbol{\xi})\mathbf{F}(\boldsymbol{\xi})^\top]$. Suppose there are pairwise disjoint measurable sets $E_1, \dots, E_L$, parameter values $\boldsymbol{\xi}_\ell$, and numbers $\beta > 0$, $\varepsilon \geq 0$ such that

$$\mu_h(E_\ell) \geq \beta, \qquad \sup_{\boldsymbol{\xi} \in E_\ell} \|\mathbf{F}(\boldsymbol{\xi}) - \mathbf{F}(\boldsymbol{\xi}_\ell)\|_2 \leq \varepsilon.$$

Let $\mathbf{V} = \big(\mathbf{F}(\boldsymbol{\xi}_1)^\top, \dots, \mathbf{F}(\boldsymbol{\xi}_L)^\top\big)^\top \in \mathbb{R}^{Ld_2 \times N}$, and assume $c_{sep} := \sigma_{\min}(\mathbf{V}) = \sqrt{\lambda_{\min}(\mathbf{V}^\top \mathbf{V})} > 0$. Then

$$\lambda_{\min}(\mathbf{K}^\infty_{(0)}) \geq \lambda_{\min}(\mathbf{K}^\infty_{out}) \geq \frac{\beta c_{sep}^2}{2} - \varepsilon^2. \tag{25}$$

In particular, $\varepsilon \leq c_{sep}\sqrt{\beta}/2$ implies $\lambda_{\min}(\mathbf{K}^\infty_{(0)}) \geq \beta c_{sep}^2/4$.

For the pointwise constraints of Definition 2.2, this statement specializes to

$$F_{iq}(\boldsymbol{\xi}) = s_i(\mathcal{A}_{iq}\sigma_{\mathbf{w},b})(\mathbf{x}_i).$$

For the weak and nonlocal rows considered below, respectively, it specializes to

$$F_{iq}(\boldsymbol{\xi}) = s_i a_i(\mathbf{e}_q \sigma_{\mathbf{w},b}, v_i), \qquad F_{iq}(\boldsymbol{\xi}) = s_i \int_D \sigma_{\mathbf{w},b}(\mathbf{y})\, d\nu_{iq}(\mathbf{y}).$$

Thus the same finite-witness lower bound applies to a combined residual–boundary feature matrix in the pointwise, weak, or nonlocal setting. No analyticity or density hypothesis is needed for this quantitative implication once the displayed features and limiting kernels exist.

See the positivity appendix for the proof.

Example C.10 (A fully numerical finite-width GD certificate). This example instantiates every constant in Theorem 4.2 for one continuum nonlocal observation. It is a deliberately conservative certificate, not a claim that the resulting width is necessary or practically optimal. Let $d_0 = d_2 = 1$, $\Omega = (-1, 1)$, $D = \overline{\Omega}$, $\gamma = 1$, $\sigma = \tanh$, $s = 1$, and

$$\lambda(f) = \frac{1}{2}\int_{-1}^{1} f(x)\,dx, \qquad v_T = 0.$$

Thus $N = 1$ and, for $z = (w, b, a)$, the per-neuron prediction is

$$h(z) = aF(w, b), \qquad F(w, b) = \frac{1}{2}\int_{-1}^{1} \tanh(wx + b)\,dx.$$

Initialize the neuron parameters independently across neurons, with mutually independent coordinates

$$w \sim \mathrm{Unif}[-0.1, 0.1], \qquad b \sim \mathrm{Unif}[1, 2], \qquad a \sim \mathrm{Unif}[-0.1, 0.1].$$

The law is width independent and compactly supported, its hidden marginal has support with nonempty interior, and $\mathbb{E}[a \mid w, b] = 0$. Set $R = 0.1$ and $\delta = 0.05$.

We now verify the constants. At initialization, $wx + b \geq 0.9$ for every $x \in [-1, 1]$. Since $\tanh$ is increasing,

$$\begin{aligned}\lambda_{\min}(\mathbf{K}^{\infty}_{(0)}) &= [\mathbf{K}^{\infty}_{(0)}]_{11} = \mathbb{E}\|\nabla h(z)\|_2^2 \\ &\geq \mathbb{E}\big[(\partial_a h(z))^2\big] = \mathbb{E}[F(w, b)^2] \geq \tanh(0.9)^2 =: \underline{\lambda} =: \lambda_0 > 0.51308.\end{aligned} \tag{26}$$

Writing subscripts for derivatives of F, differentiation under the integral gives

$$|F| \leq 1, \qquad |F_w| \leq \frac{1}{2}, \qquad |F_b| \leq 1.$$

Because $\nabla h = (aF_w, aF_b, F)$, the initialization bound $|a| \leq 0.1$ yields

$$R_K \leq 1 + 0.1^2\left(\frac{1}{4} + 1\right) = 1.0125. \tag{27}$$

Every point in a radius-R per-neuron neighborhood satisfies $|a| \leq 0.2$. The same derivative bounds therefore give the valid local choice

$$B_*^2 = 1.05. \tag{28}$$

For completeness, the gradient-Lipschitz bound follows from the Hessian rather than from an unspecified smoothness constant. The sharp global bound

$$\|\tanh''\|_{\infty} = \frac{4}{3\sqrt{3}} =: c_2$$

implies

$$|F_{ww}| \leq \frac{c_2}{3}, \qquad |F_{wb}| \leq \frac{c_2}{2}, \qquad |F_{bb}| \leq c_2.$$

Since

$$\nabla^2 h = \begin{pmatrix} aF_{ww} & aF_{wb} & F_w \\ aF_{wb} & aF_{bb} & F_b \\ F_w & F_b & 0 \end{pmatrix},$$

its operator norm is at most its Frobenius norm, and throughout the local neighborhood

$$\begin{aligned}\|\nabla^2 h\|_2^2 &\leq 0.2^2 c_2^2\left(\frac{1}{9} + \frac{1}{2} + 1\right) + 2\left(\frac{1}{4} + 1\right) \\ &= \frac{30839}{12150} < 2.5382.\end{aligned} \tag{29}$$

Hence $L_*^2 = 2.5382$ is valid. Finally, $A_{0,*} = 0.1$ and independence of a and (w, b) give

$$H_0^2 = \mathbb{E}[h(z)^2] \leq \mathbb{E}[a^2] = \frac{0.1^2}{3}.$$

Consequently, the initialization-residual quantity at $\delta = 0.05$ obeys

$$\overline{r}_{0.05} \le \frac{0.1}{\sqrt{3}} + 0.1\sqrt{2\log 40} =: \widehat{r} < 0.32936. \tag{30}$$

Here $\widehat{r}$ denotes the exact expression on the left of the last inequality; 0.32936 is its display rounding, and the bounds below are evaluated with the exact value (with the rounded value the second bound would read 1356.02 and the integer width below would become 1357).

Substitution of

$$\underline{\lambda} = \tanh(0.9)^2, \quad R_K \le 1.0125, \quad \Lambda = B_*^2 = 1.05, \quad L_*^2 = 2.5382, \quad \overline{r}_{0.05} \le \widehat{r}$$

into the three width terms in Equation (9) gives, using outward-rounded interval evaluation and $c_{\mathrm{Ch}} = \frac{1}{4} + \frac{3}{4}\log\frac{3}{4}$,

$$\frac{1.0125\log 40}{c_{\mathrm{Ch}}\underline{\lambda}} < 212.613,$$
$$(320 + 128\sqrt{6})\frac{2.5382(1.05)\widehat{r}^2}{\underline{\lambda}^3} < 1355.980,$$
$$\frac{16(1.05)\widehat{r}^2}{0.1^2\underline{\lambda}^2} < 692.252,$$

respectively. Therefore the integer width and step-size choices

$$d_1 \ge 1356, \qquad 0 < \eta \le \frac{10}{21} \tag{31}$$

satisfy the theorem. With probability at least 0.95, simultaneously for every $k \ge 0$,

$$\max_{m\in[d_1]} \|\theta_{m,(k)} - \theta_{m,(0)}\|_2 \le 0.1,$$
$$\mathbf{K}_{(k)} \succeq \frac{\tanh(0.9)^2}{2}\mathbf{I}_1,$$
$$\mathcal{L}(\theta_{(k)}) \le \left(1 - \eta\tanh(0.9)^2\right)^k \mathcal{L}(\theta_{(0)}).$$

This calculation is for the exact integral functional, not a quadrature surrogate. Its purpose is to show that the theorem's sufficient conditions can be evaluated end to end; the constants deliberately use elementary global derivative bounds and are not optimized.

C.4 General Gram and direct-sum criteria

These criteria apply when the pointwise rank criterion of Proposition 3.1 is unavailable, for example for a bias-free architecture or a constraint outside the pointwise finite-order setting.

Lemma C.11 (Positive definiteness from independent continuous parameter-derivative functions)**.** *For N sampled scalar rows, let $\Phi_i \in L^2(\mu)$ be real-valued parameter-derivative functions that are continuous on $\mathrm{supp}(\mu)$, and define $[\mathbf{K}]_{ij} = \langle\Phi_i, \Phi_j\rangle_{L^2(\mu)}$. Then $\mathbf{K}$ is positive definite if and only if $\{\Phi_i\}_{i=1}^N$ is linearly independent in $L^2(\mu)$, equivalently on $\mathrm{supp}(\mu)$.*

See the positivity appendix for the proof.

Corollary C.12 (Exact direct-sum criterion for the NTK)**.** *Given the limiting representation in the first item of Theorem C.7, let $\Phi_i^{(p)}$ denote the feature associated with constraint i in the NTK contribution from coordinate p, and define its coordinate Gram matrix by*

$$[\mathbf{K}_{(0)}^{\infty,(p)}]_{ij} = \langle\Phi_i^{(p)}, \Phi_j^{(p)}\rangle_{L^2(\mu)}.$$

Then $\mathbf{K}_{(0)}^{\infty}$ is positive definite if and only if the direct sums of parameter-derivative functions

$$\left(\Phi_i^{(1)}, \ldots, \Phi_i^{(d_\theta)}\right), \qquad i \in [N],$$

are linearly independent. Moreover,

$$\ker\mathbf{K}_{(0)}^{\infty} = \bigcap_{p=1}^{d_\theta} \ker\mathbf{K}_{(0)}^{\infty,(p)}, \qquad \mathbf{K}_{(0)}^{\infty} \succ 0 \iff \bigcap_{p=1}^{d_\theta} \ker\mathbf{K}_{(0)}^{\infty,(p)} = \{0\}.$$

In particular, a positive-definite contribution from one coordinate is sufficient, but not necessary.

Proof. The claim follows by applying Lemma G.1 to the summands in Theorem C.7. □

C.5 Finite-dimensional rank conditions for analytic functionals

Under the nonempty-interior support hypothesis below, analyticity makes the preceding Gram condition equivalent to full row rank for both a finite Taylor-coefficient matrix and one evaluated at finitely many parameter values. Without an a priori degree bound, failure at one fixed truncation does not imply that the NTK is singular.

Proposition C.13 (Taylor-coefficient and evaluation rank characterization)**.** *For the two-layer network in Definition 2.1, with its output weights trainable, let $P_i(\theta) = \lambda_i(\mathbf{u}(\cdot;\theta))$, where the λ_i are linear on a common space containing all network realizations and all single-neuron functions displayed below. Define*

$$F_{iq}(\boldsymbol{\xi}) = \lambda_i\big(\mathbf{e}_q \sigma(\mathbf{w}^\top \cdot + \gamma b)\big), \qquad \boldsymbol{\xi} = (\mathbf{w}, b).$$

Suppose every F_{iq} is real analytic on $\mathbb{R}^{d_0+1}$, the hidden law μ_h has compact support with nonempty interior, and the full limiting NTK exists. Fix $\boldsymbol{\xi}_\star \in \mathbb{R}^{d_0+1}$. For $K \in \mathbb{N}_0$, let $\mathcal{I}_K = \{\boldsymbol{\rho} \in \mathbb{N}_0^{d_0+1} : |\boldsymbol{\rho}| \le K\}$, and define

$$\mathbf{C}_K(\boldsymbol{\xi}_\star) = \left[\frac{1}{\boldsymbol{\rho}!}\partial_{\boldsymbol{\xi}}^{\boldsymbol{\rho}} F_{iq}(\boldsymbol{\xi}_\star)\right]_{\substack{i\in[N] \\ (q,\boldsymbol{\rho})\in[d_2]\times\mathcal{I}_K}}.$$

Then

$$\ker \mathbf{K}^\infty_{out} = \ker \mathbf{K}^\infty_{(0)} = \bigcap_{K\ge 0} \ker \mathbf{C}_K(\boldsymbol{\xi}_\star)^\top. \tag{32}$$

Consequently, the following are equivalent:

- *(i)* $\mathbf{K}^\infty_{(0)} \succ 0$*;*
- *(ii)* $\operatorname{rank} \mathbf{C}_K(\boldsymbol{\xi}_\star) = N$ *for some finite K;*
- *(iii) for some finite parameter set $\Xi = \{\boldsymbol{\xi}^{(1)}, \dots, \boldsymbol{\xi}^{(T)}\}$, the following evaluation matrix has row rank N:*

$$\mathbf{E}_\Xi = \left[F_{iq}(\boldsymbol{\xi}^{(t)})\right]_{i,(t,q)}.$$

Moreover, Ξ may be chosen within any prescribed nonempty open subset of $\mathbb{R}^{d_0+1}$.

Proof. For $\mathbf{c} \in \mathbb{R}^N$, the output-weight Gram representation gives

$$\mathbf{c}^\top \mathbf{K}^\infty_{\text{out}} \mathbf{c} = \sum_{q=1}^{d_2} \int \left|\sum_{i=1}^N c_i F_{iq}(\boldsymbol{\xi})\right|^2 d\mu_h(\boldsymbol{\xi}).$$

If the quadratic form vanishes, continuity forces each such sum to be null on $\operatorname{supp}\mu_h$. Because this set contains an open subset, the identity theorem makes each combination identically zero. This conclusion is equivalent to the vanishing of all Taylor coefficients at $\boldsymbol{\xi}_\star$, proving the output-weight equality in Equation (32).

The same relation among these features also gives, at every finite width and parameter value,

$$\sum_{i=1}^N c_i P_i(\theta) = \frac{1}{\sqrt{d_1}} \sum_{m=1}^{d_1} \sum_{q=1}^{d_2} (\mathbf{a}_m)_q \sum_{i=1}^N c_i F_{iq}(\boldsymbol{\xi}_m) = 0.$$

Differentiation shows that $\mathbf{c}$ annihilates every gradient block. Conversely, the output-weight kernel is a positive-semidefinite summand of the full limiting kernel. The two implications prove the identity in Equation (32).

The columns of the matrices $\mathbf{C}_K(\boldsymbol{\xi}_\star)$, $K \ge 0$, span $\mathbb{R}^N$ exactly when $\bigcap_K \ker \mathbf{C}_K(\boldsymbol{\xi}_\star)^\top = \{0\}$. Since $\operatorname{rank} \mathbf{C}_K(\boldsymbol{\xi}_\star)$ is nondecreasing in K and bounded by N, it stabilizes, so full rank is reached at a finite K. Applying the same left-nullspace argument to the evaluation vectors proves assertion (iii). If their span were deficient on a nonempty open set, a nonzero such combination would vanish there and hence everywhere, contradicting condition (i). □

Corollary C.14 (Rank conditions from monomials and moments)**.** *In the setting of Proposition C.13, suppose that σ is real analytic and nonpolynomial and $\gamma > 0$ (two conditions not contained in clause W_{pd}), and that the maps $\boldsymbol{\xi} \mapsto \mathbf{e}_q \sigma(\mathbf{w}^\top \cdot + \gamma b)$ are analytic in a common Banach trial space $\mathcal{X}$ on which every λ_i is bounded and whose norm convergence implies almost-everywhere pointwise convergence along a subsequence, such as the L^1_{loc}-embedded spaces of clause $\mathsf{W}_{\mathrm{reg}}$ or $C(D;\mathbb{R}^{d_2})$ on a compact D. For $p_\alpha(\mathbf{x}) = \mathbf{x}^\alpha$, choose, for each α, an $r_\alpha \in \mathbb{N}_0$ such that $\sigma^{(|\alpha|+r_\alpha)}(0) \neq 0$. Banach-valued differentiation gives*

$$\partial_{\mathbf{w}}^{\alpha} \partial_b^{r_\alpha} \left[\mathbf{e}_q \sigma(\mathbf{w}^\top \cdot + \gamma b)\right]\big|_{(\mathbf{w},b)=(\mathbf{0},0)} = \gamma^{r_\alpha} \sigma^{(|\alpha|+r_\alpha)}(0) \mathbf{e}_q p_\alpha \in \mathcal{X}.$$

Thus $\mathbf{e}_q p_\alpha \in \mathcal{X}$, and we may define

$$\mathbf{M}_K = \left[\lambda_i(\mathbf{e}_q p_\alpha)\right]_{\substack{i\in[N] \\ q\in[d_2],\ |\alpha|\le K}}.$$

Then

$$\ker \mathbf{K}^{\infty}_{(0)} = \bigcap_{K\ge 0} \ker \mathbf{M}_K^\top, \qquad \mathbf{K}^{\infty}_{(0)} \succ 0 \iff \operatorname{rank} \mathbf{M}_K = N \text{ for some finite } K. \tag{33}$$

For weak functionals $\lambda_i(f) = s_i a_i(f, v_i)$ and integral observations, respectively, the entries become

$$[\mathbf{M}_K^{weak}]_{i,(q,\alpha)} = s_i a_i(\mathbf{e}_q p_\alpha, v_i), \tag{34}$$

$$[\mathbf{M}_K^{nl}]_{i,(q,\alpha)} = s_i \int_D p_\alpha(\mathbf{y})\, d\nu_{iq}(\mathbf{y}). \tag{35}$$

On compact D, the nonlocal measure tuples are linearly independent if and only if $\mathbf{M}_K^{nl}$ is surjective onto $\mathbb{R}^N$ for some finite K. If instead every ν_{iq} is a combination of Dirac masses at a common finite set of distinct nodes $\mathbf{y}_1, \dots, \mathbf{y}_A$, $\nu_{iq} = \sum_{a=1}^{A} m_{iqa} \delta_{\mathbf{y}_a}$, then independence of the loss-scaled measure tuples is equivalent to row rank N of the Dirac-coefficient matrix

$$\mathbf{C}^{Dirac} = \left[s_i m_{iqa}\right]_{i,(q,a)}. \tag{36}$$

Include all residual and boundary rows jointly in these matrices.

Proof. At $\boldsymbol{\xi}_\star = (\mathbf{0}, b_\star)$, boundedness of λ_i and Banach-valued analyticity allow differentiation through the functional. For every $r \in \mathbb{N}_0$,

$$\partial_{\mathbf{w}}^{\alpha} \partial_b^{r} F_{iq}(\mathbf{0}, b_\star) = \gamma^r \sigma^{(|\alpha|+r)}(\gamma b_\star) \lambda_i(\mathbf{e}_q p_\alpha). \tag{37}$$

At every real center, a nonpolynomial analytic function has nonzero derivatives of arbitrarily high order. Hence, for each α, some $r \in \mathbb{N}_0$ makes the prefactor in Equation (37) not vanish. The Taylor-coefficient vectors and $[\lambda_i(\mathbf{e}_q p_\alpha)]_{i=1}^N$ therefore span the same subspace of $\mathbb{R}^N$. Equivalently, their common left nullspaces coincide. Proposition C.13 then proves Equation (33) and the two displayed specializations.

For compact D, the Stone–Weierstrass theorem makes polynomial restrictions dense in $C(D)$, so a finite signed regular Borel measure that annihilates all polynomials annihilates $C(D)$ and vanishes by Riesz uniqueness. Thus the infinite moment matrix has row rank N exactly when the loss-scaled tuples $\bar{\boldsymbol{\nu}}_1, \dots, \bar{\boldsymbol{\nu}}_N$ admit no nonzero linear relation; finite dimensionality makes the rank visible at some finite K. For the discrete measures above, distinct Dirac masses are linearly independent, giving Equation (36). □

The same differentiation identity verifies the analyticity and density clauses of W_{pd} on Sobolev trial spaces in every dimension, so the weak class is not restricted to one spatial dimension.

Lemma C.15 (Ridge analyticity and density in Sobolev spaces, any dimension)**.** *Let $\Omega \subset \mathbb{R}^{d_0}$ be a bounded open set with the segment property (for instance a bounded Lipschitz domain), $k \in \mathbb{N}_0$, and $\mathcal{X} = H^k(\Omega;\mathbb{R}^{d_2})$. Let σ be real analytic and nonpolynomial and $\gamma > 0$. Then every map $\boldsymbol{\xi} \mapsto \varphi_{\boldsymbol{\xi},q} = \mathbf{e}_q \sigma(\mathbf{w}^\top \cdot + \gamma b)$ is real analytic from $\mathbb{R}^{d_0+1}$ into $\mathcal{X}$, and*

$$\overline{\operatorname{span}\{\varphi_{\boldsymbol{\xi},q} : \boldsymbol{\xi} \in \mathbb{R}^{d_0+1},\ q \in [d_2]\}}^{\mathcal{X}} = \mathcal{X}.$$

Consequently the analyticity and density conditions displayed in clause W_{pd} of Assumption 2.6 hold for this trial space in every dimension d_0; the remaining condition of W_{pd}, $\operatorname{int}(\operatorname{supp}\mu_h) \neq \varnothing$, concerns μ_h alone. Clause $\mathsf{W}_{\mathrm{reg}}$ is then verified for a given finite family of rows by Corollary B.5 as soon as every row is bounded on $\mathcal{X}$.

Proof. Analyticity. Fix $\boldsymbol{\xi}_0 = (\mathbf{w}_0, b_0)$. The arguments $t_0(\mathbf{x}) = \mathbf{w}_0^\top \mathbf{x} + \gamma b_0$, $\mathbf{x} \in \overline{\Omega}$, lie in a compact interval I. A real analytic σ extends holomorphically to a complex neighborhood of I, so the Cauchy estimates give $\sup_{t\in I} |\sigma^{(n)}(t)| \le C\, n!\, \rho^{-n}$ for some $C, \rho > 0$ and all n. For $\boldsymbol{\xi} = \boldsymbol{\xi}_0 + \boldsymbol{\delta}$ with $\boldsymbol{\delta} = (\boldsymbol{\delta}_{\mathbf{w}}, \delta_b)$, Taylor's theorem in the scalar argument gives

$$\sigma(\mathbf{w}^\top \mathbf{x} + \gamma b) = \sum_{n\ge 0} \frac{1}{n!}\, \sigma^{(n)}(t_0(\mathbf{x}))\big(\boldsymbol{\delta}_{\mathbf{w}}^\top \mathbf{x} + \gamma \delta_b\big)^n.$$

The n-th term is a homogeneous polynomial of degree n in $\boldsymbol{\delta}$ with coefficients in $C^k(\overline{\Omega})$: differentiating it $j \le k$ times in $\mathbf{x}$ produces $\sigma^{(n+i)}(t_0(\mathbf{x}))$, $i \le j$, times polynomials in $\boldsymbol{\delta}_{\mathbf{w}}$ and in $\boldsymbol{\delta}_{\mathbf{w}}^\top \mathbf{x} + \gamma \delta_b$ of total degree n, so by the Cauchy estimates its $C^k(\overline{\Omega})$ norm is at most $C'(n+k)^k(\Lambda_\Omega \|\boldsymbol{\delta}\|_2/\rho)^n$ with $\Lambda_\Omega = \sup_{\mathbf{x}\in\overline{\Omega}} \|(\mathbf{x}, \gamma)\|_2$. The series therefore converges absolutely in $C^k(\overline{\Omega})$, hence in $H^k(\Omega)$ on the bounded set Ω, for $\|\boldsymbol{\delta}\|_2 < \rho/\Lambda_\Omega$, which is real analyticity into $\mathcal{X}$.

Density. Let V be the closed linear span. Derivatives of a Banach-space-valued analytic map at a point are limits of finite linear combinations of its values, hence lie in V. By the identity in Corollary C.14, $\partial_{\mathbf{w}}^\alpha \partial_b^{r_\alpha} \varphi_{\boldsymbol{\xi},q}|_{\boldsymbol{\xi}=\mathbf{0}} = \gamma^{r_\alpha} \sigma^{(|\alpha|+r_\alpha)}(0)\, \mathbf{e}_q \mathbf{x}^\alpha$, and because σ is nonpolynomial, for each α some r_α has $\sigma^{(|\alpha|+r_\alpha)}(0) \ne 0$: otherwise the Taylor series of σ at 0 would terminate and σ would be a polynomial by analyticity. Thus V contains every vector monomial $\mathbf{e}_q \mathbf{x}^\alpha$, hence all $\mathbb{R}^{d_2}$-valued polynomials. Polynomials are dense in $H^k(\Omega)$: restrictions to Ω of functions in $C_c^\infty(\mathbb{R}^{d_0})$ are dense in $H^k(\Omega)$ for domains with the segment property (Adams & Fournier, 2003, Theorem 3.22), and any such function is approximated in C^k on a compact box containing $\overline{\Omega}$ by polynomials (Weierstrass approximation of the function and its derivatives, for instance by Bernstein polynomials), hence in $H^k(\Omega)$. So $V = \mathcal{X}$. □

Remark C.16 (Quadrature and numerical rank checks). For a quadrature-based weak loss, combine repetitions and enumerate the resulting data as $(q_a, \mathbf{x}_a, \alpha_a) \in [d_2] \times \overline{\Omega} \times \mathbb{N}_0^{d_0}$, $a \in [A]$. Define

$$\mathcal{Q}_a(f) = \partial^{\alpha_a} f_{q_a}(\mathbf{x}_a), \qquad \mathbf{B} = [B_{ia}] \in \mathbb{R}^{N\times A}, \qquad \lambda_i = \sum_{a=1}^{A} B_{ia}\, \mathcal{Q}_a. \tag{38}$$

If $\{\mathcal{Q}_a\}_{a=1}^A$ is linearly independent in $\mathcal{X}^*$, then

$$\begin{aligned} \sum_{i=1}^{N} c_i \lambda_i = 0 \quad (\mathbf{c} \in \mathbb{R}^N) \quad &\Longleftrightarrow \quad \mathbf{c}^\top \mathbf{B} = 0, \\ \{\lambda_i\}_{i=1}^N \text{ is linearly independent} \quad &\Longleftrightarrow \quad \operatorname{rank} \mathbf{B} = N. \end{aligned} \tag{39}$$

The stated equivalence relies on this property in $\mathcal{X}^*$; distinct triples alone need not guarantee it for an arbitrary function space. For example, an architecture that enforces exact boundary conditions can create linear relations among the restrictions of the $\mathcal{Q}_a$ to that class. A symbolic nonzero $N \times N$ minor, or a rigorous interval bound containing its determinant that excludes zero, proves full row rank; a floating-point singular value is only a numerical diagnostic. Rescaled orthogonal polynomials are preferable to raw monomials in computations. Neither Equations (33) and (39) provides an a priori truncation degree or a mesh-uniform positive lower bound on the smallest eigenvalue, and $\operatorname{rank} \mathbf{B} = N$ concerns the discretized functionals rather than the continuum weak problem. The variation assumptions on sets of nonzero measure in Proposition C.9 yield the explicit estimate stated there.

C.6 Weak residuals and boundary constraints

Standard VPINN and hp-VPINN formulations enforce finitely many fixed test functionals (Kharazmi et al., 2021), whereas weak adversarial networks learn the tests during optimization (Zang et al., 2020). The theorem below concerns a fixed finite family. The output-weight block is a Gram matrix of functional ridge features; analyticity and density convert its kernel into exact relations among the observations. This viewpoint parallels generalized kernel collocation and bounded operator observations in physics-informed Gaussian processes (Schaback & Wendland, 2006; Pförtner et al., 2022). Recent variational-NTK analysis for coupled parabolic systems instead studies screening and conditioning for a specific matrix-valued weak formulation (Kouma & Bomisso, 2026).

Theorem C.17 (NTK existence and exact nullspace for weak-form constraints)**.** *Let $\mathcal{X}$ be a Banach space over $\mathbb{R}$ of $\mathbb{R}^{d_2}$-valued functions on $\overline{\Omega}$, continuously embedded in L^1_{loc} as in clause $\mathsf{W}_{\mathrm{reg}}$. For $i \in [N]$, let $\mathcal{V}_i$ be normed over $\mathbb{R}$, let $a_i : \mathcal{X} \times \mathcal{V}_i \to \mathbb{R}$ be a continuous bilinear form, fix $v_i \in \mathcal{V}_i$, and define the weak prediction functional and its loss-scaled version by*

$$\ell_i(f) = a_i(f, v_i), \qquad \lambda_i = s_i \ell_i \in \mathcal{X}^*, \qquad s_i > 0.$$

Use $P_i(\theta) = \lambda_i(\mathbf{u}(\cdot;\theta))$ with the network in Definition 2.1. Assume Assumption 2.5 and the weak clauses of Assumption 2.6 for this $\mathcal{X}$, the combined family $\{\lambda_i\}_{i=1}^N$, and the ridge features $\varphi_{\boldsymbol{\xi},q}$ of Definition 2.1. Define the space of exact relations among the weak constraints by

$$\mathcal{N}_{\textit{weak}} = \left\{ \mathbf{c} \in \mathbb{R}^N : \sum_{i=1}^N c_i \lambda_i = 0 \text{ in } \mathcal{X}^* \right\}. \tag{40}$$

Then $\mathbf{K}^\infty_{(0)}$ exists and $\mathbf{K}_{(0)} \to \mathbf{K}^\infty_{(0)}$ almost surely along nested widths (in probability under fresh resampling). Moreover,

$$\ker \mathbf{K}^\infty_{\textit{out}} = \ker \mathbf{K}^\infty_{(0)} = \mathcal{N}_{\textit{weak}}, \qquad \operatorname{rank} \mathbf{K}^\infty_{(0)} = \dim \operatorname{span}\{\lambda_1, \dots, \lambda_N\}. \tag{41}$$

Consequently, this NTK is positive definite if and only if the loss-scaled weak functionals are linearly independent in $\mathcal{X}^*$. In particular, suppose $\mathcal{V}_i = \mathcal{X}$, all $a_i = a$, and $a : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is coercive: $a(v, v) \geq c_a \|v\|^2_{\mathcal{X}}$ for some $c_a > 0$. Then linearly independent test functions $v_1, \dots, v_N$ imply $\mathbf{K}^\infty_{(0)} \succ 0$.

Corollary C.18 (Joint weak-residual and boundary criterion)**.** *Assume Assumption 2.5 and the weak clauses of Assumption 2.6 for the combined prediction vector. Let $\lambda^\Omega_i \in \mathcal{X}^*$, $i \in [N_\Omega]$, be loss-scaled weak PDE-residual functionals, and let $\beta^\partial_j \in \mathcal{X}^*$, $j \in [N_\partial]$, be loss-scaled boundary functionals on the same trial space. Set $N = N_\Omega + N_\partial$. A boundary functional may be a tested boundary-restriction (trace) functional or a strong-form pointwise boundary observation, provided that it is bounded on $\mathcal{X}$. Order the residual rows before the others. Use the combined predictions $P^\Omega_i(\theta) = \lambda^\Omega_i(\mathbf{u}(\cdot;\theta))$ and $P^\partial_j(\theta) = \beta^\partial_j(\mathbf{u}(\cdot;\theta))$, and define*

$$\mathcal{N}^{\textit{weak}}_{\Omega,\partial} = \left\{ (\mathbf{c}, \mathbf{d}) \in \mathbb{R}^{N_\Omega} \times \mathbb{R}^{N_\partial} : \sum_{i=1}^{N_\Omega} c_i \lambda^\Omega_i + \sum_{j=1}^{N_\partial} d_j \beta^\partial_j = 0 \text{ in } \mathcal{X}^* \right\}. \tag{42}$$

We identify $(\mathbf{c}, \mathbf{d})$ with its concatenation in $\mathbb{R}^N$ when comparing the set with a kernel. Then the mixed output-weight and full limiting NTKs obey

$$\ker \mathbf{K}^\infty_{\textit{out}} = \ker \mathbf{K}^\infty_{(0)} = \mathcal{N}^{\textit{weak}}_{\Omega,\partial}, \qquad \operatorname{rank} \mathbf{K}^\infty_{(0)} = \dim \operatorname{span}\Big(\{\lambda^\Omega_i\}_{i=1}^{N_\Omega} \cup \{\beta^\partial_j\}_{j=1}^{N_\partial} \Big).$$

Consequently, this matrix is positive definite if and only if the residual and boundary functionals are jointly *linearly independent. Checking the two families individually is not sufficient in general.*

Proof. Apply Theorem C.17 to the concatenated family $\{\lambda^\Omega_i\}_i \cup \{\beta^\partial_j\}_j$. Any functional $\beta \in \mathcal{X}^*$ not already written as $s\, a(\cdot, v)$, in particular each boundary functional, fits the theorem's bilinear notation by taking $\mathcal{V} = \mathbb{R}$, $a(f, t) = t\beta(f)$, $v = 1$, and loss scale one. The nullspace in Equation (40) is then exactly the set in Equation (42). □

For finite-width optimization, clause $\mathsf{W}_{\mathrm{reg}}$ of Assumption 2.6 and Lemma B.3 supply the bounded and Lipschitz-Jacobian constants required by Theorem 4.2. In the standard bounded-domain H^1 setting, Corollary B.4 verifies this premise directly for $\sigma \in C^3_{\mathrm{loc}}$, and Corollary B.5 (for $\sigma \in C^{k+2}_{\mathrm{loc}}$) together with Lemma C.15 (for real analytic nonpolynomial σ and $\gamma > 0$) verifies the regularity and density clauses on $H^k(\Omega; \mathbb{R}^{d_2})$ over bounded Lipschitz domains in every dimension. Energy spaces outside this Sobolev family still require the trial-space regularity and density checks in the weak clauses of Assumption 2.6; the bounded-functional transfer lemma provides a reusable proof template, but no blanket claim for arbitrary energy spaces is made here.

Corollary C.19 (Continuum weak Poisson problem with a discontinuous coefficient)**.** *Let $\Omega = (-1, 1)$, let $\kappa \in L^\infty(\Omega)$ satisfy $0 < \kappa_- \leq \kappa(x) \leq \kappa_+ < \infty$ almost everywhere, and take the common scalar trial space $\mathcal{X} = H^1(\Omega)$. Fix linearly independent tests $v_1, \dots, v_{N_\Omega} \in H^1_0(\Omega)$ and*

positive row scales s_i, t_-, t_+. Define the continuum weak-residual and soft Dirichlet prediction functionals by

$$\lambda_i^\Omega(f) = s_i \int_{-1}^{1} \kappa(x) f'(x) v_i'(x)\, dx, \qquad \beta_-^\partial(f) = t_- f(-1), \qquad \beta_+^\partial(f) = t_+ f(1). \tag{43}$$

Use the biased scalar two-layer network in Definition 2.1 with $\sigma = \tanh$ and $\gamma > 0$, and assume Assumption 2.5 with $\operatorname{int}(\operatorname{supp} \mu_h) \neq \varnothing$. Then the full limiting NTK for the joint family $\{\lambda_i^\Omega\}_{i=1}^{N_\Omega} \cup \{\beta_-^\partial, \beta_+^\partial\}$ exists and is positive definite; the independence proof below supplies its positive-definiteness certificate, and the monomial procedure in Section D.3.2 gives an alternative finite certificate for any given test family. For every fixed localization radius $R > 0$ the induced per-neuron prediction maps satisfy (G_R) by Corollary B.4. Hence, under the centering condition $\mathbb{E}[a_m \mid w_m, b_m] = 0$, for any fixed targets and $0 < \delta < 1$, the explicit width and step-size conditions in Equation (9) *imply, with probability at least $1 - \delta$, the all-iteration localization, kernel-gap preservation, and geometric empirical-loss decay in Equation* (10)*. In particular, the conclusion applies to the discontinuous coefficient $\kappa = \mathbf{1}_{(-1,0)} + 2\mathbf{1}_{[0,1)}$. For the equation $-(\kappa u')' = g$ with $g \in L^2(\Omega)$ and prescribed endpoint values $u(-1) = g_-$ and $u(1) = g_+$, the loss-scaled targets paired with the rows in Equation* (43) *are $s_i \int_{-1}^{1} g(x) v_i(x)\, dx$, $t_- g_-$, and $t_+ g_+$, respectively.*

Proof. The bilinear form

$$a(f, v) = \int_{-1}^{1} \kappa(x) f'(x) v'(x)\, dx, \qquad (f, v) \in H^1(\Omega) \times H_0^1(\Omega),$$

is continuous by Cauchy–Schwarz, and endpoint evaluation is bounded on $H^1(\Omega)$ by the one-dimensional trace theorem. Thus all rows in Equation (43) belong to $\mathcal{X}^*$.

For each $(w, b) \in \mathbb{R}^2$, the ridge $x \mapsto \tanh(wx + \gamma b)$ belongs to $H^1(\Omega)$. Moreover,

$$(w, b) \longmapsto \tanh(w \cdot + \gamma b) \tag{44}$$

is real analytic as an $H^1(\Omega)$-valued map. Indeed, at any real (w_0, b_0), the holomorphic extension of $\tanh$ has no pole on the compact segment $w_0[-1, 1] + \gamma b_0$; a sufficiently small complex neighborhood of (w_0, b_0) therefore gives a holomorphic map into $C^1([-1, 1]; \mathbb{C})$, and the embedding $C^1([-1, 1]) \hookrightarrow H^1(\Omega)$ is continuous.

The span of the ridge functions $\{\tanh(w \cdot + \gamma b) : (w, b) \in \mathbb{R}^2\}$ is dense in $H^1(\Omega)$. To see this without invoking a Sobolev universal-approximation theorem, fix $n \in \mathbb{N}_0$. Since $\tanh$ is analytic and nonpolynomial, $\tanh^{(n)}$ is not identically zero; choose b_n with $\tanh^{(n)}(\gamma b_n) \neq 0$. Banach-valued differentiation of Equation (44) gives

$$\partial_w^n \tanh(wx + \gamma b_n)|_{w=0} = \tanh^{(n)}(\gamma b_n) x^n.$$

This derivative is an H^1 limit of finite linear combinations of ridge features, so every monomial, and hence every polynomial, lies in the closure of their span. Polynomials are dense in $H^1(-1, 1)$: approximate the weak derivative of an H^1 function in L^2 by a polynomial and integrate that polynomial, matching one endpoint value. Thus the Banach-valued analyticity and ridge-density requirements in clause W_{pd} of Assumption 2.6 hold; its hidden-support requirement is assumed in the statement. For any fixed R, the closed radius-R neighborhood of the compact initialization support is compact. The complete one-neuron map $(w, b, a) \mapsto a \tanh(w \cdot + \gamma b)$ is C^∞ into $H^1(\Omega)$, so boundedness of the finitely many rows makes every induced scalar prediction map $C^{1,1}$ there. Hence clause $\mathsf{W}_{\mathrm{reg}}$ also holds, and Proposition 3.2 gives existence of the population NTK. Once joint independence is established below, conditional centering and Corollary 4.3 give the stated GD conclusion.

It remains to verify joint independence. Suppose, for every $f \in H^1(\Omega)$,

$$\sum_{i=1}^{N_\Omega} c_i \lambda_i^\Omega(f) + d_- \beta_-^\partial(f) + d_+ \beta_+^\partial(f) = 0,$$

and set $v = \sum_i c_i s_i v_i \in H_0^1(\Omega)$. Restricting to $f \in H_0^1(\Omega)$ and then choosing $f = v$ yields $\int_{-1}^{1} \kappa |v'|^2 = 0$. Coercivity and the zero trace imply $v = 0$; linear independence of the v_i and positivity of the s_i give $c_i = 0$ for every i. The remaining relation is $d_- t_- f(-1) + d_+ t_+ f(1) = 0$. Testing with the two affine endpoint cardinal functions gives $d_- = d_+ = 0$. The joint criterion in Corollary C.18 now supplies the claimed positive-definiteness certificate. □

The same route runs in higher dimension, where boundary data enter either as tested traces on H^1 or pointwise on a smoother trial space.

Corollary C.20 (Two-dimensional weak Poisson problem with tested or pointwise boundary data)**.** *Let $\Omega = (-1,1)^2$, let $\kappa \in L^\infty(\Omega)$ satisfy $0 < \kappa_- \le \kappa \le \kappa_+ < \infty$ almost everywhere, and take the scalar trial space $\mathcal{X} = H^1(\Omega)$. Fix linearly independent polynomial tests $v_1, \dots, v_{N_\Omega} \in H^1_0(\Omega)$, for instance the products $(1-x^2)(1-y^2)x^k y^l$ with distinct $(k,l) \in \mathbb{N}_0^2$, linearly independent boundary test functions $\psi_1, \dots, \psi_{N_\partial} \in L^2(\partial\Omega)$, and positive row scales s_i, t_j. Define*

$$\lambda_i^\Omega(f) = s_i \int_\Omega \kappa \, \nabla f \cdot \nabla v_i \, d\mathbf{x}, \qquad \beta_j^\partial(f) = t_j \int_{\partial\Omega} (\operatorname{Tr} f)\, \psi_j \, dS.$$

Use the biased scalar two-layer network of Definition 2.1 with a real analytic nonpolynomial σ and $\gamma > 0$, i.i.d. neurons from a width-independent compactly supported law whose hidden marginal support has nonempty interior. Then the full limiting NTK of the joint family $\{\lambda_i^\Omega\}_{i=1}^{N_\Omega} \cup \{\beta_j^\partial\}_{j=1}^{N_\partial}$ exists and is positive definite, and for every $R > 0$ the rows satisfy (G_R); if moreover $\mathbb{E}[a_m \mid w_m, b_m] = 0$, then Theorem 4.2 applies with its width and step-size conditions. The same conclusions hold with the boundary rows replaced by pointwise Dirichlet rows $\beta_j^\partial(f) = t_j f(\mathbf{x}_j)$ at distinct points $\mathbf{x}_j \in \partial\Omega$, provided the trial space is $\mathcal{X} = H^2(\Omega)$.

Proof. Ω is a bounded Lipschitz domain. The bilinear form $a(f,v) = \int_\Omega \kappa \nabla f \cdot \nabla v$ is continuous on $H^1(\Omega) \times H^1_0(\Omega)$ by Cauchy–Schwarz, and the trace $\operatorname{Tr} : H^1(\Omega) \to L^2(\partial\Omega)$ is bounded, so every row lies in $\mathcal{X}^*$. Lemma C.15 with $k = 1$ gives the analyticity and density clauses of W_{pd}, and Corollary B.5 with $k = 1$ gives the C^2 regularity of $\mathsf{W}_{\mathrm{reg}}$ and hence (G_R); the interior-support clause is assumed. For joint independence, suppose $\sum_i c_i \lambda_i^\Omega + \sum_j d_j \beta_j^\partial = 0$ on $\mathcal{X}$. Every $f \in H^1_0(\Omega)$ has zero trace, so $\sum_i c_i s_i \, a(f, v_i) = 0$ for all $f \in H^1_0(\Omega)$; taking $f = v := \sum_i c_i s_i v_i \in H^1_0(\Omega)$ gives $a(v,v) = 0$, and coercivity of a on $H^1_0(\Omega)$ (from $\kappa \ge \kappa_-$ and the Poincaré inequality) forces $v = 0$, hence $\mathbf{c} = \mathbf{0}$ because the v_i are independent and $s_i > 0$. Then $\sum_j d_j t_j \int_{\partial\Omega} (\operatorname{Tr} f) \psi_j \, dS = 0$ for all $f \in H^1(\Omega)$; since the trace map has range $H^{1/2}(\partial\Omega)$, which is dense in $L^2(\partial\Omega)$, the function $\sum_j d_j t_j \psi_j$ is orthogonal to a dense subspace and vanishes in $L^2(\partial\Omega)$; independence of the ψ_j and $t_j > 0$ give $\mathbf{d} = \mathbf{0}$. Thus Equation (42) is trivial, Corollary C.18 gives $\mathbf{K}^\infty_{(0)} \succ 0$, and Corollary 4.3 applies. For pointwise boundary rows on $H^2(\Omega)$: Corollary B.5 with $k = 2$ (admissible since a real analytic σ lies in C^4_{loc}) gives the C^2 regularity of $\mathsf{W}_{\mathrm{reg}}$ into $H^2(\Omega)$, hence (G_R), and, since $k = 2 > d_0/2$ on the Lipschitz domain Ω, bounded point evaluation; Lemma C.15 with $k = 2$ gives the analyticity and density clauses of W_{pd} in $H^2(\Omega)$; and the residual rows remain bounded on $H^2(\Omega) \hookrightarrow H^1(\Omega)$. In the independence argument, test first on $f \in H^2(\Omega) \cap H^1_0(\Omega)$, which contains the polynomial tests v_i and on which every pointwise boundary row vanishes, since such f is continuous on $\overline{\Omega}$ and its trace is its restriction to $\partial\Omega$; this gives $\mathbf{c} = \mathbf{0}$ as before. Then $\sum_j d_j t_j f(\mathbf{x}_j) = 0$ for all $f \in H^2(\Omega)$, and a polynomial taking prescribed values at the distinct points $\mathbf{x}_j$ gives $\mathbf{d} = \mathbf{0}$. □

For the nonlocal class, under N_{pd}, a two-dimensional instance is immediate: averages over pairwise disjoint closed subsquares of $\overline{\Omega}$ together with Dirac data at distinct points outside those subsquares are measure tuples with pairwise disjoint supports, hence linearly independent, and Proposition 3.3 applies on $C(\overline{\Omega})$; Dirac points inside a subsquare are covered instead by the mutual-singularity argument in the proof of Corollary C.22.

The boundary functionals β_j^∂ admitted by Corollary C.18 are arbitrary bounded functionals on $\mathcal{X}$; the two forms used in this paper are the following. For a scale $t_j > 0$, a specified boundary operator $\mathcal{B}$, and a fixed boundary test function ψ_j such that $f \mapsto \langle \mathcal{B} f, \psi_j \rangle_{\partial\Omega}$ is bounded on $\mathcal{X}$, a tested weak boundary functional has the form $\beta_j^\partial(f) = t_j \langle \mathcal{B} f, \psi_j \rangle_{\partial\Omega}$. A strong-form pointwise row may instead be

$$\beta_j^\partial(f) = t_j \sum_{q=1}^{d_2} \sum_{\alpha \in J_{jq}} b_{jq\alpha} \partial^\alpha f_q(\mathbf{x}_j), \qquad \mathbf{x}_j \in \partial\Omega.$$

Here every $J_{jq} \subset \mathbb{N}_0^{d_0}$ is finite and $b_{jq\alpha} \in \mathbb{R}$. The criterion covers both rows exactly when they are bounded on the common space $\mathcal{X}$. The independence to be checked is that of the induced functionals in $\mathcal{X}^*$, not merely of the written test functions v_i and ψ_j. For a quadrature-based variational loss,

write its distinct derivative evaluations and coefficient matrix as in Equation (38). If the retained $\mathcal{Q}_a$ are linearly independent in $\mathcal{X}^*$, then Equation (39) shows that full row rank of $\mathbf{B}$ is equivalent to independence of the discretized rows. A separate consistency argument is still needed to relate that loss to the continuum weak problem.

C.7 Nonlocal integral constraints and boundary measures

Fractional and nonlocal PINN formulations supply prominent examples of integral or integro-differential residuals (Pang et al., 2019; 2020). For the bounded integral observations treated here, uniform universal approximation turns the functional criterion into an exact statement about the defining measures. The resulting signed-measure independence criterion is related to kernel mean-embedding characterizations (Sriperumbudur et al., 2011), but it is applied here to the output-weight block of the limiting tangent kernel at initialization, whose kernel function is $k(\mathbf{x},\mathbf{y}) = \mathbb{E}_{\boldsymbol{\xi}\sim\mu_h}[\sigma(\mathbf{w}^\top\mathbf{x}+\gamma b)\,\sigma(\mathbf{w}^\top\mathbf{y}+\gamma b)]$. The scope is specifically finite signed regular Borel measures acting on $C(D)$; the theorem below does not, by itself, cover principal-value or hypersingular kernels, distributional derivatives, or continuum Caputo, Riemann–Liouville, or Riesz fractional derivatives. A quadrature formula for one of those operators that uses only point values of the unknown, such as a Grünwald–Letnikov or L1-type rule, produces a finite signed atomic measure and is therefore covered as a *discrete* constraint; a rule that samples derivatives of the unknown is instead a weak-form row on a trial space on which those evaluations are bounded. That conclusion supplies neither quadrature consistency nor a continuum fractional-operator result. A continuum operator outside the finite-measure class must instead be placed in a Banach trial space on which its tested rows are bounded, and the Banach-valued feature regularity, the density clause (or another positive-definiteness certificate), and the optimization row regularity must then be verified separately.

Theorem C.21 (NTK existence and exact nullspace for nonlocal integral constraints)**.** *Let $D \subseteq \overline{\Omega} \subset \mathbb{R}^{d_0}$ be compact and set $\mathcal{X} = C(D;\mathbb{R}^{d_2})$. For every $i \in [N]$ and $q \in [d_2]$, let ν_{iq} be a finite signed regular Borel measure on D, and define*

$$\ell_i(f) = \sum_{q=1}^{d_2} \int_D f_q(\mathbf{y})\, d\nu_{iq}(\mathbf{y}), \qquad \lambda_i = s_i\ell_i \in \mathcal{X}^*, \qquad P_i(\theta) = \lambda_i(\mathbf{u}(\cdot;\theta)), \qquad s_i > 0.$$

Assume Assumption 2.5 and the nonlocal clauses of Assumption 2.6. Set

$$\mathcal{N}_{nl} = \left\{ \mathbf{c} \in \mathbb{R}^N : \sum_{i=1}^{N} c_i s_i \nu_{iq} = 0 \text{ as a signed measure for every } q \in [d_2] \right\}.$$

Then $\mathbf{K}_{(0)}^{\infty}$ exists and $\mathbf{K}_{(0)} \to \mathbf{K}_{(0)}^{\infty}$ almost surely along nested widths (in probability under fresh resampling). Moreover,

$$\ker \mathbf{K}_{out}^{\infty} = \ker \mathbf{K}_{(0)}^{\infty} = \mathcal{N}_{nl}, \qquad \operatorname{rank} \mathbf{K}_{(0)}^{\infty} = \dim \operatorname{span}\{\bar{\boldsymbol{\nu}}_i : i \in [N]\}. \tag{45}$$

Consequently, $\mathbf{K}_{(0)}^{\infty} \succ 0$ if and only if the loss-scaled tuples $\bar{\boldsymbol{\nu}}_i = s_i(\nu_{i1},\dots,\nu_{id_2})$, $i \in [N]$, are linearly independent in $\mathcal{M}(D)^{d_2}$, where $\mathcal{M}(D) = C(D)^*$ is the space of finite signed regular Borel measures on D.

Corollary C.22 (Nonlocal residuals with nonlocal or Dirichlet boundary data)**.** *Assume Assumption 2.5 and the nonlocal clauses of Assumption 2.6 for the combined prediction map. For $i \in [N_\Omega]$, let $\bar{\boldsymbol{\nu}}_i = s_i(\nu_{i1},\dots,\nu_{id_2})$ define a nonlocal PDE-residual row. For $j \in [N_\partial]$, let $\bar{\boldsymbol{\omega}}_j = t_j(\omega_{j1},\dots,\omega_{jd_2})$, $t_j > 0$, be a vector of finite signed regular Borel measures on D, typically supported on $D \cap \partial\Omega$, defining another constraint. Set $N = N_\Omega + N_\partial$ and order the residual rows before the boundary rows. The corresponding predictions are*

$$P_i^{\Omega}(\theta) = s_i \sum_{q=1}^{d_2} \int_D u_q(\mathbf{y};\theta)\, d\nu_{iq}(\mathbf{y}), \qquad P_j^{\partial}(\theta) = t_j \sum_{q=1}^{d_2} \int_D u_q(\mathbf{y};\theta)\, d\omega_{jq}(\mathbf{y}).$$

Set

$$\mathcal{N}_{\Omega,\partial}^{nl} = \{(\mathbf{c},\mathbf{d}) \in \mathbb{R}^{N_\Omega} \times \mathbb{R}^{N_\partial} : \sum_{i=1}^{N_\Omega} c_i s_i \nu_{iq} + \sum_{j=1}^{N_\partial} d_j t_j \omega_{jq} = 0 \text{ as a signed measure for every } q \in [d_2]\}.$$

We identify $(\mathbf{c}, \mathbf{d})$ *with its concatenation in* $\mathbb{R}^N$*. Then*

$$\ker \mathbf{K}^{\infty}_{out} = \ker \mathbf{K}^{\infty}_{(0)} = \mathcal{N}^{nl}_{\Omega,\partial}, \qquad \operatorname{rank} \mathbf{K}^{\infty}_{(0)} = \dim \operatorname{span}\Big(\{\bar{\boldsymbol{\nu}}_i\}_{i=1}^{N_\Omega} \cup \{\bar{\boldsymbol{\omega}}_j\}_{j=1}^{N_\partial} \Big).$$

Thus the mixed limiting NTK is positive definite exactly when the combined family of residual and boundary measure tuples is linearly independent. Taking $\omega_{jq} = \beta_{jq}\delta_{\mathbf{x}_j}$ *with* $\mathbf{x}_j \in D$ *gives the order-zero pointwise row* $t_j \sum_{q=1}^{d_2} \beta_{jq} f_q(\mathbf{x}_j)$*, which is a Dirichlet row for component* q_0 *when* $\beta_{jq} = \mathbb{1}[q = q_0]$*.*

Suppose additionally that a positive regular Borel measure ρ *on* D *satisfies* $\nu_{iq} \ll \rho$ *(absolute continuity with respect to* ρ*) and* $\omega_{jq} \perp \rho$ *(mutual singularity with* ρ*) for every index. Then the combined family is linearly independent if and only if the residual family* $\{\bar{\boldsymbol{\nu}}_i\}_i$ *and boundary family* $\{\bar{\boldsymbol{\omega}}_j\}_j$ *are each linearly independent. Hence separate checks suffice under this mutual-singularity condition.*

Proof. Apply Theorem C.21 to the concatenated collection of measure tuples. For the final assertion, in each output coordinate q the residual part $\zeta_q^{\Omega} = \sum_i c_i s_i \nu_{iq}$ of a null combination is absolutely continuous with respect to ρ and the boundary part $\zeta_q^{\partial} = \sum_j d_j t_j \omega_{jq}$ is singular with respect to ρ. Since $\zeta_q^{\Omega} = -\zeta_q^{\partial}$ is both, it vanishes: a Borel set E with $\rho(E) = 0$ carries all of $|\zeta_q^{\partial}|$, and absolute continuity gives $|\zeta_q^{\Omega}|(E) = 0$. Hence both parts vanish separately, after which within-family independence makes every coefficient zero. The converse is immediate. □

Under Assumption 2.5 and the nonlocal clauses of Assumption 2.6, Proposition B.6 supplies explicit total-variation-based constants for every row in Corollary C.22. Hence joint measure-tuple independence and conditional output-weight centering imply the finite-width GD conclusion in Corollary 4.3. This implication is qualitative until a numerical lower bound on $\lambda_{\min}(\mathbf{K}^{\infty}_{(0)})$ is also certified.

Let ρ_i be a positive regular Borel measure on D, not necessarily the reference measure of Corollary C.22, and fix a site $\mathbf{x}_i \in D$. Define an integral-operator observation by

$$\ell_i(f) = \sum_{q=1}^{d_2} \int_D \mathfrak{k}_{iq}(\mathbf{x}_i, \mathbf{y}) f_q(\mathbf{y})\, d\rho_i(\mathbf{y}).$$

The criterion applies whenever each kernel section $\mathbf{y} \mapsto \mathfrak{k}_{iq}(\mathbf{x}_i, \mathbf{y})$ is ρ_i-integrable, by taking $d\nu_{iq}(\mathbf{y}) = \mathfrak{k}_{iq}(\mathbf{x}_i, \mathbf{y}) d\rho_i(\mathbf{y})$. A bounded-kernel nonlocal-diffusion row, in the scalar case $\ell_i(f) = \int_D \mathfrak{k}_i(\mathbf{x}_i, \mathbf{y})\,(f(\mathbf{x}_i) - f(\mathbf{y}))\, d\rho_i(\mathbf{y})$, is the finite signed measure $\nu_i = \big(\int_D \mathfrak{k}_i(\mathbf{x}_i, \cdot)\, d\rho_i\big)\delta_{\mathbf{x}_i} - \mathfrak{k}_i(\mathbf{x}_i, \cdot)\, \rho_i$ and is covered in the same way; because of its atom it is not absolutely continuous with respect to a reference measure that has no atom at $\mathbf{x}_i$, so for such rows the joint check of Corollary C.22 is required rather than the separate checks.

Remark C.23 (Derivative and hard boundary conditions). A Dirichlet evaluation is integration against a Dirac mass and therefore falls under Corollary C.22. In contrast, a Neumann or Robin normal-derivative term is a distributional derivative, not a finite measure acting on $C(D)$. Combined with finite-measure residual rows, such a row falls instead under Corollary C.18 on a common space $\mathcal{X}$, for instance $C^1(D; \mathbb{R}^{d_2})$, on which both the measure rows and the boundary derivative are bounded; combined with pointwise residual rows it is covered directly by the pointwise clauses of Assumption 2.6. The single-neuron features must also belong to $\mathcal{X}$ and satisfy the stated Banach-valued analyticity and density hypotheses. The weak formulation may alternatively incorporate tested Neumann or Robin data.

If the architecture enforces an essential boundary condition *hard* rather than penalizing boundary-loss rows, one may write $\mathbf{u}_\theta = \mathbf{g} + \mathcal{E}\mathbf{v}_\theta$, with $\mathbf{v}_\theta$ in a Banach space $\widetilde{\mathcal{X}}$. Here $\mathbf{g} \in \mathcal{X}$ is a fixed lifting satisfying the inhomogeneous data. A bounded linear operator $\mathcal{E} : \widetilde{\mathcal{X}} \to \mathcal{X}_0$ maps into a closed homogeneous-boundary tangent subspace $\mathcal{X}_0 \subseteq \mathcal{X}$ with the inherited norm. No additional penalty rows are present. The tangent features become $\mathcal{E}\varphi_{\boldsymbol{\xi},q}$, and the exact criterion concerns the restrictions of the residual functionals to $\mathcal{X}_0$:

$$\ker \mathbf{K}^{\infty}_{\text{out}} = \ker \mathbf{K}^{\infty}_{(0)} = \left\{ \mathbf{c} \in \mathbb{R}^{N_\Omega} : \sum_{i=1}^{N_\Omega} c_i \lambda_i^{\Omega} \Big|_{\mathcal{X}_0} = 0 \text{ in } \mathcal{X}_0^* \right\},$$

provided that Assumption 2.5 holds with $\operatorname{int}(\operatorname{supp}\mu_h) \neq \varnothing$, that the raw features $\varphi_{\boldsymbol{\xi},q}$ belong to $\widetilde{\mathcal{X}}$ and depend real-analytically on their parameters in that space (with boundedness of $\mathcal{E}$ this gives the C^2 regularity of clause $\mathsf{W}_{\mathrm{reg}}$ for the adapted one-neuron map, hence (G_R) by Lemma B.3), and that the adapted family $\mathcal{E}\varphi_{\boldsymbol{\xi},q}$ has dense span in $\mathcal{X}_0$. The fixed lifting changes the target but not the NTK. In particular, unconstrained single-neuron features generally do not belong to a zero-boundary-value trial space, so one cannot simply reuse a density result for the unconstrained features in a hard-boundary construction.

D Alternative positivity checks and singular examples

D.1 Sharpness of the classwise hypotheses

The clauses of Assumptions 2.5 and 2.6 are sufficient conditions, and it is fair to ask whether they are decorative. The six constructions below answer that question. Each deletes or weakens exactly one clause, retains every remaining clause of the same assumption, and verifies the retained clauses explicitly; each then exhibits a system whose rows are linearly independent yet whose limiting NTK is singular, or, in the last case, does not exist. None of them contradicts any result as stated: every construction violates a hypothesis that the corresponding theorem assumes, so that theorem simply does not apply to it. For each example we record which half of the conclusion survives the deletion and which does not.

The first two examples work by shrinking the closed span of the features the network can actually reach, so that a nonzero combination of independent rows annihilates all of them.

Example D.1 (The ridge-density clause of W_{pd} is load-bearing). Delete from clause W_{pd} of Assumption 2.6 only the ridge-density requirement

$$\overline{\operatorname{span}\{\varphi_{\boldsymbol{\xi},q} : \boldsymbol{\xi} \in \mathbb{R}^{d_0+1},\ q \in [d_2]\}}^{\mathcal{X}} = \mathcal{X},$$

and keep every other hypothesis of Proposition 3.2 and Theorem C.17. Let $\Omega = (0,1)$, so $\overline{\Omega} = [0,1]$, let $d_0 = d_2 = 1$, and let $\mathcal{X} = C([0,1])$ with the supremum norm. Use the network of Definition 2.1 with the linear activation $\sigma(z) = z$, bias scale $\gamma = 1$, and i.i.d. per-neuron initialization $\theta_m = (w_m, b_m, a_m) \sim \mathrm{Unif}([-1,1]^3)$. Take the $N = 3$ unit-scaled evaluation rows

$$\lambda_i(f) = f(x_i), \qquad (x_1, x_2, x_3) = (0, 1, \tfrac{1}{2}),$$

which lie in $\mathcal{X}^*$ and are therefore admissible in Theorem C.17 through the reduction $\mathcal{V}_i = \mathbb{R}$, $v_i = s_i = 1$, $a_i(f,t) = t\lambda_i(f)$ used in the proof of Corollary C.18. Positive loss scales s_i would only replace the null vector $\mathbf{c}$ below by $(c_i/s_i)_{i=1}^3$.

Every retained clause holds. The embedding $\mathcal{X} \hookrightarrow L^1(0,1)$ is continuous, each $\lambda_i \in \mathcal{X}^*$, and the one-neuron map $\mathcal{G}(\boldsymbol{\xi}, \mathbf{a}) = a(w \cdot + \gamma b)$ is a quadratic polynomial map $\mathbb{R}^3 \to \mathcal{X}$, hence C^∞ and real analytic on all of $\mathbb{R}^3$; clause $\mathsf{W}_{\mathrm{reg}}$ and the feature-analyticity part of W_{pd} therefore hold. The law μ is width independent and compactly supported, so Assumption 2.5 holds; $\operatorname{int}(\operatorname{supp}\mu_h) = (-1,1)^2 \neq \varnothing$, so the support part of W_{pd} holds; and $\mathbb{E}[a_m \mid w_m, b_m] = 0$, so the conditional output-weight centering that Assumption 2.6 imposes for gradient descent holds as well. Only density fails, and it fails by a definite margin: with $d_2 = 1$ the single coordinate feature is $\varphi_{(w,b)} = w \cdot + \gamma b$, so $\operatorname{span}\{\varphi_{\boldsymbol{\xi}} : \boldsymbol{\xi} \in \mathbb{R}^2\} = \operatorname{span}\{1, x\} =: \mathcal{P}_1$, a two-dimensional and therefore closed proper subspace of $\mathcal{X}$.

The three rows are linearly independent in $\mathcal{X}^*$: testing $\sum_i c_i\lambda_i = 0$ against continuous hat functions f_j with $f_j(x_j) = 1$ and $f_j(x_k) = 0$ for $k \neq j$ gives $c_j = 0$ for each j. Independence notwithstanding, $\mathbf{c} = (1, 1, -2)^\top$ and $\lambda_{\mathbf{c}} = \lambda_1 + \lambda_2 - 2\lambda_3$ satisfy

$$\lambda_{\mathbf{c}}(\alpha + \beta x) = \alpha + (\alpha + \beta) - 2\left(\alpha + \tfrac{\beta}{2}\right) = 0 \qquad \text{for all } \alpha, \beta \in \mathbb{R},$$

so the nonzero functional $\lambda_{\mathbf{c}}$ annihilates $\mathcal{P}_1$ and hence every ridge feature. The same vector quantifies the failure of density: $\lambda_{\mathbf{c}}(x^2) = \frac{1}{2}$ and $\|\lambda_{\mathbf{c}}\|_{\mathcal{X}^*} = 4$ give $\inf_{p \in \mathcal{P}_1} \|x^2 - p\|_{\mathcal{X}} \geq \frac{1}{8}$, which is exactly the Chebyshev distance attained by $x - \frac{1}{8}$.

Because $\mathbf{u}(\cdot;\theta) \in \mathcal{P}_1$ at every width and every parameter value,

$$\sum_{i=1}^{3} c_i P_i(\theta) = \lambda_{\mathbf{c}}(\mathbf{u}(\cdot;\theta)) = 0, \qquad \text{hence} \qquad \mathbf{J}(\theta)^\top \mathbf{c} = \mathbf{0}, \quad \mathbf{K}(\theta)\mathbf{c} = \mathbf{0}.$$

That is the one-line reason. Explicitly, with $h_i(w, b, a) = a(wx_i + \gamma b)$ in Equation (12), the Gram representation in Theorem C.7 gives

$$[\mathbf{K}^{\infty}_{(0)}]_{ij} = \mathbb{E}\big[a^2\big]\big(x_i x_j + \gamma^2\big) + \mathbb{E}[(wx_i + \gamma b)(wx_j + \gamma b)] = \tfrac{2}{3}\,(1 + x_i x_j)\,,$$

using $\mathbb{E}[a^2] = \mathbb{E}[w^2] = \mathbb{E}[b^2] = \frac{1}{3}$ and $\mathbb{E}[wb] = 0$. Writing $\mathbf{V}_{i,:} = (1, x_i)$, this says

$$\mathbf{K}^{\infty}_{(0)} = \tfrac{2}{3}\mathbf{V}\mathbf{V}^{\top}, \qquad \mathbf{K}^{\infty}_{\text{out}} = \tfrac{1}{3}\mathbf{V}\mathbf{V}^{\top},$$

so both matrices have rank 2 and kernel $\operatorname{span}\{(1, 1, -2)^{\top}\}$.

The weakened statement therefore fails. The relation space in Equation (40) is $\mathcal{N}_{\text{weak}} = \{\mathbf{0}\}$, whereas $\ker \mathbf{K}^{\infty}_{\text{out}} = \ker \mathbf{K}^{\infty}_{(0)} \neq \{\mathbf{0}\}$ and $\operatorname{rank} \mathbf{K}^{\infty}_{(0)} = 2$, which is smaller than $\dim \operatorname{span}\{\lambda_1, \lambda_2, \lambda_3\} = 3$. Both identities in Equation (41) fail; equivalently, the rows are linearly independent while $\mathbf{K}^{\infty}_{(0)} \not\succ 0$. The clause as actually stated excludes the construction, since $\mathcal{P}_1$ is not dense in $\mathcal{X}$, so Proposition 3.2 and Theorem C.17 are untouched.

Three scoping remarks. First, only one inclusion is lost. Density is consumed exactly once, in the step $\ker \mathbf{K}^{\infty}_{\text{out}} \subseteq \mathcal{N}_{\text{weak}}$. The opposite inclusion $\mathcal{N}_{\text{weak}} \subseteq \ker \mathbf{K}^{\infty}_{(0)}$ follows by differentiating $\sum_i c_i P_i(\theta) = 0$ and, for $\mathbf{K}^{\infty}_{\text{out}}$, is immediate from Equation (86); it survives the deletion, as do the resulting bound $\operatorname{rank} \mathbf{K}^{\infty}_{(0)} \leq \dim \operatorname{span}\{\lambda_i\}_{i=1}^{N}$ and the existence and convergence conclusions, which use only clause W_{reg}. Second, the rows are not at fault. Replacing σ by $\tanh$ with $\gamma > 0$ restores density by the ridge-monomial argument in the proof of Corollary C.19 together with the Weierstrass theorem, and the same three rows then give $\mathbf{K}^{\infty}_{(0)} \succ 0$ by Proposition 3.2, as they also do by Proposition 3.3, since the Dirac measures at x_1, x_2, x_3 are linearly independent. Third, the mechanism is not tied to an affine activation. For any polynomial σ of degree p on this one-dimensional domain the closed ridge span lies in the polynomials of degree at most p, so any $p + 2$ distinct evaluation rows are independent in $\mathcal{X}^*$ yet dependent on that span. For $\sigma = \tanh$ with the bias-free scale $\gamma = 0$ permitted by Definition 2.1 every ridge vanishes at $x = 0$, so the closed span lies in $\{f \in \mathcal{X} : f(0) = 0\}$ and the single row λ_1, which is nonzero and hence trivially independent, already yields $\mathbf{K}^{\infty}_{(0)} = 0$ for an analytic nonpolynomial activation. Clauses P_{pd} and N_{pd} exclude such configurations by explicitly requiring a nonpolynomial σ and $\gamma > 0$; clause W_{pd} states neither requirement, and its density clause is precisely what does that work on a general trial space.

What the example does not show: it establishes only that the clause cannot be deleted outright, not that density is the weakest usable replacement, and it is purely qualitative, since a feature family whose span is dense but poorly conditioned still yields a positive definite kernel whose smallest eigenvalue requires Proposition C.9. It also says nothing about the pointwise or nonlocal classes. Like Example D.8, which restricts the model space by a polynomial activation for strong-form rows, the construction is degenerate by design; the point here is the identification of the single weak-class clause that rules such degeneracies out.

Example D.2 (Deleting $\gamma > 0$ from N_{pd}: an even activation collapses two distinct Dirac rows). Take $d_0 = d_2 = 1$, $\Omega = (-1, 1)$, and $D = \overline{\Omega} = [-1, 1]$, so that $\mathcal{X} = C([-1, 1])$. Use Definition 2.1 with the bias-free scale $\gamma = 0$ and the activation $\sigma = \cosh$,

$$u_\theta(x) = \frac{1}{\sqrt{d_1}} \sum_{m=1}^{d_1} a_m \cosh(w_m x),$$

and initialize every coordinate of (w_m, b_m, a_m) i.i.d. from $\operatorname{Unif}[-1, 1]$. As in Example D.8, the formal hidden biases b_m do not enter the prediction when $\gamma = 0$, so their Jacobian columns vanish. Impose the two nonlocal rows of Theorem C.21 with $s_1 = s_2 = 1$ and

$$\nu_1 = \delta_1, \qquad \nu_2 = \delta_{-1}, \qquad P_1(\theta) = u_\theta(1), \qquad P_2(\theta) = u_\theta(-1).$$

Every clause of N_{reg} and N_{pd} in Assumption 2.6 holds except one: the rows are finite signed regular Borel measures on the compact set D, the law μ of Assumption 2.5 is compactly supported, $\cosh \in C^2_{\text{loc}}(\mathbb{R})$ is real analytic and nonpolynomial, and $\operatorname{int}(\operatorname{supp} \mu_h) = (-1, 1)^2 \neq \varnothing$. Only the clause $\gamma > 0$ is deleted, and Definition 2.1 admits $\gamma = 0$.

The deleted clause is exactly what the conclusion needs. Testing the identity $c_1\delta_1 + c_2\delta_{-1} = 0$ on the Borel sets $\{1\}$ and $\{-1\}$ gives $\mathcal{N}_{\text{nl}} = \{\mathbf{0}\}$ and $\dim \operatorname{span}\{\bar{\boldsymbol{\nu}}_1, \bar{\boldsymbol{\nu}}_2\} = 2$, so the weakened

forms of Equation (45) and Equation (7) would assert $\operatorname{rank}\mathbf{K}^{\infty}_{(0)} = 2$. It is instead 1. Because $\cosh$ is even and $\gamma = 0$, every ridge $y \mapsto \cosh(wy)$ is even, hence $u_\theta(-x) = u_\theta(x)$ for every θ and every width, and $P_1 - P_2 \equiv 0$ as a function of θ. Equivalently, the per-neuron Jacobian $\mathbf{\Psi}$ defined before Equation (4), evaluated at $z = (w, b, a)$ with columns ordered as $(\partial_w, \partial_b, \partial_a)$, has two identical rows,

$$\mathbf{\Psi}(z) = \begin{bmatrix} a\sinh w & 0 & \cosh w \\ a\sinh w & 0 & \cosh w \end{bmatrix},$$

since $\cosh(-w) = \cosh w$ and $(-1)\sinh(-w) = \sinh w$. Taking the expectation in $\mathbf{K}^{\infty}_{(0)} = \mathbb{E}_{z\sim\mu}[\mathbf{\Psi}(z)\mathbf{\Psi}(z)^\top]$ and using $\mathbb{E}[\cosh^2 w] = \frac{1}{4}(2 + \sinh 2)$, $\mathbb{E}[\sinh^2 w] = \frac{1}{4}(\sinh 2 - 2)$, and $\mathbb{E}[a^2] = \frac{1}{3}$ gives

$$\mathbf{K}^{\infty}_{(0)} = \frac{1 + \sinh 2}{3}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \qquad \ker\mathbf{K}^{\infty}_{(0)} = \operatorname{span}\{(1, -1)^\top\}, \qquad \operatorname{rank}\mathbf{K}^{\infty}_{(0)} = 1. \tag{46}$$

Both summands of Equation (4) are separately singular here, with every entry of $\mathbf{K}^{\infty}_{\text{out}}$ equal to $\frac{1}{4}(2 + \sinh 2)$ and every entry of $\mathbf{K}^{\infty}_{\text{hid}}$ equal to $\frac{1}{12}(\sinh 2 - 2)$.

The failure is localized to a single step of the proof of Theorem C.21, where $\gamma > 0$ makes the shifts γb exhaust $\mathbb{R}$ and thereby licenses the uniform universal-approximation theorem for a non-polynomial activation *with a bias* (Leshno et al., 1993). When $\gamma = 0$ the available features are $\{y \mapsto \cosh(wy) : w \in \mathbb{R}\}$, whose closed span in $C([-1, 1])$ consists of even functions only, so it is a proper closed subspace and the Riesz–Markov step is unavailable. Concretely, $\zeta = \delta_1 - \delta_{-1}$ satisfies $\int_D \cosh(wy)\, d\zeta(y) = \cosh(w) - \cosh(-w) = 0$ for every w while $\zeta \neq 0$, so Equation (91) no longer forces the corresponding signed measure to vanish. The degeneracy is exactly one dimensional: $c_1\delta_1 + c_2\delta_{-1}$ annihilates every even continuous function precisely when $c_1 + c_2 = 0$, matching Equation (46).

Restoring $\gamma > 0$ repairs the same two rows. With $\gamma = 1$ and the same law, the output-weight features $[\mathbf{\Phi}_{\text{out}}(\boldsymbol{\xi})]_{i,:} = \mathbf{f}_i(\boldsymbol{\xi})^\top$ of Equation (4) are $\mathbf{f}_1(w, b) = \cosh(w + b)$ and $\mathbf{f}_2(w, b) = \cosh(w - b)$, so $\mathbb{E}[\mathbf{f}_1\mathbf{f}_2] = \frac{1}{2}\mathbb{E}[\cosh 2w + \cosh 2b] = \frac{1}{2}\sinh 2$ and $\mathbb{E}[\mathbf{f}_1^2] = \mathbb{E}[\mathbf{f}_2^2] = \frac{1}{2} + \frac{1}{8}\sinh^2 2$, whence

$$\mathbf{K}^{\infty}_{\text{out}} = \frac{1}{8}\begin{bmatrix} 4 + \sinh^2 2 & 4\sinh 2 \\ 4\sinh 2 & 4 + \sinh^2 2 \end{bmatrix}, \qquad \operatorname{spec}(\mathbf{K}^{\infty}_{\text{out}}) = \left\{\frac{(\sinh 2 + 2)^2}{8}, \frac{(\sinh 2 - 2)^2}{8}\right\},$$

approximately $\{3.958, 0.331\}$ with eigenvectors $(1, \pm 1)^\top$. Both eigenvalues are positive because $\sinh 2 \neq 2$, so $\mathbf{K}^{\infty}_{\text{out}} \succ 0$ and therefore $\mathbf{K}^{\infty}_{(0)} \succ 0$ by Equation (5): an exact certificate in the sense of Definition 2.7, agreeing with Theorem C.21. The mechanism is general. For any $\gamma > 0$ and any μ_h whose support has nonempty interior, $\mathbf{c}^\top\mathbf{K}^{\infty}_{\text{out}}\mathbf{c} = 0$ forces the real-analytic function $g_{\mathbf{c}}(w, b) = (c_1 + c_2)\cosh w\cosh(\gamma b) + (c_1 - c_2)\sinh w\sinh(\gamma b)$ to vanish on $\operatorname{int}(\operatorname{supp}\mu_h)$, hence on all of $\mathbb{R}^2$ by the identity theorem; setting $b = 0$ gives $c_1 + c_2 = 0$ because $\cosh \geq 1$, and then $(w, b) = (1, 1/\gamma)$ gives $c_1 = c_2$. The bias is what breaks the parity.

Two limitations are worth stating. First, the construction leaves the existence half of Theorem C.21 intact: Proposition B.6 imposes no condition on γ, so $\mathbf{K}^{\infty}_{(0)}$ still exists and $\mathbf{K}_{(0)} \to \mathbf{K}^{\infty}_{(0)}$ at $\gamma = 0$. Only the nullspace and rank identity in Equation (45) fails. Second, the example isolates the hypothesis rather than the architecture: with $\gamma = 0$ and the also real-analytic nonpolynomial $\sigma = \exp$, the identity $\int_D e^{wy}\, d\zeta(y) = 0$ for all $w \in \mathbb{R}$ continues analytically to the Fourier transform of ζ and still forces $\zeta = 0$, so the conclusion survives for that instance. What the construction shows is that the weakened hypothesis set admits false instances, not that every $\gamma = 0$ model is degenerate.

The same pair also settles the pointwise class. Since $\delta_{\pm 1}$ are order-zero rows of Definition 2.2 at the distinct sites $\mathbf{x} = \pm 1 \in \overline{\Omega}$ with local coefficient vectors $\mathbf{b}_1(1) = \mathbf{b}_2(-1) = (1)$, the weakened form of Equation (6) would likewise predict $\operatorname{rank}\mathbf{K}^{\infty}_{(0)} = 1 + 1 = 2$, and $\cosh$ satisfies P_{reg} for $r_{\max} = 0$. So $\gamma > 0$ is load-bearing in P_{pd} for Proposition 3.1 exactly as it is in N_{pd} for Proposition 3.3 and Theorem C.21. The example is distinct from Example D.8, which also sets $\gamma = 0$ but additionally violates the nonpolynomiality clause; here analyticity and nonpolynomiality are retained and only $\gamma > 0$ is removed.

In the next three examples the global feature family is rich enough, and what fails is the passage from $\operatorname{supp}\mu_h$ to all of $\mathbb{R}^{d_0+1}$ on which the density clause is stated. The first two remove analyticity of the feature map; the third removes the interior hypothesis on $\operatorname{supp}\mu_h$.

Example D.3 (Smoothness without analyticity collapses the weak feature span). Delete from clause W_{pd} of Assumption 2.6 only the requirement that every $\boldsymbol{\xi} \mapsto \varphi_{\boldsymbol{\xi},q}$ be real analytic into $\mathcal{X}$, keeping $\mathrm{int}(\mathrm{supp}\,\mu_h) \neq \varnothing$, the ridge-density clause, and all of $\mathsf{W}_{\mathrm{reg}}$. Both identities in Equation (41) then fail, and with them the positive-definiteness criteria of Proposition 3.2 and Corollary C.18.

Take the row family of Corollary C.19 with $d_0 = d_2 = 1$, $\Omega = (-1, 1)$, $\mathcal{X} = H^1(\Omega)$, $\kappa \equiv 1$, $\gamma = 1$, unit row scales, the single test function $v_1(x) = 1 - x^2 \in H^1_0(\Omega)$, and the network of Definition 2.1; only the activation will be changed. The $N = 3$ rows

$$\lambda_1^{\Omega}(f) = \int_{-1}^{1} f'(x) v_1'(x)\, dx, \qquad \beta_-^{\partial}(f) = f(-1), \qquad \beta_+^{\partial}(f) = f(1)$$

are bounded on $\mathcal{X}$ by Cauchy–Schwarz and the one-dimensional trace theorem, and they are linearly independent in $\mathcal{X}^*$: testing a vanishing combination against v_1, then against $x \mapsto (1 - x)/2$, then against the constant function 1 forces its three coefficients to vanish in that order. Hence $\mathcal{N}_{\text{weak}} = \{\mathbf{0}\}$ in Equation (40), and the weakened statement would assert $\mathbf{K}^{\infty}_{(0)} \succ 0$ with rank 3. For $\sigma = \tanh$ this is exactly what Corollary C.19 proves.

Replace $\tanh$ by

$$\sigma_\star(z) = z + \chi\big(\tfrac{1}{2} - z\big), \qquad \chi(t) = e^{-1/t}\mathbf{1}_{\{t>0\}},$$

and let $\mu = \mathrm{Unif}([1, 2] \times [3, 4]) \otimes \mathrm{Unif}([-1, 1])$. Then $\sigma_\star \in C^\infty(\mathbb{R})$, it agrees with the identity on $[\frac{1}{2}, \infty)$, and it is neither polynomial nor real analytic, because $\sigma_\star(0) = e^{-2} \neq 0$ while any polynomial, or any real-analytic function, that agrees with z on a half-line equals z everywhere. The law μ is width independent, compactly supported, and sampled i.i.d., with $\mathrm{int}(\mathrm{supp}\,\mu_h) = (1, 2) \times (3, 4) \neq \varnothing$ and $\mathbb{E}[a_m \mid \boldsymbol{\xi}_m] = 0$, so Assumption 2.5 holds.

On the open cone $Q = \{\boldsymbol{\xi} = (w, b) : \gamma b - |w| > \frac{1}{2}\}$ every preactivation seen on the domain lies in the affine piece of $\sigma_\star$: for $x \in \overline{\Omega}$,

$$wx + \gamma b \geq \gamma b - |w| > \tfrac{1}{2}, \qquad \text{so} \qquad \varphi_{\boldsymbol{\xi}} = w\, p_1 + \gamma b\, p_0 \quad \text{exactly on } \overline{\Omega},$$

where $p_0 = 1$ and $p_1 = x$. Moreover $\mathrm{supp}\,\mu_h \subset Q$, because $b - |w| \geq 3 - 2 = 1$ there; any $\gamma > 0$ works after rescaling the bias support to $[3/\gamma, 4/\gamma]$. Thus $\mathrm{span}\{\varphi_{\boldsymbol{\xi}} : \boldsymbol{\xi} \in Q\} = \mathrm{span}\{p_0, p_1\}$ is two-dimensional, although the full ridge family over $\mathbb{R}^2$ is dense. Writing $F_i(\boldsymbol{\xi}) = \lambda_i(\varphi_{\boldsymbol{\xi}})$ for the three rows in the order above, and using that $v_1 \in H^1_0(\Omega)$ makes $\int_{-1}^{1} v_1' = 0$, so that λ_1^{Ω} annihilates every affine function, we obtain on Q

$$F_1 \equiv 0, \qquad F_2 = \gamma b - w, \qquad F_3 = \gamma b + w, \qquad \nabla_{\boldsymbol{\xi}} F_2 = (-1, \gamma), \qquad \nabla_{\boldsymbol{\xi}} F_3 = (1, \gamma).$$

The Gram representation in Theorem C.7, applied to the per-neuron maps $h_i(z) = aF_i(\boldsymbol{\xi})$ supplied by Lemma B.3 for $z = (\boldsymbol{\xi}, a)$, gives for every $\mathbf{c} \in \mathbb{R}^3$

$$\mathbf{c}^\top \mathbf{K}^{\infty}_{(0)} \mathbf{c} = \mathbb{E}_{\mu_h}\Big[\big((c_3 - c_2)w + (c_2 + c_3)\gamma b\big)^2\Big] + \tfrac{1}{3}\,\big\|\big(c_3 - c_2,\ \gamma(c_2 + c_3)\big)\big\|_2^2,$$

since $\mathbb{E}[a^2] = \frac{1}{3}$ and a is independent of $\boldsymbol{\xi}$. The second term vanishes precisely when $c_3 - c_2 = 0$ and $c_2 + c_3 = 0$, that is, when $c_2 = c_3 = 0$; the first term then vanishes as well, and c_1 is unconstrained, so

$$\ker \mathbf{K}^{\infty}_{(0)} = \mathrm{span}\{(1, 0, 0)^\top\} \neq \{\mathbf{0}\} = \mathcal{N}_{\text{weak}}, \qquad \mathrm{rank}\, \mathbf{K}^{\infty}_{(0)} = 2 < 3 = \dim \mathrm{span}\{\lambda_1^{\Omega}, \beta_-^{\partial}, \beta_+^{\partial}\}.$$

The annihilating functional is explicit: λ_1^{Ω} is nonzero on $\mathcal{X}$, since $\lambda_1^{\Omega}(v_1) = \int_{-1}^{1} 4x^2\, dx = \frac{8}{3}$, yet it kills every $\varphi_{\boldsymbol{\xi}}$ with $\boldsymbol{\xi} \in Q$. The degeneracy is not an infinite-width artifact: $F_1 \equiv 0$ on the open set Q and $\boldsymbol{\xi}_m \in \mathrm{supp}\,\mu_h \subset Q$ almost surely, so $\nabla_\theta P_1(\theta_{(0)}) = \mathbf{0}$ almost surely and the first row and column of $\mathbf{K}_{(0)}$ vanish at every width, while the target $\int_{-1}^{1} g v_1$ of that row is generically nonzero.

Analyticity is the only clause that fails. Clause $\mathsf{W}_{\mathrm{reg}}$ holds in full, because the rows lie in $\mathcal{X}^*$ and $\sigma_\star \in C^3_{\mathrm{loc}}(\mathbb{R})$, so Corollary B.4 makes $\mathcal{G}$ a C^2 map into $H^1(\Omega)$ on an open neighborhood of every fixed parameter tube; the existence and convergence half of Proposition 3.2 is therefore intact and only its positive-definiteness half fails. The hidden support has nonempty interior. The ridge-density clause also holds: for each $n \in \mathbb{N}_0$, nonpolynomiality gives b_n with $\sigma_\star^{(n)}(\gamma b_n) \neq 0$, and $\partial_w^n \sigma_\star(wx + \gamma b_n)|_{w=0} = \sigma_\star^{(n)}(\gamma b_n) x^n$ is an H^1 limit of finite combinations of ridge features, so

the monomials, and hence the $H^1(-1,1)$-dense polynomials, lie in the closed span, exactly as in the proof of Corollary C.19. Only Banach-valued analyticity fails, and visibly so: $\varphi_{(0,b)} = \sigma_\star(\gamma b)p_0$ is not a real-analytic function of b.

The failure occurs at exactly the step the proof needs. In the proof of Theorem C.17, analyticity is used once: a combination of the F_{iq} that vanishes μ_h-almost surely vanishes on $\mathrm{int}(\mathrm{supp}\,\mu_h)$ by continuity, and the identity theorem propagates this to all of $\mathbb{R}^{d_0+1}$, where the density clause turns it into $\sum_i c_i\lambda_i = 0$ in $\mathcal{X}^*$; the same step appears in Proposition C.13. Here the propagation is the only thing that fails: the relation $F_1 \equiv 0$ holds on Q and stops there, while the density clause quantifies over all of $\mathbb{R}^{d_0+1}$, so the two clauses never meet. The reverse inclusion $\mathcal{N}_{\text{weak}} \subseteq \ker \mathbf{K}^\infty_{(0)}$ uses only linearity and still holds. The example also shows that the clause must be read globally in $\boldsymbol{\xi}$: here $\boldsymbol{\xi} \mapsto \varphi_{\boldsymbol{\xi}}$ is real analytic on a neighborhood of $\mathrm{supp}\,\mu_h$, indeed affine there, and that is not enough.

Two remarks delimit the claim. First, $\sigma = \mathrm{ReLU}$ with the same μ produces the identical collapse, because $\gamma b - |w| \geq 1 > 0$ on $\mathrm{supp}\,\mu_h$ places the kink $x = -\gamma b/w \leq -\frac{3}{2}$ outside $\overline{\Omega}$; however, $\boldsymbol{\xi} \mapsto \varphi_{\boldsymbol{\xi}}$ is then not differentiable into $H^1(\Omega)$ once the kink enters Ω, so ReLU also violates $\mathsf{W}_{\mathrm{reg}}$ on parameter tubes of radius at least $1/\sqrt{2}$ and is not a single-clause counterexample. The smooth $\sigma_\star$ shows that neither smoothness nor nonpolynomiality substitutes for analyticity. Second, the example does not claim that non-analytic activations give degenerate kernels in general: the collapse is a property of the pair (σ, μ_h), and a hidden support over which the affine piece of $\sigma_\star$ fails to cover $\overline{\Omega}$ produces a far richer feature family. Nor is analyticity the weakest possible clause; density of $\mathrm{span}\{\varphi_{\boldsymbol{\xi},q} : \boldsymbol{\xi} \in \mathrm{supp}\,\mu_h\}$ would suffice, and analyticity is the checkable condition that transports the stated density clause from $\mathbb{R}^{d_0+1}$ to $\mathrm{supp}\,\mu_h$.

The same device applies to the other two classes with $\sigma_\star$, whose smoothness already gives $\mathsf{P}_{\mathrm{reg}}$ and $\mathsf{N}_{\mathrm{reg}}$. On Q the strong-form row $u \mapsto u''(0)$ and the nonlocal row $f \mapsto \int f\, d\nu$ with $\nu = \delta_{-1} - 2\delta_0 + \delta_1$ both annihilate $\mathrm{span}\{p_0, p_1\}$ and give an identically zero limiting NTK, although each row is nonzero and its local coefficient vector, respectively measure tuple, is nonzero. The analyticity requirements in P_{pd} and N_{pd}, used by Propositions 3.1 and 3.3, are load-bearing for the same reason. Unlike Example C.1, where the obstruction is a mismatch between activation smoothness and differential order, the activation here may be taken as smooth as desired and the obstruction is purely the loss of analytic continuation.

Example D.4 (A compactly supported C^∞ activation breaks the weak criterion). This example weakens exactly one clause of the hypotheses of Theorem C.17: in clause W_{pd} of Assumption 2.6, replace "every $\boldsymbol{\xi} \mapsto \varphi_{\boldsymbol{\xi},q}$ is real analytic into $\mathcal{X}$" by "C^∞ into $\mathcal{X}$", and keep every other requirement verbatim, including Assumption 2.5, clause $\mathsf{W}_{\mathrm{reg}}$, $\mathrm{int}(\mathrm{supp}\,\mu_h) \neq \varnothing$, and density of the ridge span over all $\boldsymbol{\xi} \in \mathbb{R}^{d_0+1}$. The weakened statement is false, and with it the weakened forms of Proposition 3.2 and Corollary C.18. It is the same clause as in Example D.3, broken by a different mechanism: there the features are locally affine on the region the row observes, here they are locally zero.

Take $d_0 = d_2 = 1$, $\Omega = (-1,1)$, $\mathcal{X} = H^1(-1,1)$, and $\gamma = 1$ in Definition 2.1, with the standard bump activation

$$\sigma(z) = \begin{cases} \exp\big((z^2-1)^{-1}\big), & |z| < 1, \\ 0, & |z| \geq 1, \end{cases} \qquad \sigma \in C^\infty_c(\mathbb{R}), \qquad \mathrm{supp}\,\sigma = [-1,1].$$

Initialize $(w_m, b_m) \sim \mathrm{Unif}\big([1,2] \times [1, \frac{3}{2}]\big)$ and $a_m \sim \mathrm{Unif}\{\pm 1\}$ independently, so μ is compactly supported and width independent, $\mathbb{E}[a_m \mid w_m, b_m] = 0$, and $\mathrm{int}(\mathrm{supp}\,\mu_h) = (1,2) \times (1, \frac{3}{2}) \neq \varnothing$. Impose the single ($N = 1$) weak residual row generated by the H^1 inner product, that is, the weak form of $-u'' + u$ tested against a fixed $v_1 \in C^\infty_c\big((\frac{1}{2}, 1)\big)$ with $v_1 \geq 0$ and $v_1 \not\equiv 0$:

$$\mathcal{V}_1 = \mathcal{X}, \qquad a(f,v) = \int_{-1}^{1} \big(f'v' + fv\big)\, dx = \langle f, v\rangle_{H^1}, \qquad \lambda_1 = a(\cdot, v_1), \qquad s_1 = 1.$$

Write $\varphi_{\boldsymbol{\xi}}(x) = \sigma(wx + b)$ and $F_1(\boldsymbol{\xi}) = \lambda_1(\varphi_{\boldsymbol{\xi}})$ for $\boldsymbol{\xi} = (w, b)$.

The weakened hypotheses hold. Since $\sigma \in C^\infty \subset C^3_{\mathrm{loc}}$, Corollary B.4 gives the required C^2 map into $\mathcal{X}$ on a neighborhood of every radius-R parameter tube, and a is continuous and bilinear with $\lambda_1 \in \mathcal{X}^*$, $\|\lambda_1\|_{\mathcal{X}^*} = \|v_1\|_{H^1}$; clause $\mathsf{W}_{\mathrm{reg}}$ therefore holds. All derivatives of σ are bounded and

uniformly continuous, so $\boldsymbol{\xi} \mapsto \varphi_{\boldsymbol{\xi}}$ is C^∞ into $C^1([-1,1]) \hookrightarrow H^1(-1,1)$. The ridge span is dense: by the Banach-valued differentiation argument in the proof of Corollary C.19, $\partial_w^n \sigma(wx+b)\big|_{w=0} = \sigma^{(n)}(b)\, x^n$ is an H^1 limit of finite linear combinations of features $\varphi_{(w,b)}$ with b fixed, and $\sigma^{(n)} \not\equiv 0$ for every n, since otherwise σ would be a polynomial vanishing on $[1,\infty)$, hence identically zero, contradicting $\sigma(0) = e^{-1}$. Choosing b_n with $\sigma^{(n)}(b_n) \neq 0$ puts every monomial, and hence every polynomial, in the closed span, and polynomials are dense in $H^1(-1,1)$. Finally a is coercive with $c_a = 1$ and $\{v_1\}$ is linearly independent, so Theorem C.17 in its weakened form, and likewise its coercive special case, would assert $\mathbf{K}^\infty_{(0)} \succ 0$.

The conclusion fails. Let $U = \{(w,b) : w > 0,\ \frac{w}{2} + b > 1\}$, an open set containing $\operatorname{supp} \mu_h = [1,2] \times [1, \frac{3}{2}]$, where $\frac{w}{2} + b \geq \frac{3}{2}$. For $\boldsymbol{\xi} \in U$ and $x \geq \frac{1}{2}$ we have $wx + b \geq \frac{w}{2} + b > 1$, and $\{x : wx + b > 1\} = \big((1-b)/w, \infty\big)$ with $(1-b)/w < \frac{1}{2}$; hence $\varphi_{\boldsymbol{\xi}}$ and all of its x-derivatives vanish on a neighborhood of $[\frac{1}{2}, 1]$ in $\overline{\Omega}$, which contains $\operatorname{supp} v_1$. Therefore

$$F_1(\boldsymbol{\xi}) = \int_{1/2}^{1} \big(\varphi'_{\boldsymbol{\xi}} v_1' + \varphi_{\boldsymbol{\xi}} v_1\big)\, dx = 0 \qquad \text{for every } \boldsymbol{\xi} \in U,$$

and, U being open, every partial derivative of F_1 vanishes on U as well. This is consistent with the density just established: F_1 is nonzero off U, for instance $F_1(0,0) = e^{-1} \int_{1/2}^1 v_1 > 0$. Because $P_1(\theta) = d_1^{-1/2} \sum_m a_m F_1(\boldsymbol{\xi}_m)$, both $\partial P_1 / \partial a_m = d_1^{-1/2} F_1(\boldsymbol{\xi}_m)$ and $\partial P_1 / \partial \boldsymbol{\xi}_m = d_1^{-1/2} a_m \nabla F_1(\boldsymbol{\xi}_m)$ vanish, so $\nabla_\theta P_1(\theta) = \mathbf{0}$ at every parameter vector whose neuron blocks lie in $\operatorname{supp} \mu$. Hence $\mathbf{K}_{(0)} = \mathbf{0}$ almost surely at every width and $\mathbf{K}^\infty_{(0)} = \mathbf{0}$. Yet $\lambda_1(v_1) = \|v_1\|^2_{H^1} > 0$, so $\lambda_1 \neq 0$ and $\mathcal{N}_{\text{weak}} = \{0\}$. Therefore

$$\ker \mathbf{K}^\infty_{(0)} = \mathbb{R} \neq \{0\} = \mathcal{N}_{\text{weak}}, \qquad \operatorname{rank} \mathbf{K}^\infty_{(0)} = 0 < 1 = \dim \operatorname{span}\{\lambda_1\},$$

contradicting Equation (41). Since $\nabla_\theta \mathcal{L}(\theta_{(0)}) = \mathbf{0}$, gradient descent is stationary at every width and step size, and by Definition 2.3 with $s_1 = 1$ the loss is frozen at $\mathcal{L}(\theta_{(0)}) = y_1^2$.

The kernel need not collapse to zero. Append the boundary row $\lambda_2 = \delta_{-1}$, that is, $\lambda_2(f) = f(-1)$, which is bounded on $H^1(-1,1)$ by the one-dimensional trace theorem and enters Theorem C.17 through the device $\mathcal{V}_2 = \mathbb{R}$, $a_2(f,t) = tf(-1)$, $v_2 = s_2 = 1$ used in the proof of Corollary C.18. Then $F_2(\boldsymbol{\xi}) = \sigma(b - w)$, and $b - w \in [-1, \frac{1}{2}]$ on $\operatorname{supp} \mu_h$, so $F_2 > 0$ μ_h-almost everywhere, with $F_2(1,1) = \sigma(0) = e^{-1}$; consequently $[\mathbf{K}^\infty_{\text{out}}]_{22} = \int F_2^2\, d\mu_h > 0$, and the full kernel dominates the output-weight block by Equation (5). The rows are independent, since $c_1\lambda_1 + c_2\lambda_2 = 0$ applied to v_1 gives $c_1 \|v_1\|^2_{H^1} = 0$ because $v_1(-1) = 0$, and applying the relation to the constant function 1 then gives $c_2 = 0$. The first row of $\mathbf{J}(\theta)$ still vanishes identically, so the first row and column of the kernel vanish and

$$\ker \mathbf{K}^\infty_{(0)} = \operatorname{span}\{\mathbf{e}_1\} \neq \{0\}, \qquad \operatorname{rank} \mathbf{K}^\infty_{(0)} = 1 < 2 = \dim \operatorname{span}\{\lambda_1, \lambda_2\}.$$

This is a weak residual row together with a boundary row on a common trial space, so the weakened form of Corollary C.18 fails as well.

The one-line reason: the proof of Theorem C.17 concludes from $F_1 \equiv 0$ on $\operatorname{supp} \mu_h$ that $F_1 \equiv 0$ on $\mathbb{R}^{d_0+1}$ by the identity theorem; a C^∞ map can be flat, so that step, and only that step, is what the weakening removes. The hypothesis as written excludes the example: σ vanishes on $|z| \geq 1$ while $\sigma(0) = e^{-1} \neq 0$, so it is not real analytic, and since $\delta_0 \in H^1(-1,1)^*$, real analyticity of $\boldsymbol{\xi} \mapsto \varphi_{\boldsymbol{\xi}}$ into $\mathcal{X}$ would force $b \mapsto \sigma(b)$ to be real analytic. The degeneracy is a different mechanism from Example C.1: there the activation violates the regularity clause and each feature is locally a nonzero polynomial that a high-order derivative row annihilates, an order mismatch, whereas here regularity is amply satisfied, the row is a bounded first-order weak row, and the features are locally *zero* on the region the row observes, a support mismatch. The example does not break the existence half of the theorem: clause W_{reg} still holds, so the first conclusion of Proposition 3.2 still yields existence of $\mathbf{K}^\infty_{(0)}$ and convergence of $\mathbf{K}_{(0)}$, here trivially. Nor does it show that analyticity is the only possible hypothesis: density over all $\boldsymbol{\xi}$ does hold, and what fails is that $\operatorname{span}\{\varphi_{\boldsymbol{\xi}} : \boldsymbol{\xi} \in \operatorname{supp} \mu_h\}$ lies in the proper closed subspace $\{f \in \mathcal{X} : f|_{[1/2,1]} = 0\}$, so requiring density of the features indexed by $\operatorname{int}(\operatorname{supp} \mu_h)$ alone would restore the conclusion. What is shown is that the analyticity clause cannot simply be downgraded to C^∞ while the density clause is quantified over all of $\mathbb{R}^{d_0+1}$, and

that the obstruction is a mismatch between $\operatorname{supp}\mu_h$ and the region a row observes rather than a defect of compactly supported activations as such.

Example D.5 (Empty-interior hidden support breaks the exact positivity criterion). Delete the clause $\operatorname{int}(\operatorname{supp}\mu_h) \neq \varnothing$ from N_{pd} and W_{pd} in Assumption 2.6 and retain every other hypothesis. The following system then satisfies all remaining clauses and has linearly independent rows, yet its limiting NTK is singular.

Let $d_0 = d_2 = 1$, $\Omega = (-\pi, \pi)$, $D = \overline{\Omega} = [-\pi, \pi]$, $\sigma = \sin$, and $\gamma > 0$. Use the network of Definition 2.1,

$$u_\theta(x) = \frac{1}{\sqrt{d_1}} \sum_{m=1}^{d_1} a_m \sin(w_m x + \gamma b_m),$$

with the three unit-scaled evaluation rows $\nu_i = \delta_{x_i}$ at $(x_1, x_2, x_3) = (-\pi, 0, \pi)$ and $s_i = 1$, so that $P_i(\theta) = u_\theta(x_i)$: two Dirichlet boundary rows and one interior measurement. Initialize the neurons i.i.d. with independent coordinates

$$w_m \sim \mathrm{Unif}\{-1, 1\}, \qquad b_m \sim \mathrm{Unif}[-1, 1], \qquad a_m \sim \mathrm{Unif}\{-1, 1\},$$

so that $\operatorname{supp}\mu_h = \{-1, 1\} \times [-1, 1]$ is compact with empty interior in $\mathbb{R}^2$; for $d_0 = 1$ this is the unit-norm hidden-weight sphere with a free bias. Every other hypothesis holds. The law μ is compactly supported and width independent as in Assumption 2.5; $\sin$ is real analytic, nonpolynomial, and smooth; $\gamma > 0$; D is compact and the rows are finite signed regular Borel measures, giving $\mathsf{N}_{\mathrm{reg}}$. Reading the same rows as $\lambda_i(f) = f(x_i)$ on $\mathcal{X} = C(D)$ gives $\mathsf{W}_{\mathrm{reg}}$, since $C(D)$ embeds continuously in L^1_{loc}, point evaluations are bounded on it, and $\boldsymbol{\xi} \mapsto \sin(w \cdot + \gamma b)$ is C^2 (indeed analytic) into $C(D)$; such rows are covered by Theorem C.17 on taking $\mathcal{V}_i = \mathbb{R}$, $v_i = s_i = 1$, and $a_i(f, t) = t\lambda_i(f)$, exactly as in the proof of Corollary C.18. The remaining clauses of W_{pd} hold as well: $\boldsymbol{\xi} \mapsto \varphi_{\boldsymbol{\xi},1} = \sin(w \cdot + \gamma b)$ is real analytic into $C(D)$ because D is compact, and the ridge span is dense by Leshno et al. (1993). Finally $\mathbb{E}[a_m \mid w_m, b_m] = 0$, so the output-weight centering used for GD also holds.

Take $\mathbf{c} = (1, 2, 1)^\top$. The associated single-neuron feature combination is

$$H(w, b) = \sum_{i=1}^{3} c_i \sin(wx_i + \gamma b) = \sin(\gamma b - \pi w) + 2\sin(\gamma b) + \sin(\gamma b + \pi w) = 4\sin(\gamma b)\cos^2\!\left(\tfrac{\pi w}{2}\right),$$

by the sum-to-product identity followed by $1 + \cos t = 2\cos^2(t/2)$. The squared factor is the whole point: $\cos^2(\pi w/2)$ has a *double* zero at every odd integer, so the per-neuron map $h_{\mathbf{c}}(w, b, a) = aH(w, b)$, which satisfies $\sum_i c_i P_i(\theta) = d_1^{-1/2} \sum_m h_{\mathbf{c}}(\theta_m)$, has gradient

$$\nabla h_{\mathbf{c}} = \left(-4\pi a \sin(\gamma b)\sin\!\left(\tfrac{\pi w}{2}\right)\cos\!\left(\tfrac{\pi w}{2}\right),\ 4a\gamma\cos(\gamma b)\cos^2\!\left(\tfrac{\pi w}{2}\right),\ 4\sin(\gamma b)\cos^2\!\left(\tfrac{\pi w}{2}\right)\right),$$

in the coordinates (w, b, a). Every component retains a factor $\cos(\pi w/2)$, so $\nabla h_{\mathbf{c}}$ vanishes identically on $\operatorname{supp}\mu$. Hence $\mathbf{J}(\theta_{(0)})^\top \mathbf{c} = \mathbf{0}$ almost surely at every width, so $\mathbf{K}_{(0)}\mathbf{c} = \mathbf{0}$ and $\mathbf{K}^\infty_{(0)}\mathbf{c} = \mathbf{0}$. Explicitly, the limit exists by $\mathsf{N}_{\mathrm{reg}}$ and Theorem C.7 and equals the per-neuron gradient expectation; writing $\mathbf{f}_1 = (-1, 1, -1)^\top$ and $\mathbf{f}_2 = (1, 0, -1)^\top$, both signs $w = \pm 1$ give $(\sin(wx_i + \gamma b))_i = \sin(\gamma b)\mathbf{f}_1$, $(\partial_w \sin(wx_i + \gamma b))_i = \pi\cos(\gamma b)\mathbf{f}_2$, and $(\partial_b \sin(wx_i + \gamma b))_i = \gamma\cos(\gamma b)\mathbf{f}_1$, so that, with $a^2 \equiv 1$ and $b \sim \mathrm{Unif}[-1, 1]$,

$$\mathbf{K}^\infty_{(0)} = \underbrace{\left(\mathbb{E}[\sin^2(\gamma b)] + \gamma^2\mathbb{E}[\cos^2(\gamma b)]\right)}_{=:A}\mathbf{f}_1\mathbf{f}_1^\top + \underbrace{\pi^2\mathbb{E}[\cos^2(\gamma b)]}_{=:B}\mathbf{f}_2\mathbf{f}_2^\top.$$

Here $A \geq \min(1, \gamma^2) > 0$ because $\sin^2 + \cos^2 = 1$, and $B > 0$ because $\cos(\gamma b)$ vanishes at only finitely many $b \in [-1, 1]$. Since $\mathbf{f}_1$ and $\mathbf{f}_2$ are independent,

$$\operatorname{rank}\mathbf{K}^\infty_{(0)} = 2, \qquad \ker\mathbf{K}^\infty_{(0)} = \operatorname{span}\{(1, 2, 1)^\top\},$$

the null direction being the one orthogonal to both $\mathbf{f}_1$ and $\mathbf{f}_2$. A referee may check this in one line from the entries $[\mathbf{K}^\infty_{(0)}]_{11} = [\mathbf{K}^\infty_{(0)}]_{33} = A + B$, $[\mathbf{K}^\infty_{(0)}]_{12} = [\mathbf{K}^\infty_{(0)}]_{23} = -A$, $[\mathbf{K}^\infty_{(0)}]_{13} = A - B$, and $[\mathbf{K}^\infty_{(0)}]_{22} = A$: the determinant is zero while the leading 2×2 minor equals $AB > 0$.

Distinct Dirac masses are linearly independent, so $\mathcal{N}_{\mathrm{nl}} = \{0\}$ and $\dim\operatorname{span}\{\bar{\boldsymbol{\nu}}_1, \bar{\boldsymbol{\nu}}_2, \bar{\boldsymbol{\nu}}_3\} = 3$. The weakened Theorem C.21 and Proposition 3.3 would therefore assert $\ker\mathbf{K}^\infty_{(0)} = \{0\}$ and

$\operatorname{rank} \mathbf{K}^{\infty}_{(0)} = 3$; both fail, and with them the implication from independence of the measure tuples to positive definiteness. Reading the same three rows in $C(D)^*$ breaks the weakened Theorem C.17 and Proposition 3.2 identically, and reading them as order-zero pointwise rows at distinct sites breaks the weakened Proposition 3.1, whose local coefficient families $\mathbf{b}_i(x_i) = (1)$ are independent. The finite-dimensional witnesses fail with them: the moment matrix $[\mathbf{M}^{\mathrm{nl}}_2]_{i,(1,\alpha)} = x_i^{\alpha}$ of Equation (35) is a Vandermonde matrix at the nodes $-\pi, 0, \pi$, with determinant $(0+\pi)(\pi+\pi)(\pi-0) = 2\pi^3 \neq 0$, so Corollary C.14 and Proposition C.13 would certify positive definiteness from a rank-3 test.

The deleted clause is exactly what licenses the analytic-continuation step. The proofs of Theorems C.17 and C.21 pass from $\mathbf{c} \in \ker \mathbf{K}^{\infty}_{\mathrm{out}}$ to vanishing of the analytic function $G_{\mathbf{c},1}$ on $\operatorname{supp} \mu_h$ — here $G_{\mathbf{c},1} = H$ — invoke the identity theorem because that support contains a nonempty open set, and only then apply ridge density as in Equation (91). In this example H does vanish on $\operatorname{supp} \mu_h$, because $\operatorname{supp} \mu_h$ lies inside the zero set $\{w \in 2\mathbb{Z}+1\} \cup \{\gamma b \in \pi\mathbb{Z}\}$, and that zero set has empty interior; there is therefore no open set to continue from, and the density step is never reached. No repair is available within the argument: it is the conclusion, not only the proof, that fails. The failure is also neither an artifact of the row count nor of a knife-edge law. Appending further rows of any class keeps $(\mathbf{c}, \mathbf{0})$ in the kernel, because $\nabla h_{\mathbf{c}}$ still vanishes on all of $\operatorname{supp} \mu$, and *every* law carried by $\{|w| = 1\}$, with any bias marginal, produces the same null vector.

Four caveats. First, only the "pd" conclusions break: the deleted clause appears nowhere in $\mathsf{N}_{\mathrm{reg}}$ or $\mathsf{W}_{\mathrm{reg}}$, so existence of $\mathbf{K}^{\infty}_{(0)}$ and the convergence $\mathbf{K}_{(0)} \to \mathbf{K}^{\infty}_{(0)}$ still hold here, and we used them. Second, the example shows that the clause cannot be *deleted*; it does not show that the clause is the weakest available hypothesis. A nontrivial real-analytic function has a Lebesgue-null zero set, so the strictly weaker requirement $|\operatorname{supp} \mu_h| > 0$ already suffices for the same proof. The exact condition under which the conclusion holds is that μ_h charge the complement of the zero set of every feature combination attached to a vector outside the relation space, but that condition is a restatement of the conclusion rather than a checkable design criterion, which is why a geometric sufficient condition is assumed instead. Third, the statement concerns the kernel at initialization, which is what these theorems characterize and what the spectral premise of Theorem 4.2 requires; the hidden weights leave $\{-1, 1\}$ under GD, and we make no claim about $\mathbf{K}_{(k)}$ for $k \geq 1$. Fourth, only the positive-dimensional support uses the periodicity of $\sin$. For an arbitrary real-analytic nonpolynomial σ, any $\gamma > 0$, and any d_0, take $\mu_h = \delta_{\boldsymbol{\xi}_0}$ and $N \geq d_0 + 3$ Dirac rows at distinct sites: the limiting NTK is then a sum of the $d_\theta = d_0 + 2$ rank-one Gram terms generated by the single parameter value $\boldsymbol{\xi}_0$, so $\operatorname{rank} \mathbf{K}^{\infty}_{(0)} \leq d_0 + 2 < N$ while the rows remain independent. Empty interior, not the choice of activation, is what the construction exploits.

The last example is of a different kind. It concerns Assumption 2.5 rather than a "pd" clause, and it breaks existence of the limiting kernel rather than its richness.

Example D.6 (A heavy-tailed initialization law destroys the deterministic limit). The initialization clause of Assumption 2.5 asks three things of the per-neuron law μ: that it be i.i.d. across neurons, width independent, and compactly supported. Delete only the third, so that μ is an arbitrary width-independent law on $\mathbb{R}^{d_\theta}$, and retain every clause of Assumption 2.6. The existence half of Proposition 3.1, equivalently statement one of Theorem C.7, then fails already for a single value row.

Take $d_0 = d_2 = 1$, $\gamma = 1$, $\Omega = (0, 1)$, and the one Dirichlet row of Definition 2.2 with $\mathbf{x}_1 = 1$, $I_{11} = \{0\}$, coefficient $a_{11\alpha} \equiv 1$ at $\alpha = 0$, and $s_1 = 1$, so that $N = 1$, $r_{\max} = 0$ and $P_1(\theta) = u_\theta(1)$. Let $\sigma(z) = \exp(z^4)$ and

$$\theta_{m,(0)} = (w_m, b_m, a_m) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_3), \qquad t_m = w_m \mathbf{x}_1 + \gamma b_m = w_m + b_m \sim \mathcal{N}(0, 2).$$

Every hypothesis other than compact support holds: σ is real analytic on $\mathbb{R}$, hence lies in $C^{r_{\max}+1,1}_{\mathrm{loc}}(\mathbb{R})$, and is nonpolynomial; $\gamma > 0$; $\operatorname{int}(\operatorname{supp} \mu_h) = \mathbb{R}^2 \neq \varnothing$; and μ is i.i.d. across neurons and width independent. Thus $\mathsf{P}_{\mathrm{reg}}$ and P_{pd} both hold, as does the structural centering requirement $\mathbb{E}[a_m \mid \boldsymbol{\xi}_m] = 0$ imposed for gradient descent, since a_m has mean zero and is independent of $\boldsymbol{\xi}_m$. Only $\operatorname{supp} \mu = \mathbb{R}^3$ is noncompact.

By Equation (20) the per-neuron map of Equation (12) is $h_1(\boldsymbol{\xi}, a) = a\, \sigma(w + \gamma b)$, so the output-weight entry of the single-neuron Jacobian $\boldsymbol{\Psi}$ defined in Definition B.1 is $\partial_a h_1 = \sigma(t)$. Since $N = 1$,

discarding the nonnegative contributions of the two hidden coordinates gives

$$[\mathbf{K}_{(0)}]_{11} = \frac{1}{d_1}\sum_{m=1}^{d_1}\big\|\mathbf{\Psi}(\theta_{m,(0)})\big\|_F^2 \;\geq\; \frac{1}{d_1}\sum_{m=1}^{d_1} X_m, \qquad X_m = \exp\big(2t_m^4\big)\,.$$

These i.i.d. nonnegative summands are not integrable: since $t^2 \leq t^4$ for $|t| \geq 1$,

$$\mathbb{E}X_1 = \frac{1}{\sqrt{4\pi}}\int_{\mathbb{R}} \exp\Big(2t^4 - \tfrac{t^2}{4}\Big)\,dt \geq \frac{1}{\sqrt{4\pi}}\int_{|t|\geq 1} \exp\big(\tfrac{7}{4}t^4\big)\,dt = \infty,$$

the last integrand being at least $e^{7/4}$ on an unbounded set. Truncation gives the one-line failure. For each $K \in \mathbb{N}$ the variables $\min\{X_m, K\}$ are i.i.d. and bounded, so the strong law yields $\liminf_{d_1\to\infty}[\mathbf{K}_{(0)}]_{11} \geq \mathbb{E}\min\{X_1, K\}$ almost surely for every K simultaneously, while $\mathbb{E}\min\{X_1, K\} \uparrow \mathbb{E}X_1 = \infty$ by monotone convergence. Hence $[\mathbf{K}_{(0)}]_{11} \to \infty$ almost surely along one coupled i.i.d. sequence, and the same truncation with the weak law gives $\Pr\{[\mathbf{K}_{(0)}]_{11} \leq M\} \to 0$ for every M under separate resampling at each width, so both convergence modes asserted in statement one fail. Moreover $\mathbb{E}[\mathbf{K}_{(0)}]_{11} = \infty$ at every finite width, so no deterministic $\mathbf{K}_{(0)}^{\infty}$ exists and the Gram representation in Theorem C.7 is undefined.

Compact support is used at exactly one point in this chain. Statement one is proved by Theorem B.9 under the row bounds of Equation (13), equivalently (G_R) in Equation (8), and for the pointwise class those bounds are not assumed but derived: the proof of Proposition B.7 extracts them from compactness of $\operatorname{supp}\mu + \overline{\mathbb{B}}_R(0)$. Here $\sup_{z\in\operatorname{supp}\mu}|h_1(z)| = \infty$, so the constant $A_{0,1}$ does not exist and the derivation has no substitute, the conclusion itself being false. Note that the divergence comes from the output-weight coordinate alone, which involves no derivative of σ and no differential order, so it is not an artifact of activation smoothness or of high-order rows.

This sharpens Example C.6, which records the infinite population moment for the same activation: the point here is that the empirical kernel itself diverges, so the conclusion fails and not merely its proof, and that every remaining clause has been checked. Three limits should be stated plainly. First, this is an existence failure and says nothing about positive definiteness; the rank criteria of Proposition 3.1 are never reached because there is no limiting matrix to test. For the same reason the identity $\mathbb{E}_{z\sim\mu}h_1(z) = 0$ delivered by conditional centering is vacuous here, h_1 not being μ-integrable; this is a symptom of the same nonintegrability rather than a second failed hypothesis. Second, the example shows that the compact-support clause cannot be deleted, not that compactness is necessary: $\mathbb{E}_{z\sim\mu}\|\mathbf{\Psi}(z)\|_F^2 < \infty$ would already suffice for the deterministic limit, which is the relaxation that Example D.7 exploits. Compact support is retained because it additionally supplies the essential-supremum constant R_K for the finite-width Chernoff bound and the uniform gradient-Lipschitz constant L_* for the local stability bound, neither of which follows from moments alone. Third, the obstruction is a joint incompatibility between the tail of μ and the growth of σ, not a defect of Gaussian initialization: a Gaussian μ with $\tanh$, or with any activation whose relevant derivatives grow polynomially, leaves the per-neuron terms integrable and the limit intact.

Taken together, these six constructions show that no clause of Assumptions 2.5 and 2.6 named above can simply be deleted, or downgraded from analytic to smooth, while the remaining clauses are left as written. They do not show that any clause is necessary in the logical sense. The clauses are sufficient conditions chosen to be checkable at design time, and in several cases a strictly weaker condition would carry the same proof: density of $\{\varphi_{\boldsymbol{\xi},q} : \boldsymbol{\xi} \in \operatorname{supp}\mu_h\}$ in place of analyticity plus global density, $|\operatorname{supp}\mu_h| > 0$ in place of nonempty interior, and a finite second moment of $\mathbf{\Psi}$ in place of compact support. Nor do the examples say anything quantitative: a feature family that satisfies every clause but is poorly conditioned still yields a positive definite kernel, and bounding its smallest eigenvalue requires Proposition C.9 rather than the exact nullspace identities tested here.

D.2 Taylor-coefficient conditions for linear independence

The parameter-derivative functions depend on the PDE, initialization, nonlinearity, architecture, and collocation data, so the linear-independence condition is problem specific outside the hypotheses of the preceding rank theorems. For bounded-order pointwise differential constraints, analytic activations make them analytic in the neuron parameters. Proposition C.13 states the exact condition $\operatorname{rank}\mathbf{C}_K(\boldsymbol{\xi}_\star) = N$ for some finite K; the calculation below evaluates the corresponding Taylor

coefficients directly. For weak and nonlocal rows, Corollary C.14 gives the matrices $\mathbf{M}_K^{\text{weak}}$ and $\mathbf{M}_K^{\text{nl}}$ in Equations (34) and (35).

Example D.7 (Positive definiteness for a Poisson PINN). Consider the following $\gamma = 0$ specialization of Definition 2.1:

$$u(x, \theta) = \frac{1}{\sqrt{d_1}} \sum_{m=1}^{d_1} a_m \tanh(w_m x).$$

Assume the marginal initialization law of w_m has compact support containing an open neighborhood of zero (for example, a compactly supported density that is positive near zero). Also take the per-neuron parameter vectors to be i.i.d. and the output weights to have finite second moment. These conditions ensure that the full limiting NTK exists; compact initialization as in Assumption 2.5 is sufficient. Use the PINN loss in Definition 2.3 with inputs $x = 1, 2, 3, 4, 5$ for the one-dimensional Poisson problem

$$-\frac{d^2u}{dx^2} = f(x), \quad x \in \Omega = (1, 5),$$
$$u(1) = 0, \quad u(5) = 0.$$

In this bias-free model the scalar inputs are all multiples of one another, so the distinct-direction condition of bias-free positivity results, which requires that no two inputs be parallel, fails. Nevertheless, $\mathbf{K}_{(0)}^{\infty}$ is positive definite because an analytic power-series coefficient calculation proves that the parameter-derivative functions in the output-weight contribution are linearly independent. This is the example-specific positive-definiteness certificate described in the pointwise procedure of Section D.3.1; see also the detailed derivation.

Example D.8 (Singular NTK with distinct constraints and zero nullspace projection). Let $\Omega = (-1, 1)$ and use the network in Definition 2.1 with one-dimensional output and analytic activation $\sigma(z) = z + z^2$:

$$u_\theta(x) = \frac{1}{\sqrt{d_1}} \sum_{m=1}^{d_1} a_m \sigma(w_m x).$$

Because $\gamma = 0$, the formal hidden-bias parameters b_m do not affect the predictions; their Jacobian columns are zero, so we omit them below. Initialize a_m and w_m independently from the uniform distribution $\mathrm{Unif}[-1, 1]$. This deliberately polynomial activation violates the activation clause used by the differential-NTK positivity result. More decisively, its limiting NTK is singular, so it admits no positive-definiteness certificate and does not satisfy the corresponding premise of the convergence theorem. Place one PDE collocation point at $x_D = 0$ and the two Dirichlet points at $x_- = -1$ and $x_+ = 1$. Every network output has the form $A_\theta x + B_\theta x^2$, and therefore

$$u_\theta''(0) = u_\theta(-1) + u_\theta(1). \tag{47}$$

Differentiate the identity in Equation (47) with respect to every trainable parameter and define

$$\mathbf{\Phi}_D := \nabla_\theta u_\theta''(0), \qquad \mathbf{\Phi}_- := \nabla_\theta u_\theta(-1), \qquad \mathbf{\Phi}_+ := \nabla_\theta u_\theta(1).$$

These gradients obey the relation

$$\mathbf{\Phi}_D = \mathbf{\Phi}_- + \mathbf{\Phi}_+.$$

Thus the NTK is singular despite the distinct physical points and the linear independence of the two boundary parameter-derivative functions. The data can obey exactly the same relation in the following well-posed problem:

$$u''(x) = 2, \qquad u(-1) = u(1) = 1.$$

The unique solution is $u^*(x) = x^2$, which the network represents for every even d_1, and its unweighted target vector $(2, 1, 1)^\top$ is orthogonal to the NTK null vector $(1, -1, -1)^\top$. Consequently, the fixed-kernel dynamics attain zero limiting loss despite singularity. The sample-group normalization in Definition 2.2 rescales both prediction components and data and preserves rank and orthogonality to the nullspace. This construction uses $\gamma = 0$ and a polynomial activation to restrict the model and tangent spaces, so it lies outside Proposition 3.1 and the positive-definiteness-certificate premise of Theorem 4.2. See Section G.1 for the full calculation.

The example shows that kernel rank and target compatibility are distinct: positive definiteness permits every sampled target, whereas a singular system can fit targets satisfying its nullspace relations.

D.3 Constructive positive-definiteness certificates

We give finite positive-definiteness certificates and constructive search procedures for turning the preceding nullspace theorems into checkable tests, using the terminology of Definition 2.7. In every procedure, residual, initial, and boundary rows are assembled in the order used by the loss. The target values do not enter an NTK rank calculation. Positive loss scales preserve exact rank, but they must be retained when conditioning or a quantitative eigenvalue bound is reported.

D.3.1 Strong-form pointwise constraints

For each distinct physical site $\mathbf{x} \in \mathcal{S}$, form the finite coefficient matrix

$$\mathbf{A}_{\mathbf{x}} := [s_i a_{iq\alpha}(\mathbf{x})]_{\substack{i \in I_{\mathbf{x}} \\ (q,\alpha) \in \mathcal{C}_{\mathbf{x}}}}. \tag{48}$$

Its i-th row is $\mathbf{b}_i(\mathbf{x})^\top$. Hence, under the hypotheses of Proposition 3.1, full row rank is exactly local linear independence.

Certificate procedure.

1. Split every vector equation into scalar loss rows, and combine repeated occurrences of the same derivative $\partial^\alpha u_q$ at the same site.
2. At each site, assemble Equation (48) jointly over all PDE, initial, and boundary rows located there.
3. Exhibit a nonzero $|I_{\mathbf{x}}| \times |I_{\mathbf{x}}|$ minor of $\mathbf{A}_{\mathbf{x}}$ for every $\mathbf{x} \in \mathcal{S}$.

These minors form a pointwise positive-definiteness certificate: they prove $\operatorname{rank} \mathbf{A}_{\mathbf{x}} = |I_{\mathbf{x}}|$ and therefore $\mathbf{K}^{\infty}_{(0)} \succ 0$. Conversely, an exact nonzero vector in $\ker \mathbf{A}_{\mathbf{x}}^\top$ is a structural NTK null vector. Checking rows separately is valid at distinct sites, but not for multiple rows at one site.

When the sitewise hypotheses of Proposition 3.1 are unavailable, the analytic alternative in Proposition C.13 applies provided that the output weights are trainable, every output-weight feature F_{iq} is real analytic, the hidden law has compact support with nonempty interior, and the full limiting NTK exists. Under these hypotheses, increase K, assemble $\mathbf{C}_K(\boldsymbol{\xi}_\star)$, and exhibit one nonzero $N \times N$ minor. Equivalently, one may exhibit a finite parameter set Ξ for which $\mathbf{E}_\Xi$ has row rank N. Failure at one Taylor order or for one finite set Ξ is inconclusive.

For the five-row Poisson PINN in Example D.7, the biased sitewise theorem is unavailable because $\gamma = 0$. The full-dimensional analytic alternative is not invoked because this example is formulated only in the active hidden parameter w and assumes interior support in that reduced parameter space. Its separate direct proof uses the coefficients of w, w^3, w^5, w^7, w^9 in the two boundary features and three residual features. The resulting 5×5 Taylor matrix, displayed in the detailed Poisson proof, has determinant $6976569344/135 \neq 0$. This reduced-parameter Taylor argument is an exact positive-definiteness certificate for that strong-form Poisson example.

Example D.9 (Exact pointwise coefficient certificate). Assume the network and initialization hypotheses of Proposition 3.1; for example, use a biased tanh network with $\gamma = 1$, i.i.d. compactly supported per-neuron initialization whose hidden marginal support has nonempty interior, and bounded conditionally centered output weights. For this scalar network on $(0, 1)$, place two marked pointwise rows at $x = 0$:

$$\ell_1(f) = f(0) + f'(0), \qquad \ell_2(f) = f(0) - f'(0).$$

The second row can, for example, be read as a Robin boundary observation; the first may be a co-located residual observation. With unit row scales and columns ordered as $(f(0), f'(0))$,

$$\mathbf{A}_0 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \qquad \det \mathbf{A}_0 = -2 \neq 0.$$

Thus the two co-located rows are jointly independent and Proposition 3.1 proves that their limiting NTK is positive definite. Replacing the second row by a copy of the first makes the same matrix rank one and produces the exact null vector $(1, -1)^\top$.

D.3.2 Mixed weak residual and boundary constraints

Concatenate the loss-scaled functionals as

$$\Lambda_r = \begin{cases} \lambda^{\Omega}_r, & r \in [N_\Omega], \\ \beta^{\partial}_{r-N_\Omega}, & r = N_\Omega + 1, \dots, N, \end{cases} \qquad N = N_\Omega + N_\partial.$$

All rows must be bounded on one common trial space $\mathcal{X}$. Under the analytic hypotheses of Corollary C.14, construct

$$\mathbf{M}_K^{\mathrm{weak,mix}} := [\Lambda_r(\mathbf{e}_q p_\alpha)]_{\substack{r\in[N]\\ q\in[d_2],\ |\alpha|\le K}} = \begin{bmatrix} [s_i a_i(\mathbf{e}_q p_\alpha, v_i)]_{i,(q,\alpha)} \\ \left[\beta_j^\partial(\mathbf{e}_q p_\alpha)\right]_{j,(q,\alpha)} \end{bmatrix}. \tag{49}$$

Certificate procedure.

1. Put every residual and boundary row on one common trial space and prove that each row is bounded there.
2. For $K = 0, 1, \ldots$, assemble Equation (49) in exact arithmetic and search its $N \times N$ minors.
3. Stop with a rigorous positive-definiteness certificate as soon as one minor is proved nonzero. If the loss is a fixed quadrature rule, use the finite jet-atom coefficient matrix described below instead.

By Equation (33), the nonzero minor is a positive-definiteness certificate and is eventually equivalent to positive definiteness of the mixed limiting NTK. For a tested boundary row $\beta_j^\partial(f) = t_j\langle \mathcal{B}f, \psi_j\rangle_{\partial\Omega}$, the lower block is evaluated by applying that expression to $\mathbf{e}_q p_\alpha$. For a strong-form pointwise boundary row $\beta_j^\partial(f) = t_j \sum_{q,\eta} b_{jq\eta}\partial^\eta f_q(\mathbf{x}_j)$, written with η for the boundary multi-index because α indexes the monomial here, it is

$$\beta_j^\partial(\mathbf{e}_q p_\alpha) = t_j \sum_{\eta\in J_{jq}} b_{jq\eta}\partial^\eta p_\alpha(\mathbf{x}_j).$$

For a fixed quadrature loss, a smaller exact positive-definiteness certificate is often available. Collect all distinct jet atoms from both residual and boundary rows, write

$$\mathcal{Q}_a(f) = \partial^{\alpha_a} f_{q_a}(\mathbf{x}_a), \qquad \Lambda_r = \sum_{a=1}^{A} B_{ra}\mathcal{Q}_a,$$

and prove that the retained $\mathcal{Q}_a$ are independent in $\mathcal{X}^*$. Then Equation (39) gives the exact finite test $\operatorname{rank}\mathbf{B} = N$. In this finite representation, rank deficiency is also conclusive. Distinct triples $(q_a, \mathbf{x}_a, \alpha_a)$ alone do not prove independence on an arbitrary trial space; if the atoms have known relations, first express them in a basis of their span and rank the resulting compressed coefficient matrix.

Example D.10 (Exact mixed weak residual–boundary certificate). Let $\mathcal{X} = H^1(-1,1)$, take $v(x) = 1 - x^2 \in H_0^1(-1,1)$, and define one weak residual and two soft Dirichlet rows by

$$\lambda^\Omega(f) = \int_{-1}^{1} f'(x)v'(x)\,dx, \qquad \beta_-^\partial(f) = f(-1), \qquad \beta_+^\partial(f) = f(1).$$

For $p_0 = 1$, $p_1 = x$, and $p_2 = x^2$, the degree-two matrix in Equation (49) is

$$\mathbf{M}_2^{\mathrm{weak,mix}} = \begin{pmatrix} 0 & 0 & -8/3 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \qquad \det \mathbf{M}_2^{\mathrm{weak,mix}} = -\frac{16}{3} \neq 0.$$

Endpoint evaluation is bounded on $H^1(-1,1)$, and the weak row is bounded by Cauchy–Schwarz. Take the biased tanh network and compact initialization specified in Corollary C.19. The explicit H^1-valued analyticity and polynomial-density argument in the proof of that corollary, together with the boundedness just checked, verifies the feature hypotheses used here. The exact minor above and Corollary C.14 therefore provide a positive-definiteness certificate for the joint three-row limiting NTK without appealing to the positive-definiteness conclusion of that corollary.

D.3.3 Mixed nonlocal residual and boundary constraints

Absorb the positive row scales into one measure tuple per row:

$$\boldsymbol{\mu}_r = \begin{cases} s_r(\nu_{r1}, \ldots, \nu_{rd_2}), & r \in [N_\Omega], \\ t_j(\omega_{j1}, \ldots, \omega_{jd_2}), & r = N_\Omega + j. \end{cases}$$

On compact D, form the mixed moment matrix

$$\mathbf{M}_K^{\mathrm{nl,mix}} := \left[\int_D p_\alpha(\mathbf{y})\,d\mu_{rq}(\mathbf{y})\right]_{\substack{r\in[N]\\ q\in[d_2],\ |\alpha|\le K}}. \tag{50}$$

Certificate procedure.

1. Encode every residual and boundary row as one loss-scaled measure tuple, combining repeated atoms before forming a matrix.
2. For finitely supported measures, rank the exact Dirac coefficient matrix below. Otherwise, for $K = 0, 1, \ldots$, assemble Equation (50) and search for a provably nonzero $N \times N$ minor.
3. A nonzero minor supplies a positive-definiteness certificate. An exact linear dependence of the complete measure tuples certifies singularity; a failed finite moment search is only inconclusive.

By Corollaries C.14 and C.22, a nonzero moment minor is an exact positive-definiteness certificate, and one exists at some finite K exactly when the combined residual and boundary measure tuples are independent.

For finitely supported measures, take the union of all residual and boundary nodes, combine repeated atoms, and write $\mu_{rq} = \sum_{a=1}^{A} c_{rqa}\delta_{\mathbf{y}_a}$. The exact test then reduces to

$$\operatorname{rank}\left[c_{rqa}\right]_{r,(q,a)} = N,$$

as in Equation (36). This includes a Dirichlet row as a Dirac mass. Neumann and Robin derivative observations are not measures on $C(D)$ and must instead be placed in the weak-functional procedure on a space where they are bounded. If residual measures are absolutely continuous and boundary measures are singular with respect to a common reference measure, the final clause of Corollary C.22 permits separate checks; otherwise the combined matrix is necessary.

Example D.11 (Exact mixed nonlocal residual–boundary certificate). Let $D = [-1, 1]$, $d_2 = 1$, and use the nonlocal residual and Dirichlet rows

$$\lambda^{\Omega}(f) = f(0) - \frac{1}{2}\int_{-1}^{1} f(y)\,dy, \qquad \beta_{-}^{\partial}(f) = f(-1), \qquad \beta_{+}^{\partial}(f) = f(1).$$

Their scaled measures, with unit scales, are $\delta_0 - \frac{1}{2}\mathrm{Leb}|_{[-1,1]}$, δ_{-1}, and δ_1. Against $p_0 = 1$, $p_1 = y$, and $p_2 = y^2$,

$$\mathbf{M}_2^{\mathrm{nl,mix}} = \begin{pmatrix} 0 & 0 & -1/3 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \qquad \det \mathbf{M}_2^{\mathrm{nl,mix}} = -\frac{2}{3} \neq 0.$$

Thus the three measure rows are jointly independent. Under the activation and initialization hypotheses of Theorem C.21, the displayed nonzero minor is a positive-definiteness certificate for the joint limiting NTK.

D.3.4 Rigorous certificates and numerical diagnostics

A symbolic nonzero minor is an exact positive-definiteness certificate. A determinant enclosure computed with outward-rounded interval arithmetic is also rigorous when the enclosure excludes zero. In contrast, floating-point rank, a sampled smallest singular value, or a Monte Carlo Gram eigenvalue is only a numerical diagnostic. Raw monomials can be severely ill-conditioned; replacing them by an exactly invertible polynomial basis, such as scaled orthogonal polynomials, preserves rank.

For a diagnostic of the complete parameter-gradient family, draw independent $z^{(1)}, \ldots, z^{(M)} \sim \mu$ and form

$$\widehat{\mathbf{F}}_M \in \mathbb{R}^{N \times M d_\theta}, \qquad [\widehat{\mathbf{F}}_M]_{i,(t,p)} = \partial_{z_p} h_i(z^{(t)}).$$

Then the normalized empirical Gram has smallest eigenvalue $\sigma_{\min}(\widehat{\mathbf{F}}_M)^2/M$. Report the sampling law, sample count, row scales, arithmetic precision, and rank tolerance. For a fixed target, also report its projection onto the estimated nullspace.

Failure to find a full-rank truncated moment or Taylor matrix does not prove singularity, because no universal truncation order is available. Exact rank deficiency of a complete finite coefficient representation, such as $\mathbf{A}_{\mathbf{x}}$, $\mathbf{B}$, or the Dirac coefficient matrix, does prove singularity. None of these qualitative positive-definiteness certificates by itself lower-bounds $\lambda_{\min}(\mathbf{K}_{(0)}^{\infty})$; the positive-measure and feature-variation data in Proposition C.9 are needed for that purpose. Finally, a quadrature rank certificate concerns the stored discrete loss, and a separate consistency argument is required to connect it to the continuum weak or nonlocal problem.

E AUXILIARY OPTIMIZATION RESULTS

E.1 INITIALIZATION ERROR

Lemma E.1 (Initial error bound)**.** *Under Assumption 2.5, suppose* (G_R) *holds and* $\mathbb{E}[\mathbf{a} \mid \boldsymbol{\xi}] = \mathbf{0}$*, set* $\mathbf{h}(z) = (h_i(z))_{i=1}^N$*,* $H_0^2 = \mathbb{E}_{z\sim\mu}\|\mathbf{h}(z)\|_2^2 \le A_{0,*}^2$*, and* $\mathbf{r}_{(0)} = \mathbf{v}_T - \mathbf{v}_P(\theta_{(0)})$*. Every initialization realization whose per-neuron parameter vectors lie in the initialization support (and hence almost every realization) satisfies*

$$\|\mathbf{r}_{(0)}\|_2 \le \|\mathbf{v}_T\|_2 + \sqrt{d_1}\, A_{0,*}. \tag{51}$$

Moreover, for every $0 < \varepsilon < 1$*,*

$$\Pr\left\{\|\mathbf{r}_{(0)}\|_2 \le \|\mathbf{v}_T\|_2 + H_0 + A_{0,*}\sqrt{2\log\frac{1}{\varepsilon}}\right\} \ge 1 - \varepsilon.$$

Thus every realized $\mathbf{r}_{(0)}$ is finite under bounded initialization, but the direct deterministic estimate can grow as $\sqrt{d_1}$. Centering the output weights gives the width-independent high-probability control used in Theorem 4.2. For a fixed number of loss groups, suppose each row is divided by the square root of its group's sample count. Under uniform bounds on the unnormalized per-neuron prediction terms and target values, $A_{0,*}$ and $\|\mathbf{v}_T\|_2$ remain $\mathcal{O}(1)$ as the discretization is refined.

E.2 STRUCTURAL DEPENDENCIES AND COMPATIBLE SINGULAR KERNELS

Definition E.2 (Structural spaces and compatibility)**.** For the pointwise, weak, and finite-measure classes, let $\mathcal{X}_{\mathrm{row}}$ be, respectively, $C^{r_{\max}}(\mathbb{R}^{d_0};\mathbb{R}^{d_2})$, the stated trial space $\mathcal{X}$, or $C(D;\mathbb{R}^{d_2})$, and let $\ell_i : \mathcal{X}_{\mathrm{row}} \to \mathbb{R}$ denote the ith loss-scaled linear row. Define

$$\mathcal{O} : \mathcal{X}_{\mathrm{row}} \to \mathbb{R}^N, \qquad [\mathcal{O}f]_i = \ell_i(f),$$

$$\mathcal{N}_{\mathrm{str}} = \left\{\mathbf{c} \in \mathbb{R}^N : \sum_{i=1}^N c_i \ell_i = 0 \text{ on } \mathcal{X}_{\mathrm{row}}\right\}, \qquad \mathcal{U}_{\mathrm{str}} = \mathcal{N}_{\mathrm{str}}^{\perp}.$$

A target is *compatible* if $\mathbf{v}_T \in \mathcal{U}_{\mathrm{str}}$, equivalently if it is orthogonal to every exact row relation. A *structural-nullspace certificate* proves $\ker \mathbf{K}_{(0)}^{\infty} = \mathcal{N}_{\mathrm{str}}$. For a nonzero prescribed active subspace $\mathcal{U}$ with orthonormal basis matrix $\mathbf{U}$, a *compressed positive-definiteness certificate* verifies $\mathbf{U}^\top \mathbf{K}_{(0)}^{\infty} \mathbf{U} \succ 0$.

Lemma E.3 (Compatible targets and exact compression)**.** *For the structural spaces in Definition E.2, under the applicable "reg" clause of Assumption 2.6, so that* $\mathbf{u}(\cdot;\theta) \in \mathcal{X}_{\mathrm{row}}$ *for every* θ*,* $\mathrm{Range}(\mathcal{O}) = \mathcal{U}_{\mathrm{str}}$*. In particular, if one admissible field* $f_\star \in \mathcal{X}_{\mathrm{row}}$ *generates all targets,* $\mathbf{v}_T = \mathcal{O}f_\star$*, then the target is compatible. Moreover, every network prediction lies in* $\mathcal{U}_{\mathrm{str}}$ *and*

$$\mathcal{N}_{\mathrm{str}} \subseteq \ker \mathbf{K}(\theta) \qquad \text{for every } \theta.$$

If a structural-nullspace certificate holds, $\mathcal{U}_{\mathrm{str}} \neq \{\mathbf{0}\}$*, and* $\mathbf{U}$ *is an orthonormal basis matrix for* $\mathcal{U}_{\mathrm{str}}$*, then* $\mathbf{U}^\top \mathbf{K}_{(0)}^{\infty} \mathbf{U} \succ 0$*, with smallest eigenvalue equal to the smallest positive eigenvalue of* $\mathbf{K}_{(0)}^{\infty}$*. For every target, the loss decomposes exactly as*

$$\mathcal{L}(\theta) = \|\mathbf{U}^\top(\mathbf{v}_T - \mathbf{v}_P(\theta))\|_2^2 + \|\Pi_{\mathcal{N}_{\mathrm{str}}}\mathbf{v}_T\|_2^2, \qquad \mathbf{K}_{\mathcal{U}}(\theta) = \mathbf{U}^\top \mathbf{K}(\theta)\mathbf{U}. \tag{52}$$

The second term is an irreducible constant and vanishes exactly for compatible targets. Hence, for a compatible target, the full residual belongs to $\mathcal{U}_{\mathrm{str}}$ *and the original loss is exactly the compressed loss; for an incompatible target, gradient descent still optimizes the same active loss but cannot remove the nullspace term. If* $\mathcal{U}_{\mathrm{str}} = \{\mathbf{0}\}$*, the loss decomposition still holds, but there is no positive active eigenvalue and no nontrivial compressed GD problem.*

Proof. The definition of $\mathcal{N}_{\mathrm{str}}$ gives $\mathcal{N}_{\mathrm{str}} = \mathrm{Range}(\mathcal{O})^{\perp}$; the range is a subspace of finite-dimensional $\mathbb{R}^N$ and hence is closed, proving the first identity. Taking $f = f_\star$ proves compatibility of a jointly generated target, while taking $f = \mathbf{u}(\cdot;\theta)$ proves that every prediction is active. Differentiating $\mathbf{c}^\top \mathbf{v}_P(\theta) = 0$ gives $\mathbf{J}(\theta)^\top \mathbf{c} = 0$, and hence the empirical-kernel inclusion. If the population kernel has kernel $\mathcal{N}_{\mathrm{str}}$, its quadratic form is strictly positive on the orthogonal complement, which proves

compressed positive definiteness. The spectral assertion follows by restricting the eigendecomposition to the positive eigenspace. Since every prediction is active, orthogonal decomposition of the residual gives Equation (52); its nullspace component is $\Pi_{\mathcal{N}_{\mathrm{str}}}\mathbf{v}_T$. Finally, the compressed Jacobian is $\mathbf{U}^\top\mathbf{J}(\theta)$, giving the final kernel identity. □

The nonsingular result in Theorem 4.2 yields contraction for arbitrary data. For a singular design, Propositions 3.1 to 3.3 identify the relevant structural nullspace $\mathcal{N}_{\mathrm{str}}$. Target compatibility is a separate data condition, not a consequence of kernel singularity; by Lemma E.3, it is automatic when one admissible exact field generates all rows of the target and otherwise must be checked as $\mathbf{v}_T \perp \mathcal{N}_{\mathrm{str}}$. The same lemma proves that the loss and NTK then compress exactly to $\mathcal{U}_{\mathrm{str}} = \mathcal{N}_{\mathrm{str}}^\perp$ and, when this space is nontrivial, that the population compressed kernel is positive definite. Under the corresponding compressed regularity, width, and step-size conditions, Corollary E.4 therefore gives the same geometric loss decay even though the full kernel is singular. In the rank-zero edge case, a compatible target has identically zero residual, so there is no optimization problem.

Corollary E.4 (Convergence when all errors lie in a fixed subspace)**.** *Fix $0 < \delta < 1$. Let $\mathcal{U} \subseteq \mathbb{R}^N$ be a fixed r-dimensional subspace, where $1 \le r \le N$, and let $\mathbf{U} \in \mathbb{R}^{N\times r}$ have orthonormal columns that span $\mathcal{U}$. Suppose*

$$\mathbf{v}_T - \mathbf{v}_P(\theta) \in \mathcal{U} \qquad \text{for every } \theta \text{ in an open neighborhood of } \mathcal{T}_R(\theta_{(0)}). \tag{53}$$

Set $\mathbf{v}_{T,\mathcal{U}} = \mathbf{U}^\top\mathbf{v}_T$, assume Assumption 2.5 and $\mathbb{E}[\mathbf{a} \mid \boldsymbol{\xi}] = \mathbf{0}$, and require (G_R) for the compressed prediction $\mathbf{U}^\top\mathbf{v}_P$. Form its per-neuron regularity constants and the corresponding initial-residual bound $\overline{r}_{\delta,\mathcal{U}}$, with target $\mathbf{v}_{T,\mathcal{U}}$, as in Theorem 4.2. Also define $\mathbf{K}_{\mathcal{U}}^\infty = \mathbf{U}^\top\mathbf{K}_{(0)}^\infty\mathbf{U}$ and require a compressed positive-definiteness certificate as in Definition E.2, namely $\lambda_{\mathcal{U}} = \lambda_{\min}(\mathbf{K}_{\mathcal{U}}^\infty) > 0$. In the width and step-size conditions of Theorem 4.2, replace $\lambda_0, \Lambda, R_K, A_{0,*}, H_0, B_*, L_*$ by their compressed counterparts, $\overline{r}_\delta$ by $\overline{r}_{\delta,\mathcal{U}}$, and N by r in the Chernoff logarithm. Assume that the resulting inequalities hold. Then, with probability at least $1-\delta$, every iterate lies in $\mathcal{T}_R(\theta_{(0)})$ and, for every $k \in \mathbb{N}_0$,

$$\mathcal{L}(\theta_{(k)}) \le (1 - \eta\lambda_{\mathcal{U}})^k \mathcal{L}(\theta_{(0)}).$$

The full $N \times N$ kernel may be singular. In particular, let $\mathcal{N}$ be the exact structural nullspace from Propositions 3.1 to 3.3, and choose $\mathcal{U} = \mathcal{N}^\perp$. If $\dim\mathcal{U} \ge 1$, then Lemma E.3 verifies Equation (53) *and the compressed certificate for every compatible target; it also shows that targets generated jointly by one admissible exact field are compatible automatically. If $\mathcal{U} = \{\mathbf{0}\}$, every compatible residual is identically zero, so there is no nontrivial optimization statement. Requiring subspace membership only at initialization, without such a global relation, is not sufficient for this corollary.*

Proof. Under Equation (53), $\|\mathbf{v}_T - \mathbf{v}_P(\theta)\|_2^2 = \|\mathbf{U}^\top(\mathbf{v}_T - \mathbf{v}_P(\theta))\|_2^2$ throughout the parameter neighborhood. Therefore the original objective and gradient are exactly those of the compressed prediction map, whose NTK is $\mathbf{U}^\top\mathbf{K}(\theta)\mathbf{U}$. Starting from the same initialization, the two update paths consequently coincide up to the first possible exit from the neighborhood. Apply Theorem 4.2 to this compressed system; its localization induction rules out leaving that set. □

E.3 Fixed-kernel limiting loss

Consider the fixed-kernel residual iteration $\mathbf{r}_{(k+1)} = (\mathbf{I}_N - 2\eta\mathbf{K}_{(0)})\mathbf{r}_{(k)}$. If $\mathbf{K}_{(0)} \ne \mathbf{0}$ and $0 < \eta < 1/\lambda_{\max}(\mathbf{K}_{(0)})$, its limiting loss is

$$\mathcal{L}_\infty^{\mathrm{lin}} = \left\|\Pi_{\ker(\mathbf{K}_{(0)})}\mathbf{r}_{(0)}\right\|_2^2.$$

When $\mathbf{K}_{(0)} = \mathbf{0}$, this expression remains valid because the update leaves every residual unchanged. For any architecture that is linear in its output weights, Lemma G.3 shows that $\mathbf{v}_P(\theta_{(0)}) \in \mathrm{Range}(\mathbf{K}_{(0)})$, so this nullspace projection equals that of $\mathbf{v}_T$. Otherwise, substitute the initial error into the displayed formula.

E.4 Necessary conditions: a loss floor and a width lower bound

The guarantees above are sufficient conditions. Two elementary necessary conditions delimit what cannot be repaired: part (i) by any width, step size, or optimizer, and part (ii) by any step size or optimizer at a width below N/d_θ. Both apply verbatim to vector-valued systems.

Proposition E.5 (Two obstructions to geometric decay)**.** *Let the rows be as in Definition 2.2 and satisfy the applicable "reg" clause of Assumption 2.6, so that* $\mathbf{u}(\cdot;\theta) \in \mathcal{X}_{\mathrm{row}}$ *for every* θ*, and let* $\mathcal{N}_{\mathrm{str}}$ *be the structural nullspace of Definition E.2.*

(i) Structural loss floor. *For every* θ*,*

$$\mathcal{L}(\theta) \geq \|\Pi_{\mathcal{N}_{\mathrm{str}}} \mathbf{v}_T\|_2^2.$$

Hence, for an incompatible target, no parameter sequence, whether produced by gradient descent at any width and step size or by any other method, drives the loss below this positive constant.

(ii) Counting bound. *Let* $B \subset \mathbb{R}^{d_1 d_\theta}$ *be bounded and suppose* $\theta \mapsto \mathbf{v}_P(\theta)$ *is Lipschitz on* $\overline{B}$*; under* (G_R) *this holds, with constant* B_**, for the closed radius-*R *tube* $B = \mathcal{T}_R(\theta_{(0)})$*. If* $N > d_1 d_\theta$*, then* $\mathbf{v}_P(\overline{B})$ *is a compact Lebesgue-null subset of* $\mathbb{R}^N$*, so for Lebesgue-almost every target* $\mathbf{v}_T \in \mathbb{R}^N$*,*

$$\inf_{\theta \in B} \mathcal{L}(\theta) > 0.$$

In particular, for every fixed initialization $\theta_{(0)}$ *and Lebesgue-almost every target, no trajectory that remains in* $\mathcal{T}_R(\theta_{(0)})$ *satisfies* $\mathcal{L}(\theta_{(k)}) \to 0$*. Since Theorem 4.2 asserts both localization in* $\mathcal{T}_R(\theta_{(0)})$ *and geometric decay, Fubini's theorem applied to the* σ*-compact set* $\{(\theta_{(0)}, \mathbf{v}_T) : \mathbf{v}_T \in \mathbf{v}_P(\mathcal{T}_R(\theta_{(0)}))\}$ *shows that, for almost every target, its conclusion has probability zero over the initialization when* $d_1 < N/d_\theta$*.*

Proof. Part (i) is the decomposition Equation (52), whose second term does not depend on θ. For part (ii), a locally Lipschitz map from an m-dimensional space into $\mathbb{R}^N$ with $m < N$ does not increase Hausdorff dimension, so its image of the bounded set $\overline{B}$ has Hausdorff dimension at most $m < N$ and therefore Lebesgue measure zero in $\mathbb{R}^N$; the image is compact because $\overline{B}$ is compact and the map is continuous. For $\mathbf{v}_T$ outside this null set, $\mathbf{v}_T \neq \mathbf{v}_P(\theta)$ for every $\theta \in \overline{B}$, so the continuous function $\theta \mapsto \mathcal{L}(\theta)$ attains a positive minimum on $\overline{B}$, which bounds the infimum over B from below. A trajectory confined to B with $\mathcal{L}(\theta_{(k)}) \to 0$ would contradict this bound. Under (G_R) every h_i has gradient norm at most B_i on each ball of the tube, so the mean-value inequality on the convex set $\mathcal{T}_R(\theta_{(0)})$ and the Cauchy–Schwarz inequality over the neurons make $\mathbf{v}_P$ Lipschitz there with constant B_*. The Fubini step uses that the set in question is the continuous image of the closed set $\{(\theta_{(0)}, \theta) : \theta \in \mathcal{T}_R(\theta_{(0)})\}$ and that each $\theta_{(0)}$-section of it is Lebesgue-null. □

The counting bound is linear in N, whereas the sufficient width in Equation (9) carries the factors λ_0^{-3}, L_*^2, Λ, and $\overline{r}_\delta^2$; the gap between the two is open. Already the matrix-Chernoff term of Equation (9) implies $d_1 \geq (N/d_\theta)\log(2N/\delta)/c_{\mathrm{Ch}}$, because $N\lambda_0 \leq \operatorname{tr} \mathbf{K}_{(0)}^{\infty} = \mathbb{E}\|\mathbf{\Psi}\|_F^2 \leq d_\theta R_K$; the counting bound is therefore a statement about every method rather than a sharpening of the theorem. Vector-valued outputs enter through $d_\theta = d_0 + d_2 + 1$ on one side and through the number N of scalar rows on the other, and they create no obstruction beyond these two: for pointwise rows, Proposition 3.1 shows that under clause P_{pd} the limiting kernel of a coupled system is singular only through a co-located dependency, which part (i) then turns into a positive loss floor for every incompatible target.

Example E.6 (A coupled system with a co-located dependency). Let $d_2 = 2$ and prescribe at one point $\mathbf{x}_0$ the two rows $u_1 + u_2 = y_1$ and $2u_1 + 2u_2 = y_2$ with loss scales $s_1, s_2 > 0$, so that $P_1(\theta) = s_1(u_1 + u_2)(\mathbf{x}_0;\theta)$ and $P_2(\theta) = 2s_2(u_1 + u_2)(\mathbf{x}_0;\theta)$. The local coefficient vectors $s_1(1,1)$ and $2s_2(1,1)$ are dependent, so $\operatorname{rank} \mathbf{K}_{(0)}^{\infty} \leq 1$, with equality under clause P_{pd} by Proposition 3.1, and the sitewise certificate of Section D.3.1 fails before any training. Indeed $2s_2 P_1(\theta) - s_1 P_2(\theta) = 0$ for every θ, so $\mathbf{c} = (2s_2, -s_1)^\top \in \mathcal{N}_{\mathrm{str}}$, and with $\mathbf{v}_T = (s_1 y_1, s_2 y_2)^\top$ part (i) of Proposition E.5 gives, for every width and every θ,

$$\mathcal{L}(\theta) \geq \frac{(\mathbf{c}^\top \mathbf{v}_T)^2}{\|\mathbf{c}\|_2^2} = \frac{s_1^2 s_2^2 (2y_1 - y_2)^2}{4s_2^2 + s_1^2},$$

which is positive whenever $y_2 \neq 2y_1$. For the compatible data $y_2 = 2y_1$ the floor vanishes, the active subspace is one-dimensional, and, under clause P_{pd}, Corollary E.4 applies to the compressed single-row problem.

F NONLINEAR CONSTRAINTS AND DEPTH: OBSTRUCTIONS TO THE PRESENT PROOF ARCHITECTURE

Every guarantee above concerns prediction components that are linear in the network output, on the two-layer network of Definition 2.1. This section records what happens when either restriction is dropped. The nonlinear part is a negative result and is proved here: with one genuinely nonlinear polynomial row the tangent Gram matrix at initialization still has a limit, but that limit is a random matrix, and no choice of width exponent repairs this. The depth part isolates one unconditional lemma, one conditional substitution proposition, and two conjectures, stated so that the hypotheses they would need are explicit. Nothing in the body depends on this section.

Relation to prior work. That the nonlinear regime is qualitatively different is not a new observation. Zhou & Yan (2024) ask when the PINN tangent kernel converges at all for possibly nonlinear differential operators and identify homogeneity-dependent width-scaling thresholds; Bonfanti et al. (2024) study the difficulties specific to the nonlinear regime. We claim no priority for the qualitative statement. What Sections F.2 and F.3 add is a form sharp enough to be used as a no-go: an exact finite-width identity that computes each kernel entry in full, so that no cancellation between contributions remains to be ruled out; an explicit $L^2(\mu)$ formula for the limiting variance rather than an assumed nonzero one; an affine-or-random dichotomy that needs no activation-specific or initialization-specific hypothesis beyond conditional centering, second-moment isotropy of the output weights, and this paper's own positive-definiteness certificate; and a trichotomy in the width exponent showing that the limiting object does not depend on the exponent at all.

F.1 COMPOSITE ROWS AND AN EXACT TANGENT GRAM IDENTITY

Definition F.1 (Width exponent)**.** For $\varrho > 0$, consider the one-parameter family

$$\mathbf{u}(\mathbf{x};\theta) = \frac{1}{d_1^{\varrho}} \sum_{m=1}^{d_1} \sigma(\mathbf{w}_m^\top \mathbf{x} + \gamma b_m)\mathbf{a}_m, \tag{54}$$

with θ_m, d_θ, and the trained parameters exactly as in Definition 2.1. The choice $\varrho = 1/2$ is Equation (1) and is the only member used elsewhere in this paper. For a nonlinear constraint of degree $p \geq 2$ we single out the *degree-matched* exponent $\varrho_p = (2p-1)/(2p)$, which equalizes the formal number of neuron sums in a degree-(p,p) kernel block.

Definition F.2 (Composite rows over a base family of linear observations)**.** Fix $n \in \mathbb{N}$ and a family $\mathcal{L} = [n]$ of scalar linear observations of the network, each satisfying the row-regularity condition of Definition B.1 at $\varrho = 1/2$, with per-neuron maps $h_\ell(\boldsymbol{\xi}, \mathbf{a}) = \mathbf{a}^\top \mathbf{f}_\ell(\boldsymbol{\xi})$ as in Equation (12) and constants $A_{0,\ell}, B_\ell, L_\ell < \infty$ as in Equation (13). Under Equation (54), the value of observation ℓ is

$$Q_\ell(\theta) = \frac{1}{d_1^{\varrho}} \sum_{m=1}^{d_1} h_\ell(\theta_m), \qquad \mathbf{Q}(\theta) = (Q_1(\theta), \ldots, Q_n(\theta))^\top. \tag{55}$$

A *composite row* is $P_i(\theta) = \pi_i(\mathbf{Q}(\theta))$ for a fixed $\pi_i \in C^1(\mathbb{R}^n)$; the loss scale s_i of Definition 2.2 is absorbed into π_i, so P_i is the full loss-scaled row and $\mathbf{K}(\theta)$ of Equation (3) is the full tangent Gram matrix. The row is *affine* when π_i is affine and *genuinely nonlinear* otherwise, and it is a *degree-p polynomial row* when π_i is a polynomial of degree p. For a multiset $J = (j_1, \ldots, j_q)$ with entries in $\mathcal{L}$ we write $\mathcal{M}_J(\theta) = \prod_{s=1}^{q} Q_{j_s}(\theta)$ for the associated monomial and call $q = |J|$ its degree; every polynomial row is a finite linear combination of monomials, and by bilinearity every kernel entry is the corresponding combination of the *monomial blocks* $K^{J,J'} = \sum_m \langle \nabla_{\theta_m} \mathcal{M}_J, \nabla_{\theta_m} \mathcal{M}_{J'} \rangle$.

An affine π_i recovers a row of Definition 2.2 whenever the base family spans the corresponding observation, and a constant term in π_i changes neither the Jacobian nor the tangent Gram matrix. A viscous scalar conservation law with quadratic flux gives the degree-two row $\pi_i(\mathbf{q}) = s_i(q_1 + q_2 q_3 - \nu q_4)$, where $Q_1, \ldots, Q_4$ are the unscaled observations $\partial_t u$, u, $\partial_x u$, $\partial_{xx} u$ at the collocation point $\mathbf{x}_i$; a p-th-power boundary row is $\pi_i(\mathbf{q}) = s_i q_\ell^p$.

Throughout this section we use two conditions on the initialization of Assumption 2.5.

(C) *Conditional output-weight centering:* $\mathbb{E}[\mathbf{a} \mid \boldsymbol{\xi}] = \mathbf{0}$ for μ-almost every neuron parameter, which implies $\mathbb{E}_{z\sim\mu} h_\ell(z) = 0$ for every $\ell \in \mathcal{L}$. This is the condition already imposed after Definition B.1 and used by Lemma E.1 and Theorem 4.2.

(C_{iso}) In addition, $\mathbb{E}[\mathbf{a}\mathbf{a}^\top \mid \boldsymbol{\xi}] = c_{\mathbf{a}} \mathbf{I}_{d_2}$ μ_h-almost surely for some $c_{\mathbf{a}} > 0$. Every initialization whose output-weight coordinates are i.i.d., symmetric, nondegenerate, and drawn independently of $\boldsymbol{\xi}$ satisfies it. This is an addition to Assumption 2.5, which imposes no second-moment structure, but it is not a restrictive one: the $\mathrm{Unif}[-1, 1]$ output weights used throughout the experiments of Section 5 satisfy it with $c_{\mathbf{a}} = 1/3$, as does any centered Gaussian or Rademacher coordinate law. It fails only when the output-weight coordinates are correlated or have unequal variances given $\boldsymbol{\xi}$.

Write $\boldsymbol{\Psi}_{\mathcal{L}}(z) \in \mathbb{R}^{n \times d_\theta}$ for the base-family analogue of $\boldsymbol{\Psi}$, namely $[\boldsymbol{\Psi}_{\mathcal{L}}(z)]_{\ell p} = \partial h_\ell(z)/\partial z_p$, and set

$$\begin{aligned} \widehat{\mathbf{K}}_{\mathcal{L}}(\theta) &= \frac{1}{d_1} \sum_{m=1}^{d_1} \boldsymbol{\Psi}_{\mathcal{L}}(\theta_m) \boldsymbol{\Psi}_{\mathcal{L}}(\theta_m)^\top, \\ [\mathbf{K}^\infty_{\mathcal{L}}]_{\ell\ell'} &= \mathbb{E}_{z\sim\mu} \langle \nabla h_\ell(z), \nabla h_{\ell'}(z) \rangle, \qquad [\boldsymbol{\Sigma}_{\mathcal{L}}]_{\ell\ell'} = \langle h_\ell, h_{\ell'} \rangle_{L^2(\mu)}. \end{aligned} \tag{56}$$

Here $\mathbf{K}^\infty_{\mathcal{L}}$ is the limiting NTK supplied by statement one of Theorem C.7 for the auxiliary *linear* design $\mathcal{L}$, and under (C) the matrix $\boldsymbol{\Sigma}_{\mathcal{L}}$ is the infinite-width prediction covariance of that design, the Gaussian-process covariance of Lee et al. (2019) evaluated on the base observations.

Lemma F.3 (Exact composite tangent Gram identity)**.** *Let the rows be composite as in Definition F.2 under the exponent ϱ. Then for every θ, every $m \in [d_1]$, and every $i, j \in [N]$,*

$$\nabla_{\theta_m} P_i(\theta) = \frac{1}{d_1^\varrho} \boldsymbol{\Psi}_{\mathcal{L}}(\theta_m)^\top \nabla \pi_i(\mathbf{Q}(\theta)), \qquad [\mathbf{K}(\theta)]_{ij} = d_1^{1-2\varrho} \nabla \pi_i(\mathbf{Q}(\theta))^\top \widehat{\mathbf{K}}_{\mathcal{L}}(\theta) \nabla \pi_j(\mathbf{Q}(\theta)). \tag{57}$$

In particular, for monomials $J = (j_1, \dots, j_q)$ and $J' = (l_1, \dots, l_{q'})$ of degrees q and q',

$$K^{J,J'}(\theta) = d_1^{1-2\varrho} \sum_{r=1}^{q} \sum_{t=1}^{q'} \Big(\prod_{s\neq r} Q_{j_s}(\theta) \Big) \Big(\prod_{s\neq t} Q_{l_s}(\theta) \Big) \big[\widehat{\mathbf{K}}_{\mathcal{L}}(\theta) \big]_{j_r l_t}. \tag{58}$$

Both identities hold at every width and every parameter value; no expectation, limit, or approximation is involved.

Proof. Only the m-th summand of Equation (55) depends on θ_m, so $\nabla_{\theta_m} Q_\ell(\theta) = d_1^{-\varrho} \nabla h_\ell(\theta_m)$, and the chain rule gives the first display. Since the neuron blocks exhaust every trainable parameter and $\mathbf{K} = \mathbf{J}\mathbf{J}^\top$ by Equation (3),

$$[\mathbf{K}(\theta)]_{ij} = \sum_{m=1}^{d_1} \langle \nabla_{\theta_m} P_i, \nabla_{\theta_m} P_j \rangle = d_1^{-2\varrho} \nabla \pi_i(\mathbf{Q})^\top \Big(\sum_{m=1}^{d_1} \boldsymbol{\Psi}_{\mathcal{L}}(\theta_m) \boldsymbol{\Psi}_{\mathcal{L}}(\theta_m)^\top \Big) \nabla \pi_j(\mathbf{Q}),$$

which is the second. For the monomial form, substitute $\nabla \mathcal{M}_J(\theta) = \sum_{r=1}^{q} \big(\prod_{s\neq r} Q_{j_s}(\theta) \big) \mathbf{e}_{j_r}$. □

The identity already displays the obstruction. The empirical average $\widehat{\mathbf{K}}_{\mathcal{L}}$ is the object the linear theory controls: it is a mean of i.i.d. bounded matrices. The factors $\nabla \pi_i(\mathbf{Q}(\theta))$ are not averages. They are fixed functions of the *predictions*, which every neuron shares, and which at initialization are normalized sums of i.i.d. terms. Such a sum is governed by the central limit theorem, not by the strong law.

Lemma F.4 (Limits of the two empirical objects)**.** *Assume Assumption 2.5 and let the base family satisfy Definition B.1. Put $\widetilde{Q}_\ell(\theta) = d_1^{-1/2} \sum_m h_\ell(\theta_m)$, so that $\mathbf{Q}(\theta) = d_1^{1/2-\varrho} \widetilde{\mathbf{Q}}(\theta)$. At initialization, as $d_1 \to \infty$:*

1. *$\widehat{\mathbf{K}}_{\mathcal{L}}(\theta_{(0)}) \to \mathbf{K}^\infty_{\mathcal{L}}$ in probability, and almost surely under the nested-width convention of Theorem C.7;*
2. *under (C), $\widetilde{\mathbf{Q}}(\theta_{(0)}) \Rightarrow \mathbf{Z} \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{\Sigma}_{\mathcal{L}})$, jointly with the convergence in item 1;*
3. *without (C), with $m_\ell = \mathbb{E}_{z\sim\mu} h_\ell(z)$, one has $d_1^{\varrho-1} Q_\ell(\theta_{(0)}) \to m_\ell$, almost surely along nested widths as in Theorem C.7 and in probability under fresh resampling.*

Moreover $|m_\ell| \le A_{0,\ell}$, $|[\mathbf{\Sigma}_\mathcal{L}]_{\ell\ell'}| \le A_{0,\ell}A_{0,\ell'}$ *and* $|[\mathbf{K}^\infty_\mathcal{L}]_{\ell\ell'}| \le B_\ell B_{\ell'}$.

Proof. By Assumption 2.5 the vectors $\theta_{m,(0)}$ are i.i.d. and lie in $\operatorname{supp}\mu$ almost surely. The matrices $\mathbf{\Psi}_\mathcal{L}(\theta_{m,(0)})\mathbf{\Psi}_\mathcal{L}(\theta_{m,(0)})^\top$ are i.i.d. and bounded entrywise by $B_\ell B_{\ell'}$ through Cauchy–Schwarz and Equation (13), so item 1 is statement one of Theorem C.7 applied to the design $\mathcal{L}$, with the same convention governing the mode of convergence. For item 2 the random vectors $(h_\ell(\theta_{m,(0)}))_\ell$ are i.i.d., bounded by $(A_{0,\ell})_\ell$, hence square integrable, and centered by (C); the multivariate Lindeberg–Lévy theorem gives the stated Gaussian limit, which is degenerate exactly when $\mathbf{\Sigma}_\mathcal{L}$ is singular. Joint convergence follows from Slutsky's theorem, which requires no independence between the two components, and indeed they are strongly dependent, being built from the same neurons; it applies because the second limit is *deterministic*. Item 3 is the strong law for the same bounded i.i.d. summands along nested widths, and the weak law otherwise. The three bounds are immediate from Equation (13). □

F.2 Nondeterminism of the infinite-width kernel

Throughout this subsection $\varrho = 1/2$, the standard parameterization of Definition 2.1.

Proposition F.5 (Random limit of the tangent Gram matrix for composite rows)**.** *Assume Assumption 2.5, let the base family satisfy Definition B.1 and* (C)*, let* $\varrho = 1/2$*, and let* $P_1, \dots, P_N$ *be composite rows. Then, as* $d_1 \to \infty$*:*

1. *jointly in all entries,*
$$\mathbf{K}_{(0)} \Longrightarrow \mathbf{K}^\infty_{\boldsymbol{\pi}}(\mathbf{Z}) := \mathbf{D}\boldsymbol{\pi}(\mathbf{Z})\,\mathbf{K}^\infty_\mathcal{L}\,\mathbf{D}\boldsymbol{\pi}(\mathbf{Z})^\top, \qquad \mathbf{Z} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{\Sigma}_\mathcal{L}), \tag{59}$$
where the rows of $\mathbf{D}\boldsymbol{\pi}(\mathbf{q}) \in \mathbb{R}^{N\times n}$ *are* $\nabla\pi_i(\mathbf{q})^\top$*;*
2. *if* π_i *and* π_j *are affine with* $\nabla\pi_i \equiv \boldsymbol{\beta}_i$ *and* $\nabla\pi_j \equiv \boldsymbol{\beta}_j$*, the corresponding limit is the deterministic constant* $\boldsymbol{\beta}_i^\top \mathbf{K}^\infty_\mathcal{L}\boldsymbol{\beta}_j$*, which is the entry* $[\mathbf{K}^\infty_{(0)}]_{ij}$ *supplied by Theorem C.7 for those rows;*
3. *if* $[\mathbf{K}^\infty_{\boldsymbol{\pi}}(\mathbf{Z})]_{ij}$ *is not almost surely constant, then* $[\mathbf{K}_{(0)}]_{ij}$ *converges in probability to no constant whatsoever, so no deterministic infinite-width kernel exists for that entry, at any rate;*
4. *if every* π_i *is a polynomial, then*
$$\mathbb{E}\big[[\mathbf{K}_{(0)}]_{ij}\big] \to \mathbb{E}\big[[\mathbf{K}^\infty_{\boldsymbol{\pi}}(\mathbf{Z})]_{ij}\big], \qquad \operatorname{Var}\big([\mathbf{K}_{(0)}]_{ij}\big) \to \operatorname{Var}\big([\mathbf{K}^\infty_{\boldsymbol{\pi}}(\mathbf{Z})]_{ij}\big). \tag{60}$$
Hence whenever the right-hand variance is positive the entrywise standard deviation of $\mathbf{K}_{(0)}$ *is* $\Theta(1)$ *in* d_1*, in contrast with the* $O(d_1^{-1/2})$ *standard deviation of the affine case.*

Proof. Item 1. At $\varrho = 1/2$ we have $\mathbf{Q} = \widetilde{\mathbf{Q}}$ and the prefactor in Equation (57) equals one. By items 1 and 2 of Lemma F.4, $\big(\mathbf{Q}(\theta_{(0)}), \widehat{\mathbf{K}}_\mathcal{L}(\theta_{(0)})\big) \Rightarrow (\mathbf{Z}, \mathbf{K}^\infty_\mathcal{L})$ in $\mathbb{R}^n \times \mathbb{R}^{n\times n}$. The map $(\mathbf{q}, \mathbf{G}) \mapsto \mathbf{D}\boldsymbol{\pi}(\mathbf{q})\mathbf{G}\mathbf{D}\boldsymbol{\pi}(\mathbf{q})^\top$ is continuous because each π_i is C^1; the continuous mapping theorem and Equation (57) give the claim.

Item 2 is immediate, $\mathbf{D}\boldsymbol{\pi}$ being constant, together with $\mathbb{E}_{z\sim\mu}\langle\nabla(\boldsymbol{\beta}_i^\top\mathbf{h})(z), \nabla(\boldsymbol{\beta}_j^\top\mathbf{h})(z)\rangle = \boldsymbol{\beta}_i^\top\mathbf{K}^\infty_\mathcal{L}\boldsymbol{\beta}_j$.

Item 3. Convergence in probability to c implies convergence in distribution to δ_c; weak limits are unique, so $[\mathbf{K}^\infty_{\boldsymbol{\pi}}(\mathbf{Z})]_{ij} = c$ almost surely, contrary to hypothesis.

Item 4. Fix $r \ge 1$. Rosenthal's inequality for independent mean-zero summands together with $|h_\ell| \le A_{0,\ell}$ on $\operatorname{supp}\mu$ gives a constant C_r with
$$\mathbb{E}|Q_\ell(\theta_{(0)})|^{2r} \le C_r\big[(\mathbb{E}h_\ell^2)^r + d_1^{1-r}\mathbb{E}|h_\ell|^{2r}\big] \le 2C_r A_{0,\ell}^{2r},$$
uniformly in d_1; since $\|\mathbf{Q}\|_2^K \le n^{K/2}\sum_\ell |Q_\ell|^K$ and $\mathbb{E}|Q_\ell|^K \le (\mathbb{E}|Q_\ell|^{2r})^{K/(2r)}$ for $2r \ge K$, every moment of $\|\mathbf{Q}(\theta_{(0)})\|_2$ is bounded uniformly in d_1. Also $\|\widehat{\mathbf{K}}_\mathcal{L}\|_2 \le \operatorname{tr}\widehat{\mathbf{K}}_\mathcal{L} \le \sum_\ell B_\ell^2$ almost surely, the first inequality because $\widehat{\mathbf{K}}_\mathcal{L} \succeq \mathbf{0}$. If π_i has degree p_i then $\|\nabla\pi_i(\mathbf{q})\|_2 \le c_i(1 + \|\mathbf{q}\|_2)^{p_i-1}$, so by Equation (57) $\sup_{d_1}\mathbb{E}|[\mathbf{K}_{(0)}]_{ij}|^4 < \infty$. A family bounded in L^4 is uniformly integrable, and so is the family of its squares; weak convergence then transfers the first two moments. In the affine case $[\mathbf{K}_{(0)}]_{ij}$ is an average of d_1 i.i.d. bounded terms, whence variance $O(d_1^{-1})$. □

The nondegeneracy hypothesis in item 3 is not an extra assumption in disguise: under this paper's own certificate it holds for every genuinely nonlinear polynomial row.

Lemma F.6 (The linear certificate controls both base matrices)**.** *Assume* ($\mathsf{C}_{\mathrm{iso}}$) *and let* $\mathbf{K}^{\infty}_{\mathcal{L},\mathrm{out}}$ *denote the output-weight block of* $\mathbf{K}^{\infty}_{\mathcal{L}}$*, that is* $[\mathbf{K}^{\infty}_{\mathcal{L},\mathrm{out}}]_{\ell\ell'} = \sum_{q=1}^{d_2} \langle [\mathbf{f}_\ell]_q, [\mathbf{f}_{\ell'}]_q \rangle_{L^2(\mu_h)}$*. Then*

$$\boldsymbol{\Sigma}_{\mathcal{L}} = c_{\mathbf{a}} \mathbf{K}^{\infty}_{\mathcal{L},\mathrm{out}}, \qquad \mathbf{K}^{\infty}_{\mathcal{L}} \succeq \mathbf{K}^{\infty}_{\mathcal{L},\mathrm{out}}. \tag{61}$$

Consequently $\mathbf{K}^{\infty}_{\mathcal{L},\mathrm{out}} \succ 0$ *implies both* $\boldsymbol{\Sigma}_{\mathcal{L}} \succ 0$ *and* $\mathbf{K}^{\infty}_{\mathcal{L}} \succ 0$*. That hypothesis is the output-block positive-definiteness certificate of Definition 2.7 applied to the* base family*, not to the composite design; when the base observations are pointwise differential rows and the clauses* $\mathsf{P}_{\mathrm{reg}}$, P_{pd} *of Assumption 2.6 hold, Proposition 3.1 certifies it exactly when the evaluated coefficient vectors of the base observations are linearly independent at every sampled location, which holds for the base family* $\{\partial_t u, u, \partial_x u, \partial_{xx} u\}$ *of the conservation-law row.*

Proof. Because $\partial h_\ell / \partial a_q = [\mathbf{f}_\ell(\boldsymbol{\xi})]_q$, the d_2 output-weight coordinates contribute exactly $\mathbf{K}^{\infty}_{\mathcal{L},\mathrm{out}}$ to Equation (56), that contribution depending on z only through $\boldsymbol{\xi}$, while the remaining $d_0 + 1$ coordinates contribute a positive-semidefinite matrix; this is the second relation. For the first,

$$[\boldsymbol{\Sigma}_{\mathcal{L}}]_{\ell\ell'} = \mathbb{E}\big[\mathbf{f}_\ell(\boldsymbol{\xi})^\top \mathbf{a}\mathbf{a}^\top \mathbf{f}_{\ell'}(\boldsymbol{\xi})\big] = \mathbb{E}_{\mu_h}\big[\mathbf{f}_\ell^\top \mathbb{E}[\mathbf{a}\mathbf{a}^\top \mid \boldsymbol{\xi}] \mathbf{f}_{\ell'}\big] = c_{\mathbf{a}} \mathbb{E}_{\mu_h}\big[\mathbf{f}_\ell^\top \mathbf{f}_{\ell'}\big]. \qquad \square$$

Corollary F.7 (Dichotomy: affine rows concentrate, nonlinear rows do not)**.** *Assume the hypotheses of Proposition F.5 with every* π_i *a polynomial, and assume* $\boldsymbol{\Sigma}_{\mathcal{L}} \succ 0$ *and* $\mathbf{K}^{\infty}_{\mathcal{L}} \succ 0$*; by Lemma F.6 the single condition* $\mathbf{K}^{\infty}_{\mathcal{L},\mathrm{out}} \succ 0$ *suffices. Then for every* $i \in [N]$*,*

$$[\mathbf{K}_{(0)}]_{ii} \text{ has a deterministic weak limit } \iff \pi_i \text{ is affine}.$$

In particular, if one row is genuinely nonlinear then $\mathbf{K}_{(0)}$ *converges in distribution to a genuinely random matrix,* $\mathrm{Var}([\mathbf{K}_{(0)}]_{ii})$ *converges to a strictly positive limit, and there is no population kernel to which a matrix Chernoff argument such as Proposition B.10 could transfer.*

Proof. If π_i is affine the limit is deterministic by item 2 of Proposition F.5. Conversely, write $\mathbf{A} = (\mathbf{K}^{\infty}_{\mathcal{L}})^{1/2}$, which is invertible. By Equation (59), $[\mathbf{K}_{(0)}]_{ii} \Rightarrow \|\mathbf{A}\nabla\pi_i(\mathbf{Z})\|_2^2$. Suppose this limit equals a constant c^2 almost surely. Since $\boldsymbol{\Sigma}_{\mathcal{L}} \succ 0$, the law of $\mathbf{Z}$ has a positive density on $\mathbb{R}^n$, and $\mathbf{z} \mapsto \|\mathbf{A}\nabla\pi_i(\mathbf{z})\|_2^2$ is continuous, so the open set on which it differs from c^2 would have positive probability unless it is empty; hence $\|\mathbf{A}\nabla\pi_i(\mathbf{z})\|_2^2 = c^2$ for *every* $\mathbf{z} \in \mathbb{R}^n$. Let $p_1, \ldots, p_n$ be the components of the polynomial map $\mathbf{A}\nabla\pi_i$, so $\sum_k p_k^2 \equiv c^2$, and let $d = \max_k \deg p_k$. If $d \geq 1$, write P_k for the degree-d homogeneous part of p_k; the degree-$2d$ homogeneous part of $\sum_k p_k^2$ is $\sum_k P_k^2$, which must vanish identically because $\sum_k p_k^2$ is constant and $2d > 0$. A finite sum of squares of real polynomials vanishes identically only if every summand does, so $P_k \equiv 0$ for all k, contradicting $d \geq 1$. Hence $\mathbf{A}\nabla\pi_i$ is constant and, $\mathbf{A}$ being invertible, π_i is affine. The final assertion combines this with item 4 of Proposition F.5. $\square$

Corollary F.8 (Product row against a linear row)**.** *Assume the hypotheses of Proposition F.5. Let* $\pi_i(\mathbf{q}) = \boldsymbol{\beta}_i^\top \mathbf{q} + q_a q_b$ *for some* $a, b \in \mathcal{L}$ *and* $\boldsymbol{\beta}_i \in \mathbb{R}^n$*, and let* π_j *be affine with* $\nabla\pi_j \equiv \boldsymbol{\beta}_j$*. Put* $\boldsymbol{\kappa}_j = \mathbf{K}^{\infty}_{\mathcal{L}} \boldsymbol{\beta}_j$*. Then*

$$[\mathbf{K}_{(0)}]_{ij} \Longrightarrow \mathcal{N}\big(\boldsymbol{\beta}_i^\top \boldsymbol{\kappa}_j,\ s_{ij}^2\big), \qquad s_{ij}^2 = \big\|[\boldsymbol{\kappa}_j]_a h_b + [\boldsymbol{\kappa}_j]_b h_a\big\|_{L^2(\mu)}^2. \tag{62}$$

The limit is nondegenerate exactly when $[\boldsymbol{\kappa}_j]_a h_b + [\boldsymbol{\kappa}_j]_b h_a \neq 0$ *in* $L^2(\mu)$*; the linear part* $\boldsymbol{\beta}_i$*, which carries the* $\partial_t u$ *and* $-\nu\partial_{xx} u$ *terms of a conservation-law residual, shifts the mean and cannot remove the fluctuation. If* $a \neq b$ *and* h_a, h_b *are linearly independent in* $L^2(\mu)$*, then* $s_{ij}^2 > 0$ *whenever* $([\boldsymbol{\kappa}_j]_a, [\boldsymbol{\kappa}_j]_b) \neq (0, 0)$*.*

Proof. Here $\nabla\pi_i(\mathbf{z}) = \boldsymbol{\beta}_i + z_b \mathbf{e}_a + z_a \mathbf{e}_b$, so Equation (59) gives $[\mathbf{K}_{(0)}]_{ij} \Rightarrow \boldsymbol{\beta}_i^\top \boldsymbol{\kappa}_j + [\boldsymbol{\kappa}_j]_a Z_b + [\boldsymbol{\kappa}_j]_b Z_a$, an affine function of a Gaussian vector and hence Gaussian. With $\mathbf{v} = [\boldsymbol{\kappa}_j]_b \mathbf{e}_a + [\boldsymbol{\kappa}_j]_a \mathbf{e}_b$, the variance is $\mathbf{v}^\top \boldsymbol{\Sigma}_{\mathcal{L}} \mathbf{v} = \mathbb{E}_{z\sim\mu}([\boldsymbol{\kappa}_j]_a h_b(z) + [\boldsymbol{\kappa}_j]_b h_a(z))^2$ by Equation (56). $\square$

Example F.9 (Explicit nondegeneracy for a convective row). Take $d_2 = 1$, assume clause P_{pd} of Assumption 2.6, so that σ is real analytic and nonpolynomial, $\gamma > 0$, and $\mathrm{int}(\mathrm{supp}\,\mu_h) \neq \varnothing$, and assume ($\mathsf{C}_{\mathrm{iso}}$). Let a and b index the value and the ∂_{x_1} observation at one interior collocation point $\mathbf{x}_i$, so that the convective term of a quadratic-flux conservation law is $Q_a Q_b$. Then h_a and h_b are linearly independent in $L^2(\mu)$. Indeed, a vanishing combination $v_a h_a + v_b h_b = 0$ reads $\mathbb{E}[a^2 G(\boldsymbol{\xi})^2] = 0$ with $G(\mathbf{w}, b) = v_a \sigma(\mathbf{w}^\top \mathbf{x}_i + \gamma b) + v_b w_1 \sigma'(\mathbf{w}^\top \mathbf{x}_i + \gamma b)$; by ($\mathsf{C}_{\mathrm{iso}}$) this forces $G = 0$ μ_h-almost surely, hence by continuity on $\mathrm{supp}\,\mu_h$, which contains a nonempty open set, and the identity theorem for real-analytic functions gives $G \equiv 0$ on $\mathbb{R}^{d_0+1}$. Since $\gamma > 0$, the substitution $v = \gamma b$ sweeps all of $\mathbb{R}$, and G is then a function of the form Equation (24) with a single physical site and index set contained in $\{\mathbf{0}, \mathbf{e}_1\}$; for a single site the conclusion of Lemma C.8 is that the associated function is not identically zero, which forbids $G \equiv 0$ unless $v_a = v_b = 0$. This settles the L^2 half of the nondegeneracy condition in Corollary F.8; the condition on $\boldsymbol{\kappa}_j$ depends on the particular affine row j and is not verified here, which is why the unconditional statement is the diagonal one in Corollary F.7.

Remark F.10 (Why the linear proof architecture is unavailable). For an affine row, $\nabla_{\theta_m} P_i = d_1^{-1/2} \nabla h_i(\theta_m)$ depends on neuron m alone, the per-neuron Gram blocks are i.i.d., $\mathbf{K}_{(0)}$ is an empirical mean, and both the law of large numbers in Theorem C.7 and the matrix Chernoff bound of Proposition B.10 apply verbatim. For a composite row, Equation (57) multiplies the same block by the global factor $\nabla \pi_i(\mathbf{Q}(\theta))$, which all d_1 neurons share. The blocks remain exchangeable but are no longer independent, and the shared factor is precisely the quantity governed by the central limit theorem: it fluctuates at scale $\Theta(1)$ and does not self-average at any width. Conditionally on $\mathbf{Q}(\theta_{(0)})$ that factor is frozen and $\widehat{\mathbf{K}}_{\mathcal{L}}(\theta_{(0)})$ is the only remaining source of randomness; the conditioning also destroys the independence of its summands, so a quenched analysis would have to re-establish its concentration rather than inherit it. Such an analysis is not excluded here, and we do not carry one out.

Remark F.11 (No deterministic spectral floor in the limit). The eigenvalue floor that drives Theorem 4.2 is lost as well, not only the population kernel. Suppose row i is quadratic with $\nabla \pi_i(\mathbf{z}) = \boldsymbol{\beta}_i + 2\mathbf{C}_i \mathbf{z}$ and $\boldsymbol{\beta}_i \in \mathrm{Range}(\mathbf{C}_i)$, and let $\boldsymbol{\Sigma}_{\mathcal{L}} \succ 0$. The hypothesis on $\boldsymbol{\beta}_i$ holds for a pure product row $\pi_i(\mathbf{q}) = q_a q_b$; it fails for a residual row whose linear part lies outside $\mathrm{Range}(\mathbf{C}_i)$, such as $\pi_i(\mathbf{q}) = s_i(q_1 + q_2 q_3 - \nu q_4)$, for which this remark says nothing. Under it, $\{\mathbf{z} : \nabla \pi_i(\mathbf{z}) = \mathbf{0}\}$ is a nonempty affine set, so for every $\varepsilon > 0$ the open set $\{\mathbf{z} : \|\nabla \pi_i(\mathbf{z})\|_2 < \varepsilon\}$ is nonempty and is charged by $\mathbf{Z}$. Since $\lambda_{\min}$ is continuous, the continuous mapping theorem and Equation (59) give $\lambda_{\min}(\mathbf{K}_{(0)}) \Rightarrow \lambda_{\min}(\mathbf{K}^{\infty}_{\boldsymbol{\pi}}(\mathbf{Z}))$, and the latter is at most $\|\mathbf{A} \nabla \pi_i(\mathbf{Z})\|_2^2$ with $\mathbf{A} = (\mathbf{K}^{\infty}_{\mathcal{L}})^{1/2}$, the smallest eigenvalue of a symmetric matrix being at most any diagonal entry. Hence 0 lies in the support of the limiting law of $\lambda_{\min}$, and the portmanteau theorem applied to an open set gives, *for each* $\varepsilon > 0$, a constant $p_\varepsilon > 0$ with

$$\liminf_{d_1 \to \infty} \Pr\{\lambda_{\min}(\mathbf{K}_{(0)}) < \varepsilon\} \geq p_\varepsilon.$$

The constant p_ε decreases to zero with ε; what matters is that it does not decrease with d_1, so no deterministic $\lambda_0 > 0$ can play the role it plays in Theorem 4.2. A p-th-power boundary row $\pi_i(\mathbf{q}) = q_\ell^p$, $p \geq 2$, is the extreme case: $[\mathbf{K}_{(0)}]_{ii} \Rightarrow p^2 Z_\ell^{2p-2} [\mathbf{K}^{\infty}_{\mathcal{L}}]_{\ell\ell}$, whose law accumulates at 0 whenever $[\boldsymbol{\Sigma}_{\mathcal{L}}]_{\ell\ell} > 0$.

F.3 No width exponent repairs the polynomial case

We now return to general ϱ and ask whether a different width exponent restores a deterministic limit. It does not. The reason is structural: by Equation (58) the exponent multiplies a single deterministic prefactor and leaves the limiting object untouched.

Theorem F.12 (Trichotomy in the width exponent)**.** *Assume Assumption 2.5, let the base family satisfy Definition B.1 and* (C)*, and fix monomials J, J' of degrees $q, q' \geq 1$. Define the homogeneous polynomial of degree $q + q' - 2$*

$$\Phi_{J,J'}(\mathbf{z}) = \sum_{r=1}^{q} \sum_{t=1}^{q'} \Big(\prod_{s \neq r} z_{j_s} \Big) \Big(\prod_{s \neq t} z_{l_s} \Big) [\mathbf{K}^{\infty}_{\mathcal{L}}]_{j_r l_t}. \tag{63}$$

Then, with $\mathbf{Z} \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{\Sigma}_{\mathcal{L}})$,

$$d_1^{-(q+q')(1/2-\varrho)} K^{J,J'}(\theta_{(0)}) \Longrightarrow \Phi_{J,J'}(\mathbf{Z}), \qquad Q_\ell(\theta_{(0)}) = d_1^{1/2-\varrho} \widetilde{Q}_\ell(\theta_{(0)}), \tag{64}$$

and $\Phi_{J,J'}$ does not depend on ϱ. Consequently:

1. *if $\varrho > 1/2$, then $Q_\ell(\theta_{(0)}) \to 0$ and $K^{J,J'}(\theta_{(0)}) \to 0$ in probability, the latter at the order $d_1^{(q+q')(1/2-\varrho)}$, which is exact whenever $\Phi_{J,J'}$ does not vanish identically on* $\mathrm{Range}(\mathbf{\Sigma}_\mathcal{L})$*;*
2. *if $\varrho = 1/2$, then $Q_\ell(\theta_{(0)}) \Rightarrow Z_\ell$ and $K^{J,J'}(\theta_{(0)}) \Rightarrow \Phi_{J,J'}(\mathbf{Z})$, both of order one, and the kernel limit is deterministic if and only if $\Phi_{J,J'}$ is constant on* $\mathrm{Range}(\mathbf{\Sigma}_\mathcal{L})$*. For $q = q' = 1$ both products in Equation* (63) *are empty, so $\Phi_{J,J'} \equiv [\mathbf{K}_\mathcal{L}^\infty]_{j_1 l_1}$ and the limit is always deterministic; for $q + q' \geq 3$ a constant $\Phi_{J,J'}$ must vanish on* $\mathrm{Range}(\mathbf{\Sigma}_\mathcal{L})$*, so the limit is random unless it is the constant* 0*;*
3. *if $\varrho < 1/2$, then $|Q_\ell(\theta_{(0)})| \to \infty$ in probability whenever $[\mathbf{\Sigma}_\mathcal{L}]_{\ell\ell} > 0$, and $|K^{J,J'}(\theta_{(0)})| \to \infty$ in probability unless $\Phi_{J,J'}$ vanishes identically on* $\mathrm{Range}(\mathbf{\Sigma}_\mathcal{L})$*.*

Proof. Substituting $Q_\ell = d_1^{1/2-\varrho}\widetilde{Q}_\ell$ in Equation (58) produces $(q-1)+(q'-1)$ factors $d_1^{1/2-\varrho}$, so the total prefactor is $d_1^{1-2\varrho+(q+q'-2)(1/2-\varrho)} = d_1^{(q+q')(1/2-\varrho)}$, and the remaining expression is $\sum_{r,t}(\prod_{s\neq r}\widetilde{Q}_{j_s})(\prod_{s\neq t}\widetilde{Q}_{l_s})[\widehat{\mathbf{K}}_\mathcal{L}]_{j_r l_t}$. By Lemma F.4 and the continuous mapping theorem applied to the polynomial map $(\mathbf{z}, \mathbf{G}) \mapsto \sum_{r,t}(\prod_{s\neq r} z_{j_s})(\prod_{s\neq t} z_{l_s})G_{j_r l_t}$, this converges weakly to $\Phi_{J,J'}(\mathbf{Z})$, giving Equation (64). Item 1 follows by Slutsky's theorem. For item 3, write $Y_n = d_1^{-(q+q')(1/2-\varrho)}K^{J,J'}(\theta_{(0)})$, so that $Y_n \Rightarrow \Phi_{J,J'}(\mathbf{Z})$. A polynomial that does not vanish identically on $V = \mathrm{Range}(\mathbf{\Sigma}_\mathcal{L})$ has a Lebesgue-null zero set in V, and the law of $\mathbf{Z}$ has a density on V, so $\Phi_{J,J'}(\mathbf{Z}) \neq 0$ almost surely. Fix T and $\varepsilon > 0$ and set $\varepsilon_n = T d_1^{-(q+q')(1/2-\varrho)} \to 0$. Once $\varepsilon_n < \varepsilon$, the portmanteau theorem for the open set $\{|\cdot| > \varepsilon\}$ gives $\liminf_n \Pr\{|K^{J,J'}(\theta_{(0)})| > T\} = \liminf_n \Pr\{|Y_n| > \varepsilon_n\} \geq \Pr\{|\Phi_{J,J'}(\mathbf{Z})| > \varepsilon\}$, and letting $\varepsilon \downarrow 0$ shows that $|K^{J,J'}(\theta_{(0)})| \to \infty$ in probability. The divergence of Q_ℓ is the same argument with Z_ℓ in place of $\Phi_{J,J'}(\mathbf{Z})$.

For item 2 the prefactor is 1. Write $V = \mathrm{Range}(\mathbf{\Sigma}_\mathcal{L})$, so the law of $\mathbf{Z}$ is supported on V and equivalent to Lebesgue measure there. If $\Phi_{J,J'} \equiv c$ on V then $\Phi_{J,J'}(\mathbf{Z}) = c$ almost surely; conversely, if $\Phi_{J,J'}(\mathbf{Z}) = c$ almost surely then the continuous function $\Phi_{J,J'} - c$ vanishes Lebesgue-almost everywhere on V and hence identically on V. When $q + q' \geq 3$, $\Phi_{J,J'}$ is homogeneous of degree $D = q + q' - 2 \geq 1$ and V is a linear subspace, so $t\mathbf{z} \in V$ and $c = \Phi_{J,J'}(t\mathbf{z}) = t^D c$ for all $\mathbf{z} \in V$ and $t \in \mathbb{R}$; taking $t = 0$ forces $c = 0$. When $q = q' = 1$ the polynomial is the stated constant. □

Corollary F.13 (No exponent yields a nonzero deterministic polynomial kernel)**.** *Under the hypotheses of Theorem F.12, let $q + q' \geq 3$ and let $\varrho \geq 1/2$ be fixed. If $K^{J,J'}(\theta_{(0)})$ converges in probability to a deterministic limit, that limit is* 0*. At the degree-matched exponent $\varrho_p = (2p-1)/(2p)$ with $q = q' = p \geq 2$ one has, at every width,*

$$K^{J,J'}(\theta_{(0)}) = d_1^{-(p-1)}\,\Phi_{J,J'}^{(d_1)}, \qquad Q_\ell(\theta_{(0)}) = d_1^{-(p-1)/(2p)}\widetilde{Q}_\ell(\theta_{(0)}), \tag{65}$$

where $\Phi_{J,J'}^{(d_1)} \Rightarrow \Phi_{J,J'}(\mathbf{Z})$; so both the kernel block and the prediction converge to zero, the displayed orders being attained whenever $\Phi_{J,J'}$ does not vanish identically on $\mathrm{Range}(\mathbf{\Sigma}_\mathcal{L})$ *and $[\mathbf{\Sigma}_\mathcal{L}]_{\ell\ell} > 0$. The remaining exponents, $\varrho < 1/2$, already make the prediction itself diverge.*

Proof. For $\varrho > 1/2$ the prefactor in Equation (64) tends to 0 while the residual factor is bounded in probability, so the limit is 0. For $\varrho = 1/2$, convergence in probability to c implies convergence in distribution to δ_c, so $\Phi_{J,J'}(\mathbf{Z}) = c$ almost surely, and the dichotomy established in item 2 of Theorem F.12 forces $c = 0$. The displayed rates are $2p(1/2 - \varrho_p) = -(p-1)$ and $1/2 - \varrho_p = -(p-1)/(2p)$. □

Proposition F.14 (The uncentered branch)**.** *Drop* (C) *and let $m_\ell = \mathbb{E}_{z\sim\mu} h_\ell(z)$. Then, almost surely along nested widths as in Theorem C.7 and in probability under fresh resampling,*

$$d_1^{1-(q+q')(1-\varrho)}K^{J,J'}(\theta_{(0)}) \longrightarrow \sum_{r=1}^{q}\sum_{t=1}^{q'}\Big(\prod_{s\neq r} m_{j_s}\Big)\Big(\prod_{s\neq t} m_{l_s}\Big)[\mathbf{K}_\mathcal{L}^\infty]_{j_r l_t}, \qquad d_1^{\varrho-1}Q_\ell(\theta_{(0)}) \to m_\ell. \tag{66}$$

At $\varrho = \varrho_p$ and $q = q' = p$ the kernel prefactor is exactly 1 *and the limit is deterministic and generically nonzero, while for a block of total degree $q + q' < 2p$ the prefactor is $d_1^{(q+q')/(2p)-1} \to 0$.*

This branch is nevertheless inadmissible on two counts: the prediction grows as $d_1^{1-\varrho_p} = d_1^{1/(2p)}$; and forcing an order-one prediction by taking $\varrho = 1$ makes every block, the affine ones included, of order d_1^{-1}. It also contradicts the centering hypothesis used by Lemma E.1 and Theorem 4.2.

Proof. Write $Q_\ell = d_1^{1-\varrho}\bar{Q}_\ell$ with $\bar{Q}_\ell = d_1^{-1}\sum_m h_\ell(\theta_m) \to m_\ell$ almost surely by item 3 of Lemma F.4. Substituting in Equation (58) gives the prefactor $d_1^{1-2\varrho+(q+q'-2)(1-\varrho)} = d_1^{(q+q')(1-\varrho)-1}$ and the stated limit, since $\widehat{\mathbf{K}}_{\mathcal{L}} \to \mathbf{K}_{\mathcal{L}}^{\infty}$ almost surely. At $\varrho = \varrho_p$ the exponent is $(q+q')/(2p) - 1$, which is 0 for $q = q' = p$. □

Remark F.15 (What degree matching does and does not achieve). Proposition F.14 isolates the one regime in which the degree-matched exponent behaves as intended. Without the p-th-power boundary transform, a residual–boundary block of a degree-p constraint has total degree $q + q' = p + 1$, hence order $d_1^{(1-p)/(2p)} \to 0$ against an order-one interior block; raising the boundary row to the p-th power restores $q + q' = 2p$ for every block and equalizes the orders. That accounting is correct *only* without centering. Under (C) the ordering reverses, because the block order $d_1^{(q+q')(1/2-\varrho)}$ *decreases* in total degree when $\varrho > 1/2$; the transform still equalizes the orders, but at the common value $d_1^{-(p-1)}$, that is, it equalizes them at zero. At $\varrho = 1/2$ with centering there is no hierarchy at all, since the prefactor is 1 for every (q, q'), so the transform is not needed for its stated purpose; what fails there is determinism, not order. Finally, if one allows a width-dependent mean $m_\ell \asymp d_1^{-\delta}$ with $\delta \in [0, 1/2]$, outside Assumption 2.5, then an order-one prediction forces $\varrho = 1 - \delta$ and the block order collapses to $d_1^{2\delta-1}$ independently of q and q'; a nonvanishing kernel then forces $\delta = 1/2$, which is exact centering, which is the pure Gaussian-fluctuation regime of Section F.2. The obstruction is centering together with degree, not the exponent. We record this last observation as a scaling heuristic only; it is used nowhere else.

Remark F.16 (Reachability of the uncentered branch). Proposition F.14 requires $m_\ell = \mathbb{E}_{z\sim\mu} h_\ell(z) \neq 0$, and $m_\ell = \mathbb{E}[\mathbf{a}]^\top \mathbb{E}[\mathbf{f}_\ell]$ when the output weight is drawn independently of $\boldsymbol{\xi}$. Giving the output weight a nonzero mean is therefore not sufficient. If the activation is odd and the hidden law μ_h is symmetric about the origin, then $\mathbb{E}[\mathbf{f}_\ell] = \mathbf{0}$ and $m_\ell = 0$ however $\mathbf{a}$ is distributed; the prediction is then pure fluctuation and, at the degree-matched exponent, decays as $d_1^{-1/(2p)}$ rather than growing. Reaching the branch requires an asymmetric hidden law or a nonodd activation.

Remark F.17 (Numerical confirmation). The three displayed rates were checked directly for $\sigma = \tanh$, $\gamma = 1$, $d_0 = d_2 = 1$, collocation points $x_1 = 0.3$ and $x_2 = -0.7$, per-neuron parameters $\mathrm{Unif}[-1, 1]^3$ in the centered runs, widths $d_1 \in \{64, 256, 1024, 4096, 16384\}$, and 100 independent initializations per width. Over that 256-fold range of widths, $d_1\,\mathrm{Var}$ of the linear-row Gram entry stays between 4.6×10^{-2} and 6.3×10^{-2}, while the variance of the product-row entry stays between 1.8×10^{-2} and 3.9×10^{-2}, with no trend comparable to the 256-fold decay of a concentrating quantity; at $d_1 = 16{,}384$ the latter exceeds the former by about 5.0×10^3. This is the dichotomy of Corollary F.7. At $p = 2$ and $\varrho_2 = 3/4$ with centered output weights, d_1 times the degree-$(2, 2)$ block stays between 1.2×10^{-1} and 1.5×10^{-1}, confirming the rate $d_1^{-(p-1)}$ of Corollary F.13. The uncentered runs draw the output weights from $\mathrm{Unif}[0, 2]$ (mean 1). With the hidden bias law also shifted to $\mathrm{Unif}[0, 2]$, so that Remark F.16 applies, the seed-mean absolute prediction divided by $d_1^{1/4}$ stays between 0.651 and 0.658, confirming the growth rate of Proposition F.14; with the symmetric hidden bias law $\mathrm{Unif}[-1, 1]$ it instead decays, from 1.7×10^{-1} at $d_1 = 64$ to 4.7×10^{-2} at $d_1 = 16{,}384$, as that remark predicts. The run is the nonlinear kernel-variance experiment of the accompanying code, under its tracked configuration.

Remark F.18 (Scope of the negative results). Proposition F.5, Corollaries F.7 and F.13, and Theorem F.12 concern the tangent Gram matrix at initialization. They say nothing about $\mathbf{K}(\theta_{(k)})$ for $k \geq 1$, where $\mathbf{Q}(\theta)$ has drifted and is no longer a sum of i.i.d. terms, and they do not assert that gradient descent fails on nonlinear PINNs; in particular they do not contradict the global convergence analysis for nonlinear PDEs of Sirignano et al. (2026), nor the stability-based error analyses of Mishra & Molinaro (2023); De Ryck et al. (2024), none of which routes through a deterministic tangent kernel. What they rule out is the specific proof architecture used here, which follows the two-layer analysis of Du et al. (2019b) and the limiting-kernel picture of Jacot et al. (2018); Lee et al. (2019) as specialized to PINNs by Wang et al. (2022): a deterministic population kernel, a width threshold obtained by transferring a population eigenvalue through a matrix concentration inequality, and a spectral floor propagated along the trajectory. Under a single genuinely nonlinear polynomial

row none of the three exists in the required form. What remains open, and is not excluded by anything above, is whether a random but uniformly positive definite limiting kernel would still yield convergence, and whether layerwise or coordinatewise scalings outside the one-parameter family of Definition F.1 behave differently.

F.4 Toward depth: the assumptions, the implication, and what remains conjectural

Every result above is proved for the two-layer parameterization of Definition 2.1. This subsection isolates, for a reader who wants to know precisely where the depth-H difficulty lies, one fact that is proved unconditionally at every depth, one regularity condition on the full network Jacobian that would suffice, the substitution proposition that this condition implies, and the two statements that remain conjectural. We claim no depth-H theorem: apart from Lemma F.20, every displayed hypothesis below is *assumed*, not derived.

F.4.1 The depth-H network

Definition F.19 (Depth-H network)**.** Fix a depth $H \geq 1$ and hidden widths $d^{[1]}, \dots, d^{[H]}$, and set $d^{[0]} = d_0$ and $d^{[H+1]} = d_2$, the input and output dimensions of Definition 2.1; the bracketed superscript is a layer index, so no $d^{[l]}$ with $l \in [H]$ is to be read as one of the two-layer symbols d_0, d_1, d_2, and $d^{[1]} = d_1$ when $H = 1$. For $l \in [H]$ let $\mathbf{W}^{[l]} \in \mathbb{R}^{d^{[l]} \times d^{[l-1]}}$ and $\mathbf{b}^{[l]} \in \mathbb{R}^{d^{[l]}}$, and let $\mathbf{W}^{[H+1]} \in \mathbb{R}^{d^{[H+1]} \times d^{[H]}}$. With the activation σ and the fixed bias scale $\gamma \geq 0$ of Definition 2.1, define

$$\begin{aligned} \mathbf{t}^{[1]}(\mathbf{x}) &= \mathbf{W}^{[1]}\mathbf{x} + \gamma \mathbf{b}^{[1]}, \\ \mathbf{t}^{[l]}(\mathbf{x}) &= \frac{1}{\sqrt{d^{[l-1]}}} \mathbf{W}^{[l]} \mathbf{y}^{[l-1]}(\mathbf{x}) + \gamma \mathbf{b}^{[l]}, \qquad l = 2, \dots, H, \\ \mathbf{y}^{[l]}(\mathbf{x}) &= \sigma\big(\mathbf{t}^{[l]}(\mathbf{x})\big), \qquad l \in [H], \\ \mathbf{u}(\mathbf{x}; \theta) &= \frac{1}{\sqrt{d^{[H]}}} \mathbf{W}^{[H+1]} \mathbf{y}^{[H]}(\mathbf{x}), \end{aligned} \tag{67}$$

with σ applied entrywise, $\theta^{[l]} = (\mathbf{W}^{[l]}, \mathbf{b}^{[l]})$ for $l \in [H]$, $\theta^{[H+1]} = \mathbf{W}^{[H+1]}$, and $\theta = (\theta^{[1]}, \dots, \theta^{[H+1]}) \in \mathbb{R}^D$ with $D = \sum_{l \in [H]} d^{[l]}(d^{[l-1]} + 1) + d^{[H+1]} d^{[H]}$. Every listed parameter is trained. The rows of Definition 2.2, the loss, the residual, and the definitions $[\mathbf{J}(\theta)]_{i,:} = \nabla_\theta P_i(\theta)^\top$ and $\mathbf{K}(\theta) = \mathbf{J}(\theta)\mathbf{J}(\theta)^\top$ of Definition 2.3 apply to Equation (67) unchanged, with $\mathbf{J}(\theta) \in \mathbb{R}^{N \times D}$ in place of the two-layer shape $\mathbb{R}^{N \times d_1 d_\theta}$ recorded there, which is the only clause of that definition that refers to the architecture. Write $d_{\min} = \min_{l \in [H]} d^{[l]}$, let $\mathbf{J}^{[l]}(\theta)$ denote the column block of $\mathbf{J}(\theta)$ indexed by the coordinates of $\theta^{[l]}$, and set

$$\mathbf{K}^{[l]}(\theta) = \mathbf{J}^{[l]}(\theta)\mathbf{J}^{[l]}(\theta)^\top \in \mathbb{R}^{N \times N}, \qquad l \in [H+1]. \tag{68}$$

At initialization the coordinates of $\theta_{(0)}$ are drawn independently across layers, those within layer l being i.i.d. from a width-independent, compactly supported scalar law $\mu^{[l]}$.

At $H = 1$, Equation (67) reads $\mathbf{u}(\mathbf{x}; \theta) = d_1^{-1/2} \mathbf{W}^{[2]} \sigma(\mathbf{W}^{[1]}\mathbf{x} + \gamma \mathbf{b}^{[1]})$, which is Equation (1) written in matrix form: the mth row of $(\mathbf{W}^{[1]}, \mathbf{b}^{[1]})$ is $\boldsymbol{\xi}_m = (\mathbf{w}_m, b_m)$, the mth column of $\mathbf{W}^{[2]}$ is $\mathbf{a}_m$, and the first layer carries no width factor, exactly as in Definition 2.1. The initialization clause is the special case of Assumption 2.5 in which $\mu = (\mu^{[1]})^{\otimes(d_0+1)} \otimes (\mu^{[2]})^{\otimes d_2}$, a law on $\mathbb{R}^{d_\theta}$ with independent coordinates, and $\int z \, d\mu^{[H+1]}(z) = 0$ is the depth analogue of the conditional output-weight centering $\mathbb{E}[\mathbf{a} \mid \boldsymbol{\xi}] = \mathbf{0}$, which the two-layer analysis imposes only for gradient descent. The partition Equation (68) at $H = 1$ is the hidden/output partition behind Equation (5).

Two structural features of Equation (67) are visible immediately. First, unit m of layer $l \geq 2$ carries the parameter block $\big((\mathbf{W}^{[l]})_{m,:}, b_m^{[l]}\big) \in \mathbb{R}^{d^{[l-1]}+1}$, whose dimension *grows* with the width of the layer below, whereas at $H = 1$ every unit carries the fixed number $d_\theta = d_0 + d_2 + 1$ of coordinates. The per-unit map h_i of Equation (12) therefore has no depth-H domain: for $H \geq 2$ the row-regularity condition of Definition B.1 cannot even be typed, let alone verified. Second, for $H \geq 2$ the prediction $P_i(\theta)$ is not a sum of independent per-unit terms, because the hidden layers couple units; the strong

law that produces $\mathbf{K}^{\infty}_{(0)}$ in Theorem C.7 and the matrix-Chernoff step of Proposition B.10 both lose their hypothesis.

What replaces Equation (12) is a bilinear contraction of two empirical Gram matrices attached to a single layer: a forward matrix built from $\mathbf{y}^{[l-1]}$ and a backward matrix built from $\partial \mathbf{u}/\partial \mathbf{t}^{[l]}$. For a scalar output ($d^{[H+1]} = 1$) and value rows only ($r_{\max} = 0$), a direct computation from Equation (67) gives, for $l \in [H]$, the Schur-product form

$$\mathbf{K}^{[l]} = \mathbf{S}\Big[\big(\widehat{\mathbf{G}}^{[l-1]} + \gamma^2 \mathbf{1}\mathbf{1}^\top\big) \odot \widehat{\mathbf{B}}^{[l]}\Big]\mathbf{S}, \qquad \begin{aligned} \widehat{\mathbf{G}}^{[l-1]}_{ij} &= \tfrac{1}{d^{[l-1]}}\big\langle \mathbf{y}^{[l-1]}(\mathbf{x}_i), \mathbf{y}^{[l-1]}(\mathbf{x}_j)\big\rangle, \\ \widehat{\mathbf{B}}^{[l]}_{ij} &= \Big\langle \tfrac{\partial u(\mathbf{x}_i)}{\partial \mathbf{t}^{[l]}}, \tfrac{\partial u(\mathbf{x}_j)}{\partial \mathbf{t}^{[l]}} \Big\rangle, \end{aligned} \tag{69}$$

where $\odot$ is the entrywise product, $\mathbf{S} = \mathrm{diag}(s_i)$, $\mathbf{1} \in \mathbb{R}^N$ is the all-ones vector, and $\widehat{\mathbf{G}}^{[0]}_{ij} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$ without the width factor; the output block, which carries no bias, is $\mathbf{K}^{[H+1]} = \mathbf{S}\widehat{\mathbf{G}}^{[H]}\mathbf{S}$. This is the structure analyzed by Du et al. (2019a), up to their bias-free scaling, in which $\gamma = 0$ and the term $\gamma^2 \mathbf{1}\mathbf{1}^\top$ is absent, their uniform hidden width, and their placement of the normalizing factors inside the hidden layers rather than on the output map. For $r_{\max} \geq 1$ both factors must be enlarged to the jet of order $r_{\max}$: the forward factor records $\partial^{\beta}_{\mathbf{x}} \mathbf{y}^{[l-1]}(\mathbf{x}_i)$ and the backward factor the corresponding sensitivities, for $|\beta| \leq r_{\max}$, and the Schur product is replaced by a contraction of the two jet blocks against the row coefficients $a_{iq\alpha}(\mathbf{x}_i)$. This jet enlargement is the PINN-specific part of the depth problem; it is absent from Du et al. (2019a); Allen-Zhu et al. (2019); Zou et al. (2020), all of which treat an ordinary supervised loss at differential order zero.

The residue of independence is thinner than this presentation may suggest, and it is worth stating exactly, because it decides which layer a concentration argument may even be about. Let $\mathcal{F}_{<l}$ be the σ-algebra generated by $\theta^{[1]}, \dots, \theta^{[l-1]}$. Conditionally on $\mathcal{F}_{<l}$, the summands of $\widehat{\mathbf{G}}^{[l]}$ indexed by the units $p \in [d^{[l]}]$ of layer l are i.i.d., for every l and for the jet enlargement as well, since $\partial^{\beta}_{\mathbf{x}} \mathbf{y}^{[l]}_p$ depends only on the parameters $\big((\mathbf{W}^{[l]})_{p,:}, b^{[l]}_p\big)$ of unit p and on the $\mathcal{F}_{<l}$-measurable jets of $\mathbf{y}^{[l-1]}$ at the sampled sites. No conditioning does the same for $\widehat{\mathbf{B}}^{[l]}$ when $l < H$. By Equation (67), $\partial u/\partial t^{[l]}_p = \sigma'(t^{[l]}_p)\, \partial u/\partial y^{[l]}_p$, and $\partial u/\partial y^{[l]}_p$ is computed through $\mathbf{t}^{[l+1]} = (d^{[l]})^{-1/2}\mathbf{W}^{[l+1]}\mathbf{y}^{[l]} + \gamma \mathbf{b}^{[l+1]}$, which depends on *every* unit of layer l; conditioning on the layers above therefore does not decouple the summands over p, and neither does any other conditioning. At $l = H$ the chain collapses: $\partial u_q/\partial t^{[H]}_p = (d^{[H]})^{-1/2}\mathbf{W}^{[H+1]}_{q,p}\sigma'(t^{[H]}_p)$ factors through the single enlarged unit block $\big((\mathbf{W}^{[H]})_{p,:}, b^{[H]}_p, \mathbf{W}^{[H+1]}_{:,p}\big)$, and the factor $(d^{[H]})^{-1/2}$ makes $\widehat{\mathbf{B}}^{[H]}$ an average over those blocks. Consequently exactly two blocks are, conditionally on $\mathcal{F}_{<H}$, averages over the units of layer H of i.i.d. terms: $\mathbf{K}^{[H+1]} = \mathbf{S}\widehat{\mathbf{G}}^{[H]}\mathbf{S}$ directly, and $\mathbf{K}^{[H]}$ because the factor $\widehat{\mathbf{G}}^{[H-1]} + \gamma^2 \mathbf{1}\mathbf{1}^\top$ is $\mathcal{F}_{<H}$-measurable and multiplies $\widehat{\mathbf{B}}^{[H]}$ entrywise. Even for those two the conditional law is not the unconditional one and the conditioning field is itself random, so neither the strong law of Theorem C.7 nor the matrix-Chernoff bound of Proposition B.10 applies to $\mathbf{K}^{[l]}$ at any l; what the structure supplies is the only visible route to clause (iv) of Definition F.21 below, and the reason to look for the certified layer among $l_\star \in \{H, H+1\}$. Du et al. (2019a) work with the block of the last hidden layer, a choice they present as a simplification costing a minor degradation of the rate.

F.4.2 What is proved at every depth: the layerwise decomposition

Lemma F.20 (Layerwise decomposition of the tangent Gram matrix)**.** *Let $\mathbf{J}(\theta) \in \mathbb{R}^{N \times D}$ be any prediction Jacobian and partition its columns into blocks $\mathbf{J}^{[1]}, \dots, \mathbf{J}^{[H+1]}$, for instance by Equation* (68)*. Then, for every θ,*

$$\mathbf{K}(\theta) = \sum_{l=1}^{H+1} \mathbf{K}^{[l]}(\theta), \qquad \mathbf{K}^{[l]}(\theta) \succeq 0, \tag{70}$$

and consequently

$$\ker \mathbf{K}(\theta) = \bigcap_{l=1}^{H+1} \ker \mathbf{K}^{[l]}(\theta), \qquad \lambda_{\min}(\mathbf{K}(\theta)) \geq \sum_{l=1}^{H+1} \lambda_{\min}(\mathbf{K}^{[l]}(\theta)) \geq \max_{l \in [H+1]} \lambda_{\min}(\mathbf{K}^{[l]}(\theta)). \tag{71}$$

In particular $\mathbf{K}(\theta) \succ 0$ *as soon as* $\mathbf{K}^{[l_\star]}(\theta) \succ 0$ *for a single* $l_\star$*: certifying one layer certifies the whole kernel. The same statements hold for the limiting objects whenever the blockwise limits exist.*

Proof. Splitting the product $\mathbf{J}\mathbf{J}^\top$ along the column partition gives Equation (70), and each $\mathbf{K}^{[l]} = \mathbf{J}^{[l]}\mathbf{J}^{[l]\top}$ is a Gram matrix, hence positive semidefinite. For $\mathbf{c} \in \mathbb{R}^N$,

$$\mathbf{c}^\top \mathbf{K}(\theta)\mathbf{c} = \sum_{l=1}^{H+1} \|\mathbf{J}^{[l]}(\theta)^\top \mathbf{c}\|_2^2,$$

a sum of nonnegative terms, so it vanishes if and only if every term vanishes; and for a positive semidefinite $\mathbf{M}$ one has $\mathbf{M}\mathbf{c} = 0$ if and only if $\mathbf{c}^\top \mathbf{M}\mathbf{c} = \|\mathbf{M}^{1/2}\mathbf{c}\|_2^2 = 0$, which turns that equivalence into the kernel identity in Equation (71). The eigenvalue bound is superadditivity of $\lambda_{\min}$ over a sum of symmetric matrices, a case of Weyl's inequality; the second inequality holds because every omitted summand is nonnegative, each $\mathbf{K}^{[l]}$ being positive semidefinite. □

The lemma is unconditional: no assumption on σ, on the initialization, on the rows, or on the widths enters it, and it holds at every iterate, not only at initialization. At $H = 1$ its sum form is exactly Equation (4), and the eigenvalue inequality Equation (5) is its consequence, so Definition 2.7 extends verbatim: a depth-H positive-definiteness certificate is a proof that $\mathbf{K}^{[l_\star]} \succ 0$ for one layer $l_\star$, and no other layer need be examined. This is also what licenses Du et al. (2019a) to work with the block of the last hidden layer alone. Two cautions. The bound transfers positivity but not conditioning: the upper bound obeys only $\lambda_{\max}(\mathbf{K}) \le \sum_l \lambda_{\max}(\mathbf{K}^{[l]})$, so the envelope Λ of Theorem H.2 remains a statement about *all* layers. And the one-layer criterion is sufficient but not necessary: $\mathbf{K}$ can be positive definite with every block singular, because the kernels intersect.

F.4.3 The assumptions

The next definition is the heart of the matter. Because Definition B.1 cannot be typed at depth, its role must be taken by a condition on the *full network Jacobian* over a neighborhood of initialization. Each clause is named and matched to the two-layer object it generalizes. Nothing in it is constructed: the event $\mathcal{E}_{(0)}$, the constants, the limiting matrix and the threshold are all posited.

Definition F.21 (Depth-H row regularity, condition (G_τ^H))**.** Consider Definition F.19 with a fixed finite family of loss-scaled rows, and fix $\tau > 0$ and $0 < \delta < 1$. Define the layerwise trust region

$$\mathcal{T}_\tau^H(\theta_{(0)}) = \Big\{\theta :\ \|\theta^{[l]} - \theta_{(0)}^{[l]}\|_F \le \tau\sqrt{d_{\min}} \ \text{ for every } l \in [H+1]\Big\}, \tag{72}$$

a convex set containing $\theta_{(0)}$. We say that the rows satisfy (G_τ^H) if there are an initialization event $\mathcal{E}_{(0)}$ with $\Pr(\mathcal{E}_{(0)}) \ge 1 - \delta/2$, constants $\Lambda_H \ge \lambda_0 > 0$, $L_H < \infty$ and $\overline{r}_\delta^H < \infty$ independent of $d^{[1]}, \dots, d^{[H]}$, a layer index $l_\star \in [H+1]$, a deterministic matrix $\mathbf{K}^{[l_\star],\infty}$, and a width threshold $\mathcal{N}_{\text{init}} = \mathcal{N}_{\text{init}}(\lambda_0, N, H, r_{\max}, \delta)$ such that clauses (i), (ii), and (iv) hold on $\mathcal{E}_{(0)}$, clause (iii) is a deterministic statement about the posited limit, and clause (v) is a separate event whose intersection with $\mathcal{E}_{(0)}$ has probability at least $1 - \delta$.

(i) (Uniform Jacobian envelope.) $\sup_{\theta \in \mathcal{T}_\tau^H(\theta_{(0)})} \|\mathbf{J}(\theta)\|_2^2 \le \Lambda_H$; in particular $\lambda_{\max}(\mathbf{K}(\theta)) \le \Lambda_H$ throughout the trust region. This generalizes the aggregate gradient bound B_*^2 of Equation (13), equivalently the choice $\Lambda = B_*^2$ of Theorem H.2, which at $H = 1$ follows from the per-unit bounds B_i by blockwise concatenation *on the per-unit tube*.

(ii) (Jacobian Lipschitz modulus with an explicit width factor.) For all $\theta, \theta' \in \mathcal{T}_\tau^H(\theta_{(0)})$,

$$\|\mathbf{J}(\theta) - \mathbf{J}(\theta')\|_2 \le \frac{L_H}{\sqrt{d_{\min}}}\|\theta - \theta'\|_2. \tag{73}$$

This generalizes Equation (97), whose modulus $L_* d_1^{-1/2}$ is *derived* from the per-unit constants L_i by blockwise separability of the Jacobian columns. At depth the columns belonging to different layers are not separable in this way, so the factor $d_{\min}^{-1/2}$ is part of the hypothesis.

(iii) (Certified limiting gap at one layer.) $\lambda_{\min}(\mathbf{K}^{[l_\star],\infty}) \geq \lambda_0 > 0$. This is the depth analogue of Definition 2.7 together with the certified bound $0 < \lambda_0 \leq \lambda_{\min}(\mathbf{K}^{\infty}_{(0)})$ of Theorem 4.2; by Lemma F.20 it certifies the full limiting kernel provided the remaining blockwise limits exist, only $\mathbf{K}^{[l_\star],\infty}$ being posited here. The clause *posits* the limiting matrix: at depth it is not produced by a strong law over units, as it is in Theorem C.7. Proposition F.22 does not consume this clause; it only ties λ_0 to a limiting object.

(iv) (Initialization concentration at the certified layer.) If $d_{\min} \geq \mathcal{N}_{\text{init}}$ then, on $\mathcal{E}_{(0)}$,

$$\lambda_{\min}\big(\mathbf{K}^{[l_\star]}_{(0)}\big) \geq \frac{3\lambda_0}{4}. \tag{74}$$

The norm bound $\big\|\mathbf{K}^{[l_\star]}_{(0)} - \mathbf{K}^{[l_\star],\infty}\big\|_2 \leq \lambda_0/4$ is one sufficient condition, and is the form that a layerwise propagation argument of the kind used by Du et al. (2019a) would deliver; we state the eigenvalue bound because that is what the proof consumes, and because the two-layer tool being replaced, Proposition B.10, produces Equation (74) directly by a matrix Chernoff bound on the independent positive-semidefinite per-unit matrices $\mathbf{G}_m$ and never passes through the norm bound. At depth that independence fails, as the discussion after Equation (69) records, so the required input is a layerwise propagation of the initialization error, conditional at each layer, rather than one concentration inequality; and by the same discussion only $l_\star \in \{H, H+1\}$ presents a conditionally independent average to propagate into.

(v) (Initial residual envelope.) $\Pr\{\|\mathbf{r}_{(0)}\|_2 \leq \overline{r}^H_\delta\} \geq 1 - \delta/2$, with $\overline{r}^H_\delta$ independent of the widths. This replaces Lemma E.1, whose McDiarmid step uses the decomposition of $\mathbf{v}_P(\theta_{(0)})$ into d_1 independent blocks, each changing the norm by at most $2A_{0,*}/\sqrt{d_1}$. At depth that decomposition does not exist and the aggregate $A_{0,*}$ has no per-unit definition.

Clauses (i), (ii) and (v) are statements about the whole network on a high-probability event; clauses (iii) and (iv) are statements about one layer. This is the sharpest way to put the difference from the two-layer case: Definition B.1 is verified by inspecting *one neuron* on the fixed compact tube $\operatorname{supp}\mu + \overline{\mathbb{B}}_R(\mathbf{0})$, and its aggregate constants $A_{0,*}, B_*, L_*$ are width-independent *by compactness*, whereas $\mathcal{T}^H_\tau(\theta_{(0)})$ has radius $\tau\sqrt{d_{\min}}$, which diverges with the width, so no compactness argument is available and width-independence of Λ_H and L_H must be *proved* on $\mathcal{E}_{(0)}$, for instance from operator-norm control of the initialization matrices, $\|\mathbf{W}^{[l]}_{(0)}\|_2 = O(\sqrt{d^{[l]}})$, as in Du et al. (2019a).

At $H = 1$ the two conditions are comparable, but the comparison goes the wrong way for a corollary. Since $\mathcal{T}_R$ is a product of radius-R per-unit balls, a θ that displaces every unit by at most R displaces each layer block by at most $R\sqrt{d_1}$ in Frobenius norm, so $\mathcal{T}_R(\theta_{(0)}) \subseteq \mathcal{T}^1_R(\theta_{(0)})$; but the inclusion is strict, and the enlargement is by the factor $\sqrt{d_1}$, since inside $\mathcal{T}^1_R$ a single unit may travel $R\sqrt{d_1}$, where the per-unit bounds of Equation (8) assert nothing. Hence (G_R) does *not* supply $\Lambda_1 = B^2_*$ and $L_1 = L_*$ on $\mathcal{T}^1_R$: the condition (G^1_R) is strictly stronger than (G_R), and Theorem 4.2 is not a corollary of Proposition F.22. What is true, and all that is claimed below, is that with $\Lambda_1 = B^2_*$, $L_1 = L_*$, $\overline{r}^1_\delta = \overline{r}_\delta$, $l_\star$ the output block and $\mathcal{N}_{\text{init}}$ the matrix-Chernoff threshold $R_K c^{-1}_{\text{Ch}} \lambda_0^{-1} \log(2N/\delta)$ of Theorem 4.2, the constants, the width conditions and the arithmetic of the two-layer proof are reproduced termwise. A layerwise-normalized trust region, with radius $\tau\sqrt{d^{[l-1]}}$ for block l, matching at uniform width the averaged Frobenius localization of Du et al. (2019a), is the natural refinement of Equation (72); we use one uniform radius because it keeps the bookkeeping below identical to the two-layer case.

F.4.4 The implication

Proposition F.22 (Depth-H substitution into the two-layer argument)**.** *Consider Definition F.19 and gradient descent as in Definition 2.4, and suppose the rows satisfy (G^H_τ) of Definition F.21, with constants $\Lambda_H \geq \lambda_0 > 0$, L_H, $\overline{r}^H_\delta$, layer $l_\star$ and threshold $\mathcal{N}_{\text{init}}$, and that $\theta \mapsto \mathbf{v}_P(\theta)$ is continuously differentiable on a convex neighborhood of $\mathcal{T}^H_\tau(\theta_{(0)})$ (for rows of type (i) this follows from clause P_{reg} of Assumption 2.6). If*

$$d_{\min} \geq \max\left\{\mathcal{N}_{\text{init}},\ (320 + 128\sqrt{6})\frac{L^2_H \Lambda_H (\overline{r}^H_\delta)^2}{\lambda_0^3},\ \frac{16\Lambda_H(\overline{r}^H_\delta)^2}{\tau^2\lambda_0^2}\right\}, \qquad 0 < \eta\Lambda_H \leq \frac{1}{2}, \tag{75}$$

then, with probability at least $1-\delta$*, simultaneously for every* $k \in \mathbb{N}_0$*,*

$$\theta_{(k)} \in \mathcal{T}_\tau^H(\theta_{(0)}), \qquad \mathbf{K}_{(k)} \succeq \frac{\lambda_0}{2}\mathbf{I}_N, \qquad \mathcal{L}(\theta_{(k)}) \le (1-\eta\lambda_0)^k \mathcal{L}(\theta_{(0)}). \tag{76}$$

We stress what this proposition is and is not. It is a substitution lemma: its hypotheses are precisely the quantities that the two-layer proof consumes, and its content is that the two-layer bookkeeping survives the substitution. It is not a depth-H convergence theorem, because every clause of Definition F.21 is assumed. In particular nothing here claims that Λ_H, L_H, $\mathcal{N}_{\mathrm{init}}$ or $\overline{r}_\delta^H$ are finite, width-independent, or of any particular size in H and $r_{\max}$; see Conjecture F.23.

Nor does it contain the two-layer theorem at $H=1$. With $\Lambda_1 = B_*^2$, $L_1 = L_*$, $\tau = R$, $d_{\min} = d_1$ and $\mathcal{N}_{\mathrm{init}}$ as above, condition Equation (75) is Equation (9) termwise, and the last two assertions of Equation (76) are the last two of Equation (10). The first assertion is strictly weaker: Equation (10) localizes every unit, $\max_{m\in[d_1]} \|\theta_{m,(k)} - \theta_{m,(0)}\|_2 \le R$, which is bought from the per-neuron one-step bound Equation (96) and is not implied by $\theta_{(k)} \in \mathcal{T}_R^1(\theta_{(0)})$. And, as noted above, the hypothesis (G_R^1) is stronger than (G_R). The proposition reproduces the two-layer bookkeeping; it does not subsume the two-layer theorem.

Proof sketch. Intersecting the two events in Definition F.21(iv) and (v) leaves probability at least $1-\delta$; work on that intersection and set $c_\eta = 1-\eta\lambda_0$ and $x_\eta = \eta\lambda_0$. Exactly as in Equation (94), we prove by strong induction that $\theta_{(j)} \in \mathcal{T}_\tau^H(\theta_{(0)})$ and $\mathcal{L}(\theta_{(j)}) \le c_\eta^j \mathcal{L}(\theta_{(0)})$ for $0 \le j \le n$; the case $n=0$ is immediate. Each step below names the two-layer ingredient and the object that replaces it.

(S1) *Taylor decomposition.* Proposition H.3 is architecture-free: the fundamental theorem of calculus along the update segment and the chain rule give $\mathbf{r}_{(k+1)} = (\mathbf{I}_N - 2\eta\mathbf{K}_{(k)})\mathbf{r}_{(k)} - \mathbf{R}_{(k)}$ with the gradient-integral remainder Equation (95). Its only structural input is that the segment $[\theta_{(k)}, \theta_{(k+1)}]$ lie in a convex set on which $\theta \mapsto \mathbf{v}_P(\theta)$ is continuously differentiable. This is the differentiability hypothesis of the proposition; for rows of type (i) it holds at every depth under clause $\mathsf{P}_{\mathrm{reg}}$ of Assumption 2.6, since finitely many compositions of $C^{r_{\max}+1,1}_{\mathrm{loc}}$ maps are again $C^{r_{\max}+1,1}$ on compact sets, so $\mathbf{J}$ exists and is continuous; clause (ii) then makes it Lipschitz on the trust region. The per-unit product tube Equation (11) is replaced by Equation (72), convex as an intersection of Frobenius balls. Nothing else changes.

(S2) *Parameter displacement.* The global one-step estimate of Lemma H.5 uses only $\|\Delta\theta_{(n)}\|_2 = 2\eta\|\mathbf{J}_{(n)}^\top \mathbf{r}_{(n)}\|_2 \le 2\eta\sqrt{\Lambda_H}\|\mathbf{r}_{(n)}\|_2$, that is, clause (i) alone; summing the geometric series as in Lemma H.6 gives $\|\theta_{(k+1)} - \theta_{(0)}\|_2 \le 4\sqrt{\Lambda_H}\|\mathbf{r}_{(0)}\|_2/\lambda_0$. The *per-neuron* one-step estimate Equation (96), which uses $\|\mathbf{\Psi}(\theta_{m,(n)})\|_2 \le B_*$ and has no depth analogue, is replaced by the trivial blockwise bound $\|\theta_{(k+1)}^{[l]} - \theta_{(0)}^{[l]}\|_F \le \|\theta_{(k+1)} - \theta_{(0)}\|_2$. The third width condition in Equation (75) is exactly $4\sqrt{\Lambda_H}\overline{r}_\delta^H/\lambda_0 \le \tau\sqrt{d_{\min}}$, so $\theta_{(k+1)} \in \mathcal{T}_\tau^H(\theta_{(0)})$, and convexity places the whole update segment there. This is the step at which the shape of the localization changes: a per-unit displacement bound of order $d_1^{-1/2}$ is traded for a relative bound on each layer's Frobenius displacement, and the per-unit conclusion is lost.

(S3) *Remainder bound.* Lemma H.7 is reproved verbatim with Equation (73) in place of Equation (97):

$$\|\mathbf{R}_{(k)}\|_2 \le \frac{L_H}{2\sqrt{d_{\min}}}\|\Delta\theta_{(k)}\|_2^2 \le \frac{2L_H}{\sqrt{d_{\min}}}\eta^2\Lambda_H\|\mathbf{r}_{(k)}\|_2^2.$$

This is the substitution that is not a free lunch: at $H=1$ the factor $d_1^{-1/2}$ is derived from blockwise separability, whereas here it is clause (ii).

(S4) *Spectral lower bound along training.* Here Lemma F.20 earns its place. Write $\sigma_{\min}(\mathbf{A})$ for the Nth largest singular value of $\mathbf{A} \in \mathbb{R}^{N\times D'}$, so that $\lambda_{\min}(\mathbf{A}\mathbf{A}^\top) = \sigma_{\min}(\mathbf{A})^2$; clause (iv) forces $D' \ge N$ for the block $l_\star$, since otherwise $\sigma_{\min}(\mathbf{J}^{[l_\star]}) = 0$. By clause (iv), $\sigma_{\min}(\mathbf{J}_{(0)}^{[l_\star]}) = \lambda_{\min}(\mathbf{K}_{(0)}^{[l_\star]})^{1/2} \ge \frac{\sqrt{3}}{2}\sqrt{\lambda_0}$. Since $\mathbf{J}^{[l_\star]}(\theta) - \mathbf{J}^{[l_\star]}(\theta')$ is a column block of $\mathbf{J}(\theta) - \mathbf{J}(\theta')$, and since for any matrix $\mathbf{M}$ with column blocks $\mathbf{M}^{[l]}$ and any $\mathbf{c} \in \mathbb{R}^N$ one has $\|\mathbf{M}^{[l]\top}\mathbf{c}\|_2 \le \|\mathbf{M}^\top\mathbf{c}\|_2$, its

spectral norm is no larger, so Equation (73) and (S2) give

$$\big\|\mathbf{J}_{(k)}^{[l_\star]} - \mathbf{J}_{(0)}^{[l_\star]}\big\|_2 \le \frac{L_H}{\sqrt{d_{\min}}} \cdot \frac{4\sqrt{\Lambda_H}}{\lambda_0} \|\mathbf{r}_{(0)}\|_2 \le \frac{\sqrt{3}-\sqrt{2}}{2}\sqrt{\lambda_0},$$

the last inequality being the second width condition in Equation (75) together with clause (v), because $64/(\sqrt{3}-\sqrt{2})^2 = 320 + 128\sqrt{6}$. Singular-value perturbation yields $\sigma_{\min}(\mathbf{J}_{(k)}^{[l_\star]}) \ge \frac{\sqrt{2}}{2}\sqrt{\lambda_0}$, hence $\lambda_{\min}(\mathbf{K}_{(k)}^{[l_\star]}) \ge \lambda_0/2$, and Equation (71) promotes this to $\lambda_{\min}(\mathbf{K}_{(k)}) \ge \lambda_0/2$. The upper bound $\lambda_{\max}(\mathbf{K}_{(k)}) \le \Lambda_H$ comes from clause (i), a statement about all layers; the layerwise lemma does not supply it. This step is the exact analogue of Proposition H.4, with a one-layer certificate in place of the full-kernel one.

(S5) *Spectral one-step descent.* With $\frac{\lambda_0}{2}\mathbf{I}_N \preceq \mathbf{K}_{(k)} \preceq \Lambda_H \mathbf{I}_N$ and $\eta\Lambda_H \le \frac{1}{2}$, every eigenvalue of $\mathbf{I}_N - 2\eta\mathbf{K}_{(k)}$ is nonnegative and its norm is at most $1 - x_\eta$. For the remainder, the induction hypothesis gives $\|\mathbf{r}_{(k)}\|_2 \le c_\eta^{k/2}\|\mathbf{r}_{(0)}\|_2 \le \|\mathbf{r}_{(0)}\|_2 \le \overline{r}_\delta^H$, so (S3) together with $\eta\Lambda_H \le \frac{1}{2}$ gives

$$\|\mathbf{R}_{(k)}\|_2 \le \frac{\eta L_H}{\sqrt{d_{\min}}}\|\mathbf{r}_{(k)}\|_2^2 \le \frac{\eta L_H \overline{r}_\delta^H}{\sqrt{d_{\min}}}\|\mathbf{r}_{(k)}\|_2.$$

Because $320 + 128\sqrt{6} > 3 + 2\sqrt{2}$ and $\Lambda_H \ge \lambda_0$, the second width condition implies $d_{\min} \ge (3+2\sqrt{2})L_H^2(\overline{r}_\delta^H)^2/\lambda_0^2$, and $\sqrt{3+2\sqrt{2}} = 1+\sqrt{2} = (\sqrt{2}-1)^{-1}$, so $L_H\overline{r}_\delta^H/\sqrt{d_{\min}} \le (\sqrt{2}-1)\lambda_0$ and therefore $\|\mathbf{R}_{(k)}\|_2 \le (\sqrt{2}-1)x_\eta\|\mathbf{r}_{(k)}\|_2$. The concluding inequality in the proof of Lemma H.8, whose only inputs are the spectral bracket and this remainder bound, then gives $\|\mathbf{r}_{(k+1)}\|_2 \le \sqrt{c_\eta}\,\|\mathbf{r}_{(k)}\|_2$.

(S6) *Closing the induction.* Squaring (S5) proves the loss assertion at $k+1$ and (S2) proves the localization assertion, which closes the strong induction and gives Equation (76). Every arithmetic constant is the two-layer one, with (B_*^2, L_*, d_1, R) replaced throughout by $(\Lambda_H, L_H, d_{\min}, \tau)$. □

F.4.5 What remains conjectural

Whether (G_τ^H) *holds* for PINN rows at depth is open. Two ingredients are missing: a quantitative bound on the constants of clauses (i), (ii) and (iv), and a positivity criterion for clause (iii). We state each as a conjecture and then record what is and is not known about it.

Conjecture F.23 (Depth and differential-order dependence of the constants)**.** *Consider Definition F.19 with the pointwise rows of Definition 2.2 of largest order $r_{\max}$, under clause $\mathsf{P}_{\mathrm{reg}}$ of Assumption 2.6 and with compactly supported $\mu^{[1]}, \dots, \mu^{[H+1]}$. For every $\tau > 0$ and $0 < \delta < 1$ there are an event $\mathcal{E}_{(0)}$ with $\Pr(\mathcal{E}_{(0)}) \ge 1 - \delta/2$ and a constant $C_0 < \infty$ depending only on σ, $r_{\max}$, τ, the supports of the $\mu^{[l]}$, $\overline{\Omega}$ and the row data, but not on H or on the widths, such that clauses (i) and (ii) of Definition F.21 hold on $\mathcal{E}_{(0)}$ with*

$$\max\{\Lambda_H,\ L_H^2\} \le C_0^{(1+r_{\max})H}, \tag{77}$$

and, given a λ_0 as in clause (iii), the threshold $\mathcal{N}_{\mathrm{init}}$ of clause (iv) obeys the same exponential rate in $(1 + r_{\max})H$, times a factor polynomial in N, λ_0^{-1} and $\log(1/\delta)$.

Remark F.24 (Why the differential order enters the exponent). The differential order is what separates Conjecture F.23 from the corresponding statement at order zero. For a strong-form row of order $|\alpha| \le r_{\max}$ the Jacobian row is $\nabla_\theta \partial_{\mathbf{x}}^\alpha u_q(\mathbf{x}_i;\theta)$, and in Equation (67) the spatial derivative obeys

$$\partial_{x_j}\mathbf{y}^{[l]} = \mathrm{diag}\big(\sigma'(\mathbf{t}^{[l]})\big)\frac{1}{\sqrt{d^{[l-1]}}}\mathbf{W}^{[l]}\,\partial_{x_j}\mathbf{y}^{[l-1]}, \qquad l = 2, \dots, H, \tag{78}$$

with $\partial_{x_j}\mathbf{y}^{[1]} = \mathrm{diag}(\sigma'(\mathbf{t}^{[1]}))\mathbf{W}^{[1]}\mathbf{e}_j$. Each of the $|\alpha|$ spatial derivatives therefore inserts one further scaled weight factor into every layer of the chain, and Faà di Bruno expands ∂^α into a sum over the ways of distributing those derivatives among the H compositions, with $\sigma^{(p)}$, $p \le r_{\max}$, at each node; the parameter gradient then appends the usual backward chain. Bounding the result term by term multiplies H layer factors, each of which now carries up to $r_{\max}$ weight insertions, which is the route to the rate in Equation (77). Three things are unproved. The bound itself: a layerwise induction of the kind used by Du et al. (2019a) at order zero is the natural route, but the jet enlargement described

below Equation (69) has no counterpart there. Width-independence: unlike the two-layer case it cannot be read off from compactness of a fixed tube, because the radius $\tau\sqrt{d_{\min}}$ of Equation (72) diverges. And sharpness: we do not know that an exponential rate in $(1 + r_{\max})H$ is necessary, nor even that an exponential rate in H is necessary at $r_{\max} = 0$. On that last point the comparison with Du et al. (2019a) must be drawn carefully. Their $2^{O(H)}$ is the constant in the propagation of the *initialization* perturbation through the layers, $\varepsilon_H \lesssim 2^{O(H)}\varepsilon_1$, and it is what makes their width requirement exponential in H; it is therefore the analogue of the threshold $\mathcal{N}_{\text{init}}$ in clause (iv), not of the envelope Λ_H or the modulus L_H, which are the quantities that Equation (78) governs. They observe that this exponential may be an artifact of the fully connected analysis, and it is replaced by $\mathrm{poly}(H)$ under residual scaling. The qualitative smoothness requirement is, by contrast, unchanged with depth: finitely many compositions of $C^{r_{\max}+1,1}_{\text{loc}}$ maps are again $C^{r_{\max}+1,1}$ on compact sets, so clause P_{reg} of Assumption 2.6 remains the right regularity to impose. What degrades with depth is quantitative, not qualitative.

Conjecture F.25 (Positivity criterion at depth)**.** *Consider Definition F.19 with the pointwise rows of Definition 2.2, under the depth-H analogue of clause P_{pd} of Assumption 2.6: σ real analytic and nonpolynomial, $\gamma > 0$, and hidden-parameter laws $\mu^{[l]}$ whose supports have nonempty interior. For every $H \geq 1$ there is a layer $l_\star \in \{H, H + 1\}$ for which $\mathbf{K}^{[l_\star],\infty}$ exists and*

$$\ker \mathbf{K}^{[l_\star],\infty} = \Big\{\mathbf{c} \in \mathbb{R}^N : \sum_{i \in I_{\mathbf{x}}} c_i \mathbf{b}_i(\mathbf{x}) = \mathbf{0} \ \textit{at every sampled site } \mathbf{x}\Big\}, \tag{79}$$

the sitewise coefficient nullspace of Equation (6)*. Consequently clause (iii) of Definition F.21 holds with some $\lambda_0 > 0$ exactly when every local family $\{\mathbf{b}_i(\mathbf{x}) : i \in I_{\mathbf{x}}\}$ is linearly independent, at every depth.*

Remark F.26 (Why the input-separation criterion is the wrong one). Du et al. (2019a) certify positivity of their limiting Gram matrix for value rows, for an analytic nonpolynomial activation, under the hypothesis that no two training inputs are parallel. That is the wrong criterion for PINN rows. Proposition 3.1 makes positivity a *sitewise coefficient rank* condition, as in Equation (6): site geometry is neither necessary nor sufficient for degeneracy, and the paper contains counterexamples in both directions. Example C.2 exhibits two rows at one shared site that are singular for one choice of operator coefficients and positive definite for another, and Example C.3 exhibits three coupled vector rows at a single site with $\nabla_\theta P_3 = \nabla_\theta P_1 + \nabla_\theta P_2$. The second example transfers to depth unconditionally, and it is worth isolating what it proves: an exact linear dependence among the loss-scaled rows *as functionals of* $\mathbf{u}$ is an identity in θ, so the associated null direction lies in $\ker \mathbf{K}(\theta)$ at every width, every depth, every architecture and every iterate, and no input-separation hypothesis can repair it. Conjecture F.25 asserts that the positive direction transfers as well. Its restriction of $l_\star$ to $\{H, H + 1\}$ is not a convenience: by the discussion after Equation (69) these are the only blocks that are, conditionally on the layers below, averages over the units of one layer of i.i.d. terms, hence the only blocks for which a limiting matrix can presently be produced at all. Of the two, $l_\star = H + 1$ is the more convenient, for the reason the two-layer certificates use $\mathbf{K}^{\infty}_{\text{out}}$: the output block's Jacobian involves only the order-$r_{\max}$ jet of the layer-H features, so the criterion is again a statement about the span of a feature family evaluated against the rows. The gap is precise. The two-layer proof separates features by analyticity in the parameters of a *single* neuron together with $\mathrm{int}(\mathrm{supp}\,\mu_h) \neq \varnothing$; at depth the layer-$H$ feature is a composition whose law is determined by the layers below, and that propagated law is controlled in the depth literature only through moments. We therefore regard Conjecture F.25 as open even at $r_{\max} = 0$, where the corresponding statement is available only under non-parallel-input hypotheses, in the bias-free scaling and for the activation class of Du et al. (2019a).

Remark F.27 (What the literature establishes at depth). For an ordinary supervised loss at differential order zero the depth-H program is complete. Du et al. (2019a) prove global linear convergence of gradient descent for deep fully connected networks, ResNets and convolutional ResNets on the squared loss over value observations, in a bias-free parameterization with a uniform hidden width m, standard Gaussian initialization, normalized inputs $\|\mathbf{x}_i\|_2 = 1$, targets $|y_i| = O(1)$, and an activation that is Lipschitz with a Lipschitz derivative, their guiding example being softplus and their positivity proposition additionally requiring analyticity and nonpolynomiality. They use exactly the layerwise decomposition of Lemma F.20, work with the block of the last hidden layer, localize by the averaged Frobenius norm $m^{-1/2}\|\mathbf{W}^{[l]}_{(k)} - \mathbf{W}^{[l]}_{(0)}\|_F$ rather than by a per-neuron tube, and require for the fully connected architecture a width $m = \Omega\big(2^{O(H)} \max\{n^4\lambda_{\min}^{-4},\, n\delta^{-1},\, n^2\lambda_{\min}^{-2} \log(Hn/\delta)\}\big)$

and a step size $\eta = O\big(\lambda_{\min} 2^{-O(H)} n^{-2}\big)$, where n is the sample size and $\lambda_{\min}$ the least eigenvalue of their limiting last-layer Gram matrix; they attribute the exponential factor to the propagation of the initialization perturbation through the layers and reduce it to $\mathrm{poly}(H)$ by the residual scaling. Allen-Zhu et al. (2019) and Zou et al. (2020) prove convergence for deep ReLU networks under separation conditions on the training inputs, again for an ordinary empirical loss. Jacot et al. (2018) and Lee et al. (2019) identify and freeze the deep NTK in the infinite-width limit, once more for value observations. None of these results covers a row of positive differential order, and none supplies the jet structure described below Equation (69). On the PINN side, the depth-H statements we are aware of assume what Definition F.21 makes explicit rather than proving it: Wang et al. (2022) derive the joint residual–boundary NTK for deep fully connected PINNs and study its infinite-width gradient flow under bounded parameters along the flow and an invertible limiting matrix, and Gan et al. (2025) prove uniform convergence in probability of the operator-transformed kernel for a deep bias-free network with a single fixed linear differential operator and state a training theorem that assumes limiting-Gram positivity and uniform gradient continuity. The kernel-convergence question for general, possibly nonlinear operators is studied by Zhou & Yan (2024), but for two-layer pointwise PINNs under a width scaling $d_1^{-\varrho}$, so that result is not a depth-H statement either. We are aware of no finite-width discrete-gradient-descent guarantee for depth-H PINNs with joint residual and boundary rows of arbitrary fixed order; that is what proofs of Conjectures F.23 and F.25, together with Proposition F.22, would deliver.

G Proofs and auxiliary results for NTK nullspaces and positive definiteness

Proof of Lemma C.11. For every $\mathbf{v} \in \mathbb{R}^N$, the Gram representation gives

$$\mathbf{v}^\top \mathbf{K} \mathbf{v} = \int \left(\sum_{i=1}^{N} v_i \Phi_i(z) \right)^2 d\mu(z) = \left\| \sum_{i=1}^{N} v_i \Phi_i \right\|_{L^2(\mu)}^2 .$$

Thus $\mathbf{K} \succ 0$ exactly when the Φ_i are independent in $L^2(\mu)$. To obtain the support formulation, let $g_{\mathbf{v}} = \sum_i v_i \Phi_i$. If the continuous function $g_{\mathbf{v}}$ is nonzero at a point of $\operatorname{supp} \mu$, it remains nonzero on a neighborhood. By definition of the support, this open set has positive μ-measure. Hence $g_{\mathbf{v}} = 0$ almost surely if and only if it vanishes on $\operatorname{supp} \mu$. □

Lemma G.1 (Linear independence across all parameter contributions)**.** *Let $\mathbf{K} = \sum_{\ell=1}^{L} \mathbf{K}^{(\ell)}$, where every $\mathbf{K}^{(\ell)}$ is a positive-semidefinite Gram matrix with parameter-derivative functions $\Phi_i^{(\ell)}$, indexed by the constraints, in a Hilbert space $\mathcal{H}_\ell$. Then*

$$\ker(\mathbf{K}) = \bigcap_{\ell=1}^{L} \ker(\mathbf{K}^{(\ell)}) = \left\{ \mathbf{c} : \sum_{i=1}^{N} c_i \Phi_i^{(\ell)} = 0 \text{ for every } \ell \right\}. \tag{80}$$

Consequently, $\mathbf{K}$ is positive definite if and only if the combined feature tuples $\left(\Phi_i^{(1)}, \dots, \Phi_i^{(L)} \right)$ are linearly independent in $\bigoplus_{\ell=1}^{L} \mathcal{H}_\ell$, equivalently if and only if the intersection in Equation (80) *is trivial; requiring one positive-semidefinite contribution $\mathbf{K}^{(\ell)}$ to be positive definite is sufficient but not necessary.*

Proof. For every $\mathbf{c} \in \mathbb{R}^N$,

$$\mathbf{c}^\top \mathbf{K} \mathbf{c} = \sum_{\ell=1}^{L} \left\| \sum_{i=1}^{N} c_i \Phi_i^{(\ell)} \right\|_{\mathcal{H}_\ell}^2 .$$

All summands are nonnegative, so the sum is zero exactly when every summand is zero. This identity proves all the stated equivalences. □

Proof of Proposition C.9. Fix $\|\mathbf{c}\|_2 = 1$. Recall that

$$\mathbf{K}_{\text{out}}^{\infty} = \mathbb{E}_{\boldsymbol{\xi} \sim \mu_h} [\mathbf{F}(\boldsymbol{\xi}) \mathbf{F}(\boldsymbol{\xi})^\top],$$
$$\|u + v\|_2^2 \geq \tfrac{1}{2} \|u\|_2^2 - \|v\|_2^2.$$

Because the sets E_ℓ are disjoint, these facts give

$$\begin{aligned}\mathbf{c}^\top \mathbf{K}_{\text{out}}^{\infty}\mathbf{c} &\geq \sum_{\ell=1}^{L}\int_{E_\ell}\|\mathbf{F}(\boldsymbol{\xi})^\top\mathbf{c}\|_2^2\,d\mu_h(\boldsymbol{\xi})\\ &\geq \frac{1}{2}\sum_{\ell=1}^{L}\mu_h(E_\ell)\|\mathbf{F}(\boldsymbol{\xi}_\ell)^\top\mathbf{c}\|_2^2-\sum_{\ell=1}^{L}\int_{E_\ell}\|(\mathbf{F}(\boldsymbol{\xi})-\mathbf{F}(\boldsymbol{\xi}_\ell))^\top\mathbf{c}\|_2^2\,d\mu_h(\boldsymbol{\xi})\\ &\geq \frac{\beta}{2}\|\mathbf{V}\mathbf{c}\|_2^2-\varepsilon^2\sum_{\ell=1}^{L}\mu_h(E_\ell)\geq\frac{\beta c_{\text{sep}}^2}{2}-\varepsilon^2.\end{aligned}$$

Taking the infimum over unit $\mathbf{c}$, and using that the limiting NTK contributions from the remaining trainable parameter groups are positive semidefinite, proves Equation (25). The final claim follows by substitution. □

The next proposition restates Proposition 3.1 together with the output-block statement Equation (81), which the main text omits; the proof below proves both.

Proposition G.2 (Exact limiting-NTK nullspace for a network with hidden biases)**.** *Use the network and pointwise constraints in Definitions 2.1 and 2.2. Assume Assumption 2.5 and the pointwise clauses of Assumption 2.6, but do not assume that the evaluated operator coefficient vectors at each location are linearly independent. Let $\mathcal{S}=\{\mathbf{x}_i : i\in[N]\}$. For every $\mathbf{x}\in\mathcal{S}$, let $I_{\mathbf{x}}=\{i:\mathbf{x}_i=\mathbf{x}\}$, and define the coordinate index set*

$$\mathcal{C}_{\mathbf{x}}=\left\{(q,\alpha): q\in[d_2],\alpha\in\bigcup_{i\in I_{\mathbf{x}}}I_{iq}\right\},$$

assigning zero to entries absent from an individual constraint on $\mathcal{C}_{\mathbf{x}}$. Define

$$\mathbf{b}_i(\mathbf{x})=\big(s_i a_{iq\alpha}(\mathbf{x})\big)_{(q,\alpha)\in\mathcal{C}_{\mathbf{x}}},\qquad \mathcal{N}_{\mathbf{x}}=\left\{\mathbf{c}_{\mathbf{x}}:\sum_{i\in I_{\mathbf{x}}}c_i\mathbf{b}_i(\mathbf{x})=\mathbf{0}\right\}.$$

The partition $[N]=\bigsqcup_{\mathbf{x}\in\mathcal{S}}I_{\mathbf{x}}$ canonically identifies $\mathbb{R}^N$ with $\bigoplus_{\mathbf{x}\in\mathcal{S}}\mathbb{R}^{I_{\mathbf{x}}}$. Then the output-weight Gram block is well defined and

$$\ker(\mathbf{K}_{out}^{\infty})=\bigoplus_{\mathbf{x}\in\mathcal{S}}\mathcal{N}_{\mathbf{x}},\qquad \operatorname{rank}(\mathbf{K}_{out}^{\infty})=\sum_{\mathbf{x}\in\mathcal{S}}\dim\operatorname{span}\{\mathbf{b}_i(\mathbf{x}):i\in I_{\mathbf{x}}\}.\tag{81}$$

The full limiting NTK $\mathbf{K}_{(0)}^{\infty}$, formed from all trainable parameters, exists and has exactly the same nullspace and rank:

$$\ker(\mathbf{K}_{(0)}^{\infty})=\ker(\mathbf{K}_{out}^{\infty}),\qquad \operatorname{rank}(\mathbf{K}_{(0)}^{\infty})=\sum_{\mathbf{x}\in\mathcal{S}}\dim\operatorname{span}\{\mathbf{b}_i(\mathbf{x}):i\in I_{\mathbf{x}}\}.\tag{82}$$

Thus, under these hypotheses, linear independence of the evaluated operator coefficient vectors at every location is necessary and sufficient for both Gram matrices to be positive definite.

Proof. For scalar constraint row i, output component q, and one hidden affine parameter $\boldsymbol{\xi}=(\mathbf{w},b)\sim\mu_h$, define the parameter-derivative function associated with $(\mathbf{a}_m)_q$:

$$\phi_{iq}(\boldsymbol{\xi})=s_i(\mathcal{A}_{iq}\sigma_{\mathbf{w},b})(\mathbf{x}_i),\qquad \sigma_{\mathbf{w},b}(\mathbf{x})=\sigma(\mathbf{w}^\top\mathbf{x}+\gamma b).$$

The d_2 output-weight families occupy disjoint coordinate blocks. Their contribution to the limiting NTK is therefore

$$[\mathbf{K}_{\text{out}}^{\infty}]_{ij}=\sum_{q=1}^{d_2}\langle\phi_{iq},\phi_{jq}\rangle_{L^2(\mu_h)},$$

and, for every $\mathbf{c}\in\mathbb{R}^N$,

$$\mathbf{c}^\top\mathbf{K}_{\text{out}}^{\infty}\mathbf{c}=\sum_{q=1}^{d_2}\left\|\sum_{i=1}^{N}c_i\phi_{iq}\right\|_{L^2(\mu_h)}^2.\tag{83}$$

The compact hidden support makes these parameter-derivative functions square-integrable, so the display also proves existence of this block. Suppose the quadratic form in Equation (83) is zero. For every output coordinate q, define the real-analytic function

$$G_{\mathbf{c},q}(\mathbf{w}, b) = \sum_{i=1}^{N} c_i s_i (\mathcal{A}_{iq} \sigma_{\mathbf{w},b})(\mathbf{x}_i).$$

It vanishes μ_h-almost surely. Continuity extends this equality to $\mathrm{supp}(\mu_h)$, which contains a nonempty open subset. Hence the identity theorem for real-analytic functions gives $G_{\mathbf{c},q} \equiv 0$ on $\mathbb{R}^{d_0+1}$. Because $\gamma > 0$, the change of variable $v = \gamma b$ ranges over all $v \in \mathbb{R}$.

Group the scalar constraints by their distinct physical sites. At a fixed $\mathbf{x}$, the corresponding part of $G_{\mathbf{c},q}$ is the following finite sum of activation derivatives multiplied by monomials in the hidden weights:

$$G_{\mathbf{c},q,\mathbf{x}}(\mathbf{w}, v) = \sum_{i:\, \mathbf{x}_i = \mathbf{x}} c_i s_i \sum_{\alpha \in I_{iq}} a_{iq\alpha}(\mathbf{x}) \sigma^{(|\alpha|)}(\mathbf{w}^\top \mathbf{x} + v) \mathbf{w}^\alpha.$$

After combining like terms, every formally nonzero $G_{\mathbf{c},q,\mathbf{x}}$ is in the standard form of Jin & Wu (2026a). Omitting trivial groups, the independence lemma in Lemma C.8 applies across the remaining distinct sites. Hence no such nonempty relation can vanish identically. Thus every grouped polynomial is formally zero, which means that, for every q, α, and $\mathbf{x} \in \mathcal{S}$,

$$\sum_{i:\, \mathbf{x}_i = \mathbf{x}} c_i s_i a_{iq\alpha}(\mathbf{x}) = 0.$$

Equivalently, $\mathbf{c}_{\mathbf{x}} \in \mathcal{N}_{\mathbf{x}}$ for every $\mathbf{x} \in \mathcal{S}$. Conversely, the displayed kernel-entry formula shows that these local coefficient cancellations make every $G_{\mathbf{c},q}$ vanish identically. The equivalence proves the first nullspace equality in Equation (81). Rank–nullity supplies its rank expression.

Proposition B.7 and Theorem B.9 give existence of the full limiting NTK. It remains to check that the other parameter coordinates neither remove nor create dependencies caused by local operator coefficient relations. If $\mathbf{c} \in \bigoplus_{\mathbf{x} \in \mathcal{S}} \mathcal{N}_{\mathbf{x}}$, then for every vector-valued function $\mathbf{u}$ sufficiently differentiable through the largest differential order among the constraints, and hence for every network realization,

$$\sum_{i=1}^{N} c_i P_i(\theta) = \sum_{\mathbf{x} \in \mathcal{S}} \sum_{(q,\alpha) \in \mathcal{C}_{\mathbf{x}}} \left(\sum_{i \in I_{\mathbf{x}}} c_i s_i a_{iq\alpha}(\mathbf{x}) \right) \partial^\alpha u_q(\mathbf{x}; \theta) = 0. \tag{84}$$

Differentiating gives $\mathbf{J}(\theta)^\top \mathbf{c} = 0$ at every finite width and every parameter value. Thus such a $\mathbf{c}$ also belongs to the kernel of the full limiting NTK whenever it exists. Conversely, that matrix is the output-weight contribution plus positive-semidefinite terms from the remaining trainable coordinates. Hence its kernel is contained in $\ker(\mathbf{K}_{\mathrm{out}}^{\infty})$. This argument establishes Equation (82). The final assertion follows because a finite family is linearly independent exactly when its relation space is zero. □

Proof of Proposition 3.1. Proposition B.7 and Theorem B.9 give existence and empirical convergence. Under clause P_{pd}, apply Proposition G.2. Its local relation spaces and span dimensions give Equation (6); the positive-definiteness criterion follows by rank–nullity. □

Proof of Theorem C.17. The complete one-neuron map

$$\mathcal{G}(\boldsymbol{\xi}, \mathbf{a}) = \sum_{q=1}^{d_2} a_q \varphi_{\boldsymbol{\xi},q}$$

is real analytic into $\mathcal{X}$, being a finite sum of the coordinates a_q times the analytic maps $\boldsymbol{\xi} \mapsto \varphi_{\boldsymbol{\xi},q}$ of clause W_{pd}; for this step the C^2 regularity of clause $\mathsf{W}_{\mathrm{reg}}$ already suffices. Compact initialization and Lemma B.3 therefore verify the condition in Definition B.1; Theorem B.9 proves existence of $\mathbf{K}_{(0)}^{\infty}$ and the stated empirical convergence.

Recall the relation map $\mathcal{R}_\lambda \mathbf{c} = \sum_{i=1}^{N} c_i \lambda_i \in \mathcal{X}^*$ of the main text. For the derivative with respect to output coordinate q of one neuron's output weight, define

$$F_{iq}(\boldsymbol{\xi}) = \lambda_i(\varphi_{\boldsymbol{\xi},q}).$$

Linearity of λ_i and the network parameterization give

$$\frac{\partial P_i(\theta)}{\partial (\mathbf{a}_m)_q} = \frac{1}{\sqrt{d_1}} F_{iq}(\boldsymbol{\xi}_m).$$

Consequently, the limiting output-weight contribution satisfies

$$[\mathbf{K}^{\infty}_{\text{out}}]_{ij} = \sum_{q=1}^{d_2} \int F_{iq}(\boldsymbol{\xi}) F_{jq}(\boldsymbol{\xi})\, d\mu_h(\boldsymbol{\xi}), \tag{85}$$

$$\mathbf{c}^{\top} \mathbf{K}^{\infty}_{\text{out}} \mathbf{c} = \sum_{q=1}^{d_2} \int |(\mathcal{R}_\lambda \mathbf{c})(\varphi_{\boldsymbol{\xi},q})|^2 \, d\mu_h(\boldsymbol{\xi}). \tag{86}$$

In particular, this matrix is positive semidefinite without any independence assumption.

Suppose first that $\mathbf{c} \in \ker \mathbf{K}^{\infty}_{\text{out}}$. Every nonnegative term in Equation (86) is then zero, so, for every q,

$$g_{\mathbf{c},q}(\boldsymbol{\xi}) = (\mathcal{R}_\lambda \mathbf{c})(\varphi_{\boldsymbol{\xi},q}) = 0 \qquad \mu_h\text{-almost surely.} \tag{87}$$

The map into $\mathcal{X}$ is real analytic by assumption, and composition with the bounded functional $\mathcal{R}_\lambda \mathbf{c}$ shows that $g_{\mathbf{c},q}$ is a real-analytic scalar function. In particular it is continuous. If it were nonzero at a point of $\operatorname{supp}(\mu_h)$, it would remain nonzero on a neighborhood having positive μ_h-measure, contradicting Equation (87). It therefore vanishes on the support, and hence on a nonempty open set. The identity theorem for real-analytic functions gives

$$(\mathcal{R}_\lambda \mathbf{c})(\varphi_{\boldsymbol{\xi},q}) = 0 \qquad \text{for every } \boldsymbol{\xi} \in \mathbb{R}^{d_0+1} \text{ and } q \in [d_2].$$

Thus $\mathcal{R}_\lambda \mathbf{c}$ vanishes on the linear span of all coordinate ridge features. This set is dense in $\mathcal{X}$, so continuity of $\mathcal{R}_\lambda \mathbf{c}$ implies $\mathcal{R}_\lambda \mathbf{c} = 0$ on $\mathcal{X}$. This approximation argument proves $\ker \mathbf{K}^{\infty}_{\text{out}} \subseteq \mathcal{N}_{\text{weak}}$. The reverse inclusion follows immediately from Equation (86); hence

$$\ker \mathbf{K}^{\infty}_{\text{out}} = \mathcal{N}_{\text{weak}}. \tag{88}$$

It remains to compare the output-weight block with the total NTK. The full limiting matrix equals this contribution plus positive-semidefinite terms from the other coordinates. Therefore $\ker \mathbf{K}^{\infty}_{(0)} \subseteq \ker \mathbf{K}^{\infty}_{\text{out}}$. Conversely, if $\mathbf{c} \in \mathcal{N}_{\text{weak}}$, then at every finite width and every parameter value,

$$\sum_{i=1}^{N} c_i P_i(\theta) = (\mathcal{R}_\lambda \mathbf{c})(\mathbf{u}(\cdot;\theta)) = 0.$$

Differentiating this relation gives $\mathbf{J}(\theta)^{\top} \mathbf{c} = 0$, so $\mathbf{c}$ lies in the empirical full NTK's kernel and hence in the kernel of its population limit. Together with Equation (88), this argument proves both nullspace identities in Equation (41). Rank–nullity then supplies

$$\operatorname{rank} \mathbf{K}^{\infty}_{(0)} = N - \dim \mathcal{N}_{\text{weak}} = \dim \operatorname{span}\{\lambda_1, \ldots, \lambda_N\}.$$

Finally, in the coercive special case, a relation among the weak functionals implies, with $v_{\mathbf{c}} = \sum_i c_i s_i v_i$,

$$a(u, v_{\mathbf{c}}) = 0 \qquad \text{for every } u \in \mathcal{X}.$$

Taking $u = v_{\mathbf{c}}$ and using coercivity yields $c_a \|v_{\mathbf{c}}\|^2_{\mathcal{X}} \le 0$, so $v_{\mathbf{c}} = 0$. Linear independence of the v_i, together with $s_i > 0$, gives $\mathbf{c} = 0$. Thus $\mathcal{N}_{\text{weak}} = \{0\}$, which proves the final assertion. □

Proof of Theorem C.21. Because analyticity implies $\sigma \in C^2_{\text{loc}}$, Proposition B.6 verifies the condition in Definition B.1; Theorem B.9 proves existence of $\mathbf{K}^{\infty}_{(0)}$ and the stated empirical convergence.

For $\boldsymbol{\xi} = (\mathbf{w}, b)$, define the scalar ridge feature $\phi_{\boldsymbol{\xi}}(\mathbf{y}) = \sigma(\mathbf{w}^{\top}\mathbf{y} + \gamma b)$ and the output-weight features

$$F_{iq}(\boldsymbol{\xi}) = s_i \int_D \phi_{\boldsymbol{\xi}}(\mathbf{y})\, d\nu_{iq}(\mathbf{y}).$$

Because D is compact and each measure has finite total variation, the integrals are finite. We next justify that $\boldsymbol{\xi} \mapsto F_{iq}(\boldsymbol{\xi})$ is real analytic. Fix $\boldsymbol{\xi}_0 = (\mathbf{w}_0, b_0)$. For a sufficiently small closed parameter

ball $B \subset \mathbb{R}^{d_0+1}$ about $\boldsymbol{\xi}_0$, all affine arguments lie in a compact interval I. A real-analytic function on $\mathbb{R}$ has compatible local holomorphic extensions, and compactness gives a holomorphic extension $\widetilde{\sigma}$ to an open set $U \subset \mathbb{C}$ containing I. There is then a neighborhood $V \subset \mathbb{C}^{d_0+1}$ of $\boldsymbol{\xi}_0$ such that every affine argument for $\boldsymbol{\xi} \in V$ and $\mathbf{y} \in D$ lies in U. The resulting integrand is holomorphic in $\boldsymbol{\xi}$ and uniformly bounded on compact subsets of V. Dominated integration against the finite total-variation measure (equivalently, Morera's theorem) preserves holomorphy near $\boldsymbol{\xi}_0$. Restriction to real arguments proves the asserted analyticity. Because $\boldsymbol{\xi}_0$ was arbitrary, the conclusion holds on all of $\mathbb{R}^{d_0+1}$.

As in Equation (85), differentiation with respect to the output weights gives

$$[\mathbf{K}^{\infty}_{\text{out}}]_{ij} = \sum_{q=1}^{d_2} \int F_{iq}(\boldsymbol{\xi}) F_{jq}(\boldsymbol{\xi})\, d\mu_h(\boldsymbol{\xi}),$$

$$\mathbf{c}^{\top} \mathbf{K}^{\infty}_{\text{out}} \mathbf{c} = \sum_{q=1}^{d_2} \int \left| \int_D \phi_{\boldsymbol{\xi}}(\mathbf{y})\, d\zeta_{\mathbf{c},q}(\mathbf{y}) \right|^2 d\mu_h(\boldsymbol{\xi}), \tag{89}$$

where

$$\zeta_{\mathbf{c},q} = \sum_{i=1}^{N} c_i s_i \nu_{iq}. \tag{90}$$

Thus this Gram matrix is always positive semidefinite.

Suppose $\mathbf{c} \in \ker \mathbf{K}^{\infty}_{\text{out}}$. For each q, define the real-analytic function

$$G_{\mathbf{c},q}(\boldsymbol{\xi}) = \int_D \phi_{\boldsymbol{\xi}}(\mathbf{y})\, d\zeta_{\mathbf{c},q}(\mathbf{y}).$$

It vanishes μ_h-almost surely by Equation (89). Continuity extends this equality to $\operatorname{supp}(\mu_h)$, and the support contains a nonempty open set. The real-analytic identity theorem therefore gives

$$\int_D \sigma(\mathbf{w}^{\top}\mathbf{y} + \gamma b)\, d\zeta_{\mathbf{c},q}(\mathbf{y}) = 0 \qquad \text{for every } (\mathbf{w}, b) \in \mathbb{R}^{d_0+1}. \tag{91}$$

Because $\gamma > 0$, the shifts γb range over all real numbers. The uniform universal-approximation theorem for a nonpolynomial activation with a bias (Leshno et al., 1993) says that the linear span of the functions in Equation (91) is dense in $C(D)$. Integration against the finite signed measure $\zeta_{\mathbf{c},q}$ defines a bounded functional on that space. It therefore vanishes on all of $C(D)$, and uniqueness in the Riesz–Markov representation theorem yields $\zeta_{\mathbf{c},q} = 0$. The conclusion holds for every q, proving $\mathbf{c} \in \mathcal{N}_{\text{nl}}$. Conversely, if $\mathbf{c} \in \mathcal{N}_{\text{nl}}$, every combined measure in Equation (90) is zero, and the quadratic form in Equation (89) is zero. Hence

$$\ker \mathbf{K}^{\infty}_{\text{out}} = \mathcal{N}_{\text{nl}}. \tag{92}$$

The full limiting NTK is a sum of positive-semidefinite parameter-coordinate contributions, so its kernel is contained in the output-weight kernel. In the reverse direction, $\mathbf{c} \in \mathcal{N}_{\text{nl}}$ implies, at every finite width and model state,

$$\sum_{i=1}^{N} c_i P_i(\theta) = \sum_{q=1}^{d_2} \int_D u_q(\mathbf{y}; \theta)\, d\zeta_{\mathbf{c},q}(\mathbf{y}) = 0.$$

Differentiating gives $\mathbf{J}(\theta)^{\top}\mathbf{c} = 0$. Thus $\mathcal{N}_{\text{nl}} \subseteq \ker \mathbf{K}^{\infty}_{(0)}$, and Equation (92) proves the two nullspace identities in Equation (45). Rank–nullity yields the displayed rank formula, and the positive-definiteness criterion follows. □

G.1 Compatible training error for a singular NTK

Calculation for Example D.8. Before the normalization assigned to each loss group in Definition 2.2, set $\widetilde{P}_D(\theta) = u''_\theta(0)$, $\widetilde{P}_-(\theta) = u_\theta(-1)$, and $\widetilde{P}_+(\theta) = u_\theta(1)$. For a scalar-output neuron with weights (a_m, w_m), define its width-rescaled parameter-gradient blocks by $\boldsymbol{\phi}_\bullet(a_m, w_m) :=$

$\sqrt{d_1}\,\nabla_{(a_m,w_m)}\widetilde{P}_\bullet(\theta)$. The hidden-bias derivative is zero because $\gamma = 0$, so we omit that coordinate below. Writing a generic block as (a, w) gives

$$\begin{aligned}\boldsymbol{\phi}_D(a,w) &= (2w^2, 4aw),\\ \boldsymbol{\phi}_-(a,w) &= (-w + w^2, a(-1+2w)),\\ \boldsymbol{\phi}_+(a,w) &= (w + w^2, a(1+2w)).\end{aligned}$$

Hence $\boldsymbol{\phi}_D = \boldsymbol{\phi}_- + \boldsymbol{\phi}_+$ pointwise for every (a, w), separately for the output-weight and hidden-weight coordinates. This identity is therefore a linear dependence among the full parameter-gradient vectors used by the NTK, rather than a dependence confined to the contribution from one parameter group.

For independent $a, w \sim \mathrm{Unif}[-1, 1]$, direct integration gives, in the order $(D, -, +)$, the unweighted population Gram matrix

$$\widetilde{\mathbf{K}}^{\infty}_{(0)} = \frac{1}{45}\begin{pmatrix}116 & 58 & 58\\ 58 & 59 & -1\\ 58 & -1 & 59\end{pmatrix}.$$

Its eigenvalues are 0, $4/3$, and $58/15$, and its nullspace is spanned by $\mathbf{n} = (1, -1, -1)^\top$. In particular, the two boundary parameter-gradient vectors are linearly independent, and the displayed identity is the only dependence among the three parameter-gradient vectors. The target $\widetilde{\mathbf{v}}_T = (2, 1, 1)^\top$ satisfies $\mathbf{n}^\top \widetilde{\mathbf{v}}_T = 0$. Every network prediction obeys the same identity, so every initial error is also orthogonal to $\mathbf{n}$ and belongs to $\mathrm{Range}(\widetilde{\mathbf{K}}^{\infty}_{(0)})$; the loss dynamics with the kernel held fixed at $\widetilde{\mathbf{K}}^{\infty}_{(0)}$ therefore contract to zero for a suitable step size.

The continuum compatibility is not merely a relation among the three target entries. The Dirichlet problem has the unique solution $u^*(x) = x^2$. For every even width, the network represents it exactly by taking half of the hidden weights equal to 1, half equal to -1, and all output weights equal to $d_1^{-1/2}$, because

$$\frac{1}{2}\left(\sigma(x) + \sigma(-x)\right) = x^2.$$

Finally, with $N_r = 1$ and $N_b = 2$, the normalization in Definition 2.2 multiplies the last two prediction components and targets by $1/\sqrt{2}$. Define the rescaled features by $\boldsymbol{\phi}_D^{\text{norm}} := \boldsymbol{\phi}_D$ and $\boldsymbol{\phi}_\pm^{\text{norm}} := \boldsymbol{\phi}_\pm/\sqrt{2}$, so $\boldsymbol{\phi}_D^{\text{norm}} = \sqrt{2}(\boldsymbol{\phi}_-^{\text{norm}} + \boldsymbol{\phi}_+^{\text{norm}})$. The null vector becomes $(1, -\sqrt{2}, -\sqrt{2})^\top$, while both the normalized target and every normalized prediction remain orthogonal to it. □

Lemma G.3 (Target projection for an architecture linear in its output weights)**.** *For the network in Definition 2.1 and linear prediction components, let $\mathbf{J}_{out}(\theta)$ be the columns of the full Jacobian associated with the output weights, and set $\mathbf{a}_{stack} = (\mathbf{a}_1^\top, \ldots, \mathbf{a}_{d_1}^\top)^\top$. At every finite width and parameter value,*

$$\mathbf{v}_P(\theta) = \mathbf{J}_{out}(\theta)\mathbf{a}_{stack}, \qquad \mathbf{v}_P(\theta) \in \mathrm{Range}(\mathbf{K}(\theta)).$$

Consequently, let $\Pi_{\ker \mathbf{K}(\theta)}$ denote the orthogonal projector onto $\ker \mathbf{K}(\theta)$; then

$$\Pi_{\ker \mathbf{K}(\theta)}\left(\mathbf{v}_T - \mathbf{v}_P(\theta)\right) = \Pi_{\ker \mathbf{K}(\theta)}\mathbf{v}_T. \tag{93}$$

For a limiting-kernel nullspace generated by the local operator coefficient relations in Proposition G.2, the same conclusion holds because Equation (84) *annihilates every network prediction. Without either structural guarantee, the generally valid formula projects the initial training error rather than the target alone.*

Proof. Linearity of every prediction component in the network output makes the prediction vector homogeneous and linear in $\mathbf{a}_{\text{stack}}$; its coefficient matrix is exactly $\mathbf{J}_{\text{out}}(\theta)$. Moreover, $\mathrm{Range}(\mathbf{J}(\theta)) = \mathrm{Range}(\mathbf{J}(\theta)\mathbf{J}(\theta)^\top) = \mathrm{Range}(\mathbf{K}(\theta))$. Hence the prediction has zero projection onto $\ker \mathbf{K}(\theta)$, which proves Equation (93). The claim for a limiting nullspace forced by local operator relations follows directly from Equation (84). □

G.2 Poisson proof by linear independence of parameter-derivative functions

Proof. The stated marginal-law hypothesis covers the Gram contribution studied below. To establish the full limiting NTK using all trainable parameters, also assume that the per-neuron parameter vectors are i.i.d. across neurons and that the output weights have finite second moment; Assumption 2.5 is one sufficient choice. These hypotheses make every individual Gram contribution integrable, so the strong-law argument in the deterministic-NTK proof gives existence. We study the parameter-derivative functions associated with a_m, whose linear independence follows from the invertibility of the coefficient matrix constructed from their Taylor coefficients. Because nonzero row scaling preserves this property, divide by the positive loss-row scales and multiply the three Poisson-residual features by -1. Denote $w = w_m$. The resulting derivatives fall into two groups. The boundary pair is $f(w) = \tanh(w)$ and $f(w) = \tanh(5w)$ on the support of the hidden-weight law. For the interior constraints, define $\Phi(x) = w^2 \tanh^{(2)}(xw)$, where $\tanh^{(2)}$ is the second derivative of the hyperbolic tangent. The three collocation sites then give $f(w) = \Phi(2)$, $f(w) = \Phi(3)$, and $f(w) = \Phi(4)$ on the support of the hidden-weight law.

All five parameter-derivative functions are analytic in w. Because the support assumption in Example D.7 contains an open neighborhood of $w = 0$, equality of a linear combination on the support implies equality of its power-series expansions there. The two boundary-feature series are

$$\tanh(w) = w - \frac{w^3}{3} + \frac{2w^5}{15} - \frac{17w^7}{315} + \frac{62w^9}{2835} + \mathcal{O}(w^{11}),$$
$$\tanh(5w) = 5w - \frac{125w^3}{3} + \frac{6250w^5}{15} - \frac{1328125w^7}{315} + \frac{121093750w^9}{2835} + \mathcal{O}(w^{11}).$$

The corresponding interior expansions are

$$\Phi(2) = -4w^3 + \frac{64w^5}{3} - \frac{1088w^7}{15} + \frac{63488w^9}{315} + \mathcal{O}(w^{11}),$$
$$\Phi(3) = -6w^3 + 72w^5 - \frac{2754w^7}{5} + \frac{120528w^9}{35} + \mathcal{O}(w^{11}),$$
$$\Phi(4) = -8w^3 + \frac{512w^5}{3} - \frac{34816w^7}{15} + \frac{8126464w^9}{315} + \mathcal{O}(w^{11}).$$

Matching the first five power-series terms of a linear combination of these parameter-derivative functions produces a five-equation system. The coefficients of w and w^3 give

$$c_1 + 5c_2 = 0,$$
$$-\frac{c_1}{3} - \frac{125c_2}{3} - 4c_3 - 6c_4 - 8c_5 = 0.$$

The remaining powers yield

$$\frac{2c_1}{15} + \frac{6250c_2}{15} + \frac{64c_3}{3} + 72c_4 + \frac{512c_5}{3} = 0,$$
$$-\frac{17c_1}{315} - \frac{1328125c_2}{315} - \frac{1088c_3}{15} - \frac{2754c_4}{5} - \frac{34816c_5}{15} = 0,$$
$$\frac{62c_1}{2835} + \frac{121093750c_2}{2835} + \frac{63488c_3}{315} + \frac{120528c_4}{35} + \frac{8126464c_5}{315} = 0.$$

The coefficient matrix of the combined system has determinant $6976569344/135 \neq 0$. Therefore it has the unique solution $\mathbf{c} = \mathbf{0}$, and the five parameter-derivative functions are linearly independent as functions of w. Hence the population output-weight contribution is nonsingular. Under the finite-moment condition above, the full limiting NTK exists and is positive definite because it is the sum of this block and positive-semidefinite terms from the other parameters. □

H Proof of gradient descent convergence

This section first outlines Theorem 4.2 and then establishes its supporting lemmas.

H.1 PROOF SKELETON

We begin with an abstract result under the row-regularity condition in Definition B.1. Proposition B.7 verifies its hypotheses for finite-order strong-form differential constraints; the theorem also covers weak and nonlocal constraints that directly satisfy the stated regularity conditions. We first prove the initial-residual bound stated in Lemma E.1.

Proof of Lemma E.1. Let $\mathbf{h}_m = (h_1(\theta_{m,(0)}), \dots, h_N(\theta_{m,(0)}))^\top$. Then $\mathbf{v}_P(\theta_{(0)}) = d_1^{-1/2} \sum_m \mathbf{h}_m$ and $\|\mathbf{h}_m\|_2 \le A_{0,*}$, which proves Equation (51) by the triangle inequality. The conditional-centering condition gives $\mathbb{E}\mathbf{h}_m = 0$; with independence of the per-neuron parameter vectors, this yields

$$\mathbb{E}\|\mathbf{v}_P(\theta_{(0)})\|_2^2 = \mathbb{E}\|\mathbf{h}_1\|_2^2 = H_0^2 \le A_{0,*}^2, \qquad \mathbb{E}\|\mathbf{v}_P(\theta_{(0)})\|_2 \le H_0.$$

If $A_{0,*} = 0$, the prediction is zero almost surely and the claim is immediate. Suppose $A_{0,*} > 0$. Changing one per-neuron parameter vector changes $\|d_1^{-1/2} \sum_m \mathbf{h}_m\|_2$ by at most $2A_{0,*}/\sqrt{d_1}$. McDiarmid's inequality therefore gives

$$\Pr\{\|\mathbf{v}_P(\theta_{(0)})\|_2 > H_0 + t\} \le \exp\left(-\frac{t^2}{2A_{0,*}^2}\right).$$

Set $t = A_{0,*}\sqrt{2\log(1/\varepsilon)}$ and use $\|\mathbf{r}_{(0)}\|_2 \le \|\mathbf{v}_T\|_2 + \|\mathbf{v}_P(\theta_{(0)})\|_2$. □

Remark H.1. For a fixed number of loss groups, divide each row by the square root of its group's sample count. If the unnormalized per-neuron contributions and target values are uniformly bounded, $A_{0,*}$ and $\|\mathbf{v}_T\|_2$ are $\mathcal{O}(1)$. Hence the initial-residual estimate in Lemma E.1 gives width- and sample-count-independent high-probability control of $\|\mathbf{r}_{(0)}\|_2$. Without per-constraint normalization, the corresponding aggregate quantities can grow with N.

Theorem H.2 (Convergence of gradient descent under abstract row regularity)**.** *Suppose the condition in Definition B.1 holds and use gradient descent as in Definition 2.4. Require a positive-definiteness certificate as in Definition 2.7, and let $0 < \lambda_0 \le \lambda_{\min}(\mathbf{K}_{(0)}^{\infty})$ be a certified lower bound. Set $\mathbf{r}_{(0)} = \mathbf{v}_T - \mathbf{v}_P(\theta_{(0)})$. Define the squared operator-norm envelope for one neuron's initialization Jacobian contribution by*

$$R_K := \operatorname*{ess\,sup}_{z\sim\mu} \|\mathbf{\Psi}(z)\|_2^2 \le B_*^2.$$

Also set

$$c_{Ch} := \frac{1}{4} + \frac{3}{4}\log\frac{3}{4} > 0 \qquad (c_{Ch}^{-1} \approx 29.207).$$

Let $\Lambda \ge \lambda_0$ be a deterministic upper bound on $\lambda_{\max}(\mathbf{K}(\theta))$ throughout the per-neuron parameter neighborhood in Definition B.1, and let $0 < \delta < 1$. If

$$d_1 \ge \frac{R_K}{c_{Ch}\lambda_0}\log\frac{N}{\delta},$$

then the matrix-Chernoff initialization event $\{\lambda_{\min}(\mathbf{K}_{(0)}) \ge 3\lambda_0/4\}$ of Proposition B.10 has probability at least $1-\delta$. Conditional on this event, suppose additionally that the realized initialization satisfies

$$d_1 \ge \max\left\{(320 + 128\sqrt{6})\frac{L_*^2\Lambda\|\mathbf{r}_{(0)}\|_2^2}{\lambda_0^3}, \frac{16B_*^2\|\mathbf{r}_{(0)}\|_2^2}{R^2\lambda_0^2}\right\}.$$

If $0 < \eta\Lambda \le 1/2$, every iterate remains in the fixed parameter set

$$\mathcal{T}_R(\theta_{(0)}) := \left\{\theta : \max_{m\in[d_1]} \|\theta_m - \theta_{m,(0)}\|_2 \le R\right\},$$

and, for every $k \in \mathbb{N}_0$,

$$\mathbf{K}_{(k)} \succeq \frac{\lambda_0}{2}\mathbf{I}_N, \qquad \mathcal{L}(\theta_{(k)}) \le (1-\eta\lambda_0)^k\mathcal{L}(\theta_{(0)}).$$

One may take $\Lambda = B_*^2$; this choice also gives $\lambda_0 \le \Lambda$, so the contraction factor is positive. The Jacobian-perturbation width condition also implies the Taylor-remainder condition

$$d_1 \ge (3 + 2\sqrt{2})\frac{L_*^2\|\mathbf{r}_{(0)}\|_2^2}{\lambda_0^2}.$$

Indeed, $\Lambda \geq \lambda_0$ *and* $320 + 128\sqrt{6} > 3 + 2\sqrt{2}$. *For a fixed collection of PINN loss groups, divide each row by the square root of its group's sample count. Then* B_* *and* L_* *are independent of the number of points when the corresponding unnormalized constraint-wise bounds are uniform.*

Proof of Theorem 4.2. Apply Lemma E.1 with $\varepsilon = \delta/2$. On the resulting event, $\|\mathbf{r}_{(0)}\|_2 \leq \overline{r}_\delta$. Invoke Theorem H.2 with its initialization failure probability set to $\delta/2$, $\Lambda = B_*^2$, the uniform per-neuron Jacobian norm R_K, and the deterministic envelope $\overline{r}_\delta$ in every condition that depends on the initial training error. The union bound gives joint confidence at least $1 - \delta$. The localization width condition ensures that every per-neuron displacement stays below R, which closes the induction. Consequently, every parameter block remains in the fixed neighborhood on which the constraint-wise constants are valid. □

Proof skeleton of Theorem H.2. Set $c_\eta = 1 - \eta\lambda_0$. We prove jointly by strong induction that

$$\theta_{(j)} \in \mathcal{T}_R(\theta_{(0)}) \quad \text{and} \quad \mathcal{L}(\theta_{(j)}) \leq c_\eta^j \mathcal{L}(\theta_{(0)}) \qquad (0 \leq j \leq n). \tag{94}$$

The assertion is immediate at $n = 0$.

Suppose it holds through $n = k$. The per-neuron part of Lemma H.6, using only the Jacobians at $\theta_{(0)}, \ldots, \theta_{(k)}$, which satisfy the displacement bound, gives

$$\max_m \|\theta_{m,(k+1)} - \theta_{m,(0)}\|_2 \leq \frac{4B_*\|\mathbf{r}_{(0)}\|_2}{\sqrt{d_1}\lambda_0} \leq R.$$

Thus $\theta_{(k+1)} \in \mathcal{T}_R(\theta_{(0)})$. Convexity places the full update segment there as well. The local gradient-Lipschitz bound is therefore available throughout this set.

The global-displacement part of Lemma H.6, the estimate for the Jacobian perturbation in Proposition H.4, and the initialization event imply

$$\frac{\lambda_0}{2}\mathbf{I}_N \preceq \mathbf{K}_{(k)} \preceq \Lambda \mathbf{I}_N.$$

Using Proposition H.3 for the residual $\mathbf{r}_{(k)}$ from Definition 2.3 gives the exact error recursion

$$\mathbf{r}_{(k+1)} = (\mathbf{I}_N - 2\eta\mathbf{K}_{(k)})\mathbf{r}_{(k)} - \mathbf{R}_{(k)}.$$

Lemma H.8 now yields

$$\|\mathbf{r}_{(k+1)}\|_2 \leq \sqrt{c_\eta}\,\|\mathbf{r}_{(k)}\|_2.$$

Squaring and applying the induction hypothesis proves the loss assertion at $k + 1$, and the per-neuron estimate proves the displacement assertion. These two conclusions close the argument. □

H.2 Taylor decomposition of consecutive predictions

The following proposition uses the fundamental theorem of calculus along the update segment to decompose the prediction increment into the Gram term and a gradient-integral remainder. The same blockwise Lipschitz constant therefore controls kernel stability and this nonlinearity.

Proposition H.3 (Prediction difference between consecutive iterates)**.** *Under the condition in Definition B.1 and the gradient descent algorithm setup in Definition 2.4, suppose the segment* $[\theta_{(n)}, \theta_{(n+1)}]$ *lies in the prescribed per-neuron parameter neighborhood. Then the difference in prediction between consecutive iterations admits the decomposition*

$$\mathbf{v}_P(\theta_{(n+1)}) - \mathbf{v}_P(\theta_{(n)}) = 2\eta\mathbf{K}_{(n)}\mathbf{r}_{(n)} + \mathbf{R}_{(n)},$$

where $\Delta\theta_{(n)} := \theta_{(n+1)} - \theta_{(n)} = -\eta\nabla\mathcal{L}(\theta_{(n)})$ *and*

$$\mathbf{R}_{(n)} = \int_0^1 \left[\mathbf{J}(\theta_{(n)} + s\Delta\theta_{(n)}) - \mathbf{J}_{(n)}\right]\Delta\theta_{(n)}\, ds. \tag{95}$$

Proof. The fundamental theorem of calculus on the update segment gives

$$\begin{aligned}\mathbf{v}_P(\theta_{(n+1)}) - \mathbf{v}_P(\theta_{(n)}) &= \int_0^1 \mathbf{J}(\theta_{(n)} + s\Delta\theta_{(n)})\Delta\theta_{(n)}\, ds \\ &= \mathbf{J}_{(n)}\Delta\theta_{(n)} + \mathbf{R}_{(n)}.\end{aligned}$$

By the chain rule,

$$\nabla \mathcal{L}(\theta_{(n)}) = 2\mathbf{J}_{(n)}^\top \left(\mathbf{v}_P(\theta_{(n)}) - \mathbf{v}_T\right).$$

Consequently,

$$\mathbf{J}_{(n)}\Delta\theta_{(n)} = 2\eta \mathbf{J}_{(n)}\mathbf{J}_{(n)}^\top \left(\mathbf{v}_T - \mathbf{v}_P(\theta_{(n)})\right),$$

which is the stated decomposition. □

H.3 Positive definiteness along gradient descent

We preserve the spectral gap by controlling the $N \times (d_1 d_\theta)$ Jacobian directly, rather than its Gram matrix. This comparison avoids one factor B_*^2/λ_0 in the overparameterization condition.

Proposition H.4 (Lower bound on the NTK eigenvalue along training)**.** *Under the condition in Definition B.1 and gradient descent in Definition 2.4, suppose $\theta_{(0)}, \ldots, \theta_{(k)}$ lie in the prescribed per-neuron parameter neighborhood and the fixed design has a positive-definiteness certificate as in Definition 2.7, with $0 < \lambda_0 \le \lambda_{\min}(\mathbf{K}_{(0)}^\infty)$ a certified lower bound. Let $\Lambda \ge \sup_{\theta \in \mathcal{T}_R(\theta_{(0)})} \lambda_{\max}(\mathbf{K}(\theta))$ be a deterministic upper bound. Fix $0 < \delta < 1$ and set $c_{Ch} = \frac{1}{4} + \frac{3}{4}\log\frac{3}{4}$. If the induction hypothesis Equation* (94) *holds through k, $0 < \eta\Lambda \le 1/2$, and*

$$d_1 \ge \max\left\{ \frac{R_K}{c_{Ch}\lambda_0} \log\frac{N}{\delta}, (320 + 128\sqrt{6}) \frac{L_*^2 \Lambda \|\mathbf{r}_{(0)}\|_2^2}{\lambda_0^3} \right\},$$

then $\lambda_{\min}(\mathbf{K}_{(k)}) \ge \lambda_0/2$ on the initialization event in Proposition B.10.

Proof of Proposition H.4. On the initialization event,

$$\sigma_{\min}(\mathbf{J}_{(0)}) = \sqrt{\lambda_{\min}(\mathbf{K}_{(0)})} \ge \frac{\sqrt{3}}{2}\sqrt{\lambda_0}.$$

The uniform Lipschitz bound for each neuron's gradient and the block structure of the Jacobian columns associated with different neurons give, as in Equation (23),

$$\|\mathbf{J}_{(k)} - \mathbf{J}_{(0)}\|_2 \le \frac{L_*}{\sqrt{d_1}} \|\theta_{(k)} - \theta_{(0)}\|_2.$$

Indeed, the squared operator norm of a horizontal concatenation is at most the sum of the squared norms of its neuronal components.

By Lemma H.6 applied through $k-1$ (the case $k = 0$ is trivial),

$$\|\theta_{(k)} - \theta_{(0)}\|_2 \le \frac{4\sqrt{\Lambda}}{\lambda_0} \|\mathbf{r}_{(0)}\|_2.$$

Because

$$320 + 128\sqrt{6} = \frac{64}{(\sqrt{3} - \sqrt{2})^2},$$

the width condition implies

$$\|\mathbf{J}_{(k)} - \mathbf{J}_{(0)}\|_2 \le \frac{\sqrt{3} - \sqrt{2}}{2}\sqrt{\lambda_0}.$$

Singular-value perturbation therefore yields

$$\sigma_{\min}(\mathbf{J}_{(k)}) \ge \frac{\sqrt{2}}{2}\sqrt{\lambda_0}.$$

Squaring proves $\lambda_{\min}(\mathbf{K}_{(k)}) \ge \lambda_0/2$. □

H.4 Consequences of strong induction and a small learning step

This subsection derives spectral controls used in Theorem H.2: a trajectory-length bound, a width-sensitive Taylor remainder, and a joint one-step contraction estimate.

H.4.1 PREPARATIONS FOR THE ANALYSIS

Lemma H.5 (One-step parameter motion)**.** *Under the row-regularity condition in Definition B.1 and gradient descent in Definition 2.4, let $n \geq 1$ be an integer. If $\theta_{(n-1)}$ lies in the prescribed per-neuron parameter neighborhood and $\lambda_{\max}(\mathbf{K}_{(n-1)}) \leq \Lambda$, then*

$$\|\theta_{(n)} - \theta_{(n-1)}\|_2^2 \leq 4\eta^2\Lambda\|\mathbf{r}_{(n-1)}\|_2^2.$$

Moreover, each unit satisfies

$$\|\theta_{m,(n)} - \theta_{m,(n-1)}\|_2 \leq \frac{2\eta B_*}{\sqrt{d_1}}\|\mathbf{r}_{(n-1)}\|_2. \tag{96}$$

Proof. The gradient update and $\mathbf{K}_{(n-1)} = \mathbf{J}_{(n-1)}\mathbf{J}_{(n-1)}^\top$ give the exact identity

$$\begin{aligned}\|\theta_{(n)} - \theta_{(n-1)}\|_2^2 &= 4\eta^2\|\mathbf{J}_{(n-1)}^\top\mathbf{r}_{(n-1)}\|_2^2 \\ &= 4\eta^2\mathbf{r}_{(n-1)}^\top\mathbf{K}_{(n-1)}\mathbf{r}_{(n-1)},\end{aligned}$$

and the spectral bound proves the global estimate. For the per-neuron claim, the block Jacobian and corresponding change are

$$\begin{aligned}\mathbf{J}_{(n-1),m} &= d_1^{-1/2}\mathbf{\Psi}(\theta_{m,(n-1)}), \\ \theta_{m,(n)} - \theta_{m,(n-1)} &= 2\eta\mathbf{J}_{(n-1),m}^\top\mathbf{r}_{(n-1)}.\end{aligned}$$

The aggregate per-constraint gradient inequality implies $\|\mathbf{\Psi}(\theta_{m,(n-1)})\|_2 \leq B_*$, which proves Equation (96). □

H.4.2 FIRST CONSEQUENCE: PARAMETER DISPLACEMENT FROM INITIALIZATION

The one-step estimate in Lemma H.5 and the strong-induction bound in Equation (94) control the parameter displacement from initialization. The proof of Proposition H.4 uses the resulting inequality.

Lemma H.6 (Parameter displacement from initialization)**.** *Under the conditions in Theorem H.2 and gradient descent in Definition 2.4, suppose $\theta_{(0)}, \ldots, \theta_{(k)}$ lie in the prescribed per-neuron parameter neighborhood. If the induction hypothesis Equation* (94) *holds through k, set $x_\eta = \eta\lambda_0$. Then*

$$\begin{aligned}\|\theta_{(k+1)} - \theta_{(0)}\|_2 &\leq \frac{2(1+\sqrt{1-x_\eta})\sqrt{\Lambda}}{\lambda_0}\|\mathbf{r}_{(0)}\|_2 \leq \frac{4\sqrt{\Lambda}}{\lambda_0}\|\mathbf{r}_{(0)}\|_2, \\ \max_m \|\theta_{m,(k+1)} - \theta_{m,(0)}\|_2 &\leq \frac{2(1+\sqrt{1-x_\eta})B_*}{\sqrt{d_1}\lambda_0}\|\mathbf{r}_{(0)}\|_2 \leq \frac{4B_*}{\sqrt{d_1}\lambda_0}\|\mathbf{r}_{(0)}\|_2.\end{aligned}$$

Proof. By Equation (94), $\|\mathbf{r}_{(n)}\|_2 \leq (1-x_\eta)^{n/2}\|\mathbf{r}_{(0)}\|_2$. Summing the global one-step estimate in Lemma H.5 gives

$$\|\theta_{(k+1)} - \theta_{(0)}\|_2 \leq 2\eta\sqrt{\Lambda}\|\mathbf{r}_{(0)}\|_2\sum_{n=0}^{k}(1-x_\eta)^{n/2}.$$

Applying the same argument to Equation (96), followed by a maximum over m, produces the analogous expression with $\sqrt{\Lambda}$ replaced by $B_*/\sqrt{d_1}$. Finally,

$$\frac{\eta}{1-\sqrt{1-x_\eta}} = \frac{1+\sqrt{1-x_\eta}}{\lambda_0} \leq \frac{2}{\lambda_0},$$

which proves both claims. All one-step estimates use only the iterates through k, each of which satisfies the per-neuron displacement bound. □

H.4.3 SECOND CONSEQUENCE: CONTROL OF THE GRADIENT-INTEGRAL REMAINDER

The uniform block-gradient Lipschitz property controls the Taylor remainder without introducing a separate Hessian constant.

Lemma H.7 (Taylor remainder bound)**.** *Under the conditions in Theorem H.2 and the update in Definition 2.4, suppose the update segment $[\theta_{(k)}, \theta_{(k+1)}]$ lies in the prescribed per-neuron parameter neighborhood. Then*

$$\left\|\mathbf{R}_{(k)}\right\|_2 \leq \frac{2L_*}{\sqrt{d_1}}\,\eta^2\Lambda\left\|\mathbf{r}_{(k)}\right\|_2^2.$$

Proof. For any two points in this convex set, the decomposition of the Jacobian by neurons and the uniform Lipschitz bound for each neuron's gradient give

$$\|\mathbf{J}(\theta') - \mathbf{J}(\theta)\|_2 \leq \frac{L_*}{\sqrt{d_1}}\|\theta' - \theta\|_2. \tag{97}$$

Indeed, after squaring, the norm of the horizontal concatenation is at most the sum of the squared operator norms of its blocks. Using Equation (95) and Equation (97), set $\Delta\theta_{(k)} = \theta_{(k+1)} - \theta_{(k)}$ and obtain

$$\|\mathbf{R}_{(k)}\|_2 \leq \int_0^1 \frac{sL_*}{\sqrt{d_1}}\|\Delta\theta_{(k)}\|_2^2\,ds = \frac{L_*}{2\sqrt{d_1}}\|\Delta\theta_{(k)}\|_2^2.$$

Finally, the one-step estimate in Lemma H.5 gives

$$\|\Delta\theta_{(k)}\|_2^2 \leq 4\eta^2\Lambda\|\mathbf{r}_{(k)}\|_2^2,$$

which proves the claim. □

Lemma H.8 (Spectral one-step descent)**.** *Suppose the induction hypothesis holds through iteration k, the update segment lies in the prescribed per-neuron parameter neighborhood, $\lambda_{\min}(\mathbf{K}_{(k)}) \geq \lambda_0/2$, and $\lambda_{\max}(\mathbf{K}_{(k)}) \leq \Lambda$. Under the width and step-size conditions of Theorem H.2,*

$$\|\mathbf{r}_{(k+1)}\|_2 \leq \sqrt{1-\eta\lambda_0}\,\|\mathbf{r}_{(k)}\|_2.$$

Proof. Set $x_\eta = \eta\lambda_0$ and $y_\eta = \eta\Lambda$. Because

$$\frac{\lambda_0}{2}\mathbf{I}_N \preceq \mathbf{K}_{(k)} \preceq \Lambda\mathbf{I}_N, \qquad y_\eta \leq \frac{1}{2},$$

all eigenvalues of $\mathbf{I}_N - 2\eta\mathbf{K}_{(k)}$ are nonnegative and

$$\|\mathbf{I}_N - 2\eta\mathbf{K}_{(k)}\|_2 \leq 1 - x_\eta. \tag{98}$$

The width condition controlling the Jacobian perturbation in Theorem H.2 implies

$$d_1 \geq (3+2\sqrt{2})\frac{L_*^2\|\mathbf{r}_{(0)}\|_2^2}{\lambda_0^2}.$$

If $\mathbf{r}_{(k)} = 0$, the update and remainder both vanish, so the result is immediate. Otherwise, because $\sqrt{3+2\sqrt{2}} = 1+\sqrt{2}$, the induction hypothesis and the Taylor-remainder bound in Lemma H.7 imply

$$\begin{aligned}\frac{\|\mathbf{R}_{(k)}\|_2}{\|\mathbf{r}_{(k)}\|_2} &\leq \frac{2L_*\|\mathbf{r}_{(0)}\|_2}{\sqrt{d_1}}\eta^2\Lambda \\ &\leq 2(\sqrt{2}-1)x_\eta y_\eta \leq (\sqrt{2}-1)x_\eta.\end{aligned} \tag{99}$$

The residual decomposition gives

$$\mathbf{r}_{(k+1)} = (\mathbf{I}_N - 2\eta\mathbf{K}_{(k)})\mathbf{r}_{(k)} - \mathbf{R}_{(k)}.$$

For $0 < x_\eta \leq 1/2$,

$$\sqrt{1-x_\eta} - (1-x_\eta) = x_\eta\frac{\sqrt{1-x_\eta}}{1+\sqrt{1-x_\eta}} \geq (\sqrt{2}-1)x_\eta.$$

Combining this inequality with Equations (98) and (99) proves the claim. □

I A SINGULAR NTK WITH A COMPATIBLE TARGET FOR $\tanh$

This example shows that a singular kernel can still yield convergence when the target satisfies the same linear relation as the predictions. It uses neither a polynomial activation nor inconsistent data. On the interval $[0, 2]$, consider the fully trainable network

$$u_\theta(x) = c + \frac{1}{\sqrt{d_1}} \sum_{m=1}^{d_1} a_m \tanh(w_m x), \qquad \theta = (c, (a_m, w_m)_{m=1}^{d_1}) \in \mathbb{R}^{2d_1+1}. \tag{100}$$

We pair this architecture with the following second-order initial-value problem for an ordinary differential equation (ODE):

$$\mathcal{P}u := u'' + u' + u = f, \qquad u(0) = 0, \qquad u'(0) = 1, \tag{101}$$

where

$$u^*(x) = \tanh x, \qquad f(x) = \tanh x + \operatorname{sech}^2 x - 2 \tanh x \operatorname{sech}^2 x. \tag{102}$$

Here $\operatorname{sech}(x) = 1/\cosh(x)$. Continuity of the ODE coefficients in (101) makes the IVP uniquely solvable. Direct substitution of the formula in (102) verifies u^*. Thus all three loss targets correspond to this exact function. Moreover, u^* belongs to the network class at every width: take $c = 0$, $w_m = 1$, and $a_m = d_1^{-1/2}$ for every neuron m.

For one ODE residual point x_f, define the example prediction and target vectors by

$$\mathbf{v}_P^{\text{ex}}(\theta; x_f) = \begin{pmatrix} (\mathcal{P}u_\theta)(x_f) \\ u_\theta(0) \\ u'_\theta(0) \end{pmatrix}, \qquad \mathbf{v}_T^{\text{ex}}(x_f) = \begin{pmatrix} f(x_f) \\ 0 \\ 1 \end{pmatrix}. \tag{103}$$

The entries in (103) correspond, in order, to the differential equation, value, and derivative constraints. Following the target-minus-prediction convention of the main text, set

$$\mathbf{r}^{\text{ex}}(\theta; x_f) = \mathbf{v}_T^{\text{ex}}(x_f) - \mathbf{v}_P^{\text{ex}}(\theta; x_f), \qquad \mathcal{L}_{\text{ex}}(\theta; x_f) = \|\mathbf{r}^{\text{ex}}(\theta; x_f)\|_2^2. \tag{104}$$

Write the three components of the residual in (104) as $r^{\text{ex}}_{\mathcal{P}}$, $r^{\text{ex}}_{\text{val}}$, and $r^{\text{ex}}_{\text{der}}$, in the displayed order. Define the three feature gradients with respect to all trainable parameters:

$$\boldsymbol{\Phi}_{\mathcal{P},x_f}(\theta) := \nabla_\theta(\mathcal{P}u_\theta)(x_f), \qquad \boldsymbol{\Phi}_{\text{val}}(\theta) := \nabla_\theta u_\theta(0), \qquad \boldsymbol{\Phi}_{\text{der}}(\theta) := \nabla_\theta u'_\theta(0),$$

$$\mathbf{J}^{\text{ex}}(\theta; x_f) := \begin{pmatrix} \boldsymbol{\Phi}_{\mathcal{P},x_f}(\theta)^\top \\ \boldsymbol{\Phi}_{\text{val}}(\theta)^\top \\ \boldsymbol{\Phi}_{\text{der}}(\theta)^\top \end{pmatrix}, \quad \mathbf{K}^{\text{ex}}(\theta; x_f) := \mathbf{J}^{\text{ex}}(\theta; x_f)\mathbf{J}^{\text{ex}}(\theta; x_f)^\top. \tag{105}$$

These vectors lie in $\mathbb{R}^{2d_1+1}$. Stacking them as in (105) gives a prediction Jacobian in $\mathbb{R}^{3\times(2d_1+1)}$ and a tangent Gram matrix in $\mathbb{R}^{3\times 3}$. Oddness of $\tanh$ yields

$$u''_\theta(0) = 0 \quad \text{and} \quad \nabla_\theta u''_\theta(0) = \mathbf{0} \qquad \text{for every } \theta \text{ and every width } d_1. \tag{106}$$

The identities in (106) imply that, when we evaluate the ODE residual, value, and derivative constraints at $x_f = 0$,

$$\boldsymbol{\Phi}_{\mathcal{P},0}(\theta) = \boldsymbol{\Phi}_{\text{val}}(\theta) + \boldsymbol{\Phi}_{\text{der}}(\theta), \qquad \mathbf{n} := \frac{1}{\sqrt{3}}(1, -1, -1)^\top \in \ker \mathbf{K}^{\text{ex}}(\theta; 0), \tag{107}$$

because $\mathbf{J}^{\text{ex}}(\theta; 0)^\top \mathbf{n} = \mathbf{0}$. This feature relation in (107) follows from the absence of hidden biases and the odd activation. The three distinct constraints evaluate the value, the first derivative, and the differential operator. Meanwhile, the target vector is $\mathbf{v}_T^{\text{ex}}(0) = (f(0), u^*(0), u^{*\prime}(0))^\top = (1, 0, 1)^\top$, so its projection onto $\mathbf{n}$ is zero. In fact, the three residual components obey $r^{\text{ex}}_{\mathcal{P}}(\theta; 0) = r^{\text{ex}}_{\text{val}}(\theta; 0) + r^{\text{ex}}_{\text{der}}(\theta; 0)$ throughout nonlinear training.

The population kernel for $x_f = 0$ is particularly transparent. Draw $a_m, w_m \overset{\text{i.i.d.}}{\sim} \text{Unif}[-1, 1]$ and set $c = 0$. Let $\mathbf{K}^{\text{ex},\infty}_{(0)}(x_f)$ denote the deterministic infinite-width limit of $\mathbf{K}^{\text{ex}}(\theta_{(0)}; x_f)$ under this initialization. At $x = 0$, the gradient of the value constraint is nonzero only in the c coordinate, whereas the gradient of the derivative constraint has per-neuron coordinates $(w_m, a_m)/\sqrt{d_1}$. Hence

$$\mathbf{K}^{\text{ex},\infty}_{(0)}(0) = \begin{pmatrix} 5/3 & 1 & 2/3 \\ 1 & 1 & 0 \\ 2/3 & 0 & 2/3 \end{pmatrix}, \qquad \operatorname{spec}\Big(\mathbf{K}^{\text{ex},\infty}_{(0)}(0)\Big) = \left\{0, \frac{5-\sqrt{7}}{3}, \frac{5+\sqrt{7}}{3}\right\}. \tag{108}$$

The calculation in (108) shows that this 3×3 kernel has rank two. Its quadratic form is strictly positive on $(\operatorname{span}\{\mathbf{n}\})^\perp$, which contains every compatible error vector.

Moving only the ODE residual point restores positive definiteness Keep the value and derivative observations at zero and move the ODE residual point to $x_f = s > 0$. It suffices to inspect the gradients with respect to the output bias and output weights. The value observation and equation row both add 1 to the former. Within the latter gradient, the derivative constraint contributes w, whereas the ODE residual constraint contributes

$$g_s(w) = w^2 \tanh''(sw) + w \tanh'(sw) + \tanh(sw). \tag{109}$$

The output feature in (109) has the following expansion at the origin:

$$g_s(w) = (1+s)w - \left(2s + s^2 + \frac{s^3}{3}\right) w^3 + O(w^5). \tag{110}$$

For $s > 0$, g_s is therefore not proportional to w. Suppose a linear combination of the three full parameter-gradient vectors with coefficients $(\chi_{\mathcal{P}}, \chi_{\text{val}}, \chi_{\text{der}})$ vanishes. The output-bias coordinate gives $\chi_{\mathcal{P}} + \chi_{\text{val}} = 0$, whereas the output-weight coordinates give $\chi_{\mathcal{P}} g_s(w) + \chi_{\text{der}} w = 0$ in $L^2(\text{Unif}[-1,1])$. The expansion in (110) forces $\chi_{\mathcal{P}} = \chi_{\text{der}} = 0$ and then $\chi_{\text{val}} = 0$. The Gram-matrix contribution from the global output bias and the per-neuron output weights is therefore positive definite, and so is the full limiting NTK. The same argument gives an empirical kernel of rank three, $\mathbf{K}^{\text{ex}}(\theta_{(0)}; s)$, almost surely for $d_1 \geq 2$: the exceptional samples form the zero set of a nonzero analytic determinant.

For a concrete population calculation, take $s = 1$. The two nonconstant entries needed below are

$$k_{\mathcal{P}\mathcal{P}}(s) := 1 + \int_0^1 \left(g_s(w)^2 + \frac{1}{3} g_s'(w)^2\right) dw, \qquad k_{\mathcal{P},\text{der}}(s) := \int_0^1 w g_s(w)\, dw + \frac{1}{3} g_s(1). \tag{111}$$

The integrals in (111) use a prime on g_s to denote differentiation with respect to w. Deterministic fixed-node Gauss–Legendre quadrature then gives

$$\mathbf{K}^{\text{ex},\infty}_{(0)}(1) = \begin{pmatrix} k_{\mathcal{P}\mathcal{P}}(1) & 1 & k_{\mathcal{P},\text{der}}(1) \\ 1 & 1 & 0 \\ k_{\mathcal{P},\text{der}}(1) & 0 & 2/3 \end{pmatrix} \simeq \begin{pmatrix} 1.68126802 & 1 & 0.48341254 \\ 1 & 1 & 0 \\ 0.48341254 & 0 & 2/3 \end{pmatrix}. \tag{112}$$

The matrix in (112) has eigenvalues

$$\text{spec}\Big(\mathbf{K}^{\text{ex},\infty}_{(0)}(1)\Big) \simeq \{0.11916210, 0.74496231, 2.48381029\}.$$

These decimals do not establish positive definiteness; the preceding linear-independence argument does so independently.

For a chosen x_f, abbreviate $\mathbf{r}^{\text{ex}}_{(0)}(x_f) := \mathbf{r}^{\text{ex}}(\theta_{(0)}; x_f)$ and $\mathbf{K}^{\text{ex}}_{(0)}(x_f) := \mathbf{K}^{\text{ex}}(\theta_{(0)}; x_f)$. Under gradient flow with this initial kernel held fixed,

$$\mathcal{L}^{\text{fixed}}_{\text{ex}}(t; x_f) = \mathbf{r}^{\text{ex}}_{(0)}(x_f)^\top \exp\Big(-4\mathbf{K}^{\text{ex}}_{(0)}(x_f) t\Big)\, \mathbf{r}^{\text{ex}}_{(0)}(x_f). \tag{113}$$

For $x_f = 0$ and $d_1 \geq 1$, the bias-coordinate value row and the nonzero derivative row give rank two with probability one. Therefore, on this event, $\ker \mathbf{K}^{\text{ex}}_{(0)}(0) = \text{span}\{\mathbf{n}\}$. The exact residual relation makes the initial residual orthogonal to this nullspace. For $x_f = 1$ and $d_1 \geq 2$, the preceding analytic-determinant argument is an almost-sure positive-definiteness certificate in the terminology of Definition 2.7; it is the analytic alternative in the pointwise procedure of Section D.3.1. Thus the fixed-kernel formula in (113) implies loss convergence to zero with probability one in both cases when $d_1 \geq 2$. For the two population matrices above, positive definiteness alone does not order their rates. The smallest nonzero eigenvalue for $x_f = 0$ is 0.785 to three decimals, whereas the smallest eigenvalue for $x_f = 1$ is 0.119. With the normalized step rule $\eta = 0.2/\lambda_{\max}$, the two limiting matrices multiply the squared residual-mode contribution in the slowest spectral direction by approximately 0.769 and 0.962 per step, respectively. For $x_f = 1$, this certificate gives a positive-definite kernel, so the loss with the kernel held fixed at initialization converges to zero for every initial error. At $x_f = 0$, the same conclusion holds for the compatible manufactured target above because the residual remains in $\mathbf{n}^\perp$.

Scope The exact singularity for $x_f = 0$ uses the hidden-bias-free form in (100); a generic trainable hidden bias removes the identity $u''_\theta(0) = 0$. This relation depends jointly on that architecture, the operator $u'' + u' + u$, and evaluation of all three constraints at zero. Repeated physical locations alone need not cause singularity. The global output bias contributes a fixed rank-one term to the kernel; equivalently, one may distribute it over i.i.d. per-neuron parameter vectors as $c = d_1^{-1/2} \sum_{m=1}^{d_1} c_m$ if all per-neuron parameter vectors must be identically distributed. Positive diagonal rescaling of the three loss components preserves the number of nonzero eigenvalues. It also leaves unchanged whether the weighted target obeys the corresponding weighted prediction relation, although it changes their magnitudes and hence the convergence rates.

J VECTOR-VALUED POINTWISE CONSTRAINTS

Example J.1 (Coupled linear advection–diffusion system). Let $\mathbf{u} = (u_1, u_2)^\top$ on $[0,1]^2$ satisfy

$$\partial_t u_1 + v_{\text{adv}} \partial_x u_1 = f_1, \qquad \partial_t u_2 - \nu_{\text{diff}} \partial_{xx} u_2 - c_{\text{coup}} u_1 = f_2,$$

where $v_{\text{adv}}, c_{\text{coup}} \in \mathbb{R}$ and $\nu_{\text{diff}} > 0$. Impose also initial conditions $u_1(x,0), u_2(x,0)$ and noncorner side conditions $u_1(0,t), u_2(1,t)$. In coordinates $(u_1, \partial_x u_1, \partial_t u_1, \partial_{xx} u_2, \partial_t u_2)$, the two interior rows at the same site are $(0, v_{\text{adv}}, 1, 0, 0)$ and $(-c_{\text{coup}}, 0, 0, -\nu_{\text{diff}}, 1)$, and they are independent for every admissible coefficient triple. At a shared physical location, the initial constraints are the standard basis vectors in (u_1, u_2), and each side constraint is nonzero. Proposition 3.1 asks for independence of the joint family at each site, not of the interior, initial, and side families separately. At a site carrying interior and data rows, the common coordinates are $(u_1, \partial_x u_1, \partial_t u_1, u_2, \partial_{xx} u_2, \partial_t u_2)$ and the rows $(0, v_{\text{adv}}, 1, 0, 0, 0)$, $(-c_{\text{coup}}, 0, 0, 0, -\nu_{\text{diff}}, 1)$, $(1, 0, 0, 0, 0, 0)$, and $(0, 0, 0, 1, 0, 0)$ have distinct pivot columns, hence are independent; the word "noncorner" excludes exactly the corner sites at which an initial row and a side row of the same component would coincide and be dependent. Thus the local coefficient vectors are independent at every site. Under Assumption 2.5 and the pointwise clauses of Assumption 2.6, the rank result in Proposition 3.1 gives $\mathbf{K}^\infty_{(0)} \succ 0$. The sitewise procedure in Section D.3.1 is therefore an explicit positive-definiteness certificate for this coupled pointwise PDE system. More generally, the same criterion applies to vector-valued pointwise systems. Their limiting NTKs are positive definite exactly when the local coefficient vectors $\mathbf{b}_i(\mathbf{x})$ are linearly independent at every sampled physical location; at a location carrying a single row this means only that its coefficient vector is nonzero.

K EXPERIMENTAL PROTOCOLS AND FINITE-WIDTH DIAGNOSTICS

This appendix documents the experiments summarized in Section 5. The study comprises 770 PDE training runs, 1,080 compatible-target two-row control runs, and 7,700 initialization measurements from 100 independent seed banks per PDE, with widths paired by nested prefixes; two small audits, a three-row joint-nullspace control and a per-iterate monotonicity check, complete the design. The tracked aggregates store the displayed summaries with content hashes, whereas per-run records of the registered campaign are not retained (Section K.9); we retain narrow or slowly converging cases. The campaign snapshot stores configuration, source, and result hashes. Table 3 summarizes the study components.

K.1 MATCHED ARCHITECTURE AND EXPERIMENTAL MATRIX

Every experiment uses

$$\mathbf{u}(\mathbf{x}; \theta) = \frac{1}{\sqrt{d_1}} \sum_{m=1}^{d_1} \mathbf{a}_m \tanh(\mathbf{w}_m^\top \mathbf{x} + b_m), \qquad (w_{m\ell}, b_m, a_{mq}) \overset{\text{i.i.d.}}{\sim} \text{Unif}[-1, 1],$$

with all hidden weights, hidden biases, and output weights trained and no global output bias. This setup is exactly Equation (1) with $\sigma = \tanh$ and $\gamma = 1$. Constraints use the group normalization in Definition 2.2, and full-batch GD minimizes $\|\mathbf{v}_P(\theta) - \mathbf{v}_T\|_2^2$ with $\eta = 0.2/\lambda_{\max}(\mathbf{K}_{(0)})$. Computation uses float64 and deterministic nested initialization prefixes.

Table 3: Experimental matrix. Independent row designs use two compatible targets with a shared initial kernel, so optimization-run counts exceed distinct-kernel counts.

Study	Instances	Runs	Grid
Rank/conditioning controls	600 kernels	1,080	six families, five row designs, widths $\{16, 256, 4{,}096, 16{,}384\}$, five paired seeds
PDE training diagnostics	11 reference problems	770	11 PDEs, widths $\{4, 16, 64, 256, 1{,}024, 4{,}096, 16{,}384\}$, ten seeds, 3,000 GD steps
Initialization sweep	7,700 kernels	7,700	11 PDEs, the same seven widths, 100 independent seed banks
Joint-nullspace audit	72 kernels	36	three designs, four perturbation sizes, widths $\{32, 128\}$, three seeds, 5,000 GD steps
Per-iterate audit	30 kernels	30	one two-row design, widths $\{256, 1{,}024\}$, three row mixings, five seeds, 3,000 GD steps

K.2 PDE BENCHMARKS AND MANUFACTURED DATA

All spatial domains are $[-1, 1]$ or $[-1, 1]^2$, and time-dependent problems use $(t, x) \in [0, 1] \times [-1, 1]$. We obtain the forcing terms and all boundary or initial data by substituting the manufactured solutions below into the listed operators. The same discrete functional defines each target and the training loss; this choice is especially relevant for the weak, nonlocal, and fractional rows. We fix the case-specific collocation locations independently of initialization and list them below. The first seven cases use automatic differentiation to evaluate the displayed pointwise derivatives at the listed sites; they introduce no PDE-operator quadrature. Table 4 gives the full suite.

Table 5: Finite-horizon GD coverage by observation class. "Loss decreased" means only $\mathcal{L}_{3,000} < \mathcal{L}_0$; it neither asserts monotonicity at unsaved iterates nor verifies the theorem's sufficient width and step-size conditions. Each problem contributes seven widths and ten initialization seeds. Weak and integral/nonlocal entries refer to the stored quadrature functionals.

Observation class	Included problems and rows	Runs	Loss decreased
Strong-form local	Seven scalar or vector strong-form systems: point evaluations and derivatives through order four, including the biharmonic and Stokes systems	490	490/490
Weak variational	Weak Poisson with four fixed test functions and endpoint traces	70	70/70
Integral/nonlocal	Nonlocal diffusion, an integro-differential equation, and transformed Caputo-time rows	210	210/210
All classes	Eleven fixed finite constraint systems	770	770/770

The strong-form row samples derivative orders only through four; it is not an experiment over every finite order. The theorem instead applies to each chosen fixed finite order when the activation has the corresponding smoothness, with constants that may depend on that order and the design. Likewise, the last two rows provide optimization evidence for their exact discrete prediction maps, not a continuum-consistency result.

The grouping in Table 5 counts observation classes: seven strong-form, one weak, and three integral or nonlocal problems. It is not the certificate-route split (seven designs certified by the pointwise procedure, three by the weak jet-atom procedure, and one by the nonlocal signed-measure procedure; Section D.3). The integro-differential and Caputo designs are observed as nonlocal but certified through their stored quadrature functionals, which accounts for the difference between the two splits. The stored route for each design is recorded in its reference manifest in the accompanying code.

Table 4: Eleven PDE benchmarks in the 770-run training study. Here N counts scalar constraints after splitting vector equations into rows, and d_2 is the network output dimension. Applying the displayed value or derivative operator to each manufactured solution yields the corresponding data.

Problem	N	d_2	PDE and sampled data
Poisson 1D	7	1	$-u_{xx} = f$; five interior and two endpoint Dirichlet rows.
Variable elliptic 2D	13	1	$-\nabla\cdot(a\nabla u) + 0.35u = f$ with $a(x,y) = 2 + 0.3x + 0.2y$; five interior rows and eight Robin rows $\partial_n u + \beta u = g$, where $\beta = 0.75 + 0.08x - 0.05y$.
Heat 1D	17	1	$u_t - 0.35u_{xx} = f$; six interior, five initial, and six endpoint Dirichlet rows.
Transport 1D	15	1	$u_t + c(t,x)u_x + 0.25u = f$ with $c(t,x) = 0.65 + 0.12x + 0.08t$; six interior, five initial, and four inflow rows at $x = -1$.
Wave 1D	22	1	$u_{tt} - 0.8^2 u_{xx} = f$; six interior rows, five co-located initial displacement/velocity pairs, and six endpoint Dirichlet rows.
Biharmonic 2D	21	1	$\Delta^2 u = f$; five interior rows and eight co-located clamped pairs $u = g_0$, $\partial_n u = g_1$.
Stokes 2D	32	3	$-0.2\Delta\mathbf{v} + \nabla p = \mathbf{f}$ and $\nabla\cdot\mathbf{v} = 0$; five shared interior locations, one pressure anchor, and both velocity components at eight boundary locations.
Weak Poisson 1D	6	1	Four rows $\int_{-1}^{1} u' v_n'\,dx$ with $v_n' = \sqrt{(2n+1)/2}\,P_n$, $n = 1,\dots,4$, plus two endpoint Dirichlet rows; 16-node Gauss–Legendre quadrature.
Nonlocal diffusion 1D	5	1	Three rows at $x_i \in \{-0.5, 0, 0.5\}$ of $\int_{-1}^{1} \kappa(x_i,y)\{u(x_i) - u(y)\}\,dy = f(x_i)$, where $\kappa(x,y) = \exp\left[-\frac{1}{2}\left(\frac{x-y}{0.5}\right)^2\right]$, plus two endpoint Dirichlet rows; 18-node Gauss–Legendre quadrature.
Integro-differential 1D	6	1	Five rows of $u'(x) + (0.4 + 0.1x)u(x) - 0.55\int_{-1}^{1} \exp\left[-\frac{1}{2}\left(\frac{x-y}{0.42}\right)^2\right] u(y)\,dy = f(x)$ and one Dirichlet row at $x = -1$; 18-node Gauss–Legendre quadrature.
Caputo diffusion 1D	17	1	${}^{C}D_t^{0.6} u - 0.3u_{xx} = f$; six fractional residual, five initial, and six endpoint Dirichlet rows; 16-node transformed quadrature for the Caputo derivative.

The scalar manufactured solutions, in the order of Table 4, are

$$
\begin{aligned}
u_{\mathrm{P}}^*(x) &= \sin(1.3x + 0.2) + 0.17\cos(2.1x - 0.1),\\
u_{\mathrm{VE}}^*(x,y) &= \sin(0.8x + 0.3)\cos(1.1y - 0.2) + 0.13\sin(xy + 0.4),\\
u_{\mathrm{H}}^*(t,x) &= e^{-0.7t}\{\sin(1.2x + 0.25) + 0.2\cos(0.7x - 0.1)\},\\
u_{\mathrm{T}}^*(t,x) &= \sin(0.9x - 0.8t + 0.35) + 0.16\cos(1.3x + 0.4t).
\end{aligned}
$$

The next four cases use

$$
\begin{aligned}
u_{\mathrm{W}}^*(t,x) &= \cos(1.1t + 0.1)\sin(0.9x + 0.4) + 0.15\sin(0.7t + 1.3x),\\
u_{\mathrm{B}}^*(x,y) &= \sin(0.8x + 0.2)\cos(1.05y - 0.3) + 0.11\cos(0.6xy + 0.2),\\
u_{\mathrm{WP}}^*(x) &= \sin(1.15x + 0.3) + 0.19\cos(0.65x - 0.2),\\
u_{\mathrm{NL}}^*(x) &= \sin(0.95x + 0.23) + 0.21\cos(1.45x - 0.17).
\end{aligned}
$$

The final pair is

$$
\begin{aligned}
u_{\mathrm{ID}}^*(x) &= e^{0.2x}\sin(1.1x + 0.4) + 0.3\cos(0.6x - 0.1),\\
u_{\mathrm{C}}^*(t,x) &= (1 + t^2)\sin(0.9x + 0.2) + 0.1e^{-0.3t}\cos(0.7x - 0.1).
\end{aligned}
$$

For Stokes, $\mathbf{u}^* = (v_1^*, v_2^*, p^*)$ is

$$
\begin{aligned}
v_1^*(x,y) &= 1.1\sin(0.8x + 0.2)\cos(1.1y - 0.3) + 0.1x,\\
v_2^*(x,y) &= -0.8\cos(0.8x + 0.2)\sin(1.1y - 0.3) - 0.1y,\\
p^*(x,y) &= 0.3\cos(0.7x - 0.4y + 0.1) + 0.08x - 0.05y.
\end{aligned}
$$

Thus every PDE target is compatible by construction. The fractional claim concerns the displayed transformed-quadrature prediction map; we infer no continuum fractional-space positivity conclusion from that experiment. The stored discrete functionals act on $C^1([-1,1])$ for weak Poisson, $C([-1,1])$ for nonlocal diffusion, $C^1([-1,1])$ for the integro-differential problem, and $C^2([0,1]\times[-1,1])$ for Caputo diffusion.

Collocation points and loss groups We scale every named group below by the inverse square root of its number of scalar rows. Time-dependent problems use coordinate order (t,x); the other pairs are (x,y). The fixed point sets are:

- **Poisson:** interior $\{-0.79,-0.34,0.07,0.43,0.82\}$ and boundary $\{-1,1\}$, forming groups of sizes 5 and 2.
- **Variable elliptic:** interior

$$\{(-0.71,-0.48),(-0.38,0.26),(0.06,-0.17),(0.41,0.63),(0.77,-0.31)\},$$

  and Robin boundary points

$$\{(-1,-0.64),(-1,0.23),(1,-0.19),(1,0.72),(-0.57,-1),(0.18,-1),(-0.24,1),(0.69,1)\},$$

  with outward coordinate normals and group sizes 5 and 8.
- **Heat:** interior points

$$\{(0.13,-0.72),(0.24,0.18),(0.39,-0.27),(0.56,0.69),(0.73,-0.51),(0.91,0.34)\},$$

  initial locations $x\in\{-0.83,-0.29,0.14,0.58,0.91\}$, and boundary points $\{(0.17,-1),(0.48,-1),(0.86,-1),(0.29,1),(0.67,1),(0.94,1)\}$, forming groups of sizes 6, 5, and 6.
- **Transport:** interior points

$$\{(0.11,-0.66),(0.22,0.09),(0.37,-0.24),(0.53,0.73),(0.76,-0.39),(0.92,0.41)\},$$

  initial locations $x\in\{-0.78,-0.31,0.12,0.49,0.87\}$, and inflow points $\{(0.16,-1),(0.43,-1),(0.71,-1),(0.93,-1)\}$, forming groups of sizes 6, 5, and 4.
- **Wave:** interior points

$$\{(0.12,-0.58),(0.28,0.21),(0.44,-0.33),(0.61,0.74),(0.79,-0.47),(0.93,0.38)\},$$

  initial locations $x\in\{-0.82,-0.36,0.09,0.53,0.88\}$, and boundary points $\{(0.18,-1),(0.52,-1),(0.87,-1),(0.31,1),(0.68,1),(0.95,1)\}$. The interior, initial-displacement, initial-velocity, and boundary groups have sizes 6, 5, 5, and 6.
- **Biharmonic:** interior points

$$\{(-0.68,-0.43),(-0.29,0.24),(0.08,-0.16),(0.39,0.67),(0.74,-0.28)\},$$

  and boundary points

$$\{(-1,-0.61),(-1,0.32),(1,-0.22),(1,0.73),(-0.52,-1),(0.19,-1),(-0.27,1),(0.66,1)\}.$$

  The interior, boundary-value, and boundary-normal groups have sizes 5, 8, and 8; the normal is the outward coordinate normal.
- **Stokes:** shared interior locations

$$\{(-0.66,-0.41),(-0.31,0.28),(0.07,-0.14),(0.42,0.64),(0.76,-0.27)\},$$

  pressure anchor $(0.07,-0.14)$, and velocity boundary locations

$$\{(-1,-0.56),(-1,0.37),(1,-0.18),(1,0.69),(-0.47,-1),(0.21,-1),(-0.25,1),(0.63,1)\}.$$

  The momentum, continuity, pressure-anchor, and velocity-boundary groups contain 10, 5, 1, and 16 scalar rows.

Table 6: Independent base functionals for the six two-row controls.

Family	Independent base functionals (ℓ_1, ℓ_2)
Pointwise Poisson	$(-u''(0.07), u(0.07))$.
Vector Stokes	$(-0.2\Delta v_1 + p_x, -0.2\Delta v_2 + p_y)$ at $(x, y) = (-0.66, -0.41)$.
Weak Poisson	The $n = 1$ and $n = 2$ rows $\int_{-1}^{1} u'v_n' \, dx$ under the 16-node Gauss–Legendre rule.
Nonlocal diffusion	The displayed nonlocal residuals at anchors $x = -0.5$ and $x = 0.5$ under the 18-node rule.
Weak boundary–residual	The Dirichlet value $u(-1)$ and the $n = 1$ weak-Poisson row.
Nonlocal boundary–residual	The Dirichlet value $u(-1)$ and the nonlocal residual at $x = -0.5$.

- **Weak Poisson:** the four weak rows form one group and use the standard 16-node Gauss–Legendre nodes on $[-1, 1]$; the two endpoint values form the boundary group. In detail,

$$v_n(x) = \int_{-1}^{x} \sqrt{\frac{2n+1}{2}} P_n(s)\, ds, \qquad \lambda_n(u) = \sum_{r=1}^{16} \omega_r u'(y_r) \sqrt{\frac{2n+1}{2}} P_n(y_r), \quad n = 1, \dots, 4. \tag{114}$$

  The weak-row formula in (114) uses the Gauss–Legendre nodes and weights (y_r, ω_r) on $[-1, 1]$; in particular, $v_n(-1) = v_n(1) = 0$.
- **Nonlocal diffusion:** the anchors $\{-0.5, 0, 0.5\}$ form the three-row residual group, with a standard 18-node Gauss–Legendre rule on $[-1, 1]$; the two endpoint values form the boundary group.
- **Integro-differential:** residual anchors $\{-0.76, -0.33, 0.08, 0.47, 0.81\}$ form one five-row group using the same 18-node rule, and $u(-1)$ is a separate one-row boundary group.
- **Caputo diffusion:** fractional-residual anchors

$$\{(0.16, -0.68), (0.29, 0.17), (0.43, -0.26), (0.61, 0.72), (0.78, -0.44), (0.93, 0.36)\},$$

  initial locations $x \in \{-0.81, -0.32, 0.11, 0.55, 0.89\}$, and boundary points $\{(0.19, -1), (0.51, -1), (0.84, -1), (0.34, 1), (0.69, 1), (0.96, 1)\}$, forming groups of sizes 6, 5, and 6.

For completeness, if $(z_r, \omega_r)_{r=1}^{16}$ are the Gauss–Legendre nodes and weights on $[0, 1]$, the stored Caputo row at (t, x) uses

$${}^{C}D_t^{\alpha} u(t, x) \approx \frac{t^{1-\alpha}}{(1-\alpha)\Gamma(1-\alpha)} \sum_{r=1}^{16} \omega_r u_t\Big(t\left(1 - z_r^{1/(1-\alpha)}\right), x\Big), \qquad \alpha = 0.6. \tag{115}$$

The change of variables in (115) cancels the endpoint singularity before Gaussian quadrature. Together with Table 4, these points, groups, and manufactured solutions specify every constraint used in the eleven-PDE study.

K.3 Compatible two-row controls

The six control constructions use the pairs in Table 6 before coefficient normalization and row mixing. We divide each coefficient vector by its Euclidean norm and apply the same scaling to its manufactured target. Weak and nonlocal rows are the exact stored quadrature functionals, so the associated conclusions concern the optimized discrete loss.

For every control family, let (ℓ_1, ℓ_2) be the normalized independent base pair. The row matrix either retains these rows or applies

$$\mathbf{C}_\varepsilon = \begin{pmatrix} 1 & 0 \\ (1+\varepsilon^2)^{-1/2} & \varepsilon(1+\varepsilon^2)^{-1/2} \end{pmatrix}, \qquad \varepsilon \in \{1, 0.1, 0.01, 0\}. \tag{116}$$

For manufactured data, we apply the same transformed functionals to the manufactured solution. For the independent baseline and for $\varepsilon > 0$, a second compatible choice adds the antisymmetric contrast

$\tau(-1,1)^\top$, with $\tau = 0.5$, to the two raw row targets. We then apply the common control-group scale $1/\sqrt{2}$. For the mixed designs, the perturbation increasingly excites the weak spectral direction as ε decreases. At $\varepsilon = 0$ the two rows and their manufactured values are identical. Consequently every reported target obeys all exact linear relations of its prediction map.

K.4 Constraint independence and numerical rank

For $\varepsilon > 0$, $\mathbf{C}_\varepsilon$ is invertible, whereas at $\varepsilon = 0$ it is singular with left-null vector $(1,-1)^\top$; equivalently, $\mathbf{C}_0^\top(1,-1)^\top = 0$. The pointwise coefficient vectors, weak functionals, and signed measures of the base pairs are independent by construction. Thus the rank formulas in Propositions 3.1 to 3.3 predict nonsingular two-row Gram matrices for the nonredundant designs and singular ones for the duplicates. These controls are deliberately constructed so that their ranks follow algebraically before any network is initialized. Their purpose is to check whether the finite-width kernels and optimization trajectories reproduce the predicted effects of exact and near row dependence, not to show that arbitrary PDE discretizations are nonredundant.

Table 7 reports the comparison.

Table 7: Structural and observed finite-width Gram status. The output-block entry reports the single design-side rank certificate; it is not a second hidden-block test. The 960 qualifying runs use 480 independent kernels and two compatible targets. The Gram-status and certificate tallies are values of the registered campaign; the tracked aggregate stores the GD-outcome counts.

Rows	Kernels	GD runs	Gram status (pred./obs.)	GD outcome
Independent	480	960	nonsingular/nonsingular in 480 of 480	Output-block certificate passed in 480 of 480; loss below its initial value in 959 of 960
Exact duplicate	120	120	singular/singular in 120 of 120	loss reduced $> 100\times$ in 120 of 120

The structural labels in the first half of the status column are algebraic and do not depend on a numerical tolerance. For the observed label, we symmetrize the Gram matrix and declare an eigenvalue nonzero when it exceeds 10^{-12} times that matrix's largest eigenvalue; this relative tolerance was fixed before aggregation and was not selected by family or optimization outcome. No output-block eigenvalue cutoff is used as a structural test; instead, the stored exact rank witness certifies that block and hence the full kernel without a hidden-block PD check. Exact duplication, in contrast, makes the two Jacobian rows identical and forces singularity independently of floating-point classification. The agreement in all 600 instances is an implementation check across six constraint classes; the analytic theorems, rather than the chosen tolerance, establish the population nullspaces.

K.5 Positive-definite and compatible singular GD

All 480 manufactured-data cases whose initial NTK passes the numerical definiteness test reduce their training loss by at least a factor of ten within 3,000 steps. Across the 960 runs with either compatible target, 959 finish below their starting loss and 812 achieve a tenfold reduction; the sole failure is a contrast run with $d_1 = 16$. All 720 cases with $d_1 \geq 256$ improve, and 612 attain $\mathcal{L}_{3,000}/\mathcal{L}_0 \leq 0.1$. At $d_1 = 16{,}384$, each of the 240 trajectories lowers its training loss. Among them, 200 achieve at least a tenfold reduction, and the median final-to-initial ratio is 6.89×10^{-7} (the counts 812, 612, and 200 are per-run tallies of the registered campaign that the tracked aggregate does not store). Contrast cases with severe ill-conditioning can nevertheless retain most of their starting loss over this finite budget, as indicated by their slow spectral modes.

For duplicate rows, both predictions and targets lie in

$$\mathcal{U}_{\text{dup}} = \{(z,z)^\top : z \in \mathbb{R}\}, \qquad \mathbf{v}_T - \mathbf{v}_P(\theta) \in \mathcal{U}_{\text{dup}} \quad \text{for every } \theta. \tag{117}$$

The subspace identity in (117) holds globally. Although the full two-by-two NTK is singular, its compression to $\mathcal{U}_{\text{dup}}$ has a positive eigenvalue. This invariant property is the global invariant-subspace hypothesis in Corollary E.4, rather than compatibility imposed only at initialization. All 120 singular

runs finish at less than one percent of their initial loss; 118 reach absolute loss below 10^{-12}, including all 90 runs with $d_1 \geq 256$. The remaining two cases with $d_1 = 16$ still improve by factors of approximately 694 and 416 (registered-campaign values outside the tracked aggregate). Under the theorem's other hypotheses, a full positive-definiteness certificate would remove the target-compatibility restriction; the duplicate-row results show that such a certificate is unnecessary when the complete error trajectory lies in an active subspace on which the compressed positive-definiteness certificate of Definition E.2 holds.

Compatibility is essential to this conclusion. If the two duplicate rows were assigned the contrast target used for the independent systems, then its antisymmetric component would be orthogonal to $\mathcal{U}_{\mathrm{dup}}$. With the experiment's contrast amplitude $\tau = 0.5$ and common $1/\sqrt{2}$ row scaling, the exact attainable-loss decomposition gives

$$\|\mathbf{v}_P(\theta) - \mathbf{v}_T\|_2^2 = \|\Pi_{\mathcal{U}_{\mathrm{dup}}}(\mathbf{v}_P(\theta) - \mathbf{v}_T)\|_2^2 + \tau^2 \geq \tau^2 = 0.25.$$

No incompatible-target trajectory is included in the reported counts. Thus the duplicate experiment demonstrates convergence for compatible singular systems; it does not empirically establish necessity of the compatibility condition, which the three-row audit below supplies through its incompatible-target pairs.

In the generated paragraph below, ϵ is the size of the third-row perturbation $A + B + \epsilon C$ of that audit and is unrelated to the row-mixing parameter ε of the two-row controls. Its rank table also shows that at $\epsilon = 10^{-5}$ the relative tolerance 10^{-12} underestimates the rank of the rank-three nonlocal design, which is why the tolerance discussion above is restricted to the 600 two-row instances.

Three-row joint-nullspace controls We tested rows A, B, and $A + B + \epsilon C$ for a co-located two-output pointwise design, a weak-quadrature residual jointly with a boundary value, and a finite-measure nonlocal residual jointly with a boundary value. The audit contains 72 initial kernels (3 cases, 4 values of ϵ, 2 widths, and 3 seeds) in float64. The parenthesized counts below are agreements with the exact coefficient-design rank $r_\star$; r_τ is numerical rank at relative tolerance τ, and $\rho_z = \|Kz\|/(\|K\|\|z\|)$. The displayed rank set records all values observed over the configured widths and seeds.

Design	ϵ	$r_\star$	$r_{10^{-6}}$	$r_{10^{-10}}$	$r_{10^{-12}}$	max. ρ_z
Co-located	0	2	2 (6/6)	2 (6/6)	2 (6/6)	2.86×10^{-16}
Co-located	10^{-1}	3	3 (6/6)	3 (6/6)	3 (6/6)	5.06×10^{-3}
Co-located	10^{-3}	3	2 (0/6)	3 (6/6)	3 (6/6)	3.48×10^{-5}
Co-located	10^{-5}	3	2 (0/6)	2 (0/6)	3 (6/6)	3.46×10^{-7}
Weak+∂	0	2	2 (6/6)	2 (6/6)	2 (6/6)	2.59×10^{-16}
Weak+∂	10^{-1}	3	3 (6/6)	3 (6/6)	3 (6/6)	9.45×10^{-3}
Weak+∂	10^{-3}	3	2 (0/6)	3 (6/6)	3 (6/6)	8.96×10^{-5}
Weak+∂	10^{-5}	3	2 (0/6)	2 (0/6)	3 (6/6)	8.96×10^{-7}
Nonlocal+∂	0	2	2 (6/6)	2 (6/6)	2 (6/6)	1.15×10^{-16}
Nonlocal+∂	10^{-1}	3	3 (6/6)	3 (6/6)	3 (6/6)	6.46×10^{-4}
Nonlocal+∂	10^{-3}	3	2 (0/6)	3 (6/6)	3 (6/6)	6.42×10^{-6}
Nonlocal+∂	10^{-5}	3	2 (0/6)	2 (0/6)	2 (0/6)	6.41×10^{-8}

For the exact relation ($\epsilon = 0$), we additionally ran 18 paired settings, hence 36 trajectories: manufactured compatible targets and targets displaced only along the known left null direction. The predicted incompatible loss floor was 6.25×10^{-2}. Across all 19 saved steps, the largest error in the predicted paired loss offset was 4.44×10^{-16}; the largest final paired parameter difference was 2.09×10^{-14}.

Design	pairs	max. $\mathcal{L}_T^{\mathrm{comp}}$	max. $\mathcal{L}_T^{\mathrm{inc}}$	max. offset error
Co-located	6	3.16×10^{-32}	6.25×10^{-2}	1.39×10^{-17}
Weak+∂	6	5.24×10^{-32}	6.25×10^{-2}	4.16×10^{-17}
Nonlocal+∂	6	1.10×10^{-10}	6.25×10^{-2}	3.47×10^{-17}

These controls diagnose the exact stored finite prediction maps. The weak and nonlocal rows use fixed 16- and 18-node quadrature, respectively; they do not establish continuum discretization error. Numerical rank depends on the displayed tolerance, and the GD runs do not certify the theorem's deterministic width or B_* step-size bounds.

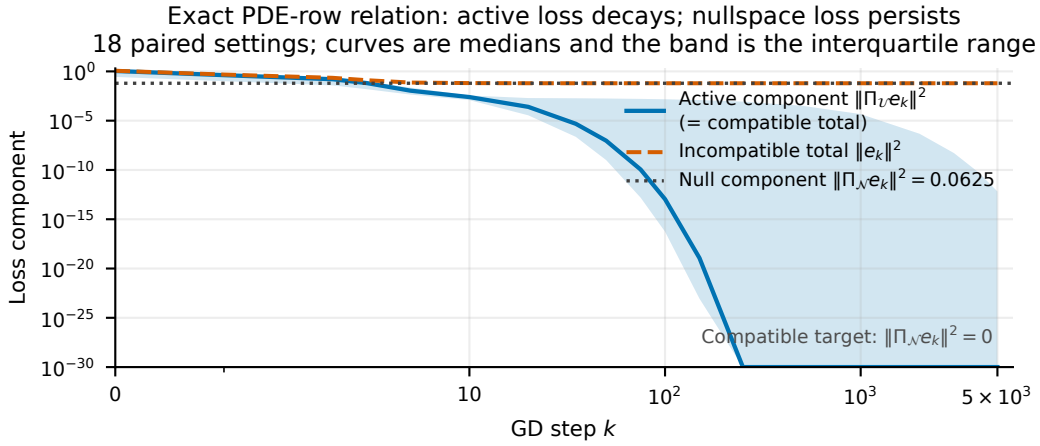


Figure 2: Exact structural nullspace control in three PDE settings. For each setting the three loss-scaled prediction rows are A, B, and $A + B$, so every prediction $\mathbf{v}_P(\theta) \in \mathbb{R}^3$ satisfies $\mathbf{z}^\top \mathbf{v}_P(\theta) = 0$ for $\mathbf{z} = (1, 1, -1)^\top/\sqrt{3}$. Define $\mathcal{N} = \mathrm{span}\{\mathbf{z}\}, \mathcal{U} = \mathcal{N}^\perp$, and the training error $\mathbf{e}_k = \mathbf{v}_P(\theta_k) - \mathbf{v}_T$; the plotted loss is $\mathcal{L}_k = \|\mathbf{e}_k\|_2^2$. A compatible target has $\Pi_\mathcal{N} \mathbf{e}_k = 0$. Its paired incompatible target is displaced by $0.25\mathbf{z}$, hence $\|\Pi_\mathcal{N} \mathbf{e}_k\|_2^2 = 0.25^2 = 0.0625$ at every step, while both targets induce the same parameter updates and therefore share the same active loss $\|\Pi_\mathcal{U} \mathbf{e}_k\|_2^2$. The 18 paired settings comprise a co-located vector PDE, a weak-residual–boundary PDE, and a nonlocal-residual–boundary PDE, with widths $d_1 \in \{32, 128\}$ and three seeds. Full-batch GD uses $\eta = 0.2/\lambda_{\max}(\mathbf{K}_{(0)})$ for 5,000 steps. Curves are medians across the 18 settings; the blue band is the interquartile range. Compatible nullspace loss is exactly zero and therefore is stated rather than drawn on the logarithmic axis. Active losses below 10^{-30} are placed at the display boundary. At all 19 saved checkpoints, the paired total-loss offset agrees with 0.0625 within 4.44×10^{-16}. The weak and nonlocal rows are the fixed stored quadrature functionals, so this is a discrete-loss diagnostic rather than a continuum-error result.

K.6 Conditioning and the GD rate

The two singular values of the coefficient transformation in Equation (116) imply

$$\lambda_{\min}(\mathbf{C}_\varepsilon \mathbf{C}_\varepsilon^\top) = 1 - \frac{1}{\sqrt{1+\varepsilon^2}} = \frac{\varepsilon^2}{2} + O(\varepsilon^4). \tag{118}$$

The expansion in (118) shows that reducing ε preserves rank for every $\varepsilon > 0$ and creates an increasingly weak row direction. Write $\rho_0 := \lambda_{\min}(\mathbf{K}_{(0)})/\lambda_{\max}(\mathbf{K}_{(0)})$ for the normalized initial minimum eigenvalue and $\kappa_0 = 1/\rho_0$ for the initial condition number plotted in Figure 1(b). In the registered campaign the median ρ_0 decreased as ε went from 1 to 0.1 to 0.01 in all 24 family–width settings, and the compatible contrast-target loss ratio worsened in 23 of 24; the only exception was the nonlocal-diffusion instance at $d_1 = 16$. Those 24 settings are recorded in the registered campaign but lie outside the tracked aggregate, which retains the largest width only. Table 8 reports that largest-width control, where the expected ordering holds in both quantities for all six families. The independent Poisson baseline without row mixing, likewise recorded in the original campaign but outside the tracked aggregate, had $\rho_0 = 3.621 \times 10^{-1}$ and $\mathcal{L}_{3,000}/\mathcal{L}_0 = 4.94 \times 10^{-32}$.

Table 8: Conditioning and finite-horizon loss at $d_1 = 16{,}384$. Entries are five-seed medians after $T = 3{,}000$ GD steps, with $\mathcal{L}_T/\mathcal{L}_0$ for the compatible contrast target; every entry is taken from the tracked aggregate that also generates Figure 1(b). Here ρ_0 is the normalized initial minimum eigenvalue defined above. Every row design is independent and every contrast target is compatible. The weak boundary–residual entry at $\varepsilon = 1$ is exactly zero in float64.

	$\varepsilon = 1$		$\varepsilon = 0.1$		$\varepsilon = 0.01$	
Family	ρ_0	$\mathcal{L}_T/\mathcal{L}_0$	ρ_0	$\mathcal{L}_T/\mathcal{L}_0$	ρ_0	$\mathcal{L}_T/\mathcal{L}_0$
Pointwise Poisson	8.778×10^{-2}	8.03×10^{-31}	2.416×10^{-3}	4.17×10^{-4}	2.667×10^{-5}	1.73×10^{-1}
Vector Stokes	2.037×10^{-1}	5.22×10^{-32}	2.665×10^{-3}	1.10×10^{-3}	2.637×10^{-5}	8.00×10^{-1}
Weak Poisson	5.007×10^{-2}	6.55×10^{-29}	5.949×10^{-4}	5.93×10^{-2}	5.956×10^{-6}	3.41×10^{-1}
Nonlocal diffusion	3.871×10^{-3}	6.36×10^{-30}	2.359×10^{-5}	5.93×10^{-1}	1.974×10^{-7}	9.33×10^{-1}
Weak boundary–residual	2.703×10^{-1}	0	3.301×10^{-3}	1.04×10^{-4}	3.195×10^{-5}	3.70×10^{-1}
Nonlocal boundary–residual	1.344×10^{-3}	1.21×10^{-2}	1.663×10^{-5}	3.51×10^{-1}	1.687×10^{-7}	3.71×10^{-1}

Two features of Table 8 qualify how these numbers may be read. First, between $\varepsilon = 0.1$ and $\varepsilon = 0.01$ the tabulated medians of ρ_0 fall by factors between 90 and 120 across the six families, consistent with the ε^2 scaling in (118); because these are medians of per-run ratios rather than a fitted exponent, we report the comparison as a diagnostic and not as an estimated rate. Second, ρ_0 does not order the finite-horizon outcome across families. At $\varepsilon = 0.01$ the pointwise Poisson, weak boundary–residual, and vector Stokes controls have median κ_0 of 3.75×10^4, 3.13×10^4, and 3.79×10^4 but loss ratios of 1.73×10^{-1}, 3.70×10^{-1}, and 8.00×10^{-1}; at the ill-conditioned end, nonlocal boundary–residual at $\kappa_0 = 5.93 \times 10^6$ finishes at 3.71×10^{-1} while nonlocal diffusion at $\kappa_0 = 5.07 \times 10^6$ finishes at 9.33×10^{-1}.

Part of the apparent trend in κ_0 is fixed by the step-size rule rather than measured. Since $\eta = 0.2/\lambda_{\max}(\mathbf{K}_{(0)})$, every mode of the frozen residual recursion $\widetilde{\mathbf{r}}_{(k+1)} = (\mathbf{I} - 2\eta\mathbf{K}_{(0)})\widetilde{\mathbf{r}}_{(k)}$ contracts by a factor in $[0.6, 1 - 0.4\rho_0]$, so the frozen loss obeys $\widetilde{\mathcal{L}}_{3{,}000}/\mathcal{L}_0 \le (1 - 0.4\rho_0)^{6{,}000}$, which is an arithmetic consequence of the normalization and not an empirical finding. It is also not a statement about the plotted quantity. The vertical axis of Figure 1(b) is the observed nonlinear loss, and three of the eighteen plotted medians—pointwise Poisson, vector Stokes, and weak Poisson at $\varepsilon = 1$—lie many orders of magnitude above this frozen bound, because those runs reach float64 roundoff long before step 3,000. Where the bound does hold informatively it does not separate the families: at $\varepsilon = 0.01$ it exceeds 0.92 for all six, while the observed medians range from 1.73×10^{-1} to 9.33×10^{-1}. We therefore draw no rate or timescale conclusion from κ_0. The five $\varepsilon = 1$ medians below the 10^{-16} display floor are drawn at the boundary by the plotting script, so the small-ε end of each curve carries the information.

K.7 Width, NTK stability, and scope

The PDE training study covers 770 cases outside the two-row controls. Every run reduces the sampled training loss, and 767 improve it by at least a factor of ten (registered-campaign tallies; per-run records are not retained). Table 9 reports pooled medians that summarize the transition toward the frozen-kernel regime. For each numerical population reference, we integrate the output-weight coordinates analytically and estimate the hidden-parameter expectation with eight independently scrambled Sobol replicates. Each uses 32,768 fine samples and its nested first 16,384 samples as a coarse approximation. The base seed is 1,729, and the chunk size is 2,048. The conservative gap subtracts 2.365 times the between-replicate standard error and the coarse–fine operator-norm difference; its relative numerical tolerance is 10^{-8}.

Let $\widehat{\mathbf{K}}^{\infty}$ denote this numerical population reference, let $\mathbf{K}_{(k)}$ be the empirical Gram matrix at GD step k, and let $\mathcal{C}$ be the set of saved checkpoints. The frozen residual starts from $\widetilde{\mathbf{r}}_{(0)} = \mathbf{r}_{(0)}$ and evolves at every GD step by $\widetilde{\mathbf{r}}_{(k+1)} = (\mathbf{I} - 2\eta\mathbf{K}_{(0)})\widetilde{\mathbf{r}}_{(k)}$. With $T = 3{,}000$, the three relative diagnostics in Table 9 are

$$\frac{\|\mathbf{K}_{(0)} - \widehat{\mathbf{K}}^{\infty}\|_{\mathrm{op}}}{\max\{\|\widehat{\mathbf{K}}^{\infty}\|_{\mathrm{op}}, \mathrm{tiny}_{64}\}}, \qquad \max_{k \in \mathcal{C}} \frac{\|\mathbf{K}_{(k)} - \mathbf{K}_{(0)}\|_{\mathrm{op}}}{\max\{\|\mathbf{K}_{(0)}\|_{\mathrm{op}}, \mathrm{tiny}_{64}\}}, \qquad \frac{\|\mathbf{r}_{(T)} - \widetilde{\mathbf{r}}_{(T)}\|_2}{\max\{\|\widetilde{\mathbf{r}}_{(T)}\|_2, \mathrm{tiny}_{64}\}},$$

respectively, where tiny_{64} is the smallest positive normal float64 number. Each ratio is computed per run before taking the pooled median.

Table 9: Width diagnostics pooled over 110 PDE runs per width, transcribed from the registered 770-run campaign; the tracked aggregate stores this transcription with its content hash. All three quantities are relative; initial-kernel error and maximum drift use operator norm. Frozen discrepancy compares the nonlinear residual with the trajectory obtained by holding the empirical Gram matrix at initialization. Gap retained counts runs that preserve at least half the numerical population-reference gap at every saved checkpoint, not at every GD iterate.

d_1	Initial kernel error	Maximum kernel drift	Final frozen discrepancy	Gap retained
16	2.23×10^{-1}	6.11×10^{-1}	9.08×10^{-1}	17/110
256	5.46×10^{-2}	4.84×10^{-2}	2.74×10^{-1}	110/110
4,096	1.56×10^{-2}	3.33×10^{-3}	2.25×10^{-2}	110/110
16,384	8.09×10^{-3}	8.62×10^{-4}	6.91×10^{-3}	110/110

Across 100 initialization seeds per PDE, log–log slopes of the median operator error over widths 64–16,384 range from -0.512 to -0.469, consistent with square-root concentration; these slopes are also registered-campaign values. These five-scale fits do not establish a sharp asymptotic exponent. More fundamentally, the experiment does not certify the deterministic derivative envelopes or the explicit sufficient-width and step-size inequalities in Equation (9). The observations cover 3,000 GD steps with broad diagnostics every 100 steps, so they support finite-horizon spectral stability and training-loss contraction rather than an all-iteration proof. Finally, the optimized quantity is sampled constraint loss; we draw no PDE generalization or continuum-solution conclusion from it.

Table 9 reports four of the seven trained widths; $d_1 \in \{4, 64, 1{,}024\}$ were trained and are counted in the 770-run totals but are not summarized here. At $d_1 = 4$ the necessary condition $d_1(d_0 + 1 + d_2) \geq N$ fails for the heat, wave, biharmonic, Stokes, and Caputo designs. The campaign admits rank-infeasible widths, so those 50 runs are trained and counted in Table 5; their Jacobians cannot have full row rank, and we make no positivity or convergence claim for them. Finally, both runners record, per run, whether the loss is nonincreasing across saved checkpoints, but the retained 770-run and 1,080-run aggregates keep no monotonicity count. Consequently we report no checkpoint-monotonicity frequency for the 770-run or 1,080-run campaigns, and producing one would require rerunning the registered campaign; the per-iterate audit of Section K.8 is a separate reduced design recorded at every step.

K.8 Per-iterate monotonicity

Theorem 4.2 asserts a decrease at every iterate, whereas the campaigns above store the loss only at saved checkpoints. This audit closes that gap on a reduced design for which every step can be recorded.

The network is the two-layer map of Definition 2.1 with $\sigma = \tanh$, $\gamma = 1$, $d_0 = d_2 = 1$, and per-neuron parameters drawn i.i.d. from $\mathrm{Unif}[-1, 1]^3$. The two rows are the value row $\ell_1(u) = u(x_\star)$ and the mixed row $\ell_2^\varepsilon = (\ell_1 + \varepsilon \ell_1')/\sqrt{1 + \varepsilon^2}$, where $\ell_1'(u) = u'(x_\star)$ is the derivative row, at $x_\star = 0.35$, so ε is the row-mixing parameter used in the control study. Training is full-batch gradient descent on the unscaled loss of Definition 2.3 for 3,000 steps at $\eta = 0.2/\lambda_{\max}(\mathbf{K}_{(0)})$, with a contrast target that excites the weak direction. Widths are $d_1 \in \{256, 1024\}$, row mixings are $\varepsilon \in \{1, 0.1, 0.01\}$, and five seeds are run per cell, giving thirty runs.

An iterate is *scored* while the loss retains double-precision significance, that is while $\mathcal{L}(\theta_{(k)}) > 10^{-26}\mathcal{L}(\theta_{(0)})$. The cutoff is necessary and not conservative: the residual is a difference of two quantities that agree to machine precision once the loss falls that far, so the recorded per-step ratios quantize to small integers and carry no information. It is the same effect that the display floor of Figure 1 accommodates.

Table 10: Per-iterate audit. Violations are steps with $\mathcal{L}(\theta_{(k+1)}) > (1 + 10^{-12})\mathcal{L}(\theta_{(k)})$ among scored iterates, pooled over five seeds. The last two columns are the largest observed per-step loss ratio and the frozen-kernel prediction $(1 - 2\eta\lambda_{\min}(\mathbf{K}_{(0)}))^2$, reported for the first seed.

d_1	ε	Violations	Scored iterates	Worst ratio	Predicted
256	1	0	2,845	—	—
256	0.1	0	15,000	0.998359	0.998371
256	0.01	0	15,000	0.999983	0.999983
1,024	1	0	3,015	—	—
1,024	0.1	0	15,000	0.998387	0.998389
1,024	0.01	0	15,000	0.999983	0.999983
Total		0	65,860		

No violation occurs in any scored iterate, and in the four cells that retain a worst per-step ratio that ratio is below one, so the 10^{-12} tolerance is inactive there. At $\varepsilon = 1$ the loss reaches the precision cutoff before step 3,000, so fewer iterates are scored and no worst-ratio figure is quoted; the ratios recorded past that point are cancellation artifacts rather than measurements. At the two smaller mixings the full horizon is scored and the largest per-step ratio matches the frozen-kernel factor to a relative discrepancy between 2.8×10^{-9} (at $d_1 = 1{,}024$, $\varepsilon = 0.01$) and 1.2×10^{-5} (at $d_1 = 256$, $\varepsilon = 0.1$), which is the sharpest agreement with a predicted rate anywhere in these experiments.

Two limits. The design is two rows at one collocation point, not one of the eleven PDE benchmarks, so this audit speaks to the per-iterate claim and not to coverage. And the agreement with $(1 - 2\eta\lambda_{\min})^2$ is agreement with the *frozen* initial kernel, not with the certified constants of Equation (9); it is consistent with the kernel drift reported above being small at these widths, and is not a verification of the theorem's sufficient constants. The run is the per-iterate monotonicity experiment of the accompanying code, under its tracked configuration.

K.9 Reproduction

The tracked aggregate of the control study and the plotting script in the accompanying code reproduce Figure 1 with one command. The script validates the displayed counts and five-seed medians and records the hash of the original 1,080-row source table. Summary statistics cannot reconstruct that source table.

The prepare and submit commands documented with the code rerun the full campaign through its campaign driver. After such a run, the control aggregation script selects and aggregates compatible controls, while the evidence-assessment script produces the operator-class, width, Jacobian-drift, neuron-motion, frozen-dynamics, and rank-tolerance summaries. The current repository does not contain the original per-run 770-PDE, 1,080-control, or 7,700-initialization records; consequently, recomputing those broad-study summaries requires rerunning the registered campaign. This limitation does not affect regeneration of the main figure from its tracked aggregate.