REPAIRING NONLINEAR MODEL MERGES: A RANK–WIDTH SEPARATION FOR GELU NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Parameter addition need not add the functions implemented by task updates. We characterize the capacity of every local repair for nondegenerate two-layer exact-GELU networks with fixed hidden biases. The optimal population error has exactly four regimes: first order when final adapter rank cannot contain the summed first-order update; second order when an explicit neuronwise compatibility condition fails; third order when compatibility holds but exact additivity fails; or zero. A unique quadratic repair rescales rows and preserves the actual summed-update ranks, so additional rank cannot improve these error orders. A quantitative extension characterizes near-compatibility: uniformly over bounded direction families, third-order local accuracy is possible exactly when a coefficient obstruction is of the order of the update scale, and a correction within the concatenated adapter budget achieves it. Globally, we determine the minimum exact hidden width. The exact-additivity condition is also necessary and sufficient for any same-width representation, even at distant parameters and with an affine output branch. Rank-one tasks can instead require width $3m$. A calibration interpolation threshold and reproducible synthetic experiments distinguish population expressivity from finite-sample fitting and task preservation.

1 INTRODUCTION

Task arithmetic combines the weight changes made by separately adapted models (Ilharco et al., 2023). Its appeal is particularly clear for low-rank adaptation: task-specific changes can be stored and composed without maintaining an entire model for each task (Hu et al., 2022). However, the sum of parameter changes and the sum of functional changes are different objects. Alignment or better scalar weights do not by themselves determine whether the remaining discrepancy is expressible by the allowed parameters.

Our objective is population fidelity to signed functional addition. Given a common base θ_0 and fixed directions a, b, define

$$g_\varepsilon = f_{\theta_0+\varepsilon a} + f_{\theta_0+\varepsilon b} - f_{\theta_0}, \qquad \Delta_\varepsilon = f_{\theta_0+\varepsilon(a+b)} - g_\varepsilon. \tag{1}$$

A repair seeks a nearby parameter realizing g_ε more accurately. The familiar expansion $\Delta_\varepsilon = \varepsilon^2 D^2 f_{\theta_0}[a, b] + O(\varepsilon^3)$ suggests projecting an interaction onto a Jacobian range. That observation alone does not answer the structural question: *which interactions are in that range, which repairs fit the final adapter rank, and what happens after the quadratic term is removed?*

Our answer separates three resources. Rank determines whether the summed first-order update is admissible; neuron structure determines whether quadratic and cubic errors can be removed; width counts the nonlinear features needed for exact representation. The local classification has no nonzero fourth-order or higher-order regime, and its lower bounds cover *all* parameters in a fixed small neighborhood. The same exact-additivity condition also characterizes *global* same-width representability. For approximate compatibility, a uniform two-sided bound identifies precisely when an inexpensive repair achieves third-order accuracy. Thus the correction illustrates a capacity classification, rather than assuming that generic task updates are repairable.

Relation to existing approaches. Tangent-space task arithmetic relates weight disentanglement to linearization (Ortiz-Jiménez et al., 2023). Core-space merging aligns low-rank updates (Panariello

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.

et al., 2025), and Fréchet averaging incorporates symmetry-aware geometry, including LoRA quotient structure (da Silva et al., 2026). Output-space projection optimizes residual fits in a chosen basis (Evans et al., 2026); HyperFix learns subset-conditioned nonlinear corrections (Kim and Wang, 2026); M-Loss measures merging–ensembling discrepancies (Wang et al., 2026). Our distinction is an explicit necessary-and-sufficient classification over the *entire local parameter space*, including rank feasibility and the first unavoidable higher-order residual. Neural identifiability (Vlačić and Bölcskei, 2021), representation merging via neuron identifiability (Bürgin et al., 2026), and GELU skip-absorption constraints (Mijoski and Karbevski, 2026) supply related representation perspectives. Here independence of *derivative features with polynomial coefficients* connects representation uniqueness to quadratic repair, a cubic obstruction, and quantitative approximation guarantees.

2 Setting and the complete local classification

We use the exact GELU $\sigma(t) = t\Phi(t)$, where Φ is the standard normal cumulative distribution function (Hendrycks and Gimpel, 2016). Let

$$f_{W,V}(x) = \sum_{i=1}^{m} v_i \sigma(z_i(x)), \qquad z_i(x) = w_i^\top x + \beta_i, \tag{2}$$

where $W \in \mathbb{R}^{m\times d}$, $V \in \mathbb{R}^{k\times m}$, $x \in \mathbb{R}^d$, and v_i is column i of V. Hidden biases β_i are fixed throughout local merging and repair. The tasks are $a = (P, Q)$ and $b = (R, S)$; $p_i^\top, r_i^\top$ are rows of P, R, and q_i, s_i are columns of Q, S. We work with effective weight matrices, so LoRA factor rescalings do not alter the problem.

Assumption 1 (Nondegeneracy and population norm)**.** Every w_i and v_i is nonzero, and $(w_i, \beta_i) \neq \pm(w_j, \beta_j)$ for $i \neq j$. The input law μ is absolutely continuous, has density positive almost everywhere on a nonempty open set, and satisfies $\mathbb{E}_\mu \|x\|^8 < \infty$. Write $\mathcal{H} = L^2(\mu; \mathbb{R}^k)$.

Neither W nor V must have full rank; overcomplete hidden layers are allowed. Open support lets population identities determine analytic functions; finite-sample fitting is treated separately in Section 6. Define two neuronwise conditions:

$$\mathsf{S}_2: \quad p_i = 0 \text{ or } r_i = 0; \quad r_i \neq 0 \Rightarrow q_i \in \operatorname{span}\{v_i\}; \quad p_i \neq 0 \Rightarrow s_i \in \operatorname{span}\{v_i\}, \quad \forall i, \tag{3}$$

$$\mathsf{S}_\infty: \quad p_i \neq 0 \Rightarrow (r_i, s_i) = (0,0); \quad r_i \neq 0 \Rightarrow (p_i, q_i) = (0,0), \quad \forall i. \tag{4}$$

Under S_∞, a neuron whose input weight changes is modified by only one task; if neither task changes its input weight, their readout changes are unrestricted.

Let $r_W = \operatorname{rank}(P + R)$ and $r_V = \operatorname{rank}(Q + S)$. For nonnegative integer budgets and a sufficiently small closed ball $\overline{B}_\eta(\theta_0)$, define

$$E_{R_W,R_V}(\varepsilon) = \min_{\substack{\theta' \in \overline{B}_\eta(\theta_0) \\ \operatorname{rank}(W'-W)\le R_W,\ \operatorname{rank}(V'-V)\le R_V}} \|f_{\theta'} - g_\varepsilon\|_{\mathcal{H}}. \tag{5}$$

The rank constraints concern the *final effective updates*, not a separately appended correction branch. The minimum exists by compactness.

Theorem 1 (Complete local error-order classification)**.** *Under Assumption 1, for fixed task directions and sufficiently small $\eta > 0$, the asymptotics as $\varepsilon \downarrow 0$ are exactly those in Table 1. All implied positive constants may depend on the base, directions, and input law.*

The asymptotic *order* saturates, while leading constants and finite-scale performance can still improve with rank. Section 5 separately characterizes global exact representation; Appendix J shows why the local classification depends on the activation.

3 Why repair is possible, and why it stops at third order

3.1 The structural ingredient

The next lemma prevents different neurons from hiding each other's quadratic interaction. Its proof, including affine biases and opposite-ridge symmetries, is in Appendix B.

Table 1: Complete local classification. "Adequate" means $R_W \geq r_W$ and $R_V \geq r_V$. Raising either budget further cannot improve the listed error order.

Final adapter budget	Neuronwise condition	Optimal error
Inadequate	Any	$\Theta(\varepsilon)$
Adequate	Not S_2	$\Theta(\varepsilon^2)$
Adequate	S_2 but not S_∞	$\Theta(\varepsilon^3)$
Adequate	S_∞	0

Lemma 2 (GELU second-derivative independence)**.** *For the affine forms in Assumption 1, suppose*

$$\sum_i \left[A_i(x)\sigma''(z_i) + B_i(x)\sigma'(z_i) + c_i\sigma(z_i)\right] + \ell(x) = 0, \tag{6}$$

where $\deg A_i \leq 2$, $\deg B_i \leq 1$, *each* c_i *is constant, and* ℓ *is affine. Then every coefficient is zero. The statement holds coordinatewise for vector coefficients and for identities in* $\mathcal{H}$*.*

Proof mechanism. Restrict to a generic line $x = tu$. Write $z_i = \alpha_i t + \beta_i$ and use $\sigma' = \Phi + z\phi$, $\sigma'' = (2 - z^2)\phi$. Two derivatives leave a sum of polynomial multiples of $\exp(-\alpha_i^2 t^2/2 - \alpha_i\beta_i t)$. Distinct affine ridge classes yield distinct quadratic–linear exponent pairs on a generic line. Asymptotic separation forces each polynomial to vanish. The remaining affine terms vanish at the negative tail; the positive tail and the roots of $2 - z^2$ then force the original coefficients to vanish. Analyticity transfers a population identity on an open support to a global identity.

Let $J = Df_{\theta_0}$ and $h = D^2 f_{\theta_0}[a, b]$. Direct differentiation gives

$$h(x) = \sum_i \left[v_i\sigma''(z_i)(p_i^\top x)(r_i^\top x) + \sigma'(z_i)(q_i r_i^\top + s_i p_i^\top)x\right], \tag{7}$$

$$J(C, D)(x) = \sum_i \left[v_i\sigma'(z_i)c_i^\top x + d_i\sigma(z_i)\right]. \tag{8}$$

Uniformly for bounded (C, D),

$$f_{\theta_0+\varepsilon(a+b)+\varepsilon^2(C,D)} - g_\varepsilon = \varepsilon^2\{h + J(C, D)\} + O_{\mathcal{H}}(\varepsilon^3). \tag{9}$$

Theorem 3 (Exact quadratic repair)**.** *The equation* $h + J(C, D) = 0$ *is solvable if and only if* S_2 *holds. Its unique solution is*

$$C_* = -\Lambda R - MP, \qquad D_* = 0, \qquad \Lambda_{ii} = \frac{v_i^\top q_i}{\|v_i\|^2}, \quad M_{ii} = \frac{v_i^\top s_i}{\|v_i\|^2}. \tag{10}$$

Moreover, $C_* = -\Gamma(P + R)$ *for a diagonal* Γ*. Thus*

$$\Delta W_{\text{fix}} = \varepsilon(I - \varepsilon\Gamma)(P + R), \qquad \Delta V_{\text{fix}} = \varepsilon(Q + S) \tag{11}$$

preserves both summed-update ranks for all sufficiently small ε*.*

Proof. Lemma 2 makes each coefficient in $h + J(C, D)$ vanish separately. Since $v_i \neq 0$, the quadratic coefficient requires $(p_i^\top x)(r_i^\top x) = 0$, hence $p_i = 0$ or $r_i = 0$. The remaining coefficients require $q_i r_i^\top + s_i p_i^\top + v_i c_i^\top = 0$ and $d_i = 0$. These equations are equivalent to (3) and uniquely give (10). Disjoint row supports make C_* a row scaling of $P + R$. The diagonal multiplier in (11) is invertible for small ε. □

3.2 A CUBIC OBSTRUCTION AND ARBITRARY LOCAL PATHS

Under S_2, consider a neuron with $p_i \neq 0$, $r_i = 0$, and $s_i = \mu_i v_i$. After (10), its cubic residual is

$$h_{3,i} = -\mu_i(q_i + \mu_i v_i)\sigma'(z_i)(p_i^\top x) - \frac{\mu_i}{2} v_i\sigma''(z_i)(p_i^\top x)^2. \tag{12}$$

For $r_i \neq 0$, swap (p_i, q_i, μ_i) with (r_i, s_i, λ_i); fixed-input neurons contribute zero. Unless S_∞ holds, a nonzero quadratic coefficient survives in $h_3 = \sum_i h_{3,i}$. Lemma 2 implies $h_3 \notin \operatorname{range} J$, so no third-order parameter correction can cancel it.

One must still exclude nonsmooth parameter paths or unusually large corrections. The lemma also makes J injective. In finite dimensions there is an $\alpha > 0$ with $\|J\delta\|_{\mathcal{H}} \geq \alpha \|\delta\|$. Continuity of Df gives, on a sufficiently small convex ball,

$$\|f_\theta - f_{\theta'}\|_{\mathcal{H}} \geq \tfrac{\alpha}{2} \|\theta - \theta'\| . \tag{13}$$

Consequently every candidate with $O(\varepsilon^2)$ error is an $O(\varepsilon^2)$ parameter correction to the direct sum. When S_2 holds, every candidate with $O(\varepsilon^3)$ error is an $O(\varepsilon^3)$ correction to the unique repaired path. The expansions therefore apply uniformly to *all* optimizing candidates. Writing

$$d_2 = \operatorname{dist}_{\mathcal{H}}(-h, \operatorname{range} J), \qquad d_3 = \operatorname{dist}_{\mathcal{H}}(-h_3, \operatorname{range} J), \tag{14}$$

yields lower bounds $\varepsilon^2 d_2 - O(\varepsilon^3)$ or $\varepsilon^3 d_3 - O(\varepsilon^4)$ in the respective nonzero regimes. An insufficient rank budget instead has positive first-order SVD-tail distance from $(P + R, Q + S)$; (13) turns it into an $\Omega(\varepsilon)$ function gap. The paths in (11) give matching upper orders. Under S_∞, the direct sum is exactly g_ε neuron by neuron. Appendix D completes all uniformity and compactness details.

Restricted repairs. Freezing the first layer after summation permits quadratic cancellation only when $C_* = 0$; under S_2 this is precisely S_∞. A scalar reweighting of the two complete task vectors permits a repair only if some $s, t \in \mathbb{R}$ satisfy $sP + tR = C_*$ and $sQ + tS = 0$. Thus linearly independent Q, S exclude every nonzero required correction of this type. This compares repairs of the same target; changing the target by discarding a task is a different objective.

4 A QUANTITATIVE CERTIFICATE AND AN INEXPENSIVE CORRECTION

Compatibility is structural: even rank-one $P = up^\top$ and $R = tr^\top$ violate S_2 when p, r and every entry of u, t are nonzero. To treat approximately compatible updates, we separate input overlap from readout mismatch quantitatively. Set $B_i = q_i r_i^\top + s_i p_i^\top$ and $\Pi_i^\perp = I - v_i v_i^\top / \|v_i\|^2$, and define

$$\kappa_2^2 = \sum_i \left[\frac{\|v_i\|^2}{2} \big(\|p_i\|^2 \|r_i\|^2 + (p_i^\top r_i)^2 \big) + \big\| \Pi_i^\perp B_i \big\|_F^2 \right]. \tag{15}$$

Then $\kappa_2 = 0$ exactly when S_2 holds. Orthogonality $p_i^\top r_i = 0$ alone does not remove the first term.

Let $\Psi(x)$ concatenate $\sigma(z_i)$, $x_j \sigma'(z_i)$, and $\operatorname{svec}(xx^\top)\sigma''(z_i)$ over i, with $\sqrt{2}$ scaling on off-diagonal coordinates. Write $G = \mathbb{E}[\Psi\Psi^\top]$. It is positive definite by Lemma 2.

Theorem 4 (Obstruction bounds)**.** *Under Assumption 1,*

$$\lambda_{\min}(G)\kappa_2^2 \leq d_2^2 \leq \lambda_{\max}(G)\kappa_2^2. \tag{16}$$

The formula (10)*, used even outside* S_2*, minimizes the Euclidean coefficient residual in this feature basis. Its function residual is at most* $\sqrt{\lambda_{\max}(G)}\kappa_2$*, hence within* $\sqrt{\operatorname{cond}(G)}$ *of the optimal population second-order residual* d_2 *when* $d_2 > 0$*.*

Coefficient projection and Gram norm comparison prove the bounds (Appendix E). The correction minimizes coefficient norm; population least squares uses the G-weighted norm. The following extension makes the approximation guarantee uniform at finite scales.

Write E_{full} for (5) without rank restrictions, and E_{cat} for budgets $R_W = \min\{m, d, \operatorname{rank} P + \operatorname{rank} R\}$ and $R_V = r_V$. Let F_ε be the error of (10) applied as an ε^2 correction.

Theorem 5 (Uniform near-compatibility criterion)**.** *Fix a base and input law satisfying Assumption 1, and* $B > 0$*. There are* $\eta, \varepsilon_0, K > 0$*, independent of directions with* $\|a\| + \|b\| \leq B$*, such that, for* $0 < \varepsilon \leq \varepsilon_0$*,*

$$\begin{aligned} \varepsilon^2 \sqrt{\lambda_{\min}(G)}\, \kappa_2 - K\varepsilon^3 \leq E_{\text{full}}(\varepsilon) \leq E_{\text{cat}}(\varepsilon) \leq F_\varepsilon, \\ F_\varepsilon \leq \varepsilon^2 \sqrt{\lambda_{\max}(G)}\, \kappa_2 + K\varepsilon^3. \end{aligned} \tag{17}$$

Consequently, for any bounded family of directions $(a_\varepsilon, b_\varepsilon)$*,*

$$E_{\text{full}}(\varepsilon) = O(\varepsilon^3) \iff \kappa_2(a_\varepsilon, b_\varepsilon) = O(\varepsilon) \iff F_\varepsilon = O(\varepsilon^3). \tag{18}$$

Uniform Taylor control gives the upper bound; local parameter control forces every optimizing candidate to have a bounded second-order coefficient, giving the lower bound. Appendix E.2 supplies the proof and explicit conditioning-dependent radii and constants. The bound applies pointwise to changing directions; Table 1 classifies *fixed* directions. Near-duplicate neurons make G ill-conditioned, so small κ_2 must be interpreted together with this geometry. Under exact S_2, Appendix E.3 also quantifies the cubic gap and gives a rank-preserving third-order coefficient refinement.

For LoRA products $P = B_a A_a$ and $R = B_b A_b$, the corrected first-layer update is

$$\varepsilon(I - \varepsilon M)B_a A_a + \varepsilon(I - \varepsilon\Lambda)B_b A_b. \tag{19}$$

The repair changes only left-factor row scalings and fits within the original concatenated rank budget. Under S_2 it also preserves the rank of the actual sum, which can be smaller than the concatenated budget. Outside S_2, cancellation in the original sum may be lost. Computing the scalars from effective readout columns costs $O(km)$; row-scaling the two left factors costs $O(m(r_a + r_b))$. Recovering the required readout columns from low-rank factors has its own matrix-product cost. The rule requires no Hessian construction, calibration fit, or additional training. Its guarantee is (17); task-loss implications are separated in Section 6.2.

5 Exact functional addition has a different width budget

The exact width minimum gives a global counterpart to the local classification. For each original neuron define

$$k_i = \begin{cases} 1, & p_i = r_i = 0, \\ 1 + \mathbf{1}\{s_i \neq 0\}, & p_i \neq 0,\ r_i = 0, \\ 1 + \mathbf{1}\{q_i \neq 0\}, & p_i = 0,\ r_i \neq 0, \\ 2, & p_i = r_i \neq 0, \\ 3, & p_i, r_i \neq 0,\ p_i \neq r_i. \end{cases} \tag{20}$$

Theorem 6 (Minimum exact width)**.** *For all sufficiently small nonzero ε, the minimum hidden width of any single-hidden-layer exact-GELU network representing g_ε is $\sum_i k_i$. Competitors may have arbitrary weights and biases and need not be near the base. The lower bound remains valid even with a free affine output branch.*

For the upper bound, compile the three terms

$$(v_i + \varepsilon q_i)\sigma(z_i + \varepsilon p_i^\top x) + (v_i + \varepsilon s_i)\sigma(z_i + \varepsilon r_i^\top x) - v_i\sigma(z_i) \tag{21}$$

and merge equal ridges. For minimality, use $\sigma(-z) = \sigma(z) - z$ to group any competing representation by affine parameters modulo sign, absorbing sign changes into an affine branch. The zero-order case of Lemma 2 requires a nonzero coefficient in every ridge class of the target. Small ε precludes collisions between distinct base classes. Each remaining class needs a neuron, giving (20) (Appendix F).

Corollary 7 (Global same-width rigidity)**.** *For every sufficiently small nonzero ε, g_ε admits a width-at-most-m exact-GELU representation if and only if S_∞ holds. The equivalence remains true with arbitrary competing weights, biases, and an affine output branch. More generally, its exact width excess is $\sum_i (k_i - 1)$.*

Proof. Each $k_i \geq 1$, and (20) gives $k_i = 1$ for every i exactly under S_∞. Apply Theorem 6; under S_∞ direct summation already realizes the target. □

Thus distant parameters cannot restore *exact* same-width representability when S_∞ fails. Take $d \geq 2$, $P = up^\top$ and $R = tr^\top$, where every entry of u, t is nonzero and p, r are independent. Both updates have rank one, but every $k_i = 3$: the target needs width $3m$, and its optimal local same-width error is $\Theta(\varepsilon^2)$ once the first-order rank budget is adequate. The readout directions can also have rank one. This compares a local approximation obstruction with a global exact-width requirement on the *same* instances; finite-tolerance width approximation is a separate question.

6 CALIBRATION AND TASK-LOSS IMPLICATIONS

6.1 INTERPOLATION IS NOT A POPULATION CERTIFICATE

Theorem 8 (A calibration interpolation threshold)**.** *For scalar output under Assumption 1, fix the base and tasks, then draw calibration inputs independently from μ. Let $p_{\mathrm{par}} = m(d+1)$. If $N \le p_{\mathrm{par}}$, the N-sample Jacobian has full row rank almost surely, so any quadratic interaction is exactly interpolatable on those inputs. For sufficiently small ε, an unrestricted nearby parameter also fits g_ε exactly on the sample. If $N \ge p_{\mathrm{par}} + 1$ and $d_2 > 0$, the quadratic interaction is almost surely not interpolatable.*

Analytic evaluation determinants and the implicit function theorem prove this unrestricted, exact-arithmetic interpolation threshold (Appendix G); statistical certification requires additional control of conditioning.

A reliable approximate certificate needs stronger assumptions. If the fixed feature basis above has bounded leverage $\Psi(x)^\top G^{-1}\Psi(x) \le L$ almost surely, a matrix Chernoff bound (Tropp, 2012) implies that

$$N \ge 3L\delta^{-2}\log(2D/\zeta), \qquad D = m\{1 + d + d(d+1)/2\}, \tag{22}$$

suffices for $(1-\delta)d_2^2 \le \widehat{d}_2^2 \le (1+\delta)d_2^2$ with probability at least $1-\zeta$, simultaneously over task directions for this fixed base. Appendix H proves the statement. Bounded leverage is additional; it does not follow from the moment assumption and is not asserted for our unbounded Gaussian experiment.

6.2 FUNCTIONAL FIDELITY IS NOT TASK PRESERVATION

Let $r = f_{\mathrm{fix}} - g_\varepsilon$. On task a,

$$f_{\mathrm{fix}} - f_{\theta_a} = (f_{\theta_b} - f_{\theta_0}) + r. \tag{23}$$

The first term is cross-task functional leakage and is unaffected by perfect repair. For an output loss with β-Lipschitz gradient, define $e_a = \|f_{\theta_b} - f_{\theta_0}\|_{L^2(\mathcal{D}_a)}$, $d_a = \|r\|_{L^2(\mathcal{D}_a)}$, and $G_a = \|\mathbb{E}[\nabla_z \ell(f_{\theta_a}(x), y) \mid x]\|_{L^2(\mathcal{D}_a)}$. Then

$$\mathcal{L}_a(f_{\mathrm{fix}}) - \mathcal{L}_a(f_{\theta_a}) \le G_a(e_a + d_a) + \tfrac{\beta}{2}(e_a + d_a)^2. \tag{24}$$

Appendix I proves this bound and gives a zero-interference example with degraded loss on both tasks. A mixture calibration law controls d_a via its mixture weight, but it does not control e_a.

7 CONTROLLED NUMERICAL VALIDATION

The synthetic experiments isolate the mechanisms in the theorems. All use exact GELU and float64 on CPU, with fixed seeds and recorded parameters. Gaussian norms use tensor Gauss–Hermite quadrature of order 80 per dimension; increasing to 96 changes the 80 primary quadratic residual estimates by at most 5.64×10^{-15}. Full protocols and the scope of numerical checks are in Appendix K.

Error orders and a cubic floor. We generate 16 nonorthogonal, biased $d = m = k = 2$ networks per case, covering overlapping updates, orthogonal overlap, readout mismatch, repairable nonexact updates, and exact additivity. At seven scales $\varepsilon = 0.2/2^j$, we compare summation, (10), and an unrestricted population-Jacobian coefficient projection. For repairable cases we additionally project the explicit cubic residual and add its optimal third-order coefficient. These baselines optimize asymptotic coefficients; Theorem 1 establishes nonlinear local optimality. An SVD-truncated first-order path illustrates rank insufficiency. Slopes are fitted on the smallest five scales, with all seeds retained. Table 2 and Figure 1 recover first-, second-, and third-order behavior. Orthogonal input directions remain quadratically obstructed. The explicit repair has slope 3.0002 ± 0.0002; even the full-J cubic projection has slope 2.9942 ± 0.0017. Exact-additivity cases have maximum summation error 2.11×10^{-16}, consistent with floating-point roundoff rather than a fitted convergence order.

Table 2: Observed error slopes: mean ± sample standard deviation over 16 seeds. Pathwise predictions match the error orders in Theorem 1.

Setting / method	Predicted order	Fitted slope
Insufficient rank / truncated path	1	1.0078 ± 0.0007
Input overlap / full-J projection	2	1.9990 ± 0.0014
Orthogonal overlap / full-J projection	2	1.9961 ± 0.0007
Readout mismatch / full-J projection	2	1.9968 ± 0.0022
Repairable / direct sum	2	2.0000 ± 0.0001
Repairable / analytic correction	3	3.0002 ± 0.0002
Repairable / cubic projection	3	2.9942 ± 0.0017

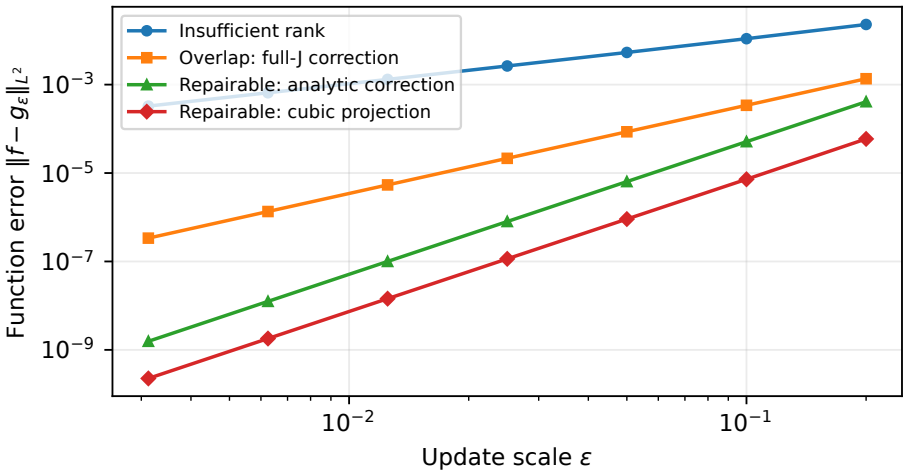


Figure 1: Median population-function errors across 16 seeds. Cubic projection reduces the third-order constant but does not remove its order. Different lines need not share the same instance category.

Capacity, conditioning, and calibration. In 16 networks with $(d, m, k) = (6, 8, 5)$, four rank-one task matrices produce rank-two sums, and the analytic repair preserves both ranks at three scales: 48 checks. A separate 80-instance compilation test covers all five rows of (20); predicted and constructed widths agree, with maximum discrepancy 4.30×10^{-15}. This checks the exact-width compiler. Across 96 rank-one direction pairs on a fixed base, every Gram bound is satisfied numerically. A near-collision sweep keeps κ_2 fixed while the interaction norm shrinks, confirming the need for conditioning constants (Appendix K).

For scalar output with six parameters, 40 independent nested calibration samples per size yield median training residual 6.31×10^{-17} at $N = 6$, but population evaluation residual 0.05442; the population coefficient optimum is 0.02113. At $N = 7$, the median training residual is already positive (0.00247). At $N = 256$, the evaluation median approaches 0.02160. Figure 2 shows the full transition. The near-zero calibration residual coexists with a strictly positive population obstruction.

Approximate compatibility. A controlled readout perturbation of the repairable instances compares a fixed incompatibility with one decreasing proportionally to ε. The resulting second- and third-order behaviors test (18); additional coefficient checks cover the cubic refinement and the distinction between actual-sum and concatenated ranks (Appendix K.6).

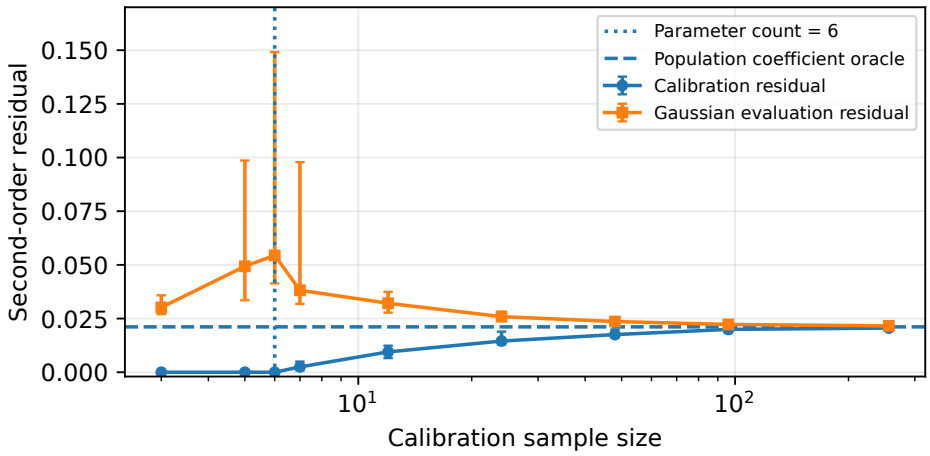


Figure 2: Calibration interpolation can hide a population obstruction. Medians over 40 independent draws; error bars span the 25th–75th percentiles, not confidence intervals. The $N = 6$ parameter threshold gives a nearly zero training residual and a larger population evaluation residual. The dashed horizontal line is the full-Jacobian population coefficient optimum.

8 Limitations and Conclusion

The classification concerns exact GELU, fixed hidden biases, a nondegenerate base, and an open-support population law in a single-hidden-layer architecture. Extending it to trainable biases, other activations, deep compositions, or lower-dimensional supports requires additional analysis. The finite-scale constants and admissible radius are explicit but conditioning-dependent, not dimension-free or uniform near neuron collisions. The global width theorem concerns exact representation, whereas the error-order theorem and near-compatibility criterion concern local approximation. The synthetic data establish implementation consistency rather than downstream benchmark superiority.

Within this scope the resource distinction is complete: rank gates first-order feasibility, neuron structure determines the quadratic and cubic obstructions, and exact width counts the missing ridge classes. Exact additivity is also the criterion for global same-width representability. Near compatibility has a quantitative interpretation: an obstruction of size $O(\varepsilon)$ is necessary and sufficient for third-order local accuracy, and a correction within the concatenated adapter budget achieves it. These results separate an expressivity failure from inadequate weight tuning without equating functional fidelity with task retention.

Reproducibility. Appendices provide full proofs, constants, data-generation rules, and evaluation conventions. The supplement includes executable code, all primary network parameters, raw CSVs, figure scripts, a fixed-seed manifest, and 45 passing unit and numerical tests. Symbolic expansion and automatic differentiation provide separate consistency checks of the implemented formulas.

AI USE STATEMENT

ChatGPT (GPT-6 Astra Pro) was used substantially in formulating the theoretical claims, developing and auditing proof arguments, designing the synthetic experiments, writing and debugging executable code, checking the literature, and drafting and formatting this manuscript. All numerical results reported here were produced by executing the included code. The mathematical arguments and executable checks are supplied for scrutiny. These AI-assisted audits are not independent peer review or formal theorem-prover verification.

ETHICS STATEMENT

This work uses analytically specified synthetic networks and generated inputs, without human-subject data, personal information, or external model checkpoints. Functional-addition fidelity is not a safety or fairness guarantee, and a repaired model need not retain either task's performance. The paper explicitly separates mathematical capacity statements from deployment claims.

REFERENCES

Vincent Bürgin, Daniel Herbst, Ya-Wei Eileen Lin, and Stefanie Jegelka. Beyond Structural Symmetries: Linear Mode Connectivity via Neuron Identifiability. *International Conference on Machine Learning*, 2026.

Marvin F. da Silva, Mohammed Adnan, Felix Dangel, and Sageev Oore. Generalizing the Geometry of Model Merging Through Fréchet Averages. *arXiv preprint arXiv:2604.27155*, 2026.

Bethan Evans, Benjamin Etheridge, Stephen Roberts, and Jared Tanner. Model Merging by Output-Space Projection. *arXiv preprint arXiv:2605.29101*, 2026.

Dan Hendrycks and Kevin Gimpel. Gaussian Error Linear Units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. *International Conference on Learning Representations*, 2022.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing Models with Task Arithmetic. *International Conference on Learning Representations*, 2023.

Hyo Seo Kim and Ren Wang. HyperFix: Combinatorial Nonlinear Correction for Task Vector Merging. *arXiv preprint arXiv:2608.11499*, 2026.

Antonij Mijoski and Marko Karbevski. Can an MLP Absorb Its Own Skip Connection?. *arXiv preprint arXiv:2604.23705*, 2026.

Guillermo Ortiz-Jiménez, Alessandro Favero, and Pascal Frossard. Task Arithmetic in the Tangent Space: Improved Editing of Pre-Trained Models. *Advances in Neural Information Processing Systems*, 2023.

Aniello Panariello, Daniel Marczak, Simone Magistri, Angelo Porrello, Bartłomiej Twardowski, Andrew D. Bagdanov, Simone Calderara, and Joost van de Weijer. Accurate and Efficient Low-Rank Model Merging in Core Space. *Advances in Neural Information Processing Systems*, 2025.

Joel A. Tropp. User-Friendly Tail Bounds for Sums of Random Matrices. *Foundations of Computational Mathematics, 12:389–434*, 2012.

Verner Vlačić and Helmut Bölcskei. Affine Symmetries and Neural Network Identifiability. *Advances in Mathematics, 376:107485*, 2021.

Tiantong Wang, Yiyang Duan, Haoyu Chen, Tiantong Wu, and Wei Yang Bryan Lim. M-Loss: Quantifying Model Merging Compatibility with Limited Unlabeled Data. *arXiv preprint arXiv:2602.08564*, 2026.

A NOTATION AND PROOF MAP

The parameter is the pair of effective matrices $\theta = (W, V)$, with norm $\|\theta\|^2 = \|W\|_F^2 + \|V\|_F^2$. The bias vector is part of the fixed architecture, not this parameter vector. A row of a first-layer direction is denoted by a column vector p_i through $P_{i:} = p_i^\top$; a second-layer column is $q_i = Q_{:i}$. All $O_{\mathcal{H}}(\varepsilon^j)$ statements use the Euclidean output norm inside $L^2(\mu)$, not a mean over output coordinates. In the local classification the base, input law, task directions, width, and rank budgets are fixed while $\varepsilon \downarrow 0$. Theorem 5 explicitly permits bounded scale-dependent directions at a fixed base and law.

The main classification uses only three ingredients: analytic independence (Appendix B), uniform Taylor estimates and local injectivity (Appendix C), and an explicit cubic obstruction (Appendix D). The quantitative bounds and width theorem are proved independently in Appendices E and F. Calibration interpolation and stable certification have distinct assumptions, treated separately in Appendices G and H. Numerical evidence is described in Appendix K; none of the mathematical proofs uses an empirical rank computation.

B PROOF OF THE GELU INDEPENDENCE LEMMA

We give full details for scalar coefficients; vector coefficients follow by applying the scalar statement to each output coordinate.

From a population identity to an analytic identity. Every term in (6) is real analytic in x. If the identity holds in $\mathcal{H}$, it holds μ-almost everywhere. On an open set where the density is positive almost everywhere, it therefore holds Lebesgue-almost everywhere. Continuity makes it hold at every point of that open set: a nonzero value would persist on a neighborhood of positive measure. The identity theorem for real analytic functions on the connected domain $\mathbb{R}^d$ extends the equality to all x.

Generic lines separate affine forms up to sign. Let $\alpha_i = w_i^\top u$. We choose u outside finitely many proper hyperplanes so that $\alpha_i \neq 0$ and the pairs $(\alpha_i^2, \alpha_i\beta_i)$ are distinct. For a pair with $w_i \neq \pm w_j$, avoid $(w_i - w_j)^\top u = 0$ and $(w_i + w_j)^\top u = 0$, making $\alpha_i^2 \neq \alpha_j^2$. If $w_j = w_i$, nondegeneracy gives $\beta_j \neq \beta_i$, so $\alpha_i\beta_i \neq \alpha_j\beta_j$. If $w_j = -w_i$, it gives $\beta_j \neq -\beta_i$, with the same conclusion. These choices form a dense open set of directions. The argument also covers $d = 1$: if the squared slopes coincide, the bias term separates the exponent pairs unless the excluded sign relation holds.

Reduction to Gaussian exponential polynomials. Restrict $x = tu$, abbreviate $A_i(tu)$ and $B_i(tu)$ as $A_i(t), B_i(t)$, and write $z_i = \alpha_i t + \beta_i$. Since

$$\sigma(z) = z\Phi(z), \quad \sigma'(z) = \Phi(z) + z\phi(z), \quad \sigma''(z) = (2 - z^2)\phi(z),$$

the contribution of neuron i is

$$F_i(t) = L_i(t)\Phi(z_i) + T_i(t)\phi(z_i), \tag{25}$$

$$L_i(t) = B_i(t) + c_i z_i, \qquad T_i(t) = A_i(t)(2 - z_i^2) + B_i(t)z_i. \tag{26}$$

Here $\deg L_i \leq 1$ and $\deg T_i \leq 4$. Twice differentiating removes the affine function $\ell(tu)$. In addition, the coefficient of Φ in F_i'' is $L_i'' = 0$. Hence

$$\sum_i P_i(t)\exp(-\alpha_i^2 t^2/2 - \alpha_i\beta_i t) = 0 \tag{27}$$

for some polynomials P_i, with the constant factors $\exp(-\beta_i^2/2)/\sqrt{2\pi}$ absorbed into them.

For completeness, the functions in (27) are linearly independent. Choose the smallest quadratic exponent coefficient $\alpha_i^2/2$, and among ties the smallest linear coefficient $\alpha_i\beta_i$. Divide by its exponential. Every other summand tends to zero as $t \to +\infty$, even after multiplication by its polynomial. The selected polynomial tends to zero and therefore is the zero polynomial. Remove that summand and repeat. Thus every $P_i = 0$, and in particular $F_i'' = 0$ for every i.

Each F_i is now affine. As $z_i \to -\infty$, both the polynomial multiple of ϕ and the polynomial multiple of Φ in (25) tend to zero. An affine function tending to zero at an infinite endpoint is identically zero.

As $z_i \to +\infty$, $\Phi(z_i) \to 1$ and $T_i\phi(z_i) \to 0$, so L_i tends to zero and is itself the zero polynomial. Hence $B_i = -c_i z_i$. Since $\phi > 0$, $T_i = 0$ gives

$$A_i(t)(2 - z_i^2) - c_i z_i^2 = 0. \tag{28}$$

As $\alpha_i \neq 0$, there is a real t with $z_i = \sqrt{2}$. Substitution implies $c_i = 0$; then $A_i(t)(2 - z_i^2) = 0$ as a polynomial, implying $A_i = 0$, and also $B_i = 0$. This holds along every generic line. A polynomial vanishing on those lines vanishes on a dense set and therefore everywhere. Finally $\ell = 0$. This proves Lemma 2. □

Why the exclusions matter. If two affine forms coincide, their output coefficients can cancel. If they are opposite, $\sigma(-z) = \sigma(z) - z$ produces an affine relation. If $w_i = 0$, a ridge is constant. Excluding these cases is essential for this independence statement. In particular the lemma is not an unconditional identifiability theorem for every GELU parameterization.

C Regularity, Taylor estimates, and local parameter control

Lemma 9 (Regularity and a uniform local lower bound)**.** *Under Assumption 1, the parameter-to-function map is C^4 into $\mathcal{H}$ on a neighborhood of θ_0. Its Jacobian at θ_0 is injective. There are $\eta, \alpha > 0$ such that* (13) *holds for all $\theta, \theta' \in \overline{B}_\eta(\theta_0)$.*

Proof. The derivatives $\sigma', \sigma'', \sigma''', \sigma^{(4)}$ are bounded; derivatives of order at least two are polynomials times ϕ. Also $|\sigma(z)| \leq |z|$. On a bounded parameter neighborhood, parameter derivatives through order four are bounded in norm by a constant times $1 + \|x\|^4$. Their squares are integrable by the eighth-moment assumption. Differentiation under the integral and dominated convergence establish existence and continuity of the Fréchet derivatives through order four.

If $J(C, D) = 0$, apply Lemma 2 to (8), with $A_i = 0$. It gives $d_i = 0$ and $v_i c_i^\top x = 0$ in each output coordinate. Since $v_i \neq 0$, $c_i = 0$. Thus J is injective. The parameter unit sphere is compact, so

$$\alpha := \min_{\|u\|=1} \|Ju\|_{\mathcal{H}} > 0.$$

Choose a convex closed ball on which $\|Df_\vartheta - J\|_{\mathrm{op}} \leq \alpha/2$. The fundamental theorem of calculus gives

$$f_\theta - f_{\theta'} = J(\theta - \theta') + \int_0^1 \big(Df_{\theta' + t(\theta - \theta')} - J\big)(\theta - \theta')\, dt.$$

The reverse triangle inequality proves (13). □

The mixed difference has the exact integral representation

$$\Delta_\varepsilon = \int_0^\varepsilon \int_0^\varepsilon D^2 f_{\theta_0 + sa + tb}[a, b]\, ds\, dt. \tag{29}$$

If the third derivative norm is at most M_3 on the relevant segments, then

$$\big\|\Delta_\varepsilon - \varepsilon^2 h\big\|_{\mathcal{H}} \leq \tfrac{M_3}{2}\varepsilon^3 \|a\| \|b\| (\|a\| + \|b\|). \tag{30}$$

Adding a bounded second-order correction $\varepsilon^2 c$ at $\theta_0 + \varepsilon(a+b)$ changes the output by $\varepsilon^2 Jc + O_{\mathcal{H}}(\varepsilon^3)$, uniformly on bounded sets of c. This proves (9). Fourth-order Taylor estimates are likewise uniform on bounded third-order correction coefficients. These are norm estimates in $\mathcal{H}$; no uniform bound over all inputs is asserted.

D Proof of the classification and the cubic gap

We first justify the cubic expression and then prove each line of Theorem 1 without assuming differentiable candidate paths.

D.1 THE SECOND-ORDER EQUATIONS

Substitution of (7) and (8) into Lemma 2 gives, for every neuron,

$$v_i(p_i^\top x)(r_i^\top x) = 0, \qquad q_i r_i^\top + s_i p_i^\top + v_i c_i^\top = 0, \qquad d_i = 0. \tag{31}$$

The polynomial ring over $\mathbb{R}$ has no zero divisors, so the first equation and $v_i \neq 0$ imply $p_i = 0$ or $r_i = 0$. If $r_i \neq 0$, a nonzero coordinate of r_i in the second equation implies that q_i is parallel to v_i; write $q_i = \lambda_i v_i$. The equation then forces $c_i = -\lambda_i r_i$. The case $p_i \neq 0$ is symmetric. If both are zero, $c_i = 0$. This supplies all necessity, sufficiency, and uniqueness statements of Theorem 3.

For a row with only p_i nonzero, set $\Gamma_{ii} = \mu_i$; for a row with only r_i nonzero set $\Gamma_{ii} = \lambda_i$; for a row with both zero take $\Gamma_{ii} = 0$. Then $C_* = -\Gamma(P + R)$ and (11) follows exactly, without a rank-manifold approximation or an additional completion term.

D.2 DERIVATION OF THE CUBIC RESIDUAL

Consider $p_i \neq 0$, $r_i = 0$, $s_i = \mu_i v_i$, and put $u = p_i^\top x$. Its target contribution is

$$(v_i + \varepsilon q_i)\sigma(z_i + \varepsilon u) + \varepsilon \mu_i v_i \sigma(z_i).$$

The repaired contribution is

$$(v_i + \varepsilon(q_i + \mu_i v_i))\sigma(z_i + \varepsilon u - \varepsilon^2 \mu_i u).$$

Expanding the activation,

$$\begin{aligned}\sigma(z_i + \varepsilon u - \varepsilon^2 \mu_i u) = \sigma(z_i) + \varepsilon\sigma'(z_i)u + \varepsilon^2\{-\mu_i\sigma'(z_i)u + \tfrac{1}{2}\sigma''(z_i)u^2\} \\ + \varepsilon^3\{-\mu_i\sigma''(z_i)u^2 + \tfrac{1}{6}\sigma'''(z_i)u^3\} + O_{\mathcal{H}}(\varepsilon^4)\end{aligned}$$

when multiplied by its bounded readout coefficient. Subtracting the target cancels orders zero, one, and two, and also the σ''' term. The remaining cubic coefficient is exactly (12). The symmetric expression for the other active-row case is

$$h_{3,i} = -\lambda_i(s_i + \lambda_i v_i)\sigma'(z_i)(r_i^\top x) - \tfrac{\lambda_i}{2} v_i \sigma''(z_i)(r_i^\top x)^2. \tag{32}$$

A neuron with no input change is affine in its readout parameters and contributes no residual. Therefore, for $\theta_2(\varepsilon) = \theta_0 + \varepsilon(a + b) + \varepsilon^2(C_*, 0)$,

$$f_{\theta_2(\varepsilon)} - g_\varepsilon = \varepsilon^3 h_3 + O_{\mathcal{H}}(\varepsilon^4). \tag{33}$$

If S_2 holds but S_∞ fails, at least one active p_i has $\mu_i \neq 0$, or one active r_i has $\lambda_i \neq 0$. Its quadratic coefficient in (12) or (32) is nonzero. Lemma 2 then gives $h_3 \notin \operatorname{range} J$. Since the range is finite dimensional and closed, $d_3 > 0$.

D.3 ARBITRARY RANK-CONSTRAINED LOCAL CANDIDATES

Choose the ball in (5) within that of Lemma 9. Let θ_ε be a minimizer. All explicit upper-bound paths below lie inside it for small ε.

Inadequate first-order budget. Define the nonnegative SVD tail

$$\tau^2 = \sum_{j > R_W} s_j(P + R)^2 + \sum_{j > R_V} s_j(Q + S)^2. \tag{34}$$

The standard best-low-rank approximation identity implies that every feasible θ is at least $\varepsilon\tau$ from $\theta_{\mathrm{sum}} = \theta_0 + \varepsilon(a + b)$. The latter lies in the small ball, though it need not be rank feasible. Equations (13) and (30) yield

$$\|f_\theta - g_\varepsilon\|_{\mathcal{H}} \geq \tfrac{\alpha}{2}\varepsilon\tau - O(\varepsilon^2).$$

If either budget is too small then $\tau > 0$. The feasible base θ_0 gives the matching $O(\varepsilon)$ upper bound. This proves the first row.

Adequate budget but no quadratic repair. The direct sum is feasible and has $O(\varepsilon^2)$ error, so the minimizing error is at most that order. Comparing f_{θ_ε} to $f_{\theta_{\mathrm{sum}}}$ and using (13) gives

$$\|\theta_\varepsilon - \theta_{\mathrm{sum}}\| = O(\varepsilon^2).$$

Thus write $\theta_\varepsilon = \theta_{\mathrm{sum}} + \varepsilon^2 c_\varepsilon$ with c_ε uniformly bounded. Equation (9), which does not require c_ε to converge, gives

$$\|f_{\theta_\varepsilon} - g_\varepsilon\|_{\mathcal{H}} \geq \varepsilon^2 \|h + Jc_\varepsilon\|_{\mathcal{H}} - K\varepsilon^3 \geq \varepsilon^2 d_2 - K\varepsilon^3.$$

By Theorem 3, $d_2 > 0$ precisely in this regime. This proves the second row.

Quadratic repair but no exact additivity. The path $\theta_2(\varepsilon)$ is rank feasible by (11) and has $O(\varepsilon^3)$ error. Hence the minimizing error has that upper order. Compare it with $f_{\theta_2(\varepsilon)}$ and apply (13): $\theta_\varepsilon = \theta_2(\varepsilon) + \varepsilon^3 e_\varepsilon$ with e_ε bounded. Continuity of the first derivative and (33) give

$$f_{\theta_\varepsilon} - g_\varepsilon = \varepsilon^3(h_3 + Je_\varepsilon) + O_{\mathcal{H}}(\varepsilon^4).$$

The resulting bound $\varepsilon^3 d_3 - K\varepsilon^4$ is positive to leading order. This proves the third row. In particular, no oscillatory, fractional-order, or merely continuous sequence of local parameters evades this lower bound.

Exact additivity. If $p_i \neq 0$, S_∞ says $r_i = s_i = 0$, so the target contribution is exactly $(v_i + \varepsilon q_i)\sigma(z_i + \varepsilon p_i^\top x)$, the direct-sum contribution. The symmetric case is identical. If $p_i = r_i = 0$, the contribution is $(v_i + \varepsilon(q_i + s_i))\sigma(z_i)$. Summing proves exact equality and the fourth row.

This completes Theorem 1. The unrestricted distances d_2 and d_3 are independent of the rank budgets and lower-bound every constrained problem. Enlarging a budget can reduce the constrained problem's leading error constant, but cannot change these positive unrestricted obstructions or improve the listed error order. □

D.4 Full-Jacobian coefficient optima

With no rank restriction, the unique second-order coefficient minimizing $\|h + Jc\|_{\mathcal{H}}$ exists because J is injective and has closed finite-dimensional range. Applying it as $\varepsilon^2 c$ gives

$$E_{\mathrm{full}}(\varepsilon) = \varepsilon^2 d_2 + O(\varepsilon^3) \quad \text{when } d_2 > 0.$$

Under S_2 but not S_∞, the analogous third-order coefficient gives $E_{\mathrm{full}}(\varepsilon) = \varepsilon^3 d_3 + O(\varepsilon^4)$. These facts justify the coefficient-oracle baselines in the experiments. They do not equate a finite-ε least-squares coefficient with an exact minimizer of the nonlinear objective.

E Proof of the quantitative certificate and repair bounds

E.1 Quadratic coefficient projection and Gram comparison

Use the orthonormal coordinate convention for a symmetric matrix A:

$$\mathrm{svec}(A) = (A_{11}, \sqrt{2}A_{12}, \ldots, A_{dd}), \qquad x^\top A x = \mathrm{svec}(A)^\top \mathrm{svec}(xx^\top).$$

The ordering of coordinates is arbitrary if used consistently. Let $T(C, D)$ denote the matrix of coefficients of $h + J(C, D)$ in the scalar feature vector Ψ. Its neuronwise blocks are

$$d_i, \qquad B_i + v_i c_i^\top, \qquad v_i\ \mathrm{svec}\left(\frac{p_i r_i^\top + r_i p_i^\top}{2}\right)^\top. \tag{35}$$

The squared population norm is $\mathrm{tr}(TGT^\top)$. Feature independence follows from Lemma 2; a nonzero coefficient vector cannot have zero L^2 norm, so $G \succ 0$. Its entries are finite by Assumption 1.

Minimizing $\|T\|_F^2$ separates across the blocks. Set $d_i = 0$ and project each column of $-B_i$ onto v_i, giving

$$c_i^\top = -\frac{v_i^\top B_i}{\|v_i\|^2} = -\lambda_i r_i^\top - \mu_i p_i^\top.$$

The remaining linear block is $\Pi_i^{\perp} B_i$. The identity

$$\left\|(p_i r_i^\top + r_i p_i^\top)/2\right\|_F^2 = \tfrac{1}{2}\{\|p_i\|^2 \|r_i\|^2 + (p_i^\top r_i)^2\}$$

gives $\min_{C,D} \|T(C,D)\|_F^2 = \kappa_2^2$. For every T,

$$\lambda_{\min}(G)\,\|T\|_F^2 \le \operatorname{tr}(TGT^\top) \le \lambda_{\max}(G)\,\|T\|_F^2 .$$

Taking minima proves (16), and evaluating the upper bound at the coefficient minimizer proves the light-repair guarantee. Since κ_2^2 is a sum of nonnegative terms, its zero set has no hidden cancellation: every input product must vanish and every residual readout block must vanish. This is exactly S_2. □

E.2 Uniform finite-scale bounds and near-compatibility

We prove Theorem 5 uniformly over $\|a\| + \|b\| \le B$. All parameter norms are the product Frobenius norm from Appendix A. Set $\nu = \min_i \|v_i\| > 0$. On a preliminary parameter ball of radius $\rho > 0$, let M_2, M_3 bound the second and third derivative operator norms. Lemma 9 gives a smaller closed ball of radius η and a constant $\alpha > 0$ satisfying (13). Define

$$L_c = \frac{B^2}{4\nu}, \qquad L_o = \frac{M_2 B^2}{\alpha}, \qquad K(L) = \frac{M_3 B^3}{8} + M_2 B L + \frac{M_2 L^2}{2}. \tag{36}$$

The correction $c_* = (C_*, 0)$ satisfies

$$\|c_*\| \le \nu^{-1}(\|Q\|_F \|R\|_F + \|S\|_F \|P\|_F) \le \nu^{-1} \|a\| \|b\| \le L_c.$$

Choose

$$\varepsilon_0 = \min\left\{1, \frac{\eta}{2B}, \sqrt{\frac{\eta}{2L_c}}\right\}, \qquad K = \max\{K(L_c), K(L_o)\}. \tag{37}$$

Then the task segments, direct sum, and corrected path stay in the chosen ball. For any $\|c\| \le L$ with its corrected endpoint in that ball, (30) and the integral estimate

$$\left\|f_{\theta_{\mathrm{sum}}+\varepsilon^2 c} - f_{\theta_{\mathrm{sum}}} - \varepsilon^2 Jc\right\|_{\mathcal{H}} \le M_2 \varepsilon^2 \|c\| \left(\varepsilon B + \tfrac{1}{2}\varepsilon^2 \|c\|\right)$$

give the uniform remainder bound

$$\left\|f_{\theta_{\mathrm{sum}}+\varepsilon^2 c} - g_\varepsilon - \varepsilon^2(h + Jc)\right\|_{\mathcal{H}} \le K(L)\varepsilon^3. \tag{38}$$

Applying this to c_* and using Theorem 4 proves the upper bound in (17). Its first-layer rank is at most $\operatorname{rank} P + \operatorname{rank} R$ by (19), and its readout rank is r_V. Thus $E_{\mathrm{full}} \le E_{\mathrm{cat}} \le F_\varepsilon$.

For the lower bound, take a minimizer θ_ε of E_{full} in the same closed ball. The direct sum is feasible and (29) yields

$$E_{\mathrm{full}}(\varepsilon) \le \|\Delta_\varepsilon\|_{\mathcal{H}} \le M_2 \|a\| \|b\| \varepsilon^2 \le M_2 B^2 \varepsilon^2/4.$$

Hence (13) implies

$$\|\theta_\varepsilon - \theta_{\mathrm{sum}}\| \le \frac{2}{\alpha}\big(E_{\mathrm{full}}(\varepsilon) + \|\Delta_\varepsilon\|_{\mathcal{H}}\big) \le L_o \varepsilon^2.$$

Write $\theta_\varepsilon = \theta_{\mathrm{sum}} + \varepsilon^2 c_\varepsilon$ and apply (38) with $L = L_o$. This gives

$$E_{\mathrm{full}}(\varepsilon) \ge \varepsilon^2 d_2 - K(L_o)\varepsilon^3 \ge \varepsilon^2 \sqrt{\lambda_{\min}(G)}\,\kappa_2 - K\varepsilon^3.$$

All constants are independent of the bounded task pair. Since G is fixed and positive definite, dividing this lower bound and the corresponding upper bound by ε^2 proves both directions of (18), even for discontinuous or oscillatory direction families. No differentiability of the directions as functions of ε is used.

Explicit radius from moments and conditioning. The constants above can be specified without a task-dependent neighborhood. Put $b_j = \sup_z |\sigma^{(j)}(z)|$, $\mathfrak{m}_j = (\mathbb{E}\,\|x\|^{2j})^{1/2}$, and $V_\rho = \|V\|_F + \rho$. Direct differentiation and Cauchy–Schwarz give valid choices

$$M_2 = b_2 V_\rho \mathfrak{m}_2 + 2 b_1 \mathfrak{m}_1, \qquad M_3 = b_3 V_\rho \mathfrak{m}_3 + 3 b_2 \mathfrak{m}_2. \tag{39}$$

Indeed, a term with j input-weight derivatives has norm bounded by $b_j \|V\|_F \|x\|^j$ times the product of the direction norms; the j terms with one readout derivative use $b_{j-1}\|x\|^{j-1}$. For the Jacobian coefficient matrix, (35) with $h = 0$ yields

$$\|J(C,D)\|_{\mathcal{H}}^2 \geq \lambda_{\min}(G)\left(\|D\|_F^2 + \sum_i \|v_i\|^2 \|c_i\|^2\right) \geq \lambda_{\min}(G)\min\{1,\nu\}^2 \|(C,D)\|^2 .$$

Thus one may take $\alpha = \sqrt{\lambda_{\min}(G)}\min\{1,\nu\}$ and $\eta = \min\{\rho, \alpha/(2M_2)\}$. The bound $\|Df_\vartheta - J\|_{\mathrm{op}} \leq M_2\eta$ then proves (13). These choices make (37) explicit in base geometry and input moments; they deteriorate as the feature Gram matrix becomes ill-conditioned.

Why the actual-sum rank guarantee is separate. Near-compatibility need not preserve the actual sum rank under the coefficient rule. Let $m = d = k = 2$, $V = I_2$, and choose any nondegenerate W, β. For nonzero scalars λ, μ, set

$$P_\delta = \begin{pmatrix} 1 & \delta \\ 0 & 0 \end{pmatrix}, \quad R_\delta = \begin{pmatrix} 0 & -\delta \\ 1 & 0 \end{pmatrix}, \quad Q = \mathrm{Diag}(0,\lambda), \quad S = \mathrm{Diag}(\mu, 0).$$

At $\delta = 0$ the directions satisfy S_2. For arbitrary δ, $\mathrm{rank}(P_\delta + R_\delta) = 1$, but (10) gives

$$\Delta W_{\mathrm{fix}} = \begin{pmatrix} \varepsilon - \mu\varepsilon^2 & -\mu\delta\varepsilon^2 \\ \varepsilon - \lambda\varepsilon^2 & 0 \end{pmatrix}, \qquad \det(\Delta W_{\mathrm{fix}}) = \mu\delta\varepsilon^3(1-\lambda\varepsilon).$$

Thus every nonzero δ gives rank two for small nonzero ε, although $\kappa_2 = O(|\delta|)$ as $\delta \to 0$. This is a limitation of actual-sum rank preservation by this rule, not a lower bound on all rank-one approximate repairs. The concatenated rank guarantee remains valid.

E.3 A CUBIC CERTIFICATE AND RANK-PRESERVING REFINEMENT

Under S_2, let $I_P = \{i : p_i \neq 0\}$ and $I_R = \{i : r_i \neq 0\}$. They are disjoint. Using the scalars in (10), define

$$\begin{aligned} \kappa_3^2 = &\sum_{i\in I_P} \mu_i^2 \left(\frac{\|v_i\|^2 \|p_i\|^4}{4} + \left\|\Pi_i^\perp q_i\right\|^2 \|p_i\|^2\right) \\ &+ \sum_{i\in I_R} \lambda_i^2 \left(\frac{\|v_i\|^2 \|r_i\|^4}{4} + \left\|\Pi_i^\perp s_i\right\|^2 \|r_i\|^2\right). \end{aligned} \tag{40}$$

Proposition 10 (Quantitative cubic obstruction)**.** *Under Assumption 1 and* S_2*,* $\kappa_3 = 0$ *if and only if* S_∞*, and*

$$\lambda_{\min}(G)\kappa_3^2 \leq d_3^2 \leq \lambda_{\max}(G)\kappa_3^2. \tag{41}$$

The unique Euclidean coefficient projection of $-h_3$ *onto the Jacobian coefficients is*

$$C_3 = (\Lambda + M)(\Lambda R + MP), \qquad D_3 = 0. \tag{42}$$

Adding $\varepsilon^3(C_3, 0)$ *after* (10) *preserves both actual summed-update ranks for sufficiently small* ε*. Its error is at most* $\varepsilon^3\sqrt{\lambda_{\max}(G)}\,\kappa_3 + O(\varepsilon^4)$*.*

Proof. For $i \in I_P$, the quadratic coefficient matrix of $h_{3,i}$ is $-(\mu_i/2)v_i\,\mathrm{svec}(p_i p_i^\top)^\top$. Its squared Frobenius norm is $\mu_i^2\|v_i\|^2\|p_i\|^4/4$. The linear coefficient is $-\mu_i(q_i + \mu_i v_i)p_i^\top$. Projection onto $v_i c_i^\top$ leaves $-\mu_i\Pi_i^\perp q_i p_i^\top$, with squared norm $\mu_i^2\left\|\Pi_i^\perp q_i\right\|^2\|p_i\|^2$, and gives $c_i = \mu_i(\lambda_i + \mu_i)p_i$. The I_R calculation swaps (p_i, q_i, μ_i) with (r_i, s_i, λ_i). Fixed-input neurons contribute zero. Summing proves that the minimum coefficient norm squared is (40), attained uniquely at (42). Gram norm comparison gives (41). Each active quadratic term is positive unless its cross-readout scalar vanishes, so the zero set is exactly S_∞.

Define $\Omega_{ii} = \mu_i(\lambda_i + \mu_i)$ on I_P, $\Omega_{ii} = \lambda_i(\lambda_i + \mu_i)$ on I_R, and zero otherwise. Then $C_3 = \Omega(P+R)$, and the refined update is

$$\Delta W_3 = \varepsilon(I - \varepsilon\Gamma + \varepsilon^2\Omega)(P+R), \qquad \Delta V_3 = \varepsilon(Q+S).$$

The diagonal multiplier is invertible for sufficiently small ε. Expanding (33) after this third-order correction gives $f_{\theta_2+\varepsilon^3(C_3,0)} - g_\varepsilon = \varepsilon^3(h_3 + J(C_3, 0)) + O_{\mathcal{H}}(\varepsilon^4)$, which proves the stated error bound. The local lower-bound argument in Appendix D also gives $\varepsilon^3\sqrt{\lambda_{\min}(G)}\kappa_3 - O(\varepsilon^4)$ for every adequate rank budget. The refinement controls the coefficient constant, while the nonzero error order remains cubic. □

E.4 Implementation and measurement bounds

A low-rank implementation of the diagnostic. The formula for $\|B_i\|_F^2$ requires only vector inner products:

$$\|B_i\|_F^2 = \|q_i\|^2 \|r_i\|^2 + \|s_i\|^2 \|p_i\|^2 + 2(q_i^\top s_i)(r_i^\top p_i),$$

and $\left\|\Pi_i^\perp B_i\right\|_F^2 = \|B_i\|_F^2 - \left\|v_i^\top B_i\right\|_2^2 / \|v_i\|^2$. Thus no $k \times d$ outer-product matrix need be stored per neuron. In floating-point code, tiny negative roundoff in these scalar differences should not be interpreted as a mathematical negative squared norm; our reference implementation forms the small blocks directly for the synthetic networks.

Proposition 11 (Sufficient repair for other smooth activations)**.** *For an arbitrary elementwise activation for which the parameter map is C^3 into $\mathcal{H}$, condition S_2 is sufficient for the repair* (10) *to give $O_{\mathcal{H}}(\varepsilon^3)$ error and preserve the summed-update ranks.*

Proof. The derivative identities (7)–(8) hold for any such activation. Disjoint row support kills the σ'' product, and readout alignment cancels the σ' term using (10). The rank identity (11) is purely algebraic. Necessity and the cubic gap, by contrast, use the GELU independence lemma. □

Approximate response measurements. Suppose $\left\|\widehat{h} - h\right\|_{\mathcal{H}} \le \eta_h$ and a bounded coefficient c satisfies $\left\|\widehat{h} + Jc\right\|_{\mathcal{H}} \le \nu$. Then the triangle inequality and (9) give

$$\|f_{\theta_{\mathrm{sum}}+\varepsilon^2 c} - g_\varepsilon\|_{\mathcal{H}} \le \varepsilon^2(\eta_h + \nu) + K\varepsilon^3. \tag{43}$$

A genuine third-order guarantee requires $\eta_h + \nu = O(\varepsilon)$. Estimating h by the four-output difference $\Delta_\varepsilon/\varepsilon^2$ magnifies an error of at most δ in each output to at most $4\delta/\varepsilon^2$. In addition to its $O(\varepsilon)$ truncation error, this estimator therefore needs $\delta = O(\varepsilon^3)$ to preserve third-order output accuracy. The experiments avoid this cancellation in response computation by using analytic derivatives, and check them independently by automatic differentiation.

F Proof of minimum exact width

We allow competing single-hidden-layer networks to use arbitrary affine ridge parameters, output coefficients, and an additional affine output function. This larger class makes the lower bound stronger. A neuron with zero input weight contributes only a constant and can be absorbed into that affine function.

For any nonzero input weight, group affine pairs by the equivalence relation $(w, \beta) \sim (-w, -\beta)$. Fix one orientation per class. Because

$$\sigma(-z) = \sigma(z) - z, \tag{44}$$

reorienting a term alters only its affine contribution, not the coefficient of the representative GELU ridge. Combine all coefficients within a class and discard classes with zero combined coefficient. If two such reduced expressions represent the same function, subtract them. Lemma 2 with $A_i = B_i = 0$ implies that the combined coefficients of every distinct nonzero ridge class agree. Therefore a target with K nonzero classes needs at least K neurons in any expression, even with the affine branch.

Now consider (21). The base has finitely many distinct classes with nonzero input weights, so for some $\varepsilon_0 > 0$, all its perturbed ridges with $0 < |\varepsilon| < \varepsilon_0$ remain nonzero and no class belonging to different base neurons collides. The same-neuron perturbed ridges cannot become opposite in that interval: their sum tends to $(2w_i, 2\beta_i)$, whose input component is nonzero. Equality of two shifted ridges within a neuron is then exactly equality of their input directions.

If both directions vanish, the combined coefficient is $v_i + \varepsilon(q_i + s_i)$, nonzero for sufficiently small ε. If only p_i is nonzero, the shifted ridge has coefficient $v_i + \varepsilon q_i$, and the base ridge has coefficient εs_i. This gives $1 + \mathbf{1}\{s_i \neq 0\}$ classes; the other one-active case is symmetric. If $p_i = r_i \neq 0$, the shifted coefficient is $2v_i + \varepsilon(q_i + s_i)$ and the base coefficient is $-v_i$, giving two classes. If both are nonzero and unequal, the three coefficients $v_i + \varepsilon q_i$, $v_i + \varepsilon s_i$, and $-v_i$ are all nonzero. This proves the class count (20).

The direct compilation with matching orientations realizes these classes using exactly $\sum_i k_i$ neurons and no affine branch. The lower bound just proved matches that construction. This completes Theorem 6. □

Corollary 12 (Rank-one directions with maximal width growth)**.** *For any nondegenerate base with $d \geq 2$, choose independent $p, r \in \mathbb{R}^d$ and vectors $u, t \in \mathbb{R}^m$ with no zero entries. The rank-one updates $P = up^\top$, $R = tr^\top$ produce a target requiring $3m$ neurons for all sufficiently small nonzero ε, regardless of the chosen fixed readout directions Q, S. In particular both readout directions may have rank one.*

Proof. Every $p_i = u_i p$ and $r_i = t_i r$ is nonzero and unequal by linear independence of p, r. Thus every $k_i = 3$. The readout coefficients in (21) remain nonzero for small ε because $v_i \neq 0$. □

An exact multi-task corollary. The ridge-class count extends directly to exact addition of more than two tasks. For T tasks, the target $\sum_{t=1}^{T} f_{\theta_0 + \varepsilon a_t} - (T-1) f_{\theta_0}$ has $T + 1$ classes per neuron when its T nonzero first-layer shifts are pairwise distinct and $T \geq 2$. It therefore needs exactly $(T+1)m$ neurons for small nonzero ε. Rank-one first-layer matrices can realize these distinct shifts. A multi-task local error classification additionally requires analyzing higher-order interactions.

G Calibration interpolation: proof and interpretation

We use a general analytic evaluation fact. If $\psi_1, \ldots, \psi_p$ are linearly independent real analytic functions on $\mathbb{R}^d$, the span of their evaluation vectors $[\psi_1(x), \ldots, \psi_p(x)]$ over all x is $\mathbb{R}^p$. Otherwise a nonzero vector orthogonal to that span would give a linear dependence. Hence there exist p points whose evaluation matrix has nonzero determinant. This determinant, viewed as a function of the p input points, is a nonzero real analytic function, whose zero set has Lebesgue measure zero. Independent samples from an absolutely continuous law therefore yield full rank almost surely. For $N < p$, an appropriate $N \times N$ minor gives the full-row-rank version of the same statement.

For scalar output, the $p_{\mathrm{par}} = md + m$ Jacobian functions are linearly independent by injectivity of J. Thus $N \leq p_{\mathrm{par}}$ gives full row rank and every vector of values $-h(x_j)$ is interpolatable. For a fixed nonrepairable h, the $p_{\mathrm{par}} + 1$ functions comprising the Jacobian coordinates and h are linearly independent. The same evaluation argument gives augmented rank $p_{\mathrm{par}} + 1$ for $N \geq p_{\mathrm{par}} + 1$, while the Jacobian rank is p_{par}. Consequently no exact quadratic fit exists. The fixed-before-sampling requirement avoids conflating this statement with an adversarial choice of h based on the sample.

Exact nonlinear sample fitting. For a sample with $N \leq p_{\mathrm{par}}$ and full row rank, choose N parameter coordinates whose evaluation derivative is invertible. Let $E : \mathbb{R}^N \to \mathbb{R}^{p_{\mathrm{par}}}$ embed those coordinates, and define

$$F(\varepsilon, u) = \left(f_{\theta_{\mathrm{sum}}(\varepsilon) + Eu}(x_j) - g_\varepsilon(x_j)\right)_{j=1}^{N}.$$

At $(\varepsilon, u) = (0, 0)$, $F = 0$ and $D_u F$ is invertible. The implicit function theorem supplies a local smooth $u(\varepsilon)$ with $F(\varepsilon, u(\varepsilon)) = 0$. First-order functional changes in the target and the direct sum agree, so $\partial_\varepsilon F(0, 0) = 0$ and $u'(0) = 0$. Hence $u(\varepsilon) = O(\varepsilon^2)$. This proves local exact sample fitting even when the population has a positive quadratic floor. The constants and allowed scale depend on the sample; no uniform conditioning claim is made, and this construction need not satisfy a low-rank budget. □

Fixed linear correction subspaces. The same quadratic interpolation argument applies to any fixed linear correction subspace of dimension s on which J is injective, with s replacing p_{par}. A freely varying matrix-rank constraint is not a linear subspace, so this substitution must not be used indiscriminately for low-rank adapters.

H A STABLE FINITE-SAMPLE POPULATION CERTIFICATE

The zero-error interpolation theorem is algebraic. The following statistical statement addresses approximate errors under an additional distributional condition.

Proposition 13 (Uniform empirical norm comparison)**.** *Let $G = \mathbb{E}\Psi\Psi^\top \succ 0$ be the D-dimensional feature Gram matrix of a fixed base. Suppose $\Psi(x)^\top G^{-1}\Psi(x) \le L$ almost surely. For independent samples, $0 < \delta < 1$, and $0 < \zeta < 1$, condition* (22) *implies, with probability at least $1 - \zeta$,*

$$(1-\delta)G \preceq \widehat{G} \preceq (1+\delta)G. \tag{45}$$

On this event, $(1-\delta)d_2^2 \le \widehat{d}_2^2 \le (1+\delta)d_2^2$ simultaneously for every pair of task directions, where $\widehat{d}_2$ minimizes the empirical L^2 coefficient residual over all (C, D).

Proof. Define $Y_j = G^{-1/2}\Psi(x_j)\Psi(x_j)^\top G^{-1/2}$. Then $Y_j \succeq 0$, $\mathbb{E}Y_j = I$, and $\|Y_j\|_{\mathrm{op}} \le L$. The upper and lower matrix Chernoff bounds (Tropp, 2012), using $0 < \delta < 1$, give

$$\Pr\left(\left\|N^{-1}\sum_j Y_j - I\right\|_{\mathrm{op}} > \delta\right) \le 2D\exp\{-N\delta^2/(3L)\}.$$

The specified N makes this at most ζ. Congruence by $G^{1/2}$ yields (45). On this event, for every coefficient matrix T,

$$(1-\delta)\operatorname{tr}(TGT^\top) \le \operatorname{tr}(T\widehat{G}T^\top) \le (1+\delta)\operatorname{tr}(TGT^\top).$$

Apply this to the coefficient matrices (35) and minimize over (C, D). The event is independent of the choice of coefficients and task directions, proving simultaneity. □

The leverage bound may be large and obeys $L \ge D$ by taking expectations. Compact input support and a positive definite fixed G suffice for a finite L, but do not make it small. An eighth-moment assumption alone does not imply bounded leverage. In particular this proposition is not used as a finite-sample guarantee for the Gaussian experiments. Its concentration inequality is a standard tool, while its role here is to state what extra assumption turns a residual fit into a reliable population certificate.

I CONNECTION TO LOSSES AND DISTRIBUTION CHOICE

Let (x, y) follow the task-a joint distribution. Assume the output loss is differentiable with β-Lipschitz gradient, and all terms below are integrable. For $z = f_{\theta_a}(x)$ and $\delta(x) = f_{\mathrm{fix}}(x) - z$, smoothness gives

$$\ell(z+\delta, y) - \ell(z, y) \le \nabla_z \ell(z, y)^\top \delta + \tfrac{\beta}{2}\|\delta\|^2.$$

Condition on x, integrate, and apply Cauchy–Schwarz to the conditional gradient. Equation (23) gives $\|\delta\|_{L^2(\mathcal{D}_a)} \le e_a + d_a$, proving (24). This upper bound is not a claim that the loss difference is positive; it bounds potential deterioration.

If the analysis norm uses $\mu = \pi_a\mathcal{D}_a + \pi_b\mathcal{D}_b$, with positive mixture weights, then

$$\|r\|_{L^2(\mathcal{D}_a)} \le \pi_a^{-1/2}\|r\|_{L^2(\mu)}.$$

No transfer bound follows solely from a small Gaussian-norm error for a different, arbitrary task distribution. The smoothness constant and conditional gradient can also differ across tasks.

Zero interference with two degraded tasks. Consider the scalar model $f_\theta(x) = \theta$ with base zero. Task a has constant label $+\varepsilon$ and expert parameter $+\varepsilon$; task b has label $-\varepsilon$ and expert parameter $-\varepsilon$. Both experts have zero half-squared loss. Parameter addition exactly realizes functional addition, so $\Delta_\varepsilon = 0$ and the functional repair error is zero. Nevertheless the merged prediction is zero and each task has loss $\varepsilon^2/2$. The interference diagnostic has done its job: there is no nonlinear merge error. The remaining harm is precisely the functional leakage in (23).

For the distillation objective $\mathcal{R}(f) = \frac{1}{2}\|f - g_\varepsilon\|_{\mathcal{H}}^2$, the nonzero orders in Theorem 1 become $\Theta(\varepsilon^2)$, $\Theta(\varepsilon^4)$, and $\Theta(\varepsilon^6)$. This direct consequence concerns a different loss from the original tasks.

J WHY RANK SATURATION IS ACTIVATION-SPECIFIC

We include a counterexample to extrapolating the GELU rank-saturation theorem to every nonlinear activation. Consider $\sigma(z) = 1 + z^2$, $W = V = I_n$, $n \geq 2$, with standard Gaussian input. Only the two weight matrices can change. Let $u = \mathbf{1} \in \mathbb{R}^n$, $e = e_1$, and

$$P = ue^\top, \qquad R = 2ue^\top, \qquad Q = 3ue^\top, \qquad S = 4ue^\top. \tag{46}$$

Every task matrix and both summed matrices have rank one.

Proposition 14 (Polynomial rank expansion)**.** *For this model, quadratic repair is possible, but every local repair achieving $O(\varepsilon^3)$ function error requires both final update matrices to have rank n for sufficiently small nonzero ε.*

Proof. The mixed response is $h(x) = 24ux_1^2$. The Jacobian is

$$J(C, D)_k = \sum_i D_{ki}(1 + x_i^2) + 2x_k \sum_i C_{ki}x_i.$$

Comparing all polynomial coefficients in $h + J(C, D) = 0$ gives the unique solution

$$C_* = -12I_n, \qquad D_* = 24(I_n - ue^\top). \tag{47}$$

For clarity, the off-diagonal cross terms force $C_{ki} = 0$ for $i \neq k$. The coefficients of x_i^2 for $i \neq k$ force $D_{k1} = -24$ when $k \neq 1$ and all other such entries zero. The zero constant coefficient forces each row sum of D to vanish, yielding $D_{kk} = 24$ for $k \neq 1$ and $D_{11} = 0$. The remaining x_k^2 coefficients force $C_{kk} = -12$.

The same coefficient comparison with $h = 0$ shows that J is injective. Gaussian moments give the required local regularity. The argument of Lemma 9 therefore implies that every local $O(\varepsilon^3)$ repair has updates

$$\Delta W = 3\varepsilon ue^\top - 12\varepsilon^2 I_n + O(\varepsilon^3), \quad \Delta V = 7\varepsilon ue^\top + 24\varepsilon^2(I_n - ue^\top) + O(\varepsilon^3).$$

Choose column-space coordinates $\mathrm{span}(u) \oplus u^\perp$ and row-space coordinates $\mathrm{span}(e) \oplus e^\perp$. The leading rank-one block is invertible on its one-dimensional support. For either update, its Schur complement on the remaining coordinates has leading order ε^2 equal, up to a nonzero scalar, to the restriction of the projection onto $u^\perp$ from $e^\perp$. This restriction is injective: a vector in $e^\perp \cap \mathrm{span}(u)$ must be zero since $u_1 = 1$. It is a square map of dimension $n - 1$, hence invertible. The $O(\varepsilon^3)$ remainder cannot destroy its full rank for small ε. Each whole matrix consequently has rank $1 + (n - 1) = n$. Conversely (47) provides a full-rank quadratic repair. □

The constant 1 in the activation removes a continuous scaling ambiguity present in the pure-square model. Parameter permissions matter: if an independent output bias can also be corrected, one can use $C = 0$, $D = -24ue^\top$, and output-bias coefficient $24u$, requiring only a rank-one weight correction. Thus this contrast is not a universal rank lower bound either. It explains why the main paper fixes both its activation and its repair permissions.

K EXPERIMENTAL PROTOCOL, ADDITIONAL RESULTS, AND AUDIT

K.1 DATA GENERATION AND POPULATION INTEGRATION

The experiments are deterministic given the documented seeds. The primary suite uses seeds $s = 0, \ldots, 15$. For each seed, independent standard normal arrays are drawn in the order specified in `code/experiments.py`, giving

$$W = \begin{pmatrix} 1.1 & 0.35 \\ -0.3 & 1.0 \end{pmatrix} + 0.08Z_W, \qquad V = \begin{pmatrix} 1.0 & 0.3 \\ -0.4 & 0.9 \end{pmatrix} + 0.06Z_V,$$
$$\beta = (-0.3, 0.4)^\top + 0.04Z_\beta, \qquad p = (0.35, 0.15)^\top + 0.03Z_p,$$
$$r = (-0.15, 0.30)^\top + 0.03Z_r, \qquad \lambda = 0.4 + 0.04Z_\lambda, \quad \mu = -0.35 + 0.04Z_\mu.$$

The five direction constructions are as follows. All unmentioned entries are zero.

Repairable, nonexact. $P_{1:} = p^\top$, $R_{2:} = r^\top$, $Q_{:2} = \lambda v_2$, $S_{:1} = \mu v_1$.

Exact. The same P, R, but $Q_{:1} = \lambda v_1$, $S_{:2} = \mu v_2$.

Readout mismatch. The repairable construction, with $0.3(-V_{22}, V_{12})^\top$ added to $Q_{:2}$, perpendicular to v_2.

Input overlap. $P = (1, 0.7)^\top p^\top$, $R = (0.6, 1)^\top r^\top$, $Q = (0.12, -0.08)^\top (0.6, 1)$, and $S = (-0.06, 0.15)^\top (1, 0.7)$.

Orthogonal overlap. The same first-layer form as input overlap, but $r = (-p_2, p_1)^\top$ and $Q = S = 0$.

The same seed deliberately supplies a common base across categories. There is no search for favorable seeds, rejection sampling, or hyperparameter selection based on slopes. Every generated primary network and all four direction matrices are included as a compressed NumPy file.

For $x \sim N(0, I_2)$, the L^2 norm is approximated using the tensor product of 80-point probabilists' Gauss–Hermite quadrature in each dimension, producing 6,400 nodes with weights summing to one. The 96×96 comparison grid is used to check all 80 quadratic-response least-squares problems. The maximum residual difference is 5.64×10^{-15} and maximum coefficient-vector difference 1.62×10^{-14}. These comparisons test convergence on this suite; they do not constitute rigorous interval enclosures or automatically certify every other integral in the package.

K.2 Methods, scales, and measured quantities

Every primary case is evaluated at $\varepsilon \in \{0.2, 0.1, 0.05, 0.025, 0.0125, 0.00625, 0.003125\}$. The methods are direct summation, the coefficient repair (10), and a full-Jacobian second-order coefficient fit. This fit solves the weighted linear least-squares problem $\min_c \|Jc + h\|_{\mathcal{H}}$ on the quadrature grid, then applies $\varepsilon^2 c$ to the original parameterization. It is neither a free per-input output correction nor a nonlinear optimizer of the entire finite-scale objective.

For repairable cases, we also compute the analytic h_3 from (12)–(32), project it against the same full Jacobian, and add the resulting coefficient at scale ε^3. For input-overlap cases, the rank-insufficient path replaces $P + R$ by its rank-one truncated SVD while retaining $Q + S$. This path demonstrates the predicted first-order rate, without claiming its leading constant is optimal. No slope is fitted to exact-additivity roundoff.

The response norms use $\sqrt{\mathbb{E} \sum_{o=1}^{k} r_o(x)^2}$. Least squares uses singular-value cutoff 10^{-13} relative to the largest singular value. Slopes regress $\log$ error against $\log \varepsilon$ on the smallest five scales. The mean and sample standard deviation in Table 2 summarize 16 slopes, not a confidence interval for a population of trained models. The raw suite has 1,904 function-error records and 224 fitted slopes; `orders.csv`, `slopes.csv`, and `responses.csv` retain all cases, including methods not shown in the main table.

K.3 Calibration experiment

The scalar-output base uses the W, β from primary seed 2026 and fixes $V = (1, -0.8)$. Its input-overlap directions use $Q = S = 0$. For each of 40 seeds $10000 + s$, draw 256 independent $N(0, I_2)$ inputs and use nested prefixes of sizes $3, 5, 6, 7, 12, 24, 48, 96, 256$. The sample-dependent correction is fitted using the empirical Jacobian. Its evaluation residual uses the independent, deterministic Gaussian quadrature law, not another subset of the training points. All sampled inputs are included. The fixed population coefficient residual is 0.0211341011.

The 360 fits have numerical rank $\min(N, 6)$ at the stated cutoff. These controlled instances are well enough conditioned to expose the exact-arithmetic transition. The theorem does not promise that every practical sample will have an easily resolved singular-value gap.

K.4 Quantitative obstruction and conditioning

For a fixed primary base generated with seed 400, draw 96 independent direction sets with seeds $20000 + s$. Each of P, Q, R, S is 0.2 times the outer product of two independent standard normal

Table 3: Calibration and evaluation residual medians over 40 draws. Values are second-order coefficient residuals, not finite-ε task losses.

N	Calibration residual	Evaluation residual	Oracle
3	$9.106e-18$	0.03018	0.02113
5	$2.453e-17$	0.04935	0.02113
6	$6.311e-17$	0.05442	0.02113
7	$2.473e-03$	0.03819	0.02113
12	$9.472e-03$	0.03213	0.02113
24	$1.453e-02$	0.02589	0.02113
48	$1.754e-02$	0.02362	0.02113
96	$1.996e-02$	0.02227	0.02113
256	$2.058e-02$	0.02160	0.02113

vectors of size two. The Gram spectrum, κ_2, full-Jacobian residual, and coefficient-repair residual are recorded in `spectral_bounds.csv`. The fixed Gram condition number is about 650.17. All 96 lower and upper inequalities pass the numerical comparison with absolute tolerance 10^{-12}; this check is not a proof of the inequalities. To vary conditioning rather than the task size, use

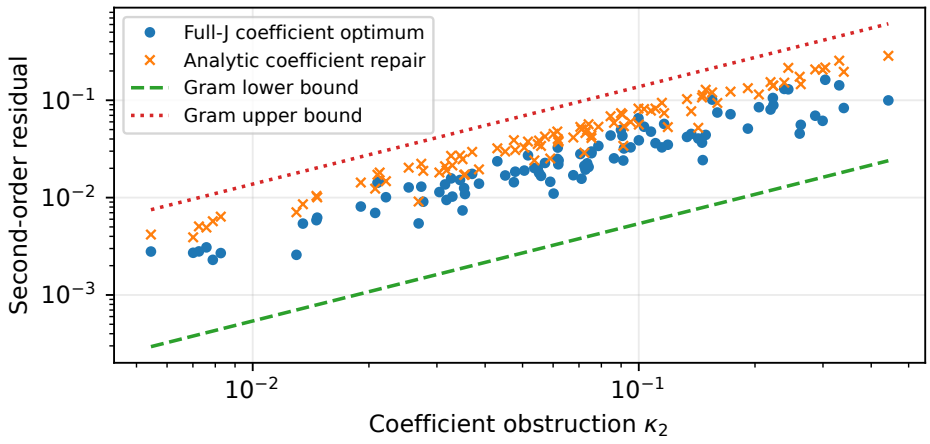


Figure 3: Ninety-six rank-one direction sets on one fixed base. The Gram bounds enclose the optimal quadratic residual; the analytic coefficient repair is not asserted to be the functional least-squares optimum.

scalar readouts $v_1 = v_2 = 1$, input weights $(1, 0.2)$ and $(1, 0.2 + \delta)$, and biases 0.2 and $0.2 + 0.2\delta$. Both rows of P equal $(0.3, 0.1)$; rows of R are $(0.1, 0.3)$ and its negative; $Q = S = 0$. For $\delta \in \{1, 0.5, 0.2, 0.1, 0.05, 0.02, 0.01\}$, $\kappa_2 = 0.116619\ldots$ is fixed, while the interaction norm decreases from approximately 0.03037 to 0.000849. Thus a positive coefficient obstruction alone is not a uniform error lower bound as the base approaches a collision. Each tested $\delta > 0$ still satisfies nondegeneracy.

K.5 Rank and exact-width construction checks

For seeds $30000, \ldots, 30015$, sample $W \in \mathbb{R}^{8\times 6}$, $V \in \mathbb{R}^{5\times 8}$, and biases from standard normals. Sample p, r as 0.15 times standard normal six-vectors. Activate row 1 of P and row 2 of R, with $Q_{:2} = 0.4v_2$, $S_{:1} = -0.3v_1$. Every task matrix has rank one. At scales $0.2, 0.05, 0.01$, both original summed-update ranks and corrected ranks are two, giving 48 rank-preservation checks.

For each base, separate direction constructions exercise all five rows of (20) at $\varepsilon = 0.05$, and compare the compiled network to the functional target at 512 generated inputs. The rowwise cases give widths $8, 8, 16, 16, 24$ respectively. The first case has no first-layer update; the second has only one task's first-layer change and no cross readout; the third adds a rank-one cross readout; the fourth uses equal nonzero first-layer shifts; the fifth uses distinct shifts from two rank-one matrices. There are 80

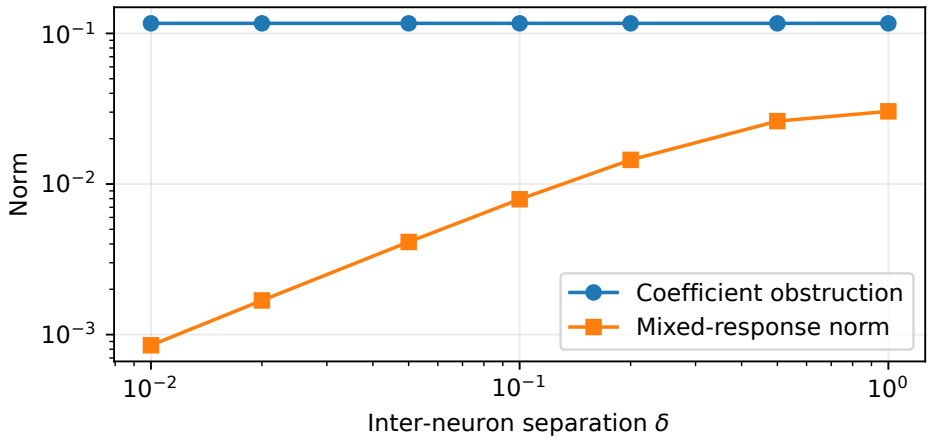


Figure 4: A fixed coefficient obstruction with shrinking function-space interaction near a neuron collision. The theorem's conditioning constants cannot be discarded.

checks and the maximum L^2 evaluation discrepancy is 4.30×10^{-15}. The implementation groups exactly equal orientations after symbolic zero-case simplification. It is not a general optimizer for approximate or sign-equivalent ridges outside the theorem's small-scale regime.

K.6 Near-compatibility and cubic-certificate checks

For each primary repairable seed $s = 0, \ldots, 15$, let $n_2 = (-V_{22}, V_{12})^\top / \|v_2\|$. Add δn_2 to the cross-task readout q_2, leaving all other directions unchanged. This gives exactly $\kappa_2 = |\delta| \, \|r_2\|$, because the input supports stay disjoint and the perturbation is perpendicular to v_2. Evaluate (10) at the same seven scales and quadrature grid as the primary suite for three choices: $\delta = 0$, fixed $\delta = 0.1$, and $\delta = \varepsilon$. Slopes use the smallest five scales and retain every seed.

Table 4: Near-compatibility checks: mean $\pm$ sample standard deviation across 16 seeds. The $\delta = \varepsilon$ row changes the task directions with scale and tests Theorem 5; it is not a new fixed-direction regime of Table 1.

Cross-readout perturbation	Predicted order	Fitted slope
$\delta = 0$	3	3.0002 ± 0.0002
$\delta = 0.1$ (fixed)	2	1.9989 ± 0.0013
$\delta = \varepsilon$ (varying directions)	3	2.9939 ± 0.0004

There are 336 function-error records and 48 fitted slopes. The fixed defect gives a second-order residual, while the scale-matched defect gives third-order accuracy despite failing exact S_2 at each nonzero scale. The recorded feature residuals obey the coefficient upper bound, and the ε^{-3}-scaled Taylor remainders are recorded rather than identified with a sharp uniform constant.

For the cubic certificate, augment the same repairable directions by an own-task readout $0.17 n_1$ in q_1 and $-0.13 n_2$ in s_2, where n_i is the corresponding unit perpendicular vector. These changes preserve S_2 and exercise the readout-mismatch terms of (40). All 16 comparisons satisfy (41) numerically; the coefficient refinement preserves the summed-update rank at all seven scales, giving 112 rank checks. The implementation, raw CSVs, source hashes, and summaries are in `code/extension_experiments.py` and the files prefixed `near_` or `cubic_` in `results/`. Additional tests independently reconstruct the cubic feature coefficients and check the actual-sum rank example in Appendix E.2.

K.7 Implementation audit and reproducibility boundary

The included 45 tests cover analytic Jacobians and mixed Hessian responses against PyTorch automatic differentiation, exact-GELU identities, symbolic derivation of the cubic coefficient, second-order cancellation and uniqueness on controlled instances, nonzero projected cubic residuals, orthogonal-overlap counterexamples, rank preservation, all width cases, calibration ranks, Gram bounds, invalid inputs, and the polynomial-activation contrast. They all pass in the recorded environment; the added tests also cover the uniform near-compatibility mechanism, cubic coefficient bounds, and actual-sum rank counterexample. The symbolic test checks the scalar algebra of the cubic expression; it is not a mechanized proof of the functional independence lemma or the asymptotic classification.

All production experiments use CPU float64 arithmetic, NumPy and SciPy. PyTorch and SymPy are used only for independent checks. The package records library versions, deterministic seeds, numerical cutoffs, source hashes, raw outputs, and executable commands. Plots and numerical tables are generated from the CSVs rather than manually entered. Hardware-dependent wall-clock measurements are not used as scientific performance claims. No real tasks, pretrained checkpoints, fine-tuning runs, benchmark accuracies, or training-efficiency comparisons were performed. The appropriate empirical conclusion is consistency of the formulas and mechanisms on the specified synthetic instances.