BEYOND BOUNDARY AGREEMENT: SHARP REPAIR COSTS FOR HARD-ROUTED MIXTURES OF EXPERTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Can a hard-routed mixture of experts be made L-Lipschitz without changing its selected expert set, while preserving its original function? We show that boundary agreement and transition-layer representation impose distinct architectural costs. A weighted projection inequality transfers approximation error of a forced, rescaled transition profile into excess repair risk. For an explicit two-expert Top-1 model, the original experts and the optimal boundary value use at most one ReLU, yet every width-Q shallow repair has an explicit leading-risk gap of at least $c_0\Lambda_0^{-2Q}$. A self-contained differential certificate gives this bound without restrictions on network parameters. Consequently, asymptotically optimal shallow repair requires diverging width; matching the $O(L^{-2})$ relative excess of a two-stage, two-ReLU repair requires width $\Omega(\log L)$. In contrast, scalar one-dimensional repair admits exact convex endpoint optimization when expert Lipschitz constants are below L, with at most six additional units per shallow ReLU branch. We also establish finite-budget scalar and vector trace oracles, a quantitative Top-k trace-width obstruction, and a scalar polyhedral expansion with a sharp-order $O(L^{-2})$ remainder. Selected expert calls, evaluated nonlinear units, and auxiliary computation are accounted for separately.

1 INTRODUCTION

Sparse mixtures of experts (MoEs) select only part of their model for each input (Shazeer et al., 2017; Fedus et al., 2022). Crossing a hard-routing interface can change the output discontinuously even when every expert is Lipschitz. We fix the router and original function, and ask for the best squared-error approximation under a global Lipschitz budget.

Boundary agreement is not enough. Continuity requires a common boundary value, but connecting even a simple value to the original experts can require a more expressive architecture. Our example has branches 0 and $1 + \sigma(y)$ and the one-ReLU optimal boundary value $(1 + \sigma(y))/2$. Nevertheless, every fixed-width shallow repair has a positive leading-risk gap, whereas two ReLU evaluations in successive nonlinear stages attain the optimal leading coefficient. The difficult target is *produced by the optimal repair problem*, not inserted into a complicated original expert.

From geometry to a quantitative resource separation. Theorem 1 bounds the transition profile's squared approximation error by excess rescaled risk. A differential certificate gives the explicit gap $c_0\Lambda_0^{-2Q}$ at shallow width Q (Theorem 2), requiring $\Omega(\log L)$ width to match the two-stage repair's $O(L^{-2})$ relative excess (Corollary 3). The original experts, router, and distribution remain fixed.

Where the obstruction disappears, and where another appears. Exact one-dimensional repair needs only convex endpoint optimization and constant additive width (Theorem 6). Missing boundary features cause a separate Top-k width deficit (Theorem 5). Clipping oracles and a scalar polyhedral expansion identify the corresponding function-space benchmarks.

These results build on gradient-constrained approximation (Buccheri et al., 2023), scalar Lipschitz extension (McShane, 1934), and established ReLU expressivity theory (Arora et al., 2018; Rueß et al., 2026). The new connection is that near-optimal stabilization forces a specific representation problem, whose obstruction can be converted into a quantitative fidelity penalty.

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.

2 PROBLEM AND COMPUTATIONAL MODEL

Let $\Omega \subset \mathbb{R}^d$ be compact and convex. A fixed affine router partitions it into finitely many polyhedral cells C_i. Its original output is $F(x) = E_i(x)$ on a cell. More generally, a cell may correspond to an active set S, with branch output

$$F_S(x) = \sum_{i \in S} w_i^S(x) E_i(x), \qquad |S| \le k. \tag{1}$$

We impose explicit regularity on the *branch output*, not merely on each summand. Norms are Euclidean. Given a probability measure μ, define

$$\mathcal{R}(G;F) = \int_\Omega \|G - F\|^2 \, \mathrm{d}\mu, \qquad \mathcal{R}_L^*(F) = \inf_{\mathrm{Lip}(G) \le L} \mathcal{R}(G;F). \tag{2}$$

The unrestricted infimum is a benchmark, not a claim of cheap implementation. Smaller L means a stronger stability constraint; our large-L expansions describe thin transition layers. Mere continuity permits arbitrarily thin layers and therefore need not entail a positive mean-square fidelity cost. Uniform approximation is different: a jump of size J imposes a maximum-error lower bound $J/2$ on every continuous repair.

Three separate resources. We distinguish (i) the selected expert indices, which remain those of the original router; (ii) the total number of *evaluated* nonlinear units, regardless of whether their output is zero; and (iii) auxiliary trace, geometry, and connection operations. An auxiliary network is not free because it is called a shared component. A compiled expert can contain copied features, but these count toward its evaluated width. The shallow lower bounds below concern affine combinations of one-hidden-layer ReLU experts, with no nonlinear mixture head and no input-dependent gate multiplication unless explicitly stated. The two-stage upper bound allows affine skip connections.

The risk in (2) measures preservation of the original function, not generalization, classification accuracy, or the outcome of retraining. All theorems hold for a fixed router. Jointly changing the router is a different optimization problem.

3 BOUNDARY-LAYER RIGIDITY TRANSFERS ARCHITECTURAL OBSTRUCTIONS

Consider a flat interface with coordinates $(y,t) \in Y \times [-T,T]$, where $Y \subset \mathbb{R}^{d-1}$ is compact and convex, and

$$F(y,t) = \begin{cases} a(y), & t < 0, \\ b(y), & t > 0, \end{cases} \qquad \mathrm{d}\mu = \rho_\pm(y)\,\mathrm{d}y\,\mathrm{d}t. \tag{3}$$

The bounded vector-valued traces a, b are Lipschitz, and the positive bounded densities are constant in the normal direction but may differ across the interface. Set $J = \|b - a\|$, $u = (b-a)/J$, and

$$r = \frac{\sqrt{\rho_+}}{\sqrt{\rho_-} + \sqrt{\rho_+}} J, \qquad P(y,s) = a(y) + u(y)\,\mathrm{clip}(s + r(y), 0, J(y)). \tag{4}$$

When $J = 0$, set $P = a$. The common value is $c_* = a + ur$, not necessarily the unweighted midpoint. Define

$$\mathcal{S} = \frac{1}{3} \int_Y \frac{\rho_- \rho_+}{(\sqrt{\rho_-} + \sqrt{\rho_+})^2} J^3 \, \mathrm{d}y. \tag{5}$$

For equal side densities ρ, this is $\int_Y \rho J^3/12$.

Theorem 1 (Weighted boundary-layer rigidity)**.** *Let $R \ge \sup_y \max\{r(y), J(y) - r(y)\}$ and $LT \ge R$. Every L-Lipschitz repair satisfies*

$$L\mathcal{R}(G;F) \;\ge\; \mathcal{S} + \int_Y \int_{-R}^{R} \rho_{\mathrm{sgn}\, s}(y) \, \|G(y, s/L) - P(y,s)\|^2 \, \mathrm{d}s \, \mathrm{d}y. \tag{6}$$

Proof idea. For fixed y, $H(s) = G(y, s/L)$ is 1-Lipschitz. The profile P is the squared-error projection of the step onto this convex constraint. This fact has a direct verification: on the ramp, the component of $H - P$ parallel to u is nonincreasing. The signed residual $P - F$ is positive before

the jump and negative after it, and its weighted integral is zero because $\rho_- r^2 = \rho_+(J-r)^2$. Hence $\int \rho\langle P-F, H-P\rangle \geq 0$. Expanding the square gives the projection inequality. Integrating in y and changing variables $s = Lt$ proves (6). Appendix A supplies all details, including vector outputs. □

Taking infima in (6) transfers an architecture's distance from P into excess risk. Thus $L\mathcal{R}(G_L; F) \to \mathcal{S}$ forces weighted L^2 convergence to the entire profile, not just agreement at $s = 0$. Only normal-slice optimality of P is used; global feasibility of P is unnecessary.

4 A STABILIZATION-INDUCED DEPTH SEPARATION

Use uniform probability measure on $[-1,1]^2$, retain the router selecting the negative or positive sign of t, and let

$$E_-(y,t) = 0, \qquad E_+(y,t) = 1 + \sigma(y), \qquad \sigma(z) = \max\{z, 0\}. \tag{7}$$

Each original expert evaluates at most one ReLU. The optimal common boundary value $(1+\sigma(y))/2$ has the same complexity. Let $\mathcal{N}_Q$ denote affine functions plus at most Q ReLU ridge functions of their inputs.

Theorem 2 (Quantitative shallow gap; constant-size compositional repair)**.** *For the model* (7)*, put*

$$S_0 = \frac{19}{192}, \qquad c_0 = \frac{1}{2^{29}3^8\pi}, \qquad \Lambda_0 = 4608. \tag{8}$$

Every globally L-Lipschitz Top-1 repair with each selected branch in $\mathcal{N}_Q$ obeys, for every integer $Q \geq 0$ and $L \geq 2$,

$$L\mathcal{R}(G; F) \geq S_0 + c_0\Lambda_0^{-2Q}. \tag{9}$$

In contrast, put $\lambda = \sqrt{L^2 - 1/4}$ and $h(y) = (1+\sigma(y))/2$. The branches

$$G_-(y,t) = \sigma(h(y) + \lambda t), \qquad G_+(y,t) = 1 + \sigma(y) - \sigma(h(y) - \lambda t) \tag{10}$$

form an L-Lipschitz repair evaluating two ReLUs per input in two successive nonlinear stages. They preserve Top-1 selection and satisfy

$$\mathcal{R}(G; F) = \frac{S_0}{\lambda}, \qquad 1 \leq \frac{\mathcal{R}(G;F)}{\mathcal{R}_L^*(F)} \leq \frac{L}{\sqrt{L^2 - 1/4}} = 1 + O(L^{-2}). \tag{11}$$

Moreover, $G = F$ whenever $|t| \geq 1/\lambda$.

Corollary 3 (Width required for oracle-level fidelity)**.** *Under the shallow architecture of Theorem 2, if $\mathcal{R}(G;F) \leq (1+\varepsilon)\mathcal{R}_L^*(F)$ for $L \geq 2$ and $\varepsilon \geq 0$, then*

$$Q \geq \frac{[\log(c_0/(S_0\eta_L))]_+}{2\log\Lambda_0}, \qquad \eta_L = (1+\varepsilon)\frac{L}{\sqrt{L^2-1/4}} - 1. \tag{12}$$

Thus $\mathcal{R}(G_L;F)/\mathcal{R}_L^*(F) \to 1$ requires $Q_L \to \infty$. If that ratio is $1 + O(L^{-2})$, then

$$Q_L \geq \frac{\log L}{\log \Lambda_0} - O(1), \tag{13}$$

whereas the two-stage construction uses two ReLUs for every $L \geq 2$.

Proof. Theorem 2 bounds $L\mathcal{R}_L^* \leq S_0 L/\sqrt{L^2-1/4}$. Combining this with (9) gives $c_0\Lambda_0^{-2Q} \leq S_0\eta_L$ and hence (12). In the two limits, respectively, $\eta_L \to 0$ and $\eta_L = O(L^{-2})$. □

Why a simple trace is not enough. On the negative normal half-plane, the forced profile is

$$P_-(y,s) = \sigma\big(1/2 + \sigma(y)/2 + s\big). \tag{14}$$

Writing $u = y/2$ and $v = s + 1/2$ gives $p(u,v) = \max\{0, v, v+u\}$. Its gradient jumps along $v = 0$, $u < 0$, but not along $v = 0$, $u > 0$. In a finite shallow ReLU sum, coincident activation lines can be combined; the resulting gradient jump is constant along each full line away from other intersections. This explains the exact representation obstruction in Figure 1.

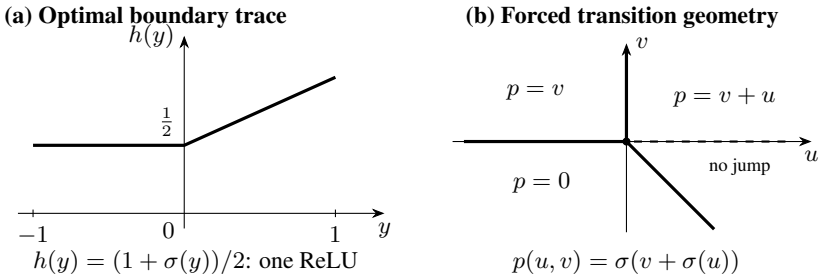

Figure 1: **Simple boundary data can force a non-shallow connection.** Left: the one-ReLU optimal trace $h(y)$. Right: the forced negative-side profile in local coordinates, $p(u,v) = \max\{0, v, v+u\}$. Solid rays are gradient jumps; the dashed continuation of $v = 0$ has no jump. A shallow ReLU's jump persists along its full activation line. The differential identity (15) turns this local geometry into a quantitative approximation certificate.

A parameter-uniform quantitative certificate. To turn the obstruction into a positive *distance*, define

$$\mathscr{T} = \partial_u \partial_v (\partial_u - \partial_v), \qquad \mathscr{T} p = \delta_{(0,0)} \tag{15}$$

in the distributional sense. Differential annihilation of ridge functions is classical (Lin and Pinkus, 1993): a derivative parallel to a ridge's level lines annihilates it. For any Q-ridge candidate, multiply these Q derivatives and then apply $\mathscr{T}$: the candidate vanishes, while p produces a derivative of a point mass. A compactly supported polynomial test function detects this mass with an explicit norm bound, giving

$$\inf_{q \in \mathcal{N}_Q} \|p - q\|^2_{L^2(B_{1/4})} \geq 2c_0 \Lambda_0^{-2Q}. \tag{16}$$

Theorem 1 and the Jacobian $\mathrm{d}y\,\mathrm{d}s = 2\,\mathrm{d}u\,\mathrm{d}v$ give (9). Appendix B proves every step and gives a sharper factorial-form bound. The certificate permits arbitrary parameter magnitudes, parallel or coincident ridges, and even arbitrary continuous univariate ridge profiles; it does not require a best-approximation existence theorem.

The upper bound follows from transition gradients of norm at most L and exact ramp integration; Appendix B gives the details and the two-ReLU accounting.

5 A CLIPPING PRIMITIVE WITH FINITE-BUDGET ORACLE BOUNDS

In the flat model (3), write the two original branches as $f_\pm(y)$, with Lipschitz constant at most B. Let a nonempty common-trace class $\mathcal{H}$ obey, uniformly over $h \in \mathcal{H}$,

$$\mathrm{Lip}(h) \leq K, \qquad \sup_{y,\pm} \|f_\pm(y) - h(y)\| \leq M. \tag{17}$$

Define the trace functional

$$\Psi(h) = \frac{1}{3} \int_Y \left[\rho_- \|f_- - h\|^3 + \rho_+ \|f_+ - h\|^3 \right] \mathrm{d}y, \qquad \Psi_* = \inf_{h \in \mathcal{H}} \Psi(h). \tag{18}$$

Let $\mathcal{R}_{L,\mathcal{H}}$ be the unrestricted L-Lipschitz optimum with boundary value in $\mathcal{H}$; this is still a function-space optimum conditioned on trace complexity.

Theorem 4 (Finite-L trace oracle)**.** *For scalar outputs, assume $L \geq B$, $L > K$, and take $\lambda = \sqrt{L^2 - K^2}$. For vector outputs, set $B_0 = \max\{B, K\}$, assume $L > 0$ and $L \geq 2B_0$, and take*

$$\lambda = L\sqrt{1 - \frac{K^2}{L^2 - B_0^2}}. \tag{19}$$

Whenever $\lambda T \geq M$,

$$\frac{\Psi_*}{L} \leq \mathcal{R}_{L,\mathcal{H}} \leq \frac{\Psi_*}{\lambda}. \tag{20}$$

The upper bound is realized up to the infimum over h by

$$G_h(y,t) = h(y) + P_{\lambda|t|}(f_\pm(y) - h(y)), \qquad P_r(v) = v\min\{1, r/\|v\|\}, \tag{21}$$

with $P_r(0) = 0$. Its pointwise squared error is exactly $(\|f_\pm - h\| - \lambda|t|)_+^2$.

Only the selected original branch is evaluated in (21). The trace and clipping are additional computation. The lower bound follows by integrating $\|G(y,\pm t) - f_\pm(y)\| \geq (\|h - f_\pm\| - Lt)_+$. The scalar upper bound uses transition gradients $(\nabla h, \pm\lambda)$. In the vector case, radial and tangential output components give the operator-norm majorant

$$\begin{pmatrix} K & \lambda \\ B_0 & 0 \end{pmatrix}, \tag{22}$$

whose norm is at most L for (19). Appendix C verifies the matrix inequality and nonsmooth gluing. For fixed B, K, the oracle ratio is $1 + O(K^2/L^2)$; it is exact when $K = 0$ under the stated hypotheses.

For equal densities and midpoint $m = (f_- + f_+)/2$, the excess trace objective is comparable, pointwise, to

$$J\|h - m\|^2 + \|h - m\|^3. \tag{23}$$

Thus imperfect boundary fitting has a quadratic leading penalty where $J > 0$, rather than a linear one. It also penalizes needless modification where $J = 0$. Appendix D gives dimension-free constants and a one-sided unbiased estimator of (18): each sample evaluates one side's selected set, not the union of both sides. Obtaining boundary samples and their density weights remains an explicit prerequisite.

6 A QUANTITATIVE TOP-k TRACE-WIDTH OBSTRUCTION

The trace can itself consume additional features. We make this precise for fixed affine Top-k routing and an arithmetic mean of shallow scalar experts. The budget P below is the maximum total number of ReLUs evaluated within any selected set; no nonlinear post-mixture head is permitted.

Theorem 5 (Evaluated-width deficit)**.** *For every $k, W \geq 1$, there is a uniform-input model on $[0,1] \times [-1,1]$ with $k+1$ experts of width W, always selecting exactly k, for which $N = (k+1)W$ and*

$$L\mathcal{R}(G;F) \geq \mathcal{S}(F)\left(1 + \frac{(N-P)_+}{432(N+1)}\right) \tag{24}$$

for every globally L-Lipschitz repair in the stated architecture and all sufficiently large L. The unrestricted optimum obeys $\mathcal{R}_L^* = \mathcal{S}/L + o(L^{-1})$. In particular, retaining the original evaluated width $P = kW$ leaves a positive relative leading-order gap.

Construction and lower-bound mechanism. There are $k-1$ always-selected experts and two experts exchanged at $t = 0$. Put $\tau = 1/(N+1)$ and place all N hidden thresholds at $\xi_r = r\tau$, allocating W thresholds to each expert. Alternating their coefficients, with appropriate offsets, makes the two branch outputs obey

$$a\tau \leq J(y) \leq 3a\tau, \tag{25}$$

while their midpoint has a slope jump of magnitude a at every threshold. Every continuous repaired output has a common boundary trace b with at most P knots. At least $(N-P)_+$ disjoint intervals of length τ therefore contain a midpoint hinge but no knot of b. The exact affine approximation error for ax_+ on $[-\tau/2, \tau/2]$ is $a^2\tau^3/192$.

A truncated normal-slice estimate, valid even for unbounded candidate trace values, yields

$$L\mathcal{R}(G;F) \geq \mathcal{S}(F) + \frac{1}{2}\int_0^1 J(y)|b(y) - m(y)|^2\,\mathrm{d}y. \tag{26}$$

Combining these inequalities proves (24); the complete coefficients and proof are in Appendix E. This obstruction differs from Theorem 2: it counts missing boundary features rather than an unrepresentable connection geometry. The bound is deliberately restricted to shallow additive branches. Multiplicative gates, deep experts, or a nonlinear head can invalidate the knot-count premise.

7 EXACT ONE-DIMENSIONAL REPAIR WITH CONSTANT WIDTH OVERHEAD

The two-dimensional obstruction has a precise one-dimensional counterpart: for scalar interval branches, boundary values determine an exact optimal repair with only a constant increase in shallow width.

Let $\tau_0 < \cdots < \tau_n$, $I_j = (\tau_{j-1}, \tau_j)$, and $F = E_j$ on I_j, with specified router tie values at endpoints. For endpoint values $c_0, \ldots, c_n$ satisfying $|c_j - c_{j-1}| \le L(\tau_j - \tau_{j-1})$, set

$$\begin{aligned} \ell_j(x) &= \max\{c_{j-1} - L(x - \tau_{j-1}),\ c_j - L(\tau_j - x)\}, \\ u_j(x) &= \min\{c_{j-1} + L(x - \tau_{j-1}),\ c_j + L(\tau_j - x)\}. \end{aligned} \tag{27}$$

Theorem 6 (Exact endpoint optimization and constant additive width)**.** *Suppose* $\mathrm{Lip}(E_j) \le B < L$ *for every scalar branch. For any finite Borel input measure, minimizing the risk of* $G_c(x) = \mathrm{clip}(E_j(x), \ell_j(x), u_j(x))$ *on* I_j*, with* $G_c(\tau_r) = c_r$*, over the feasible endpoint values is a convex problem whose attained minimum equals* $\mathcal{R}^*_L(F)$*. Endpoint atoms contribute* $\mu(\{\tau_r\})|c_r - F(\tau_r)|^2$*.*

If each original interval branch is a width-W *shallow ReLU function, an exact minimizer has at most* $W + 6$ *ReLUs per interval branch with an affine skip, or* $W + 8$ *without one. The interval router is unchanged. For an empirical measure, the optimum is the value of a quadratic program with* $O(n + n_{\mathrm{sample}})$ *variables and constraints.*

Proof idea. At fixed endpoints every feasible value lies in $[\ell_j, u_j]$. Clipping E_j to this interval is pointwise optimal and globally feasible, so the reduction is exact. The distance to the interval is a nonnegative convex function of the endpoints, hence so is its square. Each of the four slope-$\pm L$ lines crosses E_j at most once because $B < L$; the two envelopes add at most two more corners. Thus clipping adds at most six knots. Appendix F proves attainment, handles atoms, and derives the QP. □

8 SCALAR POLYHEDRAL REPAIR COSTS

The local jump cost extends beyond an isolated plane. Suppose F is scalar on a fixed finite polyhedral complex in a compact, convex, full-dimensional polytope. Assume $|E_i| \le M$, $\mathrm{Lip}(E_i) \le B$, and each cell density ρ_i is Lipschitz up to its boundary and bounded above and below by positive constants. For an internal facet Σ_{ij}, count it once and set

$$\mathcal{S}(F) = \frac{1}{3} \sum_{i<j} \int_{\Sigma_{ij}} \frac{\rho_i \rho_j}{(\sqrt{\rho_i} + \sqrt{\rho_j})^2} |E_i - E_j|^3 \, \mathrm{d}\mathcal{H}^{d-1}. \tag{28}$$

Theorem 7 (Scalar polyhedral expansion)**.** *For all sufficiently large* L*, with constants depending on the fixed geometry and regularity bounds,*

$$\left| \mathcal{R}^*_L(F) - \frac{\mathcal{S}(F)}{L} \right| \le \frac{C}{L^2}. \tag{29}$$

The L^{-2} *remainder order cannot be uniformly improved under these assumptions.*

Pairwise optimal boundary values can disagree at multi-cell junctions. The construction in Appendix G resolves these incompatible traces: remove facet strips of width $O(1/L)$ around the codimension-two skeleton, extend the remaining optimal traces with a controlled Lipschitz constant, and use scalar Lipschitz envelopes. The affected junction neighborhood has volume $O(L^{-2})$. A one-dimensional example $F(t) = \beta t + (J/2)\operatorname{sgn} t$ has the exact risk $\rho J^3/[12(L - \beta)]$, which supplies a nonzero L^{-2} term. This identifies the function-space benchmark against which architecture-dependent penalties are measured.

9 RELATION TO EXISTING APPROACHES AND COMPUTATIONAL SCOPE

Sparse and continuous routing. Shazeer et al. (2017) and Fedus et al. (2022) develop sparsely gated and Top-1 architectures. DSelect-k and COMET study differentiable selection and cardinality

L	Integrated risk of (10)	Ratio to $19/(192L)$
2	0.051101863596	1.032795559
4	0.024935156205	1.007905261
8	0.012394022428	1.001958866
16	0.006187918016	1.000488639

Table 1: Direct piecewise quadrature of the explicit two-ReLU evaluator. Ratios are to a *lower bound* on the unrestricted optimum, not to a numerically solved optimum. Over 12 tested L values, the maximum absolute difference from $19/(192\sqrt{L^2-1/4})$ was 6.94×10^{-18} in this run.

constraints (Hazimeh et al., 2021; Ibrahim et al., 2023); differentiability of selection parameters is not by itself our input–output Lipschitz guarantee. ReMoE uses ReLU routing with adaptive per-input computation (Wang et al., 2025). SmoothSMoE studies switching discontinuities and softly adds experts near boundaries, with an average-activation control (Huu et al., 2026). We keep the selected original indices fixed and quantify the cost of preserving an already specified function. Our comparison criterion is fidelity to that fixed function.

Approximation and representation. Buccheri et al. (2023) study gradient-constrained L^2 approximation, including a step example. Scalar envelopes use Lipschitz extension (McShane, 1934); distance-vanishing neural boundary constructions appear in Sukumar and Srivastava (2022). ReLU depth (Arora et al., 2018; Rueß et al., 2026) and differential ridge annihilation (Lin and Pinkus, 1993) are established subjects. Best-approximation attainment (Dereich et al., 2024) supplies a qualitative route from nonrepresentation to positive distance. Here the explicit certificate quantifies the distance to a profile *forced by optimal repair*, even for very small original experts.

What an activation cap does and does not buy. For a general multi-cell model with a common K-Lipschitz auxiliary function H, the conservative construction

$$G(x) = H(x) + P_{\lambda d_\Sigma(x)}(F(x) - H(x)), \qquad \lambda = L - B - 2K > 0, \tag{30}$$

where d_Σ is distance to the routing interface, is L-Lipschitz and preserves the selected set. Its exact squared error is $(\|F - H\| - \lambda d_\Sigma)_+^2$. Appendix H proves this and explains affine score-distance calculations. Its runtime is the original selected computation plus trace, distance, and clipping costs. If $\|F - H\| \le M$ is known, inputs with $d_\Sigma \ge M/\lambda$ bypass even the trace calculation. These costs, rather than the selected count alone, determine runtime.

10 Reproducible checks, limitations, and conclusion

The deterministic NumPy/SciPy suite passes 14 groups and 2,399 checks covering weighted steps, rigidity, clipping, depth and Top-k constructions, selected-set accounting, and interval QPs. Six additional groups and 330 checks cover the quantitative certificate, including exact distributional pairings and explicit constants. Appendix J records both protocols. Table 1 compares quadrature of the implemented two-ReLU repair with its exact risk.

Limitations. The explicit shallow lower bound is conservative and has no matching width-dependent upper bound. The Top-k trace-width theorem assumes additive shallow branches, while the polyhedral remainder theorem is scalar. The vector oracle uses a flat, normal-constant model. Large-L statements hold for fixed expert and geometry constants; the trace oracle requires an admissible trace class and enough normal space. The results concern fixed-function fidelity, not training dynamics, distribution shift, unchanged FLOPs, or learned-model accuracy. Automated checks supplement the proofs but are neither formal verification nor independent mathematical review.

Conclusion. Boundary agreement does not ensure near-optimal stabilization. Weighted rigidity exposes a transition requirement: in a fixed, simple Top-1 model, two successive ReLUs attain a fidelity rate requiring logarithmically growing shallow width. Exact one-dimensional repair identifies a constant-overhead regime; Top-k and polyhedral results distinguish trace complexity from unavoidable geometric cost.

AI USE STATEMENT

Large language models were used extensively for research ideation, theorem formulation, proof development and revision, literature discovery, code generation, tool-based symbolic and numerical checks, figure preparation, and manuscript drafting and editing. This includes the quantitative differential certificate and width-growth corollary. The checks are automated diagnostics, not proof-assistant formalization or independent human validation. The submitting human author(s) are responsible for reviewing the mathematical claims and for the final submission.

REPRODUCIBILITY STATEMENT

Complete proofs follow the references. The supplementary code runs on CPU without external datasets or network access; the executed environment and diagnostic outputs are recorded in `results/verification.json`, `results/certificate_verification.json`, and the CSV files. All reported checks were executed for this manuscript; no external datasets, checkpoints, or training runs are used. The source, finite-sample QP diagnostic, explicit model evaluators, claim-scope audit, and build instructions are supplied together.

REFERENCES

Raman Arora, Amitabh Basu, Poorya Mianjy, and Anirbit Mukherjee. Understanding deep neural networks with rectified linear units. In *International Conference on Learning Representations*, 2018. https://arxiv.org/abs/1611.01491.

Stefano Buccheri, Tommaso Leonori, and Julio D. Rossi. The best approximation of a given function in L^2-norm by Lipschitz functions with gradient constraint. *arXiv preprint arXiv:2307.12895*, 2023. https://arxiv.org/abs/2307.12895.

Steffen Dereich, Arnulf Jentzen, and Sebastian Kassing. On the existence of minimizers in shallow residual ReLU neural network optimization landscapes. *SIAM Journal on Numerical Analysis*, 62(6):2640–2666, 2024. https://doi.org/10.1137/23M1556241. Theorem 1.1 is also available in https://arxiv.org/abs/2302.14690, version 2.

William Fedus, Barret Zoph, and Noam Shazeer. Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. https://jmlr.org/papers/v23/21-0998.html.

Hussein Hazimeh, Zhe Zhao, Aakanksha Chowdhery, Maheswaran Sathiamoorthy, Yihua Chen, Rahul Mazumder, Lichan Hong, and Ed H. Chi. DSelect-k: Differentiable selection in the mixture of experts with applications to multi-task learning. In *Advances in Neural Information Processing Systems*, 2021. https://arxiv.org/abs/2106.03760.

Tho Tran Huu, Huu-Tuan Nguyen, Thien-Hai Nguyen, Nhat-Tri Ho, Viet-Hoang Tran, Tho Quan, and Tan Minh Nguyen. Geometric and stochastic analysis of discontinuities in sparse mixture-of-experts. *arXiv preprint arXiv:2606.19036*, 2026. https://arxiv.org/abs/2606.19036.

Shibal Ibrahim, Wenyu Chen, Hussein Hazimeh, Natalia Ponomareva, Zhe Zhao, and Rahul Mazumder. COMET: Learning cardinality constrained mixture of experts with trees and local search. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023. https://arxiv.org/abs/2306.02824.

Vladimir Ya. Lin and Allan Pinkus. Fundamentality of ridge functions. *Journal of Approximation Theory*, 75:295–311, 1993. https://pinkus.net.technion.ac.il/files/2021/02/lpn.pdf.

E. J. McShane. Extension of range of functions. *Bulletin of the American Mathematical Society*, 40(12):837–842, 1934. https://doi.org/10.1090/S0002-9904-1934-05978-0.

Kilian Rueß, Gennadiy Averkov, Florestan Brunck, Moritz Grillo, Christoph Hertrich, Georg Loho, Jack Stade, Moritz Stargalla, Matthew Sun, and Martin Winter. Shallower ReLU network representations via exact linear algebra. *arXiv preprint arXiv:2607.21651*, 2026. https://arxiv.org/abs/2607.21651.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. https://arxiv.org/abs/1701.06538.

N. Sukumar and Ankit Srivastava. Exact imposition of boundary conditions with distance functions in physics-informed deep neural networks. *Computer Methods in Applied Mechanics and Engineering*, 2022. https://arxiv.org/abs/2104.08426.

Ziteng Wang, Jun Zhu, and Jianfei Chen. ReMoE: Fully differentiable mixture-of-experts with ReLU routing. *arXiv preprint arXiv:2412.14711*, version 2, 2025. https://arxiv.org/abs/2412.14711.

A WEIGHTED NORMAL-SLICE PROJECTION AND RIGIDITY

We first establish the projection property without appealing to existence or closure of an infinite-dimensional optimization class.

Lemma 8 (A weighted vector step)**.** *Fix $a, b \in \mathbb{R}^m$, positive weights ρ_-, ρ_+, and $J = \|b - a\|$. Put*

$$r = \frac{\sqrt{\rho_+}}{\sqrt{\rho_-} + \sqrt{\rho_+}} J, \quad u = (b-a)/J, \quad p(s) = a + u \operatorname{clip}(s + r, 0, J).$$

For $J = 0$, set $p = a$. Let $f(s) = a$ for $s < 0$ and $f(s) = b$ for $s > 0$, and let $w(s) = \rho_-$ or ρ_+ accordingly. If $R \geq \max\{r, J - r\}$, every 1-Lipschitz $h : [-R, R] \to \mathbb{R}^m$ satisfies

$$\int_{-R}^{R} w \|h - f\|^2 \, \mathrm{d}s \geq \frac{\rho_- r^3 + \rho_+ (J-r)^3}{3} + \int_{-R}^{R} w \|h - p\|^2 \, \mathrm{d}s. \tag{31}$$

In particular, p is the unique minimizer up to measure-zero equality.

Proof. The assertion is immediate when $J = 0$. Otherwise, on the ramp interval $[-r, J - r]$ put $q(s) = \langle u, h(s) - p(s) \rangle$. For $s < t$ in this interval,

$$q(t) - q(s) = \langle u, h(t) - h(s) \rangle - (t - s) \leq 0,$$

so q is nonincreasing. Since $p - f$ vanishes outside the ramp and is parallel to u inside,

$$\begin{aligned} I &= \int_{-R}^{R} w \langle p - f, h - p \rangle \, \mathrm{d}s \\ &= \rho_- \int_{-r}^{0} (s + r) q(s) \, \mathrm{d}s + \rho_+ \int_{0}^{J-r} (s - (J - r)) q(s) \, \mathrm{d}s. \end{aligned}$$

On the first interval $q(s) \geq q(0)$ and the coefficient is nonnegative. On the second $q(s) \leq q(0)$ and the coefficient is nonpositive. Consequently,

$$I \geq \frac{q(0)}{2} \big(\rho_- r^2 - \rho_+ (J - r)^2 \big) = 0.$$

The last equality is exactly the definition of r. Expand $\|h - f\|^2 = \|p - f\|^2 + \|h - p\|^2 + 2\langle p - f, h - p \rangle$ and integrate. Direct integration of the two ramps gives the first term in (31). Strict positivity of the weights gives uniqueness in L^2; continuous representatives therefore coincide everywhere. □

The common value $c_* = p(0)$ also minimizes the cubic trace functional. To see this independently, project any proposed c onto the segment $[a, b]$. Euclidean projection onto a convex segment cannot increase its distance to either endpoint. Writing $c = a + ur'$, it remains to minimize

$$\rho_-(r')^3 + \rho_+(J - r')^3, \qquad 0 \leq r' \leq J.$$

Its derivative vanishes at $\rho_-(r')^2 = \rho_+(J - r')^2$. The minimum is

$$\frac{\rho_- \rho_+}{(\sqrt{\rho_-} + \sqrt{\rho_+})^2} J^3. \tag{32}$$

This verifies both the coefficient and the density asymmetry in the common value.

A.1 PROOF OF THEOREM 1

For each y, define $H_y(s) = G(y, s/L)$. Normal restriction and rescaling give $\|H_y(s) - H_y(t)\| \leq |s - t|$. The physical slab $|t| \leq R/L$ lies in the domain. The risk outside this slab is nonnegative, and changing variables inside it gives

$$L\mathcal{R}(G; F) \geq \int_Y \int_{-R}^{R} \rho_{\operatorname{sgn} s}(y) \|H_y(s) - F(y, s/L)\|^2 \, \mathrm{d}s \, \mathrm{d}y.$$

Apply Lemma 8 at each y. Its cubic term integrates to (5) by (32). This proves (6). All integrands are measurable and nonnegative; boundedness of the data and finite measure make the displayed profile terms integrable. The convergence assertion follows because the nonnegative squared-distance term is bounded above by $L\mathcal{R} - \mathcal{S}$.

A.2 Architecture transfer in general form

For any class $\mathcal{A}_L$ of repaired functions, let

$$\widetilde{\mathcal{A}}_{L,R} = \{(y,s) \mapsto G(y, s/L) : G \in \mathcal{A}_L,\ \mathrm{Lip}(G) \le L\}$$

over $Y \times [-R, R]$. With $\|\cdot\|_w$ denoting the weighted L^2 norm,

$$\inf_{G \in \mathcal{A}_L,\, \mathrm{Lip}(G) \le L} L\mathcal{R}(G; F) \ge \mathcal{S} + \inf_{H \in \widetilde{\mathcal{A}}_{L,R}} \|H - P\|_w^2 . \tag{33}$$

The class may depend on L. In a lower-bound argument it is valid to enlarge the class on the right or restrict its squared-error integral to a subdomain. No global feasibility of P is needed. This last point matters: tangential variation can prevent the normal-slice optimum from obeying the original isotropic Lipschitz constraint exactly.

B Complete proof of the quantitative depth separation

B.1 The forced profile is a fundamental solution

For (7), the weights in (3) are $\rho_- = \rho_+ = 1/4$, $J(y) = 1 + \sigma(y)$, and $r = J/2$. On the negative half-plane the profile is (14). Substituting $u = y/2$ and $v = s + 1/2$ gives

$$p(u, v) = \sigma(v + \sigma(u)) = \max\{0, v, v + u\}.$$

Its first derivative in u, almost everywhere, is $\partial_u p = \mathbf{1}_{\{u>0\}} \mathbf{1}_{\{v+u>0\}}$. Consequently, as distributions on $\mathbb{R}^2$,

$$\partial_v \partial_u p = \mathbf{1}_{\{u>0\}} \delta(v + u), \qquad (\partial_u - \partial_v)\partial_v \partial_u p = \delta(u)\delta(v + u) = \delta_{(0,0)}. \tag{34}$$

Here $\delta(v + u)$ is normalized by integration in v, so the final point mass has coefficient exactly one. The derivative $\partial_u - \partial_v$ is tangent to $v + u = 0$ and differentiates only the half-line cutoff. This proves (15) and identifies the endpoint of the gradient-jump ray as the source of the obstruction.

B.2 An explicit local approximation certificate

For a vector $z \in \mathbb{R}^2$, write $D_z = z_1 \partial_u + z_2 \partial_v$, and let B_a be the open Euclidean disk of radius a centered at the origin.

Lemma 9 (Quantitative ridge certificate)**.** *Let $Q \ge 0$ be an integer and $a > 0$. For every affine function plus at most Q continuous univariate ridge functions,*

$$q(x) = \ell(x) + \sum_{j=1}^{Q} g_j(w_j \cdot x + b_j),$$

one has

$$\|p - q\|_{L^2(B_a)} \ge a^2 A_Q, \qquad A_Q = \frac{Q!}{\sqrt{\pi}\, 2^{2Q+1} 3^{Q+4} (3Q + 8)^{Q+3}}. \tag{35}$$

In particular,

$$A_Q \ge A_0 \Lambda_0^{-Q}, \qquad A_0 = \frac{1}{2^{10} 3^4 \sqrt{\pi}}, \quad \Lambda_0 = 4608. \tag{36}$$

There are no bounds on the coefficients, biases, or relative ridge directions.

Proof. Annihilate the candidate. Choose unit vectors $z_j \perp w_j$; if $w_j = 0$, choose any unit vector. A ridge is constant in its z_j direction, hence $D_{z_j} g_j(w_j \cdot x + b_j) = 0$ as a distribution, including when g_j is only continuous. Pad with zero ridges when fewer than Q are present. Let

$$\mathscr{D} = D_{z_1} \cdots D_{z_Q}, \qquad \mathscr{A} = \mathscr{T}\mathscr{D}.$$

Constant-coefficient derivatives commute. Every ridge is killed by one factor, and $\mathscr{T}$ kills the affine part even for $Q = 0$ or 1. Therefore

$$\mathscr{A} q = 0, \qquad \mathscr{A} p = \mathscr{D}\delta_{(0,0)}. \tag{37}$$

Choose a direction uniformly over the ridges. There is a unit vector e with

$$\left|\prod_{j=1}^{Q} z_j \cdot e\right| \geq 2^{-Q}. \tag{38}$$

For $Q = 0$ this is the empty-product identity. Otherwise write $z_j = (\cos\theta_j, \sin\theta_j)$ and $e = (\cos\theta, \sin\theta)$. The trigonometric polynomial $\prod_j \cos(\theta - \theta_j)$ has Fourier coefficient $2^{-Q}\exp(-\mathrm{i}\sum_j \theta_j)$ at frequency Q. The absolute value of a Fourier coefficient is at most the supremum norm, proving (38).

Detect the surviving point mass. Put

$$\varphi_a(x) = (e \cdot x)^Q \left(1 - \frac{\|x\|^2}{a^2}\right)_+^{Q+4}. \tag{39}$$

This compactly supported function is C^{Q+3}: its derivatives through order $Q+3$ vanish at the circle $\|x\| = a$. At the origin its degree-Q Taylor part is $(e \cdot x)^Q$, so

$$|(\mathscr{D}\varphi_a)(0)| = Q!\left|\prod_j z_j \cdot e\right| \geq Q!2^{-Q}.$$

Pairing (37) with φ_a gives

$$Q!2^{-Q} \leq \left|\int_{B_a} (p-q)\mathscr{A}\varphi_a \,\mathrm{d}x\right| \leq \|p-q\|_{L^2(B_a)} \|\mathscr{A}\varphi_a\|_{L^2(B_a)}. \tag{40}$$

The distributional pairing is valid by smooth compactly supported approximation in C^{Q+3}; equivalently one can integrate the indicated weak derivatives. The zero extension creates no boundary distribution at the orders used here.

Bound the test norm explicitly. First take $a = 1$. Inside the unit disk, φ_1 is a polynomial of degree at most $d_Q = 3Q + 8$. Its sum of absolute monomial coefficients is at most

$$(|e_1| + |e_2|)^Q (1+1+1)^{Q+4} \leq 2^{Q/2} 3^{Q+4}.$$

The sum of absolute coefficients of the differential operator $\mathscr{A}$ is at most $2(\sqrt{2})^Q$: $\partial_u\partial_v(\partial_u - \partial_v)$ has coefficient sum two, and $|z_{j1}| + |z_{j2}| \leq \sqrt{2}$. Every derivative of total order $Q+3$ of a degree-d_Q monomial has absolute coefficient at most d_Q^{Q+3}. Since both coordinates have absolute value at most one on the disk,

$$\|\mathscr{A}\varphi_1\|_{L^2(B_1)} \leq \sqrt{\pi}\, 2^{Q+1} 3^{Q+4} (3Q+8)^{Q+3}.$$

The scaling $\varphi_a(x) = a^Q \varphi_1(x/a)$ gives $\|\mathscr{A}\varphi_a\|_2 = a^{-2}\|\mathscr{A}\varphi_1\|_2$. Substitution in (40) proves (35).

Finally, $A_0 = (2^{10}3^4\sqrt{\pi})^{-1}$. For $Q \geq 1$, use $3Q + 8 \leq 8(Q+1)$ and $Q! \geq (Q/\mathrm{e})^Q \geq ((Q+1)/(2\mathrm{e}))^Q$ to obtain

$$A_Q \geq A_0 (192\mathrm{e})^{-Q}(Q+1)^{-3} \geq A_0\, 576^{-Q} 8^{-Q} = A_0\, 4608^{-Q}.$$

The second inequality uses $\mathrm{e} < 3$ and $Q + 1 \leq 2^Q$. The factorial inequality follows by summing $\log j$ and comparing with $\int_1^Q \log x \,\mathrm{d}x$. This proves (36), also for $Q = 0$. □

The certificate applies to the larger continuous-ridge class, so in particular to shallow ReLU sums. Its uniformity handles diverging parameter sequences directly. Best-approximation attainment, such as that established by Dereich et al. (2024, Theorem 1.1) under a continuous convex-support weight, is therefore not a dependency of this lower bound.

B.3 Transfer to repaired MoE risk

Take $R = 2$ in Theorem 1. The physical slab lies in the square for $L \geq 2$, and $J \leq 2$ ensures it contains both ramps. The disk $B_{1/4}$ in (u, v) coordinates maps under $(y, s) = (2u, v - 1/2)$ into this slab's negative normal half. A width-Q negative branch remains a width-Q ridge sum after substituting $(y, t) = (2u, (v - 1/2)/L)$.

On this set the density is $1/4$ and the coordinate Jacobian is two. Thus Theorem 1 and Lemma 9 give the sharper estimate

$$L\mathcal{R}(G; F) \geq \mathcal{S} + \tfrac{1}{2} \|p - q\|^2_{L^2(B_{1/4})} \geq \mathcal{S} + \frac{A_Q^2}{512} \geq \mathcal{S} + c_0 \Lambda_0^{-2Q}, \tag{41}$$

since $A_0^2/512 = (2^{29}3^8\pi)^{-1} = c_0$. Normal rescaling may enlarge tangential derivatives; imposing no restrictions on them in the approximation class only weakens the lower bound.

The baseline coefficient is

$$\mathcal{S} = \frac{1}{48} \int_{-1}^{1} (1 + \sigma(y))^3 \, \mathrm{d}y = \frac{1}{48} \left(1 + \frac{15}{4}\right) = \frac{19}{192} = S_0.$$

This proves (9), and (35) with $a = 1/4$ also proves (16). The exact constants are conservative; their role is to make the dependence on Q explicit.

B.4 Explicit two-stage upper bound

Evaluate $r_y = \sigma(y)$ once. Set $h = (1 + r_y)/2$. On the negative branch evaluate $\sigma(h + \lambda t)$; on the positive branch evaluate $\sigma(h - \lambda t)$ and subtract it from $1 + r_y$. Thus either selected branch performs exactly two ReLU evaluations in two nonlinear stages. Affine skip connections carry t and r_y as required. No computation of another expert *module* is invoked; the copied feature calculation is nevertheless charged. The reference evaluator dispatches array inputs by masks rather than eagerly computing both branches with `np.where`; the latter would perform a third ReLU evaluation per input. The accounting tests count elementwise activation evaluations.

On $t < 0$, the output is either zero or $h + \lambda t$. On $t > 0$, it is either $1 + \sigma(y)$ or $h + \lambda t$. Both branches agree at $t = 0$. The gradients on full-dimensional affine pieces belong to

$$(0, 0), \quad (1, 0), \quad (0, \lambda), \quad (1/2, \lambda).$$

Their norms do not exceed L for $\lambda = \sqrt{L^2 - 1/4}$ and $L \geq 2$. A segment in the square splits into finitely many affine pieces, and the continuous values agree at the splitting points. Integrating the gradient bound along the segment proves global Lipschitzness, including crossings of $t = 0$ and $y = 0$.

For fixed y, put $J = 1 + \sigma(y)$. Each side is changed only for $|t| < J/(2\lambda)$, which lies inside the domain. Therefore

$$\int_{-1}^{1} |G(y, t) - F(y, t)|^2 \, \mathrm{d}t = 2 \int_0^{J/(2\lambda)} (J/2 - \lambda t)^2 \, \mathrm{d}t = \frac{J^3}{12\lambda}.$$

Multiplying by density $1/4$ and integrating in y gives $19/(192\lambda)$. Theorem 1 gives $\mathcal{R}_L^* \geq 19/(192L)$, which establishes the ratio bound. Since $J/2 \leq 1$, the output agrees with the original whenever $|t| \geq 1/\lambda$. Finally, the lower and upper bounds imply $L\mathcal{R}_L^* \to 19/192$. For fixed Q, (9) gives a limiting relative gap of at least $1 + (192/19)c_0\Lambda_0^{-2Q}$. Corollary 3 follows from the finite-L bounds, so its width may itself vary with L.

C Finite-L trace oracle: scalar and vector proofs

C.1 Projection identities and lower bound

Projection onto a fixed Euclidean ball is nonexpansive:

$$\|P_r(v) - P_r(w)\| \leq \|v - w\|, \qquad \|P_r(v) - P_s(v)\| \leq |r - s|. \tag{42}$$

For completeness, the first assertion follows from the variational inequality for closest-point projection. If $p = P_r(v)$ and $q = P_r(w)$, then $\langle v - p, q - p\rangle \le 0$ and $\langle w - q, p - q\rangle \le 0$; adding gives $\|p - q\|^2 \le \langle p - q, v - w\rangle$. Cauchy–Schwarz finishes the proof. The second follows directly from radial truncation of the length of v.

Let a feasible repair have boundary trace h. Normal Lipschitzness gives, for either side,

$$\|G(y, \pm t) - f_\pm(y)\| \ge (\|h(y) - f_\pm(y)\| - Lt)_+ .$$

Because $LT \ge \lambda T \ge M$, the full positive part is contained in the normal interval. Integrating its square yields $\mathcal{R}(G; F) \ge \Psi(h)/L$. Taking infima gives the lower half of (20).

For the proposed repair, $P_r(v)$ changes only the length of v. Hence

$$\|G_h - f_\pm\|^2 = (\|f_\pm - h\| - \lambda|t|)_+^2 .$$

Integrating again gives $\mathcal{R}(G_h; F) = \Psi(h)/\lambda$, provided that G_h is feasible. If Ψ_* is not attained, apply the construction to a minimizing sequence; attainment is not needed for the infimum bound.

C.2 Scalar Lipschitz bound

For scalar outputs,

$$G_h(y, t) = \mathrm{clip}(f_\pm(y), h(y) - \lambda|t|, h(y) + \lambda|t|).$$

The output equals either $f_\pm$ or one of the two boundary functions. On each side of $t = 0$, their respective Lipschitz constants are at most B and $\sqrt{K^2 + \lambda^2}$. Finite maxima and minima of scalar L-Lipschitz functions are L-Lipschitz: compare their values at two points using the same index that attains the maximum at one point. Thus each branch is L-Lipschitz for the stated λ. Both branches equal h at $t = 0$. A line segment crosses that plane at most once, proving the global bound.

C.3 Vector Jacobian bound

The map (21) is locally Lipschitz by (42); a coarse bound is $B + 2K + \lambda$. It suffices to bound its almost-everywhere Jacobian on the convex half-domains and then integrate along segments, or equivalently use the corresponding $W^{1,\infty}$ representative.

Inside the active clipping set put $v = f - h$, $u = v/\|v\|$, and $\alpha = \lambda|t|/\|v\| \in (0, 1)$. For a tangential input direction η,

$$\begin{aligned} D_y G\,\eta &= Dh\,\eta + \alpha(I - uu^\top)(Df - Dh)\eta, \\ uu^\top D_y G\,\eta &= uu^\top Dh\,\eta, \\ (I - uu^\top) D_y G\,\eta &= (I - uu^\top)\big((1 - \alpha)Dh + \alpha Df\big)\eta. \end{aligned}$$

The radial component has norm at most $K\|\eta\|$, while the orthogonal component has norm at most $B_0\|\eta\|$. The normal derivative is $\lambda\,\mathrm{sgn}(t)u$. For a joint input direction (η, τ), the two orthogonal output components are therefore bounded by

$$K\|\eta\| + \lambda|\tau|, \qquad B_0\|\eta\| .$$

The output norm is bounded by the norm of the matrix $M_\lambda = \left(\begin{smallmatrix} K & \lambda \\ B_0 & 0 \end{smallmatrix}\right)$ applied to $(\|\eta\|, |\tau|)$.

Write $A = L^2 - B_0^2$. With (19), $\lambda^2 = L^2(A - K^2)/A$. The diagonal terms of

$$L^2 I - M_\lambda^\top M_\lambda = \begin{pmatrix} A - K^2 & -K\lambda \\ -K\lambda & L^2 - \lambda^2 \end{pmatrix}$$

are nonnegative for $L \ge 2B_0$, and its determinant equals

$$(A - K^2)\frac{L^2 K^2}{A} - K^2\frac{L^2(A - K^2)}{A} = 0.$$

Hence this matrix is positive semidefinite and $\|M_\lambda\|_{\mathrm{op}} \le L$. The case $K = B_0 = 0$ is immediate for $L > 0$. Outside active clipping, $G = f$ and the derivative norm is at most B. Equality sets are handled by the local Lipschitz property and almost-everywhere bounds. Finally, the two side traces agree, so the global Lipschitz bound follows.

C.4 Scope of the oracle interpretation

Theorem 4 compares to every stable function with a trace in $\mathcal{H}$, not just to the chosen clipping family. Its hypotheses explicitly include enough space for both transitions and uniform trace regularity. For fixed B, K, expanding L/λ gives $1 + K^2/(2L^2) + O(L^{-4})$. When $K = 0$, $\lambda = L$ and the upper and lower bounds agree. This does not imply a free constant trace is appropriate for arbitrary experts: $\Psi(h)$ can be much larger than its unrestricted pointwise minimum.

D Trace approximation and a one-sided learning objective

Proposition 10 (Cubic trace excess)**.** *Let $m = (a + b)/2$, $J = \|a - b\|$, and $e = h - m$. Define $\Phi(h) = \|a - h\|^3 + \|b - h\|^3$. Then*

$$\max\{(3/2)J\|e\|^2,\ 2\|e\|^3\} \le \Phi(h) - J^3/4 \le 3J\|e\|^2 + 2\|e\|^3. \tag{43}$$

Proof. Put $v = (a - b)/2$. Then $\Phi(m + e) = \|v - e\|^3 + \|v + e\|^3$ and $\Phi(m) = 2\|v\|^3 = J^3/4$. The Hessian of $\|z\|^3$ has eigenvalues between $3\|z\|$ and $6\|z\|$; it extends continuously as zero at the origin. The gradient of the sum at $e = 0$ vanishes. Along te, the sum of endpoint distances lies between J and $J + 2t\|e\|$. Taylor's integral formula therefore gives

$$\frac{3}{2}J\|e\|^2 \le \Phi(m + e) - \Phi(m) \le 3J\|e\|^2 + 2\|e\|^3.$$

For the additional cubic lower bound, convexity of $x^{3/2}$ on $[0, \infty)$ and the parallelogram identity give

$$\|v - e\|^3 + \|v + e\|^3 \ge 2(\|v\|^2 + \|e\|^2)^{3/2} \ge 2\|v\|^3 + 2\|e\|^3.$$

Combining proves the result. □

For equal densities, (43) bounds $\Psi(h) - \mathcal{S}$ after multiplication by $\rho/3$ and integration. If $\|h - m\| \le \eta J$ pointwise and $\mathcal{S} > 0$, it follows that

$$\frac{\Psi(h)}{\mathcal{S}} \le 1 + 12\eta^2 + 8\eta^3. \tag{44}$$

Together with Theorem 4, this yields a finite-L risk ratio upper bound $(L/\lambda)(1 + 12\eta^2 + 8\eta^3)$ against the unrestricted lower bound $\mathcal{S}/L$. When $J = 0$, the exact excess is $2\|h - a\|^3$ before weighting. Thus an unsuitable shared zero function can create a loss even where the original branches already agree.

D.1 Boundary-only fitting with one side per sample

For multiple interfaces, define the unnormalized objective

$$\mathcal{B}(\theta) = \sum_{i<j} \int_{\Sigma_{ij}} \left[\rho_i \|E_i - H_\theta\|^3 + \rho_j \|E_j - H_\theta\|^3\right] \mathrm{d}S. \tag{45}$$

Here E_i, E_j may denote complete original branch outputs. Let a pair (i, j, z) be sampled with density $q_{ij}(z) > 0$ with respect to the disjoint union of interface surface measures, normalized to total mass one. Sample one side $s \in \{i, j\}$ with conditional probability $\pi_s(i, j, z) > 0$. Then

$$Z_\theta = \frac{\rho_s(z)}{q_{ij}(z)\pi_s(i, j, z)} \|E_s(z) - H_\theta(z)\|^3 \tag{46}$$

is an unbiased estimator of (45): multiply by the sampling probabilities, sum over the two sides, and integrate. Each draw evaluates only that side's selected experts. No statement is made that collecting z, knowing ρ_s, or evaluating the importance weights is free.

If $H_\theta = \Theta\phi$ with fixed features ϕ, the objective is convex in Θ. On a closed convex parameter set of diameter D, suppose the estimator has unbiased subgradients with second moment at most G_0^2. Projected stochastic subgradient descent with step $D/(G_0\sqrt{n})$ gives an averaged iterate satisfying

$$\mathbb{E}[\mathcal{B}(\bar{\theta}_n) - \mathcal{B}(\theta_*)] \le DG_0/\sqrt{n}.$$

Indeed, nonexpansiveness of parameter projection yields the usual squared-distance recursion with decrease $2\eta\langle g_t, \theta_t - \theta_*\rangle$ and increase $\eta^2 \|g_t\|^2$. Take conditional expectations, use convexity, sum, and divide by $2\eta n$. This is a conditional optimization guarantee, not a claim that an arbitrary nonlinear trace network can be globally trained.

E Complete Top-k width construction and lower bound

Fix $k, W \geq 1$, $a > 0$, $N = (k+1)W$, $\tau = 1/(N+1)$, and $\xi_r = r\tau$. For block $j = 1, \ldots, W$, write $q_j = (k+1)(j-1)$ and $s_j = (-1)^{j-1}$. Define $k+1$ scalar experts, all independent of t:

$$\begin{aligned} U(y) &= 2ka \sum_{j=1}^{W} s_j (y - \xi_{q_j+1})_+, \\ V(y) &= 3ka\tau + 2ka \sum_{j=1}^{W} s_j (y - \xi_{q_j+2})_+, \\ C_r(y) &= ka \sum_{j=1}^{W} s_j (y - \xi_{q_j+r+2})_+, \qquad r = 1, \ldots, k-1. \end{aligned} \tag{47}$$

Each has exactly W potentially nonzero hidden coefficients. Give every common expert score 2, and give U, V the scores $-t, t$. On $[-1, 1]$, all $k-1$ common experts are selected, along with U for $t < 0$ or V for $t > 0$. The router is affine; the tie at zero may be resolved arbitrarily. The two mean outputs are

$$F_- = (U + \textstyle\sum C_r)/k, \qquad F_+ = (V + \sum C_r)/k.$$

E.1 Jump and midpoint structure

The jump $J = F_+ - F_- = (V - U)/k$ starts at $3a\tau$. Within block j, its derivative equals $-2as_j$ between that block's first two thresholds and equals zero afterwards until the next block. Thus each odd block reduces the value from $3a\tau$ to $a\tau$, and each even block restores it; the transitions are linear and stay in this interval. Therefore

$$a\tau \leq J(y) \leq 3a\tau \quad \text{for all } y \in [0, 1].$$

The midpoint is $m = (U + V + 2\sum C_r)/(2k)$. Every threshold contributes a slope jump as_j from exactly one expert in its block. Consequently m has exactly N distinct interior knots and all of their slope jumps have magnitude a.

E.2 A missing-knot estimate

Let b be any scalar continuous piecewise-affine function with at most P knots. The N intervals $I_r = [\xi_r - \tau/2, \xi_r + \tau/2]$ have disjoint interiors. At most P contain a knot of b in their interior. On each remaining interval, b is affine almost everywhere, whereas m is an affine function plus $\pm a(y - \xi_r)_+$.

The least-squares affine projection of ax_+ on $[-h, h]$ is $ax/2 + ah/4$. This follows either by two normal equations or by writing $x_+ = (x + |x|)/2$ and projecting the even part onto constants. The squared residual integrates to $a^2h^3/24$. Thus

$$\int_0^1 |b - m|^2 \, \mathrm{d}y \geq \frac{(N - P)_+ a^2 \tau^3}{192}. \tag{48}$$

A knot lying exactly at an interval endpoint does not invalidate this argument, since it leaves the function affine on that interval's interior.

E.3 Truncated trace penalty without bounded candidate outputs

For $r \geq 0$ define

$$T_L(r) = \int_0^L (r - s)_+^2 \, \mathrm{d}s = \frac{r^3 - (r - L)_+^3}{3}.$$

Fix endpoints $A < B$ with $J = B - A \le L$, and let $\varphi(c) = T_L(|c - A|) + T_L(|c - B|)$. The even function $x \mapsto T_L(|x|)$ is twice continuously differentiable and has second derivative $2\min\{|x|, L\}$. By the triangle inequality,

$$\min\{|c - A|, L\} + \min\{|c - B|, L\} \ge J.$$

Indeed, if neither term is truncated their sum is at least J; if either is truncated, that term alone is at least J. Hence $\varphi'' \ge 2J$. The function is symmetric about $m = (A + B)/2$, with $\varphi(m) = J^3/12$ and $\varphi'(m) = 0$. Taylor's formula gives

$$\varphi(c) \ge J^3/12 + J(c - m)^2. \tag{49}$$

Now take any globally L-Lipschitz repair of the constructed model and write its common boundary value as $b(y) = G(y, 0)$. Its risk has density $1/2$ on $[0, 1] \times [-1, 1]$. Normal Lipschitzness and the substitution $s = Lt$ yield

$$L\mathcal{R}(G; F) \ge \frac{1}{2} \int_0^1 \left[T_L(|b - F_-|) + T_L(|b - F_+|)\right] \mathrm{d}y.$$

For $L \ge \sup J$, apply (49) to obtain (26), with $\mathcal{S} = \int J^3/24$. Notice that no bound on b was imposed.

E.4 Counting the repaired trace and finishing the proof

On either side, the repaired output is an affine combination of shallow experts whose total evaluated width is at most P. Restricting to $t = 0$ gives at most P one-dimensional ReLU knots. Continuity of the global repair identifies the two side traces, so (48) applies. Therefore

$$L\mathcal{R}(G; F) \ge \mathcal{S} + \frac{(N - P)_+ a^3 \tau^4}{384}.$$

Since $\mathcal{S} \le 27a^3\tau^3/24$, division yields (24). Finally, the midpoint m is Lipschitz with a finite constant for fixed k, W, a. The scalar oracle with trace m supplies a repair of risk $\mathcal{S}/\sqrt{L^2 - \mathrm{Lip}(m)^2}$ for large L. Together with the normal lower bound, this proves $\mathcal{R}_L^* = \mathcal{S}/L + o(L^{-1})$.

The theorem does not identify a tight sufficient shallow width. In particular, having enough knots on the boundary does not eliminate the transition obstruction of Theorem 2. The factor 432 is a conservative explicit constant, not asserted to be optimal.

F Scalar boundary envelopes and the exact interval problem

This section supplies a general scalar reduction used in the polyhedral upper bound, and then its finite-dimensional specialization. Unlike the density-based asymptotics, the interval reduction permits arbitrary finite Borel input measures, including atoms at routing endpoints.

F.1 Prescribing a common trace

Let Ω be compact and convex, with a finite polyhedral partition and scalar branches of Lipschitz constant at most $B \le L$. Let Σ be the nonempty closed union of its internal interfaces. A continuous repair has a single trace on this set. Fix an L-Lipschitz trace $b : \Sigma \to \mathbb{R}$ and define

$$\ell_b(x) = \sup_{z \in \Sigma} \{b(z) - L\,\|x - z\|\}, \qquad u_b(x) = \inf_{z \in \Sigma} \{b(z) + L\,\|x - z\|\}. \tag{50}$$

The same construction is valid if additional closed boundary pieces carry prescribed values; this will include both external endpoints in the interval problem.

Proposition 11 (Exact fixed-trace repair)**.** *The piecewise-defined function*

$$G_b(x) = \mathrm{clip}(E_i(x), \ell_b(x), u_b(x)), \qquad x \in C_i, \tag{51}$$

is globally L-Lipschitz, equals b on Σ, and minimizes squared fidelity error among all L-Lipschitz functions with that trace. Consequently,

$$\mathcal{R}_L^*(F) = \inf_{b \in \mathrm{Lip}_L(\Sigma)} \int_\Omega \mathrm{dist}\big(F(x), [\ell_b(x), u_b(x)]\big)^2 \,\mathrm{d}\mu(x). \tag{52}$$

The objective on the right is convex in b on its convex feasible set.

Proof. For $z, w \in \Sigma$, Lipschitzness of b and the triangle inequality give

$$b(z) - L\,\|x - z\| \le b(w) + L\,\|x - w\|\,.$$

Taking a supremum on the left and an infimum on the right proves $\ell_b \le u_b$. Both envelopes are L-Lipschitz, since they are finite-valued suprema or infima of functions sharing that Lipschitz bound. At $z \in \Sigma$, both equal $b(z)$. These are the classical scalar Lipschitz-extension envelopes (McShane, 1934); the pointwise repair argument below uses both simultaneously.

Every feasible G satisfies $b(z) - L\,\|x - z\| \le G(x) \le b(z) + L\,\|x - z\|$ for all z, so its value belongs to $[\ell_b(x), u_b(x)]$. The closest value in this interval to $F(x)$ is its scalar clipping. Thus (51) is pointwise optimal if it is feasible. On each closed cell, it is a minimum and maximum of L-Lipschitz scalar functions, hence is L-Lipschitz. On every common interface the envelopes collapse to b, so neighboring definitions agree. Any segment in the convex domain can be split into finitely many subsegments in cell closures. Summing the Lipschitz inequalities proves the global bound. If a segment lies in an interface, it already obeys the bound for b; alternatively take limits of generic segments.

These facts prove optimality and (52), including mass on interfaces: the tie-defined value of F is compared to the compulsory value b. Finally, ℓ_b is convex in b, u_b is concave, and

$$\operatorname{dist}(F, [\ell_b, u_b]) = \max\{\ell_b - F, F - u_b, 0\}.$$

This is a nonnegative convex function of b. Squaring and integrating preserve convexity. □

For bounded $F \in [-M, M]$, clipping any candidate to this range does not increase its error or its Lipschitz constant. Hence the infimum can be restricted to traces with $|b| \le M$. The unrestricted optimum is attained: the resulting family of uniformly bounded L-Lipschitz functions is compact in the uniform topology, and risk is continuous under uniform convergence for a finite input measure. This is a function-space existence statement, not a fixed-network-width statement.

F.2 Reduction to adjacent endpoint values

Let $\tau_0 < \cdots < \tau_n$ and $I_j = (\tau_{j-1}, \tau_j)$. Assume $F = E_j$ on I_j and $\operatorname{Lip}(E_j) \le B < L$. Introduce variables $c_0, \ldots, c_n$ for all endpoint values. Their feasibility constraints are exactly

$$|c_j - c_{j-1}| \le L(\tau_j - \tau_{j-1}), \qquad 1 \le j \le n. \tag{53}$$

They imply all nonadjacent endpoint constraints by the triangle inequality.

For $x \in I_j$, all other endpoint cones are redundant. For example, if $\tau_r \le \tau_{j-1}$, then

$$c_r - L(x - \tau_r) \le c_{j-1} - L(x - \tau_{j-1}),$$

and the corresponding upper cone is no more restrictive than that from τ_{j-1}. The proof for right endpoints is identical. Thus the relevant envelopes are precisely

$$\ell_j(x) = \max\{c_{j-1} - L(x - \tau_{j-1}),\ c_j - L(\tau_j - x)\}, \tag{54}$$

$$u_j(x) = \min\{c_{j-1} + L(x - \tau_{j-1}),\ c_j + L(\tau_j - x)\}. \tag{55}$$

The exact finite-dimensional problem is

$$\begin{aligned} \min_{c} \quad & \sum_{j=1}^{n} \int_{I_j} \left[(\ell_j - E_j)_+^2 + (E_j - u_j)_+^2\right] \mathrm{d}\mu \\ & + \sum_{r=0}^{n} \mu(\{\tau_r\}) |c_r - F(\tau_r)|^2, \\ \text{subject to} \quad & (53). \end{aligned} \tag{56}$$

The atom term is necessary when endpoints have positive probability. It uses the original router's specified tie convention. For atom-free measures it vanishes. Proposition 11 proves equivalence with the unrestricted Lipschitz optimum; both summands in the integrand cannot be positive simultaneously because $\ell_j \le u_j$. Bounded-range clipping permits restriction to a compact box for c, so a minimizer exists. The original experts are bounded on the finite compact intervals.

F.3 CONSTANT ADDITIVE SHALLOW WIDTH

A scalar one-dimensional shallow expert with W ReLUs has at most W knots. Clipping it between (55) can introduce only the two envelope corners and intersections with the four lines defining the envelopes. Each line has slope $+L$ or $-L$. For $x_2 > x_1$,

$$E_j(x_2) - Lx_2 < E_j(x_1) - Lx_1, \qquad E_j(x_2) + Lx_2 > E_j(x_1) + Lx_1,$$

using $B < L$. Each line therefore intersects the expert graph at most once. There are at most $W + 2 + 4 = W + 6$ knots in the repaired interval branch.

Every continuous piecewise-affine scalar function with knots $\zeta_1 < \cdots < \zeta_q$ is expressible as

$$a + bx + \sum_{r=1}^{q} \gamma_r (x - \zeta_r)_+,$$

where γ_r is its slope jump. Extend the repaired interval branch affinely outside the interval; this introduces no extra knots. Hence the exact repair has at most $W + 6$ ReLUs per selected expert with an affine skip. Without such a skip, encode the affine part using two additional ReLUs, giving the conservative bound $W + 8$. No other original expert is evaluated. Strict $B < L$ is used only for this knot count, not for Proposition 11.

F.4 FINITE-SAMPLE QUADRATIC PROGRAM

For a finite empirical measure, introduce a variable v_q for every sample x_q in the interior of I_j. Impose

$$|v_q - c_{j-1}| \le L(x_q - \tau_{j-1}), \qquad |v_q - c_j| \le L(\tau_j - x_q),$$

along with (53), and minimize $\sum_q w_q (v_q - F(x_q))^2$, including endpoint samples as their c_j values. This is a convex quadratic program with $O(n+n_{\text{sample}})$ variables and constraints. Pairwise constraints between all samples are unnecessary for the optimum value: at fixed c, the minimizing v_q is the interval clipping of $E_j(x_q)$, and Proposition 11 supplies a common globally feasible extension attaining all those values. An arbitrary nonoptimal feasible vector (v_q) need not itself have that extension; the assertion concerns optimal values and the clipped reconstruction.

G SCALAR POLYHEDRAL EXPANSION WITH A SHARP-ORDER REMAINDER

We prove Theorem 7 with the assumptions made explicit. The domain is a compact convex full-dimensional polytope. Its fixed finite partition is a polyhedral complex; an affine routing partition can be refined to such a complex without changing its output. On every closed full-dimensional cell, the scalar branch satisfies $|E_i| \le M$ and $\mathrm{Lip}(E_i) \le B$. The probability measure has density ρ_i on that cell, with

$$0 < \rho_{\min} \le \rho_i \le \rho_{\max}, \qquad \mathrm{Lip}(\rho_i) \le D.$$

Constants may depend on these bounds, the dimension, and the fixed partition, including its face angles and separations. They are not uniform over arbitrarily degenerate routers. Let Σ be the closed union of internal facets and count every facet once in the surface integral. Artificial refinement interfaces with identical branches have zero surface contribution. If Σ is empty, F is already B-Lipschitz and the assertion is immediate for $L \ge B$.

G.1 A FINITE POLYHEDRAL GEOMETRY OBSERVATION

For $d \ge 2$, let $\mathcal{K}$ be the union of all faces of dimension at most $d - 2$, including facet ends at the external boundary. For sufficiently small $r > 0$, the following statements hold with constants depending only on the fixed complex.

The r-neighborhood of $\mathcal{K}$ in Ω has volume $O(r^2)$, and its intersection with all facets has total surface measure $O(r)$. Moreover, for a sufficiently large fixed A, normal prisms of depth r about internal facets trimmed to distance at least Ar from $\mathcal{K}$ are contained in their two incident cells and have pairwise disjoint interiors. A point within distance r of Σ but farther than a sufficiently large multiple

of r from $\mathcal{K}$ sees only one incident facet in any fixed-multiple-r neighborhood, and its nearest-point projection is in that facet's relative interior.

Here is a justification that also identifies the dependence of constants. A bounded face of dimension $q \le d-2$ has a tubular neighborhood of volume at most a constant times r^{d-q}, by covering its affine span inside a fixed box and integrating in orthogonal coordinates. The analogous argument in each facet gives $O(r)$ surface measure near a codimension-one face of that facet. Sum over the finite collection. Near an incident pair of distinct affine faces, a displacement of length r can reach the other face only if the starting point is within cr of their intersection; the constant is bounded by the inverses of the finitely many nonzero principal angles. For distinct coplanar facets with disjoint relative interiors, a segment between them must leave the first facet through its boundary; trimming by distance Ar therefore gives the same locality statement without an angle denominator. If two face closures do not meet, their compactness gives a positive separation. Taking the maximum of the finitely many incident constants and then r below all nonincident separations proves the prism and locality statements. In dimension one, the internal interfaces are finitely many separated points; for small r the same prism statements hold without a skeleton or a discarded region.

G.2 Lower bound

Clip a candidate repair to $[-M, M]$, which cannot increase either risk or Lipschitz constant. Fix a trimmed facet point $z \in \Sigma_{ij}$, let $c = G(z)$, and set

$$r_i(z) = |c - E_i(z)|, \qquad r_j(z) = |c - E_j(z)|.$$

Both are at most $2M$. Along the inward unit normal ν_i of side i,

$$|G(z + t\nu_i) - E_i(z + t\nu_i)| \ge \big(r_i - (L+B)t\big)_+.$$

Take normal depth $2M/(L+B)$ and trim the facets by a sufficiently large multiple of $1/L$, so the prisms are disjoint and inside the appropriate cells. On side i, using $\rho_i(z + t\nu_i) \ge \rho_i(z) - Dt$ and increasing L so that this bound remains nonnegative in the prism, its risk is at least

$$\int_0^{2M/(L+B)} [\rho_i(z) - Dt]\big(r_i - (L+B)t\big)_+^2 \, \mathrm{d}t \tag{57}$$

$$= \frac{\rho_i(z) r_i^3}{3(L+B)} - \frac{D r_i^4}{12(L+B)^2}. \tag{58}$$

The second term is uniformly $O(L^{-2})$ per unit facet area. Minimizing the sum of the cubic terms over c gives

$$\min_c \{\rho_i |E_i - c|^3 + \rho_j |E_j - c|^3\} = \frac{\rho_i \rho_j}{(\sqrt{\rho_i} + \sqrt{\rho_j})^2} |E_i - E_j|^3. \tag{59}$$

Indeed, projection onto the endpoint interval reduces the problem to $\min_{0 \le r \le J} \rho_i r^3 + \rho_j (J-r)^3$, whose derivative vanishes when $\rho_i r^2 = \rho_j (J-r)^2$.

Integrate over all trimmed facets. The omitted surface area is $O(1/L)$ and the surface integrand is bounded, so restoring it changes the lower bound by only $O(L^{-2})$. Finally, $(L+B)^{-1} = L^{-1} + O(L^{-2})$. This proves

$$\mathcal{R}_L^*(F) \ge \mathcal{S}(F)/L - O(L^{-2}). \tag{60}$$

G.3 A compatible trace at multiple-face junctions

On each facet define its pointwise optimal trace

$$b_{ij}^*(z) = \frac{\sqrt{\rho_i(z)} E_i(z) + \sqrt{\rho_j(z)} E_j(z)}{\sqrt{\rho_i(z)} + \sqrt{\rho_j(z)}}. \tag{61}$$

It is bounded by M, and has a facetwise Lipschitz constant K_0 depending only on B, D, M, $\rho_{\min}$, and $\rho_{\max}$. Positivity of the density lower bound controls the derivatives of the square roots and of the reciprocal denominator. These facetwise traces need not agree at a junction.

Trim every facet at distance η from $\mathcal{K}$, and denote the union of the retained closed pieces by T_η. Pieces from different facets are separated by at least $c_{\mathrm{geom}}\eta$ for small η, by the geometry observation. The prescribed function b^* on T_η is therefore Lipschitz with constant at most

$$\max\{K_0, 2M/(c_{\mathrm{geom}}\eta)\}.$$

Choose $\eta = D_0/L$, with a fixed sufficiently large D_0, and then choose L large enough so this constant is at most $L/2$. Extend b^* from T_η by

$$\widetilde{b}_L(x) = \inf_{z\in T_\eta} [b^*(z) + (L/2)\,\|x - z\|],$$

and clip the extension to $[-M, M]$. Its restriction b_L to Σ is $L/2$-Lipschitz, bounded, and agrees with b^* on every retained facet. For all large L, every nonempty facet has a retained piece. In dimension one, the prescribed values on the finitely many separated interfaces already have a fixed finite Lipschitz constant; use them directly.

G.4 Locality of the scalar envelopes and upper bound

Apply Proposition 11 with trace b_L and budget L, obtaining a globally feasible G. All source and trace values lie in $[-M, M]$. A boundary point w with $\|x - w\| > 2M/L$ gives a lower cone less than $-M$ and an upper cone greater than M, so it cannot force a change to $F(x)$. In particular,

$$G(x) = F(x) \quad \text{if} \quad \operatorname{dist}(x, \Sigma) > 2M/L. \tag{62}$$

Consider a point in this thin layer but outside a sufficiently large A_0/L-neighborhood of $\mathcal{K}$. The geometry observation and the choice of A_0 ensure that every potentially constraining boundary point belongs to the same facet and to its retained portion. Let z be the orthogonal projection of x onto that facet, $t = \|x - z\|$, and $c = b^*(z)$. For any boundary point w on that facet, write $r = \|w - z\|$. The facetwise Lipschitz bound and orthogonality give

$$b^*(w) - L\,\|x - w\| \le c + K_0 r - L\sqrt{t^2 + r^2} \le c - \lambda_0 t, \quad \lambda_0 = \sqrt{L^2 - K_0^2}.$$

The last inequality follows by minimizing $L\sqrt{t^2 + r^2} - K_0 r$ for $r \ge 0$. Similarly the upper cones are at least $c + \lambda_0 t$. Incorporating the harmless far boundary points, the full envelopes satisfy

$$\ell_{b_L}(x) \le \max\{-M, c - \lambda_0 t\}, \qquad u_{b_L}(x) \ge \min\{M, c + \lambda_0 t\}.$$

Since $E_i(x) \in [-M, M]$, its distance to the allowed interval is at most

$$\big(|E_i(x) - c| - \lambda_0 t\big)_+ \le \big(|E_i(z) - c| - (\lambda_0 - B)t\big)_+. \tag{63}$$

For large L, put $a_L = \lambda_0 - B > 0$. Integrating the right-hand side squared against $\rho_i(z) + Dt$ gives the bound

$$\frac{\rho_i(z)|E_i(z) - c|^3}{3a_L} + \frac{D|E_i(z) - c|^4}{12a_L^2}.$$

The normal integrals may be enlarged to their entire positive support for this upper bound. Summing both sides and integrating over facets, (61) and (59) yield $\mathcal{S}(F)/a_L + O(L^{-2}) = \mathcal{S}(F)/L + O(L^{-2})$.

The remaining affected inputs lie in the A_0/L-neighborhood of $\mathcal{K}$, whose volume is $O(L^{-2})$. Their squared error is at most $4M^2$ and density at most $\rho_{\max}$, so they contribute $O(L^{-2})$. In dimension one there is no discarded region. This proves

$$\mathcal{R}_L^*(F) \le \mathcal{R}(G; F) \le \mathcal{S}(F)/L + O(L^{-2}).$$

Together with (60), it establishes Theorem 7. This proof uses scalar order in the exact-envelope construction. It does not assert the same remainder theorem for general vector-valued polyhedral mixtures.

G.5 THE REMAINDER ORDER CANNOT IN GENERAL BE IMPROVED

On $[-T, T]$ with constant density ρ (take $\rho = 1/(2T)$ for a probability measure), let

$$F(t) = \beta t + (J/2)\operatorname{sgn}(t), \qquad \beta, J > 0.$$

Take $L > \beta$ with $J/[2(L-\beta)] \le T$. Averaging a feasible function G with $-G(-t)$ preserves feasibility and does not increase squared risk, since F is odd. Hence the infimum may be taken over odd repairs, which have $G(0) = 0$ and $|G(t)| \le L|t|$. For positive t, the pointwise closest allowed output to $\beta t + J/2$ is $\min\{Lt, \beta t + J/2\}$. Its odd extension is feasible: its slopes are L and β, and it is continuous. Consequently the exact optimum is

$$\begin{aligned}\mathcal{R}_L^*(F) &= 2\rho \int_0^{J/[2(L-\beta)]} [J/2 - (L-\beta)t]^2 \,\mathrm{d}t = \frac{\rho J^3}{12(L-\beta)} \\ &= \frac{\mathcal{S}}{L} + \frac{\beta \mathcal{S}}{L^2} + O(L^{-3}), \qquad \mathcal{S} = \rho J^3/12. \end{aligned} \tag{64}$$

The nonzero second term rules out a general $o(L^{-2})$ remainder even with a single interface.

H SELECTED-SET-PRESERVING IMPLEMENTATION AND ITS COSTS

H.1 A GENERAL VECTOR REPAIR WITH AN EXPLICIT LIPSCHITZ BUDGET

Let the original vector-valued branches be B-Lipschitz on a finite convex polyhedral partition, let $H : \Omega \to \mathbb{R}^m$ be K-Lipschitz, and let $d_\Sigma(x) = \operatorname{dist}(x, \Sigma)$ for the nonempty internal interface set. For $\lambda > 0$, define

$$G(x) = H(x) + P_{\lambda d_\Sigma(x)}(F(x) - H(x)). \tag{65}$$

The projection identities (42) and the 1-Lipschitz property of distance give, within each cell,

$$\operatorname{Lip}(G) \le K + (B + K) + \lambda = B + 2K + \lambda.$$

At the interface the projection radius is zero, so all branches agree with H. Segment gluing proves the same global bound. Therefore $\lambda = L - B - 2K > 0$ suffices for L-Lipschitz stability. The sharper flat-model constants in Theorem 4 use additional normal invariance not assumed here.

The pointwise error remains exact:

$$\|G(x) - F(x)\|^2 = \big(\|F(x) - H(x)\| - \lambda d_\Sigma(x)\big)_+^2. \tag{66}$$

If a valid bound $\|F - H\| \le M$ is known in advance, the original output is preserved whenever $d_\Sigma \ge M/\lambda$. One can bypass evaluation of H on those inputs after the geometric test. Computing or certifying M is not free and is not supplied by the abstract construction.

H.2 DISTANCE FOR AN AFFINE TOP-k ROUTER

Write the scores as $s_i(x) = a_i^\top x + b_i$, with M_{exp} experts. In the full ambient space, an active-set cell is the intersection of half-spaces $s_i \ge s_j$ for $i \in S, j \notin S$. At an interior point of a full-dimensional nondegenerate cell,

$$d_{\mathrm{amb}}(x) = \min_{i \in S,\, j \notin S} \frac{s_i(x) - s_j(x)}{\|a_i - a_j\|}. \tag{67}$$

Constraints with identical normals and strictly positive constant gap can be omitted; permanently identical scores must first be resolved into a fixed tie convention. To verify (67), the complement of the cell is the union of the violated open half-spaces. Distance to this union is the minimum of the distances to the half-spaces. The distance to an affine half-space is its nonnegative normalized margin. The infimum is unchanged on closure, so it equals distance to the cell boundary, which is a switching interface.

Restricting the domain to Ω can remove some ambient switching pieces. Then d_{amb} need not equal distance to the restricted interface set. It is still a conservative 1-Lipschitz distance that vanishes at all actual interfaces, so replacing d_Σ by it in (65) preserves the stability proof and may only trigger extra repair. This distinction matters near the external domain boundary. Computing the minimum in (67)

involves up to $k(M_{\exp} - k)$ comparisons and the corresponding stored or computed normal norms. A raw, unnormalized score gap is not a geometric distance: multiplying all scores by a constant leaves routing unchanged but rescales that gap.

For normalized softmax weights within an active set, suppose neighboring sets exchange only i and j. On their common face $s_i = s_j$, the normalization denominator agrees on both sides. All common expert weights therefore coincide and

$$F_S(z) - F_{S'}(z) = \gamma(z)(E_i(z) - E_j(z)),$$

where γ is the exchanged expert's weight. This weighted output jump, not the unweighted expert discrepancy, enters the surface cost. The trace-width theorem does not use such input-dependent weights; its stated arithmetic-mean restriction is essential for the knot count.

H.3 Accounting rule

Evaluate only F_S using the original selected indices S, then evaluate the auxiliary trace and connection. The forward cost is

$$C_{\text{original selected}} + C_H + C_{\text{geometry}} + C_{\text{connection}}.$$

The cardinality of original expert calls remains at most k; the other terms must be reported independently. With a certified residual bound and the bypass above, the expected auxiliary cost is bounded by

$$C_{\text{geometry}} + \Pr\{d_\Sigma(X) < M/\lambda\}(C_H + C_{\text{connection}}),$$

when the component costs have the indicated uniform bounds. This formula is a cost model, not a wall-clock measurement. It cannot establish speedup without actual component costs and an implementation. A trace that internally reproduces another expert's features can cost as much as that expert even though the outer selected set has not changed.

I A margin-density variant for an isolated jump

The positive-density surface law has a simple exact alternative when mass vanishes or concentrates near a single interface. Let the two constant vector endpoints be a, b, let $J = \|b - a\|$, and take one-dimensional side densities $c_-|t|^{\alpha-1}$ and $c_+|t|^{\alpha-1}$, where $c_\pm > 0$ and $\alpha > 0$. Assume the finite interval on each side contains the optimal transition. For fixed trace distances r and $J - r$, normal Lipschitzness and a straight-line attaining construction give

$$\int_0^{r/L} (r - Lt)^2 t^{\alpha-1}\,\mathrm{d}t = \frac{2r^{\alpha+2}}{\alpha(\alpha+1)(\alpha+2)L^\alpha}.$$

The minimizer of $c_- r^{\alpha+2} + c_+(J-r)^{\alpha+2}$ satisfies

$$\frac{r}{J-r} = \left(\frac{c_+}{c_-}\right)^{1/(\alpha+1)}.$$

Projection of an arbitrary vector trace onto the endpoint segment cannot increase either distance, so this scalar minimization also gives the vector optimum. Substitution yields the exact risk

$$\mathcal{R}_L^* = \frac{2}{\alpha(\alpha+1)(\alpha+2)} \frac{c_- c_+}{\left(c_-^{1/(\alpha+1)} + c_+^{1/(\alpha+1)}\right)^{\alpha+1}} \frac{J^{\alpha+2}}{L^\alpha}. \tag{68}$$

The case $\alpha = 1$ recovers the weighted cubic law. This isolated-jump calculation does not remove the positive-density assumptions of Theorem 7 or imply its remainder for vanishing densities.

J Verification protocol and proof dependencies

The accompanying code uses no external datasets or model checkpoints. Running `python code/verify.py` from the package root writes machine-readable results with fixed seed 20260916. The checked environment uses Python 3.13.5, NumPy 2.3.5, SciPy 1.17.0, and SymPy 1.14.0. The 14 groups contain 2,399 case checks: three symbolic identities; 120 weighted-step quadratures; 100

weighted projection-inequality diagnostics; 12 two-stage-risk quadratures; 24 piecewise-gradient calculations; 496 exact-interior comparisons; 500 radial-projection tests; 500 vector majorant-matrix tests; 600 cubic trace-excess tests; 30 width-construction instances checked at all knots and endpoints; eight selected-set call-count cases; and six finite-sample endpoint quadratic programs with clipped reconstructions. These counts describe the supplied run, not independent trials supporting an empirical accuracy claim.

The additional command `python code/verify_certificate.py` writes `results/certificate_verification.json` and `results/certificate_bounds.csv`. Its six groups contain 330 checks: seven exact distributional pairings for $Q = 0, \dots, 6$; nine exact radial-test polynomial derivative bounds; 32 Fourier-coefficient diagnostics, including coincident directions; 257 log-domain comparisons of the sharper and simpler gap formulas for $Q = 0, \dots, 256$; one exact coordinate/density normalization; and 24 stable evaluations of the finite-budget width formula. The pairing tests integrate the three-piece profile against polynomial derivatives on a square, using exact rational moments and a boundary-vanishing test function. They test the same point-source identity as the disk argument; the separate radial-polynomial checks use the disk test from the proof. Finite tests do not establish the quantified all-Q statement, which follows from Lemma 9.

The interval programs are solved numerically, with an explicit 10^{-5} output-agreement tolerance and recorded solver status. Their finite-grid slope diagnostics are not proofs of a continuum Lipschitz bound; Proposition 11 supplies that proof. The two-stage numerical ratios divide by $\mathcal{S}/L$, a proved lower bound, not by a numerically estimated optimum. The observed maximal quadrature discrepancy in that check was 6.94×10^{-18}; it is a floating-point consistency result, not an interval-certified rounding bound. Full values and tolerances are in the JSON and CSV files.

The quantitative shallow lower bound is self-contained: its dependencies are the weighted projection inequality, the point-source identity (34), and the polynomial test estimate of Lemma 9. The best-approximation existence theorem of Dereich et al. (2024) is background rather than a required step. The scalar boundary envelopes are classical Lipschitz extensions; the properties needed here are proved in Appendix F. All normal-slice integrals, clipping bounds, polyhedral error estimates, and evaluated-width constants are derived in the proofs. The automated diagnostics check the stated finite identities and implementations, rather than exhaustively optimizing shallow networks or replacing the proofs.