NONLINEAR ERROR CANCELLATION AT FIXED RELU ROUNDING ENDPOINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Scalar unbiasedness does not determine the mean rounded function of a nonlinear network. We prove a sharp separation between probability-law constraints for fixed signed Gaussian ReLU teachers, with unchanged width and original adjacent dyadic endpoints. For sufficiently small dyadic h and widths $m = 2n$ with n a power of two, the optimum over all laws preserving scalar stochastic-rounding marginals has order $\Theta(h^4 + h^2/m^2)$. Preserving each neuron's complete product-rounding law instead gives optimal order $\Theta(h^4 \log(1/h))$ whenever mh is bounded below and h is sufficiently small. The separation survives realization-specific affine correction in this width regime. A legal within-neuron correlation cancels the leading boundary moment. Ordered prefix coupling then preserves the modified complete local laws, while pointwise variation of the actual endpoint contrasts is $O(h)$, yielding variance $O(h^2/m^2)$. An arithmetic projection obstruction and a nonlinear quartic floor match the joint upper order for ordinary loss and best constant correction. Thus reciprocal-step width is necessary and sufficient for quartic loss on this arithmetic family; the necessity does not extend to general affine correction or convex means. For each fixed known density pair and local laws prescribed before sampling, almost surely the separation holds for every $c_0 > 0$, all sufficiently fine dyadic h, and all even $m \geq c_0/h$; this width condition is sufficient only. All teacher scalars are off-grid, directions are distinct, and nonaffine signal remains nonzero.

1 INTRODUCTION

Rounding neural-network parameters changes both their values and the locations of activation boundaries. For a nonlinear network, unbiased scalar rounding need not preserve the mean function. We study whether dependence among legal rounding bits can reduce the *order* of the true nonlinear loss while keeping the teacher, width, adjacent endpoints and scalar rounding marginals fixed.

We prove that it can for an explicit signed Gaussian ReLU family. Two interlaced direction groups have opposite output signs and different hidden scales, and every adjacent endpoint lies outside the teacher plane. A legal correlation between two hidden rounding bits cancels the signed, scale-adjusted transverse variance. Ordered prefix rounding preserves the resulting complete local laws; the actual endpoint contrasts have pointwise variation at most $180h$, giving variance at most $32400h^2/m^2$. Together with a nonlinear quartic floor and an arithmetic obstruction, this yields the sharp optimum $\Theta(h^4+h^2/m^2)$ over *all* scalar-marginal-preserving laws. Preserving each neuron's complete product-SR law instead forces a logarithmic mean penalty, even with optimal cross-neuron dependence.

The distinction is between local and global dependence. Preserving each neuron's *complete product-SR distribution* fixes the network mean, so cross-neuron dependence acts only on variance. Preserving *scalar SR marginals* also allows within-neuron correlations to change the nonlinear mean. Deterministic endpoint selection is a third optimization problem, with no marginal constraints.

The cancellation is invisible to one standard smooth proxy. Under independent-neuron laws, the modified covariance and product SR have identical expected fixed-gate bilinear loss, including the output–hidden product, and identical expected true loss with identity activation. This comparison concerns expected objectives before cross-neuron coupling; deterministic proxy minimization is a separate question.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Related work. Adaptive rounding and block reconstruction coordinate quantization decisions (Nagel et al., 2020; Li et al., 2021). DiscQuant analyzes gradient-vector discrepancy and uses actual output KL with a rounding regularizer in its practical objective (Chee et al., 2025); YAQA studies a KL-Hessian quadratic form (Tseng et al., 2026). Their objectives differ from the fixed-gate proxy examined here. Other work bounds nonlinear quantization error and quantized approximation rates (Šíma and Vidnerová, 2025; Gonon et al., 2022; Ou et al., 2024), including guarantees based on error feedback and output-preserving preprocessing (Zhang and Saab, 2023; Maly and Saab, 2023; Zhang and Saab, 2025). Our comparison fixes the original adjacent endpoints and optimizes over explicitly defined probability-law constraints.

Prefix discrepancy via network flows is classical (Knuth, 1995). Bounded accumulated rounding error combined with variation is the first-order Sigma–Delta mechanism; frame quantization already propagates such bounds through neural networks (Czaja and Na, 2025), and one-bit neural approximation uses noise shaping (Güntürk and Li, 2023). Dependent rounding and antithetic feature couplings are also established (Gandhi et al., 2006; Chekuri et al., 2009; Ren et al., 2019; Reid et al., 2025). We use the Gaussian ReLU kernel (Cho and Saul, 2009; Alfadly et al., 2020) and its known endpoint singularity (Bietti and Bach, 2021). Fourier/Radon ridge analysis is classical (Melrose and Uhlmann, n.d.; Unser, 2022), and Mao et al. (2024) combine it with discrepancy for approximation with freely chosen network parameters. We combine these tools with a pointwise variation bound for the actual original endpoint contrasts, nonlinear mean cancellation and matching lower bounds. Together they establish a sharp separation between probability-law constraints on the same fixed teacher.

2 Comparison classes and the sharp separation

We use the unscaled-output form of a two-layer bias-free network:

$$f_\theta(X) = \frac{1}{m}\sum_{j=1}^{m} s_j \sigma(p_j^\top X), \qquad L(C) = \|\widehat{f}_C - f_\theta\|^2_{L^2(\gamma_3)}, \qquad \sigma(t) = t_+. \tag{1}$$

Equivalently, $f_\theta = m^{-1/2}\sum_j a_j \sigma(p_j^\top X)$ with $a_j = s_j/\sqrt{m}$. Hidden coordinates and s_j use the adjacent endpoints of $h\mathbb{Z}$; canonical outputs use step $h/\sqrt{m}$. An aligned scalar has one endpoint. Scalar SR selects the upper endpoint with its fractional cell position. All rounding is independent of the prediction input $X \sim N(0, I_3)$. A *legal assignment* selects one original endpoint for every scalar.

Let n be a power of two, $m = 2n$, $s = 6/5$, $L_0 = 1/4$, $a = \sqrt{85}/5$, and define

$$\begin{gathered} p(u) = (6/5, u, 7/5), \qquad u_j^+ = \frac{3}{2} + \frac{j + 1/5}{4n}, \quad u_j^- = \frac{3}{2} + \frac{j + 7/10}{4n}, \\ f_m(X) = \frac{s}{2n}\sum_{j=0}^{n-1} [\sigma(p(u_j^+)^\top X) - \sigma(2p(u_j^-)^\top X)]. \end{gathered} \tag{2}$$

At each chosen width this teacher is fixed independently of the grid step. Its signed unscaled outputs are $s, -s$. Let $\mathcal{A}$ denote affine functions, and set

$$U_{m,h} = 90000h^4 + \frac{210h^2}{m}, \qquad c_{\mathrm{SR}} = \frac{31s^2e^{-3}}{625000(2\pi)^{3/2}aL_0}. \tag{3}$$

Comparison classes. Let $\Omega_{m,h}$ be the finite set of original legal assignments, P_j neuron j's complete product-SR law, and Q_j its cancellation row defined below. All probability laws are input-independent. Table 1 specifies the optimization domains; $\Pi(P), \Pi(Q) \subseteq \mathcal{S}$, and independent SR is one member of $\Pi(P)$. Define

$$J_{\mathrm{scalar}} = \inf_{\pi\in\mathcal{S}} \mathbb{E}_\pi L, \qquad J_P = \inf_{\pi\in\Pi(P)} \mathbb{E}_\pi L, \qquad J_Q = \inf_{\pi\in\Pi(Q)} \mathbb{E}_\pi L,$$

with $L_* = \min_{C\in\Omega_{m,h}} L(C)$. A superscript const or $\mathcal{A}$ means best constant or affine correction of each realized network before taking expectations.

Theorem 1 (Sharp separation of marginal constraint classes)**.** *For every teacher* (2)*, every allowed dyadic $h \le 1/10$ and $F_{m,h} = h^4 + h^2/m^2$, there is an absolute $c > 0$ such that*

$$cF_{m,h} \le L_{*,\mathrm{const}} \le L_* \le J_{\mathrm{scalar}} \le J_Q \le 88200F_{m,h}. \tag{4}$$

Table 1: Optimization domains on the same original endpoint set. Complete-row constraints fix within-neuron dependence; scalar constraints allow it to change. The convex mean is $\mu = \sum_C \lambda_C \widehat{f}_C$.

Domain	Variable	Preserved constraints	Objective
$\Pi(P)$	Joint law π	Every complete row P_j; arbitrary cross-neuron coupling	$\mathbb{E}_\pi L$
$\Pi(Q)$	Joint law π	Every complete row Q_j; arbitrary cross-neuron coupling	$\mathbb{E}_\pi L$
$\mathcal{S}$	Joint law π	Every scalar SR marginal; arbitrary dependence	$\mathbb{E}_\pi L$
Legal set	One $C \in \Omega_{m,h}$	Original endpoints only	$L(C)$; optimum L_*
Convex hull	Weights $\lambda_C \geq 0$, $\sum_C \lambda_C = 1$	Convex mean of legal networks; no SR constraints	Mean loss $\|\mu - f_m\|_2^2$

The same order holds for $J_{\mathrm{scalar}}^{\mathrm{const}}$ and J_Q^{const}. For every fixed $c_0 > 0$, uniformly over allowed $mh \geq c_0$ and sufficiently small h,

$$J_P = \Theta_{c_0}(h^4 \log(1/h)), \qquad J_{\mathrm{scalar}} = \Theta_{c_0}(h^4), \quad J_Q = \Theta_{c_0}(h^4). \tag{5}$$

Both orders in (5) *also hold with best realization-specific constant or general affine correction. The joint $F_{m,h}$ lower order in* (4) *is asserted only for ordinary loss and constant correction, not general affine correction or convex means.*

The arithmetic and quartic lower bounds hold for every legal realization. The inclusion $\Pi(Q) \subseteq \mathcal{S}$ and the prefix construction give the upper bound (Appendix A1.3). Every law in $\Pi(P)$ has the same mean: Jensen's inequality gives the logarithmic lower bound, also after projection off affine functions, and product-row prefix coupling attains the matching order (Appendix A2). These arguments establish optimality over each full class of admissible laws.

Lemma 2 (Independent-neuron cancellation construction)**.** *For every teacher* (2) *and dyadic $h = 2^{-b} \leq 1/10$, a law Q on the original legal assignments preserves every scalar's SR marginal and satisfies*

$$\mathbb{E}_{\mathsf{Q}} L(C) \leq U_{m,h}, \qquad \min_{C \text{ legal}} L(C) \leq U_{m,h}. \tag{6}$$

For the same teacher's fully independent product SR,

$$\mathbb{E}_{\mathrm{SR}} L(C) \geq \|\mathbb{E}_{\mathrm{SR}} \widehat{f} - f_m\|_2^2 \geq c_{\mathrm{SR}} h^4 \log \frac{1}{256h}, \qquad h \leq 2^{-9}, \quad nh \geq 64. \tag{7}$$

All original scalars are off every stated dyadic grid, no pair of teacher normals is parallel, and every hidden endpoint misses the teacher plane. Moreover, $\mathrm{dist}(f_m, \mathcal{A})^2 \geq 289/(10000\pi)$. *For every integer $p \geq 0$, certified conditional expectation computes one original-width legal vector with loss at most $U_{m,h} + 2^{-p}$ in $O(m^2 \,\mathrm{poly}(B + p + \log m))$ bit operations for rational unscaled input data of bit length B.*

Appendix A5 proves the lemma. The independent-neuron construction attains $O(h^4)$ when $mh^2 \geq 1$, against SR's $\Omega(h^4 \log(1/h))$; Theorem 3 reduces the required width. Each width defines one teacher that remains fixed across grid refinements. At fixed finite width, the ordinary and best-constant legal optima and the fixed-row coupling optimum instead have order $\Theta(h^2)$ as allowed dyadic $h \downarrow 0$ (Theorem 5). The explicit constants establish asymptotic separation; finite-step gains are evaluated separately in Section 5. The affine-corrected SR lower bound loses at most $100h^4$ (Theorem 8).

The legal correlation. Rotate the first and third input coordinates by $Z_1 = (6X_1 + 7X_3)/\sqrt{85}$, $Z_2 = X_2$, and $Y = (-7X_1 + 6X_3)/\sqrt{85}$. Teacher normals become $(\lambda a, \lambda u, 0)$ at scales $\lambda = 1, 2$. Every original endpoint becomes $(\lambda a + \eta_1, \lambda u + \eta_2, \tau)$ with

$$h/\sqrt{85} \leq |\tau| \leq 9h/\sqrt{85} < h, \qquad \|\eta\| < 2h. \tag{8}$$

Thus no legal choice retains the teacher plane. The uncancelled factor five in every scalar denominator ensures nonalignment; the two shifted direction grids exclude parallel pairs.

Only the first and third hidden bits are coupled. For $N = 1/h$, let their SR upper probabilities be $\alpha_\lambda = \{6\lambda N/5\}$ and $\beta_\lambda = \{7\lambda N/5\}$. The law chooses $t_\lambda = \Pr(B_1 = B_3 = 1)$ from

$N \bmod 5$	1	2	4	3
t_1	$61/420$	$167/420$	$229/420$	$83/420$
t_2	$4/15$	$29/70$	$1/15$	$1/70$

. (9)

The other probabilities are $\alpha_\lambda - t_\lambda$, $\beta_\lambda - t_\lambda$, and $1 - \alpha_\lambda - \beta_\lambda + t_\lambda$; all four are strictly positive. The middle hidden bit and signed output remain independently SR, and different neurons are independent. All sixteen local endpoints retain positive probability. If $V_\lambda = \mathbb{E}\tau_\lambda^2/h^2$, direct calculation gives

$$V_1^{\mathsf{Q}} = \frac{11}{85}, \quad V_2^{\mathsf{Q}} = \frac{22}{85}, \quad V_1^{\mathsf{Q}} - \frac{1}{2}V_2^{\mathsf{Q}} = 0, \qquad V_1^{\mathrm{SR}} - \frac{1}{2}V_2^{\mathrm{SR}} \in \left\{\frac{193}{2125}, \frac{232}{2125}\right\}. \tag{10}$$

The signed scale factor $1/\lambda$ is essential: the product-SR moment of this same signed teacher remains positive.

3 Why the cancellation is nonlinear

The finite-step proof separates the boundary contribution from the tangential endpoint perturbations. First set $\eta = 0$ in an auxiliary mean function. With $A_\lambda = \tau_\lambda/h$, $k = hY$, the common continuous-direction profile is

$$\delta(t,k) = \frac{s}{2}[\mathbb{E}\sigma(t + A_1 k) - 2\mathbb{E}\sigma(t + A_2 k/2) + \sigma(t)]. \tag{11}$$

Its zeroth and first coefficient moments vanish by scalar unbiasedness. Its integral is a constant times $k^2(V_1 - V_2/2)$, so (10) cancels its second coefficient moment as well. Consequently δ has support in $[-|k|, |k|]$, zero integral, amplitude at most $s|k|$, and total variation at most $s|k|$. Its primitive $D(t,k) = \int_{-\infty}^{t} \delta(q,k)\,dq$ is also compactly supported.

Writing $x = aZ_1 + (3/2)Z_2$ and $b = L_0 Z_2$, continuous averaging gives, for $b \neq 0$,

$$A(x,b,k) = \int_0^1 \delta(x + br, k)\,dr = \frac{D(x+b,k) - D(x,k)}{b}.$$

The support in x has length at most $4|k|$, and $|A| \le \min(s|k|, 4sk^2/|b|)$. Integration therefore yields $\|A(x,b,hY)\|_2^2 < 100h^4$. The zero integral prevents interior crossings from accumulating a term proportional to $1/|b|$ over logarithmically many radial scales.

We now return to the original endpoints. Bounded-variation quadrature controls the two finite shifted direction grids, and the tangential perturbations satisfy

$$|\sigma(t + v + d) - \sigma(t + v) - d\mathbf{1}_{\{t>0\}}| \le |d|\mathbf{1}_{\{|t| \le |v| + |d|\}}, \qquad \mathbb{E}d = 0.$$

Explicit activation-strip counts, valid even when d and $v = \tau Y$ are dependent, then give

$$\|B_{\mathsf{Q}}\|_2 \le 210h^2 + \frac{7h}{n}, \qquad B_{\mathsf{Q}} = \mathbb{E}_{\mathsf{Q}}\widehat{f} - f_m. \tag{12}$$

Every local true-error second moment is below $5h^2$ in unscaled units, so independence across neurons completes (6). The SR lower proof instead integrates its *actual* continuously rounded function on a sector with $64h \le Z_2 \le 1/4$. Its signed mean is bounded below by a positive multiple of h^2Y^2/Z_2, and the finite-grid variation error preserves this bound when $nh \ge 64$. Squaring and integrating produces the logarithm through $\int dr/r$.

Complete local laws keep the logarithmic bias. Let P_j be neuron j's complete product-SR law and $\mu = m^{-1}\sum_j \mathbb{E}_{P_j} F_j$, where F_j is its unscaled rounded function. For any joint law $\widetilde{\mathsf{Q}}$ with local marginals P_j,

$$\mathbb{E}_{\widetilde{\mathsf{Q}}} L = \|\mu - f_m\|_2^2 + \mathbb{E}_{\widetilde{\mathsf{Q}}}\|\widehat{f} - \mu\|_2^2. \tag{13}$$

Its possible gain over independent SR is at most $5h^2/m$. When $mh^2\log(1/h) \to \infty$ in the theorem's regime, its best expected loss divided by SR's tends to one. Deterministic support selection remains a different optimization problem. Coupling only the output to an unchanged product hidden law changes the network mean by at most $\sqrt{3}h^2/4$ and likewise cannot remove its leading logarithmic norm.

An exact control using the full bilinear proxy. For either Q or SR, set $P = m^{-1}\sum_j \mathbf{1}_{\{p_j^\top X>0\}}(q_{s,j}q_j - s_j p_j)^\top X$. This retains the cross-layer product. If $v_{jk} = \mathrm{Var}(q_{jk})$ and $v_{s,j} = \mathrm{Var}(q_{s,j})$, then

$$\mathbb{E}_{\mathsf{Q}}\|P\|_2^2 = \mathbb{E}_{\mathrm{SR}}\|P\|_2^2 = \frac{1}{2m^2}\sum_j \left[s_j^2 \sum_k v_{jk} + v_{s,j}\left(\|p_j\|^2 + \sum_k v_{jk} \right) \right]. \tag{14}$$

The mean proxy is zero pointwise, cross-neuron terms vanish, and Gaussian sign symmetry halves each ungated local squared norm. Hidden-coordinate covariances do not enter the result. Identity activation gives equal true expected errors, each twice (14). The equality holds for *expected objectives under independent-neuron laws*; deterministic proxy optimization is outside this comparison.

4 Prefix variation and matching lower bounds

Ordered contrasts control the remaining error. Order each signed group by u_j. Let g_{cjk} be its sixteen actual endpoint functions, q_{cjk} the complete row probabilities, and $H_{cjk} = g_{cjk} - g_{cj0}$. An integral flow enforces, for every group, label and prefix,

$$\sum_{j\le r} \mathbf{1}_{\{C_{cj}=k\}} \in \left\{ \left\lfloor \sum_{j\le r} q_{cjk} \right\rfloor, \left\lceil \sum_{j\le r} q_{cjk} \right\rceil \right\}. \tag{15}$$

Writing $\mu = m^{-1}\sum q_{cjk} g_{cjk}$, summation by parts bounds $|\widehat{f}_C - \mu|$ by $A(X)/m$, where

$$A(X) = \sum_{c,k\neq 0} \left[|H_{c,n_c,k}(X)| + \sum_{j<n_c} |H_{c,j+1,k}(X) - H_{cjk}(X)| \right], \qquad V = \|A\|_2. \tag{16}$$

The variation is taken pointwise before the Gaussian norm. Along the ordered anchors, the scalar contrast $\sigma(t+v)-\sigma(t)$ is monotone with variation $|v|$. Retaining the output–hidden product gives $V \le 180h$ for these actual sixteen-state dictionaries. No variation or equality of the probability rows is required.

Theorem 3 (Prefix coupling after nonlinear cancellation)**.** *For every fixed teacher* (2) *and stated dyadic $h \le 1/10$, an input-independent law $\mathsf{Q}^{\mathrm{pre}}$ preserves every complete modified row Q_j and satisfies all prefixes* (15)*, with*

$$\mathbb{E}_{\mathsf{Q}^{\mathrm{pre}}}\widehat{f} = \mathbb{E}_{\mathsf{Q}}\widehat{f}, \qquad \mathrm{Var}_{\mathsf{Q}^{\mathrm{pre}}}(\widehat{f}) \le \frac{32400h^2}{m^2}, \qquad \mathbb{E}_{\mathsf{Q}^{\mathrm{pre}}} L \le 88200h^4 + \frac{32792h^2}{m^2}. \tag{17}$$

A rational flow decomposition and certified true-loss selection give one original-legal vector with the last bound plus any prescribed tolerance $\eta > 0$. Every positive-support feasible prefix assignment, without this selection, only has the bound $88200h^4 + 75272h^2/m^2$.

Appendix A1 proves the flow law, a decomposition into at most $E+1$ integral flows on an $E = O(m)$-edge network, and the pointwise variation estimate. Combining the same law's mean (12) with its variance gives $32792 = 392 + 32400$. True-loss selection attains the additive bound; every feasible flow attains the same order. Marginal preservation belongs to the randomized law, not to a selected deterministic vector.

Cancellation cannot remove all nonlinear error in the signed construction. Every endpoint gate is fixed simultaneously on $\|Z\| \le h/64$, $1 \le Y \le 2$: the endpoint base norm is below six, while its transverse magnitude is at least $h/\sqrt{85}$. Every signed combination of endpoint ridges is therefore affine in Z on that cylinder. The teacher remains nonaffine.

Corollary 4 (Sharp original-legal quartic floor)**.** *For the teachers in Lemma 2, every legal assignment, every point in their full function-space convex hull, and every affine correction obey*

$$\|\widehat{f} + \ell - f_m\|_2^2 \ge c_{\mathrm{gap}} h^4, \qquad c_{\mathrm{gap}} = \frac{289e^{-5/2}}{80000 \cdot 64^4 (2\pi)^{3/2}}. \tag{18}$$

Consequently the unrestricted legal optimum is $\Theta(h^4)$ when mh is bounded below, by Theorem 3, with or without best affine correction.

Table 2: Dependencies of the fixed-teacher result. The density extension is proved separately in Appendix A4; other comparison classes appear in the standalone supplement.

Ingredient	Consequence	Proof
Legal local covariance	$\|B_Q\|_2 \le 210h^2 + 14h/m$	App. A5
Ordered actual contrasts	Same-row variance $\le 32400h^2/m^2$	App. A1
Pointwise gap/arithmetic + Q prefix	J_{scalar}, J_Q: lower floor and attained upper order	App. A1.3
Fixed P mean + projection; P prefix	J_P: logarithmic lower bound and matching upper bound	App. A2

The complete proof (Appendix A6.5) relates affine approximation of a homogeneous function on a disk to its Gaussian affine distance, then scales the disk by $h/64$. It permits arbitrary signed coefficients and finite width within the stated endpoint-gap geometry. The floor holds realization by realization; (7) instead bounds the SR mean.

A matching discrete obstruction. Let $L_{*,\mathrm{const}}$ be the least legal loss after best constant correction, and J_Q the infimum of expected ordinary loss over all couplings with the fixed complete rows Q_j. The local rows in J_Q are fixed.

Theorem 5 (Sharp joint order on the fixed teacher)**.** *For every teacher* (2) *and allowed dyadic* $h \le 1/10$*, put* $F_{m,h} = h^4 + h^2/m^2$ *and* $c_{\mathrm{joint}} = \min\{16/10625, c_{\mathrm{gap}}/17\}$*. Then*

$$c_{\mathrm{joint}} F_{m,h} \le L_{*,\mathrm{const}} \le L_* \le J_Q \le 88200 F_{m,h}. \tag{19}$$

Along any allowed sequence $h_r \downarrow 0$*, each of these three optima is* $O(h_r^4)$ *if and only if* $\liminf_r m_r h_r > 0$*.*

The exact X_1 projection and the grid residues modulo five give every assignment and constant β the lower bound $\|\widehat{f}_C + \beta - f_m\|_2^2 \ge h^2(6 - 16mh)_+^2/(2500m^2)$. Combining it with the quartic floor proves the theorem (Appendix A1.3). The arithmetic necessity applies to ordinary loss and constant correction. The explicit finite-step SR logarithmic bound uses $mh \ge 128$ and $h \le 1/512$; the broader asymptotic regime is established below.

The same prefix control isolates the mean mechanism. Theorem 10 proves, uniformly over all allowed widths and steps, $\mathbb{E}_{\mathrm{SR}} L = \Theta(h^2/m + h^4 \log(1/h))$. Apply the identical prefix construction to the product rows P_j or the modified rows Q_j. Both have variance at most $32400h^2/m^2$, but for every fixed $c_0 > 0$, uniformly over $mh \ge c_0$ and sufficiently small h, their expected losses are respectively $\Theta_{c_0}(h^4 \log(1/h))$ and $\Theta_{c_0}(h^4)$. Thus this comparison changes the local nonlinear mean while holding the variance-control mechanism fixed. The logarithmic P lower bound survives arbitrary realization-specific affine correction (Theorem 8); the arithmetic h^2/m^2 lower bound is specific to ordinary loss and constant correction.

Exact true-function convex compression. Shared dictionaries also compress the unrestricted convex relaxation exactly:

$$R_* = \min_{\substack{\lambda_{ck} \ge 0 \\ \sum_k \lambda_{ck} = n_c}} \left\| m^{-1} \sum_{c,k} \lambda_{ck} F_{ck} - f_m \right\|_2^2 \le L_*. \tag{20}$$

Here each fine cell contains one signed group and one middle-coordinate interval, with common candidates F_{ck} and size n_c. Every aggregate row is realized by $q_{jk} = \lambda_{ck}/n_c$, giving $16B$ variables instead of $16m$, with $B \le \min\{m, 3/(4h)\}$. This exact relaxation imposes no SR marginal constraints. For rational feasible rows within $\varepsilon_{\mathrm{opt}}$ of R_*, prefix decomposition and true-loss selection give

$$L(C) \le L_* + \varepsilon + \frac{32400h^2}{m^2}, \qquad \frac{L(C)}{L_*} \le 1 + \frac{32400}{c_{\mathrm{gap}} m^2 h^2} + \frac{\varepsilon}{L_*}, \quad \varepsilon = \varepsilon_{\mathrm{opt}} + \eta. \tag{21}$$

The proof uses this proposal's own exact mean and the row-independent bound $V \le 180h$ (Corollary 7). Arbitrary integral-flow selection gives only $(\sqrt{R_* + \varepsilon_{\mathrm{opt}}} + 180h/m)^2$. The tolerance

Table 3: Preselected registered case $m = 1024, h = 1/64$. All quantities are interval midpoints divided by h^4, rounded for display. This finite comparison does not fit or certify an asymptotic rate.

Local law	Coupling	B^2/h^4	V/h^4	$\mathbb{E}L/h^4$
P	Independent	0.006128	6.467809	6.473937
P	Prefix	0.006128	0.067332	0.073460
Q	Independent	0.003367	6.467856	6.471223
Q	Prefix	0.003367	0.064432	0.067798

$\varepsilon_{\mathrm{opt}}$ must include rationalization error and a certified optimization gap. This result assumes such a certificate; solver convergence and a small moderate-width relative gap require separate analysis. The convex approximation step uses established methods (Mirrokni et al., 2017; Gautam et al., 2024).

Certified finite-support selection. The API `certified_prefix_round` takes exact rational rows in the original Cartesian endpoint-label order, with each complete signed scale group sorted by exact u. It returns a finite randomized prefix law π preserving those complete rows and a selected legal vector C satisfying

$$L(C) \le \min_{D \in \operatorname{supp} \pi} L(D) + \eta \le \mathbb{E}_\pi L + \eta.$$

An arbitrary feasible integral flow has only the variation guarantee $(\|\mu - f_m\|_2 + V/m)^2$; the displayed support-selection guarantee uses certified true-loss scores (Appendix A1). Equal candidate functions share a compressed Gram without changing rows or the law. Arb Gaussian kernels (Johansson, 2016) certify the scores. Appendices A3 and A7.4 give the cell-potential alternative and its comparisons. In general dimension, the local dictionary can contain 2^{d+1} endpoints, so its enumeration remains exponential in d.

5 Numerical results and scope

Direct comparison: local law and cross-neuron coupling. We evaluate a prespecified twelve-case 2×2 design on the same fixed teachers and original endpoints: product rows P or cancellation rows Q, each with independent or prefix coupling. Table 3 shows the preselected $m = 1024, h = 1/64$ case; Appendix A7.1 retains all 24 randomized-law rows and twelve deterministic comparisons. Each law's mean is unchanged by cross-neuron coupling, but its actual variance changes.

Q lowers the squared mean in all twelve cases, but its prefix expected loss beats P in nine and loses in three. Its selected vector beats P in nine cases, loses in two, and has no strict interval separation in one. Each prefix law beats its independent expected loss in eight cases; in the four small cases with $m = 4, 16$, its variance instead increases. The $O(h^2/m^2)$ theorem is an absolute variance bound; finite-case variance can increase. Against SR-CE, Q-CE and convex-row CE, the selected P prefix wins in $5, 6, 7$ cases and the selected Q prefix in $8, 9, 9$, respectively; the worst Q prefix/SR-CE ratio is 6.760 (Table 6). The gains are case-dependent, and these comparisons provide no global-optimality certificate. Prefix rounding of a convex proposal is not evaluated in this study.

Two three-point refinement paths examine $mh = 1$ and $m = 1/(16h^2)$ (Figure 1, Appendix A7.2). Interval evaluation separates squared mean, variance and expected loss, revealing an initial variance increase and grid-residue effects. We report finite-step values without fitting an asymptotic rate; selected deterministic loss and randomized expected loss remain distinct.

Continuous-density mean diagnostic. For a fixed continuous teacher with known one-percent density mismatch, $\|B_P\|_2^2/h^4$ rises from .00821 to .01677 over $b = 8$ to 16, while adapted Q remains between .00344 and .00355. Figure 2 and Appendix A7.3 report every law and quadrature order. This experiment evaluates the continuous mean, rather than finite-teacher prefix rounding. The 3.66% quadrature sensitivity exceeds the adaptation gain, so the numerical comparison cannot resolve that gain.

Cell rounding and conditional-expectation baselines. All 24 cell-selected vectors beat independent SR, but Q-cell beats SR-CE in only $1/12$ cases and C-cell in $7/12$, with worst ratios 6.903

and 3.965 (Table 7). Width-four cell choices remain 2.38–21.77 times the exhaustive optimum; the independent-CE followup can lose to C-cell by 13.92 times (Appendices A7.4 and A7.5). The earlier small-width study reduces mean bias in all twelve cases but improves total loss in only six, with Q/SR ratios .999876079–1.000018429. Both CE laws choose exactly the same assignment throughout, and every computed convex lower certificate is zero (Table 4). These cases lie outside the explicit SR logarithmic regime. The identical CE selections and zero convex certificates give no evidence of a covariance-selection advantage or near-optimality.

An eighteen-case wide-teacher diagnostic evaluates independent-law gains; variance is certified, while the mean is estimated. It does not perform wide-network prefix selection (Table 10, Appendix A7.6).

All 189 tests pass. The 23 prefix tests comprise seventeen combinatorial/integer-support tests, four certified-prefix tests and two native quadratic-evaluation tests; the full suite also includes five exact-law tests. Appendix A7 gives the implementation checks and artifact fields.

Supplement S2 studies complementary repair methods on sixteen trained Fashion-MNIST models. True-objective repair outperforms its bilinear counterpart, but pair search has no systematic advantage over true single-coordinate repair or the fixed nonlinear soft baseline. This study evaluates repair, not the covariance construction, and the squared SR mean accounts for a small fraction of trained error. Supplement S18.5 reports all comparisons and calibration sizes; Supplement S18.6 gives the squared-mean diagnostics.

Extension to known direction densities. Certified true-loss selection uses the exact ReLU kernel of the known Gaussian input law. For sampled directions, the extension also assumes two known, fixed densities and supplies their ratio to the local law in (68). Estimating this ratio from data is a separate statistical problem. Appendix A4 gives the bounded-measurable-density formulation, functional cancellation theorem and fixed-sequence almost-sure result. Sorting each sampled finite prefix yields expected quartic loss for $mh \geq c_0 > 0$, and almost surely for all sufficiently fine dyadic steps and all even $m \geq c_0 h^{-1}$, for every fixed $c_0 > 0$ (Theorem 15). On the same probability-one event, the scalar and modified-row optima and the deterministic legal optimum are quartic, while the product-row optimum has the logarithmic penalty, for ordinary loss and best constant or affine correction. For sampled teachers, the width condition is sufficient; the necessity proved for the arithmetic family (2) is a separate result.

Scope and conclusions. The separation shows that local rounding dependence can change the optimal order of nonlinear error even when every scalar marginal and endpoint is fixed. The construction uses signed bias-free ridges, Gaussian inputs, a common teacher plane and dyadic grids. The density extension permits bounded measurable direction densities and direction-dependent covariance while retaining the other group parameters; approximate cancellation also requires uniform bounded variation of the functional moment. Extension beyond this geometry requires new analysis: even PSD pairwise covariances can fail to admit a legal joint local law (Lemma 11). The positive first-cell models in Supplement S1 and S3 have different, step-dependent targets and yield cubic, quartic-logarithmic and quartic floors in dimensions two, three and four. Supplement S3.7 studies finite-degree Hermite certificates for that separate class. The present theorem isolates a nonlinear mean mechanism; its behavior for general trained weights and input laws remains open.

AI USE STATEMENT

The human researcher supplied the original research idea and direction. This research draft used GPT-6-family agents, including GPT-6-astra, to assist hypothesis generation and refinement, proof development and checking, literature review, implementation and debugging, experimental design, synthetic-input generation, data processing and analysis, figure preparation, and drafting and revision. Automated proof reviews and numerical tests were used during development; they are not formal verification or substitutes for independent expert review. The authors remain responsible for the claims and the final manuscript.

REPRODUCIBILITY STATEMENT

The appendices give complete proofs, including conditioning and normalization conventions. Appendix A7 documents the exact kernels, local-law coupling, implementation tests and case-level experiment protocol. Supplement S18 documents the finite-support calibration and repair studies. These sections specify the Python modules, commands, CSV fields and numerical checks used to generate the reported results.

REFERENCES

M. Alfadly, A. Bibi, E. Botero, S. Alsubaihi, and B. Ghanem. Network Moments: Extensions and Sparse-Smooth Attacks. arXiv:2006.11776, 2020. https://arxiv.org/abs/2006.11776.

E. Anacleto, C. Meneses, and S. Ravelo. Closed-form formulas for evaluating r-flip moves to the unconstrained binary quadratic programming problem. Computers & Operations Research, 113:104774, 2020. https://doi.org/10.1016/j.cor.2019.104774.

A. Bietti and F. Bach. Deep Equals Shallow for ReLU Networks in Kernel Regimes. ICLR, 2021. https://arxiv.org/abs/2009.14397.

The CGAL Project. 2D Minkowski Sums: User Manual. CGAL 6.2.1, 2026. https://doc.cgal.org/latest/Minkowski_sum_2/index.html.

Moses Charikar and Anthony Wirth. Maximizing Quadratic Programs: Extending Grothendieck's Inequality. In *45th Annual IEEE Symposium on Foundations of Computer Science*, pp. 54–60, 2004. https://doi.org/10.1109/FOCS.2004.39.

J. Chee, A. Backurs, R. Heck, L. Zhang, J. Kulkarni, T. Rothvoss, and S. Gopi. DiscQuant: A Quantization Method for Neural Networks Inspired by Discrepancy Theory. COLT, 2025. https://proceedings.mlr.press/v291/chee25a.html.

C. Chekuri, J. Vondrák, and R. Zenklusen. Dependent Randomized Rounding for Matroid Polytopes and Applications. arXiv:0909.4348, 2009. https://arxiv.org/abs/0909.4348.

Y. Cho and L. K. Saul. Kernel Methods for Deep Learning. NeurIPS, 2009. https://proceedings.neurips.cc/paper_files/paper/2009/file/5751ec3e9a4feab575962e78e006250d-Paper.pdf.

Wojciech Czaja and Sanghoon Na. Frame Quantization of Neural Networks. Journal of Fourier Analysis and Applications, 31, Article 49, 2025. https://doi.org/10.1007/s00041-025-10182-7.

E. de Klerk, M. Laurent, Z. Sun, and J. C. Vera. On the convergence rate of grid search for polynomial optimization over the simplex. Optimization Letters, 11:597–608, 2017. https://arxiv.org/abs/1510.01163.

R. Gandhi, S. Khuller, S. Parthasarathy, and A. Srinivasan. Dependent Rounding and Its Applications to Approximation Algorithms. Journal of the ACM, 2006. https://doi.org/10.1145/1147954.1147956.

Tanmay Gautam, Reid Pryzant, Ziyi Yang, Chenguang Zhu, and Somayeh Sojoudi. Soft convex quantization: revisiting Vector Quantization with convex optimization. Learning for Dynamics & Control, PMLR 242:273–285, 2024. https://proceedings.mlr.press/v242/gautam24a.html.

R. Ge, J. D. Lee, and T. Ma. Learning One-hidden-layer Neural Networks with Landscape Design. arXiv:1711.00501, 2017. https://arxiv.org/html/1711.00501.

Xi Geng. Stochastic Calculus, lecture notes, 2016. https://www.math.cmu.edu/~xig/Files/Teaching/Stochastic%20Calculus%20(Fall%202016)/Stochastic%20Calculus.pdf.

A. Gonon, N. Brisebarre, R. Gribonval, and E. Riccietti. Approximation speed of quantized vs. unquantized ReLU neural networks and beyond. arXiv:2205.11874, version 2, 2022. https://arxiv.org/abs/2205.11874.

C. Sinan Güntürk and Weilin Li. Approximation of functions with one-bit neural networks. arXiv:2112.09181, version 2, 2023 (first submitted 2021). https://arxiv.org/abs/2112.09181.

Fredrik Johansson. Arb: Efficient Arbitrary-Precision Midpoint-Radius Interval Arithmetic. arXiv:1611.02831, 2016. https://arxiv.org/abs/1611.02831.

Donald E. Knuth. Two-way rounding. SIAM Journal on Discrete Mathematics, 8(2):281–290, 1995. https://arxiv.org/abs/math/9504228.

L. C. Lau, R. Wang, and H. Zhou. Spectral Sparsification by Deterministic Discrepancy Walk. SOSA, 2025. https://arxiv.org/abs/2408.06146.

Y. Li, R. Gong, X. Tan, Y. Yang, P. Hu, Q. Zhang, F. Yu, W. Wang, and S. Gu. BRECQ: Pushing the Limit of Post-Training Quantization by Block Reconstruction. ICLR, 2021. https://arxiv.org/abs/2102.05426.

J. Maly and R. Saab. A Simple Approach for Quantizing Neural Networks. 2023. https://arxiv.org/abs/2209.03487.

Tong Mao, Jonathan W. Siegel, and Jinchao Xu. Approximation Rates for Shallow ReLU^k Neural Networks on Sobolev Spaces via the Radon Transform. arXiv:2408.10996, 2024. https://arxiv.org/abs/2408.10996.

A. Maurer and M. Pontil. Empirical Bernstein Bounds and Sample Variance Penalization. COLT, 2009. https://arxiv.org/abs/0907.3740.

Richard B. Melrose and Gunther Uhlmann. An Introduction to Microlocal Analysis. Chapter 2: Radon transform. https://math.mit.edu/~rbm/books/imaast.pdf.

Jiahao Ming, Aaditya Ramdas, Yi Shen, Ruodu Wang, and Ian Waudby-Smith. Gaffke's confidence interval for the mean of bounded data is inadmissible but asymptotically efficient. arXiv:2607.18661, 2026. https://arxiv.org/abs/2607.18661.

Vahab Mirrokni, Renato Paes Leme, Adrian Vladu, and Sam Chiu-wai Wong. Tight Bounds for Approximate Carathéodory and Beyond. ICML, PMLR 70:2440–2448, 2017. https://proceedings.mlr.press/v70/mirrokni17a.html.

Markus Nagel, Mart van Baalen, Tijmen Blankevoort, and Max Welling. Data-Free Quantization Through Weight Equalization and Bias Correction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1325–1334, 2019. https://openaccess.thecvf.com/content_ICCV_2019/html/Nagel_Data-Free_Quantization_Through_Weight_Equalization_and_Bias_Correction_ICCV_2019_paper.html.

M. Nagel, R. A. Amjad, M. Van Baalen, C. Louizos, and T. Blankevoort. Up or Down? Adaptive Rounding for Post-Training Quantization. ICML, 2020. https://proceedings.mlr.press/v119/nagel20a.html.

W. Ou, P. Schenkel, and H. Bölcskei. Three Quantization Regimes for ReLU Networks. arXiv:2405.01952, 2024. https://arxiv.org/abs/2405.01952.

I. Reid, S. Markou, K. Choromanski, R. E. Turner, and A. Weller. Variance-Reducing Couplings for Random Features. ICLR, 2025. https://arxiv.org/abs/2405.16541.

H. Ren, S. Zhao, and S. Ermon. Adaptive Antithetic Sampling for Variance Reduction. ICML, 2019. https://proceedings.mlr.press/v97/ren19b.html.

Karl H. Schlag. How to Attain Minimax Risk with Applications to Distribution-Free Nonparametric Estimation and Testing. Working paper, March 12, 2007. https://homepage.univie.ac.at/karl.schlag/research/statistics/MinimaxRisk.pdf.

J. W. Siegel. Optimal Approximation of Zonoids and Uniform Approximation by Shallow Neural Networks. Constructive Approximation, 62:441–469, 2025. https://doi.org/10.1007/s00365-025-09712-9.

J. Šíma and P. Vidnerová. Weight-Rounding Error in Deep Neural Networks. ECML PKDD, 2025. https://doi.org/10.1007/978-3-032-06078-5_23.

A. Tseng, Z. Sun, and C. De Sa. Model-Preserving Adaptive Rounding. ICML, 2026; arXiv:2505.22988, version 3. https://arxiv.org/abs/2505.22988.

Michael Unser. Ridges, Neural Networks, and the Radon Transform. arXiv:2203.02543, 2022. https://arxiv.org/abs/2203.02543.

R. Vershynin. High-Dimensional Probability: An Introduction with Applications in Data Science. Cambridge University Press, 2018. https://doi.org/10.1017/9781108231596.

H. Xiao, K. Rasul, and R. Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. arXiv:1708.07747, 2017. https://arxiv.org/abs/1708.07747.

J. Zhang and R. Saab. SPFQ: A Stochastic Algorithm and Its Error Analysis for Neural Network Quantization. arXiv:2309.10975, 2023. https://arxiv.org/abs/2309.10975.

H. Zhang and R. Saab. A Unified Stochastic Framework for Neural Network Quantization and Pruning. 2025. https://arxiv.org/abs/2412.18184v3.

Appendix A1 proves ordered complete-row rounding, the actual endpoint variation bound and finite true-loss selection. Appendix A3 proves shared-cell coupling and exact convex compression. Appendix A4 proves functional cancellation and the robust almost-sure separation. Appendix A5 proves the initial fixed-teacher signed separation, including all actual endpoint perturbations and the proxy identity. Appendix A6 proves the common-density criterion, feasible two-bit moment interval and sharp affine-resistant floor. Appendix A7 documents the signed finite study. The separately compiled supplement contains the independent comparison classes: the positive first-cell comparison classes have step-dependent targets; the continuous and fixed-dimensional families have different teacher laws; the calibration and trained-model experiments address other mechanisms. Theorems 3 and 5 are self-contained in the present appendices; the supplement develops the other models and experiments separately.

A1 Ordered endpoint laws: complete prefix and variation proofs

This appendix proves Theorem 3 and the relative guarantee (21). The construction retains the original Cartesian endpoints, width and complete modified local laws. Its flow and variation tools are established: Knuth (1995) controls partial sums using network flows, and first-order Sigma–Delta methods combine bounded accumulated quantization error with variation. In neural-network frame quantization, Czaja and Na (2025, Theorem 3.1) recall a coefficient error bound proportional to step size times frame variation, then propagate such errors through networks. Here we bound the pointwise variation of the actual legal ReLU state contrasts and combine it with nonlinear mean cancellation to obtain the fixed-endpoint separation.

A1.1 Complete rows and every ordered prefix

Let (Ω, ν) be a probability space and $\mathcal{H} = L^2(\nu; \mathbb{R})$. Partition m rows into nonempty ordered groups c, with positions $j = 1, \dots, n_c$ and labels $k = 0, \dots, K_c - 1$. Fix the orders independently of the prediction input. Each row has legal functions $g_{cjk} \in \mathcal{H}$ and prescribed complete probabilities $q_{cjk} \geq 0$, $\sum_k q_{cjk} = 1$. Define

$$\begin{gathered} Q_{crk} = \sum_{j \leq r} q_{cjk}, \qquad \mu = \frac{1}{m} \sum_{c,j,k} q_{cjk} g_{cjk}, \qquad H_{cjk} = g_{cjk} - g_{cj0}, \\ \widehat{f}_C = \frac{1}{m} \sum_{c,j} g_{cj,C_{cj}}, \qquad V = \|A\|_2, \\ A(X) = \sum_c \sum_{k \neq 0} \left[|H_{c,n_c,k}(X)| + \sum_{j < n_c} |H_{c,j+1,k}(X) - H_{cjk}(X)| \right]. \end{gathered} \tag{22}$$

The baseline label may have probability zero: it is subtracted only for an identity, and is never selected without positive support. Fix simultaneous measurable versions of these finitely many functions. Then $A \in L^2$, independently of changes on null sets. No regularity of the probability rows is assumed, and a label may represent different functions in different rows. Write the prefix constraints as

$$N_{crk} = \sum_{j \leq r} \mathbf{1}_{\{C_{cj} = k\}} \in \{\lfloor Q_{crk} \rfloor, \lceil Q_{crk} \rceil\} \quad \text{for every } c, r, k. \tag{23}$$

Theorem 6 (Ordered complete-row rounding)**.** *There exists a positive-support assignment satisfying* (23)*. Every such assignment satisfies, for every $f \in \mathcal{H}$,*

$$|\widehat{f}_C(X) - \mu(X)| \leq A(X)/m \quad \nu\text{-a.e.}, \qquad \|\widehat{f}_C - f\|_2^2 \leq (\|\mu - f\|_2 + V/m)^2. \tag{24}$$

There is also a single input-independent law supported on these assignments which preserves every complete row and obeys

$$\Pr(C_{cj} = k) = q_{cjk}, \quad \mathbb{E}\widehat{f}_C = \mu, \quad \mathrm{Var}(\widehat{f}_C) \leq V^2/m^2, \quad \mathbb{E}\|\widehat{f}_C - f\|_2^2 \leq \|\mu - f\|_2^2 + V^2/m^2. \tag{25}$$

The same law works for every target and uses only q and the fixed orders. For rational q, a constructive decomposition uses at most $E + 1$ integral flows, where $E = O(\sum_c n_c K_c)$ is the network edge count

below. If their target losses can be enclosed to any positive prescribed accuracy, selecting from this decomposition gives one legal vector with

$$\|\widehat{f}_{C^\star} - f\|_2^2 \le \|\mu - f\|_2^2 + V^2/m^2 + \eta \qquad (\eta > 0). \tag{26}$$

At $\eta = 0$ this is an existence or exact-real comparison statement. A deterministic vector has no probability-marginal assertion.

Proof. Use a source s, sink t, row vertices r_{cj}, and vertices v_{cjk} for each position and label. Each edge $s \to r_{cj}$ has both capacities one. Include $r_{cj} \to v_{cjk}$ with capacities zero and one exactly when $q_{cjk} > 0$. Include the chains

$$\begin{array}{ll} v_{crk} \to v_{c,r+1,k}, & 1 \le r < n_c, \\ v_{c,n_c,k} \to t, & r = n_c, \end{array} \qquad [\ell_{crk}, u_{crk}] = [\lfloor Q_{crk} \rfloor, \lceil Q_{crk} \rceil]. \tag{27}$$

The source supplies m, the sink receives m, and other vertices conserve flow. The fractional flow x^0 taking values q_{cjk} on row-label edges and Q_{crk} on their outgoing chain edges is feasible. Every capacity is integral and every edge interval has width zero or one. An integral flow selects exactly one positive-support label per row, and its chain flows are the prefix counts in (23).

Here integrality has a direct construction. In the undirected graph of nonintegral flow edges no vertex has degree one: conservation with integer supplies would otherwise be impossible. Any nonempty fractional graph thus has an undirected simple cycle. Traverse it and change directed edges with sign according to agreement with that traversal. This circulation preserves all supplies. Its two maximal feasible changes are $-a, b$ with $a, b > 0$; each endpoint makes at least one fractional edge integral. Leave integral edges untouched thereafter. At most E pivots give an integral flow, proving existence. Instead choose the two endpoints with probabilities

$$\Pr(\Delta = -a \mid x) = \frac{b}{a+b}, \qquad \Pr(\Delta = b \mid x) = \frac{a}{a+b}, \qquad \mathbb{E}(\Delta \mid x) = 0. \tag{28}$$

A cycle rule fixed from the current fractional graph and vertex order is independent of X and f. Every edge is a bounded martingale, and termination after at most E pivots preserves its expectation. In particular the final row-label expectation is q_{cjk}. This proves the complete-row random law on the enlarged prefix network; it is not the old fine-cell graph law.

For a feasible assignment let $\delta_{cjk} = \mathbf{1}_{\{C_{cj}=k\}} - q_{cjk}$ and $D_{crk} = \sum_{j \le r} \delta_{cjk}$. Then $|D_{crk}| < 1$, with $D_{crk} = 0$ at integral Q_{crk}, and $\sum_k \delta_{cjk} = 0$. For every fixed input, Abel summation gives

$$\begin{aligned} m(\widehat{f}_C - \mu) &= \sum_{c,j,k \ne 0} \delta_{cjk} H_{cjk}, \\ \sum_{j=1}^{n_c} \delta_{cjk} H_{cjk} &= D_{c,n_c,k} H_{c,n_c,k} + \sum_{j < n_c} D_{cjk}(H_{cjk} - H_{c,j+1,k}). \end{aligned} \tag{29}$$

Its absolute value is bounded by $A(X)$. Taking the norm and then the triangle inequality proves (24). Under the complete-row law, $\mathbb{E}\widehat{f}_C = \mu$ and every sample satisfies that pointwise bound. The variance is therefore at most V^2/m^2; the Hilbert-space bias–variance identity proves (25).

For rational data one need not expand the possibly large binary tree of martingale pivots. Let D be a common positive denominator of x^0. Maintain an integer residual vector y and positive integer mass M, initially $y = Dx^0, M = D$, with y/M feasible for the original polytope. Extract an integral feasible flow z in its current face using the deterministic pivots above, retaining every coordinate already at an original bound. On width-one edges, $z_e \in \{\ell_e, u_e\}$. Set

$$\begin{gathered} a = \min\left(\{M\} \cup \{Mu_e - y_e : z_e = \ell_e,\ u_e - \ell_e = 1\} \cup \{y_e - M\ell_e : z_e = u_e,\ u_e - \ell_e = 1\}\right), \\ y' = y - az, \qquad M' = M - a. \end{gathered} \tag{30}$$

Every entry of this minimum is a positive integer. For instance, $z_e = \ell_e$ and $y_e = Mu_e$ would violate the requirement that z remain in the current face. Width-zero edges need no constraint. Equation (30) is precisely $M'\ell_e \le y'_e \le M'u_e$; conservation and supplies scale from mass M to M' because z has the original supplies.

Record z with weight a/D. If $a = M$, the residual is zero and stop. Otherwise a previously non-bound coordinate reaches an original bound in y'/M', while coordinates already at a bound remain there. There are at most E changes, hence at most $E + 1$ recorded flows. Telescoping gives weights summing to one and weighted flow exactly x^0. Their assignments form one common complete-row law, each satisfying (24). Their weighted mean loss is bounded by (25), and the least-loss member has no greater loss. If every member's loss enclosure has width at most η, choosing the smallest upper endpoint incurs at most η above this least loss. This proves (26) with the same proposal and same decomposition. It uses no fine-cell concave potential. □

The rational flow and decomposition steps have polynomial bit complexity in the explicit input length. The product of input denominators is a valid D with bit length at most their total length. Throughout the decomposition, $0 < M \le D$ and integer residual coordinates are bounded by Dm. Integral extraction on y/M keeps denominator M and needs at most E pivots; there are at most $E + 1$ extractions and polynomial graph searches. These counts are polynomial in the explicit dictionary size, which can be 2^{d+1} in dimension d; loss-oracle evaluation is a separate cost. For the rational Gaussian ReLU candidates here, every loss is a finite kernel expression, allowing certified interval selection at any positive tolerance.

A1.2 The actual signed endpoint contrasts have variation at most $180h$

Use exactly (2) and its complete local laws from (9), including independent middle-coordinate and output SR within each neuron. Order each signed scale group by u_j, breaking equal middle anchors by original row index. Retain all sixteen labels $(\alpha, z) \in \{0,1\} \times \{0,1\}^3$, with baseline $(0,0)$. For one group $\lambda \in \{1,2\}$, signed output $s_\lambda \in \{s, -s\}$, define

$$\begin{aligned} a^- &= h\lfloor s_\lambda/h \rfloor, \qquad b_j = h\lfloor \lambda u_j/h \rfloor, \\ w_j^- &= (h\lfloor 6\lambda/(5h) \rfloor, b_j, h\lfloor 7\lambda/(5h) \rfloor), \qquad g_{j,\alpha,z}(X) = (a^- + h\alpha)\sigma((w_j^- + hz)^\top X), \\ t_j(X) &= (w_j^-)^\top X, \qquad b_1 \le \cdots \le b_n, \quad b_n - b_1 \le \lambda/4 + h \le 3/5, \quad |a^-| \le 13/10. \end{aligned} \tag{31}$$

Every endpoint $w_j^- + hz$ has norm below six, as in the endpoint-floor proof. For fixed X, $t_j(X)$ is monotone, decreasing if $X_2 < 0$ and constant if $X_2 = 0$. The exact contrast is

$$H_{j,\alpha,z}(X) = a^-[\sigma(t_j(X) + hz^\top X) - \sigma(t_j(X))] + h\alpha\sigma(t_j(X) + hz^\top X). \tag{32}$$

This includes the output–hidden product. No gate or endpoint is replaced. For real v, $R_v(t) = \sigma(t + v) - \sigma(t)$ is monotone, has amplitude at most $|v|$ and total variation $|v|$ on the real line, for either sign of v. Its three affine pieces verify these assertions directly. Thus

$$\begin{aligned} |H_{n,\alpha,z}(X)| + \sum_{j<n} |H_{j+1,\alpha,z}(X) - H_{j,\alpha,z}(X)| &\le 2|a^-|h|z^\top X| \\ &\quad + h\alpha[\sigma((w_n^- + hz)^\top X) + (b_n - b_1)|X_2|]. \end{aligned} \tag{33}$$

The last variation bound uses monotonicity and ReLU Lipschitz continuity; it does not assume $X_2 \ge 0$.

For standard Gaussian $X \in \mathbb{R}^3$, $\||z^\top X|\|_2 = \|z\|$, $\|\sigma(w^\top X)\|_2 = \|w\|/\sqrt{2}$, and $\||X_2|\|_2 = 1$. Sum the fifteen nonbaseline state bounds pointwise, then apply the norm triangle inequality. Because $\sum_{z\in\{0,1\}^3} \|z\| = 3 + 3\sqrt{2} + \sqrt{3}$ and eight labels have $\alpha = 1$, the contribution of one group satisfies

$$\begin{aligned} \|A_\lambda\|_2 &\le h[4|a^-|(3 + 3\sqrt{2} + \sqrt{3}) + 8(6/\sqrt{2} + 3/5)] \\ &< h[(26/5)(37/4) + 8(9/2 + 3/5)] = \frac{889}{10}h < 90h, \qquad V = \|A_1 + A_2\|_2 < 180h. \end{aligned} \tag{34}$$

The rational simplification uses $\sqrt{2} < 3/2$ and $\sqrt{3} < 7/4$. This estimate is uniform over all complete local probability rows; it uses neither smooth probabilities nor equality of rows or equidistribution. There are exactly two groups and sixteen labels.

Completion of Theorem 3. The random flow law preserves every complete Q_j, so the actual mean including all tangential perturbations is exactly the mean in (12). The output–hidden independence used to prove that mean remains true within each row, although neurons become dependent. In particular,

$$\begin{gathered}\|B_{\mathsf{Q}}\|_2 \le 210h^2 + 14h/m, \qquad \|B_{\mathsf{Q}}\|_2^2 \le 88200h^4 + 392h^2/m^2, \qquad V^2 \le 32400h^2, \\ 392 + 32400 = 32792, \qquad 2(14+180)^2 = 75272, \\ (\|B_{\mathsf{Q}}\|_2 + V/m)^2 \le (210h^2 + 194h/m)^2 \le 88200h^4 + 75272h^2/m^2.\end{gathered} \tag{35}$$

Equations (25), (24) and (26) now prove, respectively, the randomized bound, the arbitrary feasible-assignment bound and the selected-vector bound. All pair probabilities in (9), and all local middle and output probabilities, are rational. The finite decomposition therefore applies without approximating their prescribed marginals.

For $mh \ge c_0 > 0$ and tolerance $\eta = O(h^4)$, these upper bounds are $O_{c_0}(h^4)$. The floor (18) applies to every original-legal assignment and every affine correction, as well as their full convex hull. Choosing zero correction in the upper bound proves quartic order for both the ordinary legal optimum and its best-affine version. The same order holds for expected optimally corrected loss if correction is chosen separately after each realization. This is a joint width–step statement: at each width the teacher is fixed independently of h, while one fixed finite width eventually leaves this regime.

The ordinary SR comparison retains its separate conditions $h \le 2^{-9}$ and $nh \ge 64$, equivalently $mh \ge 128$, in (7). Thus the logarithmic ratio separation follows along the new width sequences when $mh \ge 128$, or $mh \to \infty$; $mh \ge c_0 > 0$ alone is insufficient to invoke that lower bound. Theorem 8 proves the affine-corrected SR comparison with an explicit quartic subtraction. The next subsection proves necessity of reciprocal-step width for ordinary loss and best constant correction, using this teacher's arithmetic. This cross-neuron law need not preserve the independent-laws' bilinear proxy identity (14).

A1.3 Arithmetic obstruction and sharp joint width order

This proves Theorem 5 for every $m = 2n$, with n a power of two, and every allowed dyadic $h \le 1/10$. In addition to L_*, define precisely

$$\begin{aligned}L_{*,\mathrm{const}} &= \min_{\substack{C \text{ original-legal} \\ \beta \in \mathbb{R}}} \|\widehat{f}_C + \beta - f_m\|_2^2, \\ J_Q &= \inf_{\substack{\Pi \text{ on original legal assignments} \\ \Pi(C_j = k) = Q_j(k) \text{ for all } j,k}} \mathbb{E}_\Pi \|\widehat{f}_C - f_m\|_2^2.\end{aligned} \tag{36}$$

Here Q_j is the fixed complete cancellation law of (9), including middle and output bits. Its coupling may have arbitrary cross-neuron dependence. The constant correction is external; all selected neuron parameters remain original-legal.

Exact first Gaussian projection. Write a row's original output and hidden vector as s_j, p_j, and its selected endpoints as $\widetilde{a}_j, \widetilde{w}_j$. Gaussian symmetry and $\sigma(t) = (t+|t|)/2$ give $\mathbb{E}[X_1\sigma(w^\top X)] = w_1/2$ for every w. Since $\mathbb{E}X_1 = 0$ and $\|X_1\|_2 = 1$, every constant β satisfies

$$\begin{aligned}\langle \widehat{f}_C + \beta - f_m, X_1 \rangle &= \frac{1}{2m} \sum_j (\widetilde{a}_j \widetilde{w}_{j1} - s_j p_{j1}), \\ \|\widehat{f}_C + \beta - f_m\|_2^2 &\ge \frac{1}{4m^2} \left[\sum_j (\widetilde{a}_j \widetilde{w}_{j1} - s_j p_{j1}) \right]^2 .\end{aligned} \tag{37}$$

This is the exact Gaussian projection of the actual network, with its $1/m$ normalization and unscaled outputs. Other hidden bits and all activation changes are included without any frozen-gate approximation.

Let $N = 1/h$, a power of two. The n positive rows have $(s_j, p_{j1}) = (6/5, 6/5)$ and the n negative rows have $(s_j, p_{j1}) = (-6/5, 12/5)$. Set $\epsilon_j = \mathrm{sgn}(s_j)$ and let $\lambda_j = 1, 2$ for these two groups. Both

$\widetilde{a}_j/h$ and $\widetilde{w}_{j1}/h$ are integers, including negative outputs. Define

$$\begin{aligned} \delta a_j &= \widetilde{a}_j - s_j, \qquad \delta w_{j1} = \widetilde{w}_{j1} - p_{j1}, \\ Z_C &= \sum_j \left(\epsilon_j \frac{\widetilde{w}_{j1}}{h} + \lambda_j \frac{\widetilde{a}_j}{h} \right) \in \mathbb{Z}, \qquad \sum_j s_j p_{j1} = -\frac{36n}{25}, \\ S_{\text{lin}} &:= \sum_j (s_j \delta w_{j1} + p_{j1} \delta a_j) \\ &= \frac{6h}{5} Z_C - 2 \sum_j s_j p_{j1} = \frac{6h}{5} \left(Z_C + \frac{12nN}{5} \right). \end{aligned} \tag{38}$$

Because n, N are powers of two, $12nN$ is not divisible by five. Thus $|Z_C + 12nN/5| \geq 1/5$, and $|S_{\text{lin}}| \geq 6h/25$ for every legal choice, without any probability or prefix condition.

All fractional parts of $s_j/h = \pm 6N/5$ and $p_{j1}/h = 6\lambda_j N/5$ lie in $\{1/5, 2/5, 3/5, 4/5\}$. The usual floor rule for negative outputs is included. Original adjacent endpoints therefore give $|\delta a_j|, |\delta w_{j1}| \leq 4h/5$. For the actual two-layer product error,

$$\begin{aligned} S_C &:= \sum_j (\widetilde{a}_j \widetilde{w}_{j1} - s_j p_{j1}) = S_{\text{lin}} + \sum_j \delta a_j \delta w_{j1}, \\ \left| \sum_j \delta a_j \delta w_{j1} \right| &\leq \frac{16}{25} m h^2, \qquad |S_C| \geq \frac{h}{25} (6 - 16mh)_+. \end{aligned} \tag{39}$$

Combining (37) and (39), and using the already affine-resistant quartic floor, proves

$$\|\widehat{f}_C + \beta - f_m\|_2^2 \geq \max \left\{ c_{\text{gap}} h^4, \frac{h^2 (6 - 16mh)_+^2}{2500 m^2} \right\}. \tag{40}$$

Uniform joint order and the width criterion. If $mh \leq 1/4$, then (40) is at least $h^2/(625m^2)$. Since $h^4 \leq h^2/(16m^2)$, this is at least $(16/10625)(h^4 + h^2/m^2)$. If $mh > 1/4$, then $h^2/m^2 < 16h^4$, and the quartic floor is at least $(c_{\text{gap}}/17)(h^4 + h^2/m^2)$. These cases give the uniform lower bound for $L_{*,\text{const}}$ with the stated c_{joint}. Allowing a constant can only lower the loss, so $L_{*,\text{const}} \leq L_*$. Every coupling averages losses of legal assignments and hence has expected loss at least L_*. Finally the complete-row prefix law has

$$J_Q \leq \mathbb{E}_{\mathbf{Q}^{\text{pre}}} L \leq 88200 h^4 + 32792 h^2/m^2 \leq 88200 (h^4 + h^2/m^2),$$

proving the chain in (19). Every feasible prefix assignment also attains this joint order by its 75272 upper remainder.

For any allowed joint sequence (m_r, h_r) with $h_r \downarrow 0$, division by h_r^4 yields, for each of $T \in \{L_*, L_{*,\text{const}}, J_Q\}$,

$$T(m_r, h_r) = O(h_r^4) \quad \Longleftrightarrow \quad \exists c_0 > 0, r_0 \ \forall r \geq r_0 : \ m_r h_r \geq c_0 \quad \Longleftrightarrow \quad \liminf_{r \to \infty} m_r h_r > 0. \tag{41}$$

The quartic floor makes this also the criterion for $\Theta(h_r^4)$. Thus reciprocal-step width is necessary and sufficient on this family, without claiming an optimal finite transition constant. SR's separate logarithmic lower bound still requires $nh \geq 64$ and $h \leq 2^{-9}$.

The restriction to ordinary loss and constants is substantive. General affine correction can cancel the X_1 projection, so the quartic floor remains valid but this proof gives no h^2/m^2 lower term for that objective. Convex averaging also destroys the integer obstruction: the mean of the original complete law Q has zero first-coordinate error projection, since local output–hidden independence and scalar unbiasedness give $\mathbb{E}_Q[\widetilde{a}_j \widetilde{w}_{j1}] = s_j p_{j1}$. The discrete lower term must not be transferred to this convex mean or to the full convex hull. It relies on the fixed teacher's denominator five, the dyadic grid and equal signed-group sizes. □

A1.4 True convex proposals and finite decomposition selection

The variation estimate (34) is independent of the probability rows. Let the same original dictionaries define

$$R_* = \min_{\substack{q_{jk}\geq 0\\ \sum_k q_{jk}=1}} \left\| m^{-1}\sum_{j,k} q_{jk} g_{jk} - f_m \right\|_2^2, \qquad L_* = \min_{C \text{ original-legal}} L(C). \tag{42}$$

The image of the product of row simplices is the full convex hull of original legal functions, so (18) gives $0 < c_{\mathrm{gap}} h^4 \leq R_* \leq L_*$.

Corollary 7 (True convex proposal with an inverse-square width remainder)**.** *Suppose rational feasible rows $q^\star$ have exact mean $\mu^\star$ with $\|\mu^\star - f_m\|_2^2 \leq R_* + \varepsilon_{\mathrm{opt}}$. Their prefix-flow decomposition and finite kernel loss selection give, for every $\eta > 0$, one original-legal vector satisfying*

$$\begin{aligned} &L(C^\star) \leq R_* + \varepsilon_{\mathrm{opt}} + 32400h^2/m^2 + \eta \leq L_* + \varepsilon + 32400h^2/m^2,\\ &\frac{L(C^\star)}{L_*} \leq 1 + \frac{32400}{c_{\mathrm{gap}} m^2 h^2} + \frac{\varepsilon}{L_*}, \qquad \varepsilon = \varepsilon_{\mathrm{opt}} + \eta. \end{aligned} \tag{43}$$

Proof. Apply (26) to these same rows and their own exact mean. Equation (34) gives $V \leq 180h$, and the assumed proposal accuracy and quartic floor yield the displayed bounds. □

These feasible rows may change all scalar marginals and need not satisfy the cancellation table. Their accuracy concerns the true objective of the actual rational proposal; a numerical solver status is insufficient, and rationalization error must be included. The corollary supplies no convex solver-convergence theorem. Numerically, the analytic constant is $c_{\mathrm{gap}} \approx 1.12223 \times 10^{-12}$. Even with optimization and selection errors omitted, making the displayed relative remainder at most 0.1 requires $mh \gtrsim 5.3732 \times 10^8$. This is the scale of this conservative sufficient certificate, not a necessary width for the actual algorithm. Arbitrary integral-flow extraction, without loss selection from the decomposition, only gives

$$L(C) \leq (\sqrt{R_* + \varepsilon_{\mathrm{opt}}} + 180h/m)^2. \tag{44}$$

It cannot be substituted for the additive bound above. Exact shared-cell compression in Appendix A3 remains a compatible source of proposals: rational aggregate counts λ_{ck} are realized by $q_{jk} = \lambda_{ck}/n_c$ in that cell. The prefix construction then coordinates these rows across fine cells within each signed group.

A1.5 Pointwise variation and geometric obstructions

The norm order in (22) is substantive. It uses the norm of pointwise total variation, rather than the generally larger sum of Hilbert norms of increments. For example, with independent standard Gaussians, consider

$$H(u, X) = \sigma(X_1 + uX_2 + hX_3) - \sigma(X_1 + uX_2), \qquad 1 \leq u \leq 2, \quad 0 < h \leq 1/10. \tag{45}$$

For each input its variation is at most $h|X_3|$, and its endpoint has absolute value at most $h|X_3|$, giving the analogue of $V \leq 2h$. Its strong L^2 derivative, however, is

$$\partial_u H = X_2[\mathbf{1}_{\{X_1+uX_2+hX_3>0\}} - \mathbf{1}_{\{X_1+uX_2>0\}}], \qquad ch \leq \|\partial_u H\|_2^2 \leq h/\pi \tag{46}$$

uniformly in u, for an absolute $c > 0$. Conditioning on X_2, X_3 and bounding the X_1 density proves the upper bound. For the lower bound restrict $X_2, X_3 \in [1, 2]$: the disagreement interval has length at least h and lies in $[-21/5, -1]$, where the density has a positive lower bound. Dominated difference quotients prove the derivative and absolute continuity. Its Hilbert total variation is therefore $\int_1^2 \|\partial_u H\|_2 du = \Theta(\sqrt{h})$. Taking Hilbert norms before summing gate movements loses the factor used above.

ReLU alone also does not give a universal h^2/m^2 variance bound. In dimension $d = m$, use independent Gaussian coordinates and candidates $g_{j0} = 0, g_{j1} = h\sigma(X_j)$ with probabilities $1/2, 1/2$.

These are original adjacent output endpoints $0, h$ for original output $h/2$ and aligned hidden direction e_j. With $\mu = h(2m)^{-1}\sum_j \sigma(X_j)$ and $\epsilon_j = 2C_j - 1$, every assignment satisfies

$$\begin{aligned}\|\widehat{f}_C - \mu\|_2^2 &= \frac{h^2}{4m^2}\left[\left(\frac{1}{2} - \frac{1}{2\pi}\right)m + \frac{1}{2\pi}\left(\sum_j \epsilon_j\right)^2\right] \\ &\geq \frac{1 - 1/\pi}{8}\frac{h^2}{m}.\end{aligned} \tag{47}$$

This also lower-bounds every marginal-preserving coupling's variance. This growing-dimensional example shows why the geometry in Theorem 3 matters. Beyond the two ordered groups, V remains a sufficient geometric quantity, but arbitrary noncoplanar teachers need not have $V = O(h)$.

Focused tests cover the combinatorial construction and rational decomposition, including four passing certified-prefix runtime tests. Appendix A7.1 reports all twelve paired P/Q cases for (26). These experiments evaluate the prescribed local laws, not the convex proposal in (43). Finite-case comparisons with SR conditional expectation are mixed, and the covariance construction has not been tested on trained models.

A2 SR MEAN PROJECTIONS AND A CONTROLLED PREFIX COMPARISON

All constants in this section are independent of the allowed width and dyadic step unless a subscript indicates otherwise. The teacher is exactly (2).

Theorem 8 (Affine-resistant product-law mean)**.** *Let $B_P = m^{-1}\sum_j \mathbb{E}_{P_j}F_j - f_m$ and let $\Pi_{\mathcal{A}}$ be Gaussian orthogonal projection onto $\operatorname{span}\{1, X_1, X_2, X_3\}$. For every coupling with complete local marginals P_j, and every realization-dependent affine function ℓ_C,*

$$\mathbb{E}\|\widehat{f}_C + \ell_C - f_m\|_2^2 \geq \|(I - \Pi_{\mathcal{A}})B_P\|_2^2 = \|B_P\|_2^2 - (\mathbb{E}_X B_P)^2, \qquad |\mathbb{E}_X B_P| \leq 10h^2. \tag{48}$$

In particular, when $h \leq 2^{-9}$ and $nh \geq 64$, this loss is at least $c_{\mathrm{SR}}h^4\log(1/(256h)) - 100h^4$.

Proof. The identity $\sigma(t) - \sigma(-t) = t$, local output–hidden independence and scalar unbiasedness give $\mathbb{E}_{P_j}(\widetilde{s}_j\widetilde{p}_j) = s_j p_j$. Hence $B_P(X) = B_P(-X)$ and its input-linear projections vanish. For a hidden endpoint $w + d$, every point of the segment from w to $w + d$ has first coordinate at least $6/5 - h \geq 11/10$. The Hessian of $w \mapsto \|w\|$ has operator norm at most one there. Taylor's formula, $\mathbb{E}d = 0$ and $\|d\|^2 \leq 3h^2$ therefore imply

$$|\mathbb{E}\|w + d\| - \|w\|| \leq 3h^2/2, \qquad |\mathbb{E}_X B_P| \leq \frac{3s}{2\sqrt{2\pi}}h^2 \leq 10h^2,$$

using $\mathbb{E}_X\sigma(w^\top X) = \|w\|/\sqrt{2\pi}$. For each assignment, orthogonal projection minimizes affine-corrected loss. Jensen's inequality applied to $(I - \Pi_{\mathcal{A}})(\widehat{f} - f_m)$ proves (48); infinite expected loss causes no exception. Finally (7) is a lower bound on the squared mean itself. Subtracting $100h^4$ proves the assertion. This projection argument supplies no h^2/m^2 lower term: an input-linear correction can remove the arithmetic projection in Appendix A1.3. □

Lemma 9 (Deterministic quadrature of the actual product mean)**.** *Let $B_{P,\infty}$ average the actual product-SR mean residual continuously over the uniform direction interval $I = [3/2, 7/4]$ at both scales. Then*

$$\|B_{P,\infty}\|_2^2 \leq Ch^4[1 + \log(1/h)], \qquad \|B_P - B_{P,\infty}\|_2 \leq Ch/m. \tag{49}$$

For all sufficiently small allowed h, also $\|B_{P,\infty}\|_2^2 \geq ch^4\log(1/h)$.

Proof. Apply (78)–(79) with both densities equal to $4\mathbf{1}_I$ and with the product endpoint laws. Their first/third probabilities are independent of u, so their functional moment is constant on I, uniformly bounded above and bounded away from zero by (10). Its zero extension has uniformly bounded variation. The tent estimate and the $O(h^2)$ tangential remainder give both continuum claims by the triangle and reverse triangle inequalities. This uses the continuum calculation only, with every actual tangential endpoint restored.

For the deterministic quadrature claim, fix X and one scale λ. Averaging the first/third hidden endpoints gives a convex function of the middle coordinate. Middle-coordinate SR gives its piecewise linear interpolant, still convex, with slopes between 0 and λX_2. Subtract the teacher hinge. Jensen's inequality makes the residual nonnegative. If $X_2 > 0$, its derivative is nonnegative before the teacher kink and nonpositive after it; the same statement in the reversed coordinate applies when $X_2 < 0$. At the kink, unbiasedness bounds its height by $h \sum_i |X_i|/2$, since the mean positive part of a zero-mean perturbation equals half its mean absolute value. Thus its total variation on the real u line, and on I, is at most $h \sum_i |X_i|$. If $X_2 = 0$ the residual is constant. Output SR averages to the signed original output and merely multiplies this bound by s.

Each shifted grid takes one point in each of the n equal subintervals of I. On each subinterval its point value differs from its interval average by at most the local variation; averaging the n errors bounds the discrepancy by total variation divided by n. Sum the two signed contributions, take the Gaussian norm and use $n = m/2$. This proves the second bound without a random-teacher estimate. □

Theorem 10 (Full independent-SR order and the same-prefix control)**.** *There are absolute $c, C > 0$ such that for every allowed width and dyadic $h \le 1/10$,*

$$c(h^2/m + h^4 \log(1/h)) \le \mathbb{E}_{\mathrm{SR}} L \le C(h^2/m + h^4 \log(1/h)). \tag{50}$$

The prefix construction applied respectively to P_j and Q_j gives, for $R = P, Q$,

$$\|B_R\|_2^2 \le \mathbb{E}_{R\text{-}prefix} L \le \|B_R\|_2^2 + 32400 h^2/m^2. \tag{51}$$

For every fixed $c_0 > 0$ there are positive constants and a step threshold, depending only on c_0, such that uniformly over allowed $mh \ge c_0$ below that threshold,

$$\mathbb{E}_{P\text{-}prefix} L = \Theta_{c_0}(h^4 \log(1/h)), \qquad \mathbb{E}_{Q\text{-}prefix} L = \Theta_{c_0}(h^4).$$

These two orders also hold for expected loss after an optimal affine correction chosen separately for each assignment.

Proof. The upper bound in (50) follows from Lemma 9, the local $5h^2$ second-moment bound and the bias–variance identity; $h^2/m^2 \le h^2/m$. The scalar output variance is at least $4h^2/25$, because its fractional position belongs to $\{1/5, 2/5, 3/5, 4/5\}$ even for negative outputs. Conditional independence of outputs given all hidden endpoints yields the integrated variance bound

$$\mathrm{Var}_{\mathrm{SR}}(\widehat{f}) \ge \frac{1}{m^2} \sum_j \frac{4h^2}{25} \frac{\mathbb{E}\|\widetilde{p}_j\|^2}{2} \ge \frac{2h^2}{25m},$$

since $\|\widetilde{p}_j\| \ge 11/10 > 1$. If $mh \ge 128$ and $h \le 2^{-9}$, (7) supplies the logarithmic lower term, with $\log(1/(256h)) \ge \log(1/h)/9$. If $mh < 128$, the ratio of that term to h^2/m is $mh^2 \log(1/h) \le 128h \log(1/h)$, uniformly bounded. At the finitely many coarser allowed dyadic levels, the endpoint quartic floor controls $h^4 \log(1/h)$ uniformly in m. Combining the variance and mean or gap bounds proves the lower half of (50) in every case.

The pointwise variation calculation (32)–(34) is independent of the probability rows, so the same flow construction preserves either choice with the same $32400h^2/m^2$ bound. This proves (51). For Q, Theorem 3 and the affine-resistant endpoint floor give quartic order. For P, Lemma 9 and $mh \ge c_0$ imply

$$\|B_P\|_2 \ge \sqrt{c}\, h^2 \sqrt{\log(1/h)} - C_{c_0} h^2.$$

For sufficiently small h this is a logarithmic squared-mean lower bound; the lemma and (51) give the matching upper bound. This argument explicitly covers $c_0 \le mh < 128$ by an asymptotic threshold, without enlarging the finite numerical regime of (7). Theorem 8 removes only $100h^4$ from this P lower bound. Zero correction supplies both affine upper bounds, and the gap supplies the affine Q lower bound. □

A2.1 A FINITE CURVED-DIRECTION OBSTRUCTION TO PAIRWISE MOMENT FEASIBILITY

The next result concerns the local-law feasibility needed before applying prefix rounding outside the shared-plane family. It gives no asymptotic loss or variation bound for a noncoplanar network.

Lemma 11 (PSD pairwise moments do not certify a legal local law)**.** *There is an anchor on a fixed curved direction family and a finite dyadic step at which all original hidden and output scalars are off-grid, but the signed local normal second moment cannot vanish under any legal three-bit hidden law. Nevertheless, positive definite covariance matrices with individually feasible Bernoulli pair marginals formally make that moment vanish.*

Proof. Choose the numerical constants $h_0 = 1/64$, $e = 1/10000$, $\kappa = 1/8$, $u_0 = -7/16$, and fix the teacher parameters

$$p(u) = (A, 13/8 + eh_0 + u/8, C + \kappa u^2), \qquad A = 257/128 + eh_0, \quad C = 991/2048 + eh_0.$$

The occurrence of h_0 here specifies a fixed number, not a teacher that changes during future grid refinements. Since $\det(p, p', p'') = A\kappa/4 \neq 0$, these directions are not in a common plane through the origin. At u_0 use equal proportions of scales one and two with independent output rounding and original outputs $s = 6/5$ and $-s$. All three scale-one hidden SR probabilities equal $r = 1/2 + e$; all three scale-two probabilities equal $q = 2e$. Both output scalars are also off-grid at step h_0.

Let $G = (-a, b, c) = (-9635015, 8995007, 10280008)$ and $T = \|G\|^2 = 279422229460338$. Direct differentiation gives $p(u_0) \times p'(u_0) = G/40960000$, so the unit local normal is $\nu = G/\sqrt{T}$. Writing $V_1 = \mathbb{E}[\nu^\top (B - r\mathbf{1})]^2$, enumeration of the eight vertices gives

$$V_1 \geq \min_{R \in \{-1,1\}^3} \frac{[G \cdot (5000R - \mathbf{1})]^2}{10000^2 T} = \frac{17422634965849}{279422229460338} > \frac{31}{500}.$$

At scale two, every joint law obeys $V_2 = \mathbb{E}[\nu^\top (B - q\mathbf{1})]^2 \leq 3q(1-q) < 3/5000$ by Cauchy–Schwarz and the prescribed scalar variances. Thus the signed, scale-adjusted moment satisfies

$$M = s(V_1/2 - V_2/4) > 1851/50000 > 37/1000.$$

The cells are unchanged in a sufficiently small fixed interval around u_0. Continuity of the normal, the finite vertex minimum and the scalar variances preserves a positive moment obstruction throughout that interval; the restricted direction curve still has nonzero determinant.

For a spurious second-moment solution, put $v = r(1-r)$ and let C_0 have diagonal v and off-diagonals

$$(C_0)_{12} = v\frac{a^2 + b^2 - c^2}{2ab}, \quad (C_0)_{13} = v\frac{a^2 + c^2 - b^2}{2ac}, \quad (C_0)_{23} = v\frac{a^2 - b^2 - c^2}{2bc}.$$

Strict triangle inequalities for a, b, c give positive two-by-two principal minors, while $C_0 G = 0$ and $\det C_0 = 0$; hence C_0 is positive semidefinite of rank two. Set $\lambda = q(1-q)/(2v) = 2/5001$ and $C_* = (1-\lambda)C_0 + \lambda v I$. This matrix is positive definite with diagonal v and $\nu^\top C_* \nu = q(1-q)/2$. Its off-diagonals lie strictly in the Bernoulli Fréchet interval $[-(1-r)^2, r(1-r)]$. Taking an independent scale-two law would therefore give formal $M = 0$.

However, for every actual three-bit tuple $R_i = 2B_i - 1$, $1 - R_1R_2 - R_1R_3 + R_2R_3 \geq 0$ pointwise. The proposed matrix gives

$$1 - 4\big((C_*)_{12} + (C_*)_{13} - (C_*)_{23}\big) - 4e^2 < -49/100,$$

an exact rational contradiction. Thus no joint three-bit law realizes C_*. Pairwise/PSD feasibility is therefore insufficient; feasibility must be checked in the true eight-state probability polytope. This finite obstruction concerns local-law existence. Continuous-direction mean bounds and variation rates as $h \downarrow 0$ require separate analysis. □

A3 Shared endpoint cells: complete coupling and convex-compression proof

The local covariance change in Appendix A5 removes a logarithmic mean error. A separate coupling across neurons can reduce the remaining selection variance while preserving those new complete local laws. For the fixed signed family, neurons in the same scale group and the same middle-coordinate grid cell have identical candidate functions. Coordinating their state counts improves the variance bound from $O(h^2/m)$ to $O(h/m^2)$, without adding candidates or neurons.

The rounding primitive is the established bipartite dependent-rounding method of Gandhi, Khuller, Parthasarathy, and Srinivasan, *Dependent Rounding and its Applications to Approximation Algorithms*, JACM 53(3), 324–360 (2006), (Gandhi et al., 2006), `https://doi.org/10.1145/1147954.1147956`. We use its edge-marginal and floor/ceiling degree guarantees. The function-valued potential below connects this rounding primitive to boundary cancellation at the original endpoints.

A3.1 A SHARED-CANDIDATE CELL LEMMA

Let $\mathcal{H}$ be a real Hilbert space, let $f \in \mathcal{H}$, and partition m neurons into nonempty cells c, of sizes n_c. All neurons in cell c have the same finite legal dictionary $A_c = \{F_{ck} : 1 \le k \le K_c\} \subset \mathcal{H}$. The network normalization is $1/m$; the functions F_{ck} are unscaled. Give each neuron an arbitrary local probability row $q_{jk} \ge 0$, $\sum_k q_{jk} = 1$, on its cell dictionary. The rows need not agree and may contain zeros. Define

$$\begin{aligned} \lambda_{ck} &= \sum_{j \in c} q_{jk}, \qquad \xi_{ck} = \lambda_{ck} - \lfloor \lambda_{ck} \rfloor, \qquad \mu = \frac{1}{m} \sum_{c,k} \lambda_{ck} F_{ck}, \\ D_c &\ge \operatorname{diam}(A_c), \qquad \mathcal{V}(q) = \frac{1}{2m^2} \sum_c D_c^2 \sum_k \xi_{ck}(1 - \xi_{ck}). \end{aligned} \tag{52}$$

Any known diameter upper bounds D_c can be used; they need not be sharp.

Lemma (shared-candidate rounding potential). There is a coupling of the complete local rows q_j which selects one legal state at each neuron and satisfies

$$\boxed{\begin{aligned} \mathbb{E}\widehat{f} &= \mu, \qquad \mathbb{P}(C_j = k) = q_{jk}, \\ \mathbb{E}\|\widehat{f} - f\|^2 &\le \|\mu - f\|^2 + \mathcal{V}(q), \\ \operatorname{Var}(\widehat{f}) &\le \mathcal{V}(q), \qquad \mathcal{V}(q) \le \frac{1}{2m^2} \sum_c D_c^2 \min\{n_c, K_c/4\}. \end{aligned}} \tag{53}$$

Here variance means $\mathbb{E}\|\widehat{f} - \mathbb{E}\widehat{f}\|^2$. A deterministic version returns a legal assignment with loss at most $\|\mu - f\|^2 + \mathcal{V}(q)$. It only selects states having positive initial local probability. It does not preserve marginals, which are properties of the randomized coupling.

Proof. In a fixed cell, view its current probability matrix y_{jk} as a bipartite graph with neuron rows and state columns. Only entries strictly between zero and one belong to the fractional support. Initially $y = q$. If that support contains a cycle, alternately increase and decrease its edges. Otherwise it is a forest. Every neuron vertex in the fractional support has degree at least two: its row sum is the integer one, so exactly one fractional entry is impossible. Thus a nonempty tree has two leaves on the state-column side. Use their unique path and again alternate increases and decreases. In either case move to one of the two extreme feasible values of the step, where an edge reaches zero or one. Row sums stay one, and each step removes at least one edge from the fractional support.

Cycle steps preserve all column sums. A path changes only its two endpoint column sums, by opposite amounts. Each endpoint is a leaf and has exactly one fractional incident edge; all its other entries are integral. If its current column sum lies between $\lfloor \lambda_{ck} \rfloor$ and $\lceil \lambda_{ck} \rceil$, moving that edge anywhere in $[0, 1]$ keeps the column sum in the same interval. Once a column sum is integral it cannot be a fractional-support leaf and will not change again. Induction therefore proves that the final integer column count is always one of the original floor and ceiling.

For the randomized version, if the two feasible steps are $-a$ and b, where $a, b > 0$, choose them with probabilities $b/(a+b)$ and $a/(a+b)$. Every matrix entry is then a martingale. At termination it is zero or one, proving $\mathbb{P}(C_j = k) = q_{jk}$, including preservation of every zero-probability state. Cells may be processed with independent randomness.

Let $d_{ck} = \sum_{j \in c} y_{jk}$ be the current column sums, and set $x_{ck} = d_{ck} - \lfloor \lambda_{ck} \rfloor$. Throughout the algorithm $0 \le x_{ck} \le 1$. Consider

$$\Phi(y) = \left\| \frac{1}{m} \sum_{c,k} d_{ck} F_{ck} - f \right\|^2 + \frac{1}{2m^2} \sum_c D_c^2 \sum_k x_{ck}(1 - x_{ck}). \tag{54}$$

A cycle leaves this potential unchanged. Along a path whose endpoint columns are a, b in cell c, the coefficient of the squared step in Φ is

$$\frac{\|F_{ca} - F_{cb}\|^2 - D_c^2}{m^2} \le 0. \tag{55}$$

Consequently the potential is concave on the feasible step interval. The zero-mean randomized step cannot increase its conditional expectation. Initially the potential is $\|\mu - f\|^2 + \mathcal{V}(q)$; at termination its second term is zero and its first term is the actual selected loss. This proves the loss bound in (53). The preserved mean and the Hilbert bias–variance identity prove the variance bound. Alternatively, at each step choose an endpoint with potential no larger than its current value; concavity guarantees such an endpoint. This gives the deterministic version without computing future conditional laws.

Finally, $\sum_k \xi_{ck}(1 - \xi_{ck}) \le K_c/4$, and it is also at most $\sum_k \xi_{ck} \le \sum_k \lambda_{ck} = n_c$. These inequalities prove the last bound in (53). The number of rounding steps is at most $\sum_c n_c K_c$, the initial number of matrix entries. This counts rounding steps, not all graph searches or arithmetic operations. □

A3.2 THE ORIGINAL FIXED SIGNED FAMILY

For (103), define a cell by its signed scale group and the integer $\lfloor \lambda u/h \rfloor$, where $\lambda = 1$ or 2. Its first and third hidden endpoints, output endpoints, and two middle endpoints are then identical across its neurons. Thus it has one common dictionary of sixteen unscaled functions. The different middle-coordinate SR probabilities cause no difficulty for the lemma.

For every dyadic $h \le 1/10$, the interval endpoints $3/2, 7/4, 3, 7/2$ are grid points. The two groups' middle coordinates lie strictly inside $(3/2, 7/4)$ and $(3, 7/2)$. Writing B for the number of nonempty cells gives

$$B \le \min\left\{m, \frac{3}{4h}\right\}. \tag{56}$$

For states $S\sigma(q^\top X)$ in one cell, $|S| \le 13/10$, $\|q\| \le 2\sqrt{517/80} + \sqrt{3}/10$, and any two states satisfy $|S - S'| \le h$ and $\|q - q'\| \le \sqrt{3}h$. The exact Gaussian norm $\|\sigma(q^\top X)\|_2 = \|q\|/\sqrt{2}$, together with ReLU Lipschitz continuity, therefore gives the uniform diameter bound

$$\begin{aligned} D_c &\le h\left[\frac{2\sqrt{517/80} + \sqrt{3}/10}{\sqrt{2}} + \frac{13\sqrt{3}}{10}\right] < 6h, \\ \mathcal{V}(q) &\le \frac{72h^2 B}{m^2} \le \frac{54h}{m^2}, \qquad \mathcal{V}(q) \le \frac{18h^2}{m}. \end{aligned} \tag{57}$$

The two bounds on $\mathcal{V}$ hold simultaneously for every choice of local probability rows, not only the cancellation law.

Theorem 12 (Shared-cell coupling after boundary cancellation)**.** *For every fixed signed teacher* (103)*, and every stated dyadic step, there is a law Q^{cell} on its original-width legal assignments which preserves each neuron's complete local law Q_j from Appendix A5 and satisfies*

$$\boxed{\begin{aligned} \mathbb{E}_{Q^{\mathrm{cell}}} \widehat{f} &= \mathbb{E}_Q \widehat{f}, \\ \mathrm{Var}_{Q^{\mathrm{cell}}}(\widehat{f}) &\le \frac{72h^2}{m^2} \min\left\{m, \frac{3}{4h}\right\}, \\ \mathbb{E}_{Q^{\mathrm{cell}}} \|\widehat{f} - f_m\|_2^2 &\le \left(210h^2 + \frac{14h}{m}\right)^2 + \frac{54h}{m^2} \\ &\le 88200h^4 + \frac{94h}{m^2}. \end{aligned}} \tag{58}$$

A deterministic legal assignment attains the same loss bound, or that bound plus a prescribed numerical selection tolerance under the inexact comparisons described below. No neuron or endpoint is added or refitted.

Proof. Apply the lemma to the exact sixteen-state local rows of (110), including their independent middle and output bits. Their complete local marginals are preserved, so the actual mean error

B_Q, including all tangential perturbations, is unchanged. Equation (116), with $m = 2n$, gives $\|B_Q\|_2 \le 210h^2 + 14h/m$. Add (57) to its squared bound. Finally,

$$\left(210h^2 + \frac{14h}{m}\right)^2 \le 88200h^4 + \frac{392h^2}{m^2}, \qquad 392h^2 + 54h \le 94h \quad (h \le 1/10). \tag{59}$$

The lemma also supplies the deterministic assignment. □

Thus the sufficient width condition for the quartic upper order improves to $m^2h^3 \ge c_0 > 0$. Combined with (144), the unrestricted original-legal optimum has order h^4 throughout this wider joint regime. The same order conclusion holds after best affine correction, since correction can only lower the constructed upper bound and (144) already allows it. Along dyadic joint sequences with $h \downarrow 0$ and $m^2h^3 \ge c_0 > 0$, the condition $nh \ge 64$ in (106) eventually holds. The same teacher's independent product SR consequently still has the $\Omega(h^4 \log(1/h))$ lower bound, while (58) is $O(h^4)$. The quartic rate holds in this joint-width regime. At fixed finite width the rate differs, and the explicit constants are asymptotic bounds rather than estimates of moderate-width gains.

Equation (123) applies to the original complete product-SR rows, whose mean has logarithmic bias. The cell coupling preserves the modified rows Q_j, whose leading boundary moment vanishes. Changing local covariance and coordinating different neurons are compatible operations. Cross-neuron dependence also changes the expected bilinear proxy variance in general, so the exact independent- across-neurons proxy equality (125) is not asserted for Q^{cell}.

A3.3 Exact convex compression and a sharper relative guarantee

The same shared dictionaries compress the unrestricted true-function convex relaxation exactly. With the target and candidate norm chosen consistently, including simultaneous projection away from affine functions if desired, its value is

$$\begin{aligned} R_* &= \min_{z_j \in \mathrm{conv}(A_{c(j)})} \left\| \frac{1}{m}\sum_j z_j - f_m \right\|^2 \\ &= \min_{\substack{\lambda_{ck} \ge 0 \\ \sum_k \lambda_{ck} = n_c \text{ for every } c}} \left\| \frac{1}{m}\sum_{c,k} \lambda_{ck} F_{ck} - f_m \right\|^2 . \end{aligned} \tag{60}$$

Indeed $\sum_{j \in c} \mathrm{conv}(A_c) = n_c\, \mathrm{conv}(A_c)$. Conversely, every feasible aggregate row is realized by assigning probability $q_{jk} = \lambda_{ck}/n_c$ to every neuron in that cell. There are $M = 16B$ aggregate variables instead of $16m$. This optimization imposes no SR marginal constraints.

If a feasible aggregate point has objective $F \le R_* + \varepsilon_{\mathrm{opt}}$, apply the deterministic potential rounding to any realizing local rows. For ordinary loss, or for loss in the affine quotient used throughout, one obtains

$$\boxed{\begin{aligned} &L(C) \le L_* + \varepsilon + \frac{54h}{m^2}, \\ &\frac{L(C)}{L_*} \le 1 + \frac{54}{c_{\mathrm{gap}} m^2 h^3} + \frac{\varepsilon}{L_*}, \qquad \varepsilon = \varepsilon_{\mathrm{opt}} + \varepsilon_{\mathrm{sel}}. \end{aligned}} \tag{61}$$

Here $R_* \le L_*$, (144) gives $L_* \ge c_{\mathrm{gap}} h^4$, and projection cannot increase the cell diameters. The actual fractional-count penalty $\mathcal{V}(q)$ in (52) can replace $54h/m^2$. The relative guarantee requires the stated optimization accuracy. Because c_{gap} is small, its analytic ratio bound can be loose at moderate width.

The construction has three distinct outputs. The randomized law in Theorem 12 preserves the complete prescribed Q_j rows and hence all scalar SR marginals. Its deterministic version preserves legality and the loss bound, without marginal constraints. The unrestricted convex optimization in (60) can change all local rows; this freedom is what makes its value a lower bound for the unrestricted legal optimum. Constraining that optimization to the prescribed cancellation rows would not justify the same comparison with L_*.

A3.4 Oracle costs, finite tolerance, and implementation

For the present family the matrix rounding has at most $16m$ steps. Finding a fractional-support cycle or the required forest path is a finite graph computation; even scanning the full support at every step gives polynomial work. With rational initial rows, the matrix updates and the endpoint randomization probabilities are rational. The exact-arithmetic deterministic rule additionally uses an oracle for the Hilbert inner products in (54). The randomized matrix rule needs no kernel evaluations when the analytic diameter bound $6h$ is used; those bounds enter only its guarantee.

For ordinary Gaussian loss, the M distinct candidate functions can share one M-by-M Gram matrix. Their target inner products use at most Mm candidate–teacher kernel evaluations; evaluating the target norm directly uses at most m^2 further evaluations. The target norm is common to both endpoints in each comparison and need not be recomputed. A path changes just two aggregate coefficients, so a maintained Gram-times-coefficient vector can be updated in $O(M)$ arithmetic operations per path step. These are kernel-oracle and arithmetic counts, not bounds on a convex solver's iterations or on the precision needed by its proposal. The compression exploits shared dictionaries in this fixed three-dimensional family; it does not remove the general 2^{d+1} local endpoint enumeration in dimension d.

Exact comparison of tied analytic scores is unnecessary. If each deterministic step chooses an endpoint whose potential is within η of the smaller endpoint value, the total additional loss is at most $16m\eta$. Taking $\eta = \varepsilon_{\mathrm{sel}}/(16m)$ therefore gives the selection tolerance in (61). Certified Gaussian-ReLU kernel enclosures can implement these comparisons by refining both endpoint intervals to the required positive tolerance and selecting the smaller upper endpoint. The guarantee is stated in terms of oracle costs and selection tolerance; convex-solver convergence and end-to-end bit complexity are separate questions.

The explicit constant 94 in (58) can also be retained with inexact selection. Before its final upward simplification, (59) and (57) give $88200h^4 + (466/5)h/m^2$. Choosing every endpoint within $\eta = h/(20m^3)$ of the smaller true potential, for example using score intervals of width at most η, adds at most

$$16m\eta = \frac{4h}{5m^2}, \qquad \frac{466}{5} + \frac{4}{5} = 94. \tag{62}$$

Thus a positive prescribed comparison tolerance suffices for that same displayed deterministic bound; exact comparisons of Gaussian kernel expressions are not required.

A4 Different densities: complete functional and almost-sure proofs

A4.1 Statements and information interface

The moment to cancel is a function when signed groups have different direction densities. Fix $p(u) = (A, u, C)$ on a bounded interval I, $a = \sqrt{A^2 + C^2} > 0$, positive scales λ_r, signed outputs s_r, and positive rational proportions α_r summing to one. Each group has a fixed probability density $\rho_r \le D$. Sample one infinite iid direction sequence per group, independently across groups; width m uses its first $n_r = \alpha_r m$ entries. These sequences are reused at every grid step. Measurable local laws can depend on (u, h), preserve hidden scalar SR marginals, and keep outputs independently SR. They are specified before sampling. Define

$$\Psi_h(u) = \sum_r \frac{\alpha_r s_r}{\lambda_r} \rho_r(u) \mathbb{E}(\tau_r(u,h)/h)^2, \qquad B_{m,h} = \mathbb{E}_{\mathsf{Q}} \widehat{f} - f_m. \tag{63}$$

Here $\mathbb{E}_U$ averages teacher sampling; $\mathbb{E}_{\mathsf{Q}}$ is conditional on the fixed teacher. Let $B_{\infty,h} = \mathbb{E}_U B_{m,h}$.

Theorem 13 (Functional boundary cancellation)**.** *Uniformly over these bounded densities and measurable laws, $\Psi_h = 0$ almost everywhere implies, for sufficiently small h,*

$$\|B_{\infty,h}\|_2^2 \le Ch^4, \qquad \mathbb{E}_U \|B_{m,h}\|_2^2 \le C(h^4 + h^3/m). \tag{64}$$

Independent rounding across neurons additionally gives $\mathbb{E}_U \mathbb{E}_{\mathsf{Q}} L \le C(h^4 + h^2/m)$. No bounded-variation assumption is needed. If the zero extensions instead satisfy $\|\Psi_h\|_\infty + \mathrm{TV}(\Psi_h) \le D_1$

uniformly, then

$$ch^4 \log(1/h)\|\Psi_h\|^2_{L^2(I)} - Ch^4 \le \|B_{\infty,h}\|_2^2 \le Ch^4[1 + \log(1/h)\|\Psi_h\|^2_{L^2(I)}]. \tag{65}$$

The sampling correction remains between zero and Ch^3/m.

The proof (Appendix A4) subtracts a reference tent with integral $k^2\Psi_h(u)/2$. The residual profile $q_u(t,k)$ has zero integral, support of length $O(|k|)$, and amplitude $O(|k|)$. Fourier transformation in x, then Plancherel in u, gives

$$\int_{\mathbb{R}^2} \left| \int_I q_u(x+bu,k)\,du \right|^2 dx\,db = \int_I \int_{\mathbb{R}} \frac{|\widehat{q}_u(\xi,k)|^2}{|\xi|}\,d\xi\,du \le C|k|^4. \tag{66}$$

Zero mass controls low frequencies; Plancherel controls high frequencies. This classical Fourier-slice calculation takes no direction derivative. The actual tangential endpoints are restored with the activation-strip inequality above. The separate h^3/m term is sampling variance of local *mean* residuals. For a common bounded-variation density and direction-independent local laws, the scalar criterion $M_h = \sum_r \alpha_r s_r V_r/\lambda_r$ and deterministic equal-mass quadrature are retained in Appendix A6.

A fixed open density family. Use the two signed scales of (2), but sample directions on $I = [3/2, 7/4]$ with

$$\rho_r(u) = 4(1 + g_r(u)), \qquad \int_I g_r = 0, \qquad \|g_r\|_\infty < 1/100 \quad (r = 1, 2). \tag{67}$$

This is an open family in the affine L^∞ space of normalized densities. Its densities can differ by a fixed amount and need not have bounded variation. Keep the scale-one law (9), and replace the scale-two joint probability by

$$t_2(u,h) = t_2^0(h) - \frac{11}{42}\left(\frac{\rho_1(u)}{\rho_2(u)} - 1\right), \qquad V_2(u,h) = \frac{22}{85}\frac{\rho_1(u)}{\rho_2(u)}. \tag{68}$$

Then $\Psi_h = 0$. Every scale-two pair probability exceeds $1/140$; the scale-one minimum is $1/420$. Almost surely, all sixteen local states have positive mass, every scalar is off every dyadic grid, and no teacher normals are parallel. The transverse gap and nonaffine signal are unchanged.

Corollary 14 (Almost-sure separation on fixed sampled teachers)**.** *For every fixed $c_0 > 0$, almost surely in the two infinite direction sequences, there is a finite b_0 such that simultaneously for every $b \ge b_0$, $h = 2^{-b}$, and every even width with $m^2h^3 \ge c_0$,*

$$c_{\mathrm{gap}}h^4 \le L_* := \min_{C \text{ legal}} L(C) \le \mathbb{E}_{\mathbf{Q}^{\mathrm{cell}}} L \le C_{c_0}h^4, \qquad \mathbb{E}_{\mathrm{SR}} L \ge ch^4 \log(1/h). \tag{69}$$

The cell coupling below preserves every complete modified local row. The constants are uniform over (67)*; b_0 may depend on the sampled sequences. The legal optimum after best affine correction also has quartic order. The SR comparison is ordinary loss.*

This concerns the same teacher prefixes at all steps and widths. The proof combines (64), a positive pointwise SR functional moment, and the cell variance bound below. Doob's maximal inequality controls all widths beyond $m \asymp h^{-3/2}$ with failure probability $O(h^{1/2})$; summability over dyadic levels gives the simultaneous almost-sure conclusion. The density perturbations retain the fixed planar geometry and are not an open ball of all network parameters. This sampled-density family has not been evaluated experimentally.

The common-density hypothesis of Theorem 16 can be removed. The quantity to cancel is then a function of direction. A Fourier estimate controls its zero-mass remainder even when the direction densities and the legal covariance choices are merely measurable. The result below also gives a fixed, one-percent open family of different densities for which this cancellation is feasible on all original dyadic endpoints.

A4.2 THE FUNCTIONAL BOUNDARY MOMENT

Fix a bounded interval I, $A, C \in \mathbb{R}$ with $a = \sqrt{A^2 + C^2} > 0$, a fixed number G of groups, positive scales λ_r, real outputs s_r, and positive rational proportions α_r summing to one. Each group has a fixed probability density ρ_r on I, with $\max_r \|\rho_r\|_\infty \le D < \infty$. The densities need not coincide or have bounded variation. At every direction and step, choose a measurable local law on the original adjacent hidden endpoints, preserving all three scalar SR marginals. Keep the output independently SR and use independent rounding across neurons. These local probability kernels may depend on (u, h); they are specified before the directions are sampled.

For each group sample one infinite iid sequence U_{rj} with density ρ_r, independently across groups, and use its first $n_r = \alpha_r m$ entries at an admissible width. The same sequences are used at all grid steps. With $p(u) = (A, u, C)$, let

$$\begin{aligned} f_m &= \frac{1}{m} \sum_{r,j \le n_r} s_r \sigma(\lambda_r p(U_{rj})^\top X), \qquad X \sim N(0, I_3), \\ e_{r,h}(u; X) &= s_r \left[\mathbb{E}_Q \sigma(q_{r,h}(u)^\top X) - \sigma(\lambda_r p(u)^\top X) \right], \\ B_{m,h} &= \mathbb{E}_Q \widehat{f} - f_m, \qquad B_{\infty,h} = \sum_r \alpha_r \int_I \rho_r(u) e_{r,h}(u)\, du, \\ V_r(u,h) &= \mathbb{E}_Q(\tau_r(u,h)/h)^2, \qquad \Psi_h(u) = \sum_r \frac{\alpha_r s_r}{\lambda_r} \rho_r(u) V_r(u,h). \end{aligned} \tag{70}$$

Here $q_{r,h}(u)$ is an actual original hidden endpoint and τ_r is its component along $(-C, 0, A)/a$. All norms without an indicated measure are Gaussian L^2 norms in X. The expectation $\mathbb{E}_U$ below is over teacher directions, whereas $\mathbb{E}_Q$ is conditional on the teacher and averages only its legal rounding. Outputs use step h in this unscaled representation, or $h/\sqrt{m}$ in the canonical representation, exactly as in Appendices A5, A6, and A3.

Complete statement of Theorem 13. There are constants $C_0, h_0 > 0$, uniform over the stated densities and measurable local laws, depending only on the fixed geometry, group parameters and D, with the following properties.

If $\Psi_h = 0$ almost everywhere, then for $0 < h \le h_0$,

$$\boxed{\begin{aligned} \|B_{\infty,h}\|_2^2 &\le C_0 h^4, \\ \mathbb{E}_U \|B_{m,h}\|_2^2 &\le C_0 \left(h^4 + \frac{h^3}{m} \right), \\ \mathbb{E}_U \mathbb{E}_Q \|\widehat{f} - f_m\|_2^2 &\le C_0 \left(h^4 + \frac{h^2}{m} \right). \end{aligned}} \tag{71}$$

No regularity in u of the densities or the covariance kernels is required for these inequalities. For every fixed teacher the finite legal candidate set contains an assignment whose loss is no larger than its conditional Q-expected loss; (71) bounds the average of those conditional guarantees over the sampled teachers.

For an imperfect cancellation, assume in addition that the zero-extension of Ψ_h satisfies $\|\Psi_h\|_\infty + \mathrm{TV}(\Psi_h) \le D_1$ uniformly in h. Then constants $c_1, C_1 > 0$, also depending on D_1, give

$$\boxed{\begin{aligned} c_1 h^4 \log(1/h) \|\Psi_h\|_{L^2(I)}^2 - C_1 h^4 \le \|B_{\infty,h}\|_2^2 &\\ \le C_1 h^4 \left[1 + \log(1/h) \|\Psi_h\|_{L^2(I)}^2 \right], &\\ 0 \le \mathbb{E}_U \|B_{m,h}\|_2^2 - \|B_{\infty,h}\|_2^2 \le \frac{C_1 h^3}{m}, &\\ \mathbb{E}_U \mathbb{E}_Q \|\widehat{f} - f_m\|_2^2 \le C_1 \left\{ h^4 \left[1 + \log(1/h) \|\Psi_h\|_{L^2(I)}^2 \right] + \frac{h^2}{m} \right\}. & \end{aligned}} \tag{72}$$

The full expected loss is at least the squared mean averaged over teachers. Consequently, when mh is bounded below, quartic teacher-averaged squared mean is equivalent to $\|\Psi_h\|_{L^2(I)} =$

$O(1/\sqrt{\log(1/h)})$. When mh^2 is bounded below the same equivalence holds for the full teacher-averaged expected loss. The bounded-variation hypothesis in this approximate criterion is additional; it is absent from (71).

In the two-bit geometry of (139), its feasible variance intervals $[v_r^-, v_r^+]$ do not depend on u. With $w_r = \alpha_r s_r/\lambda_r$, define the measurable functions

$$\begin{aligned} \ell_h(u) &= \sum_r \min\{w_r\rho_r(u)v_r^-, w_r\rho_r(u)v_r^+\}, \\ U_h(u) &= \sum_r \max\{w_r\rho_r(u)v_r^-, w_r\rho_r(u)v_r^+\}. \\ \Psi_h = 0 \text{ is feasible a.e.} \quad &\iff \quad \ell_h(u) \le 0 \le U_h(u) \text{ a.e.} \end{aligned} \tag{73}$$

This is a pointwise family of scalar interval problems, rather than a single integrated moment test. A measurable realizing law is obtained by mixing the two endpoint choices at each u with weight $-\ell_h(u)/(U_h(u) - \ell_h(u))$ when the denominator is positive; the zero interval needs no mixing. This establishes existence and pointwise finite arithmetic feasibility. It makes no assertion about the cost of representing arbitrary measurable density functions or about higher-dimensional covariance polytopes.

A4.3 A Fourier estimate without direction regularity

Use the rotation of Appendix A6, writing $x = aZ_1$, $b = Z_2$, $k = hY$, and $A_r(u,h) = \tau_r(u,h)/h$. Scalar unbiasedness gives $\mathbb{E}A_r = 0$. Define

$$\begin{aligned} \delta_{u,h}(t,k) &= \sum_r \alpha_r s_r \lambda_r \rho_r(u) \left[\mathbb{E}\sigma(t + A_r(u,h)k/\lambda_r) - \sigma(t)\right], \\ \int_{\mathbb{R}} \delta_{u,h}(t,k)\,dt &= \frac{k^2}{2}\Psi_h(u), \\ q_{u,h}(t,k) = \delta_{u,h}(t,k) - \Psi_h(u)\beta(t,k), &\qquad \int_{\mathbb{R}} q_{u,h}(t,k)\,dt = 0, \end{aligned} \tag{74}$$

where β is the tent in (135). The integral identity follows by integrating the unbiased-shift residual over an interval containing all its shifted hinges. Uniform endpoint bounds imply, for almost every u,

$$\operatorname{supp} q_{u,h}(\cdot,k) \subset [-C|k|, C|k|], \qquad |q_{u,h}(t,k)| \le C|k|, \qquad \|q_{u,h}(\cdot,k)\|^2_{L^2(dt)} \le C|k|^3. \tag{75}$$

For fixed nonzero k, abbreviate this profile to q_u and put $A_q(x,b) = \int_I q_u(x+bu)\,du$. With the convention $\widehat{q}_u(\xi) = \int e^{-i\xi t} q_u(t)\,dt$, Fourier transformation in x, followed by Plancherel in u, gives the exact identity

$$\begin{aligned} \int_{\mathbb{R}^2} |A_q(x,b)|^2\,dx\,db &= \frac{1}{2\pi}\int_{\mathbb{R}}\int_{\mathbb{R}} \left|\int_I e^{i\xi bu}\widehat{q}_u(\xi)\,du\right|^2 db\,d\xi \\ &= \int_I \int_{\mathbb{R}} \frac{|\widehat{q}_u(\xi)|^2}{|\xi|}\,d\xi\,du. \end{aligned} \tag{76}$$

The functions are integrable and square integrable in t on a uniform compact support, and I is bounded. Thus the first Fourier transform follows by Fubini for each fixed b; the two Plancherel equalities and Tonelli's theorem justify (76), including the possibility of an infinite value before the next bound. The factor $1/|\xi|$ comes from the substitution ξb in the second Plancherel calculation. This is the elementary Fourier slice mechanism underlying standard Radon-transform energy formulas; see Melrose and Uhlmann (n.d., Chapter 2). No direction derivative is taken here.

Zero mass gives $|\widehat{q}_u(\xi)| \le |\xi| \int |t q_u(t)| dt \le C|\xi||k|^3$. On $|\xi| \le 1/|k|$ this bounds the weighted integral by $C|k|^4$. On the complementary frequencies, $1/|\xi| \le |k|$, so Plancherel and (75) give the same bound. Since the Gaussian density of (aZ_1, Z_2) is at most $1/(2\pi a)$, (76) and $\mathbb{E}Y^4 = 3$ prove

$$\left\| \int_I q_{u,h}(aZ_1 + uZ_2, hY)\,du \right\|_2^2 \le Ch^4. \tag{77}$$

The case $k = 0$ is zero. This proof needs neither bounded variation of the densities nor regularity of the local laws in u.

Restoring the tangential perturbations still gives an $O(h^2)$ Gaussian norm. Indeed the pointwise comparison from (116), valid even for dependent tangential and transverse perturbations, is bounded by $ChR\mathbf{1}_{\{|aZ_1+uZ_2|\le Ch(|Y|+R)\}}$, where $R = \|(Z_1, Z_2)\|$. When the strip width is at least a fixed multiple of R, the amplitude is at most $Ch^2(|Y| + R)$. Otherwise a nonempty strip forces $|Z_2| \ge cR$, and its length in u is at most $Ch(|Y| + R)/R$. Integrating against any $\rho_r \le D$ gives the same bound. Consequently the actual continuous mean satisfies

$$B_{\infty,h} = T_{\Psi,h} + R_h, \qquad T_{\Psi,h} = \int_I \Psi_h(u)\beta(aZ_1 + uZ_2, hY)\, du, \qquad \|R_h\|_2 \le Ch^2. \tag{78}$$

As in Appendix A6, removing tangential coordinates was only an auxiliary proof step: the mean on the left uses every original hidden and output endpoint. Equation (78) proves the continuous part of (71) when $\Psi_h = 0$.

A4.4 Approximate cancellation and the logarithmic coefficient

For a zero-extended, uniformly bounded-variation function Ψ supported on I, the reference tent obeys

$$\|T_{\Psi,h}\|_2^2 = \frac{3}{4\pi a} h^4 \log(1/h)\|\Psi\|_{L^2(I)}^2 + O(h^4), \tag{79}$$

uniformly over the stated bounds, including signed functions that depend on h. To verify the coefficient, fix $\kappa = |k|$, put $j(v) = (1 - |v|)_+$, and write $j_\varepsilon(v) = \varepsilon^{-1} j(v/\varepsilon)$. The integral of j is one. For $b \ne 0$, the substitution $v = -x/b$ makes the tent average $\kappa^2(\Psi * j_{\kappa/|b|})(v)/(2|b|)$. Thus its Gaussian squared norm in (Z_1, Z_2) is

$$\frac{\kappa^4}{4\pi a} \int_0^\infty \frac{1}{b} \int_{\mathbb{R}} |\Psi * j_{\kappa/b}(v)|^2 e^{-b^2(1+v^2/a^2)/2}\, dv\, db.$$

For $0 < \kappa \le 1$, the region $0 < b < \kappa$ contributes $O(\kappa^4)$: Young's inequality gives $\|\Psi * j_{\kappa/b}\|_2^2 \le Cb/\kappa$. For $\kappa \le b \le 1$, bounded variation gives $\|\Psi * j_{\kappa/b} - \Psi\|_1 \le C\kappa/b$, and hence the difference of their squared L^2 norms is at most $C\kappa/b$. Their supports lie in one fixed compact interval; replacing the Gaussian factor by one costs at most Cb^2. Both errors are integrable against db/b with a bounded total. The range $b > 1$ is bounded using $e^{-b^2/2}$ and Young's inequality. The fixed-k answer is therefore $\kappa^4\|\Psi\|_2^2 \log(1/\kappa)/(4\pi a) + O(\kappa^4)$. For $\kappa > 1$, the elementary tent amplitude bound gives an $O(\kappa^4)$ upper bound. Substitution $\kappa = h|Y|$ and $\mathbb{E}[Y^4|\log|Y||] < \infty$ now prove (79), with the Gaussian tail $|Y| > 1/h$ absorbed into $O(h^4)$.

Combining (78)–(79) with $\|T + R\|^2 \le 2\|T\|^2 + 2\|R\|^2$ and $\|T + R\|^2 \ge \|T\|^2/2 - \|R\|^2$ proves the first two-sided bound of (72). The exact coefficient in (79) belongs to the auxiliary tent. An $O(h^2)$ bound on R_h alone does not give that same coefficient with an $O(h^4)$ remainder for the actual squared mean; (72) is the conclusion used here.

A4.5 Finite teachers and the distinction between the two expectations

The actual local mean residual in (70) has a stronger second-moment bound than a single randomized local error:

$$\sup_{r,u} \|e_{r,h}(u)\|_2^2 \le Ch^3. \tag{80}$$

To see this, endpoint boundedness and scalar unbiasedness give $|e_{r,h}(u; X)| \le Ch\|X\|\mathbf{1}_{\{|\lambda_r p(u)^\top X| \le Ch\|X\|\}}$. Outside this strip every endpoint and the teacher have the same activation sign, so their mean error is exactly zero. Decompose $X = Tn + W$ along the unit teacher normal. Uniformly in u, $\|\lambda_r p(u)\|$ is bounded away from zero. For sufficiently small h, the strip implies $|T| \le C'h\|W\|$. Conditioning on W, the bounded one-dimensional Gaussian density therefore bounds the squared residual integral by $Ch^3\mathbb{E}\|W\|^3$, proving (80).

Independence of the sampled directions and the Hilbert-space variance identity give the exact decomposition and its bound

$$\begin{aligned}\mathbb{E}_U\|B_{m,h}\|_2^2 &= \|B_{\infty,h}\|_2^2 + \sum_r \frac{\alpha_r^2}{n_r}\,\mathrm{Var}_{U\sim\rho_r}(e_{r,h}(U)),\\ \mathbb{E}_U\|B_{m,h}-B_{\infty,h}\|_2^2 &\le \frac{Ch^3}{m}.\end{aligned} \tag{81}$$

On the other hand, for each fixed teacher, bounded parameters, scalar SR variances, output-hidden independence and ReLU Lipschitz continuity give the original local randomized error bound Ch^2, as in (117). Independence across neurons then yields

$$\mathbb{E}_Q\|\widehat{f}-f_m\|_2^2 = \|B_{m,h}\|_2^2 + \mathrm{Var}_Q(\widehat{f}), \qquad \mathrm{Var}_Q(\widehat{f}) \le \frac{Ch^2}{m}. \tag{82}$$

Equations (78)–(82) prove the remaining loss statements of Theorem 13. The interval sum and measurable mixing in (73) prove its feasibility statement. □

The h^3/m term is sampling variance of actual mean residuals; the h^2/m term is variance of legal rounding for a fixed teacher. Neither is a discrepancy estimate for an arbitrary fixed direction grid. Applying the deterministic bounded-variation quadrature argument of Appendix A6 to direction-dependent laws would require a separate uniform variation bound on their residuals. No such assumption is implicit in Theorem 13.

A4.6 A FIXED OPEN DENSITY FAMILY WITH A UNIFORM LEGAL MARGIN

Take the geometry, signed outputs and two scales of Appendix A5, but replace its deterministic grids by two independent direction sequences with densities

$$\begin{aligned}I=[3/2,7/4],\quad s=6/5,\quad p(u)=(6/5,u,7/5),\\ \rho_r(u)=4(1+g_r(u)),\quad \int_I g_r(u)\,du=0,\quad \|g_r\|_\infty<1/100,\qquad r=1,2,\\ f_{2n}(X)=\frac{s}{2n}\sum_{j=1}^{n}[\sigma(p(U_{1j})^\top X)-\sigma(2p(U_{2j})^\top X)].\end{aligned} \tag{83}$$

This is an open set in the affine L^∞ space of normalized densities on I. The two densities can differ by a fixed amount independent of h, and they can be discontinuous or fail to have bounded variation.

For every dyadic $h=2^{-b}\le 1/10$, use the scale-one joint law from (110). If $t_2^0(h)$ denotes the scale-two entry of that table, replace it at direction u by

$$\begin{aligned}V_1(u,h)=\frac{11}{85},\qquad V_2(u,h)=\frac{22}{85}\frac{\rho_1(u)}{\rho_2(u)},\\ t_2(u,h)=t_2^0(h)-\frac{11}{42}\left(\frac{\rho_1(u)}{\rho_2(u)}-1\right),\qquad \Psi_h(u)=\frac{s}{2}\left[V_1\rho_1(u)-\frac{1}{2}V_2\rho_2(u)\right]=0.\end{aligned} \tag{84}$$

The other three joint probabilities are reconstructed from the same first/third scalar marginals. Continue to round the middle hidden bit and the output independently by SR, independently of the pair, and use independent rounding across neurons. Since the coefficient of t_2 in its variance is $-84/85$, the displayed variance is exact. Moreover,

$$\left|\frac{\rho_1}{\rho_2}-1\right|<\frac{2}{99},\quad |t_2-t_2^0|<\frac{11}{2079}<\frac{1}{140},\quad \min_{\text{four scale-two states}} Q(B_1,B_3)>\frac{1}{140}. \tag{85}$$

Indeed every probability changes by at most $|t_2-t_2^0|$, and its original minimum is $1/70$. Scale one's minimum remains $1/420$. Thus (84) is uniformly inside the legal two-bit Fréchet constraints, with a margin independent of direction, width and dyadic residue. The one-percent neighborhood does not shrink as the grid is refined.

For the same teacher under product SR, let d_1, d_2 be the two variances in (111). Its functional moment is uniformly positive:

$$\begin{aligned}\Psi_h^{\mathrm{SR}}(u) &= \frac{s}{2}\left[d_1\rho_1(u) - \frac{d_2}{2}\rho_2(u)\right] \\ &\geq \frac{2s}{2125}\left(193 - \frac{644}{100}\right) =: \psi_* > 0 \qquad \text{a.e. on } I.\end{aligned} \tag{86}$$

Here $d_1 - d_2/2 \geq 193/2125$ and $d_1 + d_2/2 \leq 644/2125$. Positivity makes a bounded-variation assumption unnecessary for its logarithmic lower bound: since the tent is nonnegative, $T_{\Psi^{\mathrm{SR}},h} \geq \psi_* T_{\mathbf{1}_I,h}$ pointwise. The lower tent estimate (137), or (79) with $\Psi = \mathbf{1}_I$, and (78) consequently give $\|B^{\mathrm{SR}}_{\infty,h}\|_2^2 \geq ch^4\log(1/h) - Ch^4$.

Corollary (robust original-endpoint separation). Uniformly over the fixed density class (83), the law (84) and the same teacher's product SR satisfy, at sufficiently small dyadic steps,

$$\begin{aligned}\mathbb{E}_U\mathbb{E}_Q\|\widehat{f} - f_m\|_2^2 &\leq C\left(h^4 + \frac{h^2}{m}\right), \\ \mathbb{E}_U\mathbb{E}_{\mathrm{SR}}\|\widehat{f} - f_m\|_2^2 &\geq ch^4\log(1/h) - Ch^4.\end{aligned} \tag{87}$$

There is also an almost-sure statement for teachers sampled once. For each fixed $c_0 > 0$, with probability one in the two infinite direction sequences, there is a finite level b_0 such that simultaneously for every $b \geq b_0$, $h = 2^{-b}$, and every even width $m = 2n$ with $mh^2 \geq c_0$,

$$\boxed{\begin{aligned}\min_{C \text{ legal}}\|\widehat{f}_C - f_m\|_2^2 \leq \mathbb{E}_Q\|\widehat{f} - f_m\|_2^2 &\leq C_{c_0}h^4, \\ \mathbb{E}_{\mathrm{SR}}\|\widehat{f} - f_m\|_2^2 &\geq ch^4\log(1/h).\end{aligned}} \tag{88}$$

The constants are uniform in the density class; the finite threshold b_0 can depend on the sampled sequences. Every finite teacher is fixed independently of h. No fresh directions are drawn when its grid is refined.

Every teacher whose directions lie in I has the nonaffine signal bound (107): the two pointwise bounds on $J(u)$ used there give $\langle f_m, H_2(Z_1)\rangle \leq -17/(100\sqrt{\pi})$ for arbitrary choices of the two sets of n directions. The endpoint gap (109) and norm bounds are also unchanged. Therefore the floor (144), including its stated explicit constant, applies:

$$\mathrm{dist}^2(f_m, \mathcal{A}) \geq \frac{289}{10000\pi}, \qquad \|g + \ell - f_m\|_2^2 \geq c_{\mathrm{gap}}h^4 \tag{89}$$

for every affine ℓ and every g in the full convex hull of this teacher's original legal assignments. Combined with (88), the unrestricted legal optimum has quartic order, almost surely in the stated joint regime. The same quartic order holds after best affine correction. The SR comparison in (88) concerns ordinary population loss.

Almost surely, all sampled directions are distinct across the two infinite sequences, and none of their scale-one or scale-two middle coordinates lies on any dyadic grid. These are countable unions of null events under continuous densities. The other hidden coordinates and both outputs have the fixed uncancelled denominator five. Hence all teacher scalars are off-grid at every stated level and no pair of teacher normals is parallel, almost surely. Every legal hidden endpoint still has $|\tau| \geq h/\sqrt{85}$, regardless of the direction-dependent probabilities in (84).

To justify the simultaneous width assertion, fix a group, a law Q or SR, and a dyadic step. Center its actual local mean residuals to obtain iid Hilbert-space random variables X_j with $\mathbb{E}\|X_j\|_2^2 \leq Ch^3$, by (80). Let $S_n = \sum_{j\leq n} X_j$. The norm of this martingale is a nonnegative submartingale. Doob's L^2 maximal inequality gives $\mathbb{E}\max_{n\leq M}\|S_n\|_2^2 \leq 4Ch^3M$; see Geng (2016, Corollary 3.3). Applying it on the dyadic blocks $N \leq n < 2N$, $2N \leq n < 4N$, and so on gives

$$\begin{aligned}&\mathbb{P}\left(\sup_{n\geq N}\|S_n/n\|_2 > t\right) \leq \sum_{\ell\geq 0}\frac{4Ch^3\,2^{\ell+1}N}{(2^\ell Nt)^2} \leq \frac{C'h^3}{Nt^2}, \\ &N = \left\lceil\frac{c_0}{2h^2}\right\rceil, \quad t = h^2 \quad\Longrightarrow\quad \mathbb{P}\left(\sup_{n\geq N}\|S_n/n\|_2 > h^2\right) \leq C_{c_0}h.\end{aligned} \tag{90}$$

Take the union of these events over the two groups and the two laws, always using the same sampled teacher sequences. Their probabilities are summable over the distinct levels $h = 2^{-b}$. Borel–Cantelli therefore makes both finite means differ from their respective continuous means by at most Ch^2, simultaneously at all the stated widths and all sufficiently fine levels, almost surely. This conclusion needs no independence between grid levels or between the two laws' residuals. Equation (78) gives the quartic Q mean; (86) gives the logarithmic SR mean; (82) controls the Q selection variance. These facts prove (88). The signal and common affine-region proof just given establish (89).

The separation is stable under fixed, unequal direction densities and direction-dependent legal covariances within this signed family. The common two-dimensional teacher plane and the two-bit covariance margin remain essential assumptions. Extending the result to general trained networks or high-dimensional endpoint laws requires additional geometric and computational analysis.

Corollary (cell coupling for the different-density family). The same fixed teachers (83) admit a coupling Q^{cell} which preserves each complete direction-dependent local row from (84), and satisfies

$$\begin{aligned} \mathbb{E}_{Q^{\mathrm{cell}}}\widehat{f} = \mathbb{E}_{Q}\widehat{f}, \qquad \mathrm{Var}_{Q^{\mathrm{cell}}}(\widehat{f}) \le \frac{54h}{m^2}, \\ \mathbb{E}_U \mathbb{E}_{Q^{\mathrm{cell}}} \|\widehat{f} - f_m\|_2^2 \le C\left(h^4 + \frac{h^3}{m}\right) + \frac{54h}{m^2}. \end{aligned} \tag{91}$$

For every fixed $c_0 > 0$, almost surely in the same two infinite teacher sequences, at every sufficiently fine dyadic level and simultaneously at all even widths with $m^2h^3 \ge c_0$,

$$\boxed{\begin{aligned} c_{\mathrm{gap}}h^4 \le \min_{C \text{ legal}} \|\widehat{f}_C - f_m\|_2^2 \le \mathbb{E}_{Q^{\mathrm{cell}}}\|\widehat{f} - f_m\|_2^2 \le C_{c_0}h^4, \\ \mathbb{E}_{\mathrm{SR}}\|\widehat{f} - f_m\|_2^2 \ge ch^4\log(1/h). \end{aligned}} \tag{92}$$

The legal optimum after best affine correction also has quartic order in this wider regime, by the same lower and upper bounds.

Proof. A group and its middle-coordinate grid cell determine the same sixteen original candidates for every neuron in that cell. There are at most $3/(4h)$ nonempty cells, and their diameters are less than $6h$, exactly as in (56)–(57). The shared-candidate lemma (53) explicitly permits different local rows within one cell. It therefore applies to the direction-dependent laws (84), preserving their actual mean and bounding the selection variance by $54h/m^2$. Equation (81) proves (91). In $m^2h^3 \ge c_0$, its two finite-width terms are at most $c_0^{-1/2}h^{9/2}$ and $54c_0^{-1}h^4$, so the teacher-averaged order is quartic.

For the almost-sure assertion, apply (90) instead with $N = \lceil\sqrt{c_0}/(2h^{3/2})\rceil$ and $t = h^2$. The failure probability is at most $Ch^{1/2}/\sqrt{c_0}$, which is still summable over the distinct dyadic levels. The same two groups and two actual mean residual laws, Q and product SR, thus satisfy the required Ch^2 approximation to their continuous means simultaneously at every stated width. Cell coupling preserves the Q mean and supplies the variance in (91); (86) gives the same teacher's SR logarithmic lower bound; (89) supplies the quartic floor. These prove (92). This improvement uses dependence between neurons under Q^{cell}; it does not improve the independent law's variance in (82). The complete modified rows are preserved, and all scalar SR marginals remain correct, but the exact proxy equality that relies on independence across neurons is not asserted for the cell coupling. □

A4.7 Sorted sampled prefixes at reciprocal-step width

The following strengthening uses the same sampled-once sequences and the known-density local laws (83)–(85). At width $m = 2n$, take the first n directions in each sequence and sort those finite lists before rounding. Sorting uses every sampled direction and preserves both the teacher and its exact local-law mean.

Theorem 15 (Expected and almost-sure density-prefix bounds)**.** *For each fixed pair of densities in* (67) *and every sufficiently small allowed dyadic h, the sorted complete-row prefix law satisfies*

$$\mathbb{E}_U\mathbb{E}_{Q\text{-}prefix}L \le C(h^4 + h^3/m) + 32400h^2/m^2 \qquad (m \text{ any even positive width}). \tag{93}$$

Consequently, for every fixed $c_0 > 0$, its teacher-averaged expected loss and the teacher-averaged legal optimum have order $\Theta_{c_0}(h^4)$ when $mh \ge c_0$, also after best realization-specific affine correction.

Fix the density pair and its measurable local laws before sampling. On a single probability-one event in the two infinite direction sequences, for every $c_0 > 0$ there is a finite b_0 such that simultaneously for every $b \geq b_0$, $h = 2^{-b}$, and every even width

$$m \geq c_0 h^{-1}, \tag{94}$$

one has

$$c_{\mathrm{gap}} h^4 \leq \mathbb{E}_{Q\text{-prefix}} L \leq C_{c_0} h^4, \qquad \mathbb{E}_{\mathrm{SR}} L \geq c h^4 \log(1/h). \tag{95}$$

The same quartic upper bound holds for every positive-support assignment satisfying the Q prefix constraints. For the probability classes defined in Section 2, evaluated conditionally on each sampled teacher, the same event gives

$$\begin{gathered} L_*^\kappa, J_{\mathrm{scalar}}^\kappa, J_Q^\kappa = \Theta_{c_0}(h^4), \qquad J_P^\kappa = \Theta_{c_0}(h^4 \log(1/h)), \\ \kappa \in \{\mathrm{ordinary}, \mathrm{const}, \mathcal{A}\}. \end{gathered} \tag{96}$$

Here ordinary denotes the uncorrected quantities, and corrections are best realization-specific corrections. Constants are uniform over the stated density class; b_0 may depend on the sequences, densities and c_0. The random-teacher width condition is sufficient. The arithmetic h^2/m^2 lower bound, and hence width necessity, belongs to the separate deterministic grid family.

Proof. The variation proof (31)–(34) needs only the ordered range $u \in [3/2, 7/4]$, not spacing or probability regularity. Conditional on any sampled teacher it therefore gives $\mathbb{E}_{Q\text{-prefix}} L \leq \|B_{m,h}\|_2^2 + 32400 h^2/m^2$. Equation (81) and functional cancellation prove (93). At $mh \geq c_0$, both remaining terms are $O_{c_0}(h^4)$. The uniform endpoint gap and nonaffine-signal floor established above apply to every sampled teacher and every affine correction, yielding all lower bounds and the expected-order assertions.

For completeness, the exact conditional prefix law exists even for irrational measurable probability entries. The bounded integral-flow polytope contains the fractional flow and is the convex hull of finitely many integral flows. At each fixed width there are finitely many possible integer prefix-capacity patterns, selected measurably by the row entries. For each pattern triangulate its fixed integral-flow polytope once and use the barycentric coordinates of the first containing simplex. This gives a measurable finite-dimensional law. Sorting breaks ties by original index. Thus the displayed conditional expectations are well defined. Only rational finite-bit implementation, not existence, would require extra representation assumptions on the supplied density ratio.

Work in $H = L^2(\gamma_3)$ before sorting, fixing one group and one law P or Q. Write $e_h(u)$ for its actual local mean residual, with bounded signed output included. Independent output rounding and unbiasedness give zero residual wherever all endpoint gates agree, because there $\mathbb{E}[\widehat{s}\widehat{w}] = sw$. Endpoint displacement is $O(h)$ and the group normal norms are bounded below. Thus, with $\nu(u) = p(u)/\|p(u)\|$,

$$|e_h(u; X)| \leq Ch \|X\| \mathbf{1}_{\{|\nu(u)^\top X| \leq Ah\|X\|\}}, \qquad \sup_u \|e_h(u)\|_H^2 \leq Ch^3. \tag{97}$$

The last bound follows by splitting a Gaussian into its normal coordinate and independent orthogonal component, as in (80). Finite measurable local laws and this domination give strongly measurable H-valued maps and Bochner expectations.

Joint strip geometry. Write $X = R_0 \omega$, with ω uniform on S^2, independent radius and $\mathbb{E}R_0^2 = 3$. Directly,

$$\delta = \sin \angle(\nu(u), \nu(v)) = \frac{\sqrt{17/5}|u - v|}{\|p(u)\| \|p(v)\|} \asymp |u - v|.$$

These angles are acute. A single strip $|\nu^\top \omega| \leq Ah$ has probability at most Ah, since $\nu^\top \omega$ is uniform on $[-1, 1]$. This suffices if $\delta \leq Dh$ for a fixed sufficiently large multiple D of A. Otherwise choose orthonormal coordinates in the normals' plane. The two strips imply $|y_1| \leq Ah$ and $|\cos(\theta) y_1 + \delta y_2| \leq Ah$, hence $|y_2| \leq 2Ah/\delta$. Choosing D large and then h small puts their intersection inside $|y| \leq 1/2$. The projected spherical density is $[2\pi\sqrt{1 - |y|^2}]^{-1}$ there, hence bounded, and the parallelogram area is $4A^2h^2/\delta$. Combining both cases gives

$$\Pr_\omega\{|\nu(u)^\top \omega|, |\nu(v)^\top \omega| \leq Ah\} \leq \frac{Ch^2}{h + |u - v|}.$$

Using (97), radial independence and $\mathbb{E}R_0^2 = 3$, without bounding the Gaussian radius, yields

$$|K_h(u,v)| := |\langle e_h(u), e_h(v)\rangle_H| \le \frac{Ch^4}{h + |u-v|}. \tag{98}$$

Centering and the squared-norm martingale. Put $\mu(du) = \rho(u)du$, $\bar{e} = \mathbb{E}_\mu e_h(U)$ and $Z(u) = e_h(u) - \bar{e}$. Bounded density and (98) imply

$$\|K_h\|^2_{L^2(\mu\otimes\mu)} \le Ch^8 \int_I \int_I (h + |u-v|)^{-2}\, du\, dv \le Ch^7.$$

Averaging operators A_u, A_v in the respective variables are commuting orthogonal projections. Exactly, $\langle Z(u), Z(v)\rangle_H = (I - A_u)(I - A_v)K_h(u,v)$. Projection contraction, the variance identity and the uniform norm bound give, for independent U, V,

$$w_h := \mathbb{E}\langle Z(U), Z(V)\rangle_H^2 \le Ch^7, \quad v_h := \mathbb{E}\|Z(U)\|_H^2 \le C_v h^3, \quad \mathbb{E}\|Z(U)\|_H^4 \le Ch^6. \tag{99}$$

For IID centered Z_j, set $S_n = \sum_{j\le n} Z_j$ and $T_n = \|S_n\|_H^2 - nv_h$. Since the conditional expectation of $2\langle S_{n-1}, Z_n\rangle_H + \|Z_n\|_H^2 - v_h$ is zero, T_n is a real square-integrable martingale. In fact,

$$\mathbb{E}T_n^2 = n\,\mathrm{Var}(\|Z_1\|_H^2) + 2n(n-1)w_h \le C(nh^6 + n^2h^7). \tag{100}$$

To check this identity, expand T_n into centered diagonal terms and twice the sum of $\langle Z_i, Z_j\rangle_H$ for $i < j$. Distinct diagonals are independent. A diagonal and a cross term have zero covariance by conditioning on the cross term's other index. Distinct cross terms also have zero covariance: disjoint pairs are independent, and overlapping pairs vanish after conditioning on their common index. Identical pairs contribute w_h, with coefficient $4\binom{n}{2} = 2n(n-1)$.

Simultaneous control of all larger widths. If $R^2Nh \ge 4C_v$ and $N \le n \le 2N$, the event $\|S_n/n\|_H > Rh^2$ implies $T_n > R^2N^2h^4 - 2NC_vh^3 \ge R^2N^2h^4/2$. Doob's L^2 inequality and (100) yield

$$\Pr\left\{\max_{N\le n\le 2N} \|S_n/n\|_H > Rh^2\right\} \le \frac{C}{R^4}\left(N^{-3}h^{-2} + N^{-2}h^{-1}\right).$$

Apply this to $N_k = 2^kN$. The threshold hypothesis remains valid, and summing the geometric series with ratios $1/8$ and $1/4$ gives

$$\Pr\left\{\sup_{n\ge N} \|S_n/n\|_H > Rh^2\right\} \le \frac{C}{R^4}\left(N^{-3}h^{-2} + N^{-2}h^{-1}\right). \tag{101}$$

For fixed $c_0 > 0$, set $N_b = \lceil c_0/(2h)\rceil$, $h = 2^{-b}$, and fix R with $R^2c_0/2 \ge 4C_v$. The right side is at most $C_{c_0}h$. Unioning over the two groups and two laws gives a summable $C_{c_0}2^{-b}$. Borel–Cantelli therefore controls all four empirical means for every $n \ge N_b$, eventually at every dyadic step. The group weights are $1/2$, so

$$\|B_{\mathsf{R},m,h} - B_{\mathsf{R},\infty,h}\|_H \le Rh^2, \qquad \mathsf{R} = P, Q, \quad m \ge c_0/h \text{ even}. \tag{102}$$

No independence across steps, overlapping widths or the two laws is required.

Mean orders and the probability classes. Functional cancellation (78) gives $\|B_{Q,\infty,h}\|_H \le Ch^2$. For the product law, (86) gives $\Psi_h^P \ge \psi_* > 0$. Since the reference tent is nonnegative, $T_{\Psi_h^P,h} \ge \psi_* \int_I \beta(aZ_1 + uZ_2, hY)\, du$ pointwise. The constant-density instance of (137), together with the remainder bound in (78), therefore gives $\|B_{P,\infty,h}\|_H \ge c_1h^2\sqrt{\log(1/h)} - Ch^2$ with $c_1 > 0$ uniform. This comparison requires no bounded variation of the actual densities or functional moment. The product upper bound needs only bounded density: integrating (98) gives for each group

$$\|\mathbb{E}_\mu e_h(U)\|_H^2 \le \iint |K_h(u,v)|\, d\mu(u) d\mu(v) \le Ch^4[1 + \log(1/h)].$$

Combine groups by the triangle inequality and use (102). The sampled Q squared mean is $O_{c_0}(h^4)$ and the sampled P squared mean is $\Theta_{c_0}(h^4\log(1/h))$ after increasing the onset. Both complete-row prefix laws have variance at most $32400h^2/m^2 = O_{c_0}(h^4)$. Every positive-support Q prefix assignment has loss at most $(\|B_{Q,m,h}\|_H + 180h/m)^2 = O_{c_0}(h^4)$.

The affine-resistant product lower bound also holds for these sampled directions. Output-hidden independence gives $\mathbb{E}_P[\widehat{s}_j\widehat{w}_j] = s_j w_j$. Since the odd part of ReLU is linear, $B_P(X) = B_P(-X)$, so its Gaussian affine projection is its constant mean. Every hidden endpoint segment has first coordinate at least $6/5 - h \geq 11/10$, so the norm Hessian has operator norm at most one. Taylor expansion using $\mathbb{E}_P d_j = 0$ and $\|d_j\|^2 \leq 3h^2$ gives

$$|\mathbb{E}_X B_P| \leq \frac{3s}{2\sqrt{2\pi}} h^2, \qquad \|(I - \Pi_{\mathcal{A}})B_P\|_H^2 = \|B_P\|_H^2 - (\mathbb{E}_X B_P)^2.$$

Every coupling preserving complete P rows has that mean. Orthogonal projection followed by Jensen therefore bounds its expected loss after any realization-specific affine correction below by this logarithmic quantity minus $O(h^4)$. Together with the ordinary P prefix upper bound this proves all three J_P orders. The uniform endpoint gap and nonaffine floor bound every legal assignment, even after affine correction, below by $c_{\mathrm{gap}} h^4$. The Q prefix law belongs to both $\Pi(Q)$ and $\mathcal{S}$, and its support supplies legal assignments with the ordinary quartic upper bound. Minimizing over corrections proves all remaining orders in (96), as well as (95).

Finally intersect the probability-one events over positive rational c_0. A smaller rational coefficient covers every real $c_0 > 0$, giving one event for each density pair and prescribed measurable local laws fixed before sampling. There is no simultaneous assertion over uncountably many density pairs or laws selected after sampling. Widths may depend on the full sampled sequences if they eventually meet the threshold, but each teacher uses the original first $m/2$ entries of each sequence. No subset selection or nested rounding laws are asserted. This proves a sufficient width, not a random-teacher necessity or an arithmetic h^2/m^2 floor. The prefix result supersedes the sufficient $m^2 h^3 \geq c_0$ regime; Corollary 14 remains a valid separate shared-cell guarantee. □

A5 Fixed signed teachers: complete legal-cancellation proof

This section gives an original-width, original-endpoint separation from the same teacher's independent product SR. Every teacher parameter is fixed as the dyadic grid is refined. Every Cartesian hidden endpoint still misses the teacher plane. The construction changes the dependence between two hidden rounding bits within each neuron, while preserving the SR marginal of every scalar. Its signed boundary second moment vanishes, whereas the corresponding product-SR moment remains positive. The two signed groups have distinct directions, and the target has a uniformly nonzero nonaffine component.

Let n be a power of two, $m = 2n$, and put

$$\begin{aligned} s = \frac{6}{5}, \qquad L = \frac{1}{4}, \qquad a = \frac{\sqrt{85}}{5}, \qquad p(u) = \left(\frac{6}{5}, u, \frac{7}{5}\right), \\ u_j^+ = \frac{3}{2} + \frac{j + 1/5}{4n}, \qquad u_j^- = \frac{3}{2} + \frac{j + 7/10}{4n}, \qquad 0 \leq j < n, \\ f_m(X) = \frac{s}{2n} \sum_{j=0}^{n-1} \left[\sigma(p(u_j^+)^\top X) - \sigma(2p(u_j^-)^\top X)\right], \qquad X \sim N(0, I_3). \end{aligned} \tag{103}$$

The unscaled output coefficients are $+s$ and $-s$. All hidden coordinates and unscaled outputs use their original adjacent $h\mathbb{Z}$ endpoints, where $h = 2^{-b} \leq 1/10$. Equivalently, the canonical outputs are $\pm s/\sqrt{m}$ on grids $(h/\sqrt{m})\mathbb{Z}$. The normalization in (103) is $1/m$. Define

$$c_{\mathrm{SR}} = \frac{31 s^2 e^{-3}}{625000(2\pi)^{3/2} aL} > 0, \qquad U_{m,h} = 90000h^4 + \frac{210h^2}{m}. \tag{104}$$

Complete statement of Lemma 2. For every teacher (103) and every stated grid step, there is a law Q on its original Cartesian assignments which preserves each scalar's original SR marginal and satisfies

$$\boxed{\begin{aligned} \mathbb{E}_Q \|\widehat{f} - f_m\|_2^2 &\leq U_{m,h}, \\ \min_{C \text{ legal}} \|\widehat{f}_C - f_m\|_2^2 &\leq U_{m,h}. \end{aligned}} \tag{105}$$

For the same teacher's fully independent product SR,

$$\boxed{\mathbb{E}_{\mathrm{SR}}\|\widehat{f} - f_m\|_2^2 \geq c_{\mathrm{SR}} h^4 \log \frac{1}{256h} \quad \text{if } h \leq 2^{-9} \text{ and } nh \geq 64.} \tag{106}$$

All scalar parameters are off every stated grid, all hidden endpoint choices have a nonzero transverse displacement, and

$$\mathrm{dist}^2(f_m, \mathcal{A}) \geq \frac{289}{10000\pi}. \tag{107}$$

The loss in (105)–(106) is ordinary population MSE. Equation (107) controls target signal; the affine-corrected SR loss bound is proved separately in Theorem 8.

Consequently, along dyadic joint sequences with $h \downarrow 0$ and $mh^2 \geq 1$, the expected loss under Q, and one legal assignment at each step, are $O(h^4)$, while the same teacher's product-SR loss is $\Omega(h^4 \log(1/h))$. Their ratio tends to zero. Each finite-width teacher is fixed independently of h; the asymptotic statement is joint in width and step, since a fixed finite width eventually violates $nh \geq 64$. The constants in (104)–(106) establish this asymptotic separation; finite-step comparisons are reported in Appendix A7.

A5.1 Fixed parameters, distinct directions, and the endpoint gap

The middle coordinates have the rational forms

$$u_j^+ = \frac{30n + 5j + 1}{20n}, \qquad u_j^- = \frac{60n + 10j + 7}{40n}.$$

Because n is a power of two, the factor five cannot cancel from either denominator, including after multiplication by two. The first and third coordinates and both output magnitudes also have uncancelled denominator five. Thus every original scalar is off every dyadic grid. Moreover, $u_j^- = u_j^+ + 1/(8n)$, and no u_j^+ equals any u_k^-. Since the first-to-third coordinate ratio is fixed, two teacher normals are parallel exactly when their unscaled middle coordinates agree. Thus no pair of normals in (103) is parallel. In particular, the signed teacher has no exact cancelling duplicate pair.

Use the fixed orthogonal rotation

$$Z_1 = \frac{6X_1 + 7X_3}{\sqrt{85}}, \qquad Z_2 = X_2, \qquad Y = \frac{-7X_1 + 6X_3}{\sqrt{85}}.$$

The rotated input remains standard Gaussian. To verify the endpoint gap, write $K_1 = q_1/h$, $K_3 = q_3/h$, and $N = 1/h$. For scale one the complete four-corner enumeration is

$N \bmod 5$	$6K_3 - 7K_1$
1	$-1,\ 5,\ -8,\ -2$
2	$-2,\ 4,\ -9,\ -3$
4	$2,\ 8,\ -5,\ 1$
3	$3,\ 9,\ -4,\ 2$

. (108)

Here K_1 is the floor or ceiling of $6N/5$, and K_3 is the floor or ceiling of $7N/5$, while $\tau = h(6K_3 - 7K_1)/\sqrt{85}$. The two groups' teacher normals become $(\lambda a, \lambda u, 0)$, with $\lambda = 1, 2$. For any original adjacent hidden endpoint, write

$$q = (\lambda a + \eta_1, \lambda u + \eta_2, \tau) \quad \text{in rotated coordinates.}$$

The four-corner enumeration (108) applies to $\lambda = 1$. For $\lambda = 2$, it applies with the residue $N \bmod 5$ replaced by $2N \bmod 5$, where $N = 1/h$. Therefore, for every Cartesian endpoint in either group,

$$\frac{h}{\sqrt{85}} \leq |\tau| \leq \frac{9h}{\sqrt{85}} < h, \qquad \|\eta\| \leq \sqrt{3}h < 2h. \tag{109}$$

The construction below retains the full Cartesian support. It does not select a hidden endpoint on the teacher plane.

For (107), let $H = H_2(Z_1) = (Z_1^2 - 1)/\sqrt{2}$. For $u \in [3/2, 7/4]$, set

$$J(u) = \langle \sigma(aZ_1 + uZ_2), H \rangle = \frac{a^2}{2\sqrt{\pi}\sqrt{a^2 + u^2}}.$$

Here $a^2 = 17/5$ and $2 < \sqrt{a^2+u^2} < 3$, so $17/(30\sqrt{\pi}) \le J(u) \le 17/(20\sqrt{\pi})$. Homogeneity gives

$$\langle f_m, H\rangle = \frac{s}{2n}\sum_j [J(u_j^+) - 2J(u_j^-)] \le -\frac{17}{100\sqrt{\pi}}.$$

Since H has unit norm and annihilates affine functions, (107) follows. This bound holds for every width in (103).

A5.2 A LEGAL SCALAR-MARGINAL LAW WITH A DIFFERENT BOUNDARY SECOND MOMENT

For a scale λ, let B_1, B_3 be the upper-endpoint bits for the first and third original hidden coordinates. Write

$$\alpha_\lambda = \{6\lambda N/5\}, \qquad \beta_\lambda = \{7\lambda N/5\}, \qquad t_\lambda = Q(B_1 = B_3 = 1).$$

Use the following exact table:

$N \bmod 5$	t_1	t_2
1	$61/420$	$4/15$
2	$167/420$	$29/70$
4	$229/420$	$1/15$
3	$83/420$	$1/70$

(110)

The other three joint probabilities are $\alpha_\lambda - t_\lambda$, $\beta_\lambda - t_\lambda$, and $1 - \alpha_\lambda - \beta_\lambda + t_\lambda$. All four are strictly positive: their minimum over the table is $1/420$ at scale one and $1/70$ at scale two. The two bits therefore have their prescribed scalar SR marginals. Round the middle hidden coordinate and the signed output independently by their own SR laws, independently of this pair. Finally, make different neurons independent. This defines Q entirely on the original endpoints.

To verify the relevant cancellation, put $c_\lambda = \mathrm{Cov}(B_1, B_3)$. The transverse displacement has mean zero and satisfies

$$V_\lambda := \frac{\mathbb{E}\tau_\lambda^2}{h^2} = \frac{49\alpha_\lambda(1-\alpha_\lambda) + 36\beta_\lambda(1-\beta_\lambda) - 84c_\lambda}{85}.$$

The table gives

$$V_1^Q = \frac{11}{85}, \qquad V_2^Q = \frac{22}{85}, \qquad V_1^Q - \frac{1}{2}V_2^Q = 0,$$

$N \bmod 5$	V_1^{SR}	V_2^{SR}	$V_1^{\mathrm{SR}} - V_2^{\mathrm{SR}}/2$
$1, 4$	$412/2125$	$438/2125$	$193/2125$
$2, 3$	$438/2125$	$412/2125$	$232/2125$

(111)

For example, the covariances at scales one and two are respectively $(137/2100, -4/75)$ in the first row type and $(163/2100, -23/350)$ in the second. They lie strictly inside their Fréchet intervals. Product SR instead has covariance zero. Thus the signed cancellation under Q does not cancel the corresponding SR moment.

A5.3 REMOVING THE LOGARITHM IN THE TRUE MEAN ERROR

First remove only the tangential hidden perturbation η from the mean quantized function. This defines an auxiliary function for the proof, not an additional candidate or rounding rule. The actual assignment continues to use every original endpoint. Let $A_\lambda = \tau_\lambda/h$ under Q, so $|A_\lambda| \le 1$ and $\mathbb{E}A_\lambda = 0$. For $k = hY$, define the continuous-direction residual profile

$$\delta(t,k) = \frac{s}{2}\left[\mathbb{E}\sigma(t + A_1 k) - 2\mathbb{E}\sigma(t + A_2 k/2) + \sigma(t)\right]. \tag{112}$$

Homogeneity accounts for the factor two and the node $A_2/2$. Its zeroth and first coefficient moments vanish. Equation (111) makes its second coefficient moment vanish as well. Consequently

$$\begin{gathered}\operatorname{supp}\delta(\cdot,k) \subset [-|k|,|k|], \qquad \int_{\mathbb{R}} \delta(t,k)\,dt = 0,\\ |\delta(t,k)| \le s|k|, \qquad \mathrm{TV}(\delta(\cdot,k)) \le s|k|.\end{gathered} \tag{113}$$

For the integral identity, integrate the ridges over $[-R, R]$ with $R > |k|$ and use $\int_{-R}^{R} \sigma(t+qk)\,dt = (R+qk)^2/2$ for $|qk| < R$. For the variation bound, an unbiased scalar-shift Jensen residual $b(t) = \mathbb{E}\sigma(t+T) - \sigma(t)$ increases up to zero and decreases thereafter, and $\mathrm{TV}(b) = 2b(0) = \mathbb{E}|T|$.

Let $D(t,k) = \int_{-\infty}^{t} \delta(q,k)\,dq$. It is supported on $[-|k|, |k|]$ and bounded by $2sk^2$. Set $x = aZ_1 + (3/2)Z_2$, $b = LZ_2$, and

$$A(x,b,k) = \int_0^1 \delta(x+br,k)\,dr.$$

For $b \neq 0$, this is $[D(x+b,k) - D(x,k)]/b$. Its support in x has total length at most $4|k|$, and $|A| \le \min\{s|k|, 4sk^2/|b|\}$. Therefore

$$\int_{\mathbb{R}^2} |A(x,b,k)|^2\,dx\,db \le 64s^2k^4, \qquad \|A(x,b,hY)\|_2^2 \le \frac{96s^2}{\pi aL}h^4 < 100h^4. \tag{114}$$

The Gaussian density after this linear change of variables is at most $1/(2\pi aL)$, and $\mathbb{E}Y^4 = 3$.

The two finite direction grids in (103) each select one point in every subinterval of length L/n. The bounded-variation quadrature bound applies to either shift of the grid. Applied separately to the two unbiased-shift residuals in (112), it gives

$$|B_{\mathrm{aux}}(Z,Y) - A(x,b,hY)| \le \frac{sh|Y|}{n}, \qquad \|B_{\mathrm{aux}}\|_2 \le 10h^2 + \frac{1.2h}{n}, \tag{115}$$

where B_{aux} is the finite auxiliary mean error.

We next bound the effect of restoring every tangential endpoint. Write $R = \|Z\|$, $t = \lambda(aZ_1 + uZ_2)$, and $d = \eta^\top Z$. Scalar unbiasedness implies $\mathbb{E}d = 0$, even though d and τ can be dependent. The elementary identity obtained by integrating the ReLU derivative gives

$$|\sigma(t+v+d) - \sigma(t+v) - d\mathbf{1}_{\{t>0\}}| \le |d|\mathbf{1}_{\{|t|\le|v|+|d|\}}.$$

With $v = \tau Y$, the expected absolute difference between the actual and auxiliary ridges is bounded by $2hR\mathbf{1}_{\{|aZ_1+uZ_2|\le W\}}$, where $W = h|Y| + 2hR$. If $W < R/4$ and this strip is nonempty, then $|Z_2| \ge R/2$, because $a = \sqrt{85}/5$ and $u \in [3/2, 7/4]$. The number of points of either direction grid in the strip is at most $4nW/(LR) + 1$. If $W \ge R/4$, use the bound by all n points instead. After including both signed groups, these two cases give

$$\begin{aligned} |B_Q - B_{\mathrm{aux}}| &\le 80h^2(R + |Y|) + \frac{3hR}{n}, \\ \|B_Q - B_{\mathrm{aux}}\|_2 &\le 200h^2 + \frac{5h}{n}, \qquad B_Q = \mathbb{E}_Q\widehat{f} - f_m. \\ \|B_Q\|_2 &\le 210h^2 + \frac{7h}{n}, \end{aligned} \tag{116}$$

Here $\|R\|_2 = \sqrt{2}$ and $\|Y\|_2 = 1$. Output-hidden independence lets each rounded output be replaced by its signed mean $\pm s$ in this mean-error calculation.

For either Q or product SR, consider an unscaled local error $e = S\sigma(q^\top X) - s_j\sigma(p_j^\top X)$, where $s_j \in \{s, -s\}$. The output is independent of the hidden rounding and has mean s_j. Scalar unbiasedness gives $\mathbb{E}\|q - p_j\|^2 \le 3h^2/4$, irrespective of hidden-bit correlations. Also $\|p_j\|^2 \le 517/20 < 26$. ReLU Lipschitz continuity and the vanishing output cross term yield

$$\begin{aligned} \mathbb{E}\|e\|_2^2 &\le \frac{3s^2h^2}{4} + \frac{h^2}{8}\left(26 + \frac{3h^2}{4}\right) < 5h^2, \\ \mathbb{E}_Q\|\widehat{f} - f_m\|_2^2 &\le \|B_Q\|_2^2 + \frac{5h^2}{m} \le 88200h^4 + \frac{98h^2}{n^2} + \frac{5h^2}{m} \le U_{m,h}. \end{aligned} \tag{117}$$

This proves the first part of (105).

A5.4 A LOWER BOUND FOR THE SAME TEACHER'S ACTUAL PRODUCT SR

For the lower bound it is useful to integrate the actual product-SR mean continuously over u, including the middle coordinate's rounding. This avoids a small-step threshold obtained by subtracting global norm bounds. Work on the input region

$$Z_2 = r > 0, \qquad Z_1 = -rv/a, \qquad \frac{25}{16} \le v \le \frac{27}{16}, \qquad 1 \le Y \le 2, \qquad 64h \le r \le \frac{1}{4}. \tag{118}$$

The interval for v is the middle half of $[3/2, 7/4]$. In the original coordinates, $X_1^2 + X_3^2 = Z_1^2 + Y^2$, and on this region $|X_1| + |X_3| < 3$.

At scale λ, first leave the middle coordinate unrounded and let $\xi_\lambda = \delta q_1 X_1 + \delta q_3 X_3$ be the first/third-coordinate perturbation. It is mean zero, and $|\xi_\lambda| < 3h$. The roots of all these perturbed ridges remain inside $[3/2, 7/4]$: their displacement from v is at most $3h/(\lambda r) \le 3/64 < 1/16$. Integrating the exact quadratic antiderivative of the hinge gives the continuous mean error $\mathbb{E}\xi_\lambda^2/(2\lambda L r)$ for this scale.

Under product SR, the diagonal entries of $h^{-2}\operatorname{Var}(\delta q_{1,1}, \delta q_{1,3}) - \frac{1}{2}h^{-2}\operatorname{Var}(\delta q_{2,1}, \delta q_{2,3})$ are either $(1/25, 4/25)$ or $(4/25, 1/25)$, and both off-diagonal entries are zero. Here the first subscript denotes the scale. Consequently the signed continuous mean with these two hidden coordinates rounded is at least

$$\frac{s}{4Lr}\left(\mathbb{E}\xi_1^2 - \frac{1}{2}\mathbb{E}\xi_2^2\right) \ge \frac{sh^2(X_1^2 + X_3^2)}{100Lr} \ge \frac{sh^2Y^2}{100Lr}. \tag{119}$$

Restoring middle-coordinate product SR adds a nonnegative convex interpolation error at each scale. For a fixed first/third endpoint, as a function of the middle coordinate the ridge is a hinge of slope r. The integral of its linear-interpolation error over one full grid cell is at most $rh^2/8$; only the cell containing its kink can contribute. The integration endpoints $\lambda(3/2)$ and $\lambda(7/4)$ lie on every grid under consideration here. Changing variables from the middle coordinate to u therefore bounds the normalized integrated excess by $rh^2/(8\lambda L)$. The positive group can only increase the signed mean, and the negative scale-two group can decrease it by at most $srh^2/(32L)$. Thus the actual continuous-direction product-SR mean error obeys

$$B_{\mathrm{SR},\infty}(X) \ge \frac{sh^2}{Lr}\left(\frac{Y^2}{100} - \frac{r^2}{32}\right) \ge \frac{sh^2Y^2}{125Lr} \quad \text{on (118).} \tag{120}$$

The last inequality uses $r \le 1/4$, $Y \ge 1$, and $1/100 - 1/512 \ge 1/125$.

There is also a direct variation bound for replacing this continuous mean by the two finite teacher grids. Let $e_\lambda(u; X)$ be the actual product-SR mean ridge minus its teacher ridge, before its signed output coefficient is applied. For fixed X in (118), the mean rounded ridge is a convex piecewise linear function of u with slopes in $[0, \lambda r]$: middle-coordinate SR is precisely linear interpolation of a convex ridge function, and averaging the other endpoints preserves these slope bounds. The teacher ridge has slope zero before $u = v$ and slope λr after it. Hence e_λ increases up to v and decreases after v. Jensen's inequality makes it nonnegative. At the peak the teacher preactivation is zero, and unbiasedness gives

$$\mathrm{TV}_{[3/2,7/4]}(e_\lambda) \le 2e_\lambda(v; X) = \mathbb{E}|\delta q_1 X_1 + \delta q_2 r + \delta q_3 X_3| \le h(|X_1| + |X_3| + r) < 4h.$$

Both shifted n-point rules therefore satisfy

$$|B_{\mathrm{SR},m} - B_{\mathrm{SR},\infty}| \le \frac{4sh}{n}, \qquad nh \ge 64 \implies B_{\mathrm{SR},m}(X) \ge \frac{sh^2Y^2}{250Lr} \quad \text{on (118).} \tag{121}$$

Indeed the last proposed lower bound is at least $8sh^2/125$ on this region, whereas $4sh/n \le sh^2/16$.

The Jacobian of $(v, r, Y) \mapsto (Z_1, Z_2, Y)$ is r/a. On (118), the Gaussian density is at least $e^{-3}/(2\pi)^{3/2}$. Squaring (121), integrating over v (length $L/2$), using $\int_1^2 Y^4\,dY = 31/5$, and integrating dr/r gives

$$\|B_{\mathrm{SR},m}\|_2^2 \ge \frac{31s^2e^{-3}}{625000(2\pi)^{3/2}aL}h^4 \log\frac{1}{256h}.$$

The independent-SR expected loss is at least its squared mean error, which proves (106). The region is nonempty for the stated $h \le 2^{-9}$. In the joint regime $mh^2 \ge 1$, this also implies $nh \ge 1/(2h) \ge 256$, so the width condition of (106) is automatic. This completes the theorem's loss and signal bounds.

A5.5 One legal vector and the finite-bit interpretation

Law Q is a product across neurons, with sixteen local Cartesian states per neuron. Sequential conditional expectation for the true Gaussian population loss chooses one local state at each step without increasing the remaining conditional average. It produces a single original-width legal assignment satisfying the second part of (105). With finite-bit score intervals, for every integer $p \ge 0$ it instead gives an explicitly computable legal C_p with

$$\|\widehat{f}_{C_p} - f_m\|_2^2 \le U_{m,h} + 2^{-p} \tag{122}$$

in $O(m^2 \operatorname{poly}(B + p + \log m))$ bit operations. Here B includes the bit length of the rational teacher and grid parameters. All unscaled vectors and local probabilities in this section are rational; hidden coordinates and their dot products remain positive and bounded away from zero. Signed output coefficients only multiply the same Gaussian-ReLU kernel terms. Thus the rational-input conditional-expectation argument applies unchanged to their finite linear combinations. Taking $2^{-p} \le h^4$ preserves the stated order bound. This is an existence and finite-bit construction, not a claim of exact global legal optimization.

For completeness, write e_{jc} for the sixteen local residual functions, $b_j = \sum_c Q_j(c) e_{jc}$ and $G_{jc,kd} = \langle e_{jc}, e_{kd} \rangle$. At step j only the score

$$H_j(c) = m^{-2} \left[G_{jc,jc} + 2 \sum_{i<j} G_{jc,ic_i} + 2 \sum_{k>j,d} Q_k(d) G_{jc,kd} \right]$$

depends on the candidate. Enclose every score to width $2^{-p}/m$ and select the smallest upper endpoint. Its exact score exceeds the minimum by at most $2^{-p}/m$; summing the m conditional-expectation increments proves (122) without testing transcendental equality. There are $O(m^2)$ Gram entries. For the positive hidden coordinates here, $d = u^\top v$ is bounded below. Compute $D = \|u\|^2\|v\|^2 - d^2$ exactly and use

$$\mathcal{K}(u, v) = \frac{\sqrt{D} + d[\pi - \arctan(\sqrt{D}/d)]}{2\pi},$$

with the exact value $d/2$ when $D = 0$. Nonzero rational D is at least $2^{-O(B)}$, and all norms are bounded, so $O(B + p + \log m)$ precision suffices. Rational square-root bounds, arctangent half-angle reduction and its alternating series, together with Machin's identity for π, give polynomial cost at that precision. This establishes the stated bit bound within this teacher class without a supplementary proof dependency.

If an implementation requires the canonical outputs $\pm s/\sqrt{m}$ themselves to be rational, one may restrict to $n = 2 \cdot 4^k$, so $m = 4^{k+1}$ is a perfect square. This is a restriction of that representation or API. The unscaled rational form (103), its proof, and (122) hold for every power-of-two n.

A5.6 Why preserving the complete local product law cannot give this improvement

Let P_j be the original complete product-SR law of neuron j, let F_j be its unscaled rounded function, and write $\mu = m^{-1} \sum_j \mathbb{E}_{P_j} F_j$ and $b_{\mathrm{SR}} = \mu - f_m$. Under any joint law $\widetilde{Q}$ which retains every complete local marginal P_j, the mean network is still μ. The exact Hilbert-space identity is

$$\begin{aligned} \mathbb{E}_{\widetilde{Q}} \|\widehat{f} - f_m\|_2^2 &= \|b_{\mathrm{SR}}\|_2^2 + \mathbb{E}_{\widetilde{Q}} \|\widehat{f} - \mu\|_2^2 \ge \|b_{\mathrm{SR}}\|_2^2, \\ \mathbb{E}_{\mathrm{SR}} \|\widehat{f} - f_m\|_2^2 &= \|b_{\mathrm{SR}}\|_2^2 + \frac{1}{m^2} \sum_j \operatorname{Var}_{P_j}(F_j). \end{aligned} \tag{123}$$

For (103), the final variance term is at most $5h^2/m$ by (117). Consequently the greatest possible expected improvement over product SR under complete-local-law preservation is at most $5h^2/m$. If

$mh^2 \log(1/h) \to \infty$ within the stated range, the best expected loss in this restricted class divided by the product-SR expected loss tends to one. In particular, its expected error cannot remove the logarithmic mean bias. This restriction on expected losses does not rule out a good deterministic assignment in the support of such a law.

Changing only output-hidden dependence is also insufficient at the leading order here. More generally, retain a neuron's whole product hidden marginal and its scalar output SR marginal, but allow arbitrary dependence between them. Relative to product SR, its change of mean function is

$$\Delta\mu_j = \mathbb{E}[(S - s_j)(\sigma(q^\top X) - \sigma(p_j^\top X))], \qquad \|\Delta\mu_j\|_2 \le \sqrt{\mathbb{E}(S - s_j)^2\, \mathbb{E}\|q - p_j\|^2} \le \frac{\sqrt{3}}{4} h^2. \tag{124}$$

The subtraction is valid because the output marginal has mean s_j; ReLU Lipschitz continuity and Cauchy–Schwarz prove the bound without output-hidden independence. Averaging neurons leaves the same $\sqrt{3}h^2/4$ bound on the network mean change. Thus a nonzero $\Omega(h^2\sqrt{\log(1/h)})$ SR mean cannot be removed by this dependence alone. Law Q instead changes hidden-hidden covariance while preserving all scalar marginals, which changes the boundary second moment by order h^2.

For the auxiliary continuous-direction profile, vanishing integral is exactly the distinction between the two orders. In a signed collection of scales λ, the integral is proportional to the signed sum of output-weighted transverse variances divided by λ. If it vanishes, its compactly supported primitive gives the quartic bound (114). If it is nonzero, every interval crossing the full support has average equal to that integral divided by its length, producing the logarithmic radial integral. The table (111) makes these alternatives different for the legal law Q and for product SR of the same fixed teacher. Tangential rounding and the finite direction grids were retained in the true-loss bounds above. This is therefore a legal signed covariance mechanism with a transverse gap, rather than a signed refit in an enlarged dictionary.

A5.7 The separation is invisible to the expected bilinear proxy

There is an exact comparison which distinguishes this mechanism from ordinary linear variance reduction. For either law, let

$$P_j(X) = \mathbf{1}_{\{p_j^\top X > 0\}} (q_{s,j} q_j - s_j p_j)^\top X, \qquad P = m^{-1} \sum_j P_j.$$

This is the raw fixed-gate proxy including the full output-hidden product. Output-hidden independence and scalar unbiasedness imply $\mathbb{E}P_j(X) = 0$ pointwise in X. Independence across neurons removes all cross terms in its expected squared loss. For a fixed rounding realization, Gaussian sign symmetry gives $\|P_j\|_2^2 = \frac{1}{2}\|q_{s,j} q_j - s_j p_j\|^2$. Writing $v_{jk} = \mathrm{Var}(q_{jk})$ and $v_{s,j} = \mathrm{Var}(q_{s,j})$, it follows that

$$\boxed{\mathbb{E}_Q \|P\|_2^2 = \mathbb{E}_{\mathrm{SR}} \|P\|_2^2 = \frac{1}{2m^2} \sum_j \left[s_j^2 \sum_k v_{jk} + v_{s,j} \left(\|p_j\|^2 + \sum_k v_{jk} \right) \right].} \tag{125}$$

Hidden-hidden covariances do not enter this expression. The equality holds for any choice of hidden-bit coupling retaining the scalar SR marginals, provided outputs remain independent of hidden rounding and different neurons remain independent. If ReLU is replaced by the identity activation, the two laws' true expected squared errors are also exactly equal, each being twice the value in (125).

Thus the expected raw bilinear proxy cannot distinguish the two laws which have the nonlinear separation (105)–(106). This is a statement about the proxy objective averaged under these laws. It does not say that every deterministic proxy minimizer selects a bad legal vector, and it does not turn underestimation of a proxy value into a proved failure of its selected solution.

A6 Common-density boundary-moment criterion: complete proof

A6.1 Statement and two-bit feasibility

The signed moment extends beyond the explicit rational example. Fix G groups with proportions $\alpha_r > 0$ summing to one, signed outputs $s_r \neq 0$, and positive bounded scales λ_r. Let $p(u) = (A, u, C)$

with $A, C > 0$. All groups sample the same probability density ρ on a compact positive interval I, using one fixed point in each equal-mass stratum; group r contains an integer $n_r = \alpha_r m$ neurons. Assume the zero extension of ρ has bounded variation and $\rho \geq \rho_* > 0$ on a nondegenerate interior interval. The teacher is $m^{-1} \sum_{r,j} s_r \sigma(\lambda_r p(u_{rj})^\top X)$. It is fixed at each width; all scalars use original adjacent endpoints.

Admissible laws preserve hidden scalar SR marginals, keep outputs independently SR, and make neurons independent. The first/third-bit law is constant across directions within each group. With τ_r the displacement normal to the teacher plane, define

$$V_r(h) = \mathbb{E}(\tau_r/h)^2, \qquad M_h = \sum_r \frac{\alpha_r s_r}{\lambda_r} V_r(h). \tag{126}$$

Theorem 16 (Common-density boundary-moment criterion)**.** *Under these fixed hypotheses, for all sufficiently small h the actual mean error satisfies*

$$B_h = M_h T_h + R_h, \qquad \|T_h\|_2^2 = \Theta(h^4 \log(1/h)), \qquad \|R_h\|_2 \leq C(h^2 + h/m), \tag{127}$$

uniformly over the admissible widths and local laws. Hence

$$\begin{aligned} cM_h^2 h^4 \log(1/h) - C(h^4 + h^2/m^2) \leq \|B_h\|_2^2 \\ \leq C[h^4\{1 + M_h^2 \log(1/h)\} + h^2/m^2], \\ \mathbb{E}L \leq C[h^4\{1 + M_h^2 \log(1/h)\} + h^2/m]. \end{aligned} \tag{128}$$

If mh is bounded below by a positive constant, quartic squared mean error is equivalent to $|M_h| = O(1/\sqrt{\log(1/h)})$. If mh^2 is bounded below by a positive constant, the same equivalence holds for full expected error. When $\sum_r \alpha_r s_r \lambda_r \neq 0$, the teacher's nonaffine energy stays bounded below at sufficiently large widths.

Appendix A6 proves the decomposition using a fixed reference tent. Subtracting M_h times that tent leaves a profile with zero integral; integration by parts against the bounded-variation density controls its primitive. The same strip counts and equal-mass quadrature then recover the actual finite teacher. This also quantifies imperfect cancellation: exact $M_h = 0$ is sufficient, while the stated inverse-square-root logarithmic tolerance is the precise order condition.

The feasible moment is computable for two bits. Let $p_{r1} = \{\lambda_r A/h\}$, $p_{r3} = \{\lambda_r C/h\}$ and $(\nu_1, \nu_3) = (-C, A)/\sqrt{A^2 + C^2}$. The classical Bernoulli Fréchet bounds give

$$c_r \in [\max(0, p_{r1} + p_{r3} - 1) - p_{r1}p_{r3},\ \min(p_{r1}, p_{r3}) - p_{r1}p_{r3}], \quad V_r = d_r + 2\nu_1\nu_3 c_r,$$

where $d_r = \nu_1^2 p_{r1}(1 - p_{r1}) + \nu_3^2 p_{r3}(1 - p_{r3})$. Sort the variance interval as $[v_r^-, v_r^+]$, and put $w_r = \alpha_r s_r/\lambda_r$. Then

$$M_h = 0 \text{ is feasible} \quad \Longleftrightarrow \quad \sum_r \min(w_r v_r^-, w_r v_r^+) \leq 0 \leq \sum_r \max(w_r v_r^-, w_r v_r^+). \tag{129}$$

The interval test and a realizing mixture take $O(G)$ real arithmetic operations given the marginals. Strict feasibility and off-grid bits allow positive mass on all four outcomes. This is a two-bit feasibility result, not an optimization algorithm for arbitrary covariance matrices.

A transverse endpoint gap forces every V_r to be positive. Same-sign outputs therefore cannot cancel M_h; signed groups can. A separation from SR further requires a nonvanishing SR moment and target signal, both verified in Lemma 2. The common density is substantive: different group densities generally require cancellation of the function $\sum_r w_r V_r \rho_r(u)$, not just one scalar. Higher-dimensional hidden-bit coupling also leaves the interval problem treated here.

The cancellation in Appendix A5 extends to finitely many fixed signed scale groups. When their continuum direction density is the same, a single scalar moment controls the logarithmic part of the mean error. For two coupled hidden bits, the feasible values of this moment form an explicit interval. This gives a criterion for constructing a legal local law; it is not an optimization of the full deterministic assignment problem.

Fix a number of groups G, positive proportions α_r summing to one, nonzero signed outputs s_r, and scales $0 < \lambda_{\min} \leq \lambda_r \leq \lambda_{\max} < \infty$. All these quantities are independent of m and h. Consider

widths for which $n_r = \alpha_r m$ are integers. Fix $A, C > 0$, $a = \sqrt{A^2 + C^2}$, a compact interval $I = [u_-, u_+] \subset (0, \infty)$, and a probability density ρ on I. Assume its extension by zero has bounded variation and that $\rho \geq \rho_* > 0$ almost everywhere on some nondegenerate interval contained in the interior of I. In particular, ρ is bounded. For each group, divide I into n_r consecutive strata of ρ-mass $1/n_r$, and choose one point u_{rj} in each stratum. Choose these points once at each width, independently of h. Define

$$p(u) = (A, u, C), \qquad X \sim N(0, I_3),$$
$$f_m(X) = \frac{1}{m} \sum_{r=1}^{G} \sum_{j=1}^{n_r} s_r \sigma(\lambda_r p(u_{rj})^\top X), \qquad K = \sum_{r=1}^{G} \alpha_r s_r \lambda_r. \tag{130}$$

Every original scalar uses its adjacent $h\mathbb{Z}$ endpoints in the unscaled representation. Canonical outputs again use step $h/\sqrt{m}$. The theorem below allows grid-aligned scalars; if an everywhere-off-grid statement or a transverse endpoint gap is desired, these are additional conditions on the fixed parameters and chosen nodes, as verified explicitly in Appendix A5.

Use the rotation with $Z_1 = (AX_1 + CX_3)/a$, $Z_2 = X_2$, and $Y = (-CX_1 + AX_3)/a$. At group r, an endpoint has rotated form $(\lambda_r a + \eta_1, \lambda_r u + \eta_2, \tau_r)$. The admissible local laws for this section preserve every hidden scalar's SR marginal, keep the output independently SR, and make different neurons independent. Require the joint law of the first and third hidden bits to be the same at every u in a given group. Thus the law of $A_r = \tau_r/h$ is independent of u, though it can depend on h. The middle bit may be independently SR; allowing further correlations is harmless for the estimates if the transverse law stays the same and all scalar marginals remain unchanged. Put

$$V_r(h) = \mathbb{E}A_r^2, \qquad M_h = \sum_{r=1}^{G} \frac{\alpha_r s_r}{\lambda_r} V_r(h), \qquad B_h = \mathbb{E}\widehat{f} - f_m. \tag{131}$$

Scalar unbiasedness gives $\mathbb{E}A_r = 0$, $\mathbb{E}\eta = 0$, and the endpoint constraints uniformly bound $|A_r|$ and $\|\eta\|/h$. The signed moment M_h is therefore uniformly bounded as well.

Complete statement of Theorem 16. Under these fixed hypotheses there are constants $c, C, h_0 > 0$, depending on the fixed family and density but not on h, the admissible width, or the chosen local laws, such that for $0 < h \leq h_0$,

$$\begin{aligned} cM_h^2 h^4 \log(1/h) - C\left(h^4 + \frac{h^2}{m^2}\right) &\leq \|B_h\|_2^2 \\ &\leq C\left[h^4\{1 + M_h^2 \log(1/h)\} + \frac{h^2}{m^2}\right], \\ \mathbb{E}\|\widehat{f} - f_m\|_2^2 &\leq C\left[h^4\{1 + M_h^2 \log(1/h)\} + \frac{h^2}{m}\right]. \end{aligned} \tag{132}$$

The expected loss is also at least the squared mean in the first line. Along sequences with mh bounded below by a positive constant, the mean squared error is $O(h^4)$ if and only if $|M_h| = O(1/\sqrt{\log(1/h)})$. If instead mh^2 is bounded below by a positive constant, the same equivalence holds for the full expected loss. In particular, exact cancellation $M_h = 0$ is sufficient; it is not necessary for quartic order.

If $K \neq 0$, the teachers have a uniformly nonzero nonaffine component at all sufficiently large admissible widths. More explicitly, with $\eta_\rho = a^2/(2\sqrt{\pi}\max_{u \in I}\sqrt{a^2 + u^2}) > 0$,

$$\operatorname{dist}(f_m, \mathcal{A}) \geq \left(|K|\eta_\rho - \frac{C}{m}\right)_+. \tag{133}$$

Distinct choices of nodes across groups can exclude all exactly parallel teacher pairs. Neither the signal conclusion nor the loss bounds require adding or refitting any neuron after the teacher has been fixed.

A6.2 A SCALAR PROFILE AND ITS ZERO-INTEGRAL COMPONENT

Remove tangential endpoint perturbations only as an auxiliary step in the proof. For $k = hY$, the common-density residual profile is

$$\delta_h(t,k) = \sum_r \alpha_r s_r \lambda_r \left[\mathbb{E}\sigma(t + A_r k/\lambda_r) - \sigma(t)\right], \qquad \int_{\mathbb{R}} \delta_h(t,k)\, dt = \frac{M_h k^2}{2}. \tag{134}$$

Its coefficient moments of degrees zero and one vanish. The profile is supported on $[-C|k|, C|k|]$, has amplitude at most $C|k|$, and has total variation at most $C|k|$, uniformly over the local laws. These facts follow from bounded scales, output mass, and transverse displacements, using the unbiased-shift calculation from (113). Define a fixed reference tent

$$\beta(t,k) = \frac{\sigma(t+k) + \sigma(t-k)}{2} - \sigma(t), \qquad \delta_{0,h} = \delta_h - M_h \beta. \tag{135}$$

The tent is nonnegative and has integral $k^2/2$. Thus $\delta_{0,h}$ has integral zero and the same uniform support, amplitude, and variation bounds. Its primitive $D_h(t,k)$ is compactly supported and satisfies $\int |D_h(t,k)|^2 dt \le C|k|^5$.

Here bounded variation of the common density is enough; differentiable density is not required. Write $x = aZ_1$, $b = Z_2$, and

$$A_{0,h}(x,b,k) = \int_I \rho(u)\delta_{0,h}(x + ub, k)\, du.$$

For $|b| \le |k|$, its amplitude is $O(|k|)$, and its support in x has length $O(|k|)$; integrating also over this range of b gives $O(k^4)$. For $|b| > |k|$, integration by parts against the bounded-variation measure of the zero-extended density gives

$$A_{0,h}(x,b,k) = -\frac{1}{b}\int_{\mathbb{R}} D_h(x + ub, k)\, d\rho(u).$$

If $T = \mathrm{TV}(\rho)$, Jensen's inequality for the total variation measure yields

$$\int_{\mathbb{R}} |A_{0,h}(x,b,k)|^2 dx \le \frac{T^2}{b^2}\int_{\mathbb{R}} |D_h(t,k)|^2 dt.$$

Integrating $|b| > |k|$, bounding the transformed Gaussian density, and using $\mathbb{E}Y^4 = 3$, proves

$$\left\| \int_I \rho(u)\delta_{0,h}(aZ_1 + uZ_2, hY)\, du \right\|_2 \le Ch^2. \tag{136}$$

The zero case $k = 0$ is immediate. Extension of ρ by zero includes its possible endpoint jumps in the integration by parts.

For the reference tent, set $T_h(Z,Y) = \int_I \rho(u)\beta(aZ_1 + uZ_2, hY)\, du$. There are positive constants such that

$$ch^4 \log(1/h) \le \|T_h\|_2^2 \le Ch^4 \log(1/h), \qquad 0 < h \le h_0. \tag{137}$$

For the upper bound, when $|b| > |k|$, boundedness of ρ gives amplitude at most $Ck^2/|b|$, with support in x of length $C(|b| + |k|)$. The Gaussian factor $e^{-b^2/2}$ makes the large-$|b|$ tail integrable. The region $|b| \le |k|$ contributes $O(k^4)$ as before. Integration in Y uses finiteness of $\mathbb{E}[Y^4 \log_+(1/|Y|)]$. For the lower bound, restrict Y to $[1,2]$, choose the crossing location inside an interval on which $\rho \ge \rho_*$, and take $C_1 h \le b \le c_1$. Whenever the full tent support lies in that interval, its average is at least $\rho_* k^2/(2b)$. The interval of admissible crossing locations has fixed positive length. Its Gaussian integral therefore contains a positive multiple of $h^4 \int_{C_1 h}^{c_1} db/b$, proving the lower bound after decreasing h_0 if necessary.

A6.3 Returning to all original endpoints and the finite teacher grids

For a function of bounded variation on I, any one-point rule on equal-ρ-mass strata differs from its ρ-integral by at most its total variation divided by the number of strata. Applying this separately to each group's unbiased-shift residual gives an auxiliary finite-grid error at most

$$Ch|Y|\sum_r \frac{\alpha_r}{n_r} = \frac{CGh|Y|}{m}.$$

The fixed number of groups is absorbed into the constants in (132).

The tangential perturbation estimate used in (116) does not require independence between tangential and transverse perturbations. With $R = \|Z\|$, $t = \lambda_r(aZ_1 + uZ_2)$, and $d = \eta^\top Z$, subtract $d\mathbf{1}_{\{t>0\}}$ before taking expectations. Its expectation is zero, and the remaining absolute difference is bounded by

$$ChR\,\mathbf{1}_{\{|aZ_1+uZ_2|\le Ch(|Y|+R)\}}.$$

If the strip width is at least a fixed multiple of R, the amplitude itself is $O(h^2(|Y| + R))$. Otherwise, a nonempty strip forces $|Z_2| \ge cR$, since $a > 0$ and I is bounded. Its length in u is then at most $Ch(|Y| + R)/R$. A direction grid with n_r equal-mass strata has at most $n_r\|\rho\|_\infty|J| + 2$ points in any interval J. Averaging groups therefore bounds the tangential correction pointwise by $C[h^2(|Y| + R) + hR/m]$. Combining this bound, (136), and the finite-grid estimate gives the useful norm identity with a controlled remainder:

$$B_h = M_hT_h + R_h, \qquad \|R_h\|_2 \le C\left(h^2 + \frac{h}{m}\right). \tag{138}$$

This is an identity between actual mean errors and an auxiliary reference function. It does not alter the legal candidate set.

The two squared-mean bounds in (132) now follow from (137)–(138), the triangle inequality, and $(x - y)_+^2 \ge x^2/2 - y^2$. Bounded teacher parameters and scalar SR variances give a local true-error second moment at most Ch^2, just as in (117). Independence across neurons adds at most Ch^2/m to the squared mean, proving the expected-loss bound. These uniform inequalities also prove both stated $O(h^4)$ equivalences.

For the signal statement, put $F_\rho = \int_I \rho(u)\sigma(p(u)^\top X)du$. Equal-mass quadrature and $\mathrm{TV}_u\,\sigma(p(u)^\top X) \le |I||X_2|$ give $\|f_m - KF_\rho\|_2 \le C/m$. The normalized $H_2(Z_1)$ pairing of F_ρ is at least η_ρ. Projection onto the orthogonal complement of affine functions proves (133). This completes Theorem 16.

A6.4 Exact two-bit feasibility and a tolerance for imperfect cancellation

For the original first and third coordinates, write

$$p_{r1} = \{\lambda_r A/h\}, \qquad p_{r3} = \{\lambda_r C/h\}, \qquad \nu_1 = -C/a, \quad \nu_3 = A/a,$$
$$c_r^- = \max(0, p_{r1} + p_{r3} - 1) - p_{r1}p_{r3}, \qquad c_r^+ = \min(p_{r1}, p_{r3}) - p_{r1}p_{r3}.$$

Every coupling of the two Bernoulli bits has covariance in $[c_r^-, c_r^+]$, and every value in that interval is attainable. Its normalized transverse variance is

$$V_r = d_r + 2\nu_1\nu_3c_r, \qquad d_r = \nu_1^2p_{r1}(1 - p_{r1}) + \nu_3^2p_{r3}(1 - p_{r3}). \tag{139}$$

Let $[v_r^-, v_r^+]$ be the interval obtained by evaluating this affine expression at the two covariance endpoints and sorting them. Set $w_r = \alpha_r s_r/\lambda_r$. Then the feasible moments are exactly

$$\boxed{\begin{gathered}\{M_h\} = [\ell_h, U_h],\\ \ell_h = \sum_r \min(w_rv_r^-, w_rv_r^+), \qquad U_h = \sum_r \max(w_rv_r^-, w_rv_r^+),\\ M_h = 0 \text{ is feasible} \quad\Longleftrightarrow\quad \ell_h \le 0 \le U_h.\end{gathered}} \tag{140}$$

The choices are independent between groups, so this is the ordinary sum of scalar intervals. Feasibility and a realizing law require $O(G)$ real arithmetic operations once the scalar marginal probabilities

are given. For example, choose the joint endpoint law in each group attaining its contribution to ℓ_h, and another attaining its contribution to U_h. If $\ell_h < 0 < U_h$, mix those two local laws in every group with the same weight $\theta = -\ell_h/(U_h - \ell_h)$. The resulting moment is zero. If both scalar bits are off-grid, this strict mixture retains all four joint outcomes with positive probability. Cases in which zero is an endpoint of the feasible interval remain legal but can require zero local probabilities. Degenerate intervals are handled by their fixed values. No bit-complexity bound is claimed by this real-arithmetic statement.

More generally, let $d_h = \operatorname{dist}(0, [\ell_h, U_h])$. Applying (132) uniformly over this family of local laws and choosing an attainable moment of smallest magnitude gives

$$\begin{aligned} c d_h^2 h^4 \log(1/h) - C\left(h^4 + \frac{h^2}{m^2}\right) &\le \inf_Q \|B_h\|_2^2 \\ &\le C\left[h^4\{1 + d_h^2 \log(1/h)\} + \frac{h^2}{m^2}\right]. \end{aligned} \tag{141}$$

Thus exact feasibility in (140) is sufficient. In the regime $mh \ge c_0 > 0$, the precise order criterion for achievable quartic mean error is $d_h = O(1/\sqrt{\log(1/h)})$. It is also the tolerance for a small residual moment in a proposed cancellation law. This optimizes the leading boundary moment within the stated local-law family; it does not determine the remaining quartic coefficient or the best deterministic legal assignment.

Under a uniform transverse endpoint gap $|\tau| \ge \gamma h$, every feasible V_r is at least γ^2. Same-sign nonzero outputs then cannot make M_h vanish. Signed outputs can make (140) feasible, but a separation from product SR additionally requires that the product moment $M_h^{\mathrm{SR}} = \sum_r w_r d_r$ stay bounded away from zero and that the target retain nonvanishing signal. A sufficient signal condition here is $K \ne 0$. Under these conditions and $mh^2 \ge c_0 > 0$, an exactly cancelling legal law has $O(h^4)$ expected loss, while the same teacher's product SR has $\Omega(h^4 \log(1/h))$ expected loss. Conditional expectation again supplies one legal deterministic vector with no larger loss than the constructed independent-across-neurons law. Appendix A5 verifies all these conditions with fixed rational parameters and a strict gap at every dyadic level.

The common density and the two-bit geometry are substantive restrictions. Different group densities generally require cancellation of the function $\sum_r w_r V_r \rho_r(u)$, not just its integral. Allowing a different transverse law at each direction likewise leaves the scalar criterion above. With more than two coupled hidden bits, the feasible covariance set can require a higher-dimensional discrete optimization; (140) is not a polynomial algorithm for that problem. Finally, all laws in the two-bit construction retain the output and cross-neuron independence used in (125). Their expected raw bilinear proxy losses, and their true expected losses under identity activation, are therefore identical even when their boundary moments differ.

A6.5 A MATCHING QUARTIC FLOOR AND A SIGNED CONVEX-RELAXATION GUARANTEE

A common affine region makes the quartic order in Appendix A5 sharp over all its original legal assignments. The argument also applies to signed mixtures of endpoint candidates, and therefore to the full function-space convex relaxation.

First, let f be any positively one-homogeneous function on $\mathbb{R}^2$ with finite Gaussian second moment. Write $f(r, \theta) = rF(\theta)$, let F_0 be its angular mean, and let $F_\perp$ be its component orthogonal to $1, \cos\theta, \sin\theta$ in normalized angular L^2. The first angular harmonics give linear functions and can be removed exactly. Independence of radius and angle gives

$$\begin{aligned} \operatorname{dist}^2_{L^2(\mathrm{Unif}(D))}(f, \mathcal{A}_2) &= \frac{F_0^2}{18} + \frac{1}{2}\|F_\perp\|^2, \\ \operatorname{dist}^2_{L^2(\gamma_2)}(f, \mathcal{A}_2) &= \left(2 - \frac{\pi}{2}\right) F_0^2 + 2\|F_\perp\|^2, \\ \inf_{\ell \in \mathcal{A}_2} \int_D |f(z) - \ell(z)|^2 dz &\ge \frac{\pi}{8} \operatorname{dist}^2_{L^2(\gamma_2)}(f, \mathcal{A}_2), \end{aligned} \tag{142}$$

where D is the unit disk and $\mathcal{A}_2$ denotes affine functions of two variables. Indeed the uniform-disk radius has variance $1/18$ and second moment $1/2$, whereas the Gaussian radius has variance $2 - \pi/2$

and second moment two. The final inequality follows from $\pi > 28/9$, and the factor π converts normalized disk measure to Lebesgue measure.

Now take $X = (Z, Y) \sim N(0, I_3)$, with a target $f(Z)$ of this type, and any finite signed candidate $g = \sum_i c_i \sigma(w_i^\top Z + \tau_i Y)$. Suppose $\|w_i\| \le B$ and $|\tau_i| \ge \gamma h > 0$ for every candidate ridge. Choose $r_0 > 0$ with $Br_0 < \gamma$ and $r_0 h \le 1$. On $\|Z\| \le r_0 h$, $1 \le Y \le 2$, every candidate gate is fixed by the sign of τ_i. Thus g, with any affine correction added, is affine in Z for each fixed Y. Scaling the disk in (142), and bounding the Gaussian density below on this cylinder, gives

$$\|g + \ell - f\|_2^2 \ge \frac{e^{-5/2}}{(2\pi)^{3/2}} \frac{\pi}{8} r_0^4 h^4 \operatorname{dist}^2_{L^2(\gamma_2)}(f, \mathcal{A}_2) \quad \text{for every affine } \ell. \tag{143}$$

There is no restriction on the signed coefficients or finite width in this argument. The target's independence of Y makes its Gaussian distance to affine functions of (Z, Y) equal to its two-dimensional distance in (143).

Complete statement of Corollary 4. For the fixed signed teachers (103), every legal assignment and every affine correction satisfy

$$\boxed{\|\widehat{f}_C + \ell - f_m\|_2^2 \ge c_{\mathrm{gap}} h^4, \qquad c_{\mathrm{gap}} = \frac{289 e^{-5/2}}{80000 \cdot 64^4 (2\pi)^{3/2}} > 0.} \tag{144}$$

The same lower bound holds for every point in the full convex hull of legal assignments. To see this, every endpoint base has norm at most $2\sqrt{517/80} + \sqrt{3}h < 6$, while (109) gives $|\tau| \ge h/\sqrt{85} > h/10$. On $\|Z\| \le h/64$, $1 \le Y \le 2$, its base contribution has magnitude less than $3h/32 < h/10$, so all endpoint gates are fixed simultaneously. Apply (143) with $r_0 = 1/64$ and the signal bound (107). Every convex combination remains affine on the same cylinder, which proves the convex-hull assertion as well.

Combining (144) with (105) shows that the unrestricted original-legal optimum has order h^4 when $mh^2 \ge 1$. The same is true with best affine correction. Thus this family has a sharp quartic legal optimum and a logarithmic product-SR obstruction; the result is stronger than an upper-bound comparison against SR alone.

For completeness, this floor also turns the convex-function-space rounding guarantee into a relative statement for a signed teacher. Let L_* be the true legal optimum, and let R_* minimize the same squared loss over the sum of the original local candidate convex hulls. Both ordinary loss and loss after projection away from affine functions can be used consistently. Equation (144) gives $c_{\mathrm{gap}} h^4 \le R_* \le L_*$. If a feasible convex point with objective at most $R_* + \varepsilon_{\mathrm{opt}}$ is obtained, independent local rounding followed by conditional expectation gives

$$\begin{aligned} L(C) &\le L_* + \varepsilon + \frac{1}{2m^2} \sum_j D_j^2 \le L_* + \varepsilon + \frac{Ch^2}{m}, \\ \frac{L(C)}{L_*} &\le 1 + \frac{C}{mh^2} + \frac{\varepsilon}{L_*}. \end{aligned} \tag{145}$$

Here D_j is the diameter of the unscaled local candidate set in the chosen Hilbert norm, $D_j \le Ch$, and ε includes optimization and final conditional-expectation numerical errors. The variance step is the exact identity $\mathrm{Var}(F) = \frac{1}{2}\mathbb{E}\|F - F'\|^2$, with F' an independent copy. Projection does not increase the diameters. This guarantee requires the stated convex objective accuracy; it does not assert that an implemented solver has achieved it. A numerical relaxation lower certificate equal to zero can be weak even though the analytic floor proves $R_* > 0$.

A7 Signed fixed-teacher experiment and certificate interpretation

The complete twelve-case record is `results/nonlinear_geometry_signed_covariance.csv`. It is generated by

```
.\.venv\Scripts\python.exe -X utf8
  experiments/nonlinear_geometry_study.py --signed-covariance-only
```

Table 4: Complete signed fixed-teacher study. B^2 is the squared mean error; losses are ordinary Gaussian MSE. All interval comparisons use exact rational laws. The same CE assignment is returned under both laws in every row; all numerical convex lower certificates are zero.

m	h	nh	B^2_{SR}/h^4	B^2_{Q}/h^4	$\mathbb{E}_{\mathrm{Q}}L/\mathbb{E}_{\mathrm{SR}}L$	$L_{\mathrm{CE}}/\mathbb{E}_{\mathrm{SR}}L$
4	$1/16$	0.125	0.02045	0.01918	1.000018429	0.027836
4	$1/32$	0.0625	0.04868	0.04695	1.000016124	0.050579
4	$1/64$	0.0313	0.10793	0.10348	1.000004864	0.028201
4	$1/128$	0.0156	0.21775	0.21063	1.000003988	0.047354
16	$1/16$	0.500	0.00399	0.00299	0.999992011	0.022006
16	$1/32$	0.250	0.00796	0.00464	0.999995240	0.047873
16	$1/64$	0.125	0.02127	0.01837	1.000000094	0.117744
16	$1/128$	0.0625	0.05145	0.04710	1.000002666	0.047281
64	$1/16$	2.00	0.00396	0.00297	0.999876079	0.014246
64	$1/32$	1.00	0.00611	0.00273	0.999923018	0.002238
64	$1/64$	0.500	0.00582	0.00306	0.999980970	0.015758
64	$1/128$	0.250	0.01021	0.00392	0.999993560	0.004093

Join these two lines with a space when invoking the command. Widths $4, 16, 64$ are squares, so the exact rational canonical API represents the unscaled signed family without introducing an irrational output coefficient. At each width the teacher parameters are identical at all four grid steps.

The study supplies the analytic exact local probabilities of (9); it does not use a fitted convex proposal. For both the constructed law and independent product SR it records Arb enclosures of expected true MSE, squared mean error, variance and the selected legal vector's MSE. It also records target total, linear and nonaffine energies, local support and smallest mass, direction count, transverse variances, actual optimization-gap upper bounds, convex lower/upper certificates and the complete endpoint assignments. Expected raw bilinear-proxy and identity-activation values are calculated from the exact marginal formula (14). The reported table ratios use the numerical values of these kernel quantities; the gain-sign columns separately use interval endpoints.

Every hidden local law has eight positive states, and adding the independently rounded output gives sixteen states. All neurons' teacher directions are distinct. Mean-bias improvement is present in all twelve cases. Six total expected improvements are certified; the other six have a small deterioration. The assignments selected by the two CE runs agree exactly in all twelve records. In particular, the numerical equality of their losses is not an inference from overlapping intervals. All external-law convex lower certificates are zero. The upper certificate is the supplied mean-function objective; the returned optimization gap is the gap actually certified, not a requested solver tolerance or proof that the supplied law is near the relaxed optimum. All cases have $nh \le 2$ and $mh^2 < 1$. They are outside the explicit SR lower-bound regime and the original independent-neuron sufficient width regime $mh^2 \ge 1$. The newer cell guarantee also covers $m = 64, h = 1/16$ at $m^2h^3 = 1$, with conservative constants.

The API `nonlinear_rounding/conditional.py` evaluates the original complete Gaussian residual Gram matrix, including within-neuron off-diagonal terms. Exact supplied probabilities may be correlated and may include zero masses. The numerical proposal path, when used, rationalizes the proposed simplex probabilities; it does not certify a floating-point optimization result by trusting solver status. Loss, variance, the convex gradient lower bound and conditional expectation are enclosed with Arb. A failed proposal reverts to a legal distribution and reports that fact. The current routine accepts rational canonical inputs and certifies raw Gaussian MSE; it does not implement the affine-projected objective of the mathematical extension. The direct complete-state method uses quadratic memory in the total number of local states and therefore does not make the wide asymptotic theorem an inexpensive dense-kernel experiment.

The historical 139-test suite, preceding the 143-test shared-cell/independent-CE suite, includes exhaustive legal-set checks, an independent 512-bit Gaussian-kernel expansion, correlated/zero-mass laws, invalid floating proposals, zero-loss and zero-variance cases, and interval precision-context restoration. The current full suite passes all 189 tests, including seventeen combinatorial prefix tests, four certified-prefix tests, two native integer-quadratic tests and five exact-law tests. These checks

Table 5: Complete registered 2×2 comparison: all 24 rows from twelve paired cases. Entries are h^4-normalized interval midpoints rounded for display, not outward certificate endpoints. Each law's squared mean B^2 is shared by its independent and prefix couplings; their actual variances differ. All columns use the final canonical CSV, including freshly recomputed independent variances and expected losses.

m	$N = 1/h$	$N \bmod 5$	Law	B^2/h^4	V_{ind}/h^4	V_{pre}/h^4	$\mathbb{E}_{\mathrm{ind}} L/h^4$	$\mathbb{E}_{\mathrm{pre}} L/h^4$
4	16	1	P	2.045e-02	1.055e+02	2.109e+02	1.055e+02	2.109e+02
4	16	1	Q	1.918e-02	1.055e+02	2.144e+02	1.055e+02	2.144e+02
4	32	2	P	4.868e-02	5.778e+02	1.302e+03	5.779e+02	1.302e+03
4	32	2	Q	4.695e-02	5.779e+02	1.504e+03	5.779e+02	1.504e+03
16	16	1	P	3.991e-03	2.613e+01	4.998e+01	2.613e+01	4.999e+01
16	16	1	Q	2.987e-03	2.613e+01	3.637e+01	2.613e+01	3.637e+01
16	32	2	P	7.955e-03	1.439e+02	2.719e+02	1.439e+02	2.719e+02
16	32	2	Q	4.644e-03	1.439e+02	2.183e+02	1.439e+02	2.183e+02
64	16	1	P	3.962e-03	6.470e+00	3.102e+00	6.474e+00	3.106e+00
64	16	1	Q	2.967e-03	6.470e+00	2.739e+00	6.473e+00	2.742e+00
64	32	2	P	6.107e-03	3.561e+01	1.998e+01	3.562e+01	1.998e+01
64	32	2	Q	2.728e-03	3.561e+01	4.893e+00	3.561e+01	4.896e+00
256	16	1	P	3.954e-03	1.616e+00	5.969e-02	1.620e+00	6.365e-02
256	16	1	Q	2.963e-03	1.616e+00	1.183e-01	1.619e+00	1.212e-01
256	32	2	P	6.168e-03	8.886e+00	7.293e-01	8.893e+00	7.355e-01
256	32	2	Q	2.782e-03	8.887e+00	3.738e-01	8.889e+00	3.766e-01
256	64	4	P	6.097e-03	2.589e+01	3.388e+00	2.589e+01	3.395e+00
256	64	4	Q	3.336e-03	2.589e+01	1.166e+00	2.589e+01	1.169e+00
1024	16	1	P	3.949e-03	4.040e-01	9.638e-03	4.080e-01	1.359e-02
1024	16	1	Q	2.959e-03	4.040e-01	9.113e-03	4.070e-01	1.207e-02
1024	32	2	P	6.170e-03	2.221e+00	6.578e-02	2.227e+00	7.195e-02
1024	32	2	Q	2.784e-03	2.221e+00	2.294e-02	2.224e+00	2.572e-02
1024	64	4	P	6.128e-03	6.468e+00	6.733e-02	6.474e+00	7.346e-02
1024	64	4	Q	3.367e-03	6.468e+00	6.443e-02	6.471e+00	6.780e-02

support their respective implementations. The mathematical claims are established by the preceding finite-step proofs, subject to the research draft's AI-use disclosure.

A7.1 Paired prefix experiment: protocol and certificate checks

Run `pythonexperiments/nonlinear_geometry_study.py--signed-prefix-only`; an interrupted registered run continues with the additional flag `--resume-signed-prefix`. The fixed twelve cases are $m \in \{4, 16, 64\}$ with $h \in \{1/16, 1/32\}$, and $m \in \{256, 1024\}$ with $h \in \{1/16, 1/32, 1/64\}$. Each case evaluates both P and Q, giving 24 completed rows across all twelve cases. The canonical file is `results/nonlinear_geometry_signed_prefix.csv`; the final file includes the independent-variance postprocessing and exact input fields.

Tables 5 and 6 show all twelve registered cases. For each row, the prefix mean enclosure overlaps the matching certified independent-CE mean, and both the historical mean-plus-variance enclosure and the shared-prefix-mean-plus-independent-variance enclosure overlap the independent expected-loss enclosure. All interval endpoints are restored as exact `Fraction(float(text))` values before these checks. Fresh independent variances and their sums with the same-law mean have also been checked against the historical certified records; the displayed independent columns now use these postprocessed canonical values.

The worst selected-loss ratios to SR-CE are 3.336060 for P prefix and 6.759680 for Q prefix (rounded for display). Thus favorable win counts coexist with substantial losses on individual cases.

At each width the teacher is identical across steps. Both laws use the same original sixteen Cartesian states per neuron. Prefixes span each entire signed scale group, sorted by the exact rational direction coordinate u; they do not restart in smaller geometric cells. The two laws share the ordering and decomposition rule, but their different probabilities generally produce different support atoms. Exact rational checks verify the complete state ordering, positive decomposition weights summing to one, every atom's legality and every label's floor/ceiling count at every prefix, exclusion of zero-probability states, and recovery of all complete local marginals. Scalar marginals then follow from the supplied complete P or Q rows.

The API evaluates the true Gaussian loss of every support atom with Arb. It encloses the randomized prefix expectation, squared mean, variance, support minimum and selected loss, and certifies selection

Table 6: Deterministic selections for all twelve registered cases, all losses divided by h^4. Values are rounded interval midpoints. The three CE baselines are re-evaluated on the identical teacher and state order; all twelve cases and all worse prefix selections are shown.

m	$N = 1/h$	SR-CE	Q-CE	Convex-CE	P prefix	Q prefix
4	16	2.936e+00	2.936e+00	6.997e-01	3.006e+00	3.006e+00
4	32	2.923e+01	2.923e+01	3.039e+01	9.751e+01	1.976e+02
16	16	5.751e-01	5.751e-01	4.839e-01	9.483e-01	3.347e-01
16	32	6.889e+00	6.889e+00	2.137e+00	1.817e+00	1.621e+00
64	16	9.223e-02	9.223e-02	6.276e-02	2.715e-02	9.752e-03
64	32	7.972e-02	7.972e-02	3.844e-01	1.032e-01	1.095e-01
256	16	8.932e-03	1.156e-02	4.743e-03	4.242e-03	3.034e-03
256	32	8.850e-03	8.850e-03	1.769e-02	1.261e-02	4.470e-03
256	64	1.514e-02	1.514e-02	9.964e-02	1.318e-02	1.021e-02
1024	16	2.624e-03	4.799e-03	1.213e-03	3.525e-03	2.630e-03
1024	32	4.574e-03	4.537e-03	5.342e-03	6.262e-03	2.477e-03
1024	64	1.238e-02	1.307e-02	9.829e-03	5.492e-03	3.649e-03

within $\eta = 2^{-64}$ of the support minimum and below the randomized expectation plus η. Selection tolerance is separate from the randomized expected loss. A separately expanded mean-function calculation must overlap the API's mean enclosure. The recorded SR-CE, Q-CE and convex-row CE assignments are each re-evaluated on this exact teacher and Cartesian ordering; their original records are used only after loss-enclosure overlap checks. These are three deterministic baselines, not additional marginal-preserving output laws.

Postprocessing recomputes each law's independent-neuron variance from exact finite-cell statistics and combines it with that same law's prefix-mean enclosure. This supplies the independent versus prefix comparison while keeping the local mean fixed. The final CSV fields include exact rational local probabilities and endpoint inputs, state ordering, assignments, loss intervals, arithmetic precision, software versions, h^4-normalized quantities, mh, $N \bmod 5$ for $N = 1/h$, and η/h^4. Reported point values are interval midpoints; interval widths and selection tolerance remain distinct. All 24 rows contain these final independent-comparison fields.

API time covers decomposition, compressed Gram construction and selection over every support atom. It excludes subsequent independent mean checks, baseline re-evaluations and independent-variance postprocessing. Memory measurements are cumulative process-lifetime peaks, not isolated case peaks or disk-write volumes; the first eighteen historical rows have no measurement. The later six rows have observed process peaks, sampled outside the prefix API timed interval. Exact readings, measurement scope and timestamps remain in the CSV. Mixed implementation versions preclude a uniform speed comparison. CSV bounds are restored through their recorded binary64 values before exact rational comparisons; the large-field reader and continuation checks are documented in the repository README. The finite case grid is not a fixed-mh refinement trajectory, and none of its cases enters the sufficient explicit SR logarithmic lower-bound regime. Accordingly the study compares finite losses and certificates without fitting an asymptotic rate.

A7.2 Finite width–step trajectories

The command `pythonexperiments/nonlinear_geometry_study.py--signed-prefix-trajectory-only` writes `results/nonlinear_geometry_signed_prefix_trajectory.csv` and the PDF of Figure 1. The fixed-mh path uses $(m, h) = (16, 1/16), (64, 1/64), (256, 1/256)$, all with $mh = 1$. The wider path uses $(16, 1/16), (64, 1/32), (256, 1/64)$, all with $m = 1/(16h^2)$. They share the first case, giving five distinct teachers/steps and ten P/Q rows: six rows reuse the registered study, while four rows execute the two additional cases. The corresponding $N \bmod 5$ sequences, with $N = 1/h$, are $1, 4, 1$ and $1, 2, 4$. Each width still specifies its own fixed teacher; connecting different widths does not turn the path into one teacher across all points.

Along $mh = 1$, the normalized P squared mean is $.003991, .005817, .007936$ and the Q squared mean $.002987, .003059, .003186$. Their prefix expected losses are $49.988684, 26.680984, 46.726815$

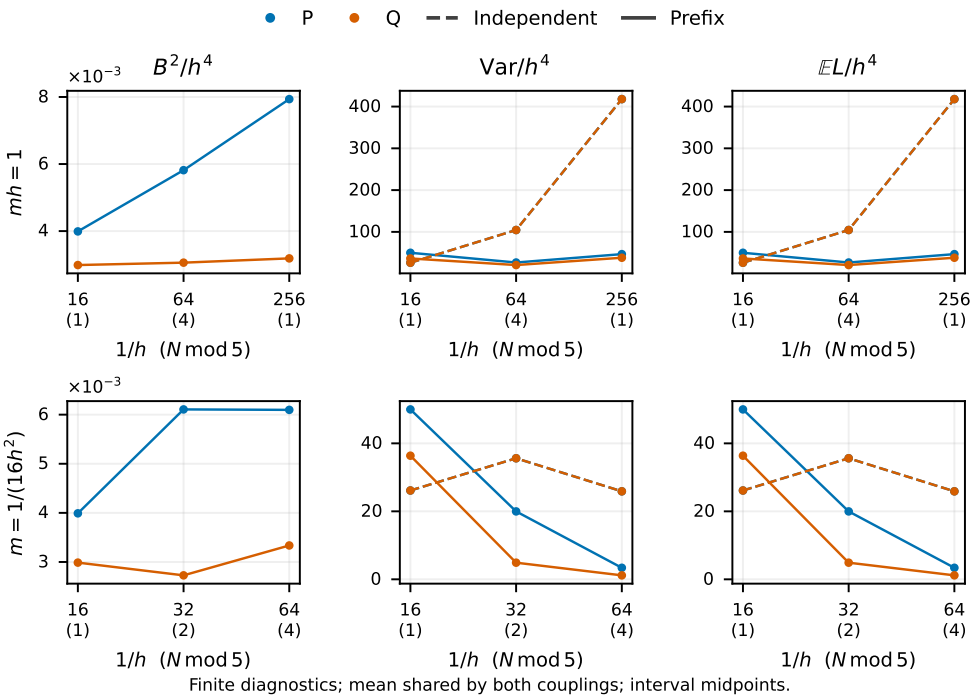


Figure 1: Finite refinement trajectories: $mh = 1$ (top) and $m = 1/(16h^2)$ (bottom). Columns show squared mean, variance and randomized expected loss, all divided by h^4. Each local law has the same mean under both couplings; dashed curves use independent and solid curves prefix coupling. Independent P/Q curves nearly coincide. Error bars are numerical certificate intervals, often smaller than the markers, not Monte Carlo uncertainty. Grid residues are displayed on the horizontal axes. Lines connect the three evaluated points without a rate fit.

and $36.370222, 20.596317, 38.087892$, respectively, showing substantial nonmonotone finite variation. Along the wider path the corresponding expected losses are $49.988684, 19.982485, 3.394525$ and $36.370222, 4.896163, 1.169492$. At the common initial case both prefix variances exceed their independent counterparts. These three-point diagnostics do not establish limiting normalized constants or asymptotic rates, and none meets the sufficient explicit finite SR logarithmic lower-bound conditions.

The figure reports randomized expectations, not deterministic selected losses. The extra cases also execute SR-CE and Q-CE, with exact input and state-order checks; they do not run a convex-row CE baseline. All selected losses and available baselines remain in the CSV. In particular, at the reused wider-path case $(64, 1/32)$, Q has lower randomized prefix expectation but its selected loss exceeds P's, and both selected losses exceed SR-CE. At $(16, 1/16)$ the selected P prefix also loses to SR-CE. Reused rows are identified as such and are not counted as fresh executions; all five cases and both laws are retained. Timing and memory fields follow the scope distinctions in Appendix A7.1.

A7.3 Actual Cartesian means for fixed continuous densities

Continuous-density mean diagnostic. Figure 2 evaluates the actual Cartesian rounding mean for the fixed known densities $\rho_{1,2}(u) = 4[1 \pm \varepsilon(8u - 13)]$ on $[3/2, 7/4]$, with $\varepsilon = .01$. The teacher is unchanged as h varies. At quadrature order 128, as b increases from 8 to 16, $\|B_P\|_2^2/h^4$ rises from .00821370 to .01676696, while the density-adapted Q stays between .00344312 and .00354031. For $R_\theta = (1-\theta)Q + \theta P$, the endpoint values are .00438503, .00656329 at $\theta = 1/2$ and .00405839, .00448341 at $\theta = 1/\sqrt{\log(1/h)}$. This is a numerical mean diagnostic consistent with cancellation and its tolerance scale; it neither samples a finite teacher nor executes prefix rounding or certifies an asymptotic rate. The 64-to-128 sensitivity reaches 3.66% for Q and exceeds the observed difference from the unadapted law at this one-percent density mismatch (Appendix A7.3).

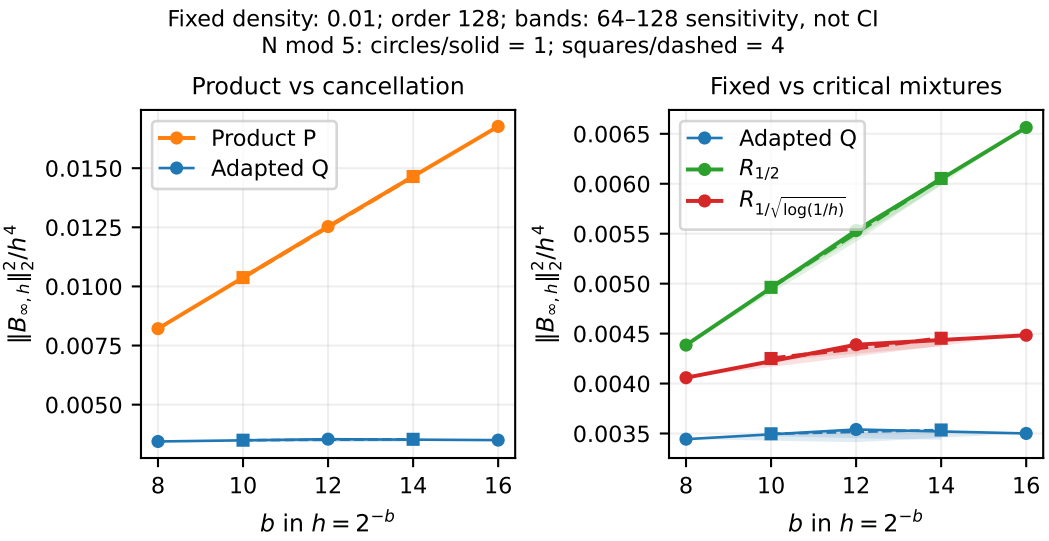


Figure 2: Actual continuous-teacher rounding means, with known density mismatch $\varepsilon = .01$. Left: product and density-adapted laws. Right: the adapted law and fixed versus critical mixtures. Both panels show $\|B\|_2^2/h^4$. Lines connect steps within the same residue class of 2^b modulo five. Bands show order-64 to order-128 quadrature sensitivity, not confidence intervals or certified error bounds.

The file `results/nonlinear_geometry_signed_density.csv` retains all 150 records: $b \in \{8, 10, 12, 14, 16\}$, $\varepsilon \in \{0, .01\}$, quadrature orders $32, 64, 128$, and five laws $P, Q, Q_{\text{fixed}}, R_{1/2}, R_{1/\sqrt{\log(1/h)}}$. The two densities are $4[1 \pm \varepsilon(8u - 13)]$ on $[3/2, 7/4]$; each integrates to one and is fixed across steps. Here P is product SR, Q adapts its two-hidden-bit probabilities to the supplied density ratio, and Q_{fixed} keeps the equal-density cancellation law. Each R_θ mixes the complete local laws, so its actual mean is the corresponding mixture. The figure uses $\varepsilon = .01$; the raw file also retains the equal-density control and every lower-order evaluation.

For each recorded (b, ε), `signed_density_prepare` reconstructs the weighted two-hidden-state masses $\rho_r(u)p_k(u)$ as exact rational affine functions: positivity, normalization, coordinate means and Q's weighted cancellation checked at $u = 3/2, 7/4$ hold throughout the interval by affine interpolation, while each conditional mass, affine divided by positive affine, is monotone or constant. Independent middle-coordinate and output SR then supply a legitimate complete joint law, including zero-probability bit states; $R_{1/2}$ and $R_{1/\sqrt{\log(1/h)}}$ are legal convex mixtures since their mixing weights lie in $[0, 1]$ on the registered grid. Thus the recorded (b, ε) and source formulas reconstruct the continuous laws without storing probability rows at every u; these exact feasibility checks do not certify the numerical sphere integral.

For each spherical input direction, the calculation integrates the actual hidden-coordinate rounding mean over u. The four first/third-coordinate endpoint states have probabilities whose products with ρ are affine in u. For each such state and hidden scale λ, middle-coordinate SR is linear interpolation of the hinge on the u grid of spacing h/λ. Its difference from the true hinge is a triangle in the unique grid cell containing the hinge root, and zero elsewhere. Piecewise-polynomial integration therefore handles the direction integral without a coarse u quadrature that could miss narrow activation intervals. Independent output rounding preserves the original signed output in this mean calculation.

Homogeneity and central symmetry then express the Gaussian squared norm as $3/(2\pi)$ times its hemisphere integral. This remaining sphere integral is numerical. A separate rational `Fraction` calculation integrates over all intersecting grid cells for direction checks; the largest recorded absolute disagreement is 1.03×10^{-16}. Probability normalization, density normalization and the Q functional-moment cancellation identity are also checked by the experiment. These checks are not an interval certificate for the sphere integral.

Table 7: All twelve shared-cell cases. B is the number of nonempty cells; $S = \mathbb{E}_{\mathrm{SR}}L$. U_Q bounds randomized Q-cell expected loss, whereas L_Q, L_C are the two deterministic selected losses. CE is independent-SR conditional expectation, evaluated with compressed candidates at all twelve widths and steps. Ratios use certified upper/lower endpoints and are rounded for display.

m	h	B	U_Q/S	L_Q/S	L_C/S	L_Q/L_{CE}	L_C/L_{CE}
4	1/16	4	4.4508	0.08295	0.02722	2.980	0.978
4	1/32	4	3.4412	0.34597	0.03783	6.840	0.748
16	1/16	12	4.3427	0.15191	0.08725	6.903	3.965
16	1/32	16	3.4540	0.07388	0.06974	1.543	1.457
64	1/16	12	2.3362	0.00193	0.03562	0.136	2.500
64	1/32	24	2.9249	0.01289	0.00078	5.761	0.346
256	1/16	12	0.8228	0.00914	0.00522	1.657	0.947
256	1/32	24	0.8629	0.00270	0.00305	2.710	3.062
256	1/64	48	2.3073	0.00379	0.00186	6.486	3.185
1024	1/16	12	0.2268	0.00808	0.00060	1.257	0.094
1024	1/32	24	0.2552	0.00448	0.00119	2.183	0.578
1024	1/64	48	0.8129	0.00646	0.00168	3.376	0.879

Across both density settings and all five steps, the largest relative change from order 64 to 128, using order 128 as denominator, is 0.775% for P, 3.66% for Q and Q_{fixed}, 2.05% for $R_{1/2}$, and 2.70% for $R_{1/\sqrt{\log(1/h)}}$. At $\varepsilon = .01$, Q's order-128 quartic normalization ends at .00350058 for $b = 16$; the discrepancy from Q_{fixed} is smaller than the quadrature sensitivity. The study therefore does not resolve a numerical advantage from adapting to this small mismatch. It measures the continuous mean alone, with no finite-teacher sampling variance, prefix execution or certified rate claim.

A7.4 Shared-cell implementation and all finite certificates

The fine-cell implementation uses $D_c < 6h$ and the fractional-count penalty

$$\mathcal{V}(q) = \frac{1}{2m^2}\sum_c D_c^2 \sum_k \xi_{ck}(1-\xi_{ck}) \le \min\{18h^2/m, 54h/m^2\}, \quad \xi_{ck} = \lambda_{ck} - \lfloor \lambda_{ck} \rfloor. \quad (146)$$

It certifies loss at most $\|\mu - f_m\|_2^2 + \mathcal{V}(q)$ plus selection tolerance. Numerical proposals receive feasible-loss and convex-gap certificates, without a promised optimization accuracy. The full coupling theorem is in Appendix A3; it also proves Corollary 14.

The command `pythonexperiments/nonlinear_geometry_study.py--signed-cell-only` evaluates the twelve fixed cases of Table 7; `results/nonlinear_geometry_signed_cells.csv` stores every case. The entry point `nonlinear_rounding/cell_rounding.py` groups identical complete local candidate dictionaries, constructs the compressed Gaussian residual Gram matrix, and applies rational cycle/path updates with certified potential comparisons. Inputs use $a_j = s_j/\sqrt{m}$ and output step $h/\sqrt{m}$, so the code's canonical $m^{-1/2}$ normalization equals the theorem's unscaled m^{-1}. All tested widths are squares. Each teacher is unchanged across its steps.

For Q-cell, the input rows are exactly the prescribed modified local laws; no optimizer is invoked. For C-cell, Clarabel proposes aggregate simplex probabilities for the unrestricted true-function objective. Every run returns `clarabel:optimal_inaccurate;rationalized_exact_cell_simplex`. Each proposal is then rationalized with 40 probability bits and recertified; this status supplies no mathematical certificate by itself. Certified matrix rounding uses requested absolute selection tolerance 2^{-64}. The arithmetic precision is 96 bits in all twelve cases. The implementation evaluates each law's independent-neuron mean, variance and expected loss, and same-teacher product SR's corresponding quantities, through the same exact compressed Gaussian kernel calculations. All recorded assignments and floor/ceiling cell counts pass exact legality checks.

The unrestricted convex lower certificate for aggregate weights λ follows from the supporting hyperplane. If $F(\lambda)$ is the feasible squared loss and $g = \nabla F(\lambda)$, then

$$R_* \ge \max\{0, F(\lambda) + \sum_c n_c \min_k g_{ck} - g^\top \lambda\}.$$

Table 8: All twelve unrestricted compressed true-convex certificates. Absolute loss columns are scaled by h^4. R_{lb} is the convex lower certificate, F the feasible objective, $\varepsilon_{\mathrm{opt}}$ the achieved optimization-gap upper bound, $\mathcal{V}_C$ the fractional-count penalty, and $\rho_{\mathrm{cert}} = L_C^+/R_{\mathrm{lb}}$ the relative certificate when $R_{\mathrm{lb}} > 0$. Ratios are rounded upward to four significant figures; N/A denotes an insufficient zero lower certificate. Other displayed values are rounded. Every solver status is `optimal_inaccurate`.

m	h	R_{lb}/h^4	F/h^4	$\varepsilon_{\mathrm{opt}}/h^4$	$\mathcal{V}_C/h^4$	L_C/h^4	ρ_{cert}
4	1/16	2.432e-03	2.448e-03	1.608e-05	3.473e+02	2.872e+00	1181
4	1/32	8.147e-03	9.334e-03	1.188e-03	1.177e+03	2.186e+01	2684
16	1/16	1.805e-04	1.836e-04	3.061e-06	1.087e+02	2.280e+00	1.263×10^4
16	1/32	1.184e-03	1.223e-03	3.877e-05	4.652e+02	1.004e+01	8476
64	1/16	7.539e-05	7.940e-05	4.015e-06	1.143e+01	2.306e-01	3059
64	1/32	2.657e-05	5.622e-05	2.966e-05	9.497e+01	2.762e-02	1040
256	1/16	7.342e-05	7.970e-05	6.275e-06	6.888e-01	8.462e-03	115.3
256	1/32	6.548e-06	4.846e-05	4.191e-05	7.585e+00	2.710e-02	4139
256	1/64	0.000e+00	1.274e-04	1.274e-04	6.380e+01	4.822e-02	N/A
1024	1/16	7.816e-05	7.926e-05	1.108e-06	4.968e-02	2.468e-04	3.158
1024	1/32	1.711e-05	4.689e-05	2.978e-05	4.580e-01	2.644e-03	154.6
1024	1/64	0.000e+00	1.389e-04	1.389e-04	4.132e+00	1.088e-02	N/A

Arb enclosures certify both sides after exact simplex repair. $F - R_{\mathrm{lb}}$ bounds the optimization gap; $L_C - R_{\mathrm{lb}}$ bounds the legal global gap. The potential guarantee also gives $L_C \le F + \mathcal{V}_C + 2^{-64}$. These are distinct bounds. A convex bracket width is the achieved optimization gap, and is not the uncertainty of a single interval evaluation.

For $R_{\mathrm{lb}} > 0$, $R_{\mathrm{lb}} \le L_*$ gives $L_C/L_* \le \rho_{\mathrm{cert}}$. The smallest displayed certificate is 3.158, so these bounds do not establish finite-width near-optimality; the two zero-lower-bound cases give no relative certificate. The original absolute selected-loss gaps G_C remain in the CSV. The common selection tolerance is $\eta = 2^{-64}$, with η/h^4 approximately $3.5527136788 \times 10^{-15}$, $5.6843418861 \times 10^{-14}$ and $9.0949470177 \times 10^{-13}$ at $h = 1/16, 1/32, 1/64$, respectively. This additive selection tolerance is distinct from the Arb enclosure width, optimization gap and variance penalty.

The prescribed-row fractional penalties $\mathcal{V}_Q$, in the row order of Table 7, are

$$7.1641\times10^{-3},\ 1.8965\times10^{-3},\ 1.7317\times10^{-3},\ 4.7398\times10^{-4},$$
$$2.3074\times10^{-4},\ 9.9347\times10^{-5},\ 2.0297\times10^{-5},\ 7.3151\times10^{-6},$$
$$3.5608\times10^{-6},\ 1.3670\times10^{-6},\ 5.3939\times10^{-7},\ 3.1347\times10^{-7}.$$

Their sum with the certified mean loss gives U_Q in the main table. No simulation or enumeration of the randomized coupling's full distribution is used; its expected loss is bounded analytically by that potential. The deterministic selections do not preserve marginal distributions. The input Q_j law is not product SR, even though every scalar has its SR marginal.

The old CE baselines are reused from the existing exact small-teacher CSV; no baseline is silently replaced. At $m = 4$, exhaustive 192-bit kernel evaluation of all 65536 original legal vectors yields global loss intervals around $7.01328239860428 \times 10^{-6}$ and $8.76035561992496 \times 10^{-6}$. The Q-cell absolute suboptimality bounds are 1.2652×10^{-4} and 1.8191×10^{-4}; the C-cell bounds are 3.6803×10^{-5} and 1.2091×10^{-5}. These expose sizable selection gaps despite accurate convex proposals. At larger widths exhaustive global optimization is not performed.

The compressed state count ranges from 64 to 768. At $m = 1024$ it is 192, 384 and 768, compared with 16384 per-neuron variables. Across all twelve cases the recorded Q-cell and C-cell runtimes total approximately 179 and 187 seconds; the largest individual times are about 55 and 62 seconds. These measurements include their case-level certified calculations on this machine and are not a general speed comparison. None of these experiments evaluates the different-density random-teacher family in Corollary 14, and the independent-law proxy equality is not asserted for the cell coupling.

Shared cells: certified gains over SR, mixed strong-baseline comparisons. A new fixed twelve-case study uses $m \in \{4, 16, 64\}$ with $h \in \{1/16, 1/32\}$, and $m \in \{256, 1024\}$ with $h \in \{1/16, 1/32, 1/64\}$. It compares potential selection from the prescribed Q_j rows (Q-cell) with

Table 9: All twelve compressed CE comparisons. $L_{\text{SR-CE}}$ starts from scalar SR; $L_{Q\text{-CE}}$ starts from the prescribed correlated rows; $L_{C\text{-CE}}$ uses exactly the fresh C-cell rows. Ratios use outward loss endpoints and are rounded for display.

m	h	$L_{Q\text{-CE}}/L_{\text{SR-CE}}$	$L_{C\text{-CE}}/L_{\text{SR-CE}}$	$L_{C\text{-CE}}/L_C$
4	1/16	1.0000	0.2383	0.2437
4	1/32	1.0000	1.0397	1.3900
16	1/16	1.0000	0.8415	0.2122
16	1/32	1.0000	0.3102	0.2130
64	1/16	1.0000	0.6805	0.2722
64	1/32	1.0000	4.8216	13.9167
256	1/16	1.2941	0.5310	0.5605
256	1/32	1.0000	1.9991	0.6529
256	1/64	1.0000	6.5803	2.0663
1024	1/16	1.8294	0.4622	4.9139
1024	1/32	0.9917	1.1678	2.0207
1024	1/64	1.0554	0.7940	0.9035

selection from the unrestricted compressed true-convex proposal (C-cell). Table 7 retains all twelve cases. All 24 selected vectors have certified lower loss than same-teacher SR's expected loss. The randomized Q-cell *upper bound* certifies improvement in five cases; it is not an exact evaluation of that coupling's expected loss.

Against SR-CE, Q-cell improves only one of twelve cases and C-cell improves seven. Their CE ratios range from 0.136 to 6.903 and from 0.094 to 3.965, respectively. The six small cases retain the earlier one-of-six and three-of-six counts. C-cell improves over Q-cell in ten of twelve cases. A paired followup applies independent CE to each fresh C-cell probability row: it improves seven of twelve C-cell outputs, but loses by up to 13.92 times (Appendix A7.5). Exhaustive enumeration of all $16^4 = 65536$ original legal assignments at each $m = 4$ step shows that the two cell selections remain respectively $19.04, 21.77$ and $6.25, 2.38$ times the global optimum. Thus a lower-loss convex proposal does not guarantee a better rounded vector or a near-optimal one.

Every convex proposal reports `optimal_inaccurate`; exact rationalization and interval bounds nevertheless certify the achieved objective gap, ranging from 7.59×10^{-12} to 1.13×10^{-9}. The convex lower bound is positive in ten cases and zero in two. Appendix A7.4 reports all achieved gaps and penalties. These gaps are optimization bounds, not Arb interval widths; the largest selected-loss enclosure width is 3.25×10^{-23}. The sufficient SR logarithmic conditions fail in every tested case, so this study evaluates finite certificates and deterministic choices, not the asymptotic logarithmic ratio.

A7.5 Compressed independent CE at all twelve cases

The separate command `pythonexperiments/nonlinear_geometry_study.py--signed-independent-ce-only` writes `results/nonlinear_geometry_signed_cells_independent_ce.csv`. It retains 36 outcomes: independent conditional expectation from scalar SR, the prescribed Q, and a freshly optimized true-convex law at each of the same twelve cases. The API option `rounding='independent'` uses the compressed candidate Gram, in original neuron order, without floor/ceiling count restrictions. It certifies selected loss at most the chosen law's independent-neuron expectation plus 2^{-64}.

For each convex pair, C-cell is run once and its exact rational probability rows are passed unchanged to independent CE. All twelve row-equality checks and mean/expectation enclosure checks pass. The fresh C-cell assignment agrees with the earlier frozen assignment in every case; this does not establish equality of the old and new probability rows, since the old CSV did not retain those rows. The six overlapping SR and Q cases also reproduce their earlier assignments. Historical numeric endpoint strings are read back as their outward binary64 values, not as exact decimal rationals. The two width-four exhaustive optima are reused for the same teachers.

All 36 outputs satisfy their independent-CE loss guarantee and use positive-support legal states. C-CE improves on C-cell in seven cases and on SR-CE in seven; neither comparison is uniform. Four whole API calls per case rebuild the compressed Gram separately. Their total recorded time is approximately

Table 10: All eighteen implicit wide-teacher comparisons: numerical $\mathbb{E}_{\mathrm{Q}}L/\mathbb{E}_{\mathrm{SR}}L$ at quadrature order 128. The last two columns report squared mean error divided by h^4 at $m = 2^{32}$. Only the variance component is interval certified. These are prescribed finite teachers, not trained or materialized networks.

h	$m = 2^{16}$	$m = 2^{24}$	$m = 2^{32}$	B^2_{SR}/h^4	B^2_{Q}/h^4
$1/16$	0.9035858	0.7509047	0.7493545	0.00395	0.00296
$1/32$	0.9171518	0.4628204	0.4510636	0.00617	0.00278
$1/64$	0.9742493	0.5769775	0.5498572	0.00613	0.00337
$1/128$	0.9888816	0.4486780	0.3195934	0.00922	0.00294
$1/256$	0.9970640	0.6714137	0.4204777	0.00821	0.00344
$1/512$	0.9989554	0.8020055	0.2507839	0.01228	0.00298

871 seconds, with no claim of equal runtime or speedup over the earlier study. All twelve convex proposals report `optimal_inaccurate`; their rows are recertified before either selection. The comparison closes the missing moderate-width deterministic baseline without treating candidate compression as evidence of loss superiority.

A7.6 Implicit finite teachers in the sufficient width range

Wide fixed teachers: numerical gains with controlled scope. A separate study uses exact arithmetic-grid aggregation to represent the same finite teachers at widths $m = 2^{16}, 2^{24}, 2^{32}$ and all six dyadic steps from $1/16$ to $1/512$. No teacher neuron is materialized or sampled; each enormous finite teacher remains fixed across its grid refinements. The variance is certified with 160-bit Arb arithmetic. The squared mean is integrated numerically over the full hemisphere, including all ratio annuli and both tails, using orders 64 and 128. Thus the reported mean and total losses are estimates, not certified enclosures.

Table 10 retains all eighteen expected-loss comparisons. Every estimated ratio is below one, ranging from 0.2507839 to 0.9989554. At $h = 1/512$, the three width ratios are 0.9989554, 0.8020055 and 0.2507839: the remaining variance can mask most of the mean improvement even at width 2^{16}. This final step meets the explicit SR lower-bound conditions at all three widths; the two larger widths also satisfy $mh^2 \geq 1$. Across all cases, the largest change in the expected-loss ratio from quadrature order 64 to 128 is 0.00120, and the largest mean-loss change divided by h^4 is 1.47×10^{-5}. These are convergence diagnostics, not error bounds. The study evaluates the two laws; it neither executes CE at these widths nor supplies wide-network global-optimality certificates.

The command option `--signed-wide-only` at the same experiment entry point produces `results/nonlinear_geometry_signed_wide.csv` and `results/nonlinear_geometry_signed_wide_annuli.csv`. All 18 combinations of $m = 2^{16}, 2^{24}, 2^{32}$ and $h = 1/16, 1/32, 1/64, 1/128, 1/256, 1/512$ are retained. The method aggregates the original finite arithmetic direction grids by occupied middle-coordinate cells and evaluates their exact affine sums. It does not replace a finite teacher with a continuous teacher law, draw teacher samples, or materialize its neurons. Its floating evaluation of this analytic aggregation is separately checked at selected rational points against an Arb reference.

The law variance is evaluated by exact finite-grid summation and 160-bit Arb Gaussian-kernel expressions. The mean-error function is positively homogeneous; its squared Gaussian norm is reduced to a sphere integral. The full hemisphere covers the complete azimuth and all ratio annuli, including the zero and infinite tails. Every case uses quadrature orders 64 and 128; both total estimates, per-annulus contributions and observed order changes are recorded. The variance intervals are rigorous arithmetic certificates. The mean and total integrals have no rigorous quadrature error bound, and neither order doubling nor selected pointwise Arb agreement certifies them.

A separate agent audit checked all 18 summary rows, 225 annulus rows, and 36 finite-grid count, mean, marginal and transverse-second-moment cases. Independent order-256 checks agreed with certified small-width mean kernels at $m = 16, h = 1/32$ to relative discrepancies below 3×10^{-4}, and with the wide $m = 2^{24}, h = 1/32$ total-loss ratio at order 128 to an absolute change below 1.5×10^{-6}. These are additional numerical cross-checks, not integral certificates.

At $h = 1/512$, the smallest width has $nh = 64$ and $mh^2 = 1/4$; the larger widths have $mh^2 = 64$ and 16384. Thus all three meet the explicit product-SR lower-bound conditions, while only the larger two meet the stated sufficient $mh^2 \geq 1$ quartic-optimality regime at this step. At coarser steps the explicit hypothesis $h \leq 1/512$ does not hold, regardless of width. All eighteen order-128 total-loss estimates favor the covariance law. The maximum absolute 64-to-128 ratio change is 0.00119013; the maximum mean-loss change divided by h^4 is 1.46329×10^{-5}. These observations are finite numerical evidence for the prescribed construction, without an asymptotic regression claim, an executed wide CE assignment, a near-optimality certificate or trained-network evidence.