STABLE MINIMA ARE IMPLICITLY ROBUST LEARNERS AGAINST SPARSE OUTLIERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Overparameterized neural networks trained in practice often do not overfit label noise. They are also surprisingly robust to label corruption, where a few responses are entirely wrong. A common belief attributes both to the stable minima that large learning rates select. Gradient descent with step size η is linearly stable only at minima whose loss Hessian has largest eigenvalue at most $2/\eta$. For label noise this belief has some theoretical support: stable minima of two-layer ReLU networks have bounded weighted variation, and this bound yields generalization from noisy labels. For label corruption much less is known. We study two-layer ReLU networks under squared loss when a small fraction of the responses is shifted by additive outliers of any amplitude, and ask what a stable minimum learns from such data. Our main result concerns the interior of a sample drawn uniformly from a ball of intrinsic dimension m. Consider a stable minimum whose outputs stay bounded on the training points, and suppose the clean signal is simple enough to be representable under the stability constraint. With high probability, fitting the outliers lowers its training loss by at most an error term of order $n^{-(m+3)/(2m+3)}$ up to log factors, at fixed corruption rate and amplitude. Conversely, a stable minimum whose loss comes down to the level set by the outliers alone has learned the clean signal, up to an error of the same order. A larger step size tightens both bounds. The geometric condition cannot be dropped: on a sphere, stable minima can interpolate arbitrary labels. Simulations with certified error bounds support these predictions. The argument extends to logistic loss with label flips.

1 INTRODUCTION

Networks with enough capacity can fit even randomly assigned labels (Zhang et al., 2017). The capacity to memorize is not used immediately: networks fit the clean patterns first and memorize noisy labels only late in training (Arpit et al., 2017). Real supervision is imperfect; ImageNet, for one, contains ambiguous and mislabeled images (Northcutt et al., 2021). The extent to which a network memorizes such labels depends on how it is trained. The step size is one lever. Gradient descent with a large constant step size η runs at the edge of stability: the largest eigenvalue of the training-loss Hessian hovers near $2/\eta$ (Cohen et al., 2021; Damian et al., 2023). A minimum at which gradient descent is linearly stable must satisfy $\lambda_{\max}(\nabla^2 L_n) \leq 2/\eta$ (Wu et al., 2018). A common belief holds that these minima are what makes large-step-size training resist memorization.

Minima stability constrains the trained function. At a linearly stable minimum a weighted path norm of the network is bounded, where each neuron is charged by how much of the data its kink splits (Wu et al., 2018; Mulayoff et al., 2021; Nacson et al., 2023). Qiao et al. (2024) turned this bound into a generalization guarantee for univariate ReLU networks relative to the ground-truth function: stable minima cannot overfit noisy labels. Liang et al. (2025; 2026) extended the guarantee to multivariate inputs and found that geometry decides the rate: data on a ball of low dimension m generalize, while data on a sphere can be memorized.

A label can be unreliable in two ways: the clean label itself may be uncertain, as when an input can carry different labels, or the observed label may depart from the clean response, as when a sample is mislabeled. The literature studies both under the name label noise. Here the clean response is a function of the input; a small fraction of the responses is shifted by additive outliers of arbitrary size.

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.

The same tools apply to corrupted data as well: the multivariate generalization bound of Liang et al. (2026) implies a floor on the corrupted training loss. It does not give a direct account of the corruption at hand: how far fitting it lowers the training loss of a stable network, and how that relates to recovering the clean signal. We therefore ask: when a small fraction of the training responses is corrupted, does stability keep a network from fitting them by more than a little, and under what conditions does it still recover the clean signal?

No single guarantee can hold for every layout of the data. On a sphere every sample sits on the boundary, so a neuron can fire on one sample alone at almost no cost in the weighted norm, and η-stable networks with bounded outputs can interpolate arbitrary labels (Section 3.4). The guarantee therefore needs a geometric condition. It holds where memorization is expensive: on samples surrounded by enough of the sample that no single kink can cut one of them off cheaply.

Our result concerns two-layer ReLU networks trained with squared loss on n points from a ball of intrinsic dimension m. The clean responses come from the teacher: the true function behind them, not a model trained on clean data. We call the well-surrounded points the deep region and write n_T for their number. Half-space depth makes the region precise (Section 3.1). For uniformly distributed inputs and a small fixed depth threshold, n_T is a constant fraction of n.

Isolating a corrupted point in the deep region means crossing many clean points, and each crossing costs weighted norm. Stability caps this norm, so on the deep region the predictions of every stable network with bounded outputs form a vector in an explicit class of bounded variation. The result compares network and teacher on the deep region.

Informal version of Theorem 5. Train a two-layer ReLU network on corrupted responses from a ball as above, and take an η-stable minimum whose outputs on the training points stay below a fixed bound. Suppose the clean signal satisfies the complexity bound that stability implies. Even the teacher has a training loss on the deep samples, and it comes entirely from the corruption. With high probability, the network's training loss on the deep samples is at least this teacher loss, plus half its squared error against the teacher, minus a remainder. The remainder decays like $n_T^{-(m+3)/(2m+3)}$ up to logarithmic factors, where n_T is the number of deep samples and the corruption rate and amplitude are held fixed.

Two things follow. The training loss has a floor: fitting the outliers can lower it below the teacher's by at most the remainder, so a stable minimum fits them only a little. A training-loss plateau on corrupted data therefore need not signal optimization failure. And when a stable minimum's loss comes down to the teacher's level, it stays close to the clean signal, because it has not overfitted the outliers. Both bounds tighten as the step size grows, since the class of stable minima shrinks (§3.3).

The argument extends in two directions. If the clean signal is too complex for the stability constraint, the network is compared with the best approximation the constraint allows. For classification with sparsely flipped labels under logistic loss the same conclusions hold, with two differences: the benchmark is a classifier that accounts for the flip rate rather than the clean signal itself, and the bound on the outputs becomes essential (§3.5).

Related work. The closest work is the line on minima stability and generalization: Wu et al. (2018), Mulayoff et al. (2021) and Nacson et al. (2023) derive the constraint that stability places on the trained function, and Qiao et al. (2024) and Liang et al. (2025; 2026) turn it into guarantees under label noise; Section 5 compares our result with the latter in detail. Appendix H covers two neighboring literatures: benign overfitting, which asks when interpolating noisy labels is harmless, and noisy-label learning and robust statistics, which build robustness into the loss or the algorithm.

Roadmap. The rest of the paper is organized as follows: Section 2 introduces the problem setup; Section 3 presents the main results for regression with sparse additive contamination; Section 4 presents numerical results; Section 5 discusses related work. Due to space limit, we defer not only technical proofs but also extensions of our main results (regression with a misspecified teacher model and classification with sparse label flips) the appendix.

2 SETUP

Design points $X_1, \dots, X_n$ are i.i.d. from a distribution on the unit ball $\mathbb{B}_1^m \subset \mathbb{R}^m$. Only the intrinsic dimension m enters our bounds; for data in an m-dimensional linear subspace of $\mathbb{R}^d$, every statement applies to the network restricted to that subspace.

We consider two-layer ReLU networks

$$f_\theta(x) = \sum_{k=1}^{K} v_k \max(w_k^\top x - b_k, 0) + \beta, \tag{1}$$

with K neurons, K arbitrary but fixed. The trainable parameters are $\theta := ((w_k, b_k, v_k)_{k=1}^K, \beta)$: input weights $w_k \in \mathbb{R}^m$, offsets b_k, output weights v_k and a bias β. A neuron with $w_k \neq 0$ is active on the half-space $w_k^\top x > b_k$; the boundary hyperplane $w_k^\top x = b_k$ is its *kink*.

2.1 DATA AND SPARSE OUTLIERS

The clean signal is the teacher $f_0 : \mathbb{B}_1^m \to \mathbb{R}$, the true function behind the responses; on the sample it is bounded, $|f_0(X_i)| \leq B_0$. The observed responses are

$$Y_i = f_0(X_i) + A\, B_i S_i, \qquad i \leq n.$$

Here $B_i \in \{0, 1\}$ marks whether response i is corrupted, $S_i \in \{\pm 1\}$ gives the direction of the corruption, and $A > 0$ is its amplitude. Every response therefore satisfies $|Y_i| \leq B_0 + A$.

The pairs (B_i, S_i) are i.i.d.; within a pair, $B_i \sim \text{Bernoulli}(\varepsilon_n)$ with $\varepsilon_n \in (0, \frac{1}{2})$ is independent of S_i, which is uniform on $\{\pm 1\}$. Thus each response is corrupted independently with probability ε_n. The subscript on ε_n is conventional: the corruption rate does not depend on n. The whole family $(B_i, S_i)_{i \leq n}$ is independent of the design, hence of the clean responses.

The training loss is the squared loss without the factor $\frac{1}{2}$,

$$L_n(\theta) := \frac{1}{n} \sum_{i \leq n} (f_\theta(X_i) - Y_i)^2.$$

To isolate the effect of sparse outliers, we do not add a separate background-noise term.

2.2 STABILITY

Definition 1 (Stability under step size η)**.** A parameter θ is η-*stable* if $\lambda_{\max}(\nabla^2 L_n(\theta)) \leq 2/\eta$, where L_n is the full-sample training loss.

This is an upper spectral constraint only; θ need not be a stationary point. The Hessian is taken where no sample lies on a kink, $w_k^\top X_i \neq b_k$ for all i and k; there L_n is twice differentiable.

3 MAIN RESULTS

We now bound how much a stable network with bounded outputs can lower its training loss by fitting the outliers. Section 3.1 fixes where the bound applies and what it compares; Section 3.2 fixes which networks it covers. For lack of space all proofs are deferred to the appendices; Appendix A lists where each statement is proved.

3.1 WHERE THE BOUND APPLIES AND WHAT IT COMPARES

Evaluation region. The bound concerns the points that are well surrounded by the rest of the sample; Section 5 discusses the others. Half-space depth measures how well a point is surrounded: the depth of a point is the smallest share of the sample that a hyperplane through it leaves on one side. Shares are taken under the empirical measure $P_{\mathcal{D}} := n^{-1} \sum_{i \leq n} \delta_{X_i}$ of the design points. For a uniform design on the ball, points near the center are deep and points near the boundary are shallow.

Definition 2 (Empirical half-space depth and deep region; cf. (Liang et al., 2026, Definition 3.1))**.**

$$\text{depth}(x, P_{\mathcal{D}}) := \inf_{u \in \mathbb{S}^{m-1}} P_{\mathcal{D}}(u^\top (X - x) \geq 0).$$

The T-deep region is $\Omega_T := \{x : \text{depth}(x, P_{\mathcal{D}}) \geq T\}$.

The deep samples are $I_T := \{i : X_i \in \Omega_T\}$, their number is $n_T := |I_T|$, and $\text{ev}_T(f) := (f(X_i))_{i \in I_T} \in \mathbb{R}^{n_T}$ collects the values of a function f on them.

Quantities compared. The bound relates three quantities. The deep-region training loss of a vector $v \in \mathbb{R}^{n_T}$ indexed by I_T is $L_T^{\mathrm{sq}}(v) := n_T^{-1} \sum_{i \in I_T} (v_i - Y_i)^2$. The clean error against the teacher is

$$E(v) := \frac{1}{n_T} \sum_{i \in I_T} (v_i - f_0(X_i))^2.$$

The realized corruption rate is $\hat{\varepsilon}_T := n_T^{-1} \sum_{i \in I_T} B_i$. Write $v_0 := \mathrm{ev}_T(f_0)$ for the teacher's vector; its deep-region training loss is $L_T^{\mathrm{sq}}(v_0) = A^2 \hat{\varepsilon}_T$. It comes entirely from the corruption; we call it the *corruption energy*.

3.2 Which networks the bound covers

A kink that reaches the deep region splits a fixed share of the sample and, under the geometric assumption below, has a cost bounded below. Stability caps the total cost of a network's kinks. On the deep region a stable network can therefore afford only a bounded amplitude of such kinks, and its prediction vector lies in an explicit class.

Kink weights. The cost of a kink is its data-dependent weight (Liang et al., 2025; 2026). For a unit direction $u \in \mathbb{S}^{m-1}$ and an offset $t \in \mathbb{R}$, the weight of the oriented hyperplane $\{u^\top x = t\}$ is

$$\widetilde{g}_{\mathcal{D}}(u,t) := \underbrace{P_{\mathcal{D}}(u^\top X > t)^2}_{\text{firing frequency}} \cdot \underbrace{\mathbb{E}_{\mathcal{D}}[u^\top X - t \mid u^\top X > t]}_{\text{average margin}} \cdot \underbrace{\sqrt{1 + \|\mathbb{E}_{\mathcal{D}}[X \mid u^\top X > t]\|^2}}_{\text{centroid norm}},$$

where $\mathbb{E}_{\mathcal{D}}[\cdot]$ averages over the design points; we set $\widetilde{g}_{\mathcal{D}}(u,t) := 0$ when the half-space contains no sample. A hyperplane has two orientations, (u,t) and $(-u,-t)$, and $g_{\mathcal{D}}(u,t) := \min\{\widetilde{g}_{\mathcal{D}}(u,t), \widetilde{g}_{\mathcal{D}}(-u,-t)\}$ is the cheaper of the two. The weighted path norm of the network (1) sums, over its neurons, the amplitude $|v_k|\|w_k\|$ times the weight of the kink:

$$\|f_\theta\|_{\mathrm{path},g} := \sum_k |v_k| \|w_k\| \, g_{\mathcal{D}}(w_k/\|w_k\|, \, b_k/\|w_k\|);$$

neurons with $w_k = 0$ contribute nothing.

Minimum geometric weight. The bound depends on the smallest weight of a kink that reaches the deep region. Each closed side of such a kink holds at least a T-fraction of the sample, so its firing frequency is at least T^2 when no sample lies on the kink. Depth controls only this factor of the weight, not the margin or the centroid norm, so a positive lower bound on the weight of such kinks is an assumption on the design. We write

$$g_{\mathcal{D}}^{\min}(T) := \inf\big\{ g_{\mathcal{D}}(u,t) : \; u \in \mathbb{S}^{m-1}, \, t \in \mathbb{R}, \; \{u^\top x = t\} \cap \Omega_T \neq \emptyset \big\} \tag{2}$$

for the minimum weight of a hyperplane meeting Ω_T; §3.4 verifies $g_{\mathcal{D}}^{\min}(T) > 0$ for the uniform ball.

Assumptions. The bound applies to networks f_θ satisfying four assumptions:

- (A0) no sample lies on a kink (§2.2);
- (A1) η-stability of the full-sample training loss (Definition 1);
- (A2) the *sample envelope* $\max_{i \le n} |f_\theta(X_i)| \le M$: the outputs stay within a bound M fixed before the corruption is seen;
- (A3) $g_{\mathcal{D}}^{\min}(T) > 0$ and $n_T \ge 1$: deep samples exist, and the weights of kinks that reach the deep region are bounded away from zero.

Budget. Under (A0)–(A2) and $0 < \eta \le 1$, which (A1) implies since $\partial^2 L_n / \partial \beta^2 = 2$, the weighted path norm is bounded,

$$\|f_\theta\|_{\mathrm{path},g} \;\le\; \bar{B}_\eta^{\mathrm{sq}} := \tfrac{1}{2}\big(1/\eta - 1 + (R+1)(M + B_0 + A)\big), \tag{3}$$

where $R := \max_i \|X_i\| \le 1$ is the data radius; the budget shrinks as the step size grows. Dividing the budget by the minimum weight gives the effective budget on the deep region, $C_T^{\mathrm{sq}} := \bar{B}_\eta^{\mathrm{sq}} / g_{\mathcal{D}}^{\min}(T)$: it caps the total amplitude of the kinks that reach the deep region.

The certified class. The class below contains the deep-region prediction vectors of all networks satisfying (A0)–(A3). Let $\mathcal{G}(C)$ be the ReLU variation ball of budget C: the two-layer networks $\sum_k c_k(u_k^\top x - t_k)_+$ with unit directions $u_k \in \mathbb{S}^{m-1}$, offsets $t_k \in [-1, 1]$ and $\sum_k |c_k| \le C$. It also contains their infinite-width limits, the integrals of such atoms against a signed measure of total variation at most C (Parhi & Nowak, 2023). Let

$$\mathcal{A}_T := \{(a^\top X_i + b)_{i \in I_T} : a \in \mathbb{R}^m,\, b \in \mathbb{R}\}$$

be the sample vectors of affine functions, of dimension at most $m + 1$. The affine part is left free: Ω_T is convex, so a kink that misses it is affine on it.

Definition 3 (Sample-vector class)**.** For $C, M > 0$ let $V_T(C) := \mathrm{ev}_T(\mathcal{G}(C)) = \{\mathrm{ev}_T(g) : g \in \mathcal{G}(C)\}$ and

$$\mathcal{K}_T(C, M) := \big(V_T(C) + \mathcal{A}_T\big) \cap [-M, M]^{n_T}.$$

The class depends on the design only. The box $[-M, M]^{n_T}$ is the sample envelope (A2) on the deep samples. $\mathcal{K}_T^{\mathrm{sq}} := \mathcal{K}_T(C_T^{\mathrm{sq}}, M)$ is the *certified class*; Lemma 4 justifies the name.

Lemma 4 (Containment)**.** *Under (A0)–(A3),* $\mathrm{ev}_T(f_\theta) \in \mathcal{K}_T(C_T^{\mathrm{sq}}, M)$.

3.3 Main result

All statements below concern vectors $v \in \mathcal{K}_T^{\mathrm{sq}}$; by Lemma 4 they apply to $\hat{v} = \mathrm{ev}_T(\hat{f})$ whenever the trained network $\hat{f}$ satisfies (A0)–(A3); no property of the optimization path is used. The theorem below holds on an event E_δ^{sq} that comes from a concentration inequality proved in Appendix B. Its remainder Δ_T^{sq} is deterministic given the design, the model parameters and the confidence level δ: it does not depend on which responses were corrupted, and (5) below bounds it explicitly.

Theorem 5 (Training-loss floor on the deep region)**.** *Assume (A3),* $0 < \eta \le 1$*, and that the teacher's deep-region vector lies in the certified class,* $v_0 \in \mathcal{K}_T^{\mathrm{sq}}$*; let* $\delta \in (0, 1)$*. On* E_δ^{sq}*, of conditional probability at least* $1 - \delta$ *given the design, simultaneously for all* $v \in \mathcal{K}_T^{\mathrm{sq}}$*,*

$$L_T^{\mathrm{sq}}(v) \;\ge\; A^2\hat{\varepsilon}_T + \tfrac{1}{2}E(v) - \Delta_T^{\mathrm{sq}}, \tag{4}$$

while $L_T^{\mathrm{sq}}(v_0) = A^2\hat{\varepsilon}_T$ *exactly.*

The assumption $v_0 \in \mathcal{K}_T^{\mathrm{sq}}$, the well-specified case, is the condition of Section 1 that the clean signal satisfy the complexity bound that stability implies; it holds if the non-affine part of f_0 has variation at most C_T^{sq} and $|f_0(X_i)| \le M$ for $i \in I_T$.

The floor. Rearranged, (4) reads $A^2\hat{\varepsilon}_T - L_T^{\mathrm{sq}}(v) \le \Delta_T^{\mathrm{sq}} - \frac{1}{2}E(v)$: fitting the outliers lowers the loss below the teacher's by at most the remainder. It bounds the total decrease of the training loss on the deep samples, not the fit of any single outlier. Exact interpolation of the deep samples, where $L_T^{\mathrm{sq}} = 0$ and $E = A^2\hat{\varepsilon}_T$, is excluded once $\Delta_T^{\mathrm{sq}} < \frac{3}{2}A^2\hat{\varepsilon}_T$.

Recovery. If $L_T^{\mathrm{sq}}(\hat{v}) \le A^2\hat{\varepsilon}_T + \varepsilon_{\mathrm{opt}}$ for an optimization tolerance $\varepsilon_{\mathrm{opt}} \ge 0$, then $E(\hat{v}) \le 2\Delta_T^{\mathrm{sq}} + 2\varepsilon_{\mathrm{opt}}$: a network whose loss comes down to the energy has learned the clean signal. A plateau, on the other hand, may come from unfitted corruption or from unfitted clean signal, and only measuring $E(v)$ separates the two (Section 4).

The remainder. With $\alpha_m := \frac{m+3}{2m+3}$, $\beta_m := \frac{2m}{2m+3} = 2 - 2\alpha_m$ and $\gamma_m := \frac{m+3}{2m+2}$, the remainder obeys

$$\Delta_T^{\mathrm{sq}} \;\le\; c_m\,\mathcal{L}_T\Big[(C_T^{\mathrm{sq}})^{\beta_m}\Big(\frac{A^2\varepsilon_n}{n_T}\Big)^{\alpha_m} + A\big(M + C_T^{\mathrm{sq}}\big)n_T^{-\gamma_m} + \big(A^2\varepsilon_n + AM\big)\frac{m + \log(1/\delta)}{n_T}\Big], \tag{5}$$

where c_m depends on m only and $\mathcal{L}_T \ge 1$ is the explicit logarithmic factor of Appendix B.4. The remainder grows with the amplitude A and with the effective budget C_T^{sq}, which carries the step size and the threshold, and it decays in the number of deep samples n_T. The first term measures the size of the certified class. The others decay faster at fixed parameters, since $\gamma_m > \alpha_m$ and $\alpha_m < 1$, but can dominate at finite sample sizes, particularly when the effective budget is large. At confidence level $\delta = n_T^{-c}$ for a fixed $c > 0$ and at fixed parameters, this gives $\Delta_T^{\mathrm{sq}} = \widetilde{O}(n_T^{-\alpha_m})$, with the parameter dependence shown in (5). For $m = 1, 2, 5$ the exponent is $\alpha_m = 4/5,\ 5/7,\ 8/13$. The exponent α_m is set by the metric entropy of the ReLU variation class, whose exponent is $2m/(m + 3)$; α_m decreases towards $1/2$ as m grows. The constant c_m is not calibrated, so the bound is read at the level of exponents. Replacing the realized rate $\hat{\varepsilon}_T$ by ε_n would cost a binomial fluctuation of order $\sqrt{\varepsilon_n/n_T}$, larger than the remainder; hence the floor is stated on the realization.

What the step size buys. The step size enters the theorem through the budget alone. The budget (3) grows like $1/(2\eta)$ as the step size decreases and is a constant of the problem near $\eta = 1$; the class and the remainder inherit it through C_T^{sq}. A larger step size therefore means a smaller certified class, a smaller remainder and a tighter floor. This holds until the budget no longer contains the teacher; then the well-specified assumption fails and the projection benchmark of §3.5 takes over. Linear stability of a gradient-descent minimum at step size η imposes the spectral threshold $2/\eta$ (Wu et al., 2018; Cohen et al., 2021), and the theorem applies once the output, geometric and teacher conditions hold too. A method that reaches sharper minima buys none of this, while a small step size gives a weaker bound, not necessarily a worse network (Section 4). The threshold plays the same role for the geometry: a deeper T raises $g_{\mathcal{D}}^{\min}(T)$ and shrinks the class, at the cost of certifying fewer samples.

3.4 Role of the geometry

Assumption (A3) concerns the data alone, and it cannot be dropped: on the uniform ball it holds with explicit constants; on the sphere it fails, and stability then permits interpolation.

Uniform ball. For a small enough fixed threshold, (A3) holds with high probability:

Proposition 6 (Uniform design)**.** *Let $X_1, \dots, X_n$ be i.i.d. uniform on $\mathbb{B}_1^m$, let $T \in (0, \frac{1}{2}]$, $\delta \in (0,1)$, and suppose $n \geq A_m\,(m + \log(1/\delta))\,T^{-2\omega_m}$ with $\omega_m := 2 + \frac{2}{m+1}$. With probability at least $1 - 2\delta$ over the design,*

$$g_{\mathcal{D}}^{\min}(T) \;\geq\; c'_m\, T^{\omega_m} \qquad \text{and} \qquad \frac{n_T}{n} \;\geq\; 1 - C_m T^{2/(m+1)} - \sqrt{\frac{2\log(1/\delta)}{n}} - \frac{\log(1/\delta)}{3n},$$

where $A_m, c'_m, C_m > 0$ depend only on m (explicit values in Appendix D).

The bounds are informative for small T: once $C_m T^{2/(m+1)} \leq \frac{1}{4}$ and the two fluctuation terms sum to at most $\frac{1}{4}$, at least half the sample is deep; the weight bound is positive at every T.

Sphere. On the sphere every sample has empirical depth $1/n$, so for $T > 1/n$ no sample is deep and the bound does not apply; the spectral constraint alone still permits interpolating any labels:

Proposition 7 (Sphere, squared loss: stable exact interpolation; adapted from the flat-interpolation construction of Liang et al. (2026))**.** *Let $n \geq 2$, let $X_1, \dots, X_n \in \mathbb{S}^{d-1}$ be pairwise distinct, and let $|y_i| \leq D$ with $D > 0$. For every fixed $\eta \in (0,1)$ and every $n \geq 2\sqrt{2}\,D\eta/(1-\eta)$ there is a width-n network with output envelope D that interpolates the labels, so that its squared training loss and gradient vanish, and that is η-stable:*

$$\lambda_{\max}(\nabla^2 L_n^{\mathrm{sq}}) \leq 2 + 4\sqrt{2}D/n \leq 2/\eta.$$

Consequently no absolute squared-loss floor can hold on the sphere at any fixed $\eta < 1$.

Here $\eta < 1$ is the whole range the spectral condition leaves: it forces $\eta \leq 1$, and $\eta = 1$ excludes every non-constant interpolant (Remark E.4). The network places each sample in its own small spherical cap, whose kink pays almost nothing in the weighted budget (3). The proposition is a statement about η-stable parameters, not about what gradient descent finds; Section 4 returns to this distinction.

3.5 Extensions

Two extensions, proved in the appendices, rest on the same concentration inequality.

A teacher outside the class. When the budget cannot represent the clean signal, the benchmark is the best approximation the class allows: the teacher's projection onto the certified class.

> **Informal version of Theorem C.5.** Keep the assumptions of Theorem 5 except that the teacher may lie outside the class. With high probability given the design, every vector in the certified class has a deep-region training loss at least that of the projection, plus half the amount by which its clean error exceeds the projection's, minus the remainder of Theorem 5. A vector whose loss is within an optimization tolerance of the projection's therefore has clean error at most the projection's, plus twice the remainder and twice the tolerance (Corollary C.6).

This bound keeps three sources of error separate: what the class cannot represent, what the realized corruption costs, and what optimization leaves on the table.

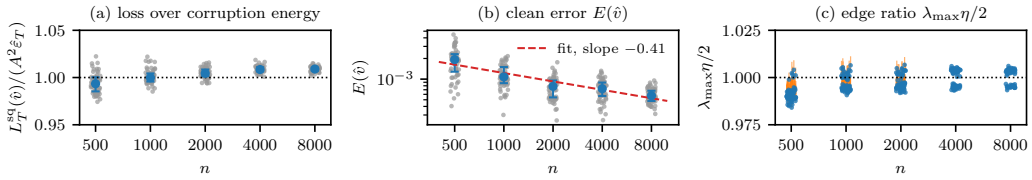


Figure 1: Trained endpoints under squared loss ($A = 1$, $T = 0.1$; 60 masks per n; dots are endpoints, blue markers medians with interquartile bars). (a) Deep-region training loss over the realized corruption energy; the dotted line at 1 is the teacher's value, the floor of Theorem 5 up to the remainder. (b) Clean error $E(\hat{v})$ against n; the dashed line is an empirical least-squares fit through the medians (slope -0.41). (c) Edge ratio $\lambda_{\max}\eta/2$: training-time power-iteration estimate (dots) and, for $n \leq 2000$, the verified upper bound (bar ends).

Logistic loss with sparse label flips. For classification the clean labels are themselves random, and the teacher's role passes to the mean-flip benchmark (Definition F.3). It minimizes the logistic loss over the class against the flip-averaged labels, the best one can do knowing the clean labels and the flip rate but not the flip locations. Bounded logits give the logistic loss a uniform curvature bound: on $[-M, M]$ its curvature is at least $\kappa_M := \sigma(M)(1 - \sigma(M))$, with σ the logistic function. This is where the output bound (A2) becomes essential.

Informal version of Theorem F.6. Condition on the inputs and the clean labels, and flip each label independently with probability ε_n. Under (A3), with high probability, every vector in the certified class built from the logistic budget of Appendix F has a deep-region logistic loss on the observed labels at least the benchmark's, plus $\kappa_M/4$ times the mean squared difference between its predictions and the benchmark's on the deep samples, minus a remainder that decays like $n_T^{-\alpha_m}$ up to logarithmic factors at fixed model and class parameters. Hence no such vector fits the flipped labels better than the benchmark by more than the remainder (Corollary F.7). A vector whose loss is within an optimization tolerance of the benchmark's has this mean squared difference at most $4/\kappa_M$ times the sum of the remainder and the tolerance (Corollary F.8).

The sphere again admits stable networks that fit the flipped labels within the output bound (Proposition F.9). The trained logistic endpoints can be checked against the bound directly. Their deep-region prediction vectors lie in the certified class for 598 of the 600 networks, even where the spectral condition fails, and that is all the bound requires. On every such record their deep-region loss on the flipped labels stays above the benchmark's, and the complete floor, quadratic term included, holds on the realization (Appendix G.8).

4 Experiments

The experiments measure the objects of Theorem 5 on a certified pipeline (Appendix G). The design is one-dimensional, so depth is a rank count and I_T, n_T and the deep-region losses involve no approximation. The inputs are embedded in $\mathbb{R}^{50}$ by zero-padding. The realized remainder Δ_{emp} is the smallest constant for which the concentration inequality of Appendix B holds on the class at the observed corruption. A convex program encloses it in a certified interval, and we report the upper end $\bar{\Delta}$. For every training endpoint we record its membership $\hat{v} \in \mathcal{K}_T^{\mathrm{sq}}$, the envelope (A2) and a three-state spectral status: certified by a verified upper bound on $\lambda_{\max}$, witnessed violation by a verified lower bound, or undetermined. Networks of width 1024 are trained by full-batch gradient descent for 20,000 epochs at $\eta = 0.2$ with $M = 2$ fixed before the corruption is drawn; the endpoint is the final parameter. The grid uses one design and 60 corruption masks per configuration, with $n \in \{500, 1000, 2000, 4000, 8000\}$ at $T = 0.1$ and $n = 2000$ at $T \in \{0.05, 0.2, 0.3\}$. The $n = 2000$ networks are re-evaluated at each T, so one network yields several evaluation records. Unless stated otherwise, $\varepsilon_n = 0.05$ and $A = 1$.

Assumptions at the endpoints. Every squared endpoint lies in the certified class $\mathcal{K}_T^{\mathrm{sq}}$ (600/600 networks), which is all that Theorem 5 requires of it, and satisfies (A2) with margin ($\max_i |\hat{f}(X_i)| \leq 1.1$). The endpoints hover at the edge of stability: of the 360 networks with a verified upper bound ($n \leq 2000$, both amplitudes), 145 are certified, 117 exceed the threshold (witnessed by the verified lower bound), and 98 lie between the two bounds. Figure 1(c) shows this per endpoint: every power-iteration estimate, and every verified upper bound, lies within two percent of the edge.

Table 1: Squared loss, trained endpoints ($A = 1$, $T = 0.1$; medians over 60 masks). Deficit $= A^2\hat{\varepsilon}_T + \frac{1}{2}E - L$; $\bar{\Delta}$ is the upper end of the certified remainder interval. Spectral counts per distinct network: certified / witnessed violation / undetermined (verified bound for $n \le 2000$).

n	$L/(A^2\hat{\varepsilon}_T)$	$E(\hat{v})$	deficit	$\bar{\Delta}/(A^2\hat{\varepsilon}_T)$	spectral	in $\mathcal{K}_T^{\mathrm{sq}}$
500	0.993	1.9×10^{-3}	$+1.1 \times 10^{-3}$	1.13	50/5/5	60/60
1000	1.000	1.1×10^{-3}	$+5.0 \times 10^{-4}$	0.66	19/17/24	60/60
2000	1.004	7.8×10^{-4}	$+1.8 \times 10^{-4}$	0.38	15/17/28	60/60
4000	1.008	7.3×10^{-4}	-1.0×10^{-4}	0.21	—	60/60
8000	1.009	5.7×10^{-4}	-1.4×10^{-4}	0.12	—	60/60

Loss and clean error. The training loss sits on the corruption energy, and the clean error $E(\hat{v})$ is one to two orders of magnitude below it (Table 1, Figure 1(a)). So the plateau reflects the corruption, not a failure to fit the clean signal. The clean error decreases with n (Figure 1(b)). At $n = 2000$ its medians are 7.8, 7.8, 5.9 and 4.6×10^{-4} for $T = 0.05, 0.1, 0.2, 0.3$, so it decreases with T above 0.1. With $A = 0.5$ the same holds (per-n medians: loss over energy 1.02–1.04, E from 4 to 8×10^{-4}).

Certificate versus observation. The deficit $A^2\hat{\varepsilon}_T + \frac{1}{2}E(\hat{v}) - L_T^{\mathrm{sq}}(\hat{v})$ stays below the certified remainder $\bar{\Delta}$ on all 960 evaluation records, as the quadratic identity (Lemma C.2) and the definition of Δ_{emp} guarantee once $\hat{v}$ and v_0 lie in the class. The check confirms the inequality on the realization; it does not verify the high-probability rate. The certificate is looser than the observation: the relative remainder $\bar{\Delta}/(A^2\hat{\varepsilon}_T)$ falls from 1.13 at $n = 500$ to 0.12 at $n = 8000$ (medians, Table 1), so at small n the certified floor lies well below the loss actually observed.

Beyond the deep region. The certificate is local but the behavior is not. On the shallow complement of the deep region, the same endpoints have a loss between 1.02 and 1.04 times the local corruption energy. Their clean error is roughly two to four times the deep-region value, still one to two orders of magnitude below the energy ($A = 1$, per-configuration medians, Appendix G.5).

Ambient dimension. Ambient dimension does not affect the floor (Appendix G.6). Changing the ambient dimension of the zero-padded design from 50 to 2 or 500 leaves the loss on the energy and every endpoint in the class. The clean error and the edge ratio change with d. This is an optimization effect: the dimension enters training only through the initialization scale $\sqrt{2/d}$ of the input weight on the one nonzero coordinate of the design.

Step size. A larger step size tightens the bound, but in our runs the fit does not improve beyond $\eta = 0.1$ (Appendix G.6). Across $\eta \in \{0.05, 0.1, 0.2, 0.4\}$ (medians, 20 masks) the loss sits on the energy up to $\eta = 0.2$, and the small-step endpoints end far below the edge with the smallest clean error. At $\eta = 0.4$ the endpoints underfit although the worst-case budget still exceeds the teacher's variation a thousandfold: they are smoother than the teacher, with a loss 1.4 times the energy, above the floor as required. The budget does not explain this; the cause would need a tighter spectral description or an analysis of the training dynamics.

Table 2: Matched geometry, $A = 1$ (six seeds, $n \in \{2000, 4000\}$; medians over twelve final endpoints; full-sample quantities, as the circle has no deep region). The cap row lists the analytic values of the Appendix E construction on the same designs (implementation residuals: Appendix G.7). Edge ratio $\lambda_{\max}\eta/2$: Lanczos lower bound for GD, analytic Gram value for the cap.

network	$L/(A^2\hat{\varepsilon})$	E	edge ratio	$\max_i \lvert f(X_i)\rvert$
disk, GD	0.99	1.7×10^{-3}	1.00	≤ 1.12
circle, GD	0.99	6×10^{-4}	0.47	≤ 1.03
circle, analytic cap	0	$A^2\hat{\varepsilon}$	0.20	$\le D$

Geometry: what stability permits and what gradient descent reaches. On the circle stability permits exact interpolation, and gradient descent does not find it. To separate the geometric assumption from the ambient dimension we compare a disk and a circle carrying the same angular design (inputs in $\mathbb{R}^2$ embedded in $\mathbb{R}^{50}$, teacher $(x_1 + x_2)/\sqrt{2}$, $M = 2$). All disk endpoints and ten of the twelve circle endpoints sit on the corruption energy (Table 2). The two seed-4 circle runs collapse

to constant predictors at the first step and are kept in the medians. On the circle the endpoints end well inside the stable region. The circle nonetheless has the capacity that Proposition 7 asserts: the analytic cap construction on the same designs interpolates exactly with $\lambda_{\max} = 2.003$ at $n = 2000$ (edge ratio 0.20) and clean error equal to the corruption energy. Its weights are of size 10^{13}–10^{14}, because the caps around a random design are that small. In the non-collapsed circle runs, gradient descent from a small initialization at width 1024 returned losses near the corruption energy instead.

5 Discussion

The assumptions in practice. In our experiments the trained endpoints satisfy the envelope (A2) with margin (Section 4), and so do those of the logistic extension except eight at $n = 500$ (Appendix G.8). The envelope is an additional assumption rather than a consequence of stability, and identifying training conditions that ensure it is the open question. By itself the envelope produces no floor: Proposition 7 is an η-stable interpolant with envelope D, and the floor needs stability and geometry together. The budget is worst-case over all η-stable parameters and grows like $1/\eta$. A certificate built from the realized $\lambda_{\max}$ rather than from $2/\eta$ would be tighter for networks that end well inside the stable region, given uniform control over the data-dependent choice of budget.

Large step sizes as implicit robustness. Training with the largest step size that stability permits is a familiar practice. The theorem shows what it buys. The step size fixes the budget, and the budget fixes the class. Wherever kinks are expensive, the class limits how far fitting sparse corruption can lower the training loss, with no robust loss and no detector. A budget of the same kind underlies the generalization bounds of Liang et al. (2025), under their own assumptions. A larger step size tightens the bound, yet in our runs the small step sizes gave the smallest clean error (Section 4).

Why the bound is local. The floor rests on $g_{\mathcal{D}}^{\min}(T) > 0$: every kink that reaches Ω_T pays at least that weight in the budget. Near the boundary this degenerates. A kink that cuts off a one-sample cap has a firing-frequency factor of order $1/n^2$, and its average-margin factor can shrink further. Its weighted cost therefore becomes arbitrarily small while its amplitude does not. This degeneration is what makes stable memorization possible on the sphere (Proposition 7 and Appendix E). In our runs the shallow complement nevertheless behaves like the deep region (Appendix G.5). On its own the bound says only that the full-sample loss is at least (n_T/n) times the deep-region loss. Full-sample refinements are out of scope here.

Scope and extensions. Uniformity of the design enters only the explicit geometric estimates of Appendix D. Beyond lying in the unit ball, the design must satisfy only (A3): deep samples exist and the minimum geometric weight is positive. The second part fails when a kink through the deep region can have vanishing weight, because depth controls only the firing-frequency factor. Examples are shell concentration, lower-dimensional support and vanishing margins. Here m is the exact dimension of the ball. For varying intrinsic dimension a componentwise or local-chart extension is natural, and the mixture analysis of Liang et al. (2026) supplies the geometry. What remains is to couple componentwise benchmarks to one global budget. The bridge from stability to the budget is specific to two-layer ReLU networks; deeper networks need a new bridge. Finally, the mechanism needs positive depth in the space where memorization happens: high-dimensional inputs often concentrate near a shell, so the relevant depth is that of learned representations.

Relation to Liang et al. (2025; 2026). The objects differ: they bound a generalization gap under bounded labels, a global bound on the network's outputs, a weighted-variation budget and an i.i.d. uniform design. Under these assumptions their bound implies, with high probability, an absolute floor $A^2\varepsilon_n - \widetilde{O}(n^{-1/(2m+4)})$ on the full-sample training loss at fixed parameters. We bound how far below the teacher's loss the deep-region training loss of a stable network can fall, on the given corruption and with the outputs bounded at the sample points only. Our bound is conditional on the clean sample; only Proposition 6 adds a design event. We borrow their geometric scaffold: depth, the affine reduction of a kink on the deep region and the sphere interpolant. New are the quantity we bound, the class of deep-region prediction vectors with its concentration inequality, and the identity that turns that inequality into a floor.

AI USE STATEMENT

The authors conceived and led this work. We chose the research question, identified the literature on minima stability that it builds on, and decided the formulation, the results and the structure of the paper. Under our direction, generative AI tools helped refine the formulation by predicting outcomes of candidate formulations and testing them in exploratory simulations. GPT (OpenAI) assisted in the theory, by formulating intermediate claims, supplying proof steps and drafting proofs, which we reworked into their present form. Claude (Anthropic) assisted in designing the experiments with us, implementing and running them, and interpreting their results, as well as in drafting the text and figures and searching the literature. We verified the results. The proofs were also audited by an expert in learning theory, and every certified quantity carries rigorous interval-arithmetic bounds. We take full responsibility for the content of this work.

REPRODUCIBILITY STATEMENT

All theorems are stated with their assumptions in Sections 2 and 3 and proved in full in Appendices B–F; Appendix A maps every claim to its proof. Appendix G specifies the experimental pipeline completely: design, teachers, corruption models, training protocol, the convex programs with their certified intervals, the verified spectral bound, and the explicit cap constructions with all constants. The supplementary material contains the pipeline code, the trained networks and inputs from which every reported certificate can be regenerated, and the full error model of both verifiers. All code used in this work has been archived and will be made publicly available.

REFERENCES

D. Arpit, S. Jastrzębski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, and S. Lacoste-Julien. A closer look at memorization in deep networks. In *ICML*, 2017.

P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *PNAS*, 117(48):30063–30070, 2020.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

J. M. Cohen, S. Kaur, Y. Li, J. Z. Kolter, and A. Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *ICLR*, 2021.

A. Damian, T. Ma, and J. D. Lee. Label noise SGD provably prefers flat global minimizers. In *NeurIPS*, 2021.

A. Damian, E. Nichani, and J. D. Lee. Self-stabilization: the implicit bias of gradient descent at the edge of stability. In *ICLR*, 2023.

S. Dirksen. Tail bounds via generic chaining. *Electronic Journal of Probability*, 20:1–29, 2015.

S. Frei, N. S. Chatterji, and P. L. Bartlett. Benign overfitting without linearity: neural network classifiers trained by gradient descent for noisy linear data. In *COLT*, 2022.

P. J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101, 1964.

M. Li, M. Soltanolkotabi, and S. Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *AISTATS*, 2020.

T. Liang, D. Qiao, Y.-X. Wang, and R. Parhi. Stable minima of ReLU neural networks suffer from the curse of dimensionality: The neural shattering phenomenon. In *NeurIPS*, 2025.

T. Liang, A. Cloninger, R. Parhi, and Y.-X. Wang. Generalization below the edge of stability: the role of data geometry. In *ICLR*, 2026.

C. Ma and L. Ying. On linear stability of SGD and input-smoothness of neural networks. *arXiv preprint arXiv:2105.13462*, 2021.

R. Mulayoff, T. Michaeli, and D. Soudry. The implicit bias of minima stability: A view from function space. In *NeurIPS*, 2021.

M. S. Nacson, R. Mulayoff, G. Ongie, T. Michaeli, and D. Soudry. The implicit bias of minima stability in multivariate shallow ReLU networks. In *ICLR*, 2023.

N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari. Learning with noisy labels. In *NeurIPS*, 2013.

C. Northcutt, L. Jiang, and I. Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021.

R. Parhi and R. D. Nowak. Near-minimax optimal estimation with shallow ReLU neural networks. *IEEE Trans. Inform. Theory*, 69(2):1125–1140, 2023.

G. Patrini, A. Rozza, A. K. Menon, R. Nock, and L. Qu. Making deep neural networks robust to label noise: a loss correction approach. In *CVPR*, 2017.

G. Pleiss, T. Zhang, E. Elenberg, and K. Q. Weinberger. Identifying mislabeled data using the area under the margin ranking. In *NeurIPS*, 2020.

D. Qiao and Y.-X. Wang. Does flatness imply generalization for logistic loss in univariate two-layer ReLU network? *arXiv preprint arXiv:2512.01473*, 2025.

D. Qiao, K. Zhang, E. Singh, D. Soudry, and Y.-X. Wang. Stable minima cannot overfit in univariate ReLU networks. In *NeurIPS*, 2024.

J. W. Siegel. Optimal approximation of zonoids and uniform approximation by shallow neural networks. *arXiv preprint arXiv:2307.15285*, 2023.

L. Wu, C. Ma, and W. E. How SGD selects the global minima in over-parameterized learning: a dynamical stability perspective. In *NeurIPS*, 2018.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017.

A How to Read the Proofs

The appendices follow the order of Section 3. Appendix B proves the two results on which the main theorem and both extensions rest. The first is containment: stable networks land in the sample-vector class (Lemma 4). The second is the offset inequality $\tau|Z(h)| \le \frac{\kappa}{4}\|h\|_{n_T}^2 + \Delta$ for the multiplier process $Z(h) := n_T^{-1}\sum_{i\in I_T}\zeta_i h_i$, where the ζ_i are the corruption multipliers (Assumption B.3) and $\tau, \kappa > 0$ are parameters that each loss sets. It holds simultaneously for all h in the difference class $\mathcal{H}_T := \mathcal{K}_T - \mathcal{K}_T$ (Definition B.1), on an event E_δ of conditional probability at least $1-\delta$. This is Theorem B.10, the concentration inequality of Section 3. Its proof combines a metric-entropy bound for $\mathcal{H}_T$ with generic chaining under Bernstein increments, and Proposition B.11 makes the remainder Δ explicit. Everything after the offset inequality is elementary. Appendix C proves Theorem 5 from the offset inequality at $(\tau, \kappa) = (2A, 2)$ through the quadratic identity, and then treats a teacher outside the class (Theorem C.5). Appendix D verifies the geometric assumption on the uniform ball (Proposition 6); Appendix E gives the sphere construction (Proposition 7). Appendix F is the logistic extension: it uses the offset inequality at $(1, \kappa_M)$ with the curvature of the mean-flip objective and ends with its own sphere construction. Appendix G details the experiments and their certification, and Appendix H covers the extended related work.

Claim	**Read in this order**	**Skip on first pass**
Lemma 4	Prop. C.1 (budget) → Lemma B.2 → the containment proof in §B.1	the value of the constant in Prop. C.1
Theorem B.10	§B.2 (entropy of $\mathcal{H}_T$) → Prop. B.9 (localized supremum) → the proof in §B.3	Proposition B.11 (explicit remainder)
Theorem 5	Lemma C.2 → Theorem B.10 at $(2A, 2)$ → the direct proof in §C.2	§C.4 (explicit remainder)
Theorem C.5	Lemma C.4 (projection benchmark) → the three-step proof in §C.3, then Corollary C.6	Lemma C.3 (well-specified case)
Proposition 6	Lemmas D.1–D.4 (Appendix D)	the constants
Proposition 7	Lemma E.1 → Lemma E.2 → Lemma E.3 → the proof in Appendix E	the endpoint remark
Theorem F.6	Lemma F.4 → Lemma F.5 → the four-step proof in §F.4, then Corollaries F.7–F.8, with Lemmas F.1–F.2 for the trained network	§F.5 (explicit remainder)
Proposition F.9	Lemma E.2 → Lemma E.3 (Appendix E) → the four-step proof in §F.6	the comparison after the proof

First-pass recommendation. Read the statement of Theorem B.10 and the translation table in §B.4, which instantiates the offset inequality for each loss. Then read the proof in §C.2; it is under a page once the offset inequality is taken as a black box. The extension to a teacher outside the class (§C.3), the entropy and chaining section (§B.2) and the explicit remainder (Proposition B.11) can wait.

Notation at a glance.

Geometry / deep region	
Ω_T, I_T, n_T	deep region, its index set, its size
$g_{\mathcal{D}}(u,t)$, $g_{\mathcal{D}}^{\min}(T)$	kink weight; minimum weight over kinks reaching Ω_T
$\bar{B}_\eta$, $C_T = \bar{B}_\eta / g_{\mathcal{D}}^{\min}(T)$	stability budget; effective budget
Class and events	
$\mathcal{K}_T = \mathcal{K}_T(C_T, M)$, $\mathcal{H}_T = \mathcal{K}_T - \mathcal{K}_T$	sample-vector class; difference class
$\|v\|_{n_T} = \|v\|_N$	empirical norm on the deep sample, $\|v\|_N^2 = N^{-1}\sum_{i\in I_T} v_i^2$
ζ_i	corruption multiplier of sample i (Assumption B.3)
$Z(h) = n_T^{-1}\sum_{i\in I_T}\zeta_i h_i$	multiplier process (also written Z_T)
E_δ, E_δ^{sq}	offset events (Theorem B.10 at $(1,\kappa_M)$ and $(2A,2)$)
$\mathcal{L}_T$	explicit logarithmic factor of the remainders
Squared loss and uniform ball	
v_0, $v^\dagger$, a_T, $E(v)$	teacher vector, its projection, approximation error, clean error
A, $\hat{\varepsilon}_T$, Δ_T^{sq}	amplitude, realized rate, remainder
$p = \frac{2m}{m+3}$; $\alpha_m = \frac{2}{2+p}$, β_m, γ_m	entropy exponent; remainder exponents
$\omega_m = 2 + \frac{2}{m+1}$, $\nu_m = \frac{2}{m+1}$	uniform-ball exponents
Logistic	
v^*	mean-flip benchmark (Appendix F)
κ_M, $\Delta_T^{\log}$, $\rho_T^2 = \Delta_T^{\log}/\kappa_M$	curvature, remainder, squared radius

B Shared Geometry-Side Proofs: the Sample-Vector Class and the Sparse Offset Inequality

This appendix proves the two statements on which the main theorem and both extensions rest. The containment Lemma 4 places the deep-region values of a stable network in the class $\mathcal{K}_T(C_T, M)$ of Definition 3. The offset inequality (10) bounds $\tau|Z(h)|$ on the difference class $\mathcal{H}_T$ by $\frac{\kappa}{4}\|h\|_N^2$ plus a remainder Δ. §B.1 records the structure of the classes, proves containment and states the assumption on the multipliers. §B.2 bounds the sup-norm entropy of $\mathcal{H}_T$ and turns Bernstein increments into a localized supremum bound by generic chaining. §B.3 proves the offset inequality and makes the remainder explicit. §B.4 reads the rates and records how the two losses instantiate the result.

Notation for the appendices. $N := n_T \geq 1$ and, for $v \in \mathbb{R}^N$ indexed by I_T, $\|v\|_{n_T}^2 = \|v\|_N^2 := N^{-1}\sum_{i\in I_T} v_i^2$, $\|v\|_\infty := \max_{i\in I_T}|v_i|$, so that $\|v\|_N \leq \|v\|_\infty \leq \sqrt{N}\|v\|_N$. For a function g on $\mathbb{B}_1^m$, $\|g\|_{\infty,\mathbb{B}} := \sup_{x\in\mathbb{B}_1^m}|g(x)|$. $\mathcal{F}_0$ is the σ-field generated by the design (and, in the logistic extension, by the clean labels); all probabilities in this appendix are conditional on $\mathcal{F}_0$. $p := 2m/(m+3) \in [1/2, 2)$ is the entropy exponent, and $\alpha_m, \beta_m, \gamma_m$ are as in Section 3: $\alpha_m = 2/(2+p)$, $2\alpha_m + \beta_m = 2$, $p\alpha_m = \beta_m$, and $\gamma_m = (m+3)/(2m+2)$ satisfies $p\gamma_m = m/(m+1) < 1$.

Scope conventions. η-stability is only the upper spectral constraint of Definition 1, taken at a parameter where no sample lies on a kink; stationarity is not assumed and negative eigenvalues are not excluded. At a twice-differentiable critical point, linear stability of the gradient-descent fixed point with step size η additionally requires $0 \preceq \nabla^2 L_n \preceq (2/\eta)I$. Independence of the corruption from the clean sample is needed jointly, not pointwise, because the index set I_T depends on all design points at once.

B.1 The classes and the containment lemma

Definition B.1 (Atoms, classes and the difference class)**.** Let $\Theta := \mathbb{S}^{m-1} \times [-1,1]$ and, for $\theta = (u,t) \in \Theta$, $\varphi_\theta(x) := (u^\top x - t)_+$. For $C \geq 0$ let

$$\mathcal{G}(C) := \Big\{ g_\mu(x) = \int_\Theta \varphi_\theta(x)\, d\mu(\theta) \;:\; \mu \text{ a finite signed Borel measure on } \Theta,\ \|\mu\|_{\mathrm{TV}} \leq C \Big\};$$

finite combinations $\sum_k c_k \varphi_{\theta_k}$ with $\sum_k |c_k| \leq C$ are the atomic case. As in Definition 3, $V_T(C) := \mathrm{ev}_T(\mathcal{G}(C))$ and $\mathcal{K}_T(C,M) := (V_T(C) + \mathcal{A}_T) \cap [-M,M]^N$, where $\mathcal{A}_T$ is the space of affine sample vectors and $q := \dim \mathcal{A}_T \leq m+1$; the affine part is never given a budget. For fixed $C > 0$ and

$M > 0$ the difference class and its constants are

$$\mathcal{H}_T := \mathcal{K}_T(C, M) - \mathcal{K}_T(C, M), \qquad B := 2M, \qquad C_H := 2C, \qquad S := B + 2C_H = 2M + 4C.$$

Lemma B.2 (Structure of the classes). *(i) For $g \in \mathcal{G}(C)$: $\|g\|_{\infty,\mathbb{B}} \le 2C$ and $\mathrm{Lip}(g) \le C$ on $\mathbb{B}_1^m$.*

(ii) $V_T(C) = C \cdot \mathrm{absconv}\{\mathrm{ev}_T(\varphi_\theta) : \theta \in \Theta\}$; it is compact, convex, symmetric and contained in $[-2C, 2C]^N$. Hence $\mathcal{K}_T(C, M)$ is compact, convex, symmetric and contains 0.

(iii) $\mathcal{H}_T = \mathcal{K}_T(2C, 2M) = (V_T(C_H) + \mathcal{A}_T) \cap [-B, B]^N$ is compact, convex, symmetric and contains 0; in particular it is star-shaped: $h \in \mathcal{H}_T$, $\lambda \in [0, 1]$ imply $\lambda h \in \mathcal{H}_T$. Every $h \in \mathcal{H}_T$ can be written $h = g + a$ with $g \in V_T(C_H)$, $\|g\|_\infty \le 2C_H$, $a \in \mathcal{A}_T$ and $\|a\|_\infty \le S$.

Proof. (i) For $x \in \mathbb{B}_1^m$ and $\theta = (u, t)$, $0 \le \varphi_\theta(x) \le |u^\top x - t| \le 2$, and $|\varphi_\theta(x) - \varphi_\theta(x')| \le |u^\top (x - x')| \le \|x - x'\|$; integrate against $|\mu|$.

(ii) The map $\theta \mapsto \mathrm{ev}_T(\varphi_\theta)$ is continuous on the compact set Θ, so $A := \{\mathrm{ev}_T(\varphi_\theta)\}$ is compact; in $\mathbb{R}^N$ the absolutely convex hull of a compact set is compact (Carathéodory), so no closure is needed. Let $\|\mu\|_{\mathrm{TV}} \le 1$ and $\mu = \mu^+ - \mu^-$. If $\mu^\pm(\Theta) > 0$, the vector $b^\pm := \mu^\pm(\Theta)^{-1} \int \mathrm{ev}_T(\varphi_\theta)\, d\mu^\pm$ is the barycenter of a probability measure on the compact set A, hence lies in $\mathrm{conv}(A)$; thus $\mathrm{ev}_T(g_\mu) = \mu^+(\Theta) b^+ - \mu^-(\Theta) b^- \in \mathrm{absconv}(A)$ because $\mu^+(\Theta) + \mu^-(\Theta) \le 1$. Conversely every element $\sum_k \lambda_k (\pm \mathrm{ev}_T(\varphi_{\theta_k}))$ of $\mathrm{absconv}(A)$, $\sum |\lambda_k| \le 1$, equals $\mathrm{ev}_T(g_\mu)$ for the atomic measure $\mu = \sum_k \pm\lambda_k \delta_{\theta_k}$. Scaling by C gives the identity; the box bound is (i). The sum of the compact set $V_T(C)$ and the closed subspace $\mathcal{A}_T$ is closed; convexity and symmetry are inherited; intersecting with the compact box gives compactness; $0 \in V_T(C) \cap \mathcal{A}_T$ lies in the box.

(iii) For a convex symmetric set $\mathcal{K}$, $v - v' = 2 \cdot \frac{1}{2}(v + (-v')) \in 2\mathcal{K}$ and $2\mathcal{K} \subset \mathcal{K} - \mathcal{K}$, so $\mathcal{K} - \mathcal{K} = 2\mathcal{K}$. All three constraints defining $\mathcal{K}_T$ are homogeneous, so $2\mathcal{K}_T(C, M) = \mathcal{K}_T(2C, 2M)$, and (ii) gives the structural properties. For the decomposition, $\|a\|_\infty \le \|h\|_\infty + \|g\|_\infty \le B + 2C_H$. □

Proof of Lemma 4 (containment). Step 0 (budget). Under (A0)–(A2) the weighted path norm satisfies $\|f_\theta\|_{\mathrm{path},g} \le \bar{B}_\eta^{\mathrm{sq}}$, the budget (3) (Proposition C.1); the proof uses the envelope only through the values $f_\theta(X_i)$, $i \le n$.

Step 1 (deep-region reduction). Let N_T be the set of neurons whose activation hyperplane $\{w_k^\top x = b_k\}$ meets Ω_T. A neuron not in N_T is either identically active or identically inactive on the convex set Ω_T, hence affine there; neurons with $w_k = 0$ are constants. So on Ω_T, $f_\theta = g_T + \ell_T$ with $g_T = \sum_{k \in N_T} v_k (w_k^\top x - b_k)_+$ and ℓ_T affine. For $k \in N_T$ the definition (2) gives $g_{\mathcal{D}}(w_k/\|w_k\|, b_k/\|w_k\|) \ge g_{\mathcal{D}}^{\min}(T)$, so the budget yields $\sum_{k \in N_T} |v_k| \|w_k\| \le \bar{B}_\eta^{\mathrm{sq}} / g_{\mathcal{D}}^{\min}(T) = C_T^{\mathrm{sq}}$.

Step 2 (normalization of the crossing neurons). For $k \in N_T$ put $\hat{u}_k := w_k/\|w_k\|$, $\hat{b}_k := b_k/\|w_k\|$ and $c_k := v_k \|w_k\|$, so that $v_k (w_k^\top x - b_k)_+ = c_k \varphi_{(\hat{u}_k, \hat{b}_k)}(x)$. Since the hyperplane $\{\hat{u}_k^\top x = \hat{b}_k\}$ meets $\Omega_T \subset \mathbb{B}_1^m$, there is x_0 with $\|x_0\| \le 1$ and $\hat{b}_k = \hat{u}_k^\top x_0$, whence $|\hat{b}_k| \le 1$ and $(\hat{u}_k, \hat{b}_k) \in \Theta$. Thus $g_T \in \mathcal{G}(C_T^{\mathrm{sq}})$ with the atomic measure $\sum_k c_k \delta_{(\hat{u}_k, \hat{b}_k)}$.

Step 3 (assembly). $\mathrm{ev}_T(g_T) \in V_T(C_T^{\mathrm{sq}})$; ℓ_T coincides on $\Omega_T \ni X_i$ ($i \in I_T$) with an affine function of x, so $\mathrm{ev}_T(\ell_T) \in \mathcal{A}_T$; and (A2) places $\mathrm{ev}_T(f_\theta)$ in the box. Hence $\mathrm{ev}_T(f_\theta) \in \mathcal{K}_T(C_T^{\mathrm{sq}}, M)$. □

From here on the corruption enters only through the following properties of the multipliers ζ_i.

Assumption B.3 (Multiplier structure). Conditionally on $\mathcal{F}_0$, the variables $(\zeta_i)_{i \in I_T}$ are independent with

$$\mathbb{E}\zeta_i = 0, \qquad |\zeta_i| \le 1, \qquad \mathbb{E}\zeta_i^2 \le \varepsilon_n.$$

The class $\mathcal{H}_T$ and the constants C, M, q, B, C_H, S are $\mathcal{F}_0$-measurable. Both corruption models satisfy this: logistic $\zeta_i = (B_i - \varepsilon_n)(1 - 2Y_i)$ with $\mathbb{E}\zeta_i^2 = \varepsilon_n(1 - \varepsilon_n)$, squared $\zeta_i = B_i S_i$ with $\mathbb{E}\zeta_i^2 = \varepsilon_n$, by the joint independence of the corruption and the clean sample stated in Section 2.

The multiplier process is $Z(h) := N^{-1} \sum_{i \in I_T} \zeta_i h_i$ for $h \in \mathbb{R}^N$; it is linear with $Z(0) = 0$, and it is written Z_T where the threshold is displayed.

B.2 Entropy and the localized supremum

This subsection proves Proposition B.9, a high-probability bound on the supremum of $|Z|$ over the localized class $\mathcal{H}_T(r) = \{h \in \mathcal{H}_T : \|h\|_N \le r\}$. The proof combines two ingredients by mixed-tail generic chaining (Input B.8). The first is a sup-norm entropy bound for $\mathcal{H}_T$ (Lemmas B.5 and B.6), derived from the approximation theorem of Input B.4. The second is the Bernstein tail of the increments of Z (Lemma B.7).

Throughout, $N(t, \mathcal{S}, d)$ is the *external* covering number of a set $\mathcal{S}$ at scale t in the (semi)metric d: the centers may lie outside $\mathcal{S}$; N_{in} counts balls centered at points of the set. Lemma B.6 converts the external bounds into the internal covers used by the chaining bound.

External input B.4 (Uniform approximation with preserved budget; Siegel (2023), Theorem 3)**.** There is $A_m < \infty$ such that for every $C > 0$, every $g \in \mathcal{G}(C)$ and every $K \ge 1$ there exist $(u_j, t_j) \in \Theta$ and coefficients c_j, $j \le K$, with

$$\sum_{j=1}^{K} |c_j| \le C, \qquad \Big\| g - \sum_{j=1}^{K} c_j \varphi_{(u_j,t_j)} \Big\|_{\infty,\mathbb{B}} \le A_m \, C \, K^{-(m+3)/(2m)}. \tag{6}$$

Source. Siegel (2023) works with the dictionary $\{(\omega^\top x + b)_+ : \omega \in \mathbb{S}^{m-1},\ b \in [-1,1]\}$ on the unit ball of $\mathbb{R}^m$ (the dimension d there is our m), which is $\{\varphi_\theta : \theta \in \Theta\}$ after $b = -t$. Theorem 3 there, with $k = 1$ (the ReLU power) and no derivatives, approximates every function in the unit ball of the variation space of this dictionary uniformly on the ball at rate $K^{-(m+3)/(2m)}$ by an element of Σ_K^1, the set of K-term combinations whose coefficients satisfy $\sum_j |c_j| \le 1$. Applied to g/C and scaled back, this is (6); A_m is the constant of that theorem. □

Lemma B.5 (Entropy of $\mathcal{H}_T$ in the sample sup norm)**.** *There are a constant A_m^{ent} depending only on m and an absolute constant A_0 (one may take $A_0 = 1 + \log 4$) such that, for $0 < t \le S$,*

$$\log N\big(t, \mathcal{H}_T, \|\cdot\|_\infty\big) \ \le \ \mathcal{E}(t) := \Big[A_m^{\text{ent}} \Big(\frac{C_H}{t}\Big)^p + A_0 \, q \Big] \log\Big(e + \frac{S}{t}\Big). \tag{7}$$

The two constants are kept apart on purpose: the affine part of the entropy carries no dependence on m beyond the dimension $q \le m + 1$ itself.

Proof. Step 1 (the nonlinear class in the sup norm on the ball). We first show that for all $C > 0$ and $0 < s \le 2C$,

$$\log N\big(s, \mathcal{G}(C), \|\cdot\|_{\infty,\mathbb{B}}\big) \ \le \ c_m \Big(\frac{C}{s}\Big)^p \log\Big(e + \frac{C}{s}\Big), \tag{8}$$

with external centers. Write $d_\Theta((u,t),(u',t')) := \|u - u'\|_2 + |t - t'|$; on $\mathbb{B}_1^m$ the atoms satisfy $\|\varphi_\theta\|_{\infty,\mathbb{B}} \le 2$ and $\|\varphi_\theta - \varphi_{\theta'}\|_{\infty,\mathbb{B}} \le d_\Theta(\theta,\theta')$. Let $K := \lceil (2A_m C/s)^p \rceil$; by Input B.4, every $g \in \mathcal{G}(C)$ is within $A_m C K^{-1/p} \le s/2$ of a K-term combination with $\sum_j |c_j| \le C$. Round each θ_j to a d_Θ-net of Θ of mesh $s/(4C)$ and each c_j to a grid of step $s/(8K)$ on $[-C, C]$: the parameter rounding changes the combination by at most $\sum_j |c_j| \cdot s/(4C) \le s/4$ and the coefficient rounding by at most $K \cdot 2 \cdot s/(8K) = s/4$, so the total error is at most s. Θ is m-dimensional, so the net has at most $(cC/s)^m$ points, and each coefficient grid has at most $1 + 16CK/s$ points; hence $\log N \le K\big[m \log(cC/s) + \log(1 + 16CK/s)\big]$, which is (8) because $K \le (cC/s)^p \le (cC/s)^2$ inside the logarithm.

Step 2 (split). By Lemma B.2(iii), $h = g + a$ with $g \in V_T(C_H)$ and $a \in \mathcal{A}_T$, $\|a\|_\infty \le S$.

Step 3 (nonlinear part). Since $\|\mathrm{ev}_T(g) - \mathrm{ev}_T(g')\|_\infty \le \|g - g'\|_{\infty,\mathbb{B}}$, $N(t/2, V_T(C_H), \|\cdot\|_\infty) \le N(t/2, \mathcal{G}(C_H), \|\cdot\|_{\infty,\mathbb{B}})$, and (8) with $C_H \le S/2$ gives $\log N \le c_m (2C_H/t)^p \log(e + 2C_H/t) \le A_m^{\text{ent}} (C_H/t)^p \log(e + S/t)$ with $A_m^{\text{ent}} := 2^p c_m$. (For $t/2 \ge 2C_H$ the set $V_T(C_H)$ lies in a single ball and the bound is trivial.)

Step 4 (affine part). $\{a \in \mathcal{A}_T : \|a\|_\infty \le S\}$ is the ball of radius S of the q-dimensional normed space $(\mathcal{A}_T, \|\cdot\|_\infty)$; the volumetric bound gives $N(t/2, \cdot, \|\cdot\|_\infty) \le (1 + 4S/t)^q$, i.e. $\log N \le q \log(1 + 4S/t) \le A_0 \, q \log(e + S/t)$ with the absolute constant $A_0 := 1 + \log 4$, because $1 + 4S/t \le 4(e + S/t)$ and $\log(e + S/t) \ge 1$.

Step 5 (sum). Sums of a $t/2$-center for g and a $t/2$-center for a form a t-net of $\mathcal{H}_T$, so the logarithms add. □

Lemma B.6 (Covering conventions). *Let $\bar{\mathcal{E}}(t) := \min\{\mathcal{E}(t),\ N\log(1+2B/t)\}$ for $0 < t \le S$ and $\bar{\mathcal{E}}(t) := 0$ for $t > S$. Then $\log N(t, \mathcal{H}_T, \|\cdot\|_\infty) \le \bar{\mathcal{E}}(t)$ for all $t > 0$, and for every $F \subseteq \mathcal{H}_T$ and $t > 0$,*

$$\log N_{\mathrm{in}}(t, F, \|\cdot\|_N) \ \le\ \log N_{\mathrm{in}}(t, F, \|\cdot\|_\infty) \ \le\ \log N(t/2, \mathcal{H}_T, \|\cdot\|_\infty) \ \le\ \bar{\mathcal{E}}(t/2).$$

Proof. $\mathcal{H}_T \subset [-B, B]^N$ is covered by $\lceil 2B/t\rceil^N \le (1+2B/t)^N$ sup-norm balls of radius t, and for $t > S \ge B$ the single ball of radius t centered at 0 suffices; with (7) this gives the first claim. For the second, from an external $t/2$-cover of $\mathcal{H}_T$ keep the balls that meet F and choose one point of F in each; by the triangle inequality these points form an internal t-cover of F in $\|\cdot\|_\infty$, and $\|\cdot\|_N \le \|\cdot\|_\infty$. □

For $m \ge 3$ the bound (7) grows like t^{-p} with $p \ge 1$ and is not integrable at 0; the finite-dimensional bound inside $\bar{\mathcal{E}}$ supplies the fine-scale truncation that Proposition B.9 needs.

Lemma B.7 (Bernstein increments). *Under Assumption B.3, for fixed $h, h' \in \mathbb{R}^N$ and $u > 0$,*

$$\mathbb{P}\Big(|Z(h) - Z(h')| > \sqrt{\tfrac{2\varepsilon_n u}{N}}\,\|h-h'\|_N + \tfrac{u}{3N}\|h-h'\|_\infty \ \Big|\ \mathcal{F}_0\Big) \le 2e^{-u}.$$

Proof. Write $Z(h) - Z(h') = \sum_i \xi_i$ with $\xi_i := \zeta_i(h_i - h'_i)/N$, independent and centered given $\mathcal{F}_0$, $|\xi_i| \le b := \|h-h'\|_\infty/N$ and $\sum_i \mathrm{Var}(\xi_i) \le (\varepsilon_n/N^2)\sum_i (h_i - h'_i)^2 = (\varepsilon_n/N)\|h-h'\|_N^2 =: \sigma^2$. Bernstein's inequality in the form $\mathbb{P}(|\sum \xi_i| \ge \sqrt{2\sigma^2 u} + bu/3) \le 2e^{-u}$ (Boucheron et al., 2013, Theorem 2.10) gives the claim. □

With $d_2(h,h') := \sqrt{2\varepsilon_n/N}\,\|h-h'\|_N$ and $d_1(h,h') := \|h-h'\|_\infty/(3N)$, Lemma B.7 reads $\mathbb{P}(|Z(h) - Z(h')| > \sqrt{u}\,d_2 + u\,d_1 \mid \mathcal{F}_0) \le 2e^{-u}$: the mixed-tail increment condition of the following input.

External input B.8 (Mixed-tail generic chaining; Dirksen (2015), Theorem 3.5 and eq. (4)). Let $(X_t)_{t\in\mathcal{T}}$ be a real process on a finite index set such that $\mathbb{P}(|X_t - X_s| \ge \sqrt{u}\,d_2(s,t) + u\,d_1(s,t)) \le 2e^{-u}$ for all s, t and $u > 0$, where d_1, d_2 are semimetrics on $\mathcal{T}$. There is an absolute constant C_D such that for every $t_0 \in \mathcal{T}$ and $u \ge 1$,

$$\mathbb{P}\Big(\sup_{t\in\mathcal{T}}|X_t - X_{t_0}| \ge C_D\big[\gamma_2(\mathcal{T}, d_2) + \gamma_1(\mathcal{T}, d_1) + \sqrt{u}\ \mathrm{diam}_{d_2}(\mathcal{T}) + u\ \mathrm{diam}_{d_1}(\mathcal{T})\big]\Big) \le e^{-u}.$$

Moreover, with internal covering numbers, $\gamma_\alpha(\mathcal{T}, d) \le c_\alpha \int_0^{\mathrm{diam}_d(\mathcal{T})} \big(\log N_{\mathrm{in}}(s, \mathcal{T}, d)\big)^{1/\alpha} ds$ for $\alpha \in \{1, 2\}$ (eq. (4) there).

Proposition B.9 (Localized supremum). *For $r > 0$ let $\mathcal{H}_T(r) := \{h \in \mathcal{H}_T : \|h\|_N \le r\}$ and define*

$$J_2(r) := \int_0^{2\min(r,B)} \sqrt{\bar{\mathcal{E}}(t)}\,dt, \qquad J_{1,N} := \int_0^{2B} \bar{\mathcal{E}}(t)\,dt.$$

There is an absolute constant C_1 such that for every $u \ge 1$, with conditional probability at least $1 - e^{-u}$,

$$\sup_{h\in\mathcal{H}_T(r)} |Z(h)| \ \le\ Q(r,u) := C_1\Big[\sqrt{\tfrac{\varepsilon_n}{N}}\,J_2(r) + \frac{J_{1,N}}{N} + r\sqrt{\tfrac{\varepsilon_n u}{N}} + \frac{Bu}{N}\Big].$$

Proof. Step 0 (finite index sets). Conditionally on $\mathcal{F}_0$ it suffices to bound $\sup_F |Z|$ for finite sets $F \subset \mathcal{H}_T(r)$ containing 0: Z is linear, hence continuous, so the suprema over an increasing sequence of finite subsets exhausting a countable dense subset of $\mathcal{H}_T(r)$ increase to $\sup_{\mathcal{H}_T(r)}|Z|$, and a bound that does not depend on F passes to the limit.

Step 1 (diameters). $\mathrm{diam}_{d_2} F \le \sqrt{2\varepsilon_n/N}\cdot 2\min(r,B)$ and $\mathrm{diam}_{d_1} F \le 2B/(3N)$.

Step 2 (γ_2). $N_{\mathrm{in}}(s, F, d_2) = N_{\mathrm{in}}\big(s\sqrt{N/(2\varepsilon_n)}, F, \|\cdot\|_N\big)$, whose logarithm is at most $\bar{\mathcal{E}}\big(\tfrac{1}{2}s\sqrt{N/(2\varepsilon_n)}\big)$ by Lemma B.6. Substituting $t = s\sqrt{N/(2\varepsilon_n)}$ and then $t = 2t'$,

$$\begin{aligned}\int_0^{\mathrm{diam}_{d_2}} \sqrt{\log N_{\mathrm{in}}(s, F, d_2)}\,ds &\le \sqrt{\tfrac{2\varepsilon_n}{N}}\int_0^{2\min(r,B)} \sqrt{\bar{\mathcal{E}}(t/2)}\,dt \\ &= 2\sqrt{\tfrac{2\varepsilon_n}{N}}\int_0^{\min(r,B)} \sqrt{\bar{\mathcal{E}}(t')}\,dt' \ \le\ 2\sqrt{\tfrac{2\varepsilon_n}{N}}\,J_2(r).\end{aligned}$$

Step 3 (γ_1). Likewise $\log N_{\mathrm{in}}(s,F,d_1) \le \bar{\mathcal{E}}(3Ns/2)$ and $\int_0^{\mathrm{diam}_{d_1}} \log N_{\mathrm{in}}(s,F,d_1)\,ds \le \frac{1}{3N}\int_0^{2B}\bar{\mathcal{E}}(t/2)\,dt = \frac{2}{3N}\int_0^{B}\bar{\mathcal{E}}(t')\,dt' \le \frac{2J_{1,N}}{3N}$.

Step 4 (assembly). Apply Input B.8 to F with $t_0 = 0$ and $Z(0) = 0$; absorb the numerical constants of Steps 1–3 and of the γ bounds into C_1, and conclude by Step 0. □

B.3 The offset inequality and the explicit remainder

Theorem B.10 (Sparse offset inequality). *Let $\tau > 0$, $\kappa > 0$, $\delta \in (0,1)$ and $u := 1 + \log(1/\delta)$. Let $r_0 > 0$ be $\mathcal{F}_0$-measurable (a function of the class $\mathcal{H}_T$ and of the constants, not of the multipliers) and suppose that*

$$\tau\, Q(r_0, u) \;\le\; \frac{\kappa r_0^2}{8}. \tag{9}$$

Then, with conditional probability at least $1-\delta$, simultaneously for all $h \in \mathcal{H}_T$,

$$\tau|Z(h)| \;\le\; \frac{\kappa}{4}\|h\|_N^2 + \Delta, \qquad \Delta := \frac{\kappa r_0^2}{8}. \tag{10}$$

Proof. Step 1 (the good event). Let $E := \{\sup_{h\in\mathcal{H}_T(r_0)}|Z(h)| \le Q(r_0,u)\}$; by Proposition B.9, $\mathbb{P}(E \mid \mathcal{F}_0) \ge 1 - e^{-u} \ge 1-\delta$. Work on E.

Step 2 (inside the radius). If $\|h\|_N \le r_0$ then $\tau|Z(h)| \le \tau Q(r_0,u) \le \kappa r_0^2/8 = \Delta \le \frac{\kappa}{4}\|h\|_N^2 + \Delta$.

Step 3 (outside the radius; scaling). If $\|h\|_N > r_0$, put $h' := (r_0/\|h\|_N)\,h$. By Lemma B.2(iii), $h' \in \mathcal{H}_T$ and $\|h'\|_N = r_0$; by linearity $Z(h) = (\|h\|_N/r_0)Z(h')$. Hence

$$\tau|Z(h)| \le \frac{\|h\|_N}{r_0}\cdot\frac{\kappa r_0^2}{8} = \frac{\kappa r_0\|h\|_N}{8} < \frac{\kappa}{8}\|h\|_N^2 \le \frac{\kappa}{4}\|h\|_N^2.$$

Steps 2 and 3 together give (10) on E. □

The event E_δ^{sq} of Section 3 is the event E of this proof at $(\tau,\kappa) = (2A, 2)$ with the radius of Proposition B.11, and the event E_δ of the logistic extension is the same at $(\tau,\kappa) = (1,\kappa_M)$; Δ_T^{sq} and $\Delta_T^{\log}$ are the corresponding values of Δ.

Conventions. $L(t) := \log(e + S/t)$ is decreasing in t, with $L(t) \le L(a) + \log(a/t)$ for $0 < t \le a$ (because $e + S/t \le (e + S/a)\,a/t$). Constants c_m depend only on m and may change from line to line; c denotes an absolute constant. Two elementary integrals are used repeatedly: for $a > 0$ and $0 \le \gamma < 1$,

$$\int_0^a t^{-\gamma}\sqrt{L(t)}\,dt \le c_\gamma\, a^{1-\gamma}\sqrt{L(a)}, \qquad \int_0^a t^{-\gamma}L(t)\,dt \le c_\gamma\, a^{1-\gamma}L(a), \tag{11}$$

with c_γ depending only on γ; both follow from $L(t) \le L(a) + \log(a/t)$, $\sqrt{L(t)} \le \sqrt{L(a)} + \sqrt{\log(a/t)}$, $L \ge 1$ and $\int_0^1 s^{-\gamma}\log^{\vartheta}(1/s)\,ds < \infty$ for $\vartheta \in \{\frac12, 1\}$.

Proposition B.11 (Explicit remainder). *Let $u := 1 + \log(1/\delta)$, $L_N := \log\big(e + S\sqrt{\kappa N/(\tau B)}\,\big)$ and*

$$R_N := c_m\Big[S\,L(B)\,N^{-\gamma_m} + \frac{qB\,L(B)}{N}\Big],$$

with $c_m \ge 1$ large enough depending on m. Define

$$r_a^2 := c_m\Big(\frac{\tau^2\varepsilon_n L_N}{\kappa^2 N}\Big)^{\alpha_m} C_H^{\beta_m}, \qquad r_b^2 := c\,\frac{\tau^2\varepsilon_n\, qL_N}{\kappa^2 N},$$
$$r_c^2 := c\,\frac{\tau^2\varepsilon_n\, u}{\kappa^2 N}, \qquad r_d^2 := c\,\frac{\tau}{\kappa}\Big(\min\{4B\log 2, R_N\} + \frac{Bu}{N}\Big),$$

and $r_0^2 := r_a^2 + r_b^2 + r_c^2 + r_d^2$, with c a large enough absolute constant and c_m large enough depending on m. Then r_0 is $\mathcal{F}_0$-measurable, satisfies (9), and the remainder of Theorem B.10 obeys, for all values of the parameters,

$$\begin{aligned}\Delta \;\le\; c_m\Big[\,&\kappa\, C_H^{\beta_m}\Big(\frac{\tau^2\varepsilon_n L_N}{\kappa^2 N}\Big)^{\alpha_m} + \frac{\tau^2\varepsilon_n\,(qL_N + u)}{\kappa N}\\ &+ \tau\min\Big\{4B\log 2,\; S\,L(B)\,N^{-\gamma_m} + \frac{qB\,L(B)}{N}\Big\} + \frac{\tau B u}{N}\Big].\end{aligned} \tag{12}$$

Proof. The proof has three parts. Steps 0–2 control the logarithm at the radius and bound the two chaining integrals J_2 and $J_{1,N}$ of Proposition B.9. Steps 3–4 split the condition (9) into four inequalities, each ensured by one of the radii $r_a, \dots, r_d$. Step 5 adds the radii to obtain (12).

Step 0 (control of the logarithm). $r_0 \ge r_d \ge \sqrt{c\tau Bu/(\kappa N)} \ge \sqrt{c\tau B/(\kappa N)}$, so $L(r_0) = \log(e + S/r_0) \le L_N$ (adjusting c).

Step 1 (the γ_2 integral). Put $\mathcal{E}_*(t) := [A_m^{\mathrm{ent}}(C_H/t)^p + A_0 q]L(t)$ for all $t > 0$. Then $\bar{\mathcal{E}} \le \mathcal{E}_*$ on $(0, \infty)$ (by (7) for $t \le S$, trivially for $t > S$), so $J_2(r) \le \int_0^{2r} \sqrt{\mathcal{E}_*(t)}\, dt$ for every $r > 0$. Since $\sqrt{\mathcal{E}_*(t)} \le \sqrt{A_m^{\mathrm{ent}}}\,(C_H/t)^{p/2}\sqrt{L(t)} + \sqrt{A_0 q}\,\sqrt{L(t)}$, (11) with $\gamma = p/2 < 1$ for the first term and $\gamma = 0$ for the second, together with $L(2r) \le L(r)$, gives

$$J_2(r) \;\le\; c_m\, C_H^{p/2}\, r^{1-p/2}\sqrt{L(r)} + c\, r\sqrt{q\, L(r)} \qquad (r > 0).$$

Step 2 (the γ_1 integral). Since $B \le S$, $\log(1 + 2B/t) \le L(t) + \log 2 \le 2L(t)$, so for $0 < t \le 2B$

$$\bar{\mathcal{E}}(t) \;\le\; L(t)\, \min\{A_m^{\mathrm{ent}}(C_H/t)^p,\, 2N\} + A_0\, q\, L(t).$$

For $x, y > 0$ and $\vartheta \in [0, 1]$, $\min\{x, y\} \le x^{\vartheta} y^{1-\vartheta}$; with $\vartheta = \gamma_m$, for which $p\gamma_m = m/(m+1) < 1$,

$$\min\{A_m^{\mathrm{ent}}(C_H/t)^p,\, 2N\} \;\le\; c_m\, N^{1-\gamma_m}\, C_H^{m/(m+1)}\, t^{-m/(m+1)}.$$

Integrating over $(0, 2B)$ with (11) at $\gamma = m/(m+1)$ and at $\gamma = 0$, and using $L(2B) \le L(B)$ and $C_H^{m/(m+1)} B^{1/(m+1)} \le \max\{C_H, B\} \le S$,

$$\frac{J_{1,N}}{N} \;\le\; c_m\, S\, L(B)\, N^{-\gamma_m} + c\, \frac{qB\, L(B)}{N} \;\le\; R_N.$$

Independently of the budget, $\bar{\mathcal{E}}(t) \le N \log(1 + 2B/t)$ and $\int_0^{2B} \log(1 + 2B/t)\, dt = 2B \int_0^1 \log(1 + 1/s)\, ds = 4B \log 2$; hence $J_{1,N}/N \le \min\{4B \log 2,\ R_N\}$.

Step 3 (four sufficient conditions). By Step 1 at r_0 and Step 0, $\tau Q(r_0, u) \le \tau C_1 \big[c_m \sqrt{\varepsilon_n/N}\, C_H^{p/2} r_0^{1-p/2}\sqrt{L_N} + c\sqrt{\varepsilon_n/N}\, r_0 \sqrt{qL_N} + r_0\sqrt{\varepsilon_n u/N}\big] + \tau C_1 \big[J_{1,N}/N + Bu/N\big]$. It suffices that each of the four pieces is at most $\kappa r_0^2/32$.

Step 4 (the four pieces). (a) $\tau c\sqrt{\varepsilon_n L_N/N}\, C_H^{p/2} r_0^{1-p/2} \le \kappa r_0^2/32$ is equivalent to $r_0^{1+p/2} \ge 32c\,\tau\sqrt{\varepsilon_n L_N/N}\, C_H^{p/2}/\kappa$, i.e. to

$$r_0^2 \;\ge\; (32c)^{2\alpha_m} \Big(\frac{\tau^2 \varepsilon_n L_N}{\kappa^2 N}\Big)^{\alpha_m} C_H^{p\alpha_m},$$

since $2/(1 + p/2) = 2\alpha_m$ and $p\alpha_m = \beta_m$; this holds because $r_0 \ge r_a$. (b) $\tau c\sqrt{\varepsilon_n qL_N/N}\, r_0 \le \kappa r_0^2/32$ iff $r_0 \ge 32c\tau\sqrt{\varepsilon_n qL_N/N}/\kappa$, implied by $r_0 \ge r_b$. (c) $\tau C_1 r_0 \sqrt{\varepsilon_n u/N} \le \kappa r_0^2/32$ is implied by $r_0 \ge r_c$. (d) $\tau C_1(J_{1,N}/N + Bu/N) \le \kappa r_0^2/32$ is implied by $r_0^2 \ge r_d^2$ and Step 2.

Step 5 (assembly). $\Delta = \kappa r_0^2/8 = \frac{\kappa}{8}(r_a^2 + r_b^2 + r_c^2 + r_d^2)$; insert the four definitions and use $L_N^{\alpha_m} \le L_N$ and, since the constant in R_N is at least 1, $\min\{4B \log 2, R_N\} \le c_m \min\{4B \log 2,\ SL(B)N^{-\gamma_m} + qBL(B)/N\}$, which pulls the constant outside the minimum in (12). Measurability: every ingredient of r_0 is a function of $(C_H, B, S, q, N, \varepsilon_n, \tau, \kappa, u)$. □

Parameter bookkeeping.

symbol	meaning	where it enters Δ
τ	process scaling (loss-specific)	every term
κ	quadratic absorption coefficient	$\kappa^{1-2\alpha_m}$ in the main term; κ^{-1} in the terms carrying $\tau^2\varepsilon_n/N$; κ^0 in $\tau J_{1,N}/N$ and $\tau Bu/N$ (algebraic dependence at fixed C_H, B, S; logarithms excluded)
ε_n	variance proxy of ζ_i (corruption rate)	$\varepsilon_n^{\alpha_m}$ main term, ε_n in (b),(c); absent from (d)
$N = n_T$	number of deep samples	$N^{-\alpha_m}$, N^{-1}, $N^{-\gamma_m}$
$C_H = 2C$	nonlinear budget of the difference class	$C_H^{\beta_m}$ main term, inside S in (d)
$B = 2M$	sup bound of the difference class	(d)
$S = B + 2C_H$	affine sup bound	(d) and logarithms
$q \le m + 1$	affine dimension	(b), (d)
$u = 1 + \log(1/\delta)$	confidence	(c), (d)

B.4 RATES AND THE TWO LOSSES

Fixed parameters. Fix $\tau, \kappa, \varepsilon_n, C_H, B, q, m$ and let $N \to \infty$ with $\delta = N^{-c}$. Then L_N, $L(B)$ and u are logarithmic and (12) reads

$$\Delta \;\le\; \widetilde{O}_m\big(N^{-\alpha_m}\big) + \widetilde{O}\big(N^{-1}\big) + \widetilde{O}_m\big(N^{-\gamma_m}\big).$$

Since $\alpha_m < 1$ and $\gamma_m - \alpha_m = \frac{m+3}{(2m+2)(2m+3)} > 0$, the main term is $N^{-\alpha_m}$.

m	α_m	β_m	γ_m
1	4/5	2/5	1
2	5/7	4/7	5/6
3	2/3	2/3	3/4
5	8/13	10/13	2/3

Validity for every ε_n. Theorem B.10 and Proposition B.11 hold for every $\varepsilon_n \in (0, 1]$ at finite N and for every value of C_H and B; only the identification of the leading term uses fixed parameters.

Remark B.12 (Sparse regime: a sufficient condition). Suppose all parameters except $\varepsilon_n = N^{-a}$, $a \ge 0$, are fixed. If $a < \gamma_m$ then $\Delta = o(\varepsilon_n)$ up to logarithms: the main term is $N^{-(1+a)\alpha_m} = o(N^{-a})$ because $(1+a)\alpha_m > a$ for all $a < (m+3)/m$; the terms with an explicit factor ε_n are $O(\varepsilon_n/N)$; and the ε_n-free terms $N^{-\gamma_m}$ and N^{-1} are $o(N^{-a})$ exactly when $a < \gamma_m$. The condition is sufficient, not necessary.

Remark B.13 (Sanity check on a sparse-multiplier example). Take the ℓ^1 ball $\mathcal{H} = \{h \in \mathbb{R}^N : \|h\|_1 \le 1\}$ ($B = 1$), the multiplier $\zeta_i = B_i S_i$ and $\varepsilon_n = N^{-1/2}$, $r = N^{-1/2}$, so that $\mathcal{H}(r) = \mathcal{H}$ and $\sup_{\mathcal{H}} |Z(h)| = N^{-1}\mathbf{1}\{\exists i : B_i = 1\}$, with expectation $(1 - (1-\varepsilon_n)^N)/N$, of order $1/N$. Its sup-norm entropy obeys $\log N(t, \mathcal{H}, \|\cdot\|_\infty) \le c\, t^{-1} \log(2N)$ for $t \in (0, 1)$, because at most $1/t$ coordinates exceed t in absolute value. Proposition B.9 then gives $Q(r, u) = \widetilde{O}(N^{-1})$: the d_2 part contributes $\sqrt{\varepsilon_n/N}\, J_2(r) \lesssim N^{-1}\sqrt{\log(2N)}$ and the d_1 part $J_{1,N}/N \lesssim N^{-1}\log^2(2N)$. A bound of the form $\sqrt{\varepsilon_n}\,\phi(r)$ with ϕ the Rademacher complexity, here $\phi(r) = 1/N$, would give $N^{-5/4}$, below the true order: the sparse-multiplier bound is not the ordinary Rademacher complexity scaled by $\sqrt{\varepsilon_n}$.

The logarithmic factor of the main text. Section 3 and the explicit remainders of §C.4 and §F.5 collect the logarithms of (12) into

$$\mathcal{L}_T := 1 + L_N + L(2M),$$

evaluated with the substitutions of the corresponding loss; $L(2M)$ is $L(B)$ at $B = 2M$. $\mathcal{L}_T$ involves n_T, the budget C_T, M and $1/M$ and, for the squared loss, $1/A$; it is logarithmic in n_T when the other parameters are fixed, and it does not involve δ. Rates are read at $\delta = n_T^{-c}$ for a fixed $c > 0$; this convention serves only to read rates and is never an assumption of a finite-sample statement.

How the two losses instantiate the result. The master inequalities, Theorem C.5 and Theorem F.6, apply (10) to the difference h between a class vector and a comparator in the class, so that $h \in \mathcal{H}_T$ by Definition B.1. What replaces curvature differs between the losses. The proofs are in §C.3 and §F.4.

	squared (Appendix C)	logistic (Appendix F)
conditioning $\mathcal{F}_0$	design	design and clean labels
multiplier ζ_i	$B_i S_i$	$(B_i - \varepsilon_n)(1 - 2Y_i)$
$\mathbb{E}\zeta_i^2$	ε_n	$\varepsilon_n(1-\varepsilon_n) \le \varepsilon_n$
τ	$2A$	1
κ	2	$\kappa_M = \sigma(M)(1-\sigma(M))$
budget C of $\mathcal{K}_T$	$C_T^{\mathrm{sq}} = \bar{B}_\eta^{\mathrm{sq}} / g_{\mathcal{D}}^{\min}(T)$	$C_T^{\log} = \bar{B}_{\eta,M}^{\log} / g_{\mathcal{D}}^{\min}(T)$
comparator	$v^\dagger$, the projection of v_0 (v_0 itself in the well-specified case)	v^*, the mean-flip benchmark on $\mathcal{K}_T^{\log}$
in place of curvature	the quadratic identity (Lemma C.2)	the curvature of the mean-flip objective (Lemma F.5)
leading term of Δ	$A^{2\alpha_m}(C_T^{\mathrm{sq}})^{\beta_m}(\varepsilon_n/N)^{\alpha_m}$	$\kappa_M^{1-2\alpha_m}(C_T^{\log})^{\beta_m}(\varepsilon_n/N)^{\alpha_m}$

The leading terms are stated up to c_m, logarithms and the factor 2^{β_m} from $C_H = 2C$. The budget-substituted form of the logistic remainder, in $D_T \geq \kappa_M C_T^{\log}$, is derived in §F.5.

C SQUARED-LOSS-SPECIFIC PROOFS

This appendix supplies the squared-loss input to Appendix B and proves the statements of §3.3: the budget behind (A1) (§C.1); the proof of Theorem 5 and the sufficient condition for the well-specified case (§C.2); a teacher outside the class (§C.3); and the explicit form of Δ_T^{sq} (§C.4). Notation is that of Appendix B ($N := n_T$). $L_T^{\mathrm{sq}}(v) = N^{-1}\sum_{i\in I_T}(v_i - Y_i)^2$, $E(v) = \|v - v_0\|_N^2$ and $\hat{\varepsilon}_T$ are as in §3.1. The multipliers are $\zeta_i = B_i S_i$, so that $Y = v_0 + A\zeta$ on I_T. All probabilities are conditional on $\mathcal{F}_0 = \sigma(X)$, since the teacher is deterministic, and the standing condition is $0 < \eta \leq 1$.

C.1 THE STABILITY BUDGET

Proposition C.1 (Squared-loss stability budget)**.** *Let f_θ satisfy (A0)–(A2) for the squared loss with the observed responses, $L_n^{\mathrm{sq}}(\theta) = n^{-1}\sum_{i\leq n}(f_\theta(X_i) - Y_i)^2$, and let $\max_{i\leq n}|f_0(X_i)| \leq B_0$. Then*

$$\|f_\theta\|_{\mathrm{path},g} \;\leq\; \bar{B}_\eta^{\mathrm{sq}} = \frac{1}{2\eta} - \frac{1}{2} + \frac{(R+1)(M+B_0+A)}{2}. \tag{13}$$

The bound depends on (R, M, B_0, A, η) only, not on the realization of the corruption, and it is nonnegative under $\eta \leq 1$, so $C_T^{\mathrm{sq}} = \bar{B}_\eta^{\mathrm{sq}}/g_{\mathcal{D}}^{\min}(T)$ is a legitimate budget.

Proof. Step 1 (Hessian of the unhalved loss). Under (A0) the loss is twice differentiable at θ and, with $J_i := \nabla_\theta f_\theta(X_i)$, $F_i := \nabla_\theta^2 f_\theta(X_i)$ and $G_\theta := n^{-1}\sum_i J_i J_i^\top$,

$$\nabla_\theta^2 L_n^{\mathrm{sq}}(\theta) = 2G_\theta + \frac{2}{n}\sum_{i\leq n}\big(f_\theta(X_i) - Y_i\big)F_i.$$

Step 2 (residual factor). By (A2) and $|Y_i| \leq B_0 + A$ (deterministic, because $|\zeta_i| \leq 1$), $|f_\theta(X_i) - Y_i| \leq M + B_0 + A$ for every $i \leq n$ and every realization.

Step 3 (second derivatives of the network). F_i is block diagonal over neurons: for a neuron active at X_i its block, in the coordinates $(v_k; (w_k, b_k))$, is $\begin{pmatrix} 0 & \xi_i^\top \\ \xi_i & 0 \end{pmatrix}$ with $\xi_i = (X_i, -1)$; inactive neurons contribute 0; β has no second derivatives. Hence $\|F_i\|_{\mathrm{op}} \leq \sqrt{\|X_i\|^2 + 1} \leq R + 1$ and $\big\|\frac{2}{n}\sum_i (f_\theta(X_i) - Y_i)F_i\big\|_{\mathrm{op}} \leq 2(R+1)(M + B_0 + A)$.

Step 4 (Weyl). $2\lambda_{\max}(G_\theta) \leq \lambda_{\max}(\nabla^2 L_n^{\mathrm{sq}}) + 2(R+1)(M+B_0+A) \leq 2/\eta + 2(R+1)(M+B_0+A)$, so $\lambda_{\max}(G_\theta) \leq 1/\eta + (R+1)(M + B_0 + A)$.

Step 5 (Jacobian inequality). It remains to bound $\lambda_{\max}(G_\theta)$ from below by the path norm. Let $\bar{J} := n^{-1}\sum_i J_i$. Since $G_\theta - \bar{J}\bar{J}^\top = n^{-1}\sum_i (J_i - \bar{J})(J_i - \bar{J})^\top \succeq 0$, $\lambda_{\max}(G_\theta) \geq \|\bar{J}\|^2$. The β coordinate of $\bar{J}$ equals 1. For a neuron with $w_k \neq 0$ put $u_k := w_k/\|w_k\|$, $t_k := b_k/\|w_k\|$, $p_k := P_{\mathcal{D}}(u_k^\top X > t_k)$, $s_k := \mathbb{E}_{\mathcal{D}}(u_k^\top X - t_k)_+$ and $c_k := \mathbb{E}_{\mathcal{D}}[X \mid u_k^\top X > t_k]$. Its block of $\bar{J}$, in the coordinates $(v_k; (w_k, b_k))$, is $(\|w_k\| s_k;\, v_k p_k c_k,\, -v_k p_k)$, so by $a^2 + b^2 \geq 2ab$ its squared norm is at least

$$2|v_k|\|w_k\|\, p_k s_k \sqrt{1 + \|c_k\|^2} = 2|v_k|\|w_k\|\, \widetilde{g}_{\mathcal{D}}(u_k, t_k) \geq 2|v_k|\|w_k\|\, g_{\mathcal{D}}(u_k, t_k),$$

where every term vanishes if $p_k = 0$; neurons with $w_k = 0$ have nonnegative blocks. Summing gives $\lambda_{\max}(G_\theta) \geq 1 + 2\|f_\theta\|_{\mathrm{path},g}$, the inequality of Liang et al. (2025, Appendix C, eq. (27)); hence $\|f_\theta\|_{\mathrm{path},g} \leq \frac{1}{2}(\lambda_{\max}(G_\theta) - 1) \leq \bar{B}_\eta^{\mathrm{sq}}$. □

Three facts about the budget are used by the other proofs. The proof evaluates f_θ and f_0 at the training points only, so the sample envelope (A2) suffices. Since $\partial^2 L_n^{\mathrm{sq}}/\partial\beta^2 \equiv 2$, (A1) forces $\eta \leq 1$, and under this standing condition $\bar{B}_\eta^{\mathrm{sq}} \geq 0$. Because $|Y_i| \leq B_0 + A$ holds for every realization, $\bar{B}_\eta^{\mathrm{sq}}$, C_T^{sq} and $\mathcal{K}_T^{\mathrm{sq}}$ are $\mathcal{F}_0$-measurable, as Assumption B.3 requires; the argument uses this predetermined budget, not residuals or path norms measured after the corruption is observed.

C.2 THE QUADRATIC IDENTITY AND THE PROOF OF THEOREM 5

Lemma C.2 (Quadratic identity on vectors)**.** *For every $v \in \mathbb{R}^N$, $L_T^{\mathrm{sq}}(v) = E(v) - 2A\,Z(v - v_0) + A^2\hat{\varepsilon}_T$.*

Proof. $Y - v_0 = A\zeta$ coordinatewise, so $(v_i - Y_i)^2 = (v_i - v_{0,i})^2 - 2A\zeta_i(v_i - v_{0,i}) + A^2\zeta_i^2$ and $\zeta_i^2 = B_i$; average over I_T. □

Throughout, E_δ^{sq} is the event of Theorem B.10 for $\mathcal{H}_T = \mathcal{K}_T^{\mathrm{sq}} - \mathcal{K}_T^{\mathrm{sq}}$ at $(\tau, \kappa) = (2A, 2)$, with the $\mathcal{F}_0$-measurable radius of Proposition B.11. It has conditional probability at least $1 - \delta$, and on it

$$2A|Z(h)| \;\le\; \tfrac{1}{2}\|h\|_N^2 + \Delta_T^{\mathrm{sq}} \qquad \text{for all } h \in \mathcal{H}_T. \tag{14}$$

Proof of Theorem 5. Let $v \in \mathcal{K}_T^{\mathrm{sq}}$ and $h := v - v_0$; since $v_0 \in \mathcal{K}_T^{\mathrm{sq}}$, $h \in \mathcal{H}_T$ and $\|h\|_N^2 = E(v)$. By Lemma C.2 and (14), on E_δ^{sq},

$$L_T^{\mathrm{sq}}(v) = A^2\hat{\varepsilon}_T + E(v) - 2A\,Z(h) \;\ge\; A^2\hat{\varepsilon}_T + E(v) - \tfrac{1}{2}E(v) - \Delta_T^{\mathrm{sq}} = A^2\hat{\varepsilon}_T + \tfrac{1}{2}E(v) - \Delta_T^{\mathrm{sq}}.$$

At $v = v_0$, $E(v_0) = 0$ and $Z(0) = 0$, so $L_T^{\mathrm{sq}}(v_0) = A^2\hat{\varepsilon}_T$. □

Section 3.3 gives a sufficient condition for the assumption $v_0 \in \mathcal{K}_T^{\mathrm{sq}}$ in terms of the variation of the teacher, and Lemma C.3 below proves it. For a function g on $\mathbb{B}_1^m$, $|g|_{\mathcal{V}(\mathbb{B}_1^m)}$ denotes the infimum of $\|\mu\|_{\mathrm{TV}}$ over all representations $g = \ell + g_\mu$ on $\mathbb{B}_1^m$ with ℓ affine and g_μ as in Definition B.1.

Lemma C.3 (A sufficient condition for the well-specified case)**.** *If $f_0 = \ell_0 + g_0$ on $\mathbb{B}_1^m$ with ℓ_0 affine and $|g_0|_{\mathcal{V}(\mathbb{B}_1^m)} \le C_T^{\mathrm{sq}}$, and $|f_0(X_i)| \le M$ for $i \in I_T$, then $v_0 \in \mathcal{K}_T^{\mathrm{sq}}$.*

Proof. Write $C := C_T^{\mathrm{sq}}$. For every $j \ge 1$ there is a representation $g_0 = \ell_j + g_{\mu_j}$ on $\mathbb{B}_1^m$ with ℓ_j affine and $\|\mu_j\|_{\mathrm{TV}} \le C + 1/j$, so $v_0 = (C + 1/j)\,z_j + a_j$ with $z_j := \mathrm{ev}_T(g_{\mu_j})/(C + 1/j) \in V_T(1)$ and $a_j := \mathrm{ev}_T(\ell_0 + \ell_j) \in \mathcal{A}_T$. By Lemma B.2(ii), $V_T(1)$ is compact, so along a subsequence $z_j \to z \in V_T(1)$; then $a_j = v_0 - (C + 1/j)z_j \to v_0 - Cz$, which lies in the closed subspace $\mathcal{A}_T$. Hence $v_0 = Cz + (v_0 - Cz) \in V_T(C) + \mathcal{A}_T$, and the envelope on the deep samples places v_0 in the box $[-M, M]^N$; so $v_0 \in \mathcal{K}_T^{\mathrm{sq}}$. □

C.3 A TEACHER OUTSIDE THE CLASS

When $v_0 \notin \mathcal{K}_T^{\mathrm{sq}}$, the comparison point is the projection of v_0 onto the class.

Lemma C.4 (Projection benchmark)**.** *There is a unique $v^\dagger \in \mathcal{K}_T^{\mathrm{sq}}$ minimizing $\|v - v_0\|_N$ over $\mathcal{K}_T^{\mathrm{sq}}$; it is $\mathcal{F}_0$-measurable. With $a_T := E(v^\dagger) = \min_{v \in \mathcal{K}_T^{\mathrm{sq}}} E(v)$,*

$$E(v) - a_T \;\ge\; \|v - v^\dagger\|_N^2 \qquad \textit{for all } v \in \mathcal{K}_T^{\mathrm{sq}}. \tag{15}$$

In the well-specified case $v_0 \in \mathcal{K}_T^{\mathrm{sq}}$, in particular under Lemma C.3, $v^\dagger = v_0$ and $a_T = 0$.

Proof. $\mathcal{K}_T^{\mathrm{sq}}$ is nonempty, compact and convex (Lemma B.2), so the $\|\cdot\|_N$-projection of v_0 exists and is unique; it is a measurable function of (X, f_0). The projection is characterized by $\langle v_0 - v^\dagger,\, v - v^\dagger\rangle_N \le 0$ for all $v \in \mathcal{K}_T^{\mathrm{sq}}$, hence $E(v) = \|v - v^\dagger\|_N^2 + \|v^\dagger - v_0\|_N^2 - 2\langle v - v^\dagger,\, v_0 - v^\dagger\rangle_N \ge \|v - v^\dagger\|_N^2 + a_T$. If $v_0 \in \mathcal{K}_T^{\mathrm{sq}}$, the minimizer is v_0 itself. □

Theorem C.5 (Squared master inequality on the same corruption realization)**.** *Assume (A3) and $0 < \eta \le 1$, and let $\delta \in (0, 1)$. On E_δ^{sq}, of conditional probability at least $1 - \delta$ given the design, simultaneously for all $v \in \mathcal{K}_T^{\mathrm{sq}}$,*

$$L_T^{\mathrm{sq}}(v) - L_T^{\mathrm{sq}}(v^\dagger) \;\ge\; \tfrac{1}{2}\big(E(v) - a_T\big) - \Delta_T^{\mathrm{sq}}. \tag{16}$$

Proof. Step 1 (identity). By Lemma C.2 at v and at $v^\dagger$ the term $A^2\hat{\varepsilon}_T$ cancels: $L_T^{\mathrm{sq}}(v) - L_T^{\mathrm{sq}}(v^\dagger) = E(v) - a_T - 2A\,Z(h)$ with $h := v - v^\dagger \in \mathcal{H}_T$.

Step 2 (offset). On E_δ^{sq}, $2A|Z(h)| \le \frac{1}{2}\|h\|_N^2 + \Delta_T^{\mathrm{sq}}$ by (14).

Step 3 (keep the projection information). By (15), $L_T^{\mathrm{sq}}(v) - L_T^{\mathrm{sq}}(v^\dagger) \ge E(v) - a_T - \frac{1}{2}\|v - v^\dagger\|_N^2 - \Delta_T^{\mathrm{sq}} \ge E(v) - a_T - \frac{1}{2}(E(v) - a_T) - \Delta_T^{\mathrm{sq}}$. □

Corollary C.6 (Oracle inequality for the clean error)**.** *On E_δ^{sq}, if $\hat{v} \in \mathcal{K}_T^{\mathrm{sq}}$ satisfies $L_T^{\mathrm{sq}}(\hat{v}) \le L_T^{\mathrm{sq}}(v^\dagger) + \varepsilon_{\mathrm{opt}}$ for some $\varepsilon_{\mathrm{opt}} \ge 0$, then*

$$E(\hat{v}) = \|\hat{v} - v_0\|_{n_T}^2 \;\le\; a_T + 2\Delta_T^{\mathrm{sq}} + 2\varepsilon_{\mathrm{opt}}, \qquad \|\hat{v} - v^\dagger\|_{n_T}^2 \;\le\; 2\Delta_T^{\mathrm{sq}} + 2\varepsilon_{\mathrm{opt}}. \tag{17}$$

Proof. Insert the assumption into (16): $\frac{1}{2}(E(\hat{v}) - a_T) \le \Delta_T^{\mathrm{sq}} + \varepsilon_{\mathrm{opt}}$; the second bound uses (15). □

Why the event covers the trained network. E_δ^{sq} is constructed conditionally on the design and is uniform over the class. It therefore covers a predictor chosen with the help of the corrupted responses, such as $\hat{v} = \mathrm{ev}_T(\hat{f})$, as long as that predictor lies in the class; by Lemma 4 this is the case whenever $\hat{f}$ satisfies (A0)–(A3). The probability bound is not claimed conditionally on the event that the trained network satisfies the assumptions.

C.4 The explicit remainder

Δ_T^{sq} is the remainder Δ of Theorem B.10 at

$$\tau = 2A, \qquad \kappa = 2, \qquad C = C_T^{\mathrm{sq}}, \qquad B = 2M,$$
$$C_H = 2C_T^{\mathrm{sq}}, \qquad S = 2M + 4C_T^{\mathrm{sq}}, \qquad q = \dim \mathcal{A}_T \le m + 1,$$

variance proxy ε_n, sample size N and $u = 1 + \log(1/\delta)$. With $\tau^2/\kappa^2 = A^2$ and $\kappa^{1-2\alpha_m} = 2^{1-2\alpha_m}$ absorbed into c_m, (12) reads

$$\begin{aligned}\Delta_T^{\mathrm{sq}} \;\le\; c_m\Big[\, (C_T^{\mathrm{sq}})^{\beta_m}\Big(\frac{A^2\varepsilon_n L_N}{N}\Big)^{\alpha_m} + \frac{A^2\varepsilon_n\,(qL_N + u)}{N} \\ + A\min\Big\{8M\log 2,\; (2M + 4C_T^{\mathrm{sq}})\,L(2M)\,N^{-\gamma_m} + \frac{2qM\,L(2M)}{N}\Big\} + \frac{AMu}{N}\Big].\end{aligned} \tag{18}$$

Three bounds turn this into the main-text display (5). The minimum is bounded by its second entry. Each logarithm is bounded by $\mathcal{L}_T = 1 + L_N + L(2M)$, evaluated at $\tau = 2A$, $\kappa = 2$, $B = 2M$ and $S = 2M + 4C_T^{\mathrm{sq}}$, so that $L_N^{\alpha_m} \le L_N \le \mathcal{L}_T$ and $L(2M) \le \mathcal{L}_T$. The affine dimension $q \le m + 1$ and the confidence term $u = 1 + \log(1/\delta)$ are collected into the displayed factor $m + \log(1/\delta)$. The amplitude A enters twice, as the multiplier scaling $\tau = 2A$ and through the response envelope $B_0 + A$ inside $\bar{B}_\eta^{\mathrm{sq}}$, hence inside C_T^{sq}.

D Geometric Inputs on the Uniform Ball: Depth, Weights and Design Events

This appendix proves Proposition 6 of §3.4. It has two parts. On the population side, the minimum kink weight over hyperplanes through the deep region of the uniform ball is bounded below by a power of the depth threshold. On the sample side, the half-space deviation of the design transfers this bound to the empirical weight and controls the deep fraction. Throughout, $X_1, \dots, X_n$ are i.i.d. uniform on $\mathbb{B}_1^m$, P denotes their law and $P_{\mathcal{D}}$ the empirical measure. For $Q \in \{P, P_{\mathcal{D}}\}$, $\mathrm{depth}_Q(x) := \inf_{u \in \mathbb{S}^{m-1}} Q\{u^\top(X - x) \ge 0\}$ is the half-space depth, population or empirical, and $\Omega_s(Q) := \{x : \mathrm{depth}_Q(x) \ge s\}$ is the deep region at level s. Let $b_k := \mathrm{vol}(\mathbb{B}_1^k)$ ($b_0 := 1$) and

$$\omega_m := 2 + \frac{2}{m+1} = \frac{2m+4}{m+1}, \qquad \nu_m := \frac{2}{m+1},$$
$$c_m^{\mathrm{cap}} := \frac{2\,b_{m-1}}{(m+1)\,b_m}, \qquad C_m^{\mathrm{cap}} := 2^{(m-1)/2} c_m^{\mathrm{cap}}.$$

D.1 Depth and kink weights on the uniform ball

Weights. For $Q \in \{P, P_{\mathcal{D}}\}$, $u \in \mathbb{S}^{m-1}$, $t \in \mathbb{R}$: $p_Q(u,t) := Q\{u^\top X > t\}$, $s_Q(u,t) := \mathbb{E}_Q(u^\top X - t)_+$, and $c_Q(u,t) := \mathbb{E}_Q[X \mid u^\top X > t]$ when $p_Q(u,t) > 0$, $c_Q(u,t) := 0$ otherwise. The kink weight of §3.2 is

$$\begin{aligned}\widetilde{g}_Q(u,t) &= p_Q(u,t)^2\,\mathbb{E}_Q[u^\top X - t \mid u^\top X > t]\,\sqrt{1 + \|c_Q(u,t)\|^2} \\ &= p_Q(u,t)\,s_Q(u,t)\sqrt{1 + \|c_Q(u,t)\|^2},\end{aligned}$$

with the convention $\widetilde{g}_Q(u,t) := 0$ when $p_Q(u,t) = 0$, under which both sides vanish and the identity holds for all (u,t). As in §3.2, $g_Q(u,t) = \min\{\widetilde{g}_Q(u,t), \widetilde{g}_Q(-u,-t)\}$ is the cheaper orientation, and $g_Q^{\min}(s) := \inf\{g_Q(u,t) : \{u^\top x = t\} \cap \Omega_s(Q) \neq \emptyset\}$. Since c_Q is an average of points of $\mathbb{B}_1^m$ (or 0), $\|c_Q\| \le 1$ and, for all (u,t),

$$p_Q\, s_Q \;\le\; \widetilde{g}_Q \;\le\; \sqrt{2}\, p_Q\, s_Q. \tag{19}$$

On the uniform ball, the depth and the weight factors p_P and s_P reduce to masses of caps.

Lemma D.1 (Caps, depth and shell mass on the uniform ball)**.** *Let* $\mathrm{cap}(h) := P\{X_1 \ge 1-h\}$, $h \in [0,2]$*, the mass of a cap of height* h *(the same for every direction).*

(a) For $h \in [0,1]$,

$$\mathrm{cap}(h) = \frac{b_{m-1}}{b_m} \int_0^h (2s - s^2)^{(m-1)/2}\, ds,$$
$$c_m^{\mathrm{cap}}\, h^{(m+1)/2} \;\le\; \mathrm{cap}(h) \;\le\; C_m^{\mathrm{cap}}\, h^{(m+1)/2};$$

cap *is continuous and strictly increasing on* $[0,1]$ *with* $\mathrm{cap}(1) = \frac{1}{2}$.

(b) For $x \in \mathbb{B}_1^m$, $\mathrm{depth}_P(x) = \mathrm{cap}(1 - \|x\|)$. *Consequently, for* $s \in (0, \frac{1}{2}]$, $\Omega_s(P)$ *is the centered closed ball of radius* $1 - h_s$*, where* $h_s \in (0,1]$ *solves* $\mathrm{cap}(h_s) = s$*, and* $(s/C_m^{\mathrm{cap}})^{\nu_m} \le h_s \le (s/c_m^{\mathrm{cap}})^{\nu_m}$.

(c) For every $s > 0$, $p(s) := P\{\mathrm{depth}_P(X) < s\} \le m\,(s/c_m^{\mathrm{cap}})^{\nu_m}$.

Proof. (a) The section of $\mathbb{B}_1^m$ at $x_1 = 1-s$ is an $(m-1)$-ball of radius $\sqrt{1-(1-s)^2} = \sqrt{2s-s^2}$; integrate its volume $b_{m-1}(2s-s^2)^{(m-1)/2}$ over $s \in (0,h)$ and divide by b_m. For $s \in [0,1]$, $s \le 2s - s^2 \le 2s$, and $\int_0^h s^{(m-1)/2} ds = \frac{2}{m+1} h^{(m+1)/2}$, which gives the lower bound with c_m^{cap} and the upper bound with $C_m^{\mathrm{cap}} = 2^{(m-1)/2} c_m^{\mathrm{cap}}$. Monotonicity and continuity are clear; $\mathrm{cap}(1)$ is the mass of a half-ball.

(b) For a unit vector u, $P\{u^\top(X-x) \ge 0\} = P\{u^\top X \ge u^\top x\} = \mathrm{cap}(1 - u^\top x)$ by rotation invariance; since cap is nondecreasing and $u^\top x \le \|x\|$, the infimum over u is attained at $u = x/\|x\|$ (any u if $x = 0$), giving $\mathrm{depth}_P(x) = \mathrm{cap}(1 - \|x\|)$. Thus $\mathrm{depth}_P(x) \ge s$ iff $1 - \|x\| \ge h_s$; h_s exists and is unique by (a) for $s \in (0, \frac{1}{2}]$, and the two bounds on h_s are (a) evaluated at $h = h_s$.

(c) For $s \le \frac{1}{2}$, $\{\mathrm{depth}_P(X) < s\} = \{\|X\| > 1 - h_s\}$ has mass $1 - (1-h_s)^m \le m h_s \le m(s/c_m^{\mathrm{cap}})^{\nu_m}$ by (b). For $s > \frac{1}{2}$ the right-hand side exceeds $m \ge 1 \ge p(s)$ because $c_m^{\mathrm{cap}} \le \mathrm{cap}(1) = \frac{1}{2} < s$. □

Lemma D.2 (Population weight lower bound)**.** *For* $s \in (0, \frac{1}{2}]$,

$$g_P^{\min}(s) \;\ge\; c_m^w\, s^{\omega_m}, \qquad c_m^w := \frac{2c_m^{\mathrm{cap}}}{m+3}\,(C_m^{\mathrm{cap}})^{-(m+3)/(m+1)} \;\le\; 1.$$

Proof. Let $\{u^\top x = t\}$ meet $\Omega_s(P)$, the ball of radius $r_s := 1 - h_s$ (Lemma D.1(b)); then $|t| \le r_s$. For the orientation (u,t): $p_P(u,t) \ge P\{u^\top X > r_s\} = \mathrm{cap}(h_s) = s$ (the boundary is P-null), and by the layer-cake formula $s_P(u,t) = \int_t^1 p_P(u,z)\, dz$ (as $u^\top X \le 1$),

$$\begin{aligned} s_P(u,t) \;&\ge\; \int_{r_s}^1 P\{u^\top X > z\}\, dz = \int_0^{h_s} \mathrm{cap}(h')\, dh' \;\ge\; c_m^{\mathrm{cap}} \int_0^{h_s} h'^{(m+1)/2} dh' \\ &= \frac{2c_m^{\mathrm{cap}}}{m+3}\, h_s^{(m+3)/2} \;\ge\; \frac{2c_m^{\mathrm{cap}}}{m+3} \Big(\frac{s}{C_m^{\mathrm{cap}}}\Big)^{\frac{m+3}{m+1}}, \end{aligned}$$

using $h_s \ge (s/C_m^{\mathrm{cap}})^{\nu_m}$. By (19), $\widetilde{g}_P(u,t) \ge p_P s_P \ge c_m^w\, s^{1+(m+3)/(m+1)} = c_m^w\, s^{\omega_m}$. The orientation $(-u,-t)$ also satisfies $|-t| \le r_s$, so the same bound holds for it, hence for the minimum $g_P(u,t)$; take the infimum over such (u,t). Finally $c_m^{\mathrm{cap}} \le \mathrm{cap}(1) = \frac{1}{2}$ and $C_m^{\mathrm{cap}} \ge \mathrm{cap}(1) = \frac{1}{2}$ give $c_m^w \le \frac{1}{m+3}\, 2^{(m+3)/(m+1)} \le \frac{4}{m+3} \le 1$. □

D.2 From the population to the sample

Lemma D.3 (One-sided comparison of the empirical weight). *Let $\epsilon_n > 0$ and suppose that*

$$\sup_{u\in\mathbb{S}^{m-1},\, t\in[-1,1]} |P_{\mathcal{D}}\{u^\top X \geq t\} - P\{u^\top X \geq t\}| \leq \epsilon_n. \tag{20}$$

Then $\sup_{u,\, t\in[-1,1]}|p_{\mathcal{D}} - p_P| \leq \epsilon_n$, $\sup_{u,\, t\in[-1,1]}|s_{\mathcal{D}} - s_P| \leq 2\epsilon_n$, *and for all* $u \in \mathbb{S}^{m-1}$, $|t| \leq 1$,

$$g_{\mathcal{D}}(u,t) \;\geq\; \frac{1}{\sqrt{2}}\, g_P(u,t) - 4\epsilon_n. \tag{21}$$

If moreover $T \in (0, \frac{1}{2}]$ *and* $\epsilon_n \leq T/2$, *then* $\Omega_T(\mathcal{D}) \subseteq \Omega_{T/2}(P)$ *and*

$$g^{\min}_{\mathcal{D}}(T) \;\geq\; \frac{1}{\sqrt{2}}\, g^{\min}_P(T/2) - 4\epsilon_n. \tag{22}$$

Proof. Open half-spaces. $p_Q(u,t) = Q\{u^\top X > t\} = 1 - Q\{(-u)^\top X \geq -t\}$, and $(-u,-t)$ ranges over the same index set $\mathbb{S}^{m-1}\times[-1,1]$; hence (20) gives $|p_{\mathcal{D}}(u,t) - p_P(u,t)| \leq \epsilon_n$ directly.

Layer cake. $s_Q(u,t) = \int_t^\infty Q\{u^\top X > z\}\, dz = \int_t^1 p_Q(u,z)\, dz$, so $|s_{\mathcal{D}} - s_P| \leq (1-t)\epsilon_n \leq 2\epsilon_n$.

Products. With $p_{\mathcal{D}} \leq 1$ and $s_P \leq 1 - t \leq 2$, $|p_{\mathcal{D}}s_{\mathcal{D}} - p_P s_P| \leq p_{\mathcal{D}}|s_{\mathcal{D}} - s_P| + s_P|p_{\mathcal{D}} - p_P| \leq 4\epsilon_n$. By (19), $\widetilde{g}_{\mathcal{D}}(u,t) \geq p_{\mathcal{D}}s_{\mathcal{D}} \geq p_P s_P - 4\epsilon_n \geq \widetilde{g}_P(u,t)/\sqrt{2} - 4\epsilon_n$; the same holds at $(-u,-t)$ (also $|-t| \leq 1$), and if $a \geq a' - c$, $b \geq b' - c$ then $\min\{a,b\} \geq \min\{a',b'\} - c$; this is (21).

Inclusion. Every x with $\mathrm{depth}_{\mathcal{D}}(x) \geq T > 0$ lies in the convex hull of the sample, hence in $\mathbb{B}_1^m$, so the thresholds $u^\top x$ lie in $[-1,1]$ and (20) gives, for every u, $P\{u^\top(X-x) \geq 0\} \geq P_{\mathcal{D}}\{u^\top(X-x) \geq 0\} - \epsilon_n \geq T - \epsilon_n \geq T/2$; thus $x \in \Omega_{T/2}(P)$. A hyperplane meeting $\Omega_T(\mathcal{D})$ therefore meets $\Omega_{T/2}(P) \subseteq \mathbb{B}_1^m$, so $|t| \leq 1$ and (21) applies: $g_{\mathcal{D}}(u,t) \geq g_P(u,t)/\sqrt{2} - 4\epsilon_n \geq g_P^{\min}(T/2)/\sqrt{2} - 4\epsilon_n$. Take the infimum. □

Lemma D.4 (Two design events). *Fix $T \in (0, \frac{1}{2}]$ and $\delta \in (0,1)$, and let $\epsilon_n := C_{\mathrm{VC}}\sqrt{(m + \log(1/\delta))/n}$, where C_{VC} is the universal constant of the Vapnik–Chervonenkis uniform deviation inequality for the class of half-spaces $\{x : u^\top x \geq t\}$, $u \in \mathbb{S}^{m-1}$, $t \in [-1,1]$ (VC dimension $m+1$; stated for open half-spaces in Liang et al. (2026, Lemma E.5)). With probability at least $1 - 2\delta$ over the design, simultaneously, the half-space deviation bound* (20) *holds and*

$$\frac{1}{n}\sum_{i\leq n} \mathbf{1}\{\mathrm{depth}_P(X_i) < T + \epsilon_n\} \;\leq\; p(T+\epsilon_n) + \sqrt{\frac{2p(T+\epsilon_n)\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{3n}. \tag{23}$$

Proof. The bound (20) holds with probability at least $1 - \delta$ by the VC inequality; the closed version follows from the open one through the identity $Q\{u^\top X \geq t\} = 1 - Q\{(-u)^\top X > -t\}$, with the same constant and the same probability. The threshold T and the precision ϵ_n are fixed before the design is drawn, so $\{x : \mathrm{depth}_P(x) < T + \epsilon_n\}$ is a fixed set of P-mass $p(T+\epsilon_n)$, and (23) is Bernstein's inequality for the sum of its i.i.d. indicators, with probability at least $1 - \delta$. A union bound combines the two. □

Proof of Proposition 6. Work on the event of Lemma D.4, of probability at least $1 - 2\delta$.

Weight bound. The sample-size condition $n \geq A_m(m + \log(1/\delta))T^{-2\omega_m}$ with $A_m := \big(8\sqrt{2}\, 2^{\omega_m} C_{\mathrm{VC}}/c_m^w\big)^2$ gives $\epsilon_n \leq C_{\mathrm{VC}}T^{\omega_m}/\sqrt{A_m} = c_m^w (T/2)^{\omega_m}/(8\sqrt{2})$, which is at most $T/2$ because $c_m^w \leq 1$, $T/2 \leq 1$ and $\omega_m \geq 2$. Hence (22) applies, and with Lemma D.2 at $s = T/2$,

$$g^{\min}_{\mathcal{D}}(T) \;\geq\; \frac{1}{\sqrt{2}}\, g^{\min}_P(T/2) - 4\epsilon_n \;\geq\; c_m^w \Big(\frac{T}{2}\Big)^{\omega_m}\Big(\frac{1}{\sqrt{2}} - \frac{1}{2\sqrt{2}}\Big) = c'_m\, T^{\omega_m}, \qquad c'_m := \frac{c_m^w\, 2^{-\omega_m}}{2\sqrt{2}}.$$

Deep fraction. Write $\pi_T := n_T/n$ for the deep fraction. If $\mathrm{depth}_P(x) \geq T + \epsilon_n$ then $x \in \mathbb{B}_1^m$ and, by (20), for every u, $P_{\mathcal{D}}\{u^\top(X-x) \geq 0\} \geq P\{u^\top(X-x) \geq 0\} - \epsilon_n \geq T$; so $\Omega_{T+\epsilon_n}(P) \subseteq \Omega_T(\mathcal{D})$, and every sample outside $\Omega_T(\mathcal{D})$ lies in the fixed set counted in (23). Therefore

$$1 - \pi_T \;\leq\; p(T+\epsilon_n) + \sqrt{\frac{2p(T+\epsilon_n)\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{3n}.$$

Since $\epsilon_n \leq T/2$, Lemma D.1(c) gives $p(T+\epsilon_n) \leq p(2T) \leq m(2T/c_m^{\mathrm{cap}})^{\nu_m} =: C_m T^{2/(m+1)}$, and $p(T+\epsilon_n) \leq 1$ inside the square root. □

The constants are not optimized; the bound establishes an explicit polynomial dependence on T, not a sharp numerical estimate.

E SPHERE CONSTRUCTIONS

This appendix proves Proposition 7 and develops the cap network reused in Proposition F.9: one neuron per sample, each active only on its own cap, and the caps are pairwise disjoint. Throughout, $n \geq 2$, $X_1, \ldots, X_n \in \mathbb{S}^{d-1}$ are pairwise distinct and $f_\theta(x) = \sum_{k \leq K} v_k (w_k^\top x - b_k)_+ + \beta$ with the full parameter vector $\theta = ((v_k, w_k, b_k)_{k \leq K}, \beta)$.

E.1 THE CAP LEMMA AND THE CAP NETWORK

Lemma E.1 (Disjoint caps). *Let $\rho := \max_{i \neq j} X_i^\top X_j \in [-1, 1)$ and $\delta := (1-\rho)/4 \in (0, \frac{1}{2}]$. Define $\mathcal{C}_i := \{x \in \mathbb{B}_1^d : \ X_i^\top x > 1 - \delta\}$. Then:*

(i) $X_i \in \mathcal{C}_i$ and $X_j \notin \mathcal{C}_i$ for $j \neq i$; more precisely $X_i^\top X_j - (1-\delta) \leq \rho - 1 + \delta = -3\delta$.
(ii) The caps are pairwise disjoint.
(iii) For every $x \in \mathbb{B}_1^d$, $0 \leq (X_i^\top x - (1-\delta))_+ \leq \delta$.

Proof. (i) $X_i^\top X_i = 1 > 1 - \delta$, and $X_i^\top X_j \leq \rho = 1 - 4\delta < 1 - \delta$. (ii) Suppose $x \in \mathcal{C}_i \cap \mathcal{C}_j$ with $\|x\| \leq 1$. Then $\|x - X_i\|^2 = \|x\|^2 + 1 - 2X_i^\top x < 2 - 2(1-\delta) = 2\delta$, and likewise for j, so $\|X_i - X_j\| < 2\sqrt{2\delta}$ and $\|X_i - X_j\|^2 < 8\delta = 2(1-\rho)$. But $\|X_i - X_j\|^2 = 2 - 2X_i^\top X_j \geq 2 - 2\rho = 2(1-\rho)$, a contradiction. (iii) $X_i^\top x \leq \|x\| \leq 1$. □

The cap network. Given a scale $s > 0$ and output weights $v \in \mathbb{R}^n$, the cap network $f_{s,v} := f_\theta$ has $K = n$, $w_i := sX_i$, $b_i := s(1-\delta)$ and $\beta := 0$, so that

$$f_{s,v}(x) = \sum_{i=1}^{n} v_i \, s \left(X_i^\top x - (1-\delta)\right)_+.$$

All derivatives below are taken in the full, unconstrained parameter space. These relations specify the point of differentiation, not constraints imposed while differentiating; in particular β remains a trainable coordinate.

Lemma E.2 (Values, envelope, derivatives). *For the cap network:*

(i) $f_{s,v}(X_j) = v_j \, s\delta$ for every j; at X_j neuron j has preactivation $s\delta > 0$ and every other neuron has preactivation $\leq -3s\delta < 0$. In particular f_θ is twice continuously differentiable in θ in a neighborhood of the data points, and no ReLU kink sits on a sample.
(ii) $\sup_{x \in \mathbb{B}_1^d} |f_{s,v}(x)| \leq \max_i |v_i| \, s\delta$.
(iii) The Jacobian $J_j := \nabla_\theta f_\theta(X_j)$ is supported on the block of neuron j and on β: $\partial_{v_j} f = s\delta$, $\partial_{w_j} f = v_j X_j$, $\partial_{b_j} f = -v_j$, $\partial_\beta f = 1$, all other entries zero. Writing $J_j = (\mathsf{b}_j, 1)$ with b_j the neuron-j block, $\|\mathsf{b}_j\|^2 = s^2\delta^2 + 2v_j^2$.
(iv) The Hessian $\nabla_\theta^2 f_\theta(X_j)$ is supported on the block of neuron j and equals $\begin{pmatrix} 0 & \xi_j^\top \\ \xi_j & 0 \end{pmatrix}$ in the coordinates $(v_j; (w_j, b_j))$ with $\xi_j := (X_j, -1) \in \mathbb{R}^{d+1}$; its operator norm is $\|\xi_j\| = \sqrt{2}$.

Proof. (i) is Lemma E.1(i); (ii) is Lemma E.1(ii)–(iii): at most one neuron is active at any $x \in \mathbb{B}_1^d$ and its activation is at most δ. (iii), (iv): differentiate $v_j(w_j^\top x - b_j)_+$ at a point where the argument is positive, and use that inactive neurons vanish identically near X_j; β enters linearly, so it has no second derivatives; $\|X_j\| = 1$. □

Lemma E.3 (Gram bound). *Let $\mathsf{A} := [J_1 \ \cdots \ J_n]$ (columns). Then $\lambda_{\max}(\mathsf{A}\mathsf{A}^\top) = \lambda_{\max}(\mathsf{A}^\top\mathsf{A}) \leq n + \max_j \|\mathsf{b}_j\|^2$.*

Proof. $(\mathsf{A}^\top\mathsf{A})_{jk} = \langle J_j, J_k \rangle = \langle \mathsf{b}_j, \mathsf{b}_k \rangle + 1 = \mathbf{1}\{j = k\}\|\mathsf{b}_j\|^2 + 1$ because the blocks of distinct neurons are disjoint and the shared coordinate β contributes 1. Hence $\mathsf{A}^\top\mathsf{A} = \mathbf{1}\mathbf{1}^\top + \mathrm{diag}(\|\mathsf{b}_j\|^2)$ and $\lambda_{\max} \leq \lambda_{\max}(\mathbf{1}\mathbf{1}^\top) + \max_j \|\mathsf{b}_j\|^2$. □

E.2 Squared loss: exact interpolation for every fixed $\eta \in (0,1)$

Proposition 7 is stated in Section 3.4; its proof follows.

Proof of Proposition 7. Let $s > 0$ be such that $s^2\delta^2 = \sqrt{2}D$ and $v_i := y_i/(s\delta)$, and let $f := f_{s,v}$ with parameter θ. By Lemma E.2(i)–(ii), $f(X_i) = v_i s\delta = y_i$ for all i and $\sup_{\mathbb{B}_1^d}|f| \le D$; hence $L_n^{\mathrm{sq}}(\theta) = 0$, θ is a global minimizer and $\nabla L_n^{\mathrm{sq}}(\theta) = 0$. For the Hessian, $\nabla^2 L_n^{\mathrm{sq}} = \frac{2}{n}\sum_j J_j J_j^\top + \frac{2}{n}\sum_j (f(X_j) - y_j)\nabla^2 f(X_j)$ and the second sum vanishes at interpolation. By Lemma E.3, $\lambda_{\max} \le \frac{2}{n}(n + \max_j \|\mathsf{b}_j\|^2)$ and $\|\mathsf{b}_j\|^2 = s^2\delta^2 + 2y_j^2/(s^2\delta^2) \le \sqrt{2}D + 2D^2/(\sqrt{2}D) = 2\sqrt{2}D$, so $\lambda_{\max}(\nabla^2 L_n^{\mathrm{sq}}(\theta)) \le 2 + 4\sqrt{2}D/n$. Positive semidefiniteness holds because θ is a minimizer of a function that is C^2 near θ (Lemma E.2(i)). Finally $2 + 4\sqrt{2}D/n \le 2/\eta$ if and only if $n \ge 2\sqrt{2}D\eta/(1-\eta)$. □

The construction is adapted from the flat-interpolation theorem of Liang et al. (2026), whose squared loss carries the factor $\frac{1}{2n}$ and whose bound reads $\lambda_{\max} \le 1 + (D^2+2)/n$. With the normalization $L_n^{\mathrm{sq}} = \frac{1}{n}\sum_i (f(X_i) - y_i)^2$ used here and the balancing $s^2\delta^2 = \sqrt{2}D$, the bound becomes $2 + 4\sqrt{2}D/n$.

Remark E.4 (The endpoint $\eta = 1$). At $\eta = 1$ the spectral condition reads $\lambda_{\max}(\nabla^2 L_n^{\mathrm{sq}}) \le 2$. It is not vacuous, but it excludes non-constant interpolation in the present smooth setting: $\partial^2 L_n^{\mathrm{sq}}/\partial\beta\,\partial v_k = \frac{2}{n}\sum_i (w_k^\top X_i - b_k)_+$, and if $2I - \nabla^2 L_n^{\mathrm{sq}} \succeq 0$ then its (β,β) entry is 0, so the entire β row of the positive semidefinite matrix $2I - \nabla^2 L_n^{\mathrm{sq}}$ vanishes; hence every neuron is inactive on every sample and the sample predictions all equal the constant β. The proposition therefore cannot be extended to $\eta = 1$.

F Extension: Logistic Loss with Sparse Label Flips

This appendix proves the logistic extension summarized in §3.5. It first establishes the logistic stability budget and the containment of trained networks in the class $\mathcal{K}_T^{\log}$ (§F.2) and defines the mean-flip benchmark (§F.3). It then inserts the offset inequality of Appendix B at $(\tau,\kappa) = (1,\kappa_M)$ to obtain the master inequality and its corollaries (§F.4), makes the remainder explicit (§F.5) and ends with the sphere construction (§F.6). Notation is that of Appendix B ($N := n_T$).

F.1 The logistic model

The teacher $f_0 : \mathbb{B}_1^m \to \mathbb{R}$ generates clean labels $Y_i \mid X_i \sim \mathrm{Bern}(\sigma(f_0(X_i)))$, where $\sigma(u) = 1/(1+e^{-u})$. No regularity or representation assumption on the teacher is used. The logistic case is known to be more delicate than the squared one (Qiao & Wang, 2025). Observed labels are sparse random flips:

$$\widetilde{Y}_i = \begin{cases} Y_i, & B_i = 0, \\ 1 - Y_i, & B_i = 1, \end{cases} \qquad B_i \sim \mathrm{Bern}(\varepsilon_n),\ \varepsilon_n \in (0, \tfrac{1}{2}). \tag{24}$$

The flip indicators $(B_i)_{i\le n}$ are i.i.d. and independent of the whole clean sample $(X_i, Y_i)_{i\le n}$ jointly. The rate ε_n is a fixed number whose subscript only marks the sparse regime. The logistic loss is $\ell(u,y) := \psi(u) - yu$ with $\psi(u) := \log(1+e^u)$ the softplus function. Stability (Definition 1) is taken with respect to the full-sample training loss $L_n^{\log}(\theta) := n^{-1}\sum_{i\le n}\ell(f_\theta(X_i), \widetilde{Y}_i)$, and the deep-region training loss of a vector $v \in \mathbb{R}^N$ is $\widehat{L}_{T,\mathrm{corr}}^{\log}(v) := N^{-1}\sum_{i\in I_T}[\psi(v_i) - \widetilde{Y}_i v_i]$. All probabilities are conditional on $\mathcal{F}_0 = \sigma(X,Y)$: the design and the clean labels are given, and only the flips are random. Throughout, $M > 0$ is an output level fixed before the corruption is seen; on $[-M, M]$ the loss has curvature $\psi'' \ge \kappa_M := \sigma(M)(1-\sigma(M)) > 0$.

F.2 The stability budget and containment

Lemma F.1 (Logistic stability budget bound)**.** *Let f_θ be a two-layer ReLU network* (1) *satisfying (A0)–(A2) for the corrupted-label logistic loss $L_n^{\log}(\theta) = n^{-1}\sum_{i\le n}[\psi(f_\theta(X_i)) - \widetilde{Y}_i f_\theta(X_i)]$. Then*

$$\|f_\theta\|_{\mathrm{path},g} \ \le\ \bar{B}_{\eta,M}^{\log} \ =\ \frac{1}{\kappa_M}\Big(\frac{1}{\eta} + R + 1\Big) - \frac{1}{2}. \tag{25}$$

Proof. Under (A0) the loss is twice differentiable at θ.

Step 1 (Hessian decomposition). With $f_i := f_\theta(X_i)$, $J_i := \nabla_\theta f_i$ and $F_i := \nabla^2_\theta f_i$,

$$\nabla^2_\theta L_n^{\log}(\theta) = \underbrace{\frac{1}{n}\sum_i \psi''(f_i)\, J_i J_i^\top}_{\text{(I)}} + \underbrace{\frac{1}{n}\sum_i \big(\sigma(f_i) - \widetilde{Y}_i\big) F_i}_{\text{(II)}}.$$

Step 2 (lower bound on (I)). By (A2), $|f_i| \le M$ for all $i \le n$, so $\psi''(f_i) \ge \kappa_M = \min_{|u|\le M} \psi''(u)$ and (I) $\succeq \kappa_M G_\theta$ with $G_\theta := n^{-1}\sum_i J_i J_i^\top$.

Step 3 (uniform bound on (II)). $|\sigma(f_i) - \widetilde{Y}_i| \le 1$. For a two-layer ReLU network the only nonzero second derivatives are the (v_k, w_k) and (v_k, b_k) cross terms, so F_i is block diagonal over neurons with blocks of operator norm $\sqrt{\|X_i\|^2 + 1} \le R + 1$; hence $\|\text{(II)}\|_{\mathrm{op}} \le 2(R+1)$ (we keep $2(R+1)$ rather than $R+1$ for a cleaner constant).

Step 4 (stability gives a Jacobian bound). By Weyl's inequality and (A1), $\kappa_M \lambda_{\max}(G_\theta) \le 2/\eta + 2(R+1)$.

Step 5 (Jacobian bound gives a path-norm bound). The Jacobian inequality $\lambda_{\max}(G_\theta) \ge 1 + 2\|f_\theta\|_{\mathrm{path},g}$ of Step 5 in the proof of Proposition C.1, which involves G_θ only, and Step 4 give $\|f_\theta\|_{\mathrm{path},g} \le \frac{1}{2}\big((2/\eta + 2(R+1))/\kappa_M - 1\big) = \bar{B}^{\log}_{\eta,M}$. □

The proof uses the labels only through $|\sigma(f_i) - \widetilde{Y}_i| \le 1$; the spectral constraint (A1) is imposed on the corrupted-label loss used for training. The deep-region budget is $C_T^{\log} := \bar{B}^{\log}_{\eta,M} / g^{\min}_{\mathcal{D}}(T)$ and the class is $\mathcal{K}_T^{\log} := \mathcal{K}_T(C_T^{\log}, M)$; both are determined by the design and the model parameters and do not depend on the flips.

Lemma F.2 (Containment, logistic)**.** *Let f_θ satisfy (A0)–(A3) for the corrupted-label logistic loss. Then* $\mathrm{ev}_T(f_\theta) \in \mathcal{K}_T^{\log}$.

Proof. Lemma F.1 gives the budget (25). The proof of Lemma 4 in Appendix B.1 goes through with this budget in place of (3): together with (A2) and (A3) it gives $\mathrm{ev}_T(f_\theta) \in \mathcal{K}_T(C_T^{\log}, M) = \mathcal{K}_T^{\log}$. □

F.3 The mean-flip benchmark

The mean-flip objective. The *mean-flip objective* uses the soft labels $Y_i^{(\varepsilon)} := \varepsilon_n + (1 - 2\varepsilon_n) Y_i$: for $v \in \mathbb{R}^{n_T}$,

$$\widehat{L}^{\log}_{T,\varepsilon}(v) := \frac{1}{n_T}\sum_{i \in I_T} \big[\psi(v_i) - Y_i^{(\varepsilon)} v_i\big]. \tag{26}$$

Because the logistic loss is affine in the label, $\mathbb{E}[\widehat{L}^{\log}_{T,\mathrm{corr}}(v) \mid X, Y] = \widehat{L}^{\log}_{T,\varepsilon}(v)$: the mean-flip objective is the conditional expectation of the corrupted one given the clean sample.

Definition F.3 (Mean-flip benchmark)**.** $v^* := \arg\min_{v \in \mathcal{K}_T^{\log}} \widehat{L}^{\log}_{T,\varepsilon}(v)$.

The benchmark is determined by the design and the clean labels and does not depend on the realized flips. It is defined within the class, which is why no regularity of the teacher is needed. Two facts drive the proof: the corrupted loss is the mean-flip objective minus a linear multiplier term (Lemma F.4), and the benchmark satisfies a first-order condition with curvature κ_M (Lemma F.5).

Lemma F.4 (Flip linearization on vectors)**.** *For every $v \in \mathbb{R}^N$, $\widehat{L}^{\log}_{T,\mathrm{corr}}(v) = \widehat{L}^{\log}_{T,\varepsilon}(v) - Z(v)$, with Z the multiplier process of Appendix B for $\zeta_i = (B_i - \varepsilon_n)(1 - 2Y_i)$.*

Proof. $\widetilde{Y}_i - Y_i^{(\varepsilon)} = (B_i - \varepsilon_n)(1 - 2Y_i) = \zeta_i$, and the loss is affine in the label: $\psi(v_i) - \widetilde{Y}_i v_i = \psi(v_i) - Y_i^{(\varepsilon)} v_i - \zeta_i v_i$. Average over I_T. □

Lemma F.5 (Benchmark: existence, uniqueness, first-order condition, curvature)**.** *$\widehat{L}^{\log}_{T,\varepsilon}$ has a unique minimizer v^* over $\mathcal{K}_T^{\log}$, which is $\mathcal{F}_0$-measurable, and for every $v \in \mathcal{K}_T^{\log}$*

$$\frac{1}{N}\sum_{i \in I_T} \big(\sigma(v_i^*) - Y_i^{(\varepsilon)}\big)(v_i - v_i^*) \;\ge\; 0, \qquad \widehat{L}^{\log}_{T,\varepsilon}(v) - \widehat{L}^{\log}_{T,\varepsilon}(v^*) \;\ge\; \frac{\kappa_M}{2}\|v - v^*\|_N^2. \tag{27}$$

Proof. $\widehat{L}^{\log}_{T,\varepsilon}$ is C^2 on $\mathbb{R}^N$ with Hessian $N^{-1}\operatorname{diag}(\psi''(v_i)) \succ 0$, hence strictly convex; $\mathcal{K}^{\log}_T$ is nonempty, compact and convex (Lemma B.2), so the minimizer exists and is unique; it is $\mathcal{F}_0$-measurable because $\widehat{L}^{\log}_{T,\varepsilon}$ and $\mathcal{K}^{\log}_T$ are. The variational inequality in (27) is the first-order optimality condition on the convex set. For the curvature bound, expand coordinatewise: $\psi(v_i) - \psi(v_i^*) - \sigma(v_i^*)(v_i - v_i^*) = \int_0^1 (1-t)\,\psi''\big(v_i^* + t(v_i - v_i^*)\big)\,dt \cdot (v_i - v_i^*)^2 \geq \frac{\kappa_M}{2}(v_i - v_i^*)^2$, because $v_i, v_i^* \in [-M, M]$ places the whole segment in $[-M, M]$, where $\psi'' \geq \kappa_M$. Averaging and adding the nonnegative first-order term gives the claim. □

F.4 The master inequality and its consequences

The multiplier process is $Z(h) = N^{-1}\sum_{i\in I_T} \zeta_i h_i$ with $\zeta_i = (B_i - \varepsilon_n)(1 - 2Y_i)$, which satisfies Assumption B.3 with $\mathbb{E}\zeta_i^2 = \varepsilon_n(1-\varepsilon_n) \leq \varepsilon_n$. The event E_δ is the event of Theorem B.10 for $\mathcal{H}_T = \mathcal{K}^{\log}_T - \mathcal{K}^{\log}_T$ at $(\tau, \kappa) = (1, \kappa_M)$, with the $\mathcal{F}_0$-measurable radius of Proposition B.11. It has conditional probability at least $1 - \delta$, and on it

$$|Z(h)| \;\leq\; \frac{\kappa_M}{4}\|h\|_N^2 + \Delta^{\log}_T \qquad \text{for all } h \in \mathcal{H}_T. \tag{28}$$

Theorem F.6 (Logistic master inequality on the same corruption realization)**.** *Assume (A3) and let $\delta \in (0,1)$. On the event E_δ, of conditional probability at least $1-\delta$ given (X, Y), simultaneously for all $v \in \mathcal{K}^{\log}_T$,*

$$\widehat{L}^{\log}_{T,\mathrm{corr}}(v) - \widehat{L}^{\log}_{T,\mathrm{corr}}(v^*) \;\geq\; \frac{\kappa_M}{4}\|v - v^*\|_{n_T}^2 - \Delta^{\log}_T. \tag{29}$$

Proof. Step 1 (linearize). By Lemma F.4 at v and at v^*, $\widehat{L}^{\log}_{T,\mathrm{corr}}(v) - \widehat{L}^{\log}_{T,\mathrm{corr}}(v^*) = \big[\widehat{L}^{\log}_{T,\varepsilon}(v) - \widehat{L}^{\log}_{T,\varepsilon}(v^*)\big] - Z(v - v^*)$; the fixed term $Z(v^*)$ has canceled.

Step 2 (curvature). By Lemma F.5, $\widehat{L}^{\log}_{T,\varepsilon}(v) - \widehat{L}^{\log}_{T,\varepsilon}(v^*) \geq \frac{\kappa_M}{2}\|h\|_N^2$ with $h := v - v^* \in \mathcal{H}_T$.

Step 3 (offset). On E_δ, $|Z(h)| \leq \frac{\kappa_M}{4}\|h\|_N^2 + \Delta^{\log}_T$ by (28).

Step 4 (assemble). On E_δ: $\widehat{L}^{\log}_{T,\mathrm{corr}}(v) - \widehat{L}^{\log}_{T,\mathrm{corr}}(v^*) \geq \frac{\kappa_M}{2}\|h\|_N^2 - \frac{\kappa_M}{4}\|h\|_N^2 - \Delta^{\log}_T$. □

Corollary F.7 (Adaptive gain)**.** *On E_δ, for every $v \in \mathcal{K}^{\log}_T$, and in particular for $\hat{v} = \mathrm{ev}_T(\hat{f})$ whenever $\hat{f}$ satisfies (A0)–(A3),*

$$\widehat{L}^{\log}_{T,\mathrm{corr}}(v^*) - \widehat{L}^{\log}_{T,\mathrm{corr}}(v) \;\leq\; \Delta^{\log}_T.$$

Proof. Drop the nonnegative quadratic term in (29); membership of $\mathrm{ev}_T(\hat{f})$ in $\mathcal{K}^{\log}_T$ is Lemma F.2. □

Below, $a \lesssim_{\log,m} b$ means $a \leq c_m \mathcal{L}_T\, b$, with c_m depending only on m and $\mathcal{L}_T$ the logarithmic factor of Appendix B.4 evaluated at the logistic substitutions of §F.5.

Corollary F.8 (Proximity to the benchmark)**.** *On E_δ, if $\hat{v} \in \mathcal{K}^{\log}_T$ satisfies $\widehat{L}^{\log}_{T,\mathrm{corr}}(\hat{v}) \leq \widehat{L}^{\log}_{T,\mathrm{corr}}(v^*) + \varepsilon_{\mathrm{opt}}$ for some $\varepsilon_{\mathrm{opt}} \geq 0$, then*

$$\|\hat{v} - v^*\|_{n_T}^2 \;\leq\; 4\Big(\rho_T^2 + \frac{\varepsilon_{\mathrm{opt}}}{\kappa_M}\Big), \qquad \rho_T^2 := \frac{\Delta^{\log}_T}{\kappa_M},$$

$$\rho_T^2 \;\lesssim_{\log,m}\; \frac{1}{\kappa_M^2}\, D_T^{\beta_m}\Big(\frac{\varepsilon_n}{n_T}\Big)^{\alpha_m} + \frac{1}{\kappa_M}\big(\textit{lower-order terms of } (32)\big).$$

Proof. Insert the assumption $\widehat{L}^{\log}_{T,\mathrm{corr}}(\hat{v}) \leq \widehat{L}^{\log}_{T,\mathrm{corr}}(v^*) + \varepsilon_{\mathrm{opt}}$ into the left side of (29) and rearrange: $\frac{\kappa_M}{4}\|\hat{v} - v^*\|_N^2 \leq \Delta^{\log}_T + \varepsilon_{\mathrm{opt}}$. The displayed bound on ρ_T^2 is (32) divided by κ_M. □

Why the event covers the trained network. E_δ is constructed conditionally on the clean sample and is uniform over the class. It therefore covers a predictor selected with the help of the corrupted labels, such as $\hat{v} = \mathrm{ev}_T(\hat{f})$, as long as that predictor lies in the class; by Lemma F.2 this is the case whenever $\hat{f}$ satisfies (A0)–(A3). The probability bound is not claimed conditionally on the event that the trained network satisfies the assumptions. Where Corollary F.8 requires near-optimality, it remains an assumption on $\hat{v}$: full-sample gradient descent does not automatically deliver deep-region near-optimality against a benchmark that is not observable.

The envelope alone does not produce the decay. The sample envelope enters the proof at several places: the budget $\bar{B}^{\log}_{\eta,M}$, the curvature κ_M, the box of $\mathcal{K}^{\log}_T$ and the increment bounds of the multiplier process. By itself, together with the optimality of v^* on the class, it gives only the following. For $v \in \mathcal{K}^{\log}_T$, Lemma F.4 at v and v^* and $\widehat{L}^{\log}_{T,\varepsilon}(v^*) \le \widehat{L}^{\log}_{T,\varepsilon}(v)$ give $\widehat{L}^{\log}_{T,\mathrm{corr}}(v^*) - \widehat{L}^{\log}_{T,\mathrm{corr}}(v) \le Z(v - v^*)$ and $\widehat{L}^{\log}_{T,\varepsilon}(v^*) - \widehat{L}^{\log}_{T,\mathrm{corr}}(v) \le Z(v)$; since $|v_i - v_i^*| \le 2M$, $|v_i| \le M$ and $|\zeta_i| = |B_i - \varepsilon_n|$,

$$\widehat{L}^{\log}_{T,\mathrm{corr}}(v^*) - \widehat{L}^{\log}_{T,\mathrm{corr}}(v) \;\le\; \frac{2M}{N}\sum_{i\in I_T}|B_i - \varepsilon_n|, \qquad \widehat{L}^{\log}_{T,\varepsilon}(v^*) - \widehat{L}^{\log}_{T,\mathrm{corr}}(v) \;\le\; \frac{M}{N}\sum_{i\in I_T}|B_i - \varepsilon_n|, \tag{30}$$

whose right-hand sides have conditional expectations $4M\varepsilon_n(1-\varepsilon_n)$ and $2M\varepsilon_n(1-\varepsilon_n)$ given (X, Y), because $\mathbb{E}[|B_i - \varepsilon_n| \mid X, Y] = 2\varepsilon_n(1-\varepsilon_n)$. At fixed M and ε_n these upper bounds do not vanish as N grows. The decay $\Delta^{\log}_T = \widetilde{O}_m(N^{-\alpha_m})$ of Theorem F.6 uses, in addition, the complexity of the difference class $\mathcal{H}_T$, which the stability budget and the deep-region geometry provide. Proposition F.9 gives nonvanishing gains under a fixed envelope when the deep-region guarantee is unavailable; its comparison with (30) is in conditional expectation.

F.5 The explicit remainder

$\Delta^{\log}_T$ is the remainder Δ of Theorem B.10 at

$$\tau = 1, \qquad \kappa = \kappa_M, \qquad C = C^{\log}_T, \qquad B = 2M,$$
$$C_H = 2C^{\log}_T, \qquad S = 2M + 4C^{\log}_T, \qquad q = \dim \mathcal{A}_T \le m + 1,$$

variance proxy ε_n, sample size N and $u = 1 + \log(1/\delta)$. Inserting these values into (12) (with L_N and $L(\cdot)$ as in Appendix B.3 and the constants absorbed into c_m),

$$\Delta^{\log}_T \;\le\; c_m\Big[\,\kappa_M\,(C^{\log}_T)^{\beta_m}\Big(\frac{\varepsilon_n L_N}{\kappa_M^2 N}\Big)^{\alpha_m} + \frac{\varepsilon_n\,(qL_N + u)}{\kappa_M N} + \min\Big\{8M\log 2,\; (2M + 4C^{\log}_T)\,L(2M)\,N^{-\gamma_m} + \frac{2qM\,L(2M)}{N}\Big\} + \frac{2Mu}{N}\Big]. \tag{31}$$

Its first term carries the curvature power $\kappa_M^{1-2\alpha_m}$ of the fixed class. Let $D_T := (1/\eta + R + 1)/g^{\min}_{\mathcal{D}}(T)$ be the curvature-free part of the budget, so that $C^{\log}_T \le \kappa_M^{-1} D_T$. Four bounds turn (31) into the form used in Corollary F.8. In the first term and in the minimum, $C^{\log}_T$ is bounded by $\kappa_M^{-1} D_T$. With $2\alpha_m + \beta_m = 2$, the curvature power of the first term becomes $\kappa_M^{1-\beta_m-2\alpha_m} = \kappa_M^{-1}$. The minimum is bounded by its second entry. Each logarithm is bounded by $\mathcal{L}_T = 1 + L_N + L(2M)$. The affine dimension $q \le m + 1$ and the confidence term $u = 1 + \log(1/\delta)$ are collected into the factor $m + \log(1/\delta)$. Together, these bounds give

$$\Delta^{\log}_T \;\le\; c_m\,\mathcal{L}_T\Big[\frac{1}{\kappa_M}\,D_T^{\beta_m}\Big(\frac{\varepsilon_n}{n_T}\Big)^{\alpha_m} + \Big(M + \frac{D_T}{\kappa_M}\Big)n_T^{-\gamma_m} + \Big(\frac{\varepsilon_n}{\kappa_M} + M\Big)\frac{m + \log(1/\delta)}{n_T}\Big], \tag{32}$$

with c_m depending on m only; $\mathcal{L}_T$ depends on N, M and $C^{\log}_T$ but not on δ.

How to read the rate. At fixed class and model parameters the leading term carries the factor $\varepsilon_n^{\alpha_m}$, because the sparse multipliers drive the process, and it decays as $n_T^{-\alpha_m}$ up to logarithmic factors. Substituting the stability budget gives the displayed dependence $D_T^{\beta_m}/\kappa_M$ of this leading term: a smaller step size or a smaller minimum weight weakens it as $\eta^{-\beta_m}$ and $g^{\min}_{\mathcal{D}}(T)^{-\beta_m}$, while the algebraic dependence of the remaining terms on D_T is at most linear. The prediction-error remainder $\rho_T^2 = \Delta^{\log}_T/\kappa_M$ of Corollary F.8 incurs one more factor $1/\kappa_M$, so its leading term is $\kappa_M^{-2} D_T^{\beta_m}(\varepsilon_n/n_T)^{\alpha_m}$ up to logarithms.

F.6 SPHERE CONSTRUCTION FOR THE LOGISTIC LOSS

On the full sample, $\widehat{L}^{\log}_{n,\mathrm{corr}}(v) = n^{-1}\sum_i[\psi(v_i) - \widetilde{Y}_i v_i]$ and $\widehat{L}^{\log}_{n,\varepsilon}(v) = n^{-1}\sum_i[\psi(v_i) - Y_i^{(\varepsilon)} v_i]$ for a prediction vector $v \in \mathbb{R}^n$, with $\widetilde{Y}_i = Y_i + B_i(1-2Y_i)$, $Y_i^{(\varepsilon)} = \varepsilon + (1-2\varepsilon)Y_i$ and $\hat{\varepsilon}_n = n^{-1}\sum_i B_i$. For any pairwise distinct sphere design each sample has empirical depth exactly $1/n$, so $n_T = 0$ whenever $T > 1/n$ and the deep-region benchmark is undefined; the comparator is the box $[-M, M]^n$ of the full sample. The construction is the cap network of Appendix E with output weights $\pm M/(s\delta)$.

Proposition F.9 (Sphere, logistic loss: memorization under a fixed envelope and spectral constraint)**.** *Fix $\varepsilon \in (0, \frac{1}{2})$, $0 < M \le \log\frac{1-\varepsilon}{\varepsilon}$, $\eta \in (0, 2/\kappa_M)$ with $\kappa_M = \sigma(M)(1-\sigma(M))$, and $n \ge (2\sqrt{2}M\kappa_M + \sqrt{2})/(2/\eta - \kappa_M)$. For $n \ge 2$ pairwise distinct $X_1, \dots, X_n \in \mathbb{S}^{d-1}$ there are two width-n networks f_{clean} and f_{corr} whose sample logits are $\pm M$ with the signs of the clean and of the corrupted labels respectively; both have global envelope M and both are η-stable for the corrupted-label logistic loss. The sample vector of f_{clean} minimizes the mean-flip objective over the box $[-M, M]^n$, and, with $\hat{\varepsilon}_n$ the realized flip rate,*

$$\widehat{L}^{\log}_{n,\mathrm{corr}}(f_{\mathrm{clean}}) - \widehat{L}^{\log}_{n,\mathrm{corr}}(f_{\mathrm{corr}}) = M\hat{\varepsilon}_n, \qquad \widehat{L}^{\log}_{n,\varepsilon}(f_{\mathrm{clean}}) - \widehat{L}^{\log}_{n,\mathrm{corr}}(f_{\mathrm{corr}}) = M\varepsilon.$$

Proof of Proposition F.9. Choose $s > 0$ with $s^2\delta^2 = \sqrt{2}M$ and, for a sign vector $\varsigma \in \{-1, 1\}^n$, let $f_\varsigma := f_{s,v}$ with $v_i := \varsigma_i M/(s\delta)$; put $f_{\mathrm{clean}} := f_{2Y-1}$ and $f_{\mathrm{corr}} := f_{2\widetilde{Y}-1}$.

Step 1 (values and envelope). Lemma E.2(i)–(ii) with $|v_i|s\delta = M$ give $f_\varsigma(X_i) = \varsigma_i M$ and $\sup_{\mathbb{B}_1^d}|f_\varsigma| \le M$ for every ς.

Step 2 (Hessian). At θ_ς every logit equals $\pm M$, so $\psi''(f(X_i)) = \kappa_M$ for all i. For any label vector $\tilde{y} \in \{0, 1\}^n$ the logistic Hessian decomposes as

$$\nabla^2_\theta L^{\log}_n(\theta_\varsigma; \tilde{y}) = \frac{1}{n}\sum_j \psi''(f(X_j))\, J_j J_j^\top + \frac{1}{n}\sum_j \big(\sigma(f(X_j)) - \tilde{y}_j\big)\nabla^2_\theta f(X_j) = \frac{\kappa_M}{n}\mathsf{A}\mathsf{A}^\top + \mathsf{R}.$$

By Lemma E.3 and $\|\mathsf{b}_j\|^2 = s^2\delta^2 + 2M^2/(s^2\delta^2) = 2\sqrt{2}M$, $\lambda_{\max}(\frac{\kappa_M}{n}\mathsf{A}\mathsf{A}^\top) \le \kappa_M(1 + 2\sqrt{2}M/n)$. By Lemma E.2(iv), R is block diagonal over the neuron blocks (and zero on the β coordinate) with blocks $\frac{1}{n}(\sigma(f(X_j)) - \tilde{y}_j)\begin{pmatrix} 0 & \xi_j^\top \\ \xi_j & 0 \end{pmatrix}$, each of operator norm at most $\frac{\sqrt{2}}{n}|\sigma(f(X_j)) - \tilde{y}_j| \le \frac{\sqrt{2}}{n}$; hence $\|\mathsf{R}\| \le \sqrt{2}/n$ and

$$\lambda_{\max}\big(\nabla^2_\theta L^{\log}_n(\theta_\varsigma; \tilde{y})\big) \le \kappa_M\Big(1 + \frac{2\sqrt{2}M}{n}\Big) + \frac{\sqrt{2}}{n} \le \frac{2}{\eta},$$

the last inequality being equivalent to $n(2/\eta - \kappa_M) \ge 2\sqrt{2}M\kappa_M + \sqrt{2}$. The bound is uniform in $\tilde{y}$ and in ς; in particular f_{clean} and f_{corr} are both η-stable for the corrupted-label loss.

Step 3 (box minimizer of the mean-flip objective). $\widehat{L}^{\log}_{n,\varepsilon}$ is a sum of coordinatewise terms $\phi_i(u) = \psi(u) - Y_i^{(\varepsilon)}u$ with $\phi_i'(u) = \sigma(u) - Y_i^{(\varepsilon)}$. If $Y_i = 1$ then $Y_i^{(\varepsilon)} = 1 - \varepsilon$ and, for $u \in [-M, M]$, $\phi_i'(u) \le \sigma(M) - (1-\varepsilon) \le 0$ because $M \le \log\frac{1-\varepsilon}{\varepsilon}$ is equivalent to $\sigma(M) \le 1 - \varepsilon$; so ϕ_i is nonincreasing on the box and is minimized at $u = M = f_{\mathrm{clean}}(X_i)$. If $Y_i = 0$ the symmetric argument gives the minimizer $-M$. Hence the sample vector $M(2Y - 1)$ of f_{clean} minimizes $\widehat{L}^{\log}_{n,\varepsilon}$ over $[-M, M]^n$, and therefore over every class that contains it and is contained in the box.

Step 4 (gains). For f_{corr} every logit has the sign of its corrupted label, so each coordinate contributes $\psi(M) - M = \psi(-M) = \log(1 + e^{-M})$; thus $\widehat{L}^{\log}_{n,\mathrm{corr}}(f_{\mathrm{corr}}) = \log(1 + e^{-M}) > 0$. For f_{clean} the unflipped coordinates contribute the same amount, while a flipped coordinate contributes $\psi(M)$ (if $\widetilde{Y}_i = 0$) or $\psi(-M) + M$ (if $\widetilde{Y}_i = 1$), both equal to $\log(1 + e^{-M}) + M$; summing gives $\widehat{L}^{\log}_{n,\mathrm{corr}}(f_{\mathrm{clean}}) - \widehat{L}^{\log}_{n,\mathrm{corr}}(f_{\mathrm{corr}}) = M\hat{\varepsilon}_n$. For the averaged gain, each coordinate of $\widehat{L}^{\log}_{n,\varepsilon}(f_{\mathrm{clean}})$ equals $\psi(M) - (1-\varepsilon)M = \log(1 + e^{-M}) + \varepsilon M$, whence $\widehat{L}^{\log}_{n,\varepsilon}(f_{\mathrm{clean}}) - \widehat{L}^{\log}_{n,\mathrm{corr}}(f_{\mathrm{corr}}) = M\varepsilon$. □

These networks are not stationary points of the unconstrained training loss, since $\partial_{v_j} L^{\log}_n = \frac{s\delta}{n}\big(\sigma(f(X_j)) - \widetilde{Y}_j\big) \ne 0$; the word "minimum" is reserved for the squared construction of Proposition 7. f_{clean} and f_{corr} satisfy the same output envelope, the same logit curvature κ_M at every sample

and a common spectral upper bound; their gain $M\hat{\varepsilon}_n$ comes from adapting the signs to the realized flips. In conditional expectation given (X, Y) both gains of the proposition equal $M\varepsilon$. Up to constant factors this matches the envelope-only bounds (30) with n in place of n_T, whose right-hand sides have conditional expectations $4M\varepsilon(1-\varepsilon)$ and $2M\varepsilon(1-\varepsilon)$.

G Experiment Details

All experiments of Section 4 run on one pipeline, which computes deep regions by exact depth and benchmarks and remainders by convex programs whose values carry a certified interval. The spectral condition has verified bounds for the one-dimensional designs with $n \le 2000$; elsewhere the spectral values of trained networks are diagnostics. Networks are trained once and stored; every evaluation re-reads the stored parameters in double precision, so no number below depends on retraining.

G.1 Design, teachers and corruption

The main grid uses a one-dimensional design: $x_1, \dots, x_n$ i.i.d. uniform on $[-1, 1]$, zero-padded to $\mathbb{R}^{50}$, one design per n. The theorems are applied in the intrinsic coordinate: the network restricted to the line has the same sample outputs, and by the block structure of §G.4 the same largest Hessian eigenvalue. With $m = 1$ the half-space depth of the i-th order statistic is $\min\{i, n+1-i\}/n$, so I_T consists of the ranks $\lceil nT \rceil, \dots, n+1-\lceil nT \rceil$ exactly, Ω_T is the interval between the two extreme deep points, and the minimum weight $g_{\mathcal{D}}^{\min}(T)$ of Section 3.2 has a closed form on the line.

Logistic model: $f_0(x) = 1.5\sin(2x_1)$ and clean labels $Y_i \sim \mathrm{Bernoulli}(\sigma(f_0(X_i)))$, drawn once from a random stream independent of the design (the theorems condition on the clean sample); flips $B_i \sim \mathrm{Bernoulli}(\varepsilon_n)$ with $\varepsilon_n \in \{0.05, 0.1\}$ and a separate mask for each rate. Squared model: $f_0(x) = (x_1 + 0.8)_+ - (x_1)_+ + (x_1 - 0.8)_+$, three kinks of unit amplitude with $B_0 = \max|f_0| = 1$; responses $f_0(X_i) + AB_iS_i$ with $A \in \{1, 0.5\}$, $\varepsilon_n = 0.05$ and one mask (B, S) shared by both amplitudes. Configurations: 60 masks at each $n \in \{500, 1000, 2000, 4000, 8000\}$ with $T = 0.1$; the 240 networks trained at $n = 2000$ are re-evaluated at the thresholds $T \in \{0.05, 0.2, 0.3\}$, since neither the design nor the masks depend on T.

The budgets (3) and (25) are evaluated from η, $M = 2$, B_0, A, $\max_i|x_i|$ and $g_{\mathcal{D}}^{\min}(T)$. Membership $\hat{v} \in \mathcal{K}_T$ is the pair of checks "slope variation of the deep sample vector $\le C_T$" and "$\max_{i \in I_T}|\hat{f}(X_i)| \le M$". The full-sample envelope (A2) is recorded separately.

G.2 Training

Two-layer ReLU networks of width 1024 (Section 2), one random initialization per design shared by all networks of a configuration, full-batch gradient descent for 20,000 epochs with gradient-norm clipping at 50, at $\eta = 1$ for the logistic loss and $\eta = 0.2$ for the squared loss in the normalization of Section 2. The endpoint is the final parameter; nothing during training refers to M, and the envelope (A2) is assessed afterwards. Training runs in single precision; all evaluation is in double precision.

G.3 Evaluation and certified convex programs

For each stored network and threshold the evaluation records the deep sample vector $\hat{v}$, the deep losses, $\hat{\varepsilon}_T$, the clean error $E(\hat{v}) = \|\hat{v} - v_0\|_N^2$, the deficit $A^2\hat{\varepsilon}_T + \frac{1}{2}E(\hat{v}) - L_T^{\mathrm{sq}}(\hat{v})$ (squared) or the gain $G_T(\hat{v}) = \widehat{L}_{T,\mathrm{corr}}^{\log}(v^*) - \widehat{L}_{T,\mathrm{corr}}^{\log}(\hat{v})$ (logistic), the envelope, and the membership checks, in the notation $N = n_T$, $\|\cdot\|_N$ of Appendix B. The deep vector is the double-precision output of the stored network together with a rounding bound on each coordinate, and every check below holds for all vectors within that bound, hence for the real-arithmetic output of the stored network.

The class on the line. Order the deep inputs as $x_1 < \cdots < x_N$, let $s_j := (v_{j+1} - v_j)/(x_{j+1} - x_j)$ and let $D \in \mathbb{R}^{(N-2)\times N}$ map v to its slope changes $(s_{j+1} - s_j)_{j \le N-2}$ (for $N \le 2$ it has no rows). Then

$$\mathcal{K}_T(C, M) = \{v \in \mathbb{R}^N : \ \|Dv\|_1 \le C, \ \|v\|_\infty \le M\}.$$

The discrete slopes of a unit atom are monotone with total change at most 1, so $\|Dv\|_1 \le C$ on $V_T(C)$ by the triangle inequality, and D vanishes on affine vectors; conversely, the piecewise-linear interpolant of v is an affine function plus $\sum_{j=2}^{N-1}(s_j - s_{j-1})(x - x_j)_+$, a combination of unit atoms with total coefficient $\|Dv\|_1$. The two checks of §G.1 therefore decide membership exactly.

Programs and brackets. The benchmarks and remainders are programs over this set: the projection benchmark $v^\dagger$ with a_T (equal to 0 when $v_0 \in \mathcal{K}_T$, which holds for the squared teacher above), the mean-flip benchmark v^*, the corrupted empirical risk minimizer over $\mathcal{K}_T^{\log}(C, M)$ and hence $G_{\max} = \widehat{L}^{\log}_{T,\mathrm{corr}}(v^*) - \min_{\mathcal{K}} \widehat{L}^{\log}_{T,\mathrm{corr}}$, and the realized remainder $\Delta_{\mathrm{emp}}(C)$ below. Each is $\min_{v\in\mathcal{K}} F(v)$ for a separable convex $F(v) = N^{-1}\sum_i \ell_i(v_i)$ over $\mathcal{K} = \mathcal{K}_T(C, M)$, or over $\mathcal{K}_T(2C, 2M)$ for the remainder, in which case C and M are doubled in the bound below. For any $y \in \mathbb{R}^{N-2}$ and any anchor $t \in \mathbb{R}^N$, convexity of the ℓ_i and $y^\top Dv \le C\|y\|_\infty$ on $\mathcal{K}$ give

$$\min_{\mathcal{K}} F \;\ge\; \sum_{i\le N}\Big[\frac{\ell_i(t_i) - \ell_i'(t_i)\,t_i}{N} - M\,\Big|(D^\top y)_i + \frac{\ell_i'(t_i)}{N}\Big|\Big] - C\|y\|_\infty,$$

which is the Fenchel dual value when t_i minimizes $\ell_i(t)/N + (D^\top y)_i t$ over $[-M, M]$, and every $v \in \mathcal{K}$ gives $\min_{\mathcal{K}} F \le F(v)$. A primal–dual interior-point method supplies v and y; a separate verifier evaluates both bounds and the constraints $\|Dv\|_1 \le C$, $\|v\|_\infty \le M$ in 120-bit interval arithmetic, with D rebuilt from the nodes, rounds every end outward to double precision, and accepts a bracket only when v is certified feasible. Solver tolerances target a contribution below 10^{-4} to each reported quantity, and the achieved widths are carried into every comparison.

The realized remainder. With $\mathcal{H} = \mathcal{K}_T(2C, 2M) = \mathcal{K} - \mathcal{K}$ (Lemma B.2(iii)), the smallest constant for which (10) holds on the class at the observed ζ is

$$\Delta_{\mathrm{emp}}(C) = \sup_{h\in\mathcal{H}}\Big\{\tau Z(h) - \frac{\kappa}{4}\|h\|_N^2\Big\} = \frac{\tau^2}{\kappa}\|\zeta\|_N^2 - \frac{\kappa}{4}\min_{h\in\mathcal{H}}\Big\|h - \frac{2\tau}{\kappa}\zeta\Big\|_N^2,$$

where the absolute value of (10) is dropped because $\mathcal{H}$ is symmetric, and $(\tau, \kappa) = (2A, 2)$ for squared and $(1, \kappa_M)$ for logistic loss. The direct program (the negative of the supremum) gives the bracket used in every judgment. The projection program is bracketed separately as a cross-check, with the rounding of its center $2\tau\zeta/\kappa$ propagated through $|\mathrm{dist}_N^2(c, \mathcal{H}) - \mathrm{dist}_N^2(\hat{c}, \mathcal{H})| \le \|c - \hat{c}\|_N(2\|\hat{c}\|_N + \|c - \hat{c}\|_N)$, and the two intervals overlap on every record. For the squared loss $\|\zeta\|_N^2 = \hat{\varepsilon}_T$, so $\Delta_{\mathrm{emp}} = 2A^2\hat{\varepsilon}_T - 2\min_{v\in\mathcal{K}}\|v - A\zeta\|_N^2$; when $\hat{v}, v_0 \in \mathcal{K}$ the quadratic identity (Lemma C.2) writes the deficit as $2AZ(\hat{v} - v_0) - \frac{1}{2}\|\hat{v} - v_0\|_N^2$, which is at most Δ_{emp}.

Gains. If $\tilde{v}$ is the returned mean-flip optimizer and g the width of its bracket, then $0 \le \widehat{L}^{\log}_{T,\varepsilon}(\tilde{v}) - \widehat{L}^{\log}_{T,\varepsilon}(v^*) \le g$, and the curvature of $\widehat{L}^{\log}_{T,\varepsilon}$ on the box (Lemma F.5) gives $\|\tilde{v} - v^*\|_N \le e_* := \sqrt{2g/\kappa_M}$; the tolerance is $g \le \kappa_M \cdot 10^{-8}/2$, so that $e_* \le 10^{-4}$. The logistic loss is 1-Lipschitz in the logit, so $|\widehat{L}^{\log}_{T,\mathrm{corr}}(\tilde{v}) - \widehat{L}^{\log}_{T,\mathrm{corr}}(v^*)| \le e_*$; since $\widehat{L}^{\log}_{T,\mathrm{corr}} = \widehat{L}^{\log}_{T,\varepsilon} - Z$ (Lemma F.4), also $\widehat{L}^{\log}_{T,\mathrm{corr}}(v^*) \in [\widehat{L}^{\log}_{T,\mathrm{corr}}(\tilde{v}) - g - \|\zeta\|_N e_*,\ \widehat{L}^{\log}_{T,\mathrm{corr}}(\tilde{v}) + \|\zeta\|_N e_*]$, and the two enclosures are intersected. The gain $G_T(\hat{v})$ subtracts $\widehat{L}^{\log}_{T,\mathrm{corr}}(\hat{v})$ over the rounding box of $\hat{v}$; the master inequality adds $\frac{\kappa_M}{4}\|\hat{v} - v^*\|_N^2$, with $\|\hat{v} - v^*\|_N$ enclosed by $\|\hat{v} - \tilde{v}\|_N \pm e_*$; and $G_{\max}(C)$ combines the enclosure of $\widehat{L}^{\log}_{T,\mathrm{corr}}(v^*(C))$ with the bracket of $\min_{\mathcal{K}_T(C,M)} \widehat{L}^{\log}_{T,\mathrm{corr}}$ for the same budget C. All losses are evaluated in interval arithmetic. The largest achieved values are $g \le 2.2 \times 10^{-7}$ and $e_* \le 2.1 \times 10^{-3}$, for a few benchmarks at $n = 8000$ where the solver stops short of its tolerance. Every reported gain interval, including those at these benchmarks, is narrower than 1.3×10^{-3}.

Exact constants. Budgets and constants ($g_{\mathcal{D}}^{\min}(T)$, C_T, κ_M) are enclosed in 120-bit interval arithmetic, $C_- \le C_T \le C_+$ and $\kappa_- \le \kappa_M \le \kappa_+$, with relative widths below 10^{-14}. The programs are solved at computed values and the results are transferred to the exact ones. For $a \ge 1$, $\mathcal{K}_T(aC, M) \subseteq a\mathcal{K}_T(C, M)$ gives $\Delta_{\mathrm{emp}}(C) \le \Delta_{\mathrm{emp}}(aC) \le a\Delta_{\mathrm{emp}}(C)$; the realized remainder changes by at most $M^2(\kappa_+ - \kappa_-)$ between κ_- and κ_+; and scaling a minimizer toward 0 shows that minima of the logistic objectives over $\mathcal{K}_T(C_-, M)$ and $\mathcal{K}_T(C_+, M)$ differ by at most $M(1 - C_-/C_+)$. Every judgment reported below holds for the exact C_T and κ_M.

Judgments. An inequality between certified quantities is *supported* when it holds with the interval ends taken against it, *contradicted* when it fails with the ends taken in its favor, and *undecided* otherwise; Table 5 counts these states.

G.4 SPECTRAL STATUS

Training tracks the largest eigenvalue of the Hessian by a power iteration in single precision. This value is a diagnostic: it is a Rayleigh quotient, hence a lower bound in exact arithmetic, but it carries no rounding control and certifies nothing.

The verified upper bound uses the block structure: for the zero-padded design the full-parameter Hessian is the direct sum of the intrinsic block in $(v_k, w_{k1}, b_k, \beta)$ and a zero block, so $\lambda_{\max}$ is that of the intrinsic block. For the squared loss this block is $H = 2G + R$ with $G = n^{-1}\sum_i J_i J_i^\top$ and $R = \frac{2}{n}\sum_i (f_i - Y_i)F_i$ (Proposition C.1, Step 1); for the logistic loss $2G$ and $\frac{2}{n}(f_i - Y_i)$ become $n^{-1}\sum_i \psi''(f_i) J_i J_i^\top$ and $\frac{1}{n}(\sigma(f_i) - \widetilde{Y}_i)$. The residual R is a direct sum of 3×3 blocks $\left(\begin{smallmatrix} 0 & z_k^\top \\ z_k & 0 \end{smallmatrix}\right)$, one per neuron, with $z_k = \frac{2}{n}\sum_i (f_i - Y_i)\mathbb{1}\{w_{k1}x_i > b_k\}(x_i, -1)$, so $\|R\|_2 = \max_k \|z_k\|$. The Gauss–Newton part has the nonzero eigenvalues of the $n \times n$ matrix $W = n^{-1}BB^\top$, where B has the rows $\sqrt{\ell_i''}\, J_i^\top$ with ℓ_i'' the second derivative of the loss in the output; for a shift μ and any computed Cholesky factor L of $\mu I - W$, $\lambda_{\max}(W) \le \mu + \|\mu I - W - LL^\top\|_2$. The verifier bounds this residual, the error of forming W and the errors of the z_k in the standard model of binary64 arithmetic, with absolute bounds for every quantity that can cancel. It encloses the logistic weights $\sigma(f_i)$ and $\psi''(f_i)$ in interval arithmetic over the rounding box of f_i. It certifies each activation sign by $|w_{k1}x_i - b_k|$ exceeding its rounding bound and abstains otherwise. With $\|W - \widehat{W}\|_2 \le e_W$ for the computed matrix $\widehat{W}$, $\|\mu I - \widehat{W} - LL^\top\|_2 \le r_{\text{chol}}$ and $\|R\|_2 \le r_R$, Weyl's inequality gives $\lambda_{\max}(H) \le U_H := \mu + r_{\text{chol}} + e_W + r_R$, where H is the Hessian of the real-arithmetic network defined by the stored parameters; U_H and its comparison with $2/\eta$ are rounded outward. U_H is computed for $n \le 2000$. The supplementary material contains the complete error model.

The verified lower bound L_H is the Rayleigh quotient $x^\top H x / x^\top x$ at a computed vector x, bounded from below under the same rounding model. A network is *certified* when $U_H \eta/2 \le 1$, a *witnessed violation* when $L_H \eta/2 > 1$, and *undetermined* otherwise. On every network with a verified tier, $L_H \le U_H$, and L_H witnesses the same violations as the training-time value. Counts are by distinct network; the four thresholds at $n = 2000$ share their networks.

G.5 DEEP REGION VERSUS SHALLOW COMPLEMENT

The theorem certifies the deep region only, but the stored checkpoints allow the same objects to be recomputed on its complement without retraining. For every squared endpoint the design, teacher and corruption mask are rebuilt from the training seeds, and the network is evaluated in double precision on the full sample. The training loss, the realized corruption rate and the clean error are then taken on I_T, on $\{1, \dots, n\} \setminus I_T$ and on the full sample. The recomputed deep-region values agree with the stored ones to 3×10^{-17}. Table 3 reports medians for $A = 1$. On the shallow complement the loss lies slightly above the local corruption energy. The clean error is roughly two to four times larger than on the deep region, as expected near the boundary of the design, but it stays one to two orders of magnitude below the energy. So the plateau at the corruption energy extends beyond the deep region.

Table 3: Deep region versus shallow complement, squared loss, $A = 1$ (medians over the trained masks). Ratio $= L/(A^2\hat{\varepsilon})$ on the stratum; E = clean error on the stratum.

n	T	n_T/n	deep ratio	shallow ratio	full ratio	deep E	shallow E
500	0.1	0.80	0.993	1.041	1.001	1.9×10^{-3}	3.4×10^{-3}
1000	0.1	0.80	1.000	1.031	1.006	1.1×10^{-3}	2.2×10^{-3}
2000	0.1	0.80	1.004	1.032	1.010	7.8×10^{-4}	1.9×10^{-3}
4000	0.1	0.80	1.008	1.027	1.013	7.3×10^{-4}	1.6×10^{-3}
8000	0.1	0.80	1.009	1.024	1.012	5.7×10^{-4}	1.3×10^{-3}
2000	0.05	0.90	1.006	1.041	1.010	7.8×10^{-4}	2.8×10^{-3}
2000	0.2	0.60	1.003	1.021	1.010	5.9×10^{-4}	1.5×10^{-3}
2000	0.3	0.40	0.998	1.018	1.010	4.6×10^{-4}	1.3×10^{-3}

G.6 AMBIENT DIMENSION AND STEP SIZE

Two sweeps at $n = 2000$, $T = 0.1$, $A = 1$ reuse the first 20 corruption masks of the main grid, so that every row of Table 4 is paired with the same masks of the baseline ($d = 50$, $\eta = 0.2$); training and evaluation are otherwise identical to §G.2–G.4. *Ambient dimension.* The design is zero-padded,

so d enters neither the geometry nor the theory. It does enter the optimization: the initialization $w_k \sim \mathcal{N}(0, \frac{2}{d} I)$ gives the input weight on the one nonzero coordinate the scale $\sqrt{2/d}$. At all three d the loss sits on the corruption energy and every endpoint lies in the class. The clean error and the edge ratio change with d: the $d = 50$ runs end at the edge with the largest clean error, while the $d = 2$ and $d = 500$ runs end inside the stable region with a clean error about four times smaller. *Step size.* At every step size the endpoints lie in the class. Up to $\eta = 0.2$ the loss sits on the energy, and the small-step runs end far below the edge (edge ratio 0.28 and 0.60) with a smaller clean error than the edge-of-stability run at $\eta = 0.2$. At $\eta = 0.4$ the endpoints underfit: they are smoother than the teacher (variation 1.6 against the teacher's 2.0 on the deep region), the clean error rises to 0.019 and the loss to 1.39 times the energy, above the floor of Theorem 5 as the theorem requires. The worst-case budget $C_T^{\mathrm{sq}} \approx 4 \times 10^3$ at this step size is still far above the teacher's variation, so the well-specified assumption holds and the budget does not explain the underfitting. The verified bounds certify none of these endpoints and witness a violation on eight of them. Their training-time edge ratios have median 0.93 and exceed one on the same eight. Identifying the cause would need a tighter spectral description or an analysis of the training dynamics.

Table 4: Ambient-dimension and step-size sweeps (squared loss, $n = 2000$, $T = 0.1$, $A = 1$, masks 0–19; medians). Edge ratio: training-time power-iteration estimate. Certified: networks whose verified upper bound is at most $2/\eta$.

configuration	$L/(A^2\hat{\varepsilon}_T)$	$E(\hat{v})$	edge ratio	certified	$\mathrm{Var}(\hat{v})$
$d = 50$, $\eta = 0.2$ (baseline)	1.006	8.2×10^{-4}	1.00	5/20	5.5
$d = 2$, $\eta = 0.2$	0.995	2.0×10^{-4}	0.31	20/20	4.4
$d = 500$, $\eta = 0.2$	0.996	2.2×10^{-4}	0.84	20/20	2.9
$d = 50$, $\eta = 0.05$	0.997	3.4×10^{-4}	0.28	20/20	6.1
$d = 50$, $\eta = 0.1$	0.996	2.9×10^{-4}	0.60	20/20	6.7
$d = 50$, $\eta = 0.4$	1.386	1.9×10^{-2}	0.93	0/20	1.6

G.7 Matched geometry and the cap constructions

For Table 2 each seed draws one set of random streams, separated by role (angle, radius, label, flip, sign, initialization) and shared by the two geometries. The designs are $X_i = \sqrt{U_i}\, u_i$ on the disk and $X_i = u_i$ on the circle, with U_i uniform on $[0, 1]$ and u_i uniform on the unit circle, both in the first two coordinates of $\mathbb{R}^{50}$. The teacher is $f_0(x) = (x_1 + x_2)/\sqrt{2}$ on both ($B_0 = 1$ on the disk), and labels, flips and responses are as above with $\varepsilon_n = 0.05$ and $A \in \{1, 0.5\}$. There are six seeds and $n \in \{2000, 4000\}$; training is as in §G.2. On the disk the deep region consists of the samples with exact half-plane depth at least T. On the circle every point has depth $1/n$, so the deep region is empty for $T > 1/n$ and full-sample quantities are reported. One circle seed (seed 4, at both n and both amplitudes) collapsed to a constant network at the first gradient step; it is kept in the medians of Table 2. Edge ratios of trained networks are Lanczos values, a diagnostic, because the verified route of §G.4 applies only to one-dimensional designs.

The cap networks are the constructions of Appendix E instantiated on the same designs: $\delta = (1-\rho)/4$ with $\rho = \max_{i \neq j} X_i \cdot X_j$, weights sX_i and biases $s(1-\delta)$ with $s^2\delta^2 = \sqrt{2}D$ (squared, $D = B_0 + A$) or $\sqrt{2}M$ (logistic), outputs $y_i/(s\delta)$ or $\pm M/(s\delta)$. Sample values and the envelope follow analytically from the construction. The implementation checks them with 50-digit residuals on a subset and in double precision on all samples, where the cancellation in the preactivations loses about $s \cdot 10^{-16}$. The double-precision forward pass misses the labels by up to 0.16 and exceeds D on the samples by up to 0.15, so Table 2 reports the analytic values, not this implementation. For the analytic construction the residual vanishes and the squared Hessian is twice a Gram matrix; its largest eigenvalue is computed from that exactly derived $n \times n$ matrix, which does not involve δ, so no small cap separation enters the numerical eigenvalue problem. This certifies the analytic construction, not the rounded parameters, for which the Lanczos value is a diagnostic; the logistic Hessian is bounded analytically and cross-checked by Lanczos on the same implementation.

G.8 Runs under the logistic extension

The same grid was trained under the logistic loss with flips ($\varepsilon_n \in \{0.05, 0.1\}$, $\eta = 1$, $M = 2$), and every endpoint was evaluated with the certified programs of §G.3, at the budget $C_T^{\log}$ and with membership decided for the true deep vector. Of the 600 networks, 598 have their deep vectors

in $\mathcal{K}_T^{\log} = \mathcal{K}_T(C_T^{\log}, 2)$; the two exceptions, at $n = 500$, exceed M on the deep sample (2.02 and 2.07). The full-sample envelope (A2) fails on eight networks, all at $n = 500$ (largest output 2.76), and holds on the rest. The spectral condition (A1) is certified for 196, witnessed violated for 129 and undetermined for 35 of the 360 networks with a verified tier ($n = 500$: $19/84/17$; $n = 1000$: $66/37/17$; $n = 2000$: $111/8/1$). The in-class endpoints that violate (A1) or (A2) fail the sufficient condition of Lemma F.2, but Theorem F.6 and Corollary F.7 still apply to their deep vectors.

On every in-class record the deep-region corrupted loss of the endpoint lies above the benchmark's: $G_T(\hat{v}) < 0$ on all 958 records, with median -0.13 at $n = 500$ and -0.014 at $n = 8000$, so the gain bound of Corollary F.7 holds on every record. The master inequality of Theorem F.6 holds on every record as well: on each record, $\Delta_{\text{emp}}^{\log}$ exceeds the left side $G_T(\hat{v}) + \frac{\kappa_M}{4}\|\hat{v} - v^*\|_N^2$ by at least 0.024. The left side is negative on all but one record, where it lies in $[5.9, 6.1] \times 10^{-4}$ ($n = 2000$, $T = 0.3$). The remainder $\Delta_{\text{emp}}^{\log}$ is at least 0.013, with median 0.22 at $n = 500$ and 0.024 at $n = 8000$. These values are for the exact $C_T^{\log}$ and κ_M (§G.3). The full-sample corrupted loss of the endpoints (median 0.58–0.59 at $\varepsilon_n = 0.05$) is slightly below the teacher's (0.59–0.60). As for the squared loss, these check the inequalities on the realization and do not verify the high-probability rate.

The class-level pass tests the inequality on the whole class instead of one vector. For 500 corruption masks at $n = 2000$, $T = 0.1$, $\varepsilon_n = 0.05$ and the budgets 1, 64, 2048 and $C_T^{\log} \approx 2.36 \times 10^4$, the convex programs return the largest gain over the benchmark that any vector of $\mathcal{K}_T(C, M)$ achieves, $G_{\max}$, and the realized remainder $\Delta_{\text{emp}}^{\log}$, each with the benchmark of the same budget. The first 60 masks are those of the trained networks. Corollary F.7 requires $0 \le G_{\max} \le \Delta_{\text{emp}}^{\log}$; it is supported on all 2000 mask–budget comparisons, with $\Delta_{\text{emp}}^{\log}$ about five times $G_{\max}$ (Table 5).

Table 5: Class-level comparison, $n = 2000$, $T = 0.1$, $\varepsilon_n = 0.05$, 500 masks (medians over masks; every certified interval of $G_{\max}$ is narrower than 3×10^{-5}); $C_T^{\log} \approx 2.36 \times 10^4$ for this design.

budget C	$G_{\max}$	$\Delta_{\text{emp}}^{\log}$	supported
1	1.9×10^{-4}	1.1×10^{-3}	500/500
64	1.0×10^{-3}	5.5×10^{-3}	500/500
2048	4.2×10^{-3}	2.3×10^{-2}	500/500
$C_T^{\log}$	1.1×10^{-2}	6.3×10^{-2}	500/500

On the matched disk–circle designs of §G.7, under logistic loss the two-dimensional endpoints stay within the envelope on both geometries ($22/24$) and at the teacher's loss, whereas the $\pm M$ cap networks of Proposition F.9 satisfy the spectral bound $0.106 \ll 2$ and reach the losses 0.127 (f_{corr}) and 0.22 (f_{clean}, whose sample vector is the box benchmark; Table 6).

Table 6: Matched geometry under the logistic extension (six seeds, $n \in \{2000, 4000\}$, medians; final endpoints; edge ratio $= \lambda_{\max}\eta/2$, Lanczos lower bound for GD, analytic bound for the caps).

network	L_{corr}	edge ratio	$\max_i \lvert f(X_i)\rvert$
disk, GD	0.66	0.99	≤ 2.08
circle, GD	0.64	1.00	≤ 1.35
circle, cap f_{corr} / f_{clean}	0.127 / 0.22	≤ 0.053	2

G.9 Reporting

Tables report medians over masks or seeds and counts over distinct networks. The evaluation of the trained networks is complete for every configuration, and the class-level pass covers 500 masks at $n = 2000$, $T = 0.1$, $\varepsilon_n = 0.05$ for the four budgets of Table 5.

H Extended Related Work

Large learning rates and minima stability. Wu et al. (2018) characterize the minima that (stochastic) gradient descent can select by their linear stability. The edge-of-stability phenomenon was identified empirically by Cohen et al. (2021) and analyzed theoretically by Damian et al. (2023). Ma & Ying (2021) connected linear stability to input-space Sobolev regularization. The function-space consequences of minima stability were studied for univariate interpolation by Mulayoff et al. (2021),

for multivariate interpolation by Nacson et al. (2023), and for univariate non-interpolation by Qiao et al. (2024). A contemporaneous univariate logistic analysis by Qiao & Wang (2025) shows that flatness-implied generalization under logistic loss is more delicate than under squared loss: positive results hold within solution-dependent uncertain regions, while flatness alone is insufficient in general. We extend this line to multivariate non-interpolation with corruption, under squared and logistic loss.

Data geometry and shatterability. Liang et al. (2025) proved that stable minima suffer from a curse of dimensionality on the uniform ball, and introduced the neural shattering phenomenon. Liang et al. (2026) showed that generalization rates adapt to intrinsic dimension on mixtures of low-dimensional balls, and provided the deep-region / half-space-depth framework that our proof builds upon. Section 5 states the full-sample floor that their bounds imply under corruption and compares it with our deep-region bound.

Benign overfitting and memorization. Zhang et al. (2017) showed that deep networks can fit random labels, and Arpit et al. (2017) documented that networks learn patterns before memorizing noise. The benign-overfitting literature for linear models (Bartlett et al., 2020) and neural networks (Frei et al., 2022) studies when interpolation generalizes. Our setting is different. Theorem 5 limits how far fitting the corruption can lower the training loss, and it excludes exact interpolation of the deep samples once its remainder is below $\frac{3}{2}A^2\hat{\varepsilon}_T$; on the sphere stable interpolation is possible (Proposition 7).

Robust estimation and noisy labels. Classical robust statistics (Huber, 1964) designs explicit loss functions to resist outliers. In the noisy-label literature, Natarajan et al. (2013) and Patrini et al. (2017) construct corrected surrogates for classification under random label flips. Label-error detection methods such as Confident Learning (Northcutt et al., 2021) and the area-under-the-margin ranking (Pleiss et al., 2020) ask which examples are corrupted rather than how much a trained network fits them. Optimization-induced robustness has also been studied: Li et al. (2020) prove that early-stopped GD on overparameterized networks fits clean labels before overfitting corrupted ones, and Damian et al. (2021) show that SGD with label noise prefers flatter global minimizers. Our contribution is complementary. We design no corrected loss and use no early stopping. We prove that under plain squared (or logistic) loss, stable networks with bounded outputs can lower their training loss on the deep region of a low-dimensional ball only a little by fitting the corruption.

Comparison summary. Table 7 summarizes how our results differ from the closest prior work in setting, quantity and technique.

Table 7: Theorem-object comparison with prior work on stable minima. Section 5 lists the shared geometric scaffold: depth, the affine reduction of a kink on the deep region and the sphere interpolant.

Paper	Setting	Quantity	Distinctive technique
This work	Logistic; sparse flips; m-ball deep region; sample envelope	Same-realization gain over the mean-flip benchmark	Sample-vector class, sparse offset inequality
This work	Squared; additive outliers; m-ball deep region; sample envelope	Training-loss floor; clean-error oracle inequality	Projection benchmark, quadratic identity
Liang et al. (2026)	Squared; bounded labels; mixture of m-balls; global envelope; i.i.d. uniform design	Gen. gap	Half-space-depth quantile partition
Liang et al. (2025)	Squared; bounded labels; ball / sphere; global envelope; i.i.d. uniform design	MSE / gap	Neural shattering
Qiao et al. (2024)	Squared; noisy labels; univariate interior	Interior MSE	Block-barrier
Qiao & Wang (2025)	Logistic; no label corruption; univariate	Gen. in uncertain region	Solution-dependent uncertain sets