REMOVING SPURIOUS MINIMA FOR PLANAR FEATURES BY SKIP CONNECTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding loss landscapes is central to explaining neural-network training, yet their structure remains only partially understood even in simple models. We study the Gaussian population loss of shallow, bias-free ReLU networks in the teacher–student setting. This provides a simple model for studying essential aspects such as feature learning and overparameterization. For teacher networks with positive output weights and planar features, we show that including a learned linear skip removes all spurious local minima with non-negative student output weights once the student network is at least as wide as the teacher network. In contrast, without the skip, we construct a fixed teacher network with positive output weights and only three hidden neurons in input dimension two whose spurious local minima persist at every student width at least three. Thus, a learned linear skip can remove spurious minima that persist under arbitrary overparameterization. Furthermore, we show that a positive output weight student network always learns the subspace spanned by the teacher features: student features at local minima with non-negative student output weights lie in the span of the teacher features. For ReLU networks in two dimensions, even heavily overparameterized student networks have effective width controlled by the teacher width: every critical point with positive student output weights has at most twice as many distinct student feature directions as teacher neurons. Finally, we transfer the benignity result to empirical minima over parameter balls of any prescribed radius, with the required sampling accuracy depending on that radius.

1 INTRODUCTION

Training a neural network typically requires solving a highly non-convex optimization problem, and understanding the corresponding loss landscape is a central problem in machine learning. Overparameterization and adding residual connections are both associated with easier training. Sufficient overparameterization enables convergence guarantees for gradient descent from random initialization under suitable assumptions (Du et al., 2019). Empirical studies show improved training with residual connections (He et al., 2016a;b) and two-dimensional loss visualization suggest less rugged loss landscapes with residual connections (Li et al., 2018). We ask when increasing width or adding a linear skip connection actually excludes spurious local minima.

Teacher–student models provide a standard setting for studying loss landscapes and training dynamics (Tian, 2017; Safran & Shamir, 2018; Arjevani & Field, 2021; Zhou et al., 2021). When the student network architecture contains the teacher network's, the problem is realizable: zero population loss is attainable. Moreover, it is a simple model to study essential aspects such as feature learning and overparameterization. The latter can be modeled by allowing the student network to have more hidden neurons than the teacher network. For a shallow teacher network without a skip, labels depend only on the input's projection onto the span of its *feature directions*, namely the normalized non-zero input-weight vectors of its hidden neurons. It is therefore a particular multi-index model and provides a controlled setting for asking whether training identifies the target-relevant subspace. Prior work studies such subspace learning through gradient-based training (Damian et al., 2022; Mousavi-Hosseini et al., 2023); this leaves open to what extent subspace learning is a property of the optimization dynamics or of the loss landscape itself.

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.

Overparameterization also raises the question whether additional neurons learn distinct features or become redundant. We measure *effective width* by the number of distinct feature directions among neurons with non-zero contribution, and ask whether criticality alone bounds it, even when the student network does not fit the teacher network.

Realizability alone does not exclude spurious local minima: for shallow ReLU networks with Gaussian inputs they exist even for teacher networks with orthogonal feature directions (Safran & Shamir, 2018), in families whose Hessian spectra are known (Arjevani & Field, 2021), and increasing width turns only certain of them into saddles (Safran et al., 2021; Arjevani & Field, 2022). It remains natural to ask whether an architectural change can remove minima that survive arbitrary overparameterization.

We study Gaussian population squared loss for bias-free, scalar-output, one-hidden-layer ReLU networks. The population setting retains the non-convex geometry of feature learning while separating it from finite-sample effects, making it a useful model for isolating the roles of width and architecture. At the same time, population-level landscape guarantees can inform the finite-data setting. We compare the respective loss landscapes of one-hidden-layer ReLU networks with and without a linear input-to-output skip connection. We refer to these models as *plain network* and *skip network*.

1.1 Our Contributions

We establish five results relating the loss landscape to the architecture, the teacher geometry, and the student width.

- *Main result:* A linear skip eliminates spurious local minima under the following assumptions: Suppose that the teacher network and the student network are both skip networks and the teacher network has positive output weights and feature directions in a two-dimensional linear subspace. Whenever the student network is at least as wide as the teacher network, every local minimum of the Gaussian population loss with non-negative student output weights has zero loss (Theorem 10).
- *Limits of overparameterization:* Increasing width alone need not eliminate spurious local minima. More precisely, we construct as teacher network a plain network in input dimension two with three hidden neurons and positive output weights. At every student width at least three, its population loss admits a spurious local minimum with positive student output weights (Theorem 11). Our main result (Theorem 10) excludes such minima once a linear skip is added.
- *Feature learning:* Local minimality confines student features to the teacher subspace. More precisely, with or without a linear skip, at every local minimum with non-negative student output weights, each student neuron with non-zero output weight has its feature direction in the span of the teacher network's feature directions (Proposition 8).
- *Effective width:* At critical points, the teacher width limits the number of student features. More precisely, for a plain network in input dimension two, every critical point of the population loss with positive student output weights has at most twice as many distinct feature directions as the teacher network has hidden neurons. This bound is sharp (Proposition 12).
- *Finite sample consequence:* We extend the main result from population loss to empirical loss given some additional local structure. More precisely, assume that the same conditions are fulfilled as in the 'main result' and that the teacher-generated samples are sufficiently accurate. Then, a locally optimal student network has zero empirical loss if the locality depends on the accuracy (Theorem 13).

The key to our analysis of skip networks is the identity $[z]_+ = \frac{1}{2}|z| + \frac{1}{2}z$: a plain network splits into a *centered* network with activation $|z|/2$ and a linear map that a learned skip can absorb (Zhou et al., 2021). Centering removes exactly the first angular Fourier harmonic and leaves the even ones unchanged, so the centered network is π-periodic in the direction angle; for non-negative masses the split is a translation of zonotopes (Figure 3 in Section A.1).

Modern paper format. We decided on a novel structure for this paper to achieve various desirable aspects in the age of AI-generated mathematics. The first 9 pages give an introduction and overview

of the main results, explaining the fundamental concepts and some of the interesting proof ideas. The appendix does not contain full proofs but all proof ideas supporting human understanding involved in the deduction of our main results extracted by the authors from LLM-generated proofs. For proper independent verification, we link our repository containing an auto-formalization of the results; these Lean certificates follow the same proof structure and calculations omitted from the exposition (Section B). We complement this by adding the formalization of the main results and relevant definitions in the appendix so that one can check that they agree with the natural language statements given in the first 9 pages.

Human–AI collaboration and proof certification. Generative AI played a substantial role in the mathematical discovery and proof search in Lean underlying this work. For perspectives on AI in mathematics, see Tao (2026); Avigad (2026); Commelin et al. (2026). The problem was particularly amenable to this approach: it is basic yet timely, has a natural hierarchy of subproblems and levels of generality, and involves substantial computation whose conceptual consequences are more informative than its details. Numerical exploration and formal proof search thus helped reveal low-level patterns that were progressively distilled into general statements. Section C details the discovery process, unsuccessful approaches, human and AI contributions, and verification procedures; the supplementary material also includes the final agent skills as Markdown (`.md`) files. Our contribution is mathematical rather than an evaluation of an agentic proof-search system.

1.2 Related Work

Teacher–student models. For shallow ReLU networks with Gaussian inputs and fixed output weights, Tian (2017) derives explicit population gradients and uses them to analyze critical points and convergence. Safran & Shamir (2018) give a computer-assisted construction of spurious local minima even for teacher networks with orthogonal teacher feature directions. Li & Yuan (2017) prove SGD convergence with a *pre-activation* identity connection under Gaussian inputs and a prescribed random-initialization regime; unlike our learned input-to-output skip, it acts before the activation. More recently, Tinaz et al. (2026) study the learning of a linear target by a ReLU network, showing that gradient descent from a moderately small random initialization converges globally at a linear rate despite the presence of non-strict saddle points.

Population loss with a learned linear term. Zhou et al. (2021) study Gaussian population regression for one-hidden-layer networks with absolute-value activation and non-negative output weights, and derive its equivalence to ReLU regression with an additional learned linear term. They prove gradient-descent convergence from a sufficiently small-loss initialization for teacher networks of any feature rank with sufficiently separated feature directions and any student width at least the teacher network's. In contrast, for rank-at-most-two teacher networks we characterize the full landscape: every local minimum with non-negative output weights is an exact fit, without a small-loss or teacher-separation assumption, thereby ruling out positive-loss local minima left open by their local convergence analysis. Their proof is perturbative around the teacher configuration, using low loss to aggregate student neurons around teacher neurons and construct a descent direction, whereas our residual formulation globally encodes the loss and its first derivatives and enables structural arguments from local minimality. For other residual architectures, local-minimum guarantees compare risk with that of the best linear predictor under architectural and regularity assumptions (Shamir, 2018; Kawaguchi & Bengio, 2019; Yun et al., 2019). Such bounds need not imply exact fitting of a non-linear teacher network, unlike our zero-loss conclusion under the stated assumptions. Liang et al. (2018) remove bad local minima for suitable empirical classification losses using a regularized exponential neuron; we instead use an unregularized linear skip.

Learning the relevant subspace. Damian et al. (2022) show that gradient descent can learn representations depending only on the relevant subspace of low-dimensional polynomial targets, yielding a sample-complexity advantage over kernel methods. For Gaussian multi-index models, Mousavi-Hosseini et al. (2023) prove that online SGD with weight decay drives the first-layer weights toward the target-relevant subspace and use this to obtain width-independent generalization bounds.

Duplicated neurons. Şimşek et al. (2021) analyze how neuron replication and permutation symmetries generate connected sets of global minima and families of critical points in overparameterized

networks. When a network is widened by duplicating neurons without changing the function it represents, a local minimum of the original network may remain a local minimum or become a saddle point in the wider network (Fukumizu & Amari, 2000; Fukumizu et al., 2019; Wu et al., 2025).

Empirical loss landscapes. Sample-dependent width bounds can exclude spurious valleys without excluding non-strict spurious minima (Venturi et al., 2019). Convex reformulations characterize optima and stationary points of regularized two-layer ReLU networks (Pilanci & Ergen, 2020; Wang et al., 2022; Kim et al., 2025). Our transfer uses fixed-radius minimality and uniform convergence of loss values, rather than of gradients (Foster et al., 2018).

2 PRELIMINARIES

Unless stated otherwise, $d \geq 2$ is the input dimension; statements that also hold in dimension one, among them our main results, say so explicitly. Throughout, $X \sim \mathcal{N}(0, I_d)$, and $q \in \{C, R\}$ labels the two *features* $\varphi_C(z) = |z|/2$ (centered) and $\varphi_R(z) = [z]_+$ (plain).

Definition 1 (Units, networks, and parameter classes)**.** A *one-hidden-layer network of width* $n \geq 1$ has a weight vector $a_i \in \mathbb{R}^d$ and an output weight $o_i \in \mathbb{R}$ per hidden neuron and *represents* the function

$$f_{q;o,a}(x) = \sum_{i=1}^{n} o_i \, \varphi_q(\langle a_i, x\rangle), \qquad (o, a) \in \mathbb{R}^n \times (\mathbb{R}^d)^n. \tag{1}$$

We write $\mathcal{N}_{\mathrm{skip}}(d, n)$ for the class of functions of the form $F_{b,o,a}(x) = \langle b, x\rangle + f_{R;o,a}(x)$ represented by a plain network together with a *linear skip vector* $b \in \mathbb{R}^d$.

Both features are positively homogeneous, so $o_i\varphi_q(\langle a_i, x\rangle) = o_i\|a_i\| \, \varphi_q(\langle a_i/\|a_i\|, x\rangle)$ whenever $a_i \neq 0$: only the product $o_i\|a_i\|$ and the direction $a_i/\|a_i\|$ enter the represented function. A *unit* is accordingly a pair $(s, w) \in \mathbb{R} \times \mathbb{S}^{d-1}$ of a *mass* and a *(feature) direction*, the vector $sw \in \mathbb{R}^d$ is its *generator* (Figure 1a), and the same classes are parameterized in mass–direction coordinates by

$$f_{q;s,w}(x) = \sum_{i=1}^{n} s_i \, \varphi_q(\langle w_i, x\rangle), \qquad (s, w) \in \mathbb{R}^n \times (\mathbb{S}^{d-1})^n, \tag{2}$$

and $F_{b,s,w}(x) = \langle b, x\rangle + f_{R;s,w}(x)$. We call parameters that are not restricted to form unit directions the *raw weights*. On the open set $\mathcal{U}_{\neq 0}(d, n)$ of raw weights (o, a) with every $a_i \neq 0$, the *reparameterization map* $\Pi(o, a) = (o_i\|a_i\|, \, a_i/\|a_i\|)_i$ satisfies $f_{q;o,a} = f_{q;\Pi(o,a)}$; it is a continuous surjection that removes the positive-scaling redundancy of the raw weights.

We fix a *teacher network* by fixing parameters $(b^{\mathbf{t}}, t, v) \in \mathbb{R}^d \times \mathbb{R}^m \times (\mathbb{S}^{d-1})^m$, representing the function $F_{b^{\mathbf{t}},t,v} \in \mathcal{N}_{\mathrm{skip}}(d, m)$. Its width m may differ from the width n of the trainable *student network* with parameters $(b^{\mathbf{s}}, s, w) \in \mathbb{R}^d \times \mathbb{R}^n \times (\mathbb{S}^{d-1})^n$. The weight tuples (t_k, v_k) and (s_i, w_i), are the *teacher units* and the *student units*. A unit with $s_i = 0$ is a *zero-mass unit* and its respective hidden neuron contributes nothing to the represented function. Widths are *underparameterized* ($n < m$), *matched* ($n = m$), or *overparameterized* ($n > m$).

We consider the following parameter regimes for the mass–direction vector pairs (s, w), with the same notation for student and teacher parameters,

$$\mathcal{P}^{+}(d, n) = \{s_i \geq 0 \; \forall i\}, \quad \mathcal{P}^{++}(d, n) = \{s_i > 0 \; \forall i\}, \quad \mathcal{P}_{\mathrm{pl}}(d, n) = \{\dim \operatorname{span}\{w_i\} \leq 2\}, \tag{3}$$

comprising parameters with non-negative masses, with positive masses, and with coplanar directions.

To measure the distance between the teacher network and a student network, we introduce three loss functions, depending on the inclusion of linear skips and the choice of feature.

Definition 2 (Losses)**.** For a teacher network $(b^{\mathbf{t}}, t, v) \in \mathbb{R}^d \times \mathbb{R}^m \times (\mathbb{S}^{d-1})^m$ and a student parameter $(b^{\mathbf{s}}, s, w) \in \mathbb{R}^d \times \mathbb{R}^n \times (\mathbb{S}^{d-1})^n$ the *skip loss* is

$$\mathscr{L}_{\mathrm{skip}}(b^{\mathbf{s}}, s, w) = \tfrac{1}{2} \, \mathbb{E}\big[(F_{b^{\mathbf{s}},s,w}(X) - F_{b^{\mathbf{t}},t,v}(X))^2\big], \tag{4}$$

and for a teacher network (t, v) without linear skip and student parameter $(s, w) \in \mathbb{R}^n \times (\mathbb{S}^{d-1})^n$ the *q-loss* is

$$\mathscr{L}_q(s, w) = \tfrac{1}{2} \, \mathbb{E}\big[(f_{q;s,w}(X) - f_{q;t,v}(X))^2\big], \tag{5}$$

where we suppress the dependence of the loss on the teacher parameters in the notation to keep formulas light. The two instances of the q-loss are the *centered loss* $\mathscr{L}_C$ and the *plain loss* $\mathscr{L}_R$.

Definition 3 (Critical points and minima)**.** Let L be one of the losses above, with parameter space $M = \mathbb{R}^n \times (\mathbb{S}^{d-1})^n$ for the q-losses and $M = \mathbb{R}^d \times \mathbb{R}^n \times (\mathbb{S}^{d-1})^n$ for the skip loss, whose elements we write $\theta = (s, w)$, respectively $\theta = (b^{\mathbf{s}}, s, w)$. A parameter $\theta \in M$ is a *critical point* if $\nabla_s L = 0$, every spherical gradient $\nabla_{\mathbb{S},w_i} L = (I_d - w_i w_i^\top)\nabla_{w_i} L$, i.e. the projection of the Euclidean gradient onto the tangent space $T_u\mathbb{S}^{d-1}$, vanishes and additionally for the skip loss, $\nabla_{b^{\mathbf{s}}} L = 0$; a *local minimum* if no parameter of some neighborhood has smaller loss, *strict* if every other one has strictly larger loss and *flat* otherwise; an *exact fit* if $L(\theta) = 0$; and *spurious* if it is a local minimum but not a global minimum. A local minimum *in* a subset $P \subseteq M$ is a local minimum of L on all of M that lies in P, and the landscape of L is *benign* on P if it has no spurious local minimum in P.

For $n \geq m$, by padding teacher parameters with $n - m$ zeros one obtains parameters for an exact-fit student network, so each loss has global minimum 0 and a local minimum is spurious exactly when its loss is positive. For $n < m$, a global minimum may have positive loss, and we give an underparameterized example (Lemma 29). We analyze the loss landscape in the mass–direction parameter space $\mathbb{R}^n \times (\mathbb{S}^{d-1})^n$. On raw weights the losses are the compositions with the reparameterization map: for $(o, a) \in \mathcal{U}_{\neq 0}(d, n)$ we have $\frac{1}{2}\mathbb{E}[(f_{q;o,a}(X) - f_{q;t,v}(X))^2] = \mathscr{L}_q(\Pi(o, a))$, and likewise for the skip loss. The following lemma transfers our results to the raw parameter space $\mathbb{R}^n \times (\mathbb{R}^d)^n$.

Lemma 4 (`Lean_1`, `Lean_2`)**.** *Let $(o, a) \in \mathcal{U}_{\neq 0}(d, n)$. Then (o, a) is a critical point (a local minimum, a spurious local minimum) of $\mathscr{L}_q \circ \Pi$ if and only if $\Pi(o, a)$ is a critical point (a local minimum, a spurious local minimum respectively) of $\mathscr{L}_q$; the same holds for the skip loss.*

The proof (Section A.1) uses that Π has a continuous local right inverse, so neighborhoods and loss values correspond, and applies the chain rule.

How the losses are connected. The identity $[z]_+ = \frac{1}{2}|z| + \frac{1}{2}z$ splits a skip network into a centered network plus a linear map,

$$F_{b^{\mathbf{s}},s,w} = \langle b^{\mathbf{s}} + \tau(s, w), \cdot \rangle + f_{C;s,w}, \text{ where } \tau(s, w) = \tfrac{1}{2}\sum_i s_i w_i = \tfrac{1}{2}\, s^\top w. \tag{6}$$

Figure 3 pictures this split for $d = 2$. By central symmetry of the distribution of X, the random variables $f_{C;s,w}(X)$ and $\langle u, X\rangle$ are uncorrelated for any $u \in \mathbb{R}^d$: their defining functions are even and odd, respectively. Since $\mathrm{Cov}(X) = I_d$, the linear residual contributes $\frac{1}{2}\mathbb{E}[\langle u, X\rangle^2] = \frac{1}{2}\|u\|_2^2$ to the loss. The proof of the following Lemma based on these ideas is given in Section A.1.

Lemma 5 (Loss decomposition, `Lean_3`)**.** *For every teacher parameter $(b^{\mathbf{t}}, t, v) \in \mathbb{R}^d \times \mathbb{R}^m \times (\mathbb{S}^{d-1})^m$ and every student parameter $(b^{\mathbf{s}}, s, w) \in \mathbb{R}^d \times \mathbb{R}^n \times (\mathbb{S}^{d-1})^n$ we have*

$$\mathscr{L}_{\mathrm{skip}}(b^{\mathbf{s}}, s, w) = \mathscr{L}_C(s, w) + \tfrac{1}{2}\|b^{\mathbf{s}} - b_{\mathrm{match}}(s, w)\|_2^2 \tag{7}$$

$$\mathscr{L}_R(s, w) = \mathscr{L}_C(s, w) + \tfrac{1}{8}\|s^\top w - t^\top v\|_2^2, \tag{8}$$

where $b_{\mathrm{match}}(s, w) = b^{\mathbf{t}} + \frac{1}{2}(t^\top v - s^\top w)$. Moreover, $(b^{\mathbf{s}}, s, w)$ is a (strict) local minimum (or critical point) of $\mathscr{L}_{\mathrm{skip}}$ iff $b^{\mathbf{s}} = b_{\mathrm{match}}(s, w)$ and (s, w) is a (strict) local minimum (or critical point) of $\mathscr{L}_C$.

From Gaussian loss to one residual. The following closed form of the pair moment goes back to Cho & Saul (2009) and underlies the computations of Tian (2017) and Safran & Shamir (2018). For unit vectors $\xi, \zeta \in \mathbb{S}^{d-1}$ we consider the *pair moment* $M_q(\xi, \zeta) = \mathbb{E}[\varphi_q(\langle \xi, X\rangle)\varphi_q(\langle \zeta, X\rangle)]$.

Lemma 6 (Pair moments, `Lean_4`)**.** *Let $\xi, \zeta \in \mathbb{S}^{d-1}$ with $\rho = \langle \xi, \zeta\rangle$ and $q \in \{C, R\}$. Then,*

$$2\pi M_q(\xi, \zeta) = \widetilde{\Phi}_q(\rho), \text{ where } \widetilde{\Phi}_C(\rho) = \rho \arcsin \rho + \sqrt{1 - \rho^2}, \text{ and } \widetilde{\Phi}_R(\rho) = \widetilde{\Phi}_C(\rho) + \frac{\pi}{2}\rho. \tag{9}$$

The proof (Section A.1) reduces to the plane of ξ, ζ and integrates in polar coordinates over the arc on which both ReLUs are active; the centered case follows from the analysis of Φ_R via $\varphi_R = \varphi_C + z/2$. We call $\widetilde{\Phi}_C$ and $\widetilde{\Phi}_R$ the *kernels*; they differ only by the degree-one harmonic $\frac{\pi}{2}\rho$ (Figure 2 (c)).

The next proposition expresses the loss and its first derivatives through the *residual* $\mathcal{F}_q \colon \mathbb{S}^{d-1} \to \mathbb{R}$ with student and teacher parameters (s, w) and (t, v) respectively,

$$\mathcal{F}_q(u) = \sum_i s_i \widetilde{\Phi}_q(\langle u, w_i \rangle) - \sum_k t_k \widetilde{\Phi}_q(\langle u, v_k \rangle), \tag{10}$$

the *student potential* minus the *teacher potential*. We suppress the dependence of the residual on the student network and the teacher network in the notation for light formulas.

Proposition 7 (Residual representation of the loss, `Lean_5`)**.** *For student and teacher parameters (s, w) and (t, v) respectively, the loss and the directional derivatives of the loss satisfy*

$$\mathscr{L}_q = \frac{1}{4\pi}\Big(\sum_i s_i \mathcal{F}_q(w_i) - \sum_k t_k \mathcal{F}_q(v_k)\Big), \tag{11}$$

$$\partial_{s_i} \mathscr{L}_q = \frac{1}{2\pi} \mathcal{F}_q(w_i), \qquad \nabla_{\mathbb{S}, w_i} \mathscr{L}_q = \frac{1}{2\pi} s_i \, \nabla_{\mathbb{S}, w_i} \mathcal{F}_q(w_i). \tag{12}$$

The proof (Section A.1) expands the square with Lemma 6 into a quadratic form in the masses with the kernels as coefficients and regroups it by units. This gives the loss representation; the derivative representations follow from simple calculations.

3 Properties of Local Minima

Two properties of every local minimum with non-negative masses drive the analysis of Section 4: each student direction of positive mass lies in the span of the teacher directions, and a student unit of zero mass forces an exact fit.

Proposition 8 (Feature Learning, `Lean_6`)**.** *Let $d \geq 2$. For any teacher parameter (t, v), let $\mathcal{T} = \operatorname{span}\{v_k\}$ be the linear span of the teacher directions. For $q \in \{C, R\}$, let $(s, w) \in \mathcal{P}^+(d, n)$ be a local minimum of $\mathscr{L}_q$. Then $w_i \in \mathcal{T}$ for every $i \in [n]$ with $s_i > 0$.*

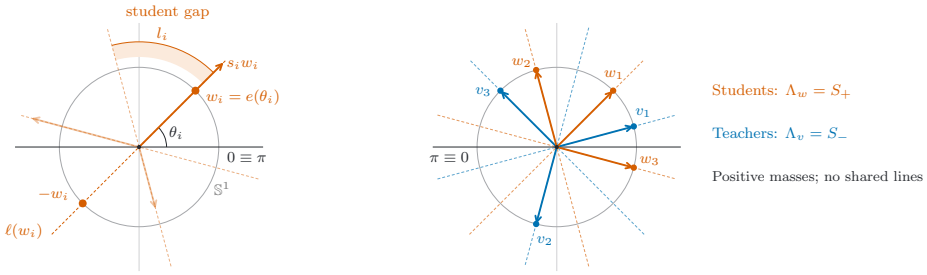
The proof (Section A.2) decomposes each student direction as $w_i = p_i + q_i$, with $p_i \in \mathcal{T}$ and $q_i \perp \mathcal{T}$, and shows that $s_i q_i = 0$. We conclude this by analyzing the variation of $\mathscr{L}_q$ under two specific motions, that is infinitesimal changes of student generators $s_i w_i$. The *normal motion* $z_N = (s_1 q_1, \ldots, s_n q_n)$ stretches the normal components, while the *radial motion* $z_A = (s_1 w_1, \ldots, s_n w_n)$ scales all generators uniformly. For two motions z, z', write $z^\top H z'$ for the mixed second derivative of $\mathscr{L}_q$ in directions z, z'. First, we show that local minimality of (s, w) implies that the normal motion causes no change in $\mathscr{L}_q$ up to second order, $z_N^\top H z_N = 0$. Consequently, Cauchy–Schwarz yields $z_A^\top H z_N = 0$. Second, we interpret this mixed derivative in terms of variation of student's second moment under the normal motion, $z_A^\top H z_N = \frac{1}{2} \frac{d}{d\tau}\big|_{\tau=0} \mathbb{E}[f_{z_A + \tau z_N}(X)^2]$, where f_V denotes the student output with generators V. An argument using convexity of this second moment in the generators shows that the mixed derivative $z_A^\top H z_N$ would be positive if some $s_i q_i \neq 0$, contradicting the first step.

A zero-mass unit also contradicts being a spurious local minimum.

Proposition 9 (Zero student mass implies exact fit, `Lean_7`, `Lean_8`)**.** *Let $d \geq 1$ and assume only that the teacher masses are non-negative when $d = 2$. If (s, w) with unrestricted masses is a local minimum of $\mathscr{L}_C$ (for $d \geq 1$) or $\mathscr{L}_R$ (for $d \geq 3$) with a zero-mass unit, then it is an exact fit.*

The proof (Section A.2) uses the representation of the loss via the residual in Proposition 7: to show an exact fit, it suffices to show that the residual vanishes everywhere. Analyzing how the loss changes when changing the direction and mass of the zero-mass unit (s_i, w_i), using eq. (10), we show that there would be a descent direction if the residual would not vanish on an open set around w_i. We prove that the residual is real analytic away from finitely many breakpoints at the student and teacher directions and their reflections w.r.t. the origin. Depending on the dimension, different techniques propagate this zero set of the residual to prove the residual vanishes everywhere, with the same proof for the plain and centered residual in case $d \geq 3$.

Surprisingly, the $d = 2$ non-negative-teacher hypothesis cannot be dropped, as the example in Lemma 29(b) demonstrates.



(a) Mass, direction, and line

(b) Strict interlacing

Figure 1: Unit notation and strict interlacing in the centered model. (a) A student unit has mass s_i, direction $w_i = e(\theta_i)$ and generator $s_i w_i$; the antipodes $\pm w_i$ determine one line $\ell(w_i) \in \mathbb{R}/\pi\mathbb{Z}$, and the shaded arc marks a student gap of length l_i. (b) Every gap between student lines (orange) contains exactly one teacher line (blue); with positive masses and distinct lines the student lines are S_+ and the teacher lines S_-. The configuration illustrates the order condition of interlacing, not stationarity.

4 Benign Loss Landscape for Coplanar Teacher Networks

This section presents the ideas behind our main result and the role of the residual eq. (10) in our analysis.

Theorem 10 (Main Result, `Lean_9`)**.** *Let $d \geq 1$, let the teacher network satisfy $(t, v) \in \mathcal{P}^{++}(d, m) \cap \mathcal{P}_{\mathrm{pl}}(d, m)$ with any linear skip $b^{\mathbf{t}} \in \mathbb{R}^d$, and let $n \geq m$. Then every local minimum $(b^{\mathbf{s}}, s, w)$ of $\mathscr{L}_{\mathrm{skip}}$ on $\mathbb{R}^d \times \mathbb{R}^n \times (\mathbb{S}^{d-1})^n$ with $(s, w) \in \mathcal{P}^+(d, n)$ is an exact fit.*

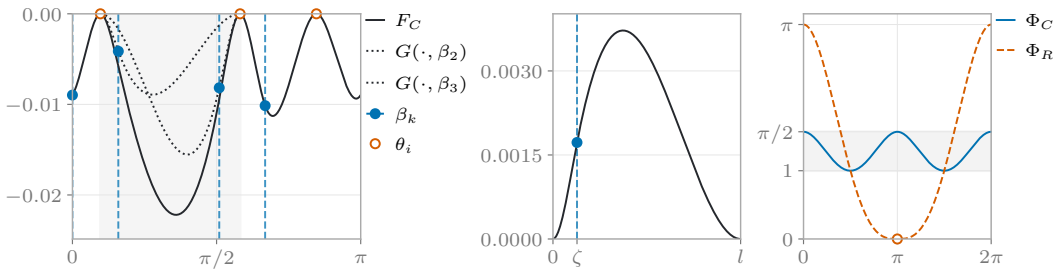
The proof first reduces the statement to the centered model: the skip loss is the centered loss plus a term in the linear skip alone (Lemma 5), so a spurious local minimum with non-negative masses of the skip model yields one of the centered model, and the main result follows from its centered formulation, Theorem 28 in Section A.3. The case $d = 1$ is a direct argument in the masses. Let $d \geq 2$ and let (s, w) be a local minimum of $\mathscr{L}_C$. A zero student mass gives an exact fit (Proposition 9); otherwise feature learning (Proposition 8) places student units and teacher units in a common plane P, and since centered features of directions in P see only the projection of X onto P, (s, w) is a local minimum of the planar loss with the same value. For convenience let us identify $P = \mathbb{R}^2$.

In the plane, write a direction as $e(\gamma) = (\cos\gamma, \sin\gamma)$, with student and teacher angles θ_i, β_k. The centered feature $|\langle u, x\rangle|$ does not distinguish u from $-u$, so only the *line* $\ell(u) = \gamma \bmod \pi \in \mathbb{R}/\pi\mathbb{Z}$ of $u = e(\gamma)$ matters (Figure 1a); collecting the units on a line α gives its *net mass* and the lines of positive and negative net mass,

$$\omega(\alpha) = \sum_{\ell(w_i)=\alpha} s_i - \sum_{\ell(v_k)=\alpha} t_k, \qquad S_\pm = \{\alpha \in \mathbb{R}/\pi\mathbb{Z} : \pm\omega(\alpha) > 0\}. \tag{13}$$

A spurious local minimum with positive masses is excluded in two steps. First, we demonstrate (Lemma 20, Proposition 25) that its lines of positive and negative net mass alternate around the circle: every *gap* of S_+, a connected component of $(\mathbb{R}/\pi\mathbb{Z}) \setminus S_+$, contains exactly one point of S_-; we call this *interlacing* (Figure 1b); moreover S_+ is the set of the n student lines and S_- that of the $m = n$ teacher lines, all distinct (Proposition 25), so for $n > m$ the first step already closes the theorem by contradiction, while the following second step is only needed for the matched case. At an interlacing critical point we construct a direction of strictly negative second variation of the centered loss (Proposition 26), so it is not a local minimum. Both steps rest on the residual of Proposition 7 read as a function of the angle, and together they prove Theorem 28 and with it the main result.

The residual as a bent ring. The residual (10) is a weighted sum of shifted copies of the kernel. We observe that between student and teacher lines, it solves a certain ODE (Lemma 18), whose analysis is the key to our main result. Surprisingly, the same formulas were studied in the physical discipline of elasticity: the residual follows the shape of a ring bent by one pair of opposite forces per line, of strength the net mass (Remark 19; compare to Love (1927)). The response to a single unit load at ζ is the Green function $G(\,\cdot\,;\zeta)$ (Figure 2b), and the centered residual is also a sum of Green functions between the lines of S_+ (Figure 2a), from which the interlacing Lemma 20 follows combinatorially.



(a) Centered angular residual $\mathcal{F}_C(\gamma)$ (b) Green function $G_l(\cdot;\zeta)$ (c) Angular kernels $\Phi_q(\tau)$

Figure 2: Two representations of the residual: kernel sums and Green-function solutions (Lemma 18). (a) Centered angular residual (31) for three student units at numerical double zeros (orange markers; uncertified Newton approximation) and the teacher units $\beta = (0, 0.5, 1.6, 2.1)$, $t = (1, 0.6, 0.8, 1.2)$ (blue markers). On the shaded student gap, which contains two teacher units, the dotted Green functions $-2.4G_1$, $-3.2G_2$ of the two loads sum to the solid residual as in (34). (b) The Green function $G(\,\cdot\,;\zeta)$: non-negative, with zero value and slope at both ends and a unit third-derivative jump at ζ. (c) The angular kernels $\Phi_q(\tau) = \widetilde{\Phi}_q(\cos\tau)$ of (29); Φ_C is π-periodic, Φ_R is 2π-periodic and vanishes at $\tau = \pi$.

5 Spurious local minima for ReLU

In this section, we construct a spurious local minimum for plain networks with positive second-layer weights that persists under arbitrary overparameterization.

Theorem 11 (Plain spurious local minima at arbitrary overparameterization, `Lean_10`)**.** *Let $d = 2$ and take the teacher network with unit masses at the angles $(0, 5\pi/6, 4\pi/3)$. For every $n \geq 3$ the plain loss $\mathscr{L}_R$ has a spurious local minimum with all student masses positive. For $n = 3$, the local minimum is strict and for $n > 3$ it is flat.*

Spurious local minima of this loss were first established by Safran & Shamir (2018), for orthonormal teacher networks of width at least six, by a computer-assisted certificate around a candidate proposed by gradient descent: a small gradient, a Hessian floor, and a bound on the third derivative confine a local minimum to a small ball. We follow this paradigm with three changes (full proof ideas in Section A.4). Strong convexity on a whole box around a rational candidate, together with a small gradient at its center, yields a unique strict local minimum inside the box (Lemma 30). Positive loss follows from a conceptual argument analyzing the kinks of the student and teacher networks. This yields the strict spurious local minimum for $n = 3$. At it, the third student unit has positive residual curvature, $\mathcal{F}_R''(\theta_3) > 0$, and positive residual curvature lets a student unit be split into several units with the same direction on which the mass is distributed without losing local minimality (Lemma 31).

The construction achieves persistence through collisions of student features. Are such collisions necessary at large student width? Proposition 12 bounds the number of distinct student directions at critical points.

Proposition 12 (Effective width at overparameterization, `Lean_11`)**.** *Let $d = 2$. A critical point of $\mathscr{L}_R$ with all masses positive with respect to a teacher of width m and unrestricted masses has at most $2m$ pairwise distinct student directions, and for every odd $m \geq 3$ this bound is attained.*

The proof (Section A.4) studies the sign pattern of $\mathcal{W} = \frac{1}{2}(\mathcal{F}_R'' + \mathcal{F}_R)$, a piecewise-sinusoid function with slope jumps at the student and teacher angles. We assign each student direction an interval on which $\mathcal{W} > 0$ and a combinatorial argument gives the desired bound. The bound is sharp for every odd $m \geq 3$, as witnessed by an explicit critical point with $2m$ distinct student directions.

6 Finite sample consequence

We leverage the main result Theorem 10 in the empirical loss setting of a finite dataset $\mathcal{D} = \{(x_j, y_j)\}_{j=1}^N$ with inputs $x_j \in \mathbb{R}^d$ and *teacher-generated* labels $y_j = F_{b^\mathrm{t},t,v}(x_j)$, where $F_{b,s,w}(x) = \langle b, x\rangle + f_{R;s,w}(x)$ is the skip network: every empirical local minimum that persists on

a parameter ball of fixed radius has zero empirical loss once the inputs approximate the Gaussian distribution well enough. The empirical loss is the unweighted sample average

$$\widehat{L}_{\mathcal{D}}(b^{\mathbf{s}}, s, w) = \frac{1}{2N} \sum_{j=1}^{N} \big[F_{b^{\mathbf{s}},s,w}(x_j) - y_j\big]^2, \tag{14}$$

and the approximation is measured on the inputs alone, through their average squared radius $R_2(\mathcal{D})$: the dataset has *accuracy* δ if for every continuous $H\colon \mathbb{R}^d \to \mathbb{R}$ with $H(\tau x) = \tau^2 H(x)$ for $\tau \geq 0$ and every Lipschitz constant L of H on the closed unit ball, with $G \sim \mathcal{N}(0, I_d)$,

$$R_2(\mathcal{D}) = \tfrac{1}{N} \textstyle\sum_j \|x_j\|^2, \qquad \Big|\tfrac{d}{N} \textstyle\sum_j H(x_j) - R_2(\mathcal{D})\, \mathbb{E}H(G)\Big| \leq d\, R_2(\mathcal{D})\, L\delta. \tag{15}$$

For independent, identically distributed inputs from an orthogonally invariant distribution with finite positive second moment, accuracy δ holds with high probability for all sufficiently large N (Remark 32 in Section A.5). Minima are compared on balls of a fixed radius: let $d_J(p, p')$ be the maximum of the distances of the masses, of the directions, and of the linear skips measured from the matching skip b_{match} of Lemma 5 ((79) in Section A.5); a local minimum p is *r-stable* if it minimizes $\widehat{L}_{\mathcal{D}}$ on the closed d_J-ball of radius r about itself.

Theorem 13 (Fixed-radius stability transfer, `Lean_12`)**.** *Let $d \geq 1$, let the teacher network be positive and coplanar, and let the student width satisfy $n \geq m$, as in Theorem 10. For every $r > 0$ there is a $\delta > 0$ such that, for every finite dataset with teacher-generated labels and accuracy δ, every r-stable local minimum with non-negative masses has zero empirical loss. For $d = 1$ the conclusion holds for every finite dataset, with no accuracy requirement.*

We emphasize that this result only concerns r-stable minima and is therefore an application, not a one-to-one transfer of the main result to the empirical loss. The proof ideas (Section A.5) handle $d = 1$ by a direct scaling argument, while for $d \geq 2$ we argue by contradiction. Suppose there are datasets $\mathcal{D}_j$ of accuracy $1/j$ and r-stable minima p_j of positive empirical loss. The minimizer stability theorem of Bastin et al. (2006), Theorem 33, states that cluster points of fixed-radius minima under uniform convergence are local minima of the limit. A scaling identity and the accuracy bound place all p_j in a common compact set K, while on its fixed-radius enlargement the rescaled empirical losses converge uniformly to $\mathscr{L}_{\mathrm{skip}}$. Hence a cluster point p is a population local minimum and thus an exact fit by Theorem 10. The fit is pointwise, so $\widehat{L}_{\mathcal{D}_j}(p) = 0$ for every j, while for large j, p lies in the r-ball about p_j, contradicting $0 < \widehat{L}_{\mathcal{D}_j}(p_j) \leq \widehat{L}_{\mathcal{D}_j}(p) = 0$.

7 Conclusion and Limitations

We expect the planarity assumption to be removable: it enters only through the reduction to the circle in Section 4, while confinement (Proposition 8) and the zero-mass argument (Proposition 9) hold in every dimension.

Conjecture 14. *Let $d \geq 1$, let the teacher network satisfy $(t, v) \in \mathcal{P}^{++}(d, m)$ with any linear skip $b^{\mathbf{t}} \in \mathbb{R}^d$ and let $n \geq m$. Then every local minimum $(b^{\mathbf{s}}, s, w)$ of $\mathscr{L}_{\mathrm{skip}}$ on $\mathbb{R}^d \times \mathbb{R}^n \times (\mathbb{S}^{d-1})^n$ with $(s, w) \in \mathcal{P}^{+}(d, n)$ is an exact fit.*

What is missing is a replacement for the angular residual analysis: the interlacing and descent arguments rely heavily on the one-dimensional order of lines on the circle, and no way to extend these techniques to higher dimensions is known to us. Like most of the loss landscape theory it builds on, our analysis is confined to one hidden layer: the residual representation of Proposition 7 rests on the closed form of the pair moments, which has no counterpart for deeper networks, so the effect of a learned skip on the landscape of deep networks remains open. Moreover, our result on the empirical loss landscape concerns minima that persist on a prescribed parameter ball. Whether one can rule out ordinary positive-loss empirical local minima, without a fixed-radius stability assumption, remains open. Understanding when population-landscape benignity transfers quantitatively to finite samples is an important direction for future work.

Our results are landscape statements and do not imply convergence of a particular optimization algorithm. A natural next question is whether our insights into the geometry of the loss landscape can be used to analyze gradient-based dynamics.

REFERENCES

Yossi Arjevani and Michael Field. Analytic study of families of spurious minima in two-layer ReLU neural networks: A tale of symmetry II. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/806d926414ce19d907700e23177ab4ff-Abstract.html.

Yossi Arjevani and Michael Field. Annihilation of spurious minima in two-layer ReLU networks. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/f3da4165893c2465fd7e8df453c41ffa-Abstract-Conference.html.

Jeremy Avigad. Mathematicians in the age of AI. arXiv preprint arXiv:2603.03684, 2026. URL https://arxiv.org/abs/2603.03684.

Fabian Bastin, Cinzia Cirillo, and Philippe L. Toint. Convergence theory for nonconvex stochastic programming with an application to mixed logit. *Mathematical Programming*, 108(2–3), 2006. doi: 10.1007/s10107-006-0708-6.

Youngmin Cho and Lawrence K. Saul. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems*, volume 22, 2009. URL https://proceedings.neurips.cc/paper/2009/hash/5751ec3e9a4feab575962e78e006250d-Abstract.html.

Johan Commelin, Mateja Jamnik, Rodrigo Ochigame, Lenny Taelman, and Akshay Venkatesh. Shaping the future of mathematics in the age of AI. *Notices of the American Mathematical Society*, 73(6), 2026. URL https://arxiv.org/abs/2603.24914.

Alexandru Damian, Jason D. Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Proceedings of the 35th Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, 2022. URL https://proceedings.mlr.press/v178/damian22a.html.

Leonardo de Moura and Sebastian Ullrich. The Lean 4 theorem prover and programming language. In *Automated Deduction – CADE 28*, volume 12699 of *Lecture Notes in Computer Science*. Springer, 2021. doi: 10.1007/978-3-030-79876-5_37.

Simon S. Du, Xiyu Zhai, Barnabás Póczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1eK3i09YQ.

Dylan J. Foster, Ayush Sekhari, and Karthik Sridharan. Uniform convergence of gradients for non-convex learning and optimization. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/59ab3ba90ae4b4ab84fe69de7b8e3f5f-Abstract.html.

Kenji Fukumizu and Shun-ichi Amari. Local minima and plateaus in hierarchical structures of multilayer perceptrons. *Neural Networks*, 13(3), 2000. doi: 10.1016/S0893-6080(00)00009-5.

Kenji Fukumizu, Shoichiro Yamaguchi, Yoh-ichi Mototake, and Mirai Tanaka. Semi-flat minima and saddle points by embedding neural networks to overparameterization. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/a4ee59dd868ba016ed2de90d330acb6a-Abstract.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016a. doi: 10.1109/CVPR.2016.90.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision – ECCV 2016*, volume 9908 of *Lecture Notes in Computer Science*. Springer, 2016b. doi: 10.1007/978-3-319-46493-0_38.

Kenji Kawaguchi and Yoshua Bengio. Depth with nonlinearity creates no bad local minima in ResNets. *Neural Networks*, 118, 2019. doi: 10.1016/j.neunet.2019.06.009.

Sungyoon Kim, Aaron Mishkin, and Mert Pilanci. Exploring the loss landscape of regularized neural networks via convex duality. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4xWQS2z77v.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/a41b3bb3e6b050b6c9067c67f663b915-Abstract.html.

Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with ReLU activation. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/2017/hash/a96b65a721e561e1e3de768ac819ffbb-Abstract.html.

Shiyu Liang, Ruoyu Sun, Jason D. Lee, and R. Srikant. Adding one neuron can eliminate all bad local minima. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/a012869311d64a44b5a0d567cd20de04-Abstract.html.

A. E. H. Love. *A Treatise on the Mathematical Theory of Elasticity*. Cambridge University Press, fourth edition, 1927. URL https://archive.org/details/in.ernet.dli.2015.462644. Unabridged Dover reprint, 1944.

Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A), 2018. doi: 10.1214/17-AOS1637.

Alireza Mousavi-Hosseini, Sejun Park, Manuela Girotti, Ioannis Mitliagkas, and Murat A. Erdogdu. Neural networks efficiently learn low-dimensional representations with SGD. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6taykzqcPD.

Mert Pilanci and Tolga Ergen. Neural networks are convex regularizers: Exact polynomial-time convex optimization formulations for two-layer networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 2020. URL https://proceedings.mlr.press/v119/pilanci20a.html.

Elijah Polak. On the use of consistent approximations in the solution of semi-infinite optimization and optimal control problems. *Mathematical Programming*, 62, 1993. doi: 10.1007/BF01585175.

Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer ReLU neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 2018. URL https://proceedings.mlr.press/v80/safran18a.html.

Itay Safran, Gilad Yehudai, and Ohad Shamir. The effects of mild over-parameterization on the optimization landscape of shallow ReLU neural networks. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, 2021. URL https://proceedings.mlr.press/v134/safran21a.html.

Ohad Shamir. Are ResNets provably better than linear predictors? In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/26e359e83860db1d11b6acca57d8ea88-Abstract.html.

Berfin Şimşek, François Ged, Arthur Jacot, Francesco Spadaro, Clément Hongler, Wulfram Gerstner, and Johanni Brea. Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 2021. URL https://proceedings.mlr.press/v139/simsek21a.html.

Terence Tao. Mathematics in the age of AI. arXiv preprint arXiv:2608.16753, 2026. URL https://arxiv.org/abs/2608.16753. Submitted to the Proceedings of the International Congress of Mathematicians 2026.

The mathlib Community. The Lean mathematical library. In *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs*, 2020. doi: 10.1145/3372885.3373824.

Yuandong Tian. An analytical formula of population gradient for two-layered ReLU network and its applications in convergence and critical point analysis. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 2017. URL https://proceedings.mlr.press/v70/tian17a.html.

Berk Tinaz, Changzhi Xie, and Mahdi Soltanolkotabi. When both layers learn: Training dynamics of representing linear models via ReLU networks. In *Proceedings of the 39th Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, 2026. URL https://proceedings.mlr.press/v336/tinaz26a.html.

Aad W. van der Vaart. *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 1998. doi: 10.1017/CBO9780511802256.

Luca Venturi, Afonso S. Bandeira, and Joan Bruna. Spurious valleys in one-hidden-layer neural network optimization landscapes. *Journal of Machine Learning Research*, 20(133), 2019. URL https://www.jmlr.org/papers/v20/18-674.html.

Yifei Wang, Jonathan Lacotte, and Mert Pilanci. The hidden convex optimization landscape of regularized two-layer ReLU networks: An exact characterization of optimal solutions. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Z7Lk2cQEG8a.

Frank Zhengqing Wu, Berfin Şimşek, and François Ged. Loss landscape of shallow ReLU-like neural networks: Stationary points, saddle escape, and network embedding. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ogKE7LcvW6.

Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Are deep ResNets provably better than linear predictors? In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/661c1c090ff5831a647202397c61d73c-Abstract.html.

Mo Zhou, Rong Ge, and Chi Jin. A local convergence theory for mildly over-parameterized two-layer neural network. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, 2021. URL https://proceedings.mlr.press/v134/zhou21b.html.

A PROOFS

In this appendix we give the proof ideas and underlying theory of the results and formulas presented in our main part. The following subsections reflect the structure of the main paper:

- Section A.1 (*Auxiliary Lemmas*) proves the results of Section 2: the critical-point transfer under the reparameterization map, the loss decomposition, the centered and plain kernel, and the residual representation of the loss.
- Section A.2 (*Properties of Local Minima*) proves the results of Section 3: feature learning (teacher-span confinement) and the zero-mass exact-fit proposition.
- Section A.3 (*Benign Loss Landscape for Coplanar Teacher Networks*) proves the main result, Theorem 10, through its centered formulation Theorem 28, as presented in Section 4.
- Section A.4 (*Spurious Local Minima for ReLU*) proves the results of Section 5: the trap for plain networks at arbitrary overparameterization and the effective-width bound, via the box certificate and the splitting lemma.
- Section A.5 (*Finite sample consequence*) proves the result of Section 6: the transfer of the main result to empirical loss, via the minimizer stability theorem.

A.1 AUXILIARY LEMMAS

Recall from Definition 3 that a *saddle* is a critical point that is neither a local minimum nor a local maximum. In this subsection we give the proofs and proof ideas of Section 2, namely of the critical-point transfer Lemma 4, the loss decomposition Lemma 5, the pair-moment Lemma 6, and the residual representation Proposition 7.

Proof ideas of Lemma 4. The map $\sigma(s', w') = (s'_i/\|a_i\|,\ \|a_i\| w'_i)_i$ is continuous, satisfies $\Pi \circ \sigma = \mathrm{id}$ and $\sigma(\Pi(o, a)) = (o, a)$. Hence Π maps every neighborhood of (o, a) onto a set containing a neighborhood of $\Pi(o, a)$, and since Π is continuous, the preimage of a neighborhood of $\Pi(o, a)$ is a neighborhood of (o, a). As $\mathscr{L}_q \circ \Pi$ and $\mathscr{L}_q$ take the same values on corresponding parameters, local minimality and global minimality transfer in both directions. For critical points, the chain rule with $s_i = o_i \|a_i\|$ and $w_i = a_i/\|a_i\|$ gives

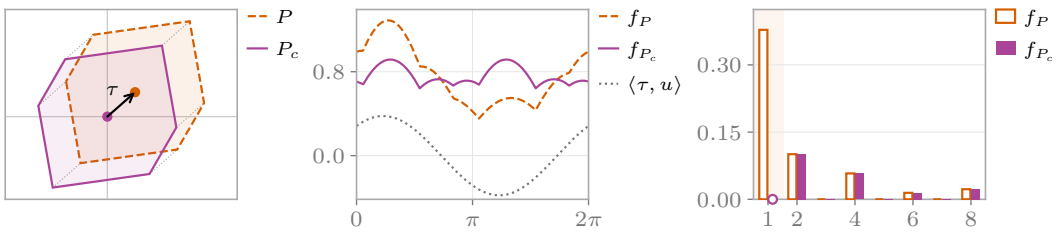
$$\partial_{o_i}(\mathscr{L}_q \circ \Pi) = \|a_i\|\, \partial_{s_i} \mathscr{L}_q, \qquad \nabla_{a_i}(\mathscr{L}_q \circ \Pi) = o_i\, \partial_{s_i} \mathscr{L}_q\ w_i + \frac{1}{\|a_i\|} \nabla_{\mathbb{S}, w_i} \mathscr{L}_q, \tag{16}$$

evaluated at $\Pi(o, a)$, where the two summands of the second gradient are orthogonal. Since $a_i \neq 0$, the raw gradient vanishes if and only if $\partial_{s_i} \mathscr{L}_q = 0$ and also the spherical gradient $\nabla_{\mathbb{S}, w_i} \mathscr{L}_q = 0$ for all i. □

Proof of Lemma 5. Expanding the square in the definition of the skip loss (4) after substituting (6), the cross term between the difference of centered model outputs and the linear residual has zero expectation: the difference of centered model outputs is even in the input and the linear residual is odd, so their product is odd and hence has zero expectation under the centrally symmetric Gaussian law. The squared linear residual has expectation equal to the squared Euclidean norm of its coefficient vector, since the Gaussian has identity covariance. This gives (7); setting both skip vectors to zero gives (8).

The smooth coordinate change $(b^{\mathbf{s}}, s, w) \leftrightarrow (u, s, w)$ reduces the loss to $\mathscr{L}_C(s, w) + \|u\|_2^2/2$. Differentiation gives the critical-point equivalence. Local minimality forces $u = 0$ because moving non-zero u toward zero lowers the loss; restriction to $u = 0$ gives a centered minimum and preserves strictness. Conversely, adding the non-negative quadratic preserves a centered minimum. If that minimum is strict, any different nearby parameter increases either the centered term or the quadratic, proving strictness for the skip loss. □

Figure 3 shows the split (6) behind this proof for a planar student network with non-negative masses, where both networks represent support functions of zonotopes and the linear term is the translation by the Steiner point.



(a) Polytope translation (b) Support functions (c) Fourier amplitudes

Figure 3: The split (6) as a translation of a zonotope, for one planar student network (s, w) with all $s_i > 0$. A network with non-negative masses is the *support function* $u \mapsto \max_{x \in K} \langle u, x \rangle$ of a zonotope K: the plain network of $P = \sum_i s_i[0, w_i]$, the centered network of $P_c = \sum_i s_i[-\frac{1}{2}w_i, \frac{1}{2}w_i]$. (a) The two zonotopes and the shift $\tau = \tau(s, w) = \frac{1}{2}\sum_i s_i w_i$, the *Steiner point*, with $P = P_c + \tau$ (arrow). (b) The support functions $f_P = f_{R;s,w}$ and $f_{P_c} = f_{C;s,w}$ on the unit circle $u = (\cos\theta, \sin\theta)$; their difference is the odd term $\langle \tau, u \rangle$, and f_{P_c} is π-periodic. (c) Angular Fourier amplitudes of $\theta \mapsto f_P(u)$ and $\theta \mapsto f_{P_c}(u)$: centering removes the degree-one coefficient and leaves the even degrees unchanged. Support-function language needs $s_i \geq 0$; (6) and Lemma 5 hold for potentially negative masses.

Proof of Lemma 6. As only the projection of the Gaussian X on $\operatorname{span}(\xi, \zeta)$ enters, we may assume $d = 2$ and by rotational invariance we may assume $\xi = (1, 0)$ and $\zeta = (\cos\vartheta, \sin\vartheta)$. The two ReLU factors $[R\cos a]_+$ and $[R\cos(a - \vartheta)]_+$ are both positive exactly for $a \in (\vartheta - \pi/2, \pi/2)$, so, writing the projected input in polar coordinates $(R\cos a, R\sin a)$ with a uniform on $[0, 2\pi)$ independent of R and $\mathbb{E}R^2 = 2$,

$$M_R(\xi, \zeta) = \frac{\mathbb{E}R^2}{2\pi} \int_{\vartheta - \pi/2}^{\pi/2} \cos a \cos(a - \vartheta)\, da = \frac{\sin\vartheta + (\pi - \vartheta)\cos\vartheta}{2\pi}, \tag{17}$$

where $2\cos a\cos(a - \vartheta) = \cos\vartheta + \cos(2a - \vartheta)$ has the antiderivative $a\cos\vartheta + \frac{1}{2}\sin(2a - \vartheta)$, and $\sin(\pi - \vartheta) - \sin(\vartheta - \pi) = 2\sin\vartheta$, yielding the derived equation. Substituting $\sin\vartheta = \sqrt{1 - \rho^2}$ and $\pi - \vartheta = \pi/2 + \arcsin\rho$ gives $2\pi M_R = \widetilde{\Phi}_C(\rho) + \frac{\pi}{2}\rho$. For the centered feature, $\varphi_R(z) = \varphi_C(z) + z/2$ and the cancellation of the even–odd cross terms under $X \mapsto -X$ give $M_R = M_C + \frac{1}{4}\mathbb{E}[\langle \xi, X \rangle \langle \zeta, X \rangle] = M_C + \rho/4$, yielding $2\pi M_C = \widetilde{\Phi}_C(\rho)$. □

Proof of Proposition 7. Expanding the square in (5) and applying Lemma 6 to every pair of unit directions gives

$$4\pi \mathscr{L}_q = \sum_{i,j} s_i s_j \widetilde{\Phi}_q(\langle w_i, w_j \rangle) - 2\sum_{i,k} s_i t_k \widetilde{\Phi}_q(\langle w_i, v_k \rangle) + \sum_{k,l} t_k t_l \widetilde{\Phi}_q(\langle v_k, v_l \rangle). \tag{18}$$

Regrouping the right side as

$$\begin{aligned} &\sum_i s_i \Big[\sum_j s_j \widetilde{\Phi}_q(\langle w_i, w_j \rangle) - \sum_k t_k \widetilde{\Phi}_q(\langle w_i, v_k \rangle) \Big] \\ &- \sum_k t_k \Big[\sum_j s_j \widetilde{\Phi}_q(\langle v_k, w_j \rangle) - \sum_l t_l \widetilde{\Phi}_q(\langle v_k, v_l \rangle) \Big] \end{aligned} \tag{19}$$

and using the definition of the residual (10) gives (11). The right side of (18) is a quadratic polynomial in the student masses; differentiating it in s_i collects the two symmetric student sums $\sum_j s_j \widetilde{\Phi}_q(\langle w_i, w_j \rangle)$ and the teacher sum $2\sum_k t_k \widetilde{\Phi}_q(\langle w_i, v_k \rangle)$ into $2\mathcal{F}_q(w_i)$, which is (12).

For the angle identity of (12), move w_i along the sphere, $w_i(\tau) = w_i + \tau h$ with $h \perp w_i$. The diagonal term $s_i^2 \widetilde{\Phi}_q(1)$ is constant, and every other term containing w_i appears twice in the symmetric student sum and once in the teacher sum, so

$$4\pi \frac{d}{d\tau}\Big|_{\tau=0} \mathscr{L}_q = 2s_i \Big[\sum_{j \neq i} s_j \widetilde{\Phi}_q'(\langle w_i, w_j \rangle) \langle h, w_j \rangle - \sum_k t_k \widetilde{\Phi}_q'(\langle w_i, v_k \rangle) \langle h, v_k \rangle \Big]$$

$$= 2s_i\, D\mathcal{F}_q(w_i)[h], \tag{20}$$

where $D\mathcal{F}_q(w_i)[h]$ is the derivative of the probe $u \mapsto \mathcal{F}_q(u)$ at $u = w_i$ in the tangential direction h. This is the stated spherical-gradient identity. □

A.2 Properties of Local Minima

In this subsection we give the proof ideas of Section 3, namely of the feature-learning Proposition 8 and the zero-mass Proposition 9.

Proof ideas of Proposition 8. Fix $q \in \{C, R\}$, the input dimension $d \geq 2$, teacher parameters (t, v) with arbitrary masses, and student parameters (s, w) with non-negative masses $s_i \geq 0$ constituting a local minimum of the q-loss $\mathscr{L}_q$. Let us denote the linear span of the teacher directions, the *teacher span*, by $\mathcal{T} = \operatorname{span}\{v_k\}$. Decompose each student direction as $w_i = p_i + q_i$, with $p_i \in \mathcal{T}$ and $q_i \perp \mathcal{T}$. We aim to show that $q_i = 0$ for every positive-mass student unit.

We may assume all student masses to be positive: zero-mass units may be pruned without changing the represented function or the local minimality assumption. If no units remain, the claim is immediate. Absorb the masses into the student weights and work with the generators $V_i = s_i w_i$, interpreted as raw first-layer weights with fixed unit output weights. This preserves the network function and local minimality by Lemma 4. In these coordinates, write

$$V = (s_1 w_1, \ldots, s_n w_n), \qquad f_V(x) = \sum_i \varphi_q(\langle V_i, x\rangle).$$

The idea of the proof is to compare the local curvature of the loss along two motions: the *normal motion* stretches each $s_i q_i$ while keeping the projection of the generator onto $\mathcal{T}$ fixed, whereas the *radial motion* scales all generators uniformly. Their velocities and paths are

$$\begin{aligned} z_N &= (s_1 q_1, \ldots, s_n q_n), & V + \tau z_N &= (s_i p_i + (1+\tau) s_i q_i)_i, \\ z_A &= V, & V + a z_A &= (1+a)V. \end{aligned}$$

We call the first and second variations along these motions the *normal and radial derivatives and curvatures*, respectively.

We take special care to justify the second and mixed derivatives used below, without assuming that the full loss admits a Hessian everywhere. The proof shows that its restriction to $V + \mathcal{W}$, where $\mathcal{W} = \operatorname{span}\{z_A, z_N, z_{e,b}\}$ and the shear velocities $z_{e,b}$ are defined below, is C^2 near V. Write H for the Hessian of this restriction at V, acting on $\mathcal{W}$. Thus, for $z, z' \in \mathcal{W}$,

$$\begin{aligned} \partial_z \mathscr{L}_q(V) &:= \frac{d}{d\tau}\Big|_{\tau=0} \mathscr{L}_q(V + \tau z), \\ z^\top H z' &= \partial_a \partial_\tau \Big|_{(a,\tau)=(0,0)} \mathscr{L}_q(V + az + \tau z'). \end{aligned} \tag{21}$$

Local minimality implies $\partial_z \mathscr{L}_q(V) = 0$ and $z^\top H z \geq 0$ for every $z \in \mathcal{W}$.

The first step shows that the normal curvature is zero. The argument evaluates its sum with the curvatures along *shear motions*: these move each generator within $\mathcal{T}$, with velocity determined by its normal component. For e and b ranging over orthonormal bases of $\mathcal{T}$ and $\mathbb{R}^d$, respectively, define

$$z_{e,b} = \big(s_1 \langle b, q_1\rangle e, \ldots, s_n \langle b, q_n\rangle e\big).$$

For $q = C$, Parseval's identity and cancellations in the explicit curvature formulas show that the sum of the normal and shear curvatures equals $(\dim \mathcal{T} + 1)$ times the normal derivative of the loss:

$$z_N^\top H z_N + \sum_{e,b} z_{e,b}^\top H z_{e,b} = (\dim \mathcal{T} + 1)\, \partial_{z_N} \mathscr{L}_C(V). \tag{22}$$

For $q = R$, the additional first-moment quadratic satisfies the same identity:

$$\Delta(V) := \mathscr{L}_R(V) - \mathscr{L}_C(V) = \frac{1}{8} \left\| \sum_i V_i - \sum_k t_k v_k \right\|_2^2,$$

$$\partial^2_{z_N}\Delta(V) + \sum_{e,b} \partial^2_{z_{e,b}}\Delta(V) = (\dim \mathcal{T} + 1)\, \partial_{z_N}\Delta(V). \tag{23}$$

Adding this identity proves (22) with $\mathscr{L}_C$ replaced by $\mathscr{L}_R$. In either case, the normal derivative vanishes at the local minimum, while all curvatures on the left are non-negative. Consequently, $z_N^\top H z_N = 0$.

This also forces the mixed normal–radial curvature to vanish: by Cauchy–Schwarz for the positive-semidefinite restricted Hessian, its square is bounded by the product of the normal and radial curvatures,

$$(z_A^\top H z_N)^2 \le (z_A^\top H z_A)(z_N^\top H z_N) = 0.$$

Next, we identify the mixed curvature with half the normal derivative of the *second student moment* $\mathbb{E}[f_V(X)^2]$, a quantity independent of the teacher network. To see this, first differentiate the loss along the normal motion: this gives the expected product of the student–teacher difference and the normal derivative of the student output,

$$\frac{d}{d\tau}\mathscr{L}_q(V + \tau z_N) = \mathbb{E}\left[\big(f_{V+\tau z_N} - f_{q;t,v}\big)(X)\frac{d}{d\tau} f_{V+\tau z_N}(X)\right], \tag{24}$$

$$\frac{d}{d\tau} f_{V+\tau z_N}(x) = \sum_i \varphi_q'\big(\langle V_i + \tau (z_N)_i, x\rangle\big)\, \langle (z_N)_i, x\rangle\,. \tag{25}$$

These feature derivatives exist almost everywhere. Now differentiate this product along the radial motion, keeping the normal displacement z_N fixed: positive homogeneity makes the radial derivative of the student–teacher difference equal to f_V, while the student network's normal response remains unchanged almost everywhere under positive radial scaling. Dominated convergence justifies differentiating under the expectation. The remaining product therefore gives

$$z_A^\top H z_N = \mathbb{E}\left[f_V(X)\,\frac{d}{d\tau}\Big|_{\tau=0} f_{V+\tau z_N}(X)\right] = \frac{1}{2}\,\frac{d}{d\tau}\Big|_{\tau=0}\mathbb{E}\big[f_{V+\tau z_N}(X)^2\big]. \tag{26}$$

Let us now assume that some $q_i \neq 0$. Write $P = V - z_N = (s_i p_i)_i$ for the projected generators and parameterize the second student moment by their normal scale:

$$M(c) := \mathbb{E}\big[f_{P+cz_N}(X)^2\big], \qquad\qquad V + \tau z_N = P + (1+\tau) z_N.$$

Thus $M(0)$ is the projected student network's second moment, $M(1)$ is the original student network's second moment, and $z_A^\top H z_N = \frac{1}{2}M'(1)$.

To show that this normal derivative is positive, first observe that the second moment is convex along the normal motion. Indeed, for fixed input the pre-activations depend affinely on the normal scale:

$$f_{P+cz_N}(x) = \sum_i \varphi_q\big(\langle P_i, x\rangle + c\,\langle (z_N)_i, x\rangle\big).$$

Since both activations are convex and non-negative, this output and its square are convex in c. As integration preserves convexity, so $c \longmapsto M(c)$ is a convex function. Geometrically, the derivative at the original generators is at least the secant slope from the projected generators: $M'(1) \ge M(1) - M(0)$. It therefore suffices to show that the normal parts of the generators increase the second student moment, this is $M(1) \ge M(0)$.

For this comparison, reflect each input x across the teacher span $\mathcal{T}$, and write Rx for its reflection. Their midpoint is the projection of x onto $\mathcal{T}$. Since the student output as a function of the network input is convex and non-negative, its squared output at this midpoint is at most the average of its squared outputs at x and Rx. Evaluating the student network at the projected input is equivalent to evaluating the student network with projected generators P at the original input.

If some $s_i q_i \neq 0$, this comparison is strict at $x_0 = s_i q_i$: the projected student output vanishes there, whereas the i-th original student unit contributes positively. By continuity, strictness persists on an open neighborhood, which has positive Gaussian probability by full support. Since reflection preserves the Gaussian law, integrating the comparison shows that the projected second moment $M(0)$ is strictly smaller than the original second moment $M(1)$. As this leads to $z_A^T H z_N = \frac{1}{2}M'(1) > 0$, a contradiction to the first step, this closes the proof of Proposition 8. □

Proof ideas of Proposition 9. Let the input dimension $d \geq 1$ and let the student masses and directions (s, w) constitute a local minimum of the centered loss $\mathscr{L}_C$ (for $d \geq 1$) or the plain loss $\mathscr{L}_R$ (for $d \geq 3$), and let us assume that the teacher masses are non-negative in case $d = 2$ (for $\mathscr{L}_C$). Moreover, let us assume that there is a zero-mass unit (s_i, w_i). We want to prove that the student network is an exact fit.

In the first step we show for both $q \in \{C, R\}$ that the residual $\mathcal{F}_q$, see (10), vanishes on an open set $U \subseteq \mathbb{S}^{d-1}$, see `Lean`. The argument analyzes the change of loss when altering the mass s_i and direction w_i of the zero-mass unit. In case the residual does not vanish on an open set around the direction w_i, we would be able to find student parameters arbitrarily close to (s, w) with smaller loss, a contradiction. The argument works using the representation of the loss as a quadratic polynomial in the student mass coordinates with the values of the residual at the respective student and teacher directions as the coefficients (see Proposition 7, `Lean`).

This representation also comes in handy to show the final statement of the theorem, namely that the loss is zero: by Proposition 7 it suffices to show that the residual vanishes at the student and teacher directions, see `Lean`. Hence, in our second step we show that the residual vanishes on the whole sphere, see `Lean`. For $d \geq 3$ this follows from an elegant argument using the fact that an analytic function vanishing on an open set vanishes everywhere. The centered kernel (9) is analytic on $(-1, 0) \cup (0, 1)$ and the plain kernel is analytic on $(-1, 1)$; in either case it is easy to see that the residual is analytic away from the finitely many points $\{\pm w_i, \pm v_k \mid i, k\}$, which we call its *atoms*. Hence, it is also analytic restricted to *great circles*, the intersection of the sphere with a hyperplane. We observe that for any point x on the sphere $\mathbb{S}^{d-1}$, there is a great circle through it that intersects the zero set U in an open set and avoids the finitely many residual atoms. As the residual restricted to this great circle has to vanish, it vanishes at x. As the residual vanishes at all points on the sphere except the atoms, its continuity implies it also vanishes at the atoms. This concludes the proof for the case $d \geq 3$.

For $d = 1$, the residual is centrally symmetric, so the zero set from the first step extends to the whole sphere. For $d = 2$, where the unit circle $\mathbb{S}^{d-1}$ is chopped into pieces by the residual atoms, we use the machinery developed in Section A.3 and non-negative teacher weights to propagate the zero set inductively around the circle.

This finishes the proof of Proposition 9. □

A.3 Benign Loss Landscape for Coplanar Teacher Networks

In this subsection we give the proof ideas of the main result, Theorem 10 through its centered formulation, Theorem 28.

Denote by $(t, v) \in \mathcal{P}^{++}(d, m) \cap \mathcal{P}_{\mathrm{pl}}(d, m)$ the teacher network's mass–direction vector pair throughout, and let the student width satisfy $n \geq m$. Whenever we call a student parameter (s, w) a local minimum or critical point of $\mathscr{L}_C$, it is with respect to these fixed teacher parameters (t, v). The positivity of the teacher masses enters through one fact only: a line of positive net mass carries a student unit of positive mass, compare to the definition of the net mass ω in (13). Lemmas 16 to 18, 20 and 21 do not use teacher mass non-negativity and hold for unrestricted teacher masses.

A hypothetical spurious local minimum is progressively restricted by four facts. A zero-mass student unit forces an exact fit (Proposition 9). Otherwise, confinement (Proposition 8) places every student direction in the teacher span, which reduces the problem to the plane. There, the necessary positive-mass configuration of matched width and strictly interlacing student and teacher lines already resolves the overparameterized case (Proposition 25). For the matched width case, we need to take one step further: in the given setting Proposition 26 constructs a direction in parameter space certified to decrease the loss, so also this last case cannot be a local minimum. Finally, Lemma 5 carries the result about the centered to the skip network (Theorem 10). For the analysis below take $d \geq 2$; the case $d = 1$ is treated directly in the proof of Theorem 28. By the assumption, the teacher directions $\mathcal{T} = \operatorname{span}\{v_k\}$ are contained in a two-dimensional linear subspace that we identify with $\mathbb{R}^2$.

With student and teacher directions in $\mathbb{R}^2$, it is convenient to work with angles $e(\gamma)$, θ_i, β_k and lines $\ell(u)$ of Section 4 (Figure 1a); every direction $u \in \mathbb{S}^1$ is $e(\gamma)$ for a unique angle $\gamma =: \theta(u) \in \mathbb{R}/2\pi\mathbb{Z}$, and the line of an angle γ is $\ell(\gamma) = \gamma \bmod \pi$, so that $\ell(u) = \ell(\theta(u))$. Throughout, α denotes a line and γ, θ, β denote angles.

We call (t, β) and (s, θ) the *angular representation* of the teacher network and the student network, respectively, and write $\mathscr{L}_q(s, \theta)$ for the loss in these coordinates. Let $\Lambda_v \coloneqq \{\ell(v_k) \mid k \in [m]\}$ and $\Lambda_w \coloneqq \{\ell(w_i) \mid i \in [n]\} \subset \mathbb{R}/\pi\mathbb{Z}$ be the finite sets of teacher and student lines and their union $\Lambda_{v,w} = \Lambda_v \cup \Lambda_w$. The net mass ω (13) vanishes off $\Lambda_{v,w}$, so $S_\pm \subseteq \Lambda_{v,w}$.

Our planar theorems mostly rely on the connection of the loss with the residual presented in Proposition 7; the key property we exploit is the residual's relation to the *angular kernel*, the kernel (9) evaluated at the angle τ between two directions. As $\arcsin(\cos\tau) = \frac{\pi}{2} - |\tau|$ and $\sqrt{1-\cos^2\tau} = |\sin\tau|$ for $|\tau| \le \pi$, it is explicitly given by

$$\Phi_q(\tau) = \widetilde{\Phi}_q(\cos\tau), \tag{27}$$

$$\Phi_C(\tau) = \Big(\frac{\pi}{2} - |\tau|\Big)\cos\tau + |\sin\tau|, \tag{28}$$

$$\Phi_R(\tau) = (\pi - |\tau|)\cos\tau + |\sin\tau| \tag{29}$$

extended 2π-periodically. The key identity is

$$\Phi_q''(\tau) + \Phi_q(\tau) = 2|\sin\tau| \quad \text{for} \quad q \in \{C, R\}, \tag{30}$$

which we will use to determine other representations of the angular residual. Before this work, let us observe that the angular kernel is even and 2π-periodic, and Φ_C is moreover π-periodic with $1 \le \Phi_C \le \pi/2$, while $\Phi_R = \Phi_C + \frac{\pi}{2}\cos\tau$ satisfies $\Phi_R(\pi) = 0$.

The *student potential* $S_q(\gamma) = \sum_i s_i \Phi_q(\gamma - \theta_i)$ and *teacher potential* $P_q(\gamma) = \sum_k t_k \Phi_q(\gamma - \beta_k)$ give the *angular residual* $\mathcal{F}_q \colon \mathbb{R}/2\pi\mathbb{Z} \to \mathbb{R}$, namely $\mathcal{F}_q = S_q - P_q$:

$$\mathcal{F}_q(\gamma) = \sum_i s_i \Phi_q(\gamma - \theta_i) - \sum_k t_k \Phi_q(\gamma - \beta_k). \tag{31}$$

Compare this to the non-angular residual definition in (10) and take a look at Figure 4 for intuition about the connection of the mass and angular gradient of the student units to the angular residual. For $q = C$ the angular residual is π-periodic, hence a function of lines, and collecting the units on each line gives $\mathcal{F}_C(\gamma) = \sum_{\alpha \in \Lambda_{v,w}} \omega(\alpha)\Phi_C(\gamma - \alpha)$, where $\Phi_C(\gamma - \alpha)$ is well defined because Φ_C is π-periodic.

Remark 15 (`Lean_13`)**.** For $\gamma \in [0, 2\pi]$ with $u = e(\gamma)$, the angular residual satisfies $\mathcal{F}_q(\gamma) = \mathcal{F}_q(u)$ as well as $\partial_{s_i}\mathscr{L}_q = \frac{1}{2\pi}\mathcal{F}_q(\theta_i)$ and $\partial_{\theta_i}\mathscr{L}_q = \frac{1}{2\pi}s_i\mathcal{F}_q'(\theta_i)$.

A point $\alpha \in \mathbb{R}/\pi\mathbb{Z}$ is a *double zero* of $\mathcal{F}_C$ if $\mathcal{F}_C(\alpha) = \mathcal{F}_C'(\alpha) = 0$.

Lemma 16 (Double zeros of $\mathcal{F}_C$ at critical points, `Lean_14`)**.** *If (s, w) is a critical point of $\mathscr{L}_C$, then $\ell(w_i)$ is a double zero of $\mathcal{F}_C$ for all i with $s_i \neq 0$.*

Let us define the beam operator

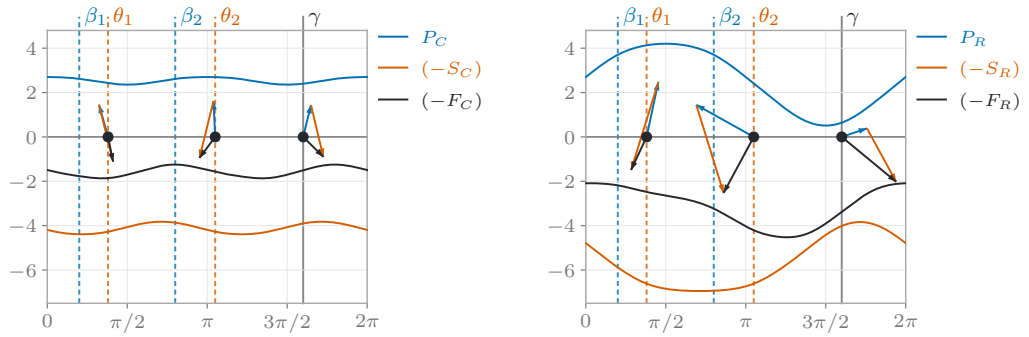
$$L = (D^2 + 1)^2. \tag{32}$$

In the following lemma we prove that the kernel of L on every interval is the following four-dimensional space of functions on $\mathbb{R}$.

$$\mathcal{V} = \operatorname{span}\{\cos x, \sin x, x\cos x, x\sin x\} \tag{33}$$

Lemma 17 (The kernel of the beam operator, `Lean_15`)**.** *Let $I \subseteq \mathbb{R}$ be a non-empty open interval. A function $f \in C^4(I)$ satisfies $Lf = 0$ on I if and only if f is the restriction to I of an element of $\mathcal{V}$.*

Proof ideas of Lemma 17. Each function of $\mathcal{V}$ solves $Lf = 0$ by direct differentiation; the four generators correspond to the double roots $\pm i$ of the characteristic polynomial $(\lambda^2 + 1)^2$. They are linearly independent on I: the combination $g = a\cos x + b\sin x + cx\cos x + dx\sin x$ has the jet $(g, g', g'', g''')(0) = (a,\ b + c,\ -a + 2d,\ -b - 3c)$ at 0, which vanishes only for $a = b = c = d = 0$, and a real-analytic function vanishing on an interval vanishes on $\mathbb{R}$. Conversely, L is a linear differential operator of order four with continuous coefficients, so by the uniqueness theorem for linear ordinary differential equations a solution on I is determined by its values (f, f', f'', f''') at one point of I; hence the kernel has dimension at most four and is spanned by the four restricted generators. □



(a) Centered: mass rate $2\pi\dot{s}$ (b) Plain: mass rate $2\pi\dot{s}$

Figure 4: Pull, push, and descent for an added probe of mass $s = 1$. Background teacher units lie at $\beta = (0.2\pi, 0.8\pi)$ (blue dashed guides) and student units at $\theta = (0.38\pi, 1.05\pi)$ (orange dashed guides), all with mass 1. Solid curves show P_q (blue), $-S_q$ (orange), and $-\mathcal{F}_q = P_q - S_q$ (black), evaluated at γ for the configuration with the probe adjoined there (see (31) for the definition of $\mathcal{F}_q$). By Remark 15 the value of the residual gives the gradient of the student masses while the slope gives the gradient for the student angles. Thus angular pull climbs the teacher potential, while angular push descends the student potential. The three dots at the student angles and a third angle each consider one probe separately; blue and orange arrows add head to tail to the black descent arrow.

Lemma 18 (The angular residual on a gap, `Lean_16`)**.** *Let (a, b) be a gap of the set S_+ of lines of positive net mass* (13)*, of length $l = b - a \in (0, \pi]$, whose endpoints a, b are double zeros of the centered angular residual $\mathcal{F}_C$* (31)*. Then there is a function $G(\cdot\,; \zeta)\colon [0, l] \to [0, \infty)$, depending on l and on a parameter $\zeta \in (0, l)$, such that the residual on (a, b) is a sum of contributions of the points of $(a, b) \cap S_-$ only,*

$$\mathcal{F}_C(a + z) = 4 \sum_{p \in S_- \cap (a,b)} \omega(p)\, G(z;\, p - a) \qquad (0 < z < l). \tag{34}$$

Consequently, since $\omega(p) < 0$ for $p \in S_-$,

1. *$\mathcal{F}_C \equiv 0$ on $[a, b]$ if and only if $(a, b) \cap S_- = \varnothing$, and otherwise $\mathcal{F}_C < 0$ on (a, b);*
2. *if $(a, b) \cap S_- \neq \varnothing$, then $\mathcal{F}_C''(a) < 0$ and $\mathcal{F}_C''(b) < 0$.*

Figure 2 illustrates the double zeros and the Green superposition.

Proof ideas of Lemma 18. The key observation is that $\mathcal{F}_C$ is a piecewise-$\mathcal{V}$ function whose breakpoints are given by $S_+ \cup S_-$. Furthermore, since $\mathcal{F}_C \in C^2$, the left- and the right-hand pieces agree up to second order at each breakpoint; only the third derivative $\mathcal{F}_C'''$ exhibits a jump discontinuity. Consequently, the main idea is to construct a C^2, piecewise-$\mathcal{V}$ function $G(\cdot\,; \zeta)$ for each breakpoint ζ of $\mathcal{F}_C(a + \cdot)\big|_{[0,l]}$. This function $G(\cdot\,; \zeta)$ is designed to have a unique breakpoint at ζ with a specified jump in $G'''(\zeta)$, while simultaneously satisfying the boundary conditions at the endpoints of $[0, l]$. A dimension count argument for the underlying function space then guarantees that $\mathcal{F}_C(a + \cdot)\big|_{[0,l]}$ can be uniquely represented as a weighted sum (34) of these functions $G(\cdot\,; \zeta)$. Finally, claims 1.–2. reduce to verifying two key properties of $G(\cdot\,; \zeta)$: its non-negativity on $[0, l]$ and its strict convexity at the endpoints 0 and l.

To demonstrate that $\mathcal{F}_C$ is piecewise-$\mathcal{V}$, recall from (13) that $\mathcal{F}_C(\gamma) = \sum_{\alpha \in S_+ \cup S_-} \omega(\alpha) \Phi_C(\gamma - \alpha)$. According to (30), the angular kernel Φ_C satisfies $(D^2 + 1)\Phi_C(\gamma) = 2|\sin(\gamma)|$. Applying this differential operator to $\mathcal{F}_C$ yields $(D^2 + 1)\mathcal{F}_C(\gamma) = 2 \cdot \sum_{\alpha \in S_+ \cup S_-} \omega(\alpha) |\sin(\gamma - \alpha)|$. Because functions of the form $\pm\sin(\cdot - \alpha)$ and their linear combinations form the kernel of $D^2 + 1$, it follows that $(D^2 + 1)^2 \mathcal{F}_C\big|_I \equiv 0$ on any interval $I \subset \mathbb{R}/\pi\mathbb{Z}$ with $I \cap (S_+ \cup S_-) = \emptyset$. By Lemma 17, the restriction of $\mathcal{F}_C$ to I coincides with that of some function from $\mathcal{V}$. The breakpoints of $\mathcal{F}_C$ are therefore given by $S_+ \cup S_-$; they are characterized by jump of the third derivative: an explicit

calculation shows that $\Phi_C'''(0^+) - \Phi_C'''(0^-) = 4$, hence at each breakpoint $\alpha \in S_+ \cup S_-$ we have $\mathcal{F}_C'''(\alpha^+) - \mathcal{F}_C'''(\alpha^-) = 4\omega(\alpha)$.

Formula (34) follows from a more general decomposition that we establish below. Specifically, any piecewise $\mathcal{V}$ function $f \in C^2([0,l])$ satisfying the boundary conditions $0 = f(0) = f'(0) = f(l) = f'(l)$ admits the representation

$$f(\,\cdot\,) = \sum_{\zeta \text{ is a breakpoint of } f} c_{\zeta,f} G(\,\cdot\,;\zeta), \quad \text{where} \quad c_{\zeta,f} := f'''(\zeta^+) - f'''(\zeta^-). \tag{35}$$

Here, a breakpoint refers to a point of discontinuity in the third derivative of f. The basis functions $G(\,\cdot\,;\zeta) \in C^2([0,l])$ are constructed below as the unique functions satisfying the double-zero boundary conditions at 0 and l, while possessing a unique breakpoint at ζ with a unit third-derivative jump, i.e. $c_{\zeta,G(\,\cdot\,;\zeta)} = 1$, and belonging to $\mathcal{V}$ on each side of ζ. For any partition $0 =: \zeta_0 < \zeta_1 < \ldots < \zeta_k < \zeta_{k+1} := l$, these piecewise-$\mathcal{V}$ functions with specified breakpoints form a linear space. Explicitly, this space consists of all functions $f \in C^2([0,l])$ satisfying $0 = f(0) = f'(0) = f(l) = f'(l)$ such that for each $i = 0, \ldots, k$ there exists $f_i \in \mathcal{V}$ with $f_i\big|_{[\zeta_i,\zeta_{i+1}]} \equiv f\big|_{[\zeta_i,\zeta_{i+1}]}$. We prove that this space is k-dimensional and $G(\,\cdot\,;\zeta_i)$ indeed form its basis.

First, we show that if $c_{\zeta_i,f} = 0$ for all $i = 1, \ldots, k$, then $f \equiv 0$. Indeed, the assumption implies $f_{i-1}'''(\zeta_i) = f'''(\zeta_i^-) = f'''(\zeta_i^+) = f_i'''(\zeta_i)$. Since f_{i-1} and f_i both lie in the kernel of the differential operator $(D^2+1)^2$ and coincide up to third order at ζ_i, the existence and uniqueness theorem yields $f_{i-1} \equiv f_i$. Consequently, f globally belongs to $\mathcal{V}$; it remains to show that double zeros of f at 0 and l imply $f \equiv 0$. The conditions $f(0) = f'(0) = 0$ leave the two-parameter family $f(x) = \lambda_1(x\cos x - \sin x) + \lambda_2 x \sin x$, while the double zero at l is equivalent to the linear system:

$$M_l \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = \begin{pmatrix} \kappa_1 \\ \kappa_2 \end{pmatrix} \quad \text{with} \quad M_l := \begin{pmatrix} l\cos l - \sin l & l\sin l \\ -l\sin l & l\cos l + \sin l \end{pmatrix}, \tag{36}$$

where $\kappa_1 = \kappa_2 = 0$. Since $\det M_l = l^2 - \sin^2 l > 0$ for all $l \neq 0$, the matrix M_l is invertible, thus $\lambda_1 = \lambda_2 = 0$ is the unique solution.

Second, we demonstrate that for each $i = 1, \ldots, k$ there exists a piecewise-$\mathcal{V}$ function $G(\,\cdot\,;\zeta_i)$ satisfying the boundary conditions at 0 and l, having its unique breakpoint at ζ_i with a unit jump $c_{\zeta_i,G(\,\cdot\,;\zeta_i)} = 1$. Consider the auxiliary function $\psi(\gamma) := \sin\gamma - \gamma\cos\gamma$. A direct calculation verifies that $(\psi, \psi', \psi'', \psi''')(0) = (0,0,0,2)$. Consequently, we construct

$$\begin{aligned} G(\gamma;\zeta_i) &:= \frac{1}{2}\psi(\gamma - \zeta_i)\mathbf{1}_{\gamma > \zeta_i} + g(\gamma) \quad \text{with} \\ g(\gamma) &:= \begin{pmatrix} \gamma\cos\gamma - \sin\gamma \\ \gamma\sin\gamma \end{pmatrix}^\top M_l^{-1} \begin{pmatrix} -\psi(l-\zeta_i)/2 \\ -\psi'(l-\zeta_i)/2 \end{pmatrix}. \end{aligned} \tag{37}$$

Here $g \in \mathcal{V}$ is determined by the unique solution to the linear system (36) with appropriate parameters κ_1, κ_2 chosen so that $G(\,\cdot\,;\zeta_i)$ satisfies the boundary conditions.

Applying the above construction (35) to a gap (a,b) and the piecewise-$\mathcal{V}$ function $\mathcal{F}_C$ yields (34). An explicit expression for the function $G(\,\cdot\,;\zeta)$ illustrated in Figure 2(b) comes from (37). Direct calculations show that $G(\gamma;\zeta) > 0$ for all $\gamma \in (0,l)$, and that the second derivatives satisfy $G''(0;\zeta), G''(l;\zeta) > 0$ for any $l \in (0,\pi]$. This finishes the proof of Lemma 18. □

Remark 19 (The angular residual as a bent ring)**.** The beam operator L (32) has a classical interpretation in elasticity. The equations for in-plane bending of a thin elastic circular ring under a radial load X lead to the fourth-order displacement equation (Love, 1927, Art. 292)

$$(D^2+1)^2 u \propto X + C, \tag{38}$$

where u is the radial displacement and C is a constant corresponding to uniform radial loading and $\propto$ denotes equality up to constants. Thus, up to physical scaling and this uniform load C which we assume to be zero, the point forces

$$X = \sum_\alpha \omega(\alpha)\delta_\alpha$$

produce the same differential equation as our residual,

$$(D^2+1)^2\mathcal{F}_C = 4\sum_\alpha \omega(\alpha)\delta_\alpha. \tag{39}$$

Accordingly, $\mathcal{F}_C$ may be viewed as the radial displacement of a circular ring under point forces at the student and teacher lines, with strengths proportional to their net masses.

Lemma 20 (Interlacing, `Lean_17`)**.** *Suppose that $|S_+| \geq 2$ and that every point of S_+ is a double zero of $\mathcal{F}_C$. Then every gap of S_+ contains a point of S_-. Consequently, $|S_+| \leq |S_-|$.*

Proof ideas of Lemma 20. Let (a,b) be a gap of S_+. Its endpoints are double zeros, so Lemma 18 gives

$$(a,b)\cap S_- = \varnothing \implies F_C''(a) = F_C''(b) = 0 \text{ and } (a,b)\cap S_- \neq \varnothing \implies F_C''(a), F_C''(b) < 0. \tag{40}$$

Then, since $\mathcal{F}_C''$ is continuous and neighboring gaps share an endpoint, it follows that either all gaps must contain a point of S_- or none does. The latter option is impossible, since $S_- = \emptyset$ would imply that $\mathcal{F}_C > 0$, contradicting the assumption that the points in S_+ are double zeros. Hence all gaps must contain a point of S_-. The $|S_+|$ gaps are pairwise disjoint, so $|S_+| \leq |S_-|$. □

Lemma 21 (Colliding student units provide descent in the centered case, `Lean_18`)**.** *Let (s,θ) be a planar student parameter in which two units $i \neq j$ lie on a common line $\ell(\theta_i) = \ell(\theta_j) =: \alpha$, and define $z = (\delta s, \delta\theta)$ by $\delta s = 0$, $\delta\theta_i = s_j$, $\delta\theta_j = -s_i$, and $\delta\theta_r = 0$ for $r \notin \{i,j\}$. Then*

$$D^2\mathscr{L}_C(s,\theta)[z,z] = \frac{s_i s_j (s_i+s_j)}{2\pi}\,\mathcal{F}_C''(\alpha). \tag{41}$$

In particular, if $s_i, s_j > 0$ and $\mathcal{F}_C''(\alpha) < 0$, then z is a direction of strictly negative curvature, so (s,w) is not a local minimum.

Proof ideas of Lemma 21. Since Φ_C is π-periodic we may assume $\theta_i = \theta_j$, and then $\mathcal{F}_C''(\alpha) = \mathcal{F}_C''(\theta_i)$. By Proposition 7 and Remark 15 the angle gradient is $2\pi\,\partial_{\theta_k}\mathscr{L}_C = s_k\mathcal{F}_C'(\theta_k)$, where $\mathcal{F}_C'(\gamma) = \sum_l s_l\Phi_C'(\gamma-\theta_l) - \sum_m t_m\Phi_C'(\gamma-\beta_m)$ depends on θ_l only through its l-th summand. Differentiating once more, for $k \neq l$ the only θ_l-dependent summand is $s_l\Phi_C'(\theta_k-\theta_l)$, so $2\pi\,\partial^2_{\theta_k\theta_l}\mathscr{L}_C = -s_k s_l\,\Phi_C''(\theta_k-\theta_l)$, while for $k=l$ the chain rule gives

$$\begin{aligned}
2\pi\,\partial^2_{\theta_k\theta_k}\mathscr{L}_C &= \frac{d}{d\theta_k}\Big[s_k\,\mathcal{F}_C'(\theta_k)\Big] \\
&= s_k\,\frac{d}{d\theta_k}\Big[\sum_{l\neq k} s_l\,\Phi_C'(\theta_k-\theta_l) + s_k\,\Phi_C'(0) - \sum_m t_m\,\Phi_C'(\theta_k-\beta_m)\Big] \\
&= s_k\Big[\sum_{l\neq k} s_l\,\Phi_C''(\theta_k-\theta_l) \;-\; \sum_m t_m\,\Phi_C''(\theta_k-\beta_m)\Big] \\
&= s_k\big[\mathcal{F}_C''(\theta_k) - s_k\,\Phi_C''(0)\big].
\end{aligned} \tag{42}$$

Using this, let us calculate

$$\begin{aligned}
2\pi\,D^2\mathscr{L}_C(s,\theta)[z,z] &= 2\pi\sum_{k,l}\partial^2_{\theta_k\theta_l}\mathscr{L}_C\;\delta\theta_k\,\delta\theta_l \\
&= 2\pi\Big[\partial^2_{\theta_i\theta_i}\mathscr{L}_C\,(\delta\theta_i)^2 + 2\,\partial^2_{\theta_i\theta_j}\mathscr{L}_C\,\delta\theta_i\,\delta\theta_j + \partial^2_{\theta_j\theta_j}\mathscr{L}_C\,(\delta\theta_j)^2\Big] \\
&= s_i\big[\mathcal{F}_C''(\alpha) - s_i\Phi_C''(0)\big]\,s_j^2 \;+\; 2\big(-s_i s_j\Phi_C''(0)\big)\,s_j(-s_i) \\
&\qquad +\; s_j\big[\mathcal{F}_C''(\alpha) - s_j\Phi_C''(0)\big]\,s_i^2 \\
&= s_i s_j(s_i+s_j)\,\mathcal{F}_C''(\alpha),
\end{aligned} \tag{43}$$

which closes the proof. □

Lemma 22 (At most one positive line implies exact fit, `Lean_19`)**.** *If $(s,w) \in \mathcal{P}^+(d,n)$ with $n \geq m$ is a local minimum of $\mathscr{L}_C$ and $|S_+| \leq 1$, then it is an exact fit.*

Proof of Lemma 22. If $n = 0$, then $m = 0$ by $n \geq m$, and the loss is zero. We may therefore assume $n \geq 1$. If some student mass is zero, Proposition 9 gives an exact fit; hence we may also assume positive student masses.

First suppose that there are no lines of positive net mass $S_+ = \varnothing$, this is $\omega(\alpha) \leq 0$ for every line α. Due to Proposition 7 to show zero loss it suffices to show that the centered angular residual (31) vanishes everywhere. As by its definition $\mathcal{F}_C(x) = \sum_\alpha \omega(\alpha)\Phi_C(x - \alpha)$ and $\Phi_C > 0$ it is easy to see that the residual is negative if $S_- \neq \varnothing$ and otherwise vanishes everywhere. However, as student angles are double zeros by Lemma 16, the former cannot be the case and the residual vanishes everywhere.

Now, suppose there is exactly one line of positive mass $S_+ = \{\alpha_0\}$. Since the teacher masses are positive, and the net mass calculates as aggregated student masses minus aggregated teacher masses, there must be a student direction w_i on the given line $\ell(w_i) = \alpha_0$. By Lemma 16, α_0 is a double zero of $\mathcal{F}_C$. Moreover, $S_- \neq \varnothing$, because otherwise $\mathcal{F}_C(\alpha_0) = \omega(\alpha_0)\Phi_C(0) > 0$, which contradicts criticality. Thus the unique gap of S_+ contains a point of S_-. By Lemma 18(1), $\mathcal{F}_C(\alpha) < 0$ for every line $\alpha \neq \alpha_0$ and since $\mathcal{F}_C$ vanishes at every student line, all student units lie on α_0.

Moreover, Lemma 18(2) gives $\mathcal{F}_C''(\alpha_0) < 0$. If $n \geq 2$, two positive-mass student units therefore lie on α_0, and Lemma 21 supplies a direction of strictly negative curvature, contradicting local minimality.

It remains to consider the case of exactly one student unit (s_1, θ_1). Since $n \geq m$, we have at most one teacher unit $m \leq 1$. If $m = 0$, the target function is the zero function and the loss is a convex function in the single student unit mass s_1 and the only local minimum is where the student unit has mass zero $s_1 = 0$.

The last case is that of one teacher unit (t_1, β_1). On a high level angular criticality forces angle collision or orthogonal angles, but in the latter case masses cannot be optimal. Hence the angles collide and the problem is a convex problem in the masses and any local minimum is an exact fit: The representation of the loss in Proposition 7 evaluates to

$$\mathscr{L}_C(s_1, \theta_1) = \frac{s_1^2 + t_1^2}{8} - \frac{s_1 t_1}{2\pi}\Phi_C(\theta_1 - \beta_1). \tag{44}$$

Now criticality in the angle and $s_1, t_1 > 0$ requires $\Phi_C'(\theta_1 - \beta_1) = 0$. Calculating $\Phi_C'(\eta) = -\sin\eta \arcsin(\cos\eta)$ implies $\sin\eta = 0$ or $\cos\eta = 0$. In the latter case, $\Phi_C''(\eta) = 1$, so $\partial_{\theta_1}^2 \mathscr{L}_C = -s_1 t_1/(2\pi) < 0$, contradicting local minimality. Hence teacher and student unit directions lie on the same line and $\Phi_C(\theta_1 - \beta_1) = \pi/2$. Mass criticality now gives $0 = \partial_{s_1}\mathscr{L}_C = (s_1 - t_1)/4$, so $s_1 = t_1$. We conclude, that the student network is an exact fit. □

Proposition 23 (Positive lines are student lines, `Lean_20`)**.** *If $(s, w) \in \mathcal{P}^{++}(d, n)$ is a critical point of $\mathscr{L}_C$ for the positive teacher network (t, v) and $|S_+| \geq 2$, then $S_+ = \Lambda_w$.*

Proof of Proposition 23. Since the student network is a critical point, it follows by Lemma 16 that every student line is a double zero of $\mathcal{F}_C$. Since the masses of the teacher network are positive we have that $S_+ \subseteq \Lambda_w$. Hence, by Lemma 20, every gap of S_+ contains a point of S_-, and Lemma 18(1) gives $\mathcal{F}_C(x) < 0$ on every gap of S_+. This in turn, again by Lemma 16 implies that every student line must be contained in S_+ and hence $S_+ = \Lambda_w$. □

Lemma 24 (Student line count at centered spurious local minima, `Lean_21`)**.** *If $(s, w) \in \mathcal{P}^{++}(d, n)$ is a spurious local minimum of $\mathscr{L}_C$, then $|\Lambda_w| = n$.*

Proof of Lemma 24. Since (s, w) is a spurious local minimum, Lemma 22 gives $|S_+| \geq 2$. Hence, by Lemma 20, every gap of S_+ contains a point of S_-, so Lemma 18 applies to every gap. Since the masses are positive, by Lemma 18(1), $S_+ = \Lambda_w$, and therefore by Lemma 18(2) $\mathcal{F}_C''(\alpha) < 0$ for every $\alpha \in \Lambda_w$.

Assume for the sake of contradiction that there are w_i, w_j with $\ell(w_j) = \ell(w_i)$. Then, since the masses are positive, Lemma 21 applies. Since $\mathcal{F}_C''(\alpha) < 0$, the direction constructed in Lemma 21 contradicts the positive semidefiniteness of the Hessian at a local minimum. Hence, all student directions lie on different lines, equivalently, $|\Lambda_w| = n$. □

We call the student network and the teacher network *line-separated* if the lines are pairwise distinct, equivalently, $n = |\Lambda_w|, m = |\Lambda_v|$ and $\Lambda_w \cap \Lambda_v = \emptyset$. We say that S_+ and S_- *strictly interlace* if every gap of S_+ contains exactly one point of S_-. When the student network and the teacher network are line-separated and have positive masses, we say that they strictly interlace if $\Lambda_w = S_+$ and $\Lambda_v = S_-$ strictly interlace. Figure 1b shows three teacher and three student lines that strictly interlace.

Proposition 25 (Interlacing and Line-separation at centered spurious local minima, `Lean_22`)**.** *If $(s, w) \in \mathcal{P}^{++}(d, n)$ is a spurious local minimum of $\mathscr{L}_C$, then $n = m$ and the student network and the teacher network are line-separated and strictly interlace.*

Proof of Proposition 25. Since the local minimum is spurious, by Lemma 22, we have that $|S_+| \geq 2$. By Lemma 24, we have that $|S_+| = |\Lambda_w| = n$ and consequently by Lemma 20, $n = |S_+| \leq |S_-| \leq |\Lambda_v| \leq m$. Since also $n \geq m$, it follows that $n = m$ and that equality holds everywhere in the preceding chain of inequalities. In particular, the student network and the teacher network are line-separated. Equality in Lemma 20 implies that every gap of S_+ contains exactly one point of S_-, so they strictly interlace. □

Proposition 26 (Descent direction in the matched width case, `Lean_23`)**.** *Let the student network and the teacher network be line-separated and strictly interlacing, with $n \geq 2$. If $(s, w) \in \mathcal{P}^{++}(d, n)$ is a critical point of $\mathscr{L}_C$, then there exists $z \in \mathbb{R}^n \times \mathbb{R}^n$ such that*

$$D^2\mathscr{L}_C(s, \theta)[z, z] < 0. \tag{45}$$

Proof ideas of Proposition 26. We give explicit rational expressions involving sines and cosines for the construction of the negative-curvature direction, and outline the identities certifying it. In what follows we focus on first giving the precise definition of the descent direction and give the conceptual ideas and steps used to show negative loss curvature along the constructed direction.

Fix teacher parameters (t, β) and student parameters (s, θ), which we assume to constitute a critical point of $\mathscr{L}_C$, with student and teacher width $n = m \geq 2$, positive masses, and pairwise distinct, strictly interlacing student and teacher lines. We may assume that $\theta_1 < \cdots < \theta_n < \theta_{n+1} := \theta_1 + \pi$; let us index the teacher unit in student gap (θ_g, θ_{g+1}), let us call it *gap* g, by $\beta_g = \theta_g + p_g$, and denote the gap length $l_g := \theta_{g+1} - \theta_g \in (0, \pi)$, so $p_g \in (0, l_g)$. Write $z = (\delta s, \delta\theta)$ for a *mass–angle velocity* and $x_g = s_g(\delta\theta)_g$ for its *mass-weighted angular components*. Define

$$\Lambda(l) = l^2 - \sin^2 l, \qquad \psi(l) = \sin l - l\cos l,$$
$$u_g = -\frac{4}{\mathcal{F}_C''(\theta_g)}, \qquad \varrho_g = -\frac{\mathcal{F}_C''(\theta_g)}{s_g}.$$

These divisions are justified. The Green function representation of the centered angular residual in Lemma 18 and our interlacing assumption show $\mathcal{F}_C''(\theta_g) < 0$ at every student unit. In particular, $u_g, \varrho_g > 0$, while $\Lambda(l_g) > 0$ and $s_g > 0$ justify the remaining divisions.

Define coefficients

$$a_g = \frac{2(l_g - \sin l_g \cos l_g)}{\Lambda(l_g)}, \qquad b_g = \frac{2\psi(l_g)}{\Lambda(l_g)},$$
$$d_g = \frac{8\sin^2 l_g}{\Lambda(l_g)}, \qquad e_g = \frac{8 l_g \sin l_g}{\Lambda(l_g)}. \tag{46}$$

Now we are ready to give the formulas for the descent direction $(\delta s, \delta\theta)$:

$$(\delta\theta)_g = \frac{x_g}{s_g}, \text{ where } x_g = (a_g + a_{g-1})u_g + b_g u_{g+1} + b_{g-1}u_{g-1} \tag{47}$$

$$(\delta s)_g = \frac{1}{4}\big[(d_g - d_{g-1})u_g + e_g u_{g+1} - e_{g-1}u_{g-1}\big]. \tag{48}$$

The explicit representation of the centered angular residual in (34) makes these rational expressions in the gap lengths, teacher offsets, masses, and their trigonometric data: only the two neighboring gap geometries and student unit masses enter the descent direction for each student unit. It can be

verified that $a_g, b_g, u_g, s_g > 0$-we remark that indeed all student units rotate in the same direction, $(\delta\theta)_g > 0$.

To understand the construction, consider first an arbitrary velocity $z = (\delta s, \delta\theta)$, and let $\xi(y)$ be the almost-everywhere first variation of twice the student network output for unit direction input at angle y along z. At criticality, the curvature along z is the non-negative feature energy minus a term for each student unit, containing mass-weighted angular components and mass-divided curvature of the residual at the specific student angle:

$$2\pi D^2 \mathscr{L}_C(s,\theta)[z,z] = E - \sum_g \varrho_g x_g^2, \tag{49}$$

$$E = \int_0^\pi \xi(y)^2 \, dy \geq 0. \tag{50}$$

We show that its shifted restriction $S_g(y) := 2\xi(\theta_g + y + \pi/2)$ is a sinusoid, that is, an amplitude-scaled and phase-shifted sine function of unit frequency, and its squared integral gives the gap's contribution to the feature energy:

$$E = \int_0^\pi \xi(y)^2 \, dy = \frac{1}{4} \sum_g \int_0^{l_g} S_g(y)^2 \, dy. \tag{51}$$

We choose z as in (47)–(48) so that these squared-sinusoid integrals admit an explicit evaluation suited to proving negative curvature. The proof progresses by showing that the chosen descent direction z makes S_g equal to an auxiliary sinusoid B_g: their value and slope jumps agree, so their difference extends to one global sinusoid. It is shown that the difference is π-periodic, whereas every unit-frequency sinusoid is π-antiperiodic, and the proof concludes equality (`periodicity argument`). Evaluating the integrals in (51) using the formulation of the auxiliary sinusoid B_g, the calculation arrives at

$$E = 4 \sum_g u_g x_g. \tag{52}$$

Substituting this energy identity into (49), for the specifically constructed descent direction z, the proof arrives at

$$\begin{aligned} 2\pi D^2 \mathscr{L}_C(s,\theta)[z,z] &= 4 \sum_g u_g x_g - \sum_g \varrho_g x_g^2 \\ &= 4 \left[\sum_g \frac{4u_g^2}{\varrho_g} - \sum_g u_g x_g \right] - \sum_g \frac{(\varrho_g x_g - 4u_g)^2}{\varrho_g}. \end{aligned} \tag{53}$$

The rest of the proof certifies negativity of the given formula and since $\varrho_g > 0$, it suffices to prove the `summed inequality`

$$\sum_g \frac{4u_g^2}{\varrho_g} < \sum_g u_g x_g. \tag{54}$$

Lean establishes this by `regrouping into gap contributions` and proving each `gap inequality` through an `exact trigonometric factorization` and monotonicity of ψ/Λ. The `completed-square identity` then gives strictly negative curvature, closing the proof of Proposition 26. □

Remark 27 (Scope and presentation of the matched-width certificate)**.** Proposition 26 is needed only for the matched-width case $n = m$. The overparameterized case $n > m$ is already excluded by Proposition 25: any hypothetical positive-loss local minimum must satisfy

$$n = |\Lambda_w| = |S_+| \leq |S_-| \leq |\Lambda_v| \leq m,$$

which contradicts $n > m$. Thus neither the descent construction of Proposition 26 nor its gap inequality enters the proof in the overparameterized regime.

The remaining matched-width completion uses an exact but technically intricate trigonometric certificate specific to the constructed direction. This certificate arose from AI-assisted formal

proof search rather than from an argument originally designed for concise human exposition. We subsequently worked through the algebra, reconstructed its logical dependencies, and made repeated attempts to simplify and translate it into a conventional proof. Those attempts did not yield an exposition that we considered reasonably readable or independently checkable at conference-paper length.

We therefore separate exposition from complete verification. The paper gives the exact negative-curvature direction, the curvature decomposition, the energy identity, and the precise summed inequality whose strictness implies negative curvature. The accompanying Lean development supplies the full derivation of that inequality, including the auxiliary identities and exact trigonometric calculations omitted from the prose. Every displayed equation and named proof step has a corresponding Lean declaration. Our aim is to make the theorem and the structural ideas behind it accessible to a broad audience without replacing the exposition by a long transcription of a machine-discovered algebraic certificate.

Theorem 28 (Main Result, centered formulation, `Lean_24`)**.** *Let $d \geq 1$, let the teacher network satisfy $(t, v) \in \mathcal{P}^{++}(d, m) \cap \mathcal{P}_{\mathrm{pl}}(d, m)$, and let $n \geq m$. Then every $(s, w) \in \mathcal{P}^{+}(d, n)$ that is a local minimum of $\mathscr{L}_C$ on $\mathbb{R}^n \times (\mathbb{S}^{d-1})^n$ is an exact fit.*

Proof of Theorem 28. For $d = 1$, every unit direction is ± 1, so every centered unit function equals $|x|/2$. Put $D = \sum_i s_i - \sum_k t_k$. The loss is κD^2 with the fixed coefficient $\kappa = \mathbb{E}[|X|^2]/8$; varying one mass by τ changes it to $\kappa(D + \tau)^2$. Stationarity gives $2\kappa D = 0$, hence $\mathscr{L}_C = 0$. This argument needs no mass-sign assumption (`Lean`). If empty networks are allowed, $n = 0$ forces $m = 0$ and the centered loss vanishes identically.

Now let $d \geq 2$ and let $(s, w) \in \mathcal{P}^{+}(d, n)$ be a local minimum. If some student mass is zero, the parameter is an exact fit by Proposition 9, whose planar case applies since the teacher masses are positive. Otherwise all masses are positive, and Proposition 8 places every student direction in the teacher span $\mathcal{T}$, which has dimension at most two. Hence all student and teacher directions lie in a common two-dimensional subspace $P \subseteq \mathbb{R}^d$. Restricting the perturbations to parameters with directions in P preserves local minimality, and since every feature $\varphi_C(\langle w, x\rangle)$ with $w \in P$ depends on x only through its orthogonal projection onto P, which is standard Gaussian in P, the loss of such parameters equals the planar loss of the same teacher network and student network in $P \cong \mathbb{R}^2$. This reduces the claim to $d = 2$. Assume, for the sake of contradiction, that this planar local minimum is spurious. By Proposition 25, the student network and the teacher network are line-separated with $n = m$ and strictly interlace, so Proposition 26 supplies a direction of strictly negative second variation, contradicting local minimality. □

The loss decomposition in Lemma 5 gives the main result for $d \geq 2$; a direct residual-scaling argument covers $d = 1$.

Proof ideas of Theorem 10. For $d \geq 2$, apply Lemma 5 and Theorem 28. For $d = 1$, keep the student directions fixed and choose masses with total $\sum_k t_k$ and their matching linear skip. This is a pointwise exact fit, since all centered unit functions equal $|x|/2$. Interpolating the masses and linear skip linearly from the candidate to this fit multiplies the residual by $1 - \tau$, and the loss by $(1 - \tau)^2$. Stationarity at $\tau = 0$ forces the original loss to vanish (`Lean`). For $n = m = 0$, interpolate just the linear skip to $b^{\mathbf{t}}$. □

The hypothesis $d \geq 3$ of Proposition 9 cannot be dropped: Lemma 29 exhibits spurious local minima with a zero-mass unit in dimensions 1 and 2, the planar one for a teacher network with negative masses. Whether Proposition 9 holds for $d = 2$ and non-negative teacher masses, as Proposition 9 does for the centered loss, is open.

Lemma 29 (Zero-mass spurious minima in $d \leq 2$, `Lean_25`)**.** *There are spurious local minima of $\mathscr{L}_R$ with a positive-mass teacher unit and a zero-mass student unit in $d = 1$, and spurious local minima of $\mathscr{L}_R$ and $\mathscr{L}_C$ for a teacher network with negative-mass units and a student network consisting of a single zero-mass unit in $d = 2$.*

- *(a) In $d = 1$ for the one-unit teacher network $(t, v) = (1, v)$ and one student unit at $w = -v$ with mass s, $\mathscr{L}_R(s, -v) = \frac{1}{4}(s^2 + 1)$, so $(0, -v)$ is a local minimum of $\mathscr{L}_R$ with loss $\frac{1}{4}$, whereas $(1, v)$ has loss 0.*

(b) In $d = 2$ take the teacher network with angles $\beta = (\pi/6, \pi/3, \pi/2, 2\pi/3, 5\pi/6)$ and masses $t = (1, -2\sqrt{3}, 5, -2\sqrt{3}, 1)$, and one student unit of mass 0 at angle τ. For every $|\tau| < \pi/6$ the parameter $(0, e(\tau))$ is a local minimum of $\mathscr{L}_R$ and of $\mathscr{L}_C$ with positive loss.

Proof ideas of Lemma 29. (a) In $d = 1$ let us observe $\langle \pm v, X\rangle = \pm X$, $[X]_+[-X]_+ = 0$ and $\mathbb{E}[X]_+^2 = \mathbb{E}[-X]_+^2 = \frac{1}{2}$. This gives $\mathscr{L}_R(s, -v) = \frac{1}{2}\mathbb{E}[(s[-X]_+ - [X]_+)^2] = \frac{1}{4}(s^2 + 1)$. Since $\mathbb{S}^0 = \{\pm v\}$ is discrete, a neighborhood of $(0, -v)$ only varies s, and $s = 0$ is the minimizer.

(b) For one student unit of mass s at angle θ, Proposition 7 gives the exact quadratic

$$\mathscr{L}_q(s, \theta) - \mathscr{L}_q(0, \theta) = \kappa_q s^2 - \frac{s}{2\pi} P_q(\theta), \qquad P_q(\theta) = \sum_k t_k \Phi_q(\theta - \beta_k) \tag{55}$$

with P_q the teacher potential, and $\kappa_q = \widetilde{\Phi}_q(1)/(4\pi)$ resulting in $\kappa_C = \frac{1}{8}$ and $\kappa_R = \frac{1}{4}$. Since a zero-mass student unit contributes nothing, $\mathscr{L}_q(0, \theta) = \frac{1}{2}\mathbb{E}[f_{q;t,v}(X)^2]$ does not depend on θ. Let us show that for all s and all $|\theta| < \pi/6$ the loss is

$$\mathscr{L}_q(s, \theta) = \mathscr{L}_q(0, \tau) + \kappa_q s^2, \tag{56}$$

so that on the neighborhood $\{|\theta| < \pi/6\}$ of $(0, \tau)$ the loss is at least $\mathscr{L}_q(0, \tau)$, with equality exactly on $s = 0$: $(0, e(\tau))$ is a non-strict local minimum in mass and angle together. By the quadratic, the claim is $P_q \equiv 0$ on $|\theta| < \pi/6$. Let us first show this for $q = R$. There every $\theta - \beta_k$ lies in $(-\pi, 0)$, where $\arcsin(\cos z) = \frac{\pi}{2} + z$ and $\sqrt{1 - \cos^2 z} = -\sin z$ give $\Phi_R(z) = (\pi + z)\cos z - \sin z$. Expanding $\cos(\theta - \beta_k)$ and $\sin(\theta - \beta_k)$ by the addition theorems,

$$P_R(\theta) = (\pi + \theta)\big(A\cos\theta + B\sin\theta\big) - \big(C\cos\theta + D\sin\theta\big) - \big(A\sin\theta - B\cos\theta\big), \tag{57}$$

with $A = \sum_k t_k \cos\beta_k$, $B = \sum_k t_k \sin\beta_k$, $C = \sum_k t_k\beta_k \cos\beta_k$, $D = \sum_k t_k\beta_k \sin\beta_k$. For the displayed teacher network all four vanish for the given teacher angles and masses—the simple calculation uses $\cos\frac{\pi}{6} = \frac{\sqrt{3}}{2}$, $\cos\frac{\pi}{3} = \frac{1}{2}$ and their companions. To show $P_q \equiv 0$ on the arc for $q = C$ observe that $\Phi_C = \Phi_R - \frac{\pi}{2}\cos$ and the argument runs through analogously.

The minimum is not global: as $\mathscr{L}_q(0, \tau) = \frac{1}{2}\mathbb{E}[f_{q;t,v}(X)^2] > 0$, the teacher function is non-zero, and since $P_q(\theta) = 2\pi\mathbb{E}[f_{q;t,v}(X)\varphi_q(\langle e(\theta), X\rangle)]$ by Lemma 6, $P_q \equiv 0$ would give $\mathbb{E}[f_{q;t,v}(X)^2] = \frac{1}{2\pi}\sum_k t_k P_q(\beta_k) = 0$. Hence $P_q(\theta) \neq 0$ for some angle θ, and by (55) a small mass s of the sign of $P_q(\theta)$ at this angle lowers the loss below $\mathscr{L}_q(0, \tau)$. □

A.4 Spurious Local Minima for ReLU

In this subsection we give the proof ideas of Theorem 11 and Proposition 12 of Section 5, preceded by the box certificate and the splitting lemma on which the trap rests.

Spurious local minima of this population loss were first established by Safran & Shamir (2018), for teacher networks with orthonormal directions of width $k \geq 6$ at matched student width, by a computer-assisted proof of the following shape. Gradient descent proposes a candidate w; certificates using variable-precision floating-point arithmetic for a small gradient $\|\nabla F(w)\| \leq \varepsilon$, a Hessian floor $\lambda_{\min}(\nabla^2 F(w)) \geq \lambda_{\min} > 0$, and a uniform bound B on the third-order Taylor remainder are combined to $F(w + \tau u) \geq F(w) + \tau\big(\frac{\lambda_{\min}}{2}\tau - \frac{B}{6}\tau^2 - \varepsilon\big) > F(w)$ for unit u and τ in an interval: a small closed ball around w has values above $F(w)$ and contains an interior local minimum; a Lipschitz bound on F and $F(w) > 0$ show it is not global. We follow the same paradigm with three changes. First, the Hessian is bounded below on a whole box, so no third-derivative remainder enters: strong convexity on the box gives a single and strict interior local minimum. Second, every enclosure is exact rational interval arithmetic checked in Lean, the Hessian floor is certified as a sum-of-squares identity, rather than floating-point arithmetic. Third, positive loss follows from a conceptual rather than numerical argument analyzing the differentiability of the student network and the teacher network on a certain input line.

Lemma 30 (Box certificate for a strict minimum; `Lean_26`)**.** *Let $\ell : \mathbb{R}^D \to \mathbb{R}$ be C^2, let $r, \lambda > 0$ and $g \geq 0$, and put $B = \{z : \|z - z_0\|_\infty \leq r\}$. Suppose*

$$D^2\ell(z)[h, h] \geq \lambda\|h\|_2^2 \quad (z \in B,\ h \in \mathbb{R}^D), \qquad \|\nabla\ell(z_0)\|_2 \leq g, \qquad 2g < \lambda r. \tag{58}$$

Then ℓ has a unique minimizer $z_\star$ on B, this point lies in the interior of B, and it is a strict local minimum.

Proof ideas of Lemma 30. For $z, z+v \in B$, the segment function $\varphi(\tau) = \ell(z+\tau v)$ satisfies $\varphi''(\tau) \geq \lambda\|v\|_2^2$ on $[0,1]$. Integrating twice gives

$$\begin{aligned}\ell(z+v) - \ell(z) - D\ell(z)[v] &= \int_0^1 \big(\varphi'(\tau) - \varphi'(0)\big)\, d\tau \\ &= \int_0^1 \int_0^\tau \varphi''(\sigma)\, d\sigma\, d\tau \\ &\geq \lambda\|v\|_2^2 \int_0^1 \tau\, d\tau \\ &= \frac{\lambda}{2}\|v\|_2^2. \end{aligned} \tag{59}$$

Pulling $D\ell(z)[v]$ to the right and applying Cauchy–Schwarz yields

$$\ell(z_0+v) - \ell(z_0) \geq \|v\|_2 \left(\frac{\lambda}{2}\|v\|_2 - g\right). \tag{60}$$

At a boundary point $z_0 + v$ with $\|v\|_2 \geq r$ this is positive by the assumption $2g < \lambda r$. Compactness therefore gives an interior minimizer $z_\star$, where $D\ell(z_\star) = 0$. Applying (59) there gives strict comparison with every other point of B, proving uniqueness and strict local minimality. □

Lemma 31 (Positive residual curvature permits finite splitting; `Lean_27`)**.** *Fix a planar plain-ReLU teacher network. Let (s,θ) be a local minimum of $\mathscr{L}_R$, and suppose that some student unit i satisfies $s_i > 0$ and $\mathcal{F}_R''(\theta_i) > 0$. For any $k \geq 1$ and any positive partition*

$$s_i = \sum_{j=1}^{k} s'_j, \qquad s'_j > 0, \tag{61}$$

replace that unit by k children with masses s'_j, all at angle θ_i, leaving every other unit unchanged. The resulting parameter is a local minimum with the same represented function and loss. For $k \geq 2$ it is non-strict: transferring mass between two children gives a non-constant constant-loss curve through it.

Proof ideas of Lemma 31. Merging nearby children will give a loss lower bound through Jensen's inequality. For a fixed teacher network, set the normalized kernel loss $\ell_R(s,\theta)$ to be

$$\begin{aligned}\ell_R(s,\theta) &= \frac{1}{2}\sum_{i,j} s_i s_j \Phi_R(\theta_i - \theta_j) - \sum_{i,k} s_i t_k \Phi_R(\theta_i - \beta_k) \\ &= 2\pi\mathscr{L}_R(s,\theta) - \frac{1}{2}\sum_{k,l} t_k t_l \Phi_R(\beta_k - \beta_l). \end{aligned} \tag{62}$$

Write $q_\star = (s,\theta)$ for the given width-n local minimum and $p_\star$ for its split: unit i is replaced by k children of masses $s'_1, \ldots, s'_k$, all at the angle θ_i, and the other units are copied. The children together are the old unit, so $p_\star$ represents the same function and $\mathscr{L}_R(p_\star) = \mathscr{L}_R(q_\star)$. We must show $\mathscr{L}_R(p) \geq \mathscr{L}_R(p_\star)$ for every width-$(n+k-1)$ parameter p near $p_\star$. Near $p_\star$ the child masses $s'_j(p)$ remain positive, and we read the child angles $\vartheta_j(p)$ as real numbers near θ_i. The merge $M(p)$ is the width-n parameter that replaces the children by a single unit of mass $A(p) = \sum_j s'_j(p)$ at the barycentric angle

$$\bar{\vartheta}(p) = \frac{\sum_j s'_j(p)\vartheta_j(p)}{A(p)}, \tag{63}$$

and copies the other units. It is continuous with $M(p_\star) = q_\star$, so local minimality of $q_\star$ gives $\mathscr{L}_R(M(p)) \geq \mathscr{L}_R(q_\star)$ for p near $p_\star$. It remains to compare p with its merge.

Since $\partial_\gamma^2 \mathcal{F}_R(q_\star, \theta_i) > 0$, there are a neighborhood U of $q_\star$ and an open interval $I \ni \theta_i$ with $\partial_\gamma^2 \mathcal{F}_R(q,\gamma) > 0$ for all $q \in U$ and $\gamma \in I$; that is, for each fixed $q \in U$ the function $\gamma \mapsto \mathcal{F}_R(q,\gamma)$ is convex on the same interval I. For p near $p_\star$, continuity of M puts $M(p) \in U$, and the child angles

$\vartheta_j(p)$, hence their weighted mean $\bar{\vartheta}(p)$, lie in I. Thus $F := \mathcal{F}_R(M(p), \cdot)$ is convex on I, an interval containing every child angle of p.

Expanding the loss square, $\ell_R(p) - \ell_R(M(p))$ equals the non-negative term $\pi\,\mathbb{E}[(f_{R;p}(X) - f_{R;M(p)}(X))^2]$ plus

$$\sum_{j=1}^{k} s'_j(p) F(\vartheta_j(p)) - A(p) F(\bar{\vartheta}(p)) \geq 0, \tag{64}$$

by Jensen with weights $s'_j(p)/A(p)$. Consequently,

$$\mathscr{L}_R(p) \geq \mathscr{L}_R(M(p)) \geq \mathscr{L}_R(q_\star) = \mathscr{L}_R(p_\star), \tag{65}$$

where the equality follows because the split children represent the original unit. This proves local minimality and preserves the function and loss for any finite partition. For $k \geq 2$, the masses $s'_1 + \tau, s'_2 - \tau$ remain positive for small $|\tau|$ and define a non-constant constant-loss curve, proving non-strictness. □

Proof ideas of Theorem 11. The box certificate gives a strict seed; positive residual curvature then permits splitting by Lemma 31.

Step 11.1 (Verify the box certificate). Use ℓ_R from the splitting proof for the stated teacher network, in coordinate order $(s_1, s_2, s_3, \theta_1, \theta_2, \theta_3)$. Angles and box radii are measured in radians, not in units of π. The center $x_0 = (\bar{s}, \bar{\theta})$ has exact rational masses and angle ratios:

$$\bar{s} = (1.000437388123,\ 1.390797880326,\ 0.604352649479) \tag{66}$$

$$\frac{\bar{\theta}}{\pi} = (1.777333876184,\ 1.056213642576,\ 0.363169886907). \tag{67}$$

On $B = \{x_0 + v : \|v\|_\infty \leq r\}$ with $r = 1/50000$, we certify curvature, gradient, and positive loss.

(i) *Curvature floor.* Let $H(z) = D^2\ell_R(z)$ and use the upper-triangular rational matrix

$$C = 10^{-6} \begin{pmatrix} 1744016 & 118937 & 353914 & 0 & 536774 & -434587 \\ 0 & 1739956 & 132152 & -387017 & -36692 & 304924 \\ 0 & 0 & 1702608 & 766945 & -755979 & 66668 \\ 0 & 0 & 0 & 1414697 & -907422 & -512806 \\ 0 & 0 & 0 & 0 & 1844853 & -717181 \\ 0 & 0 & 0 & 0 & 0 & 339667 \end{pmatrix}. \tag{68}$$

For this `exact matrix`, the `entrywise enclosure` is

$$\left| H_{ij}(z) - (C^\mathsf{T} C)_{ij} - \frac{1}{10}\delta_{ij} \right| \leq \frac{1}{250} \qquad (z \in B,\ 1 \leq i, j \leq 6), \tag{69}$$

with the certified form identified as the actual Hessian by `the derivative identification`. For $E = H - C^\mathsf{T}C - \frac{1}{10}I$, Cauchy–Schwarz gives $|v^\mathsf{T} E v| \leq \|v\|_1^2/250 \leq 6\|v\|_2^2/250$. Hence $H \succeq Q := C^\mathsf{T}C + \frac{19}{250}I$, and

$$v^\mathsf{T}\left(Q - \frac{1}{16}I\right)v = \sum_{j=1}^{6}(Cv)_j^2 + \frac{27}{2000}\sum_{i=1}^{6} v_i^2 \geq 0 \tag{70}$$

certifies $D^2\ell_R(z)[v, v] \geq \underline{\lambda}\|v\|_2^2$ with $\underline{\lambda} = 1/16$.

(ii) *Gradient bound.* The `center-gradient enclosures` bound the six absolute coordinates by $10^{-14}(14571, 14180, 14985, 8537, 11601, 5258)$; thus $g^2 := \|\nabla\ell_R(x_0)\|_2^2 \leq 873046800 \cdot 10^{-28} < 10^{-19}$. For (i) and (ii), rational center–radius enclosures propagate through sums, products, and absolute values. Trigonometric argument reduction, degree-12 cosine and degree-13 sine `Taylor remainder bounds` on $|x| \leq 1$, and $3.141592653589 < \pi < 3.14159265359$ enclose the inputs. The `enclosure soundness proof` turns these finite rational checks into inequalities for the real derivatives.

(iii) *Positive loss.* For every parameter in B the student angles lie in $(\pi, 2\pi)$, $(\pi, 2\pi)$, and $(0, \pi)$, so $\sin\theta_i \neq 0$ for all i. On the input line $x = (\xi, 1)$ a unit at angle γ contributes $[\sin\gamma + \xi\cos\gamma]_+$. The three student units and the teacher units at $5\pi/6$ and $4\pi/3$ have $\sin\gamma \neq 0$ and are therefore differentiable in ξ at $\xi = 0$, whereas the teacher unit at angle 0 contributes the kink $[\xi]_+$. Hence the two networks differ at some point of this line (`no finite differentiable sum equals a kink`); both are continuous and the Gaussian density is positive, so their squared difference is positive on a non-empty open set of positive measure. Thus

$$\mathscr{L}_R(z) > 0 \qquad (z \in B). \tag{71}$$

Conclusion. The finite kernel expression ℓ_R is C^2 because Φ_R is, and $2g < \underline{\lambda}r$ since $4g^2 < 4 \cdot 10^{-19} < (1/800000)^2 = (\underline{\lambda}r)^2$. Lemma 30 therefore gives a strict local minimum $z_\star$ of ℓ_R in the interior of B. In angle coordinates $\mathscr{L}_R = \ell_R/(2\pi) + \text{const}$, and $\theta \mapsto e(\theta)$ is a local chart of the circle, so $z_\star$ is a strict local minimum of $\mathscr{L}_R$ in mass–direction coordinates. Its masses are positive, since every point of B has masses above 0.6, and it is spurious, since $\mathscr{L}_R(z_\star) > 0$ by (iii) while the teacher network itself has loss zero.

Step 11.2 (Certify positive residual curvature). For splitting the third student unit of $z_\star = (s_\star, \theta_\star)$, we certify $\mathcal{F}_R''(\theta_{\star,3}) > 1/8$, although only positivity is needed. The identity $\Phi_R'' + \Phi_R = 2|\sin\xi|$ gives

$$\mathcal{F}_R''(\gamma) = 2\Big[\sum_i s_i\,|\sin(\gamma - \theta_i)| - \sum_k t_k\,|\sin(\gamma - \beta_k)|\Big] - \mathcal{F}_R(\gamma). \tag{72}$$

At $\gamma = \theta_{\star,3}$ the third student term vanishes, as does $\mathcal{F}_R(\theta_{\star,3})$ by criticality. Throughout B, the angle sectors give $|\sin(\theta_3 - \theta_1)| \geq 24/25$, $|\sin(\theta_3 - \theta_2)| \geq 163/200$, $|\sin\theta_3| \leq 37/40$, and $|\sin(\theta_3 - 4\pi/3)| \leq 1/10$. With $s_1 \geq 999/1000$, $s_2 \geq 1389/1000$, and $|\sin(\theta_3 - 5\pi/6)| \leq 1$, the lower bound is

$$2\left(\frac{999}{1000}\frac{24}{25} + \frac{1389}{1000}\frac{163}{200} - \frac{37}{40} - 1 - \frac{1}{10}\right) = \frac{2643}{20000} > \frac{1}{8}, \tag{73}$$

so

$$\mathcal{F}_R''(\theta_{\star,3}) > \frac{1}{8}. \tag{74}$$

Step 11.3 (Apply finite splitting). For $n > 3$, Lemma 31 splits the third student unit into $n - 2$ children of mass $s_{\star,3}/(n-2)$ at its angle, leaving the other two unchanged. This preserves positive masses and positive loss and supplies a constant-loss mass-transfer curve. The padded teacher copy fits exactly, so the minimum is spurious and non-strict. For $n = 3$, use the strict seed from Step 11.1. □

Proof ideas of Proposition 12. As we work in input dimension $d = 2$, we are working with angles instead of directions. Fix teacher and student weight-angle parameter pairs (t, β) and (s, θ) of widths m and n respectively such that all student masses are positive and assume this constitutes a critical point of the plain loss $\mathscr{L}_R$. First, we aim to show that there are at most $2m$ pairwise distinct student angles.

Let us merge student units that share a direction; this changes neither the represented function nor criticality, so we may assume the n student angles to be pairwise distinct and have to show $n \leq 2m$. For $n = 1$ this is immediate, as the double zero $\mathcal{F}_R(\theta_1) = 0$ forces a teacher unit, so let $n \geq 2$. Unlike in the centered case, plain units with opposite directions compute different functions, so we order the student angles on the full circle $\mathbb{R}/2\pi\mathbb{Z}$, $\theta_1 < \cdots < \theta_n < \theta_{n+1} := \theta_1 + 2\pi$, and call $[\theta_j, \theta_{j+1})$ the *gap* of student unit j; a gap is shorter than 2π but may be longer than π. The *test interval* of student unit j is the closed gap if the gap has length at most π, and otherwise its middle piece $[\theta_{j+1} - \pi, \theta_j + \pi]$, which has length less than π and lies strictly inside the gap.

The proof studies the *sine load* $\mathcal{W} = \frac{1}{2}(\mathcal{F}_R'' + \mathcal{F}_R)$, where $\mathcal{F}_R$ is the plain angular residual (31), which by the kernel identity (30) is simply $\sum_i s_i|\sin(\cdot - \theta_i)| - \sum_k t_k|\sin(\cdot - \beta_k)|$. We call the student and teacher angles together with their antipodes the *atoms*, and those on a teacher line the *teacher atoms*;

there are at most $2m$ teacher atoms. A *sinusoid* is a function of the form $\gamma \mapsto \alpha \cos\gamma + \beta \sin\gamma$; sinusoids form a two-dimensional space, and a sinusoid is determined by its value and its slope at any single point. The sine load is π-periodic and a sinusoid between consecutive atoms, and at an atom τ its slope jumps by

$$\mathcal{W}'(\tau^+) - \mathcal{W}'(\tau^-) = 2\omega(\ell(\tau)), \tag{75}$$

with the net mass ω of (13): upward where the student mass on the line dominates, downward where the teacher mass does.

The goal is the following count that we achieve via a combinatorial argument about *positive humps* of $\mathcal{W}$, the maximal intervals where $\mathcal{W} > 0$. We show that in every student gap $[\theta_i, \theta_{i+1})$ either a positive hump ends, and every such hump contains a teacher atom with $\mathcal{W} > 0$, or the student angle θ_i itself is a teacher atom with $\mathcal{W} = 0$. As distinct humps are disjoint and each hump ends in exactly one gap, this gives

$$\begin{aligned} n &= \#\{\text{student units in whose gap a hump ends}\} + \#\{\text{student units in whose gap no hump ends}\} \\ &\le \#\{\text{teacher atoms with } \mathcal{W} > 0\} + \#\{\text{teacher atoms with } \mathcal{W} = 0\} \le 2m. \end{aligned} \tag{76}$$

It remains to explain the two alternatives.

Criticality in the mass and angle of each student unit makes every student angle a double zero of $\mathcal{F}_R$, as $s_j > 0$ (Remark 15). Since $2\mathcal{W} = \mathcal{F}_R'' + \mathcal{F}_R$, integrating by parts twice over a gap $[\theta_j, \theta_{j+1}]$ between consecutive student angles leaves only boundary terms in $\mathcal{F}_R$ and $\mathcal{F}_R'$, which vanish by our criticality assumption, so $\mathcal{W}$ is orthogonal to every sinusoid on the gap in the following sense

$$\int_{\theta_j}^{\theta_{j+1}} \mathcal{W}(\gamma) \sin(\gamma - a)\, d\gamma = 0 \qquad \text{for every phase } a. \tag{77}$$

For a gap longer than π, its two end pieces $[\theta_j, \theta_{j+1} - \pi]$ and $[\theta_j + \pi, \theta_{j+1}]$ are translates of each other by π, so their contributions cancel by π-periodicity of $\mathcal{W}$. Hence $\mathcal{W}$ is orthogonal to every sinusoid on each test interval. On an interval of length at most π, such a function either vanishes or changes sign at least twice: if it changed sign only once, at ξ_0, then $\mathcal{W}(\gamma)\sin(\gamma - \xi_0)$ would have one sign on the interval and a non-zero integral. Hence, for every student unit, either $\mathcal{W}$ vanishes on its test interval or a positive hump of $\mathcal{W}$ ends inside the test interval, and thus inside the gap.

Both alternatives now come from the way $\mathcal{W}$ bends at atoms: as student masses are positive, only teacher atoms can bend it downward. Positive humps are shorter than π, since a hump of length at least π would, by π-periodicity, make $\mathcal{W}$ positive everywhere up to isolated zeros, while $\mathcal{W}$ vanishes or changes sign on every test interval. A positive hump must bend downward somewhere to return to zero, since a sinusoid that only bends upward stays positive for half a period, longer than any hump; hence every hump contains a teacher atom with $\mathcal{W} > 0$. If instead no hump ends in the gap of student unit j, then $\mathcal{W}$ vanishes on its test interval and cannot bend upward at θ_j, so the teacher mass on this line is at least the student mass there, which is at least $s_j > 0$. Here positivity of s_j is what forces a teacher onto the line, and θ_j is itself a teacher atom, with $\mathcal{W}(\theta_j) = 0$.

The bound is attained for odd $m \ge 3$: teacher units of mass 2 at the vertices of a regular m-gon and student units of mass 1 at these vertices and their antipodes fit exactly, because $2[z]_+ = [z]_+ + [-z]_+ + z$ and the vertices sum to zero. This finishes the proof of Proposition 12. □

A.5 Finite sample consequence

In this subsection we give the proof of Theorem 13 of Section 6 with the definitions, the sampling remark, and the minimizer stability theorem it uses.

In the mismatch coordinate $u = b^{\mathbf{s}} - b_{\mathrm{match}}(s, w)$ of Lemma 5, the identity $[z]_+ = (|z| + z)/2$ and teacher-generated labels write the empirical loss (14) as

$$\widehat{L}_{\mathcal{D}}(b^{\mathbf{s}}, s, w) = \frac{1}{2N} \sum_{j=1}^{N} \big[f_{C;s,w}(x_j) - f_{C;t,v}(x_j) + \langle u, x_j \rangle \big]^2. \tag{78}$$

For $p = (b^{\mathbf{s}}, s, w)$ and $p' = (b^{\mathbf{s}'}, s', w')$, with the corresponding mismatch coordinates u, u', the max-product distance of Section 6 is

$$d_J(p, p') = \max \left\{ \|s - s'\|_\infty,\ \max_{i \in [n]} \|w_i - w_i'\|_2,\ \|u - u'\|_2 \right\}. \tag{79}$$

and the closed ball of an r-stable local minimum is $B_J(p, r) = \{p' \in \mathbb{R}^d \times \mathbb{R}^n \times (\mathbb{S}^{d-1})^n : d_J(p, p') \le r\}$.

Remark 32 (Lean_28)**.** Let $d \ge 2$. For independent, identically distributed inputs from an orthogonally invariant law, i.e. a distribution invariant under orthogonal transformations of $\mathbb{R}^d$, with finite positive second moment, the accuracy condition (15) holds with arbitrarily high probability for all sufficiently large sample sizes (pairwise independence suffices, as the Lean statement records). Indeed, after normalization by its Lipschitz constant, every test function in (15) belongs to the class $\mathcal{G}$ of functions $x \mapsto \|x\|^2 h(x/\|x\|)$ with h ranging over the 1-Lipschitz functions on the sphere bounded by one. This index set is compact in the uniform norm by Arzelà–Ascoli, $h \mapsto \|x\|^2 h(x/\|x\|)$ is continuous for every x, and $\|x\|^2$ is an integrable envelope, so the bracketing numbers of $\mathcal{G}$, the number of pairs $l \le u$ with $\mathbb{E}[u(X) - l(X)] \le \varepsilon$ needed to sandwich every member of $\mathcal{G}$ between some l and u, are finite for every $\varepsilon > 0$, and the Glivenko–Cantelli theorem under bracketing gives almost sure convergence of the sample averages, uniformly over $\mathcal{G}$ (van der Vaart, 1998, Theorem 19.4 and Example 19.8). Since the distribution is orthogonally invariant, the weighted direction $Z/\|Z\|$ under the weight $\|Z\|^2$ is uniform on the sphere, so $\mathbb{E}[\|Z\|^2 h(Z/\|Z\|)] = \mathbb{E}\|Z\|^2 \int h \, d\sigma$, and for the Gaussian G likewise $\mathbb{E}H(G) = d \int h \, d\sigma$ with $h = H|_{\mathbb{S}^{d-1}}$; as also $R_2(\mathcal{D}) \to \mathbb{E}\|Z\|^2$, the uniform convergence yields $\frac{d}{N} \sum_j H(Z_j) - R_2(\mathcal{D})\, \mathbb{E}H(G) \to 0$ uniformly over the normalized tests, which is (15). The accuracy statement needs $d \ge 2$; for every $d \ge 1$ the same event makes every qualifying fixed-radius candidate an exact fit of the sampled data (Lean), the event being the whole sample space when $d = 1$.

By Theorem 10, every population local minimum with non-negative masses is a pointwise exact fit. To transfer this conclusion to accurate datasets, we use the following stability fact.

Theorem 33 (Bastin et al., 2006, Assumption A.4 and Theorem 4.1, Lean_29)**.** *Let ℓ be a continuous function on a metric space, let K be a compact subset, let $r > 0$, and let $\ell_j \to \ell$ uniformly on the union K_r of the closed balls of radius r about the points of K. If each member of a sequence in K minimizes the corresponding ℓ_j on the closed ball of radius r about itself, then every cluster point of the sequence, that is, every limit of a subsequence, is a local minimum of ℓ.*

Proof of Theorem 33. Let p be a cluster point; pass to a subsequence $p_j \to p$ and relabel the sequence l_j accordingly. Since K is closed, $p \in K$. Let y with $d(y, p) < r/2$. For all large j, $d(p_j, p) < r/2$, hence y lies in the closed r-ball about p_j and $\ell_j(p_j) \le \ell_j(y)$. Both p_j and y lie in K_r, so uniform convergence on K_r and the continuity of ℓ give $\ell(p) \le \ell(y)$ in the limit. Thus p minimizes ℓ on the open $r/2$-ball about itself. □

The radius must be fixed before the approximation is refined: the counterexample in Bastin et al. (2006) of minima with shrinking neighborhoods converging to a saddle shows that this is not an artifact. Polak (1993, Theorem 2.3(b)) has the epi-convergence version for uniformly strict local minima, and Mei et al. (2018, Theorems 1 and 2) the smooth version under gradient and Hessian control and a Morse population risk, neither of which we have.

Proof ideas of Theorem 13. For $d = 1$ every unit direction is ± 1, so a student network with the fixed directions, masses of total $\sum_k t_k$, and the matching linear skip fits every input exactly, as in the proof of Theorem 10. Moving masses and mismatch linear skip linearly from the candidate towards this fit multiplies every empirical residual by $1 - \tau$ and stays in $B_J(p, r)$ for small $\tau > 0$; minimality on the ball gives $\widehat{L}_{\mathcal{D}}(p) \le (1 - \tau)^2 \widehat{L}_{\mathcal{D}}(p)$, hence zero loss (Lean). Let now $d \ge 2$. If $R_2(\mathcal{D}) = 0$, every input and teacher-generated label is zero, so the empirical loss is zero at every parameter. We may therefore assume $R_2(\mathcal{D}) > 0$. Multiplication by $d/R_2(\mathcal{D}) > 0$ preserves all ball minima of the actual empirical loss. We compare this rescaled objective with $\mathscr{L}_{\mathrm{skip}}$ in the coordinates (s, w, u) with $u = b^{\mathbf{s}} - b_{\mathrm{match}}(s, w)$: since b_{match} is continuous, this change of coordinates is a homeomorphism with inverse $(s, w, u) \mapsto (u + b_{\mathrm{match}}(s, w), s, w)$, and d_J becomes the maximum of the three distances in s, w and u, so the balls $B_J(p, r)$ are balls of this product metric.

Step 13.1 (parameters of r-stable local minima lie in compact set.). Let $p = (b^{\mathbf{s}}, s, w)$ be r-stable with $s_i \ge 0$, and put $C = \sum_i s_i$, $S = \sum_k |t_k| = \sum_k t_k$, $U = \|u\|$, $A(x) = f_{C;s,w}(x) + \langle u, x \rangle$, and $T(x) = f_{C;t,v}(x)$. Scaling masses and mismatch linear skip together, $(s, u) \mapsto (as, au)$, stays in

$B_J(p, r)$ for a sufficiently close to 1 and changes the residual to $aA - T$. Differentiating this finite quadratic at its minimum $a = 1$ gives

$$\frac{1}{N}\sum_j A(x_j)^2 = \frac{1}{N}\sum_j A(x_j)T(x_j). \tag{80}$$

Substituting this identity into the squared residual gives

$$\widehat{L}_{\mathcal{D}}(p) = \frac{1}{2}\left(\frac{1}{N}\sum_j T(x_j)^2 - \frac{1}{N}\sum_j A(x_j)^2\right). \tag{81}$$

Non-negativity of the loss therefore implies

$$\frac{1}{N}\sum_j A(x_j)^2 \le \frac{1}{N}\sum_j T(x_j)^2 \le \frac{R_2(\mathcal{D})S^2}{4}. \tag{82}$$

If $m = 0$, this already forces zero empirical loss. Otherwise $n \ge m \ge 1$, so the following denominators are positive. For the Gaussian moment $P = \mathbb{E}A(G)^2$, the even centered student network is orthogonal to the odd linear skip, and non-negative masses prevent cancellation between centered units. Hence

$$P \ge \frac{C^2}{4n} + U^2. \tag{83}$$

The function A^2 is continuous and homogeneous of degree two. On the unit ball its Lipschitz constant is at most $2(C/2 + U)^2 \le 4nP$. Applying (15) with this test and $\delta \le 1/(8dn)$ therefore gives

$$P \le \frac{2d}{R_2(\mathcal{D})}\,\frac{1}{N}\sum_j A(x_j)^2 \le \frac{dS^2}{2}. \tag{84}$$

Consequently,

$$C^2 \le 2dnS^2, \qquad U^2 \le \frac{dS^2}{2}. \tag{85}$$

Since $0 \le s_i \le C$ and all directions are unit vectors, these bounds place the minima in one compact parameter set K, independent of the dataset and its radial scale.

Step 13.2 (Uniform comparison on the enlarged set). On the union K_r of closed d_J-balls of radius r about K, all masses, now possibly negative, and mismatch linear skips are uniformly bounded. Thus every residual $F_{b^{\mathbf{s}},s,w} - F_{b^{\mathbf{t}},t,v}$ is bounded by one constant L and is L-Lipschitz on the unit ball. Its squared half-residual is continuous, homogeneous of degree two, and L^2-Lipschitz there. Applying (15) gives, uniformly on K_r,

$$\left|\frac{d}{R_2(\mathcal{D})}\widehat{L}_{\mathcal{D}} - \mathscr{L}_{\mathrm{skip}}\right| \le dL^2\delta. \tag{86}$$

Step 13.3 (A common zero rules out positive loss). If the theorem failed for some fixed r, there would be datasets $\mathcal{D}_j$ of accuracy at most $1/j$ and r-stable minima p_j of positive empirical loss. Their second moments are positive, and, after discarding finitely many indices, the compactness step places p_j in K. Pass to a convergent subsequence. The rescaled objectives $d\widehat{L}_{\mathcal{D}_j}/R_2(\mathcal{D}_j)$ converge uniformly to $\mathscr{L}_{\mathrm{skip}}$ on K_r, so Theorem 33 makes the limit $p = (b^{\mathbf{s}}, s, w)$ a population local minimum with non-negative masses. By Theorem 10, $\mathscr{L}_{\mathrm{skip}}(p) = 0$, and since the residual $F_{b^{\mathbf{s}},s,w} - F_{b^{\mathbf{t}},t,v}$ is continuous and the Gaussian density is positive, it vanishes pointwise; hence $\widehat{L}_{\mathcal{D}_j}(p) = 0$ for every dataset in the sequence. For all sufficiently large j, the limit lies in $B_J(p_j, r)$, and stability gives $0 < \widehat{L}_{\mathcal{D}_j}(p_j) \le \widehat{L}_{\mathcal{D}_j}(p) = 0$, a contradiction. This finishes the proof of Theorem 13. □

B LEAN CERTIFICATES

Every theorem, proposition, lemma, displayed equation, and named proof step of this paper has a counterpart in the Lean 4 development at the root of the accompanying repository, which compiles against Mathlib (The mathlib Community, 2020) without admitted proofs; `verify.py` there checks the five contribution theorems against their recorded statement types, and the build certificate confirms that no paper statement uses an axiom beyond propositional extensionality, choice, and quotient soundness. Throughout the paper, an equation number and a step heading are hyperlinks to the Lean declaration that states, respectively proves, that formula or step; the linked declaration may state a formula in the generality its proof needs, for instance without the unit-norm restriction of the display. The map from paper labels to Lean declarations is shipped as `paper_tools/links.json`; this appendix and every equation link are generated from it by `paper_tools/gen_lean_links.py`. Section B.1 prints the five contribution theorems as Lean source; Section B.2 lists every paper statement with its Lean contract and proof.

B.1 THE FIVE CONTRIBUTIONS IN LEAN

The five theorems of `Overview_theorems.lean` are printed below in the order of Section 1.1, each preceded by the definitions from `Overview_definitions.lean` that its statement introduces. Every proof in that file is a one-line delegation to the statement index `Theorems.lean`; the listings omit these proofs and link to them instead. The code is reproduced verbatim, with Lean's Unicode notation typeset as the editor shows it: `⟪w i, x⟫_ℝ` is the inner product $\langle w_i, x\rangle$, `‖x‖` the Euclidean norm, `Fin n → ℝ` a real vector indexed by $\{0, \dots, n-1\}$, and `∫ x : Vec d, f x` the Lebesgue integral over $\mathbb{R}^d$.

B.1.1 THE MODEL

Both features, the Gaussian input law, and the two population losses (5) are fixed once; every contribution is stated in these terms. Masses are signed real numbers: nonnegativity is an assumption at a candidate, never a constraint on its neighborhood. The losses are literal Bochner integrals, whose value is 0 for a non-integrable integrand, so a reader should check that their integrands are integrable: the loss–kernel identity behind Lemma 6 and Proposition 7 is proved from `a bridge` that assumes integrability of every feature pair, discharged by `centered_feature_pair_integrable` and `gaussian_plain_pair_integrable`; the squared centered residual and its product with the linear skip are integrable by `centered_residual_square_integrable` and `linear_centered_residual_integrable`; and the strict inequality $0 < \mathscr{L}_R$ in Theorem 11 is a Lean-checked witness that the losses are not identically zero.

```
/-- The two hidden-unit features of the paper: the centered feature
`C(z) = |z| / 2` and the plain ReLU `R(z) = max 0 z`. -/
inductive Model where
  | centered
  | plainRelu
  deriving DecidableEq

/-- Standard Gaussian density with respect to Euclidean Lebesgue measure. -/
def stdGaussianDensity {d : ℕ} (x : EuclideanSpace ℝ (Fin d)) : ℝ :=
  (sqrt (2 * π))⁻¹ ^ d * exp (-‖x‖ ^ 2 / 2)

/-- Input space `ℝ^d`. -/
abbrev Vec (d : ℕ) := EuclideanSpace ℝ (Fin d)

/-- A width-`n` student: its masses `s` and its directions `w`. -/
abbrev Parameters (d n : ℕ) := (Fin n → ℝ) × (Fin n → Vec d)

/-- Centered population loss: the full formula in one place. -/
def CenteredLoss {d n m : ℕ} (s : Fin n → ℝ) (w : Fin n → Vec d)
    (t : Fin m → ℝ) (v : Fin m → Vec d) : ℝ :=
  (1 / 2 : ℝ) * ∫ x : Vec d,
    ((∑ i, s i * (|⟪w i, x⟫_ℝ| / 2)) -
      ∑ k, t k * (|⟪v k, x⟫_ℝ| / 2)) ^ 2 * stdGaussianDensity x

/-- Plain-ReLU population loss: the full formula in one place. -/
def PlainLoss {d n m : ℕ} (s : Fin n → ℝ) (w : Fin n → Vec d)
```

```
    (t : Fin m → ℝ) (v : Fin m → Vec d) : ℝ :=
  (1 / 2 : ℝ) * ∫ x : Vec d,
    ((∑ i, s i * max 0 ⟪w i, x⟫_ℝ) −
      ∑ k, t k * max 0 ⟪v k, x⟫_ℝ) ^ 2 * stdGaussianDensity x
```

B.1.2 Contribution 1: Benign loss landscape with linear skips

Theorem 10 is the population statement of the paper's title: with a learned linear skip, a non-negative local minimum of the skip loss against a positive coplanar teacher network has zero loss whenever the student network is at least as wide. The skip loss is the only new definition.

```
/-- Population loss of a plain-ReLU student with a learned linear skip
`bStudent` against a plain-ReLU teacher with linear skip `bTeacher`. -/
def SkipLoss {d n m : ℕ} (bStudent : Vec d) (s : Fin n → ℝ) (w : Fin n → Vec d)
    (bTeacher : Vec d) (t : Fin m → ℝ) (v : Fin m → Vec d) : ℝ :=
  (1 / 2 : ℝ) * ∫ x : Vec d,
    (⟪bStudent, x⟫_ℝ + (∑ i, s i * max 0 ⟪w i, x⟫_ℝ) −
      (⟪bTeacher, x⟫_ℝ + ∑ k, t k * max 0 ⟪v k, x⟫_ℝ)) ^ 2 *
        stdGaussianDensity x

/-- 1. Linear skip: a positive coplanar teacher has zero loss at every
nonnegative local minimum once the student is at least as wide (`thm:headline`). -/
theorem Main {d n m : ℕ} (hd : 1 ≤ d) (hwidth : m ≤ n)
    {bStudent bTeacher : Vec d} {s : Fin n → ℝ} {w : Fin n → Vec d}
    {t : Fin m → ℝ} {v : Fin m → Vec d}
    (hplane : FiniteDimensional.finrank ℝ (Submodule.span ℝ (Set.range v)) ≤ 2)
    (hv : ∀ k, ‖v k‖ = 1) (hw : ∀ i, ‖w i‖ = 1)
    (hs : ∀ i, 0 ≤ s i) (ht : ∀ k, 0 < t k)
    (hmin : IsLocalMinOn
      (fun p : Vec d × Parameters d n => SkipLoss p.1 p.2.1 p.2.2 bTeacher t v)
      {p | ∀ i, ‖p.2.2 i‖ = 1} (bStudent, (s, w))) :
    SkipLoss bStudent s w bTeacher t v = 0
```

The proof of `Main` applies `thm_headline` (proof: Geometry.lean:5218).

B.1.3 Contribution 2: Spurious minima at arbitrary overparameterization

Theorem 11 needs the fixed planar trap teacher network, the planar direction $e(\theta)$, and strict local minimality on the unit-direction parameter space; its three clauses are exactly the theorem's three conclusions.

```
/-- The unit direction `e(θ) = (cos θ, sin θ)` of the plane. -/
def Angle (θ : ℝ) : Vec 2 := ![cos θ, sin θ]

/-- The fixed planar teacher of the trap: three units at angles
`0`, `5π/6`, and `4π/3`, ... -/
def TrapTeacherAngle : Fin 3 → ℝ := ![0 * π, (5 / 6 : ℝ) * π, (4 / 3 : ℝ) * π]

/-- ... each of unit mass. -/
def TrapTeacherMass : Fin 3 → ℝ := fun _ => 1

/-- Strict local minimality within a set: all distinct sufficiently nearby
feasible parameters have strictly larger loss. -/
def IsStrictLocalMinOn {X : Type*} [TopologicalSpace X]
    (f : X → ℝ) (s : Set X) (x : X) : Prop :=
  ∀ᶠ y in 𝓝[s] x, y ≠ x → f x < f y

/-- 2. Spurious local minima at arbitrary overparameterization (`thm:plain-trap`).
The fixed teacher has unit masses at angles `0`, `5π/6`, and `4π/3`.

The three clauses are exactly the theorem's conclusions: at every width at
least three there is a positive-mass spurious minimum; at width three one is
strict; at larger widths one is non-strict. The first clause expresses
spuriousness by exhibiting a feasible point of strictly smaller loss.
The strict and non-strict witnesses have positive loss; copying the teacher
and adding zero-mass units gives a zero-loss competitor at these widths. -/
theorem PlainTrap :
    (∀ {n : ℕ} (_ : 3 ≤ n),
      ∃ s : Fin n → ℝ, ∃ w : Fin n → Vec 2,
        (∀ i, ‖w i‖ = 1) ∧ (∀ i, 0 < s i) ∧
```

```
      IsLocalMinOn (fun p : Parameters 2 n =>
        PlainLoss p.1 p.2 TrapTeacherMass (fun k => Angle (TrapTeacherAngle k)))
        {p | ∀ i, ‖p.2 i‖ = 1} (s, w) ∧
      ∃ s₀ : Fin n → ℝ, ∃ w₀ : Fin n → Vec 2,
        (∀ i, ‖w₀ i‖ = 1) ∧
        PlainLoss s₀ w₀ TrapTeacherMass (fun k => Angle (TrapTeacherAngle k)) <
          PlainLoss s w TrapTeacherMass (fun k => Angle (TrapTeacherAngle k))) ∧
  (∃ s : Fin 3 → ℝ, ∃ w : Fin 3 → Vec 2,
      (∀ i, ‖w i‖ = 1) ∧ (∀ i, 0 < s i) ∧
      IsStrictLocalMinOn (fun p : Parameters 2 3 =>
        PlainLoss p.1 p.2 TrapTeacherMass (fun k => Angle (TrapTeacherAngle k)))
        {p | ∀ i, ‖p.2 i‖ = 1} (s, w) ∧
      0 < PlainLoss s w TrapTeacherMass (fun k => Angle (TrapTeacherAngle k))) ∧
  (∀ {n : ℕ} (_ : 3 < n),
    ∃ s : Fin n → ℝ, ∃ w : Fin n → Vec 2,
      (∀ i, ‖w i‖ = 1) ∧ (∀ i, 0 < s i) ∧
      IsLocalMinOn (fun p : Parameters 2 n =>
        PlainLoss p.1 p.2 TrapTeacherMass (fun k => Angle (TrapTeacherAngle k)))
        {p | ∀ i, ‖p.2 i‖ = 1} (s, w) ∧
      0 < PlainLoss s w TrapTeacherMass (fun k => Angle (TrapTeacherAngle k)) ∧
      ¬ IsStrictLocalMinOn (fun p : Parameters 2 n =>
        PlainLoss p.1 p.2 TrapTeacherMass (fun k => Angle (TrapTeacherAngle k)))
        {p | ∀ i, ‖p.2 i‖ = 1} (s, w))
```

The proof of `PlainTrap` applies `thm_plain_trap_spurious` (proof: Plain.lean:10254), `thm_plain_trap_strict` (proof: Plain.lean:10271), `thm_plain_trap_flat` (proof: Plain.lean:10292).

B.1.4 Contribution 3: Learning the teacher subspace

Proposition 8 holds for both features, so the loss is selected by the `Model`. The conclusion is containment of every positive-mass student direction in the teacher span.

```
/-- The population loss of either feature, selected by the `Model`. -/
def Loss {d n m : ℕ} : Model → (Fin n → ℝ) → (Fin n → Vec d) →
    (Fin m → ℝ) → (Fin m → Vec d) → ℝ
  | .centered => CenteredLoss
  | .plainRelu => PlainLoss

/-- 3. Teacher-span confinement for both features
(`prop:teacher-span-confinement`).
Every student direction with positive mass lies in the teacher span.
This is containment; it does not assert that the student spans the whole space. -/
theorem Confinement {d n m : ℕ} (hd : 2 ≤ d) (q : Model)
    {s : Fin n → ℝ} {w : Fin n → Vec d} {t : Fin m → ℝ} {v : Fin m → Vec d}
    (hv : ∀ k, ‖v k‖ = 1) (hw : ∀ i, ‖w i‖ = 1) (hs : ∀ i, 0 ≤ s i)
    (hmin : IsLocalMinOn (fun p : Parameters d n => Loss q p.1 p.2 t v)
      {p | ∀ i, ‖p.2 i‖ = 1} (s, w)) :
    ∀ i, 0 < s i → w i ∈ Submodule.span ℝ (Set.range v)
```

The proof of `Confinement` applies `prop_teacher_span_confinement` (proof: Geometry.lean:4286).

B.1.5 Contribution 4: Effective width

Proposition 12 is stated in the plane with directions `Angle` θ and needs no further definition: a critical point is a zero of the Fréchet derivative in the mass and angle coordinates, and oriented directions are counted with period 2π. The hypothesis is written with Mathlib's `fderiv`, which is 0 at a point of non-differentiability; it is not vacuous here, because in mass–angle coordinates the plain loss is an affine rescaling of the finite kernel loss (`plain_loss_eq_excess_div`), which is globally C^2 (`loss_contDiff_two`); the hypothesis therefore states that the genuine Fréchet derivative vanishes, equivalently `HasFDerivAt ... 0 (s, θ)`.

```
/-- 4. Effective width (`prop:plain-collision-ceiling`).
In dimension two, every positive-mass plain-ReLU critical point has at most
`2m` distinct oriented student directions. The second clause supplies
attainment: for every odd teacher width at least three there is a critical
point with exactly `2m` directions. Opposite directions count separately.
The `fderiv` hypothesis is a genuine derivative condition: see the conventions
```

```
in the module docstring (`Plain.plain_loss_eq_excess_div`,
`Plain.ParameterCalculus.loss_contDiff_two`). -/
theorem EffectiveWidth :
    (∀ {n m : ℕ} {s θ : Fin n → ℝ} {t β : Fin m → ℝ}
        (_ : ∀ i, 0 < s i)
        (_ : fderiv ℝ (fun p : (Fin n → ℝ) × (Fin n → ℝ) =>
          PlainLoss p.1 (fun i => Angle (p.2 i)) t (fun k => Angle (β k))) (s, θ) =
   0),
      (Finset.univ.image (fun i => Angle (θ i))).card ≤ 2 * m) ∧
    (∀ (m : ℕ) (_ : 3 ≤ m) (_ : Odd m),
      ∃ (s θ : Fin (m + m) → ℝ) (t β : Fin m → ℝ),
        (∀ i, 0 < s i) ∧ (∀ k, 0 < t k) ∧
        fderiv ℝ (fun p : (Fin (m + m) → ℝ) × (Fin (m + m) → ℝ) =>
          PlainLoss p.1 (fun i => Angle (p.2 i)) t (fun k => Angle (β k))) (s, θ) =
   0 ∧
        (Finset.univ.image (fun i => Angle (θ i))).card = 2 * m)
```

The proof of `EffectiveWidth` applies `prop_plain_collision_ceiling` (proof: Plain.lean:10531), `prop_plain_collision_ceiling_sharp` (proof: Plain.lean:10602).

B.1.6 Contribution 5: Transfer to empirical loss

Theorem 13 replaces the Gaussian distribution by a finite dataset. The definitions fix the empirical loss, the joint parameter distance through the matching skip, the accuracy condition against the Gaussian functional on degree-two homogeneous tests, and the fixed-radius minimization hypothesis; the theorem then produces one tolerance δ before the dataset, the labels, the skips, and the student network are chosen.

```
/-- The linear skip that cancels the first Gaussian moment of the residual;
`JointDistance` measures the student's skip relative to it. -/
def MatchingSkip {d n m : ℕ} (bTeacher : Vec d) (t : Fin m → ℝ)
    (v : Fin m → Vec d) (s : Fin n → ℝ) (w : Fin n → Vec d) : Vec d :=
  bTeacher + (1 / 2 : ℝ) • ((∑ k, t k • v k) - ∑ i, s i • w i)

/-- The observed dataset's average squared radius. No directional weights occur. -/
def DataSecondMoment {d N : ℕ} (x : Fin N → Vec d) : ℝ :=
  (∑ j, ‖x j‖ ^ 2) / (N : ℝ)

/-- Ordinary unweighted empirical loss for input vectors and observed labels.
The literal formula also covers zero input vectors; results requiring a
nonempty dataset state `0 < N` explicitly. -/
def DataLoss {d n N : ℕ} (x : Fin N → Vec d) (y : Fin N → ℝ)
    (bStudent : Vec d) (s : Fin n → ℝ) (w : Fin n → Vec d) : ℝ :=
  ((1 / 2 : ℝ) * ∑ j,
    (⟪bStudent, x j⟫_ℝ + (∑ i, s i * max 0 ⟪w i, x j⟫_ℝ) - y j) ^ 2) / (N : ℝ)

/-- Maximum metric in masses, directions, and matching-skip mismatch. -/
def JointDistance {d n m : ℕ} (bTeacher : Vec d) (t : Fin m → ℝ) (v : Fin m → Vec
   d)
    (p q : Vec d × Parameters d n) : ℝ :=
  max (max (dist p.2.1 q.2.1) (dist p.2.2 q.2.2))
    (dist (p.1 - MatchingSkip bTeacher t v p.2.1 p.2.2)
      (q.1 - MatchingSkip bTeacher t v q.2.1 q.2.2))

/-- The Gaussian expectation of a test function. -/
def gaussianFunctional (d : ℕ) (f : Vec d → ℝ) : ℝ :=
  ∫ x : Vec d, f x * stdGaussianDensity x

/-- Positive homogeneity of degree two, the scaling law shared by every
squared residual of the networks. -/
def IsPosHomogeneousDegTwo {d : ℕ} (F : Vec d → ℝ) : Prop :=
  ∀ τ : ℝ, 0 ≤ τ → ∀ x : Vec d, F (τ • x) = τ ^ 2 * F x

/-- Accuracy of the actual unweighted input dataset against the Gaussian
functional for continuous degree-two homogeneous tests. The observed second
moment supplies the radial normalization. This does not construct or assume
a spherical quadrature rule. -/
def DataAccuracy {d N : ℕ} (x : Fin N → Vec d) (δ : ℝ) : Prop :=
  ∀ H : Vec d → ℝ, Continuous H → IsPosHomogeneousDegTwo H →
    ∀ L : ℝ,
```

```
      (∀ z z' : Vec d, ‖z‖ ≤ 1 → ‖z'‖ ≤ 1 → |H z − H z'| ≤ L * ‖z − z'‖) →
      |(d : ℝ) * ((∑ j, H (x j)) / (N : ℝ)) −
          DataSecondMoment x * gaussianFunctional d H| ≤
        (d : ℝ) * DataSecondMoment x * L * δ

/-- The observed labels come exactly from the teacher's plain-ReLU network
with its linear skip. Here `x j` is input number `j`, `y j` its label,
`t k` the mass of teacher unit `k`, and `v k` its direction.
There is no observation noise in this hypothesis. -/
def TeacherGeneratedLabels {d m N : ℕ} (x : Fin N → Vec d) (y : Fin N → ℝ)
    (bTeacher : Vec d) (t : Fin m → ℝ) (v : Fin m → Vec d) : Prop :=
  ∀ j, y j = ⟪bTeacher, x j⟫_ℝ + ∑ k, t k * max 0 ⟪v k, x j⟫_ℝ

/-- The candidate minimizes empirical loss over the entire closed ball of
radius `r`, intersected with the unit-direction parameter space.

A parameter is `(linear skip, (masses, directions))`. The distance is
`JointDistance`, defined above: the maximum of the mass distance, direction
distance, and distance between matching-skip mismatches. Thus this is the
paper's joint parameter ball, with its stated choice of coordinates.

`IsMinOn` means the candidate's loss is no larger than the loss at *every*
feasible point in this ball. This is stronger than ordinary local minimality,
which only promises some possibly smaller neighborhood. The ball permits
signed mass perturbations; nonnegative masses are required only at the
candidate in `ZeroLossForFixedRadiusMinima` below. -/
def MinimizesOnJointBall {d n m N : ℕ} (x : Fin N → Vec d) (y : Fin N → ℝ)
    (bTeacher : Vec d) (t : Fin m → ℝ) (v : Fin m → Vec d) (r : ℝ)
    (candidate : Vec d × Parameters d n) : Prop :=
  IsMinOn (fun p : Vec d × Parameters d n => DataLoss x y p.1 p.2.1 p.2.2)
    {p | (∀ i, ‖p.2.2 i‖ = 1) ∧ JointDistance bTeacher t v p candidate ≤ r}
    candidate

/-- The desired deterministic conclusion for one dataset, one teacher's
hidden units, one student width, and one radius.

For every teacher skip and candidate student: if the labels come from that
teacher, the candidate has nonnegative masses and unit directions, and it
minimizes on the prescribed ball, then its empirical loss is zero.
For nonempty data, zero loss means every observed label is fitted exactly. -/
def ZeroLossForFixedRadiusMinima {d m N : ℕ} (n : ℕ)
    (x : Fin N → Vec d) (y : Fin N → ℝ)
    (t : Fin m → ℝ) (v : Fin m → Vec d) (r : ℝ) : Prop :=
  ∀ (bTeacher bStudent : Vec d) (s : Fin n → ℝ) (w : Fin n → Vec d),
    TeacherGeneratedLabels x y bTeacher t v →
    (∀ i, 0 ≤ s i) → (∀ i, ‖w i‖ = 1) →
    MinimizesOnJointBall x y bTeacher t v r (bStudent, (s, w)) →
    DataLoss x y bStudent s w = 0

/-- 5. Transfer to empirical loss (`thm:fixed-radius-bridge`).

Fix the dimensions, teacher hidden units, and a positive radius `r`.
The teacher has positive masses and unit directions spanning at most two
dimensions; the student is at least as wide as the teacher. Every positive
input dimension is covered; in dimension one the conclusion holds for every
dataset, so the tolerance is not needed there.

There is an accuracy tolerance `δ > 0` that works for *every* nonempty dataset
and every candidate student satisfying the displayed conditions. In particular,
`δ` is chosen before the sample size, inputs, labels, skip vectors, and student
parameters. `DataAccuracy` is the explicit Gaussian approximation condition
in `Overview_definitions.lean`, using the observed average squared input radius. -/
theorem EmpiricalTransfer {d n m : ℕ} (hd : 1 ≤ d) (hwidth : m ≤ n)
    {t : Fin m → ℝ} {v : Fin m → Vec d}
    (hplane : FiniteDimensional.finrank ℝ (Submodule.span ℝ (Set.range v)) ≤ 2)
    (hteacherDirections : ∀ k, ‖v k‖ = 1) (hteacherMasses : ∀ k, 0 < t k)
    {r : ℝ} (hr : 0 < r) :
    ∃ δ : ℝ, 0 < δ ∧
      ∀ (N : ℕ) (x : Fin N → Vec d) (y : Fin N → ℝ),
        0 < N → DataAccuracy x δ → ZeroLossForFixedRadiusMinima n x y t v r
```

The proof of `EmpiricalTransfer` applies `thm_fixed_radius_bridge` (proof: Empirical.lean:1599).

B.2 Statements and proofs

The statement column is the paper-facing contract in `Theorems.lean`; the proof column is the chapter declaration it delegates to. Separate rows distinguish clauses of one statement.

Paper statement / clause	Statement in Lean	Proof	Scope
Contribution 1: Theorem 10	Overview_ theorems.lean:57	Section B.1.2	Benign loss landscape with linear skips; exact contract of Theorem 10.
Contribution 2: Theorem 11	Overview_ theorems.lean:78	Section B.1.3	Spurious minima at arbitrary overparameterization; exact contract of Theorem 11.
Contribution 3: Proposition 8	Overview_ theorems.lean:111	Section B.1.4	Learning the teacher subspace; exact contract of Proposition 8.
Contribution 4: Proposition 12	Overview_ theorems.lean:130	Section B.1.5	Effective width; exact contract of Proposition 12.
Contribution 5: Theorem 13	Overview_ theorems.lean:157	Section B.1.6	Transfer to empirical loss; exact contract of Theorem 13.
Definition 1	—	—	Definition components and parameter conventions; no separate theorem binding. See (1).
Definition 2	—	—	Definition components and parameter conventions; no separate theorem binding. See (4), (5).
Definition 3	—	—	Definition components and parameter conventions; no separate theorem binding. See (3).
Lemma 4 q	Theorems.lean:40	Preliminaries.lean: 2703	Exact contract and local proof.
Lemma 4 linear skip	Theorems.lean:59	Preliminaries.lean: 2727	Exact contract and local proof.
Lemma 5 critical	Theorems.lean:116	Skip.lean:604	Exact contract and local proof.
Lemma 5 identities	Theorems.lean:83	Skip.lean:513	Exact contract and local proof.
Lemma 5 local minimum	Theorems.lean:93	Skip.lean:564	Exact contract and local proof.
Lemma 5 strict local minimum	Theorems.lean:104	Skip.lean:583	Exact contract and local proof.
Lemma 6	Theorems.lean:133	Preliminaries.lean: 1649	Exact contract and local proof.
Proposition 7	Theorems.lean:146	Preliminaries.lean: 2297	Energy pairing, actual mass derivative, and the actual spherical directional derivative along a great circle; no replacement by an arbitrary named gradient expression.
Proposition 8	Theorems.lean:189	Geometry.lean:4286	Input dimension at least two; both activation models; non-negative student masses, arbitrary signed teacher masses, unit directions and local minimality. Active student directions lie in the teacher span.
Proposition 9	Theorems.lean:199	Geometry.lean:4774	Exact contract and local proof.
Proposition 9	Theorems.lean:211	Geometry.lean:4815	Exact contract and local proof.

Paper statement / clause	Statement in Lean	Proof	Scope
Lemma 29 line, (a)	Theorems.lean:228	Plain.lean:11111	The opposed zero-mass student unit on the line is a local minimum on the unit-direction parameter space, with the explicit loss $(s^2+1)/4$ of the opposed student unit and loss 0 for the teacher copy.
Lemma 29 plane, (b)	Theorems.lean:228	Plain.lean:10907	For the fixed five-unit signed teacher network and either activation, every zero-mass one-student parameter with angle in $(-\pi/6, \pi/6)$ is a local minimum on the unit-direction space with positive loss and a lower-loss one-student competitor.
Remark 15	Theorems.lean:170	Planar.lean:5025	Exact contract and local proof.
Lemma 16	Theorems.lean:258	Planar.lean:5100	Exact contract and local proof.
Lemma 17	Theorems.lean:268	Beam.lean:169	Exact contract and local proof.
Lemma 18	Theorems.lean:275	Planar.lean:4755	Canonical representatives in [0,pi), translated left cut at zero. The Green representation, empty-gap equivalence, sign and both endpoint curvatures are all included.
Lemma 20	Theorems.lean:299	Planar.lean:4860	Canonical real line representatives, positive/negative net supports, double zeros and cyclic weak interlacing; the support cardinality consequence is included.
Lemma 21	Theorems.lean:314	Planar.lean:5192	Exact contract and local proof.
Lemma 22	Theorems.lean:347	Geometry.lean:5021	Exact contract and local proof.
Proposition 23	Theorems.lean:325	Planar.lean:5213	Exact contract and local proof.
Lemma 24	Theorems.lean:336	Planar.lean:5241	Exact contract and local proof.
Proposition 25	Theorems.lean:358	Planar.lean:5256	Exact contract and local proof.
Proposition 26	Theorems.lean:370	Planar.lean:5575	Actual Gaussian affine second derivative is strictly negative. The fresh proof derives interlacing and constructs its direction from the stated positive equal-width separated critical configuration.
Theorem 28	Theorems.lean:397	Geometry.lean:5198	All $d \geq 1$; the scalar total-mass proof handles $d = 1$, and the planar reduction handles $d \geq 2$. Empty widths are included.
Theorem 10	Theorems.lean:410	Geometry.lean:5218	All $d \geq 1$; a fixed-direction residual-scaling path handles $d = 1$, with the original loss decomposition for $d \geq 2$.
Theorem 11 every width	Theorems.lean:504	Plain.lean:10233	Exact contract and local proof.
Theorem 11 flat above three	Theorems.lean:540	Plain.lean:10292	Exact contract and local proof.
Theorem 11 non-global	Theorems.lean:516	Plain.lean:10254	Explicit feasible same-width competitor with strictly smaller loss, establishing genuine non-global spuriousness.
Theorem 11 strict three	Theorems.lean:530	Plain.lean:10271	Exact contract and local proof.

Paper statement / clause	Statement in Lean	Proof	Scope
Lemma 30	Theorems.lean:428	Plain.lean:582	A C2 objective on arbitrary finite coordinate blocks, with a Euclidean Hessian floor and dual Euclidean center-gradient bound, has a unique strict minimum in the interior of the max-norm box.
Lemma 30 spurious	Theorems.lean:448	Plain.lean:645	Positive affine normalization, positive actual loss on the box and an exact-fit competitor imply strict spuriousness. Positive mass coordinates are an optional separate consequence, not a premise for spuriousness.
Lemma 31	Theorems.lean:465	Plain.lean:9968	Any positive finite partition of the chosen positive mass preserves the represented network, actual loss and local minimality when its residual curvature is positive. Two or more children give an injective constant-loss curve and non-strictness. The local provider indexes k+1 children; the central entry point explicitly adapts to every $k \geq 1$.
Proposition 12 bound	Theorems.lean:557	Plain.lean:10531	Exact contract and local proof.
Proposition 12 sharp	Theorems.lean:566	Plain.lean:10602	Sharpness is witnessed for odd teacher widths $m \geq 3$, as in the printed explicit family; not a claim of attainment for every m.
Remark 32 accuracy	Theorems.lean:613	Empirical.lean:2901	Model-independent DataAccuracy for identically distributed inputs with an orthogonally invariant common distribution and finite positive second moment; one measurable high-probability event works for every sufficiently large sample size and every test.
Theorem 33	Theorems.lean:581	Empirical.lean:351	Exact contract and local proof.
Theorem 13	Theorems.lean:593	Empirical.lean:1599	All $d \geq 1$; dimension one needs no accuracy hypothesis. Ordinary finite DataLoss with teacher-generated labels. The accuracy threshold is uniform over non-empty datasets, teacher linear skips and non-negative fixed-radius minima. Zero-input datasets are included; signed competitors remain in the ball.

C AI USE STATEMENT

Generative AI played a substantial role in the research underlying this paper, extending beyond language editing or transcription into formal syntax. We used language models for literature exploration, conjecture generation, proof and counterexample search, informal mathematical reasoning, numerical experimentation, development of supporting software, auto-formalization, Lean 4 proof development, and proof simplification. We also used AI assistance for manuscript formulations and for locating and interpreting relevant declarations in the Lean development. During July–September 2026, we used Codex with Sol and Claude Code with Opus and Fable; the final re-formalization described below

used Codex with Astra Ultra. Lean 4 (de Moura & Ullrich, 2021) provided the formal proof-checking environment.

The authors selected and refined the research questions, guided the exploration, reviewed the mathematical meaning of proposed statements, and reorganized the resulting arguments. In particular, the proof ideas in the main paper were reconstructed and written largely manually by the authors, with AI assistance for formulations and navigation of the formal development. This does not imply that their underlying ideas were exclusively human-generated: AI was heavily utilized for idea generation and contributed substantially to the discovery and development of the mathematics. The authors take responsibility for the final statements, arguments, citations, code, and accompanying artifacts.

As a rough measure of computational usage, the project consumed approximately 40 full weekly Codex Pro20 usage allowances and 20 full weekly Claude Max20 usage allowances in the time span from the beginning of July to the end of September 2026. We report usage in these terms because complete token counts are not available and, moreover, token counts alone do not capture the heterogeneous computational resources involved in long-horizon agentic workflows.

C.1 Scope of the contribution and reproducibility

This paper contributes mathematical results, not an empirical evaluation of an AI research system. We do not introduce the agent orchestration as a separately evaluated method, nor do we claim fully autonomous discovery, comparative model performance, proof-search success rates, or improvements in research efficiency. The workflow account below documents the provenance of the results, the extent of AI assistance, and the limitations encountered during the project.

C.2 Progression of the research workflow

The workflow evolved through the following phases. The phases describe the progression of the project rather than a single fixed pipeline that would have been planned or applied perfectly in serial.

- **Phase 1: Text/TeX-only exploration.** We developed an agentic proof-search setup with specialized roles for literature exploration, conjecturing, proving, and verification/utility judging, and experimented with multiple language models and orchestration pipelines. Mathematical knowledge was organized as a directed hypergraph: statements formed the vertices, while proofs formed hyperedges connecting their required ingredients to their conclusions. Statements and proofs were stored in individual Markdown files. Python code orchestrated the workflow by invoking OpenCode with different models as subprocesses. Verification remained informal, relying on judge agents to examine arguments step by step or attempt to reproduce results. The physical analogy behind Section 4, that the angular residual is the radial displacement of a thin elastic ring under point forces and hence obeys the classical ring equation (Remark 19), was found in this phase, when an agent searching the literature for the differential equation of the kernel matched it with the elasticity of curved rods. Hallucinations and failures of this verification process motivated the transition to formal proof checking.
- **Phase 2: Auto-formalization and consolidation.** We used language-model-assisted auto-formalization to translate selected exploratory results and supporting mathematics from the literature into a shared Lean 4 library. Compilation was supplemented by semantic review: we examined whether the formal statements expressed the intended mathematical claims, including their definitions, implicit assumptions, and quantifiers. We also checked the existence of objects satisfying relevant assumptions to guard against vacuous formulations. Prompts and reusable agent instructions were revised in response to formalization errors and discrepancies between formal statements and their intended meaning.
- **Phase 3: Agentic proof search with a shared Lean library.** Prompt- and skill-based agents explored focused subproblems, developed candidate proofs, reused verified lemmas across related settings, and were incentivized to combine intermediate results into stronger statements. Human guidance included selecting and decomposing research questions, reviewing proposed claims, and transferring useful ideas between subproblems. Any time the search would stall, the authors would ask the agents to read relevant literature or find new related work, brainstorm ideas and directions, or evaluate ideas given by the authors.

When the proof search was at a branching point, the authors would discuss significance, difficulties and generalizability of proof techniques with the agents in order to decide where to focus next. Numerical experiments helped test conjectures and identify counterexamples, while failed approaches were preserved to inform subsequent searches. Candidate Lean proofs were compiled, and incorporation into the verified library required both a completed formal proof and review of the statement's mathematical meaning. Integration into the Lean library would also include a rebuild of its changed parts. When a major new result would have been integrated into the Lean library, the authors would check the respective Lean proposition and let the agents generalize from unnecessary theorem requirements.

- **Phase 4: Proof simplification and fresh searches.** We conducted additional searches to simplify arguments and reduce the complexity and build times of the Lean development. For selected statements, fresh searches received the statement and necessary definitions either in Lean or in natural language, but not the existing proof. These searches sought alternative or simpler arguments; proposed replacements were subject to the same formal and semantic checks.
- **Phase 5: Attempted proof translation and author-led exposition.** We used Codex and Claude Code to translate the earlier Lean proofs into an appendix of conventional mathematical proofs. Despite substantial prompting effort and repeated instructions and revisions concerning mathematical presentation, the generated drafts remained difficult to read and did not provide satisfactory exposition. We therefore extracted the proof ideas, formulas, and logical dependencies and substantially reorganized the arguments. The authors wrote the resulting proof ideas largely manually, while continuing to use AI for formulations and assistance in navigating and interpreting the Lean theorems. These reconstructed proof ideas, rather than the earlier automatically generated proof appendix, are the exposition presented in the main paper. The assessment of the generated drafts describes our experience in this project; it is not a systematic evaluation of the models' mathematical writing capabilities.
- **Phase 6: Re-formalization of the simplified proof structure.** The author-led reconstruction substantially simplified the organization of the arguments. We then instructed Codex with Astra Ultra to produce Lean proofs following the structure of the revised proof ideas. The resulting development forms the supplementary Lean certificate accompanying the current manuscript.

C.3 Verification and author responsibility

Admission into the certified development required complete Lean proofs without `sorry`/`admit`, additional project-specific axioms, or unchecked computational shortcuts. Numerical certificates used in the proofs were justified by exact Lean-checked arithmetic. Compilation and transitive axiom audits were supplemented by review of the relationship between the formal statements and the mathematics stated in the paper.

For the five contribution theorem entry points, the accompanying `verify.py` script checks the recorded statement types and audits their axiom dependencies against propositional extensionality, choice, and quotient soundness. The coverage and correspondence of the formal development are documented in Section B. The authors remain responsible for the mathematical exposition and for the accuracy of all claims about the formal certificates.

C.4 Released skills and limits of the workflow record

The supplementary material includes the final versions of the agent skills used during the project as Markdown (`.md`) files. These reusable instructions document how agents were directed during the research and make the workflow inspectable.

The skills evolved during the project in response to failed approaches, verification problems, and changing research needs. The released files therefore represent their final versions; they should not be interpreted as instructions that were used unchanged in every earlier run. They are also not a complete chronological record of all prompts, model interactions, intermediate states, or rejected candidates.

We provide these materials to disclose substantive AI involvement and document the procedures used, not as evidence that the workflow is superior to alternatives or sufficient for autonomous discovery. Accordingly, the absence of a benchmark or an ablation study of the agent workflow limits conclusions about that workflow, but is separate from the verification of the mathematical results claimed in this paper. Those results are supported by the mathematical arguments and the accompanying certificates, with the authors responsible for their correctness and correspondence.