Teacher Geometry Shapes Learnability in Teacher-Student Networks

Anonymous authors
Paper under double-blind review

Abstract

Teacher-student systems, in which a teacher neural network generates targets so that a student neural network can learn to implement the same function, are widely used to study learning. However, the structure of the teacher weights is often overlooked by assuming normally-distributed parameters. This hides substantial variations in how learnable different teachers are. We formalize learnability as the success rate of converging to the global minimum and study how it is affected by overparameterization, teacher geometry, initialization distribution, in the context of a bounded data domain. We analytically reduce and systematically characterize the loss landscape to identify two mechanisms through which teacher geometry shapes learnability in small, tractable ReLU systems: (i) In single-neurons systems, the convergence to this minimum is governed by the initial similarity between teacher and student neurons. If the neurons are initially too dissimilar, the student settles into a sub-optimal minimum, which we name out-of-bounds, where the training dynamics push the ReLU activation threshold outside of the bounded data domain. (ii) In two-neuron systems, the alignment between the two teacher neurons determines the existence of suboptimal local minima and shapes their basins of attraction. Motivated by these findings, we empirically study the trainability of larger systems by constructing teacher distributions with different types of alignment between nodes. We discover that increasing alignment between teacher nodes not only reduces learnability but also increases the rate of convergence to out-of-bounds minima. Moreover, convergence rates to these sub-optimal minima can be reduced by increasing the output-layer learning rate, improving learnability. Finally, by training networks on a set of physics regression problems, we confirm that problems that require aligned-nodes solutions tend to have an increased rate of convergence to out-of-bounds minima.

1 Introduction

How well a function can be learned by a neural network, also known as its learnability, is strongly determined by its expressibility [1–3]: how many units, and what architecture, does a neural network need to express it? If a neural network does not have sufficient capacity to represent a function, it will never be able to learn it. Famously, XOR requires two nodes and two layers to be solved successfully [4, 5]. In higher dimensions, parity on multiple bits places requirements on architecture and input data distribution to be learned successfully [6–8]. Functions containing periodic components are hard for neural networks to approximate and require a large number of neurons [9]. However, this is not the complete picture. Even when a function can be efficiently implemented by a neural network – i.e. there exist a finite-size teacher network that can generate the data – gradient descent may still fail to find the solution and converge to suboptimal local minima [10, 11].

We therefore ask: Apart from expressibility, what factors influence the learnability of a function? To make this question precise, we make use of the teacher-student framework in which a student is trained to match the outputs generated by another network [12–14]. This makes a minimal parameterization known by construction. Teachers in many theoretical studies are drawn with normally distributed weights [12, 14–17], or assume simplified structures on

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.

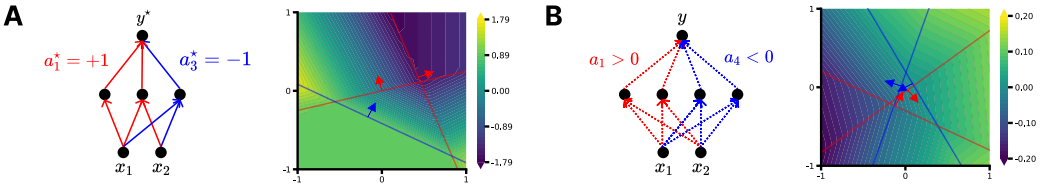

Figure 1: (**A**) A teacher network with input dimension $D = 2$ and $M = 3$ hidden ReLU neurons. The color of the lines indicates whether a node has a positive (*red*) or negative (*blue*) readout weight. The color in the contour plot indicates the network output as a function of the input $\mathbf{x}$; the arrows indicate the non-zero side of the ReLU. (**B**) Same as (A), except showing an overparameterized student network at initialization with $N = 4$ hidden neurons.

weights or data (e.g. diagonal [18], orthogonal [19]); these modeling choices may overlook differences in the difficulty between teachers. For example, teachers of different input and hidden dimensions show a high diversity in learnability, measured by the amount of student overparameterization needed to converge to the global minimum [20]. A complete picture of what factors determine the probability of a student network reaching zero loss is missing.

We investigate these factors through the study of three teacher distributions with highly different success rates, through an analytical reduction to study the loss landscape of small networks and through a number of controlled experiments to identify specific factors of variance. The paper makes the following contributions:

- We introduce and formalize teacher-student learnability as a function of teacher distribution, initialization, data, overparameterization, and optimizer.
- We demonstrate analytically that teacher structure affects the critical points in simple, tractable teacher-student systems. We visualize and characterize the amount of minima in single- and two-neuron teacher-student networks, as well as the importance of starting similarity in single-neuron systems.
- We define teacher distributions that maximize and minimize pairwise differences in orientation and readout signs between neurons, and show that these exhibit large differences in learnability across input dimensions and number of neurons.
- We describe the minima reached by students in the different teacher distributions and discover that more self-similar teacher geometries lead to more local minima with out-of-bounds units (dead or linear units), while more dissimilar teacher geometries lead to more in-bounds local minima (all units operate in the non-linear regime).
- We reveal that increasing the learning rate of the readout weights or decreasing that of the inner biases substantially improves success rates across all distributions.
- We relate these results to real-world functions by training neural networks on a dataset of physics equations. We find that the directed curvature of the target function, which bears resemblance to similar teacher networks, results in both greater directional alignment and a larger number of out-of-bounds units in trained networks.

2 Teacher-student learnability

2.1 Setting

We consider a standard teacher-student framework [14] with networks containing one hidden layer. Let $D \in \mathbb{N}$ denote the input dimensionality, and $X \sim \mathcal{X}$ a set of data points $\mathbf{x}_i \in \mathbb{R}^D$ drawn from the input distribution $\mathcal{X}$. In this setting, a *teacher network* is defined through embedding weights $\mathbf{W}^\star \in \mathbb{R}^{M \times D}$, readout weights $\mathbf{a}^\star \in \mathbb{R}^M$, inner biases $\mathbf{b}^\star \in \mathbb{R}^M$, and an

outer bias $c^\star \in \mathbb{R}$, where M is the number of neurons, and a teacher activation function $\sigma^\star$ (e.g. ReLU, softplus, or tanh; see Figure 1A for an example). The weights of the teacher are drawn according to a distribution $\mathcal{T}$. For each $\mathbf{x}_i \in X$, the teacher then generates output labels

$$y_i^\star = \mathbf{a}^{\star\top}\, \sigma(\mathbf{W}^\star \mathbf{x}_i + \mathbf{b}^\star) + c^\star\,. \tag{1}$$

The teacher-student problem consists of training a *student network* to match the outputs of the teacher. The student is defined by the embedding weights $\mathbf{W} \in \mathbb{R}^{N\times D}$, inner biases $\mathbf{b} \in \mathbb{R}^N$, readout weights $\mathbf{a} \in \mathbb{R}^N$, and outer bias $c \in \mathbb{R}$, where N is the number of neurons in the student. The parameters are drawn from an initial distribution $\mathcal{I}$ and change during learning (Figure 1B). Given a student activation function σ, the output of the student network is $y_i = \mathbf{a}^\top\, \sigma(\mathbf{W}\mathbf{x}_i + \mathbf{b}) + c\,.$

The overparameterization ratio is $\rho = N/M$. The parameters of the student neural network are optimized on a large dataset of K input-output pairs generated by the teacher. Using a given optimizer, the goal is to minimize mean-squared error loss.

2.2 Points of convergence

For a unit with ReLU activation functions, the signed distance from the origin (in the direction of its weight vector $\mathbf{w}$) describes the position of its *kink* $k = -\frac{b}{||w||}$. We describe a neuron as *in-bounds* if datapoints lie on both sides of its kink, described by $\min_i \mathbf{w}\cdot\mathbf{x}_i < k < \max_i \mathbf{w}\cdot\mathbf{x}_i$; if not, the neuron is *out-of-bounds*. This allows us to distinguish the following points to which the system can converge:

- **Global minima,** $\mathcal{L} = 0$: Each teacher node is matched by at least one student node (potentially with symmetries [20, 21]).
- **Out of bounds (OOB) local minima** $\mathcal{L} > 0$: A student fails to match a teacher because either one or more nodes are out of bounds pointing outwards (dead neurons [22, 23]), out of bounds pointing inwards (linear neurons [24, 25]), or both.
- **Interior local minima** $\mathcal{L} > 0$: A student fails even though all nodes are in-bounds, but have converged to a suboptimal local minimum.

As the probability of a network to reach a saddle point exactly is vanishingly small, it will reach a point belonging to one of the three classes above if trained to convergence. Dead relu neurons cannot be rescued during training because they prevent gradient backpropagation. On the contrary, linear units allow gradients to backpropagate, but training dynamics may prevent their kink to re-enter the input data domain. We derive the conditions for linear neurons re-entry in Appendix D.

2.3 Learnability

We define teacher-student learnability for a given teacher network T, overparameterization ratio $\rho = N/M$, and optimizer as the expected conditional probability of reaching the global minimum

$$\mathbb{P}(\text{global minimum}) = \mathbb{E}_{I\sim\mathcal{I}, X\sim\mathcal{X}}\left[\mathbb{I}(\mathcal{L} = 0 \mid I, X, T, \rho, \text{optimizer})\right]\,, \tag{2}$$

for student initializations I drawn from $\mathcal{I}$ and training data X drawn from $\mathcal{X}$, where $\mathbb{I}(\text{true}) = 1$ is the indicator function. We estimate this probability empirically by training multiple student networks (each with a different random initialization seed) on the same teacher, and measuring the fraction that converge to a loss below 10^{-18} (see Figure 7). For each teacher, we fit student networks with independent seeds, yielding a binomial estimate of the learnability for that teacher.

3 Loss landscapes of one- and two-node ReLU systems

3.1 Starting similarity in single-neuron teacher-student system

We can find a closed form of the single-node teacher-student systems for dimensions $D > 1$ under population loss, Gaussian data and setting the linear readout weights to their instanta-

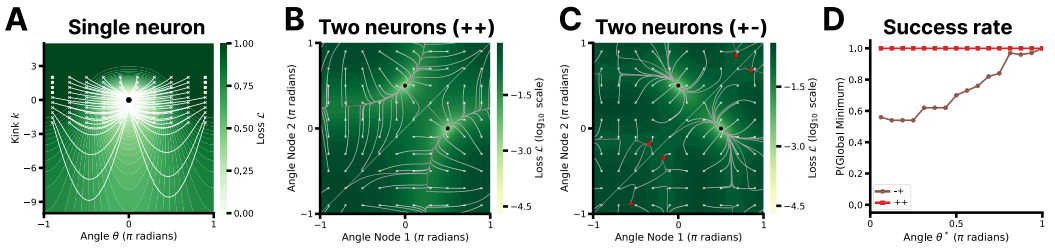

Figure 2: Importance of initial similarity between student and teacher neuron. (**A**) Loss landscape of a single-neuron teacher-student system assuming infinite data, fast readout weights, and $k^\star = 1$. The loss varies from low (*light green*) to high (*dark green*) and has a single global minimum (*black dot*). The landscape is overlaid with with individual trajectories (*white lines*) simulated over two-dimensional inputs, 20'000 Gaussian datapoints, and equal readout weights. (**B**) Same as in (A), but for a two-neuron bias-free teacher-student system with same-sign readouts showing the presence of local (*red dots*) and global minima (*black dots*). (**C**) Same as in (B), but for a system with mixed-sign readouts, additionally showing the presence of spurious local (*red dots*) minima. (**D**) Fraction of the loss landscape in (D) leading to a global minimum evaluated for 100 starting trajectories as a function of the angle $\theta^\star$ between the two teacher neurons. The setting from (D) corresponds to $\theta^\star = \frac{1}{2}\pi$. Remarkably, there are no local minima when both teacher nodes have the same sign.

neous optimal value (see Appendix E.1). Making the readout weights adapt instantaneously corresponds to a change of metric of the gradient flow. While this affects the dynamics, importantly, the critical points of the loss landscape remain the same [26]. With the change of variables $z = \mathbf{w}^\top \mathbf{x}/\|\mathbf{w}\|$ and $z^\star = \mathbf{w}^{\star\top}\mathbf{x}/\|\mathbf{w}^\star\|$, and using the homogeneity of ReLU, the population loss landscape can be rewritten as a function of the angle between student and teacher weights, $\cos\theta = \frac{\mathbf{w}^\top \mathbf{w}^\star}{\|\mathbf{w}\|\,\|\mathbf{w}^\star\|}$, and the kink, $k = -b/\|\mathbf{w}\|$, which represents the signed distance between the origin and the hyperplane of the neuron:

$$\frac{\mathcal{L}(\theta, k)}{\text{const.}} = 1 - \frac{\langle \sigma(z-k)\,\sigma(z^\star - k^\star)\rangle^2_{z,z_\star}}{\langle \sigma(z-k)^2\rangle_z \,\langle \sigma(z^\star - k^\star)^2\rangle_{z^\star}}. \tag{3}$$

The dependency on θ enters through the correlation $\langle z, z^\star\rangle = \cos\theta$. Due to the Gaussian data assumption $z, z^\star$ are bivariate Gaussian with zero mean and unit variance. The Gaussian expectations $\langle \sigma(\cdot)\sigma(\cdot)\rangle$ can be computed analytically [27]. With this, we can prove the following proposition (see Appendix E.1 for the proof):

Proposition 1. *For a single-node teacher student system with ReLU activation function, instantaneous readout weights, and Gaussian inputs* $\mathbf{x} \sim \mathcal{N}(0, I)$*, parameterized as above by an overlap angle* $\cos\theta$*, teacher and student preactivations* $z^\star$ *and* z*, and kinks* $k^\star$ *and* k*, the only critical points are the global minimum at* $\cos\theta = 1, b^\star = b$ *and global maxima at* $\cos\theta = -1$*.*

Remarkably, under these assumptions, the loss landscape has no spurious local minima (see Proposition 1 and Figure 2A). For a finite data domain and equal learning rates, however, around half of the trajectories can fail to converge [22]. We investigate these failure modes in Appendix F.4.1: Success is highly conditioned on starting similarity (Figure 12A), with dissimilar student nodes being rotated away from the teacher (Figure 12B) and driven out of bounds (Figure 12C). If the student begins with the wrong readout sign, none of the nodes converge (Figure 12D).

3.2 Teacher structure in two-neuron teacher-student setting

The two-neuron teacher-student setting has $2D + 5$ parameters, composed of $2D$ input weights, two bias terms, two readout weights, and an outer bias. To again be able to visualize the loss landscape in two dimensions with key assumptions, we introduce the following lemma (see Appendix E.1 for full derivation):

Lemma 1. *Consider a two neuron teacher-student system in $D = 2$ without teacher and student biases. Fix the first teacher neuron at angle 0, let the second teacher neuron have angle $\theta^\star$, and denote the student angles by θ_1 and θ_2. Then, assuming the student readout weights adapt instantaneously, the loss landscape is*

$$\mathcal{L}(\theta_1, \theta_2) = \frac{1}{2}\mathbf{a}^{\star\top}\Big[\mathbf{K}_{TT} - \mathbf{K}_{TS}\,\mathbf{K}_{SS}^{-1}\,\mathbf{K}_{TS}^{\top}\Big]\mathbf{a}^\star, \tag{4}$$

where $\mathbf{K}_{TT} = \langle\sigma(\mathbf{z}^\star)\sigma(\mathbf{z}^\star)^\top\rangle_{\mathbf{z}^\star}$, $\mathbf{K}_{SS} = \langle\sigma(\mathbf{z})\sigma(\mathbf{z})^\top\rangle_{\mathbf{z}}$, $\mathbf{K}_{TS} = \langle\sigma(\mathbf{z}^\star)\sigma(\mathbf{z})^\top\rangle_{\mathbf{z}^\star,\mathbf{z}}$ are the teacher–teacher, student–student, and teacher–student kernel matrices.

In this expression, $\mathbf{z}, \mathbf{z}^\star$ are jointly Gaussian with zero mean, unit variance, and covariances defined by the angles between the respective embedding weights; using the rotational invariance of the setup we set the first teacher as the reference such that all angles can be expressed via θ_1, θ_2, and the angle between the two teacher weights $\theta^\star$. The teacher's readout weights $\mathbf{a}^\star$ and the angle of their weights $\theta^\star$ jointly control the geometry of the landscape.

For teachers with *same-sign readouts* ($\mathbf{a}^\star = (1,1)^\top$) and students with instantaneous readout weights, the loss landscape has no local minima other than the global minimum: all gradient flow trajectories converge regardless of initialization (see Figure 2B). The fact that no spurious minima exist in same-signed two-teacher systems has been shown formally [28, 29]. In contrast, for teachers with *mixed-sign readouts* ($\mathbf{a}^\star = (-1,1)^\top$), the landscape does present local minima (see Figure 2D). At the local minima, one student neurons partially aligns to a teacher neuron, while the other student unit is mis-aligned; likely similar to descriptions of local minima in [30, 31]. The basin of attraction of these local minima decreases with increasing $\theta^\star$ (Figure 2E).

Collectively, this shows that even in simple two-neuron bias-free systems, the structure of the teachers not only determines the existence of interior local minima, but also the sizes of their relative basins of attraction. When we simulate trajectories for systems with bias, finite, bounded data, Glorot-initialized parameters, and equal learning rates for both layers again in the same empirical setting we use for larger networks, overall success rates are smaller, with learnability increasing for both teachers as the teacher angle grows (Figure 13).

4 Comparing teacher distributions

4.1 Defining maximally dissimilar, standard, and minimally dissimilar distributions of teachers

Based on the factors we identified in the previous section, we now define different teacher distributions. Under the standard distribution (common in studies of teacher-student setups [12, 14, 15, 17, 32–34]), weights are drawn independently from a normal distribution and biases are set to zero. To create our distributions, we instead vary similarity between the different nodes along two different axes (see Appendix C.1 for procedure):

- **Directions:** For *similar directions* (sim.), weight vectors $\mathbf{w}_i$ of all teacher nodes are chosen to point in the same "cone" of angle $\pi/8$. This means that all nodes are pairwise highly aligned. For *dissimliar directions* (dissim.), weight vectors $\mathbf{w}_i$ are chosen to have directions as dissimilar from each other as possible.
- **Signs:** For *same-signed readouts* (++), all nodes are given the same readout weight $a_i = 1$. For *mixed-sign readouts* (+-), the readout weights are randomly chosen so that half the nodes have readout weight $+1$ and the other half -1.

These two dimensions independently vary the expected starting similarity (see Propositions 2 and 3). Biases are sampled from uniformly spaced points between $[-0.5, 0.5]$.

4.2 The different distributions lead to markedly different success rates

We vary the width of the teacher network $M \in \{2, 4, 8, 16\}$ and the input dimensionality $D \in \{2, 4, 8, 16\}$, changing one parameter at a time while keeping the other constant at

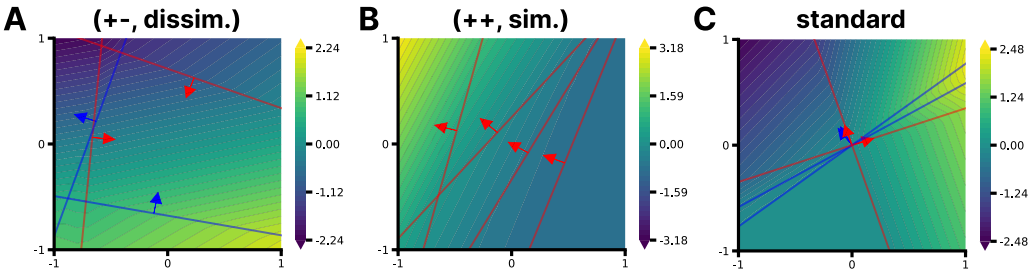

Figure 3: Example teachers with four nodes from the three different distributions. (**A**) A teacher drawn from the distribution with dissimilar directions and mixed-sign readouts. The lines correspond to the individual nodes' hyperplanes, i.e. where $\mathbf{w}^\top \mathbf{x} + b = 0$. The arrow indicates the region where the node has a positive output. The color of the arrow and the hyperplane indicates whether the readout sign a of the node is positive or negative. (**B**) Same as (A), but for teachers from the distribution with similar directions and mixed-sign readouts. (**C**) Same as (B), but for teachers from the standard distribution.

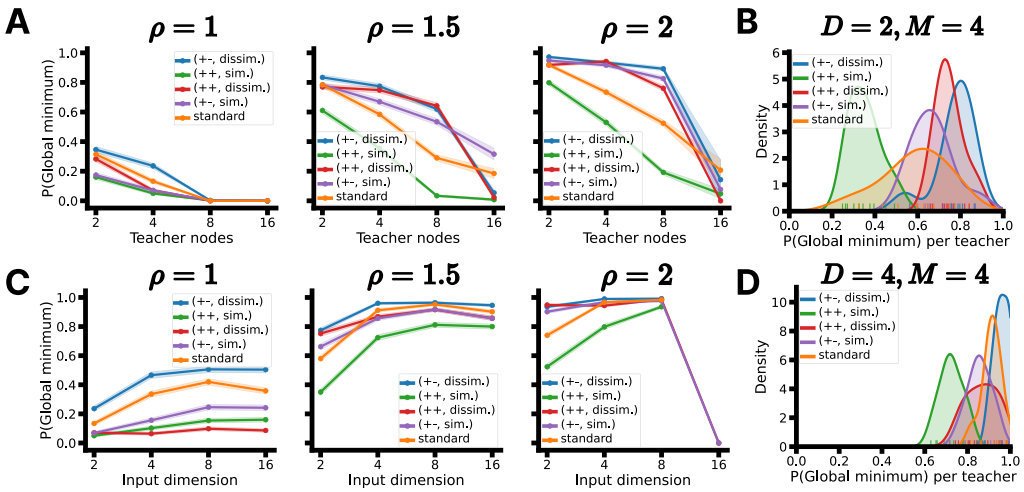

Figure 4: The probability of student networks reaching the global minimum depends on the teacher geometry. (**A**) Fraction of students reaching the global minimum as the number of teacher nodes are increased for $D = 2$. The shaded region indicates the standard error of the binomial proportion. (**B**) Frequency of students with $\rho = 1$ reaching the global minimum for teachers with $D = 2$ and $M = 4$, smoothed using Gaussian kernels. (**C**) Same as (A), but for increasing input dimensionality D and for fixed number of teacher nodes $M = 4$. (**D**) Same as (B), but for teachers with $D = 4$ and $M = 4$.

$M = 4$ or $D = 2$. We fit the students networks using the empirical mean squared error loss using many samples with full-batch gradient flow (see Appendix B.1 for details of the training procedure). Fitting the student networks results in markedly different levels of convergence to the global minimum. We see that, as the number of teacher nodes increases, the proportion of successful fits decreases (Figure 4A); conversely, as the input dimensionality increases, fitting becomes easier (Figure 4C). In almost all simulations across these settings, the teacher distribution with dissimilar directions and mixed signs has the highest success rates, followed by the other distributions roughly in order of difference, with the standard distribution in the middle. Within each distribution, individual teachers span a spectrum of difficulty (see Figure 4B,D).

To ensure that the findings generalize beyond the ReLU activation function, we run the same experiment for both the softplus activation function (Figure 8) and tanh (Figure 9), reaching similar results in both cases.

5 Analysis of the local minima

Global minima, out-of-bounds (OOB) local minima, and in-bounds local minima are quantified from different student initializations (see Figure 5A for examples). Importantly, the types of minima reached varied significantly even between teachers that had similar rates of global minima. Same-sign and similar-node systems tended to have low rates of OOB minima, while higher rates of interior minima compared to dissimilar-nodes with opposite-signs (see Figure 5B for average across all settings and Figure 14 for individual runs).

Because similarity at initialization plays a large role in determining success in single-node systems, we ask how much it can explain for these larger networks. The differences introduced on the readout signs and weight similarity all induce different expectations in initial overlap between teachers and students (see Appendix C.2), given by the quantity

$$S(\mathcal{T},\mathcal{I}) = \mathbb{E}_{T\sim\mathcal{T},I\sim\mathcal{I}}\left[\max_{p\in P_N}\sum_{j=1}^{M} a_{p(j)}a_j^\star \mathbf{w}_{p(j)}^\top \mathbf{w}_j^\star\right] \tag{5}$$

where P_N denotes the symmetric group on $N \geq M$ elements, i.e. the set of all permutations of student nodes. Similar overlap terms play an important role in the learning dynamics of teacher-student settings [12, 17, 35]. To study the effect of initialization on success, we computed the overlap term $\max_{p\in P_N}\sum_{j=1}^{M} a_{p(j)}a_j^\star\, \mathbf{w}_{p(j)}^\top \mathbf{w}_j^\star$ for each individual teacher-student pair, excluding teacher-student systems where $a_{p(j)}a_j^\star < 0$ and $\mathbf{w}_{p(j)}^\top \mathbf{w}_j^\star < 0$ to avoid cases where the system was highly aligned to a symmetry [20, 21], since we focus on cases where $\rho = 1$ (see Appendix B.4). This initial overlap is highest for the maximally dissimilar teacher distribution, and tends to increase with increasing number of teacher nodes (Figure 5C), and roughly trends towards zero for increasing dimensionality (Figure 15). We hypothesized that a higher initial overlap would lead to a higher success rate in finding the global minimum (see Figure 16A for a scatter plot showing the relationship for a single seed). To test this hypothesis, we used the initial overlap as the decision score in an AUROC calculation to see how well we could use it to classify what minimum type that teacher reaches (in a 1-v-all manner). We computed a separate AUROC score for each teacher across the 50 trained students. We averaged the results first within an M, D setting (see Figure 16), and then across settings (see Figure 5D). Consistently across settings, the initial similarity predicts success rate better than chance for global and OOB minima. However, it performs only at around chance level for interior minima, indicating a complicated relationship in this latter type.

6 Differential learning rates

The loss landscape analyses have shown that local minima disappear under fast readout weights; the empirical simulations have shown that neurons go frequently out of bounds for incorrect starting positions. These two observations suggest that simulation performance could potentially improve with differential learning speeds, in which the readout weights are learned more quickly than the embedding weights and the inner bias. We hypothesized that success rates could be increased by applying different learning rates to different parameters, so as to align the readout sign before the neurons go out of bounds. We create versions of the optimization with different learning rates for different parameters. When we increase the learning rate of the readout weights for sample networks, success rates increase. In contrast, success rates are decreased when the learning rate of the inner bias b is increased, as more neurons are driven out of bounds (Figure 6A). These changes are driven by corresponding de- and increases in the number of OOB neurons (Figure 6B). In contrast, changing the optimizer from gradient flow to Adam does not increase the success rates (Figure 6C). This indicates that the momentum effect plays other roles and may help mitigate the impact of other critical points or noise in stochastic gradient descent, but does not prevent neurons from going out of bounds.

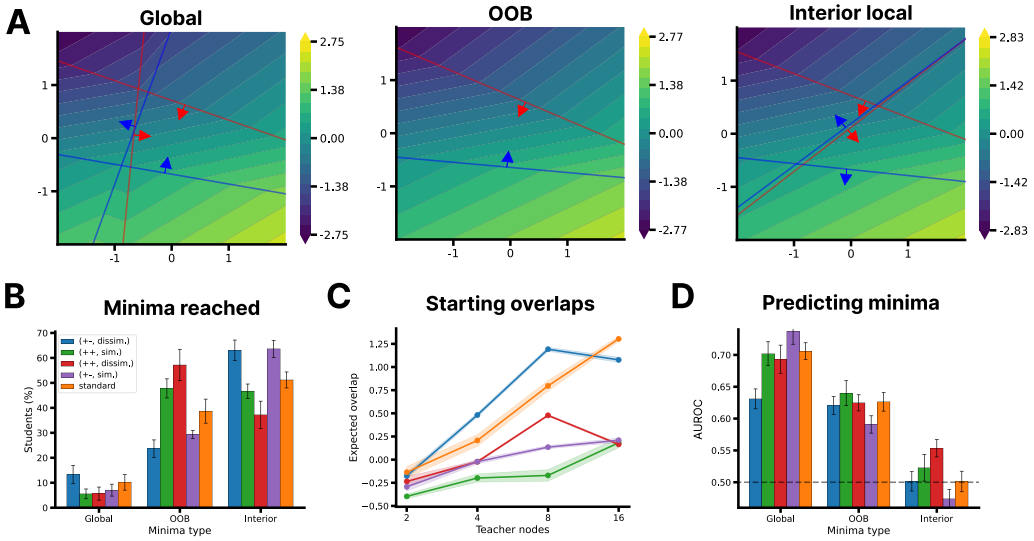

Figure 5: Analyzing the local minima reached by the system for $\rho = 1$. (**A**) Example solutions reached by neurons in (*left*) the global minimum, (*middle*) a local minimum with two out-of-bounds (OOB) neurons pointing inwards (beyond the domain visible in the figure), (*right*) an interior local minimum in which all nodes are in bounds ($\rho = 1, D = 2, M = 4$). (**B**) The frequency of types of minima – global, OOB, and interior – across teacher distributions. The percentage across the three different types of minima are computed within each setting for M and D, and then aggregated across settings. Error bars represent standard error of mean across the eight settings $M \in \{2, 4, 8, 16\}$ with $D = 2$ fixed and $D \in \{2, 4, 8, 16\}$ with $M = 4$ fixed. (**C**) Initial overlap between teacher and student as a function of the teacher distribution and number of teacher nodes. (**D**) AUROC scores for 1-v-all classification of the minima type based on the initial overlap as the decision score for a single teacher, and aggregated across all teachers and settings. Error bars represent standard error of mean across different teachers.

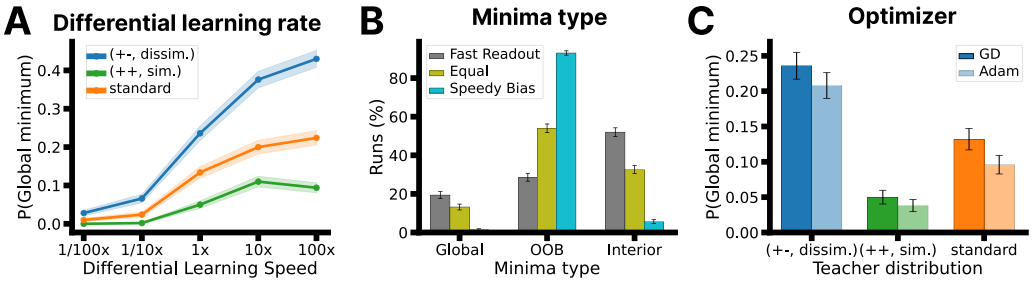

Figure 6: Differential learning rates for $\rho = 1$. (**A**) Percentage of student networks that reach the global minimum as we differentially change, in opposite directions, the speed of the hidden layer bias and of the readout weights and bias. A differential learning speed of 10x corresponds to increasing the speed of the readout weight and bias by a factor of 10, and decreasing the speed of the inner bias by a factor of 10. The shaded region indicates the standard error of the binomial proportion. (**B**) Frequency of the different types of minima – global, out of bounds, and interior – for (*grey*) fast readout weights, (*light green*) equal learning rates, and (*light blue*) fast biases in the $M = 4$, $D = 2$ setting. Error bars represent standard error of mean across the 10 different teachers. (**C**) Fraction of networks finding the global minimum under (*dark blue*) gradient flow and (*light blue*) the Adam optimizer. The error bars indicate the standard error of the binomial proportion.

7 Training on a real-world physics-inspired dataset

To validate our results on real-world functions, we trained neural networks on data generated using AI Feynman, a benchmark of equations drawn from physics problems, including periodic ones [36]. The different ReLU teacher distributions we consider represent different classes of piecewise linear functions. We consider metrics that relate the AI Feynman equations to these distributions. Thus, equations that score highly in the top Hessian share, computed as $\max_i \frac{|\lambda_i|}{\sum_i |\lambda_i|}$ where λ_i are the eigenvalues of the expected Hessian, resemble teachers with similar nodes, as their variance is concentrated along a single axis (see Figure 17). In an analogy to the readout signs, we consider the sign ratio of the top shares computed over Hessian and absolute Hessian, since this indicates how much a function oscillates. Finally, as a baseline, we consider linearity, the R^2 score of the best linear fit. Across these equations, the top Hessian share is the best predictor of directional alignment between nodes ($\rho = 0.55$, $p < 0.001$), the sign ratio the best predictor when slopes are taken into account ($\rho = 0.56$, $p < 0.001$), and both Hessian top share ($\rho = 0.38$, $p < 0.001$) and linearity ($\rho = 0.55$, $p < 0.001$) predict the amount of out-of-bounds units (see Figures 18, 19). Linearity is most determinant of the final loss; however, unlike in the teacher-student systems, this does not relate to network capacity in this setting as directly.

8 Discussion

We studied teacher-student learnability as a function of the teacher geometry. Although one could imagine a consistent signal helping convergence, our results show that randomly initialized networks benefit strongly from having diverse targets, which in expectation gives higher starting similarity (see Appendix C.2). This indicates a complicated interplay of teacher geometry and student initialization. Using the popular notion of lottery tickets in the context of neural networks [37–39], and interpreting tickets as a set of units (instead of a subgraph), one could see the maximally dissimilar teacher distribution together with standard student initialization as a lottery with more winning tickets than the other distributions. However, the expected initial teacher-student overlap does not fully explain the rate of finding the global minimum. Convergence to interior local minima seems impossible to predict based on starting similarity (Figure 5), but also depends on teacher geometry. For both kinds of minima, however, diversity in teacher nodes could have the advantage of decoupling the student neurons' trajectories from each other by pulling them in different directions, leading to a "neural race" in which each student neuron converges to the closest target feature independently [19, 40, 41]. Overall, this study provides a framework for understanding why some functions are easier to learn for neural networks than others, and to being able to train smaller networks by limiting rates of network failure.

AI USE STATEMENT

In this work, we used generative AI tools for formulating mathematical claims and providing ingredients for proving mathematical claims, and assistance in writing proofs, although not for the reduction of the loss landscape. We also used generative AI tools to implement the minima classification presented in Section 5, the recreation of the AI Feynman dataset and training, and for plotting and displaying of results generally, as well as possible minor assistance on other coding tasks, such as from chatbots. We have not used generative AI tools for help developing theoretical models or conceptual frameworks, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, or interpreting results. Assisting with translation and supporting qualitative and thematic data analysis are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

8.1 REPRODUCIBILITY STATEMENT

The code used to produce these results is viewable anonymously during the review period at https://anonymous.4open.science/r/nncriticalpoints-trainability-2A7B/README.md and will be made available deanonymized afterwards.

References

[1] Dmitry Yarotsky. "Error Bounds for Approximations with Deep ReLU Networks". In: *Neural Networks* 94 (Oct. 2017), pp. 103–114. ISSN: 0893-6080. DOI: 10.1016/j.neunet.2017.07.002.

[2] Guido Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. "On the Number of Linear Regions of Deep Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 27. Curran Associates, Inc., 2014.

[3] Raman Arora, Amitabh Basu, Poorya Mianjy, and Anirbit Mukherjee. "Understanding Deep Neural Networks with Rectified Linear Units". In: *International Conference on Learning Representations*. Feb. 2018.

[4] Marvin Minsky and Seymour A. Papert. *Perceptrons: An Introduction to Computational Geometry*. Cambridge, MA, USA: MIT Press, Jan. 1969. ISBN: 978-0-262-63022-1.

[5] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors". In: *Nature* 323 (Oct. 1986). ADS Bibcode: 1986Natur.323..533R, pp. 533–536. ISSN: 0028-0836. DOI: 10.1038/323533a0.

[6] Amit Daniely and Eran Malach. *Learning Parities with Neural Networks*. arXiv:2002.07400 [cs]. July 2020. DOI: 10.48550/arXiv.2002.07400.

[7] Itamar Shoshani and Ohad Shamir. *Hardness of Learning Fixed Parities with Neural Networks*. arXiv:2501.00817 [cs]. Jan. 2025. DOI: 10.48550/arXiv.2501.00817.

[8] Boaz Barak, Benjamin Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. "Hidden Progress in Deep Learning: SGD Learns Parities Near the Computational Limit". In: *Advances in Neural Information Processing Systems*. Vol. 35. Curran Associates, Inc., 2022, pp. 21750–21764. DOI: 10.52202/068431-1581.

[9] Liu Ziyin, Tilman Hartwig, and Masahito Ueda. "Neural Networks Fail to Learn Periodic Functions and How to Fix It". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 1583–1594.

[10] Itay Safran and Ohad Shamir. "Spurious local minima are common in two-layer relu neural networks". In: *International conference on machine learning*. PMLR. 2018, pp. 4433–4441.

[11] Shai Shalev-Shwartz, Ohad Shamir, and Shaked Shammah. "Failures of Gradient-Based Deep Learning". In: *Proceedings of the 34th International Conference on Machine Learning*. PMLR, July 2017, pp. 3067–3075.

[12] David Saad and Sara A. Solla. "Exact Solution for On-Line Learning in Multilayer Neural Networks". In: *Physical Review Letters* 74.21 (May 1995), pp. 4337–4340. DOI: 10.1103/PhysRevLett.74.4337.

[13] David Saad and Sara A. Solla. "On-Line Learning in Soft Committee Machines". In: *Physical Review E* 52.4 (Oct. 1, 1995), pp. 4225–4243. DOI: 10.1103/PhysRevE.52.4225.

[14] H. S. Seung, H. Sompolinsky, and N. Tishby. "Statistical mechanics of learning from examples". In: *Physical Review A* 45.8 (Apr. 1992), pp. 6056–6091. DOI: 10.1103/PhysRevA.45.6056.

[15] Madhu S. Advani, Andrew M. Saxe, and Haim Sompolinsky. "High-Dimensional Dynamics of Generalization Error in Neural Networks". In: *Neural Networks* 132 (Dec. 2020), pp. 428–446. ISSN: 0893-6080. DOI: 10.1016/j.neunet.2020.08.022.

[16] Benjamin Aubin, Antoine Maillard, Florent Krzakala, Nicolas Macris, Lenka Zdeborová, et al. "The committee machine: Computational to statistical gaps in learning a two-layers neural network". In: *Advances in Neural Information Processing Systems* 31 (2018).

[17] Sebastian Goldt, Madhu Advani, Andrew M Saxe, Florent Krzakala, and Lenka Zdeborová. "Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019.

[18] Scott Pesme and Nicolas Flammarion. "Saddle-to-saddle dynamics in diagonal linear networks". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 7475–7505.

[19] Etienne Boursier, Loucas Pillaud-Vivien, and Nicolas Flammarion. "Gradient flow dynamics of shallow relu networks for square loss and orthogonal inputs". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 20105–20118.

[20] Flavio Martinelli, Berfin Simsek, Wulfram Gerstner, and Johanni Brea. "Expand-and-Cluster: Parameter Recovery of Neural Networks". In: *Proceedings of the 41st International Conference on Machine Learning*. PMLR, July 2024, pp. 34895–34919.

[21] Berfin Simsek, François Ged, Arthur Jacot, Francesco Spadaro, Clement 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*. PMLR, July 2021, pp. 9722–9732.

[22] Gal Vardi, Gilad Yehudai, and Ohad Shamir. "Learning a Single Neuron with Bias Using Gradient Descent". In: *Proceedings of the 35th International Conference on Neural Information Processing Systems*. NIPS '21. Red Hook, NY, USA: Curran Associates Inc., Dec. 2021, pp. 28690–28700. ISBN: 978-1-7138-4539-3.

[23] Lu Lu, Yeonjong Shin, Yanhui Su, and George Em Karniadakis. "Dying ReLU and Initialization: Theory and Numerical Examples". In: *Communications in Computational Physics* 28.5 (June 2020), pp. 1671–1706. ISSN: 1815-2406, 1991-7120. DOI: 10.4208/cicp.OA-2020-0165. arXiv: 1903.06733 [stat].

[24] Mark Ainsworth and Yeonjong Shin. *Plateau Phenomenon in Gradient Descent Training of ReLU Networks: Explanation, Quantification and Avoidance*. July 2020. DOI: 10.48550/arXiv.2007.07213. arXiv: 2007.07213 [cs.LG].

[25] Constantin Christof and Julia Kowalczyk. "On the Omnipresence of Spurious Local Minima in Certain Neural Network Training Problems". In: *Constructive Approximation* 60.2 (Oct. 2024), pp. 197–224. ISSN: 1432-0940. DOI: 10.1007/s00365-023-09658-w.

[26] Simone Carlo Surace, Jean-Pascal Pfister, Wulfram Gerstner, and Johanni Brea. "On the Choice of Metric in Gradient-Based Theories of Brain Function". In: *PLOS Computational Biology* 16.4 (Apr. 2020), e1007640. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1007640.

[27] Flavio Martinelli, Alexander van Meegen, Berfin Şimşek, Wulfram Gerstner, and Johanni Brea. "Flat Channels to Infinity in Neural Loss Landscapes". In: *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 2025.

[28] Chenwei Wu, Jiajun Luo, and Jason D. Lee. "No Spurious Local Minima in a Two Hidden Unit ReLU Network". In: (Feb. 2018).

[29] 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*. PMLR, July 2017, pp. 3404–3413.

[30] Yossi Arjevani and Michael Field. "On the principle of least symmetry breaking in shallow relu models". In: *arXiv preprint arXiv:1912.11939* (2019).

[31] Jie Huang, Bruno Loureiro, and Stefano Sarao Mannelli. "Sharp description of local minima in the loss landscape of high-dimensional two-layer ReLU neural networks". In: *arXiv preprint arXiv:2604.09412* (2026).

[32] Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. "Modeling the Influence of Data Structure on Learning in Neural Networks: The Hidden Manifold Model". In: *Physical Review X* 10.4 (Dec. 2020), p. 041044. DOI: 10.1103/PhysRevX.10.041044.

[33] Francesco Camilli, Daria Tieplova, and Jean Barbier. "Fundamental Limits of Overparametrized Shallow Neural Networks for Supervised Learning". In: *Bollettino dell'Unione Matematica Italiana* 19.3 (Sept. 2026), pp. 775–812. ISSN: 2198-2759. DOI: 10.1007/s40574-025-00506-2.

[34] Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. *The Gaussian equivalence of generative models for learning with shallow neural networks*. arXiv:2006.14709 [stat]. May 2021. DOI: 10.48550/arXiv.2006.14709.

[35] Berfin Şimşek, Amire Bendjeddou, Wulfram Gerstner, and Johanni Brea. *Should Under-parameterized Student Networks Copy or Average Teacher Weights?* arXiv:2311.01644 [cs, stat]. Jan. 2024. DOI: 10.48550/arXiv.2311.01644.

[36] Silviu-Marian Udrescu and Max Tegmark. *AI Feynman: A Physics-Inspired Method for Symbolic Regression*. Apr. 2020. DOI: 10.48550/arXiv.1905.11481. arXiv: 1905.11481 [physics.comp-ph].

[37] Jonathan Frankle and Michael Carbin. "The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks". In: *International Conference on Learning Representations*. Sept. 2018.

[38] Stefano Sarao Mannelli, Eric Vanden-Eijnden, and Lenka Zdeborová. "Optimization and Generalization of Shallow Neural Networks with Quadratic Activation Functions". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 13445–13455.

[39] Yuandong Tian, Tina Jiang, Qucheng Gong, and Ari Morcos. *Luck Matters: Understanding Training Dynamics of Deep ReLU Networks*. June 2019. DOI: 10.48550/arXiv.1905.13405. arXiv: 1905.13405.

[40] Etienne Boursier and Nicolas Flammarion. "Simplicity Bias and Optimization Threshold in Two-Layer ReLU Networks". In: *Proceedings of the 42nd International Conference on Machine Learning*. PMLR, Oct. 2025, pp. 5241–5275.

[41] Hannah Pinson. "It's not a Lottery, it's a Race: Understanding How Gradient Descent Adapts the Network's Capacity to the Task". In: *arXiv preprint arXiv:2602.04832* (2026).

[42] K. Fukumizu and S. Amari. "Local minima and plateaus in hierarchical structures of multilayer perceptrons". In: *Neural Networks* 13.3 (Apr. 2000), pp. 317–327. ISSN: 0893-6080. DOI: 10.1016/S0893-6080(00)00009-5.

[43] Zhengqing Wu, Berfin Simsek, and Francois Ged. *Loss Landscape of Shallow ReLU-like Neural Networks: Stationary Points, Saddle Escaping, and Network Embedding*. arXiv:2402.05626 [cs]. June 2024. DOI: 10.48550/arXiv.2402.05626.

[44] Yuchen Zhang, Jason Lee, Martin Wainwright, and Michael I. Jordan. "On the Learnability of Fully-Connected Neural Networks". In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. PMLR, Apr. 2017, pp. 83–91.

[45] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. *Online stochastic gradient descent on non-convex losses from high-dimensional inference*. arXiv:2003.10409 [stat]. May 2021. DOI: 10.48550/arXiv.2003.10409.

[46] Emmanuel Abbe, Enric Boix-Adsera, Matthew S Brennan, Guy Bresler, and Dheeraj Nagaraj. "The staircase property: How hierarchical structure can guide deep learning". In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 26989–27002.

[47] Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. "The merged-staircase property: a necessary and nearly sufficient condition for SGD learning of sparse functions on two-layer neural networks". In: *Proceedings of Thirty Fifth Conference on Learning Theory*. PMLR, June 2022, pp. 4782–4887.

[48] Razvan Pascanu, Guido Montufar, and Yoshua Bengio. *On the Number of Response Regions of Deep Feed Forward Networks with Piece-Wise Linear Activations*. https://arxiv.org/abs/1312.6098v5. Dec. 2013.

[49] Thiago Serra, Christian Tjandraatmadja, and Srikumar Ramalingam. "Bounding and Counting Linear Regions of Deep Neural Networks". In: *Proceedings of the 35th International Conference on Machine Learning*. PMLR, July 2018, pp. 4558–4566.

[50] 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*. PMLR, July 2017, pp. 3404–3413.

[51] Sangmin Lee, Byeongsu Sim, and Jong Chul Ye. "Magnitude and Angle Dynamics in Training Single ReLU Neurons". In: *Neural Networks* 178 (Oct. 2024), p. 106435. ISSN: 0893-6080. DOI: 10.1016/j.neunet.2024.106435.

[52] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. "How Transferable Are Features in Deep Neural Networks?" In: *Advances in Neural Information Processing Systems*. Vol. 27. Curran Associates, Inc., 2014.

[53] Jeremy Howard and Sebastian Ruder. "Universal Language Model Fine-tuning for Text Classification". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Iryna Gurevych and Yusuke Miyao. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 328–339. DOI: 10.18653/v1/P18-1031.

[54] Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weinberger, and Yoav Artzi. "Revisiting Few-sample BERT Fine-tuning". In: *International Conference on Learning Representations*. Oct. 2020.

[55] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. "XLNet: Generalized Autoregressive Pretraining for Language Understanding". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019.

[56] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators". In: *International Conference on Learning Representations*. Sept. 2019.

[57] Pierre Marion and Raphaël Berthier. "Leveraging the Two-Timescale Regime to Demonstrate Convergence of Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 36. Curran Associates, Inc., 2023, pp. 64996–65029. DOI: 10.52202/075280-2836.

[58] Sihan Zeng, Sujay Bhatt, and Sumitra Ganesh. *Rethinking Neural Network Learning Rates: A Stackelberg Perspective*. May 29, 2026. DOI: 10.48550/arXiv.2605.15530. arXiv: 2605.15530 [cs.LG]. URL: http://arxiv.org/abs/2605.15530 (visited on 08/16/2026). Pre-published.

[59] Alexandre Galashov, Nathaël Da Costa, Liyuan Xu, Philipp Hennig, and Arthur Gretton. *Closed-Form Last Layer Optimization*. May 8, 2026. DOI: 10.48550/arXiv.2510.04606. arXiv: 2510.04606 [cs.LG]. URL: http://arxiv.org/abs/2510.04606 (visited on 08/16/2026). Pre-published.

[60] Johanni Brea, Flavio Martinelli, Berfin Şimşek, and Wulfram Gerstner. *MLPGradientFlow: going with the flow of multilayer perceptrons (and finding minima fast and accurately)*. arXiv:2301.10638 [cs]. Jan. 2023. DOI: 10.48550/arXiv.2301.10638.

[61] Youngmin Cho and Lawrence K. Saul. "Kernel Methods for Deep Learning". In: *Advances in Neural Information Processing Systems*. Vol. 22. 2009.

A Related works

A.1 Characterizing fixed points in loss landscapes for teacher-student systems

Fukumizu and Amari described the formation of lines of critical points that emerge from adding neurons to a smaller network at a critical point of its loss [42]. Simsek et al. [21] built on this in the teacher-student setting, showing that permutation symmetries generate a precisely characterised hierarchy of symmetry-induced saddles, so that the global minima manifold grows faster than the saddle manifold as width increases, making overparameterization a systematic mechanism for landscape smoothing. Interestingly, the ratio of all symmetry-induced critical subspaces to global minimum subspaces drops below one when the student network has approximately 1.5 times the number of hidden neurons as the teacher network, independently of the distribution from which the teacher weights are drawn. Wu et al. sharpened this picture for shallow ReLU-like networks, introducing the notion of "escape neurons" that satisfy specific first-order conditions and allow overparameterized networks to escape from saddle points [43]. An important factor influencing convergence rates is the number of permutation-induced symmetries, which are particularly prevalent for ReLU-activation functions [20]. In this paper, we build on these studies to investigate how these factors are affected by differences between different teachers. However, we don't primarily focus on symmetries, considering teacher distributions without parallel hyperplanes and students with small amounts of overparameterization.

A.2 Learnability of functions

Most prior work on learnability focuses on how quickly neural networks can learn to approximate a given function [44]. The computational cost of learning has been shown to depend on the kind of function. The index of the first non-zero Hermite coefficient of the target function, known as the information exponent, governs sample complexity [45]. A class of quickly learnable functions for two-layer networks are those that satisfy the so-called staircase [46] or merged-staircase [47] properties, whose higher-order Fourier functions cannot be built from lower-order ones requiring superpolynomially more updates.

In this paper, however, we are primarily interested in the limit of how well a function will be implemented at convergence, and how many neurons this requires. For this reason, we use a two-step optimization procedure to ensure that the student networks converge to the local minima whose basin of attraction they fall under. The most classic finding of this category is that XOR cannot be implemented by a single perceptron [4], but instead requires a larger network [5]. Its generalization to higher dimensions, the parity function, is still considered the classic example of a difficult function today [6–8], as are periodic functions [9]. We consider a more abstract setting, but periodic functions have some similarity to our minimally dissimilar nodes.

A.3 The ReLU activation function

The ReLU activation function, on which this paper focuses, partitions input space into linear regions [2, 48, 49]. Tian [50] exploits this piecewise linearity to derive a population gradient characterizing the evolution of ReLU networks. In later work, they show that student neurons initialized close to teachers converge faster [39]. We in part build on these results in formalizing the different teacher distributions, but we focus on the implications of this finding for the learnability of different kinds of teachers. Lee et al [51] have analyzed the dynamics of single-neuron ReLU networks in detail. In this paper, we take a similar view, but focus on studying the loss landscape.

A.4 Differential learning rates

Surprisingly, we find that the rate of OOB neurons can be decreased substantially by increasing the speed of the readout weights, and slowing down the speed of the biases. This simple intervention increases success rates substantially. In the literature, differential learn-

ing rates across layers are frequently encountered in the domain of transfer learning, with the intuition that early layers will already have learned useful features during pretraining, so that readout weights are given a higher learning rate [52–56]. However, only few works have considered applying this same intervention already during initial feature learning [57–59]. In the teacher-student setting we consider, we see how allowing the bias term to adapt too quickly in difficult learning settings hinders the formation of good features, in particular when there is an initial mismatch in the readout sign; the solution is slowing down the bias term while increasing the speed of the readout weights. Applying this to real-world domains represents a promising direction for future research.

B Training details

B.1 Training

We train neural networks by solving the gradient flow equation with the MLPGradientFlow package [60]. For each teacher, we fit 50 different student seeds. The input consists of 10,000 points chosen from a standard multivariate Gaussian distribution of dimension D.

Training proceeds in two phases: First, an ODE solver is run until convergence, for up to two hours of CPU time (see Appendix B.5); second, a second-order optimizer optimizes the parameters for an additional two hours.

Some runs, in particular in the larger settings, did not terminate successfully within the timeframe. When no file was generated, the runs were excluded from analysis; Table 1 reports the number of completed runs for each setting. All runs that were successfully saved were included in the analysis without application of an additional convergence criterion.

B.2 Weight normalization procedure

For the maximally and minimally dissimilar distributions, each embedding weight vector $\mathbf{w}_j^\star$ is constructed to lie on the unit hypersphere by design. For the standard distribution, weight vectors are drawn with scale $1/\sqrt{D}$, giving expected unit norm.

In the Python analysis pipeline, teacher networks are additionally normalized for visualization and overlap computation: each embedding weight vector is rescaled to unit norm and multiplied by 3. The readout weights are then iteratively adjusted so that the standard deviation of the teacher output on a 30,000-sample standard Gaussian input equals 1 and the mean equals 0. This normalization is applied only in the analysis scripts and does not affect the training experiments.

B.3 Convergence criteria and success classification

A student run is classified as having reached the global minimum if its final training loss satisfies $\mathcal{L} < 10^{-18}$. Training itself terminates when the loss reaches 10^{-30}, when a patience criterion is met (no improvement over 10^6 iterations), or when the time limit is reached (2 hours for ODE-based training, 2 hours for second-order optimization).

For student runs that do not reach $\mathcal{L} < 10^{-30}$, the minimum eigenvalue of the Hessian is computed at the stopping point. Runs with minimum eigenvalue $> 10^{-5}$ are classified as having converged to a strict local minimum; all others are classified as non-strict critical points or saddle regions. Runs where the optimizer's line search failed are recorded separately.

B.4 Initial overlap computation

The overlap between a student neuron i and a teacher neuron j is computed as

$$s_{ij} = a_i \, a_j^\star \, \left(\mathbf{w}_i^\top \mathbf{w}_j^\star\right), \tag{6}$$

the product of the two readout weights and the inner product of the embedding weight vectors. A higher score indicates that the student neuron has a similar weight direction and the same readout sign as the teacher neuron. Note that a double-negative case (both $a_i \, a_j^\star < 0$ and $\mathbf{w}_i^\top \mathbf{w}_j^\star < 0$) is treated as no overlap for ReLU networks, since the two sign flips do not cancel functionally.

The optimal assignment between student and teacher neurons is found using the Hungarian algorithm, maximizing the total overlap score $\sum_{j=1}^{M} s_{p(j),j}$ over all permutations $p \in P_N$. This is exactly the quantity $\max_{p \in P_N} \sum_{j=1}^{M} s_{p(j),j}$ whose expectation over $\mathcal{T}$ and $\mathcal{I}$ defines $S(\mathcal{T}, \mathcal{I})$ in Equation 5.

B.5 Compute used

The experiments were run on a CPU across 20 parallel cores. The paper requires about 100,000 individual random seeds. Each seed was capped at 4 hours, though most finished much faster, giving an estimated total runtime of about 2000 hours (on AMD EPYC 9454 48-Core Processor CPU) with this setup.

C Teacher distributions

C.1 Teacher generation details

We sample the teachers from the distributions with a fixed random seed (0). We generate 10 teachers per combination of input dimensionality $D \in \{2, 4, 8, 16\}$ and number of teacher nodes $M \in \{2, 4, 8, 16\}$.

Throughout this subsection $j = 1, \ldots, M$ indexes the teacher nodes, so that $\mathbf{w}_j^\star$, $b_j^\star$ and $a_j^\star$ denote the embedding weight vector, the inner bias, and the readout weight of teacher node j, and $c^\star$ denotes the teacher outer bias.

Below, we describe the sampling procedure we used for readout weights, inner weights, and biases, according to the corresponding distribution.

Dissimilar weights and biases The weight vectors $\mathbf{w}_j^\star$ are initialized on a quasi-uniform hyperspiral to be approximately equidistant on the unit hypersphere. They are then refined by 10^6 gradient descent steps with learning rate 0.1 minimizing the sum of squared pairwise cosine similarities, with renormalization to the unit sphere after each step. Gaussian noise with standard deviation 0.1 is added and the vectors are renormalized. Biases are uniformly spaced between $[-0.5, 0.5]$ and assigned randomly to the different nodes.

Similar weights and biases A base direction $\bar{\mathbf{w}}$ is drawn uniformly from the unit hypersphere. The weight vectors $\mathbf{w}_j^\star$ are then drawn uniformly at random and accepted only if the angle θ to $\bar{\mathbf{w}}$ satisfies $\theta < \pi/8$, repeating until M vectors are accepted. The biases $b_j^\star$ are set to M evenly spaced values in $[-0.5, 0.5]$ and assigned to nodes in random order.

Same-signed readout weights All readout weights are $a_j^\star = +1$ and the outer bias is $c^\star = 0$.

Mixed-signed readout weights Half the neurons are randomly assigned a readout weight of $a_j^\star = +1$ and the other half $a_j^\star = -1$.

Standard. The entries of the weight vectors are sampled as $w_{jd}^\star \sim \mathcal{N}(0, 1/D)$ independently for each node j and dimension d, i.e. with standard deviation $1/\sqrt{D}$. All biases are zero, $b_j^\star = 0$. The readout weights are sampled as $a_j^\star \sim \mathcal{N}(0, 1)$ independently; the outer bias is $c^\star = 0$.

C.2 The different teacher distributions maximize and minimize expected initial similarity

In this section, we argue that the expected initial similarity across students for teachers depends on the teacher distribution. We separately consider the effect of aligning teacher nodes (Proposition 2 for the simplified setting with two teacher nodes) and of using mixed- or same-sign teacher nodes (Proposition 3).

This expression is minimized at $\theta = 0$ and maximized at $\theta = \pi$. These two settings correspond to the orientations between nodes of the minimally dissimilar teacher distribution and maximally dissimilar teacher distribution, respectively.

In the settings we consider in the paper, the readout signs also differ between the maximally dissimilar distribution of teachers (where they are evenly split between $\{+1, -1\}$) and the minimally dissimilar distribution (where every node has $a_j^\star = +1$). Randomly-initialized student weights will randomly be positive or negative. By a similar argument, the simi-

larity at initialization will therefore also be greater with the maximally dissimilar teacher distribution.

We describe this result only for a simplified setting in $D = 2$, the starting similarity can be seen to differ empirically between the different distributions Figure 5C and Figure 15.

We show below for two teacher nodes that maximizing/minimizing the angle between them maximizes/minimizes expected starting similarity between teacher and student nodes, currently leaving the generalization to a larger number of nodes as a conjecture. We separately show that random student signs have larger expected best matched similarity to a balanced teacher sign pattern than to a teacher whose readouts all have the same sign.

Proposition 2. *Let $\mathbf{w}_1^\star, \mathbf{w}_2^\star \in \mathbb{S}^1$ have angle $\theta \in [0, \pi]$, and let $\mathbf{w}_1, \mathbf{w}_2 \overset{\text{iid}}{\sim} \text{Unif}(\mathbb{S}^1)$. The expected best-matched initial similarity $s(\theta) = \mathbb{E}\left[\max_{p \in P_2} \sum_{j=1}^{2} \mathbf{w}_{p(j)}^\top \mathbf{w}_j^\star\right]$. can then be written as $s(\theta) = C \sin\left(\frac{\theta}{2}\right)$, where $C = \mathbb{E}|w_{1,x} - w_{2,x}| > 0$.*

Consequently, $s(\theta)$ is strictly increasing on $[0, \pi]$. In particular, the expected initial similarity is minimized when the teacher nodes coincide, $\theta = 0$, and maximized when they are antipodal, $\theta = \pi$.

Proof. The two permutations give

$$S_{\text{id}} = \mathbf{w}_1^\top \mathbf{w}_1^\star + \mathbf{w}_2^\top \mathbf{w}_2^\star, \tag{7}$$

$$S_{\text{swap}} = \mathbf{w}_2^\top \mathbf{w}_1^\star + \mathbf{w}_1^\top \mathbf{w}_2^\star, \tag{8}$$

so that $s(\mathbf{w}_1^\star, \mathbf{w}_2^\star) = \mathbb{E}[\max(S_{\text{id}}, S_{\text{swap}})]$. By rotational invariance of the student distribution, s does not depend on the absolute positions of the teacher vectors but only on the angle θ between them. We therefore set $\mathbf{w}_1^\star = (1, 0)^\top$ and $\mathbf{w}_2^\star = (\cos\theta, \sin\theta)^\top$ without loss of generality.

Applying $\max(a, b) = \frac{a+b}{2} + \frac{|a-b|}{2}$ and using

$$\mathbb{E}[S_{\text{id}} + S_{\text{swap}}] = (\mathbf{w}_1^\star + \mathbf{w}_2^\star)^\top \mathbb{E}[\mathbf{w}_1 + \mathbf{w}_2] = 0, \tag{9}$$

we obtain

$$s(\theta) = \tfrac{1}{2}\,\mathbb{E}\big|S_{\text{id}} - S_{\text{swap}}\big|. \tag{10}$$

Direct expansion yields

$$S_{\text{id}} - S_{\text{swap}} = (1 - \cos\theta)(w_{1,x} - w_{2,x}) - \sin\theta\,(w_{1,y} - w_{2,y}) = \mathbf{v}^\top \mathbf{d}, \tag{11}$$

where $\mathbf{v} = (1 - \cos\theta, -\sin\theta)^\top$ and $\mathbf{d} = \mathbf{w}_1 - \mathbf{w}_2$, and $w_{i,x}$ and $w_{i,y}$ denote the x and y components of $\mathbf{w}_i$, respectively. Then

$$\|\mathbf{v}\|^2 = (1 - \cos\theta)^2 + \sin^2\theta = 2(1 - \cos\theta) = 4\sin^2(\theta/2), \tag{12}$$

so $\|\mathbf{v}\| = 2\,|\sin(\theta/2)|$. The distribution of $\mathbf{d}$ is rotationally invariant, so $\mathbb{E}|\mathbf{u}^\top \mathbf{d}|$ is independent of $\mathbf{u}$ for the unit vector $\mathbf{u} := \mathbf{v}/\|\mathbf{v}\|$. The expectation $\mathbb{E}|\mathbf{u}^\top \mathbf{d}|$ does not depend on θ since $\mathbf{d}$ is rotationally invariant. We therefore rotate $\mathbf{u}$ to align with the x-axis:

$$\mathbb{E}|\mathbf{u}^\top \mathbf{d}| = \mathbb{E}|\mathbf{e}_1^\top \mathbf{d}| = \mathbb{E}|w_{1,x} - w_{2,x}| =: C > 0. \tag{13}$$

Combining with $\|\mathbf{v}\| = 2|\sin(\theta/2)|$ gives

$$s(\theta) = C\,|\sin(\theta/2)|. \tag{14}$$

□

We leave the generalization to a larger number of nodes as a conjecture.

Proposition 3. *Let the number of teacher nodes, M, be even, and let m be the amount of teacher nodes with positive readout sign. The sign similarity $s_a = \mathbb{E}\left[\max_{p \in P_M} \sum_{j=1}^{M} a_{p(j)} a_j^\star\right]$ is maximized when the teacher signs are evenly split, $m = \frac{M}{2}$, and is minimized when all teacher signs are equal, $m = 0$ or $m = M$.*

Proof. Let k denote the number of positive student readouts, so $k \sim \text{Binomial}\left(M, \frac{1}{2}\right)$. For fixed k, the optimal permutation matches as many equal signs as possible. Exactly $|k-m|$ signs must remain mismatched, and hence $s_a(m) = M - 2\,\mathbb{E}|k-m|$. The distribution of k is symmetric about $M/2$. It is therefore minimized at $m = M/2$ and maximized at the endpoints $m = 0$ and $m = M$. □

In these propositions, we do not consider possible interactions between both kinds of alignment. However, they are an indication that each individual change affects overall expected similarity in the given direction in more complicated teacher-student settings.

D Existence of different types of out-of-bounds minima

We consider training a neural network with ReLU activation function on a finite dataset with a maximum distance of R to the origin; that is, $X \in \{\mathbf{x} : \|\mathbf{x}\| \leq R\}$. Recall that for a neuron $\mathbf{w}$ the kink $k = -\frac{b}{||w||}$ is the signed distance from the origin to its hyperplane along its weight vector $\mathbf{w}_j$, and that the neuron is considered out-of-bounds if $k \leq \min_i \mathbf{w} \cdot \mathbf{x}_i$ or $k \geq \max_i \mathbf{w} \cdot \mathbf{x}_i$. If $k \geq \max_i \mathbf{w} \cdot \mathbf{x}_i$, the neuron is a *dead neuron*; that is, it is not active for any data point. This is a local minimum because the neuron carries no gradient signal.

In contrast, if $k \leq \min_i \mathbf{w} \cdot \mathbf{x}_i$, it is a *linear neuron* that is active for every data point. Writing $\varepsilon_i = y_i - y_i^\star$ for the residual, $\partial\mathcal{L}/\partial b = a\,\langle\varepsilon\rangle$ and $\partial\mathcal{L}/\partial\mathbf{w} = a\,\langle\varepsilon\,\mathbf{x}\rangle$ since $\sigma'(\mathbf{w}^\top\mathbf{x} + b) = 1$ everywhere in the always-on regime. Differentiating $k = -b/\|\mathbf{w}\|$ along gradient flow yields

$$\dot{k} = \frac{\eta}{\|\mathbf{w}\|}\left(\frac{\partial\mathcal{L}}{\partial b} + k\,\frac{\partial\mathcal{L}}{\partial\|\mathbf{w}\|}\right), \qquad \frac{\partial\mathcal{L}}{\partial\|\mathbf{w}\|} = \frac{\mathbf{w}^\top}{\|\mathbf{w}\|}\,\frac{\partial\mathcal{L}}{\partial\mathbf{w}} = a\left\langle\varepsilon\,\frac{\mathbf{w}^\top\mathbf{x}}{\|\mathbf{w}\|}\right\rangle. \tag{15}$$

The tangential component of $\dot{\mathbf{w}}$ rotates the neuron's orientation but drops out of $\dot{k}$, which depends only on b and $\|\mathbf{w}\|$.

Setting $k = -R$, the neuron re-enters the data domain ($\dot{k} > 0$) iff

$$a\left(\langle\varepsilon\rangle - R\left\langle\varepsilon\,\frac{\mathbf{w}^\top\mathbf{x}}{\|\mathbf{w}\|}\right\rangle\right) > 0. \tag{16}$$

An error term $\langle\varepsilon\rangle$ and a weight-norm term $R\,\langle\varepsilon\,\mathbf{w}^\top\mathbf{x}/\|\mathbf{w}\|\rangle$ compete; the readout sign a flips both. A neuron can therefore be trapped not because the residual lacks structure along $\mathbf{w}$, but because its readout has the wrong sign to exploit that structure, making this a trapping region. For large R the bias term is negligible and re-entry is controlled by the sign of $-a\,\partial\mathcal{L}/\partial\|\mathbf{w}\|$ alone.

The connection between fixed ReLU activation patterns and local least squares problems has previously been studied in the context of ReLU training dynamics [24]. Both the argumentation described above and the proposition below only show the existence of trapping regions for linear neurons. The experiments show that neurons do end in these regions in practice.

Proposition 4. *Consider a finite dataset $\{(\mathbf{x}_i, y_i^\star)\}_{i=1}^K$ and a single ReLU neuron with an outer bias,*

$$y_i = a\,\sigma(\mathbf{w}^\top\mathbf{x}_i + b) + c. \tag{17}$$

Then there exists a local minimum of the ReLU loss in the always active region,

$$\mathbf{w}^\top\mathbf{x}_i + b > 0, \qquad i = 1, \ldots, K, \tag{18}$$

whose loss is equal to the minimum affine least squares loss. If a lower loss is attainable outside the always active region, then this local minimum is spurious.

Proof. Let $(\beta^\star, \gamma^\star)$ be a solution of the affine least squares problem

$$\min_{\beta,\gamma}\frac{1}{2K}\sum_{i=1}^{K}\left(y_i^\star - \beta^\top\mathbf{x}_i - \gamma\right)^2. \tag{19}$$

Choose $a^\star = 1$ and $\mathbf{w}^\star = \beta^\star$ and choose $b^\star$ such that $b^\star > -\min_{1\le i\le K} \beta^{\star\top}\mathbf{x}_i$. Then, $\mathbf{w}^{\star\top}\mathbf{x}_i + b^\star > 0$, for $i = 1, \dots, K$. Finally, set $c^\star = \gamma^\star - b^\star$. The neuron is therefore active on every training sample and satisfies

$$a^\star\sigma(\mathbf{w}^{\star\top}\mathbf{x}_i + b^\star) + c^\star = \beta^{\star\top}\mathbf{x}_i + \gamma^\star. \tag{20}$$

Since the dataset is finite and all preactivations are strictly positive, there exists an open neighborhood of $(\mathbf{w}^\star, b^\star, a^\star, c^\star)$ in which the neuron remains active on every sample. Throughout this neighborhood,

$$a\,\sigma(\mathbf{w}^\top\mathbf{x}_i + b) + c = a\mathbf{w}^\top\mathbf{x}_i + ab + c. \tag{21}$$

Thus the ReLU loss coincides locally with the affine least squares loss under $\beta = a\mathbf{w}$ and $\gamma = ab + c$. Since $(\beta^\star, \gamma^\star)$ minimizes the affine least squares loss, no sufficiently small perturbation of $(\mathbf{w}^\star, b^\star, a^\star, c^\star)$ can decrease the loss. Hence this point is a local minimum of the ReLU loss. □

E Complete picture of the loss landscape of one- and two-node ReLU systems

E.1 Reducing the expression of the loss landscape

Generally, it is very difficult to completely characterize the loss landscape of even small neural networks. Here, we achieve a full characterization of the loss landscape of specific ReLU students. To this end, we systematically reduce the number of parameters by making use of simplifying assumptions and symmetry arguments to just two variables. Thus, we are able to fully characterize the loss landscape of these systems as the teacher changes structure.

The assumptions we make to reduce the number of parameters are:

- **Instantaneous readout weights**: the readout weights $\mathbf{a}$ evolve instantaneously. By assuming the embedding weights fixed, we can compute the optimal readout weights as their optimization is a convex problem. This eliminates a total of N degrees of freedom from the student.
- **Homogeneity of the activation function**: due to the positive homogeneity of the ReLU activation ($\lambda\,\sigma(\mathbf{w}^\top\mathbf{x}) = \sigma(\lambda\,\mathbf{w}^\top\mathbf{x})$ for any $\lambda > 0$, so that $\|\mathbf{w}\|\,\sigma(\hat{\mathbf{w}}^\top\mathbf{x}) = \sigma(\mathbf{w}^\top\mathbf{x})$ with $\hat{\mathbf{w}} = \mathbf{w}/\|\mathbf{w}\|$) we can fix the norm of the student weight vectors to 1 without loss of generality, absorbing $\|\mathbf{w}\|$ into the readout weight. By parameterizing the student weights in terms of their angles, N other degrees of freedom are eliminated.
- **Gaussian inputs** ($\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$): the infinite support of the Gaussian distribution removes boundary effects that arise from finite data domains, and allows us to take the limit $K \to \infty$ so that the loss is computed exactly in expectation. This eliminates the case of out-of-bounds minima.

Gaussian data average. The starting point is the finite-sample square loss over K data points,

$$\mathcal{L} = \frac{1}{2K}\sum_{i=1}^{K}\left[\mathbf{a}^\top\sigma(\mathbf{W}\mathbf{x}_i + \mathbf{b}) - \mathbf{a}^{\star\top}\sigma(\mathbf{W}^\star\mathbf{x}_i + \mathbf{b}^\star)\right]^2, \tag{22}$$

where $\sigma(z) = \max(0, z)$ is the ReLU activation applied element-wise. For $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, in the limit $K \to \infty$ the empirical average converges to an expectation over the input distribution:

$$\mathcal{L} \;\to\; \frac{1}{2}\Big\langle\left[\mathbf{a}^\top\sigma(\mathbf{W}\mathbf{x} + \mathbf{b}) - \mathbf{a}^{\star\top}\sigma(\mathbf{W}^\star\mathbf{x} + \mathbf{b}^\star)\right]^2\Big\rangle_{\mathbf{x}}. \tag{23}$$

Expanding the square and introducing the preactivations $\mathbf{z} = \mathbf{W}\mathbf{x}$ and $\mathbf{z}^\star = \mathbf{W}^\star\mathbf{x}$, which are jointly Gaussian with covariances $\langle\mathbf{z}\mathbf{z}^\top\rangle = \mathbf{W}\mathbf{W}^\top$, $\langle\mathbf{z}\mathbf{z}^{\star\top}\rangle = \mathbf{W}\mathbf{W}^{\star\top}$, and $\langle\mathbf{z}^\star\mathbf{z}^{\star\top}\rangle =$

$\mathbf{W}^\star\mathbf{W}^{\star\top}$, the loss separates into

$$\mathcal{L}(\mathbf{a},\mathbf{W},\mathbf{b}) = \frac{1}{2}\mathbf{a}^\top \underbrace{\langle\sigma(\mathbf{z}+\mathbf{b})\sigma(\mathbf{z}+\mathbf{b})^\top\rangle_{\mathbf{z}}}_{\mathbf{K}_{SS}}\mathbf{a} \;-\; \mathbf{a}^{\star\top}\underbrace{\langle\sigma(\mathbf{z}^\star+\mathbf{b}^\star)\sigma(\mathbf{z}+\mathbf{b})^\top\rangle_{\mathbf{z},\mathbf{z}^\star}}_{\mathbf{K}_{TS}}\mathbf{a} \;+\; C, \quad (24)$$

where $C = \frac{1}{2}\mathbf{a}^{\star\top}\mathbf{K}_{TT}\,\mathbf{a}^\star$ depends only on the teacher. The key observation is that each entry of these kernel matrices $[\mathbf{K}]_{ij} = \langle\sigma(\mathbf{w}_i^\top\mathbf{x}+b_i)\,\sigma(\mathbf{w}_j^\top\mathbf{x}+b_j)\rangle_{\mathbf{x}}$ depends on $\mathbf{w}_i$ and $\mathbf{w}_j$ only through their norms and their relative angle. For ReLU without bias, this expectation has the closed form [61]

$$\langle\sigma(z_i)\sigma(z_j)\rangle = \frac{\|\mathbf{w}_i\|\|\mathbf{w}_j\|}{2\pi}\left[\sin\theta + (\pi-\theta)\cos\theta\right], \qquad \cos\theta = \frac{\mathbf{w}_i^\top\mathbf{w}_j}{\|\mathbf{w}_i\|\|\mathbf{w}_j\|}, \quad (25)$$

which for unit-norm weights simplifies to $\frac{1}{2}f(r) = \frac{1}{2\pi}\left[\sqrt{1-r^2} + (\pi - \cos^{-1}r)\,r\right]$ with $r = \cos\theta$. For ReLU with bias, the expectation can also be determined in closed form [27].

Fast readout weights. Assuming the readout weights $\mathbf{a}$ evolve infinitely faster than the hidden-layer weights, they instantaneously converge to the optimum

$$\mathbf{a}(\mathbf{W},\mathbf{b}) = \langle\sigma(\mathbf{z}+\mathbf{b})\sigma(\mathbf{z}+\mathbf{b})^\top\rangle_{\mathbf{z}}^{-1}\,\langle\sigma(\mathbf{z}+\mathbf{b})\sigma(\mathbf{z}^\star+\mathbf{b}^\star)^\top\rangle_{\mathbf{z},\mathbf{z}^\star}\,\mathbf{a}^\star \;=\; \mathbf{K}_{SS}^{-1}\mathbf{K}_{TS}^\top\,\mathbf{a}^\star. \quad (26)$$

Substituting back, the loss reduces to a Schur complement:

$$\mathcal{L}(\mathbf{W},\mathbf{b}) = \frac{1}{2}\mathbf{a}^{\star\top}\Big[\mathbf{K}_{TT} - \mathbf{K}_{TS}\,\mathbf{K}_{SS}^{-1}\,\mathbf{K}_{TS}^\top\Big]\mathbf{a}^\star, \quad (27)$$

where $\mathbf{K}_{TT}$, $\mathbf{K}_{SS}$, $\mathbf{K}_{TS}$ are the teacher–teacher, student–student, and teacher–student kernel matrices, respectively. For ReLU, the student weight norms cancel due to homogeneity, so the loss depends *only on angles and biases.*

E.2 Derivation of the single-neuron normalized loss Equation 3

With a single teacher and a single student neuron in $D > 1$, the only angle in the problem is the angle θ between teacher and student weight. Thus, there are two degrees of freedom: the teacher-student angle θ and the student bias b.

The kernel matrices simplify to scalars, such that the loss reduces to

$$\mathcal{L}(\theta,b) = \frac{1}{2}(a^\star)^2\Big[\langle\sigma(z^\star+b^\star)^2\rangle_{z^\star} - \frac{\langle\sigma(z+b)\sigma(z^\star+b^\star)\rangle_{z,z^\star}^2}{\langle\sigma(z+b)^2\rangle_z}\Big]. \quad (28)$$

Using homogeneity and introducing the kink $k = -b/\|\mathbf{w}\|$, we get

$$\frac{\mathcal{L}(\theta,k)}{\frac{1}{2}(a^\star)^2\langle\sigma(z^\star+b^\star)^2\rangle_{z^\star}} = 1 - \frac{\langle\sigma(z-k)\sigma(z^\star-k^\star)\rangle_{z,z^\star}^2}{\langle\sigma(z-k)^2\rangle_z\langle\sigma(z^\star-k^\star)^2\rangle_{z^\star}} \quad (29)$$

where $z, z^\star$ are zero-mean and unit-variance with correlation $\langle zz^\star\rangle = \cos\theta$, and the denominator on the right hand side is a teacher-specific constant.

E.3 No spurious local minima exist in single-neuron systems with fast readout weights and infinite data

We give the proof for Proposition 1 in this section.

Proof. Beginning from Equation 3, we have the normalized loss as

$$\frac{\mathcal{L}(\theta,k)}{\frac{1}{2}(a^\star)^2\mathbf{K}_{TT}} = 1 - \frac{\mathbf{K}_{TS}^2}{\mathbf{K}_{SS}\mathbf{K}_{TT}}.$$

Since $\mathbf{K}_{SS}$ and $\mathbf{K}_{TT}$ do not depend on the overlap $\cos\theta$,

$$\frac{\partial}{\partial(\cos\theta)}\left(\frac{\mathcal{L}(\theta,k)}{\frac{1}{2}(a^\star)^2\mathbf{K}_{TT}}\right) = -\frac{2\mathbf{K}_{TS}}{\mathbf{K}_{SS}\mathbf{K}_{TT}}\frac{\partial\mathbf{K}_{TS}}{\partial(\cos\theta)}.$$

By Price's theorem,

$$\frac{\partial\mathbf{K}_{TS}}{\partial(\cos\theta)} = \langle\sigma'(z-k)\sigma'(z^\star-k^\star)\rangle_{z,z^\star}\,.$$

Since $\sigma'(z-k) = \mathbf{1}_{\{z>k\}}$ for all $z \neq k$, $\frac{\partial\mathbf{K}_{TS}}{\partial(\cos\theta)} = \mathbb{P}\left(z > k,\, z^\star > k^\star\right) > 0$ for finite $k, k^\star$ and $-1 < \cos\theta < 1$. Furthermore, since $\mathbf{K}_{TS} > 0$, $\mathbf{K}_{SS} > 0$, $\mathbf{K}_{TT} > 0$,

$$\frac{\partial\mathcal{L}}{\partial(\cos\theta)} < 0 \qquad \text{for } -1 < \cos\theta < 1.$$

Thus the derivative with respect to the overlap never vanishes in the interior. Consequently, $\nabla_{(\theta,k)}\mathcal{L} \neq 0$ for every point with $-1 < \cos\theta < 1$. Hence the loss has no interior critical points, and therefore no interior local minima, maxima, or saddle points. □

E.4 Two neurons without bias and $D = 2$.

In the two-neuron case, both teacher and student biases are set to zero for the analytical derivation of the loss landscape.

We first give the proof of Lemma 1.

Proof. In $D = 2$, each unit norm embedding weight is specified by its angle on the unit circle. Rotational invariance of the Gaussian input distribution allows the first teacher neuron to be fixed at angle 0. The teacher teacher, teacher student, and student student kernel matrices in Equation (27) therefore depend only on the angular differences between their corresponding weight vectors. The equation follows from substituting in these angular kernel matrices into Equation (27),

$$2\mathbf{K}_{TT} = \begin{pmatrix} 1 & f(\cos\theta^\star) \\ f(\cos\theta^\star) & 1 \end{pmatrix}, \qquad 2\mathbf{K}_{TS}^\top = \mathbf{T}_S, \qquad \frac{1}{2}\mathbf{K}_{SS}^{-1} = \mathbf{G}(\theta_1 - \theta_2), \tag{30}$$

Here, $f(\cos\theta) = \frac{1}{\pi}\left[\sin\theta + (\pi-\theta)\cos\theta\right], \theta \in [0,\pi]$ is the normalized Gaussian ReLU kernel [20, 61].

□

Prior work has shown that two-neuron ReLU teacher networks with equal positive teacher signs do not admit any spurious minima [28, 29]. As adding instantaneous readout weights aid convergence and our simulations support the fact that two-neuron teacher-student systems with equal teacher signs do not allow spurious local minima, we do not show this again but assume that this result holds also for instantaneous readout weights. Remarkably, analyzing the loss landscape reveals that for opposite-signed readouts the system exhibits true local minima (Figure 2C).

F Additional simulation results

F.1 Overall convergence statistics for the different settings

Table 1: Overall convergence statistics across for experimental settings (+-, dissim.), (++. sim.), and standard. Cumulatively, these settings start with 1500 runs. The D *sweep* varies the input dimensionality $D \in \{2, 4, 8, 16\}$ with the number of teacher nodes fixed at $M = 4$; the M *sweep* varies the number of teacher nodes $M \in \{2, 4, 8, 16\}$ with input dimensionality fixed at $D = 2$. (We repeated the shared $M = 4, D = 2$ setting both times independently.) In all settings the student has the same number of nodes and the same activation function as the teacher ($\rho = 1$). *Maximal* = maximally dissimilar distribution; *Minimal* = minimally dissimilar distribution.

Distribution	Activation	Sweep	Value	Completed	Global min
Maximal	ReLU	D sweep	2	1500	973
			4	1500	1208
			8	1500	1231
			16	1000	725
		M sweep	2	1500	1076
			4	1500	972
			8	1500	757
			16	267	8
	Softplus	D sweep	2	1500	1204
			4	1500	1341
			8	1500	1360
			16	1003	875
		M sweep	2	1500	1333
			4	1500	1204
			8	1500	509
			16	1172	0
	Tanh	D sweep	2	1500	1230
			4	1500	1485
			8	1500	1472
			16	1001	961
		M sweep	2	1500	1491
			4	1500	1230
			8	312	167
			16	65	0
Minimal	ReLU	D sweep	2	1500	462
			4	1500	812
			8	1500	951
			16	1000	480
		M sweep	2	1500	784
			4	1500	467
			8	1500	114
			16	302	3
	Softplus	D sweep	2	1500	142
			4	1500	589
			8	1500	503
			16	1026	159
		M sweep	2	1500	816
			4	1500	145
			8	1500	0
			16	1497	0
	Tanh	D sweep	2	1500	573
			4	1500	1358

Continued on next page

Continued from previous page

Distribution	Activation	Sweep	Value	Completed	Global min
			8	1500	1413
			16	1000	950
		M sweep	2	1500	1423
			4	1500	573
			8	702	43
			16	422	0
Standard	ReLU	D sweep	2	1500	727
			4	1500	1107
			8	1500	1178
			16	1000	630
		M sweep	2	1500	1008
			4	1500	725
			8	1500	407
			16	289	30
	Softplus	D sweep	2	1500	453
			4	1500	988
			8	1500	1161
			16	1017	670
		M sweep	2	1500	1162
			4	1500	457
			8	1262	109
			16	823	3
	Tanh	D sweep	2	1500	1001
			4	1500	1421
			8	1500	1464
			16	1003	948
		M sweep	2	1500	1479
			4	1500	1000
			8	1129	259
			16	432	2

F.2 The chosen loss threshold separates global from local minima

We use 1×10^{-18} as the loss threshold to classify global from local minima in the main text. As shown in Figure 7, there is a large gap in losses below this threshold and those above for converged student runs. This indicates that the threshold is comfortably above the precision the system reaches, and that the higher loss values therefore do not correspond to global minima.

F.3 Additional hyperparameter settings

To ensure generalizability of the results, we run the simulations on additional settings.

Additional activation functions We rerun all teacher sizes and input dimensionalities across the three different teacher distributions for the softplus activation function in Figure 8 and the tanh activation function in Figure 9. In both cases, the results are qualitatively very similar to those for ReLU.

Adam optimizer We evaluated how the results change when the optimization is performed with Adam, rather than gradient flow (Figure 10). Again, the results are qualitatively similar.

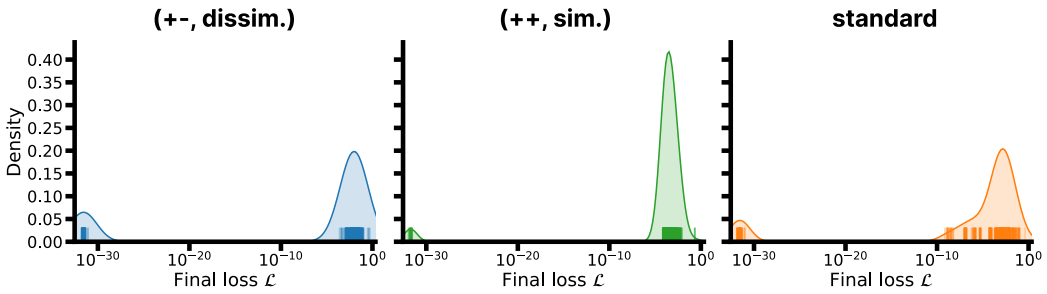

Figure 7: Distribution of losses for converged student runs for all teacher networks with $M = 4, D = 2$. (**A**) Distribution for teachers drawn from the maximal distribution. (**B**) Same as (A), but for teachers drawn from the standard distribution. (**C**) Same as (B), but for teachers drawn from the minimal distribution.

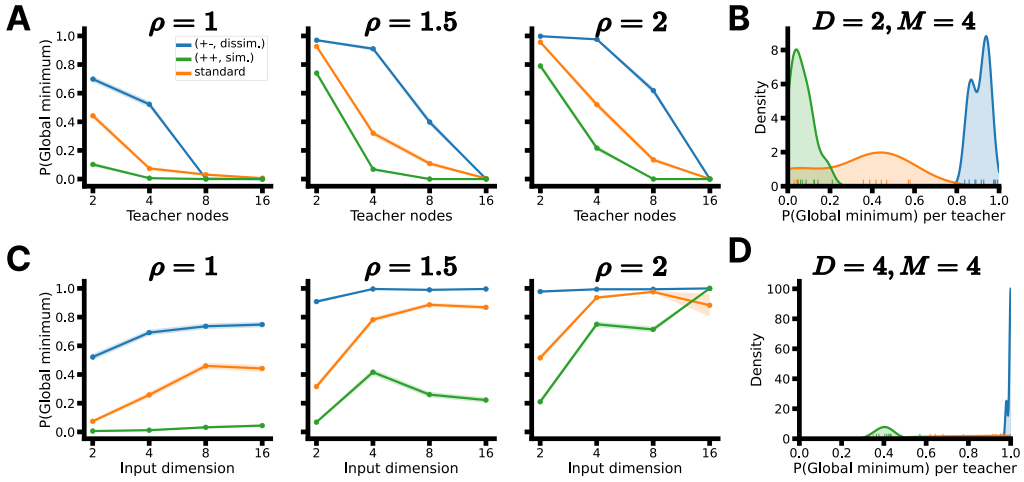

Figure 8: Comparing the teachers reaching global minimum across all three distributions for the softplus activation function. (**A**) Fraction of students reaching the global minimum as the number of teacher nodes are increased for $D = 2$. The shaded region indicates the standard error of the binomial proportion. (**B**) Frequency of students reaching the global minimum for teachers with $D = 2$ and $M = 4$, smoothed using Gaussian kernels. (**C**) Same as (A), but for increasing input dimensionality D and for fixed number of teacher nodes $M = 4$. (**D**) Same as (B), but for teachers with $D = 4$ and $M = 4$.

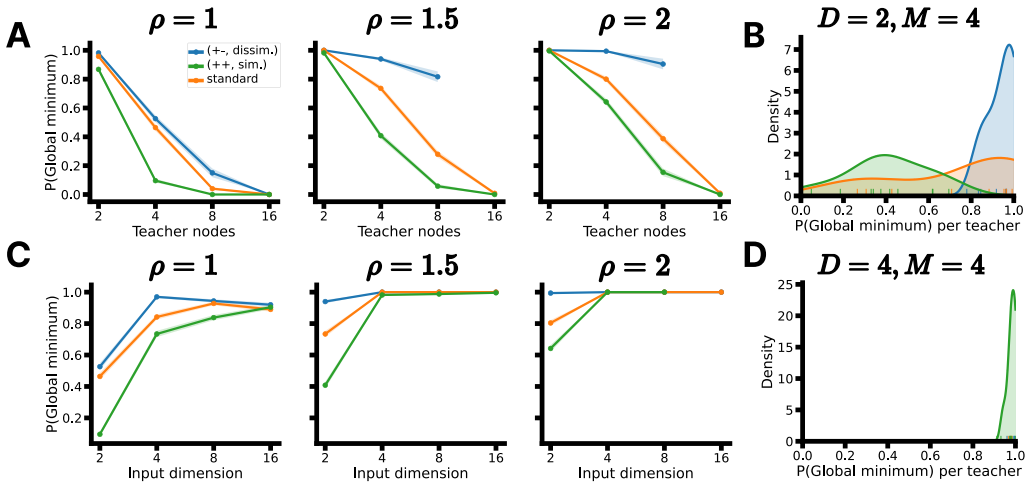

Figure 9: Comparing the teachers reaching global minimum across all three distributions for the sigmoidal tanh activation function. (**A**) Fraction of students reaching the global minimum as the number of teacher nodes are increased for $D = 2$. The shaded region indicates the standard error of the binomial proportion. (**B**) Frequency of students reaching the global minimum for teachers with $D = 2$ and $M = 4$, smoothed using Gaussian kernels. (**C**) Same as (A), but for increasing input dimensionality D and for fixed number of teacher nodes $M = 4$. (**D**) Same as (B), but for teachers with $D = 4$ and $M = 4$.

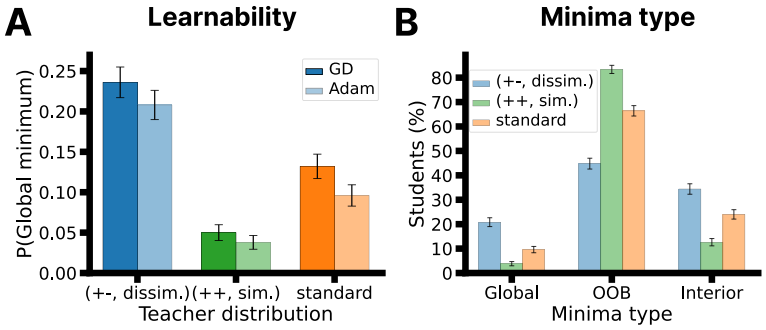

Figure 10: Results for students optimized using Adam for $M = 4$, $D = 2$. (**A**) Success results for students optimized with (*light*) Adam, compared with (*dark*) gradient flow for each of the three different teacher distributions. (**B**) Types of minima reached using Adam. Error bars represent standard error of the binomial proportion.

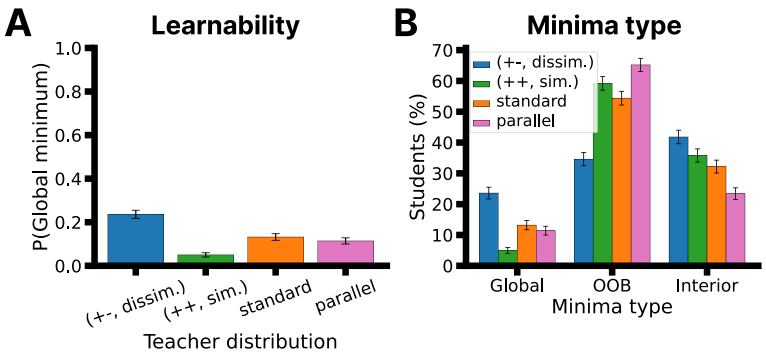

Figure 11: Results for teachers with perfectly parallel hyperplanes. (**A**) Success results for students on teachers drawn from (*pink*) the distribution with parallel hyperplanes, compared with the other three teacher distributions. (**B**) Types of minima reached on teachers with parallel hyperplanes, compared with the other three distributions. Error bars represent standard error of the binomial proportion.

Parallel hyperplanes We evaluated the students on a teacher distribution with parallel hyperplanes (Figure 11).

F.4 Empirical results for small networks

F.4.1 Single-node teacher-student systems

We ask how important starting similarity is between student and teacher network in determining success in finding the global minimum.

For this, we begin with targeted experiments in which we initialize a single teacher node ($M = 1$) with

$$\mathbf{w}^\star = \frac{1}{\sqrt{D}}(1, \ldots, 1)^\top, \qquad b^\star = 0, \qquad a^\star = 1. \tag{31}$$

We initialize single-node students ($N = 1$) along a line on the sphere $\|\mathbf{w}\|_2 = 0.1$ with $b = 0$. For both values $a = 0.1$ and $a = -0.1$, we vary the angle θ between $\mathbf{w}$ and $\mathbf{w}^\star$ over 20 linearly spaced values between 0 and π. We repeat the experiment for dimensionalities in $\{1, 2, 4, 8, 16\}$. The results are shown in Figure 12.

If instead of constructing the student initializations in this way, we instead randomly sample the initial student weights according to the Glorot distribution, we see a slight positive trend in success rates as a function of dimensionality (Figure 12C).

F.5 Two-neuron teacher-student systems

Next, we perform an experiment where we fix a teacher node, place a second teacher node with a given offset, and rotate it once around. We calculate the student success rate at each level. We find that the success rate roughly increases as the student node is rotated around to point in the exact opposite direction of the teacher node. Further, success rates are larger across the board when the two teacher nodes in the network have different signs than when they have the same sign (Figure **??**F).

F.6 Extended analysis of the local minima

Minima types We plot the types of minima reached by the different teacher distributions at each level in Figure 14.

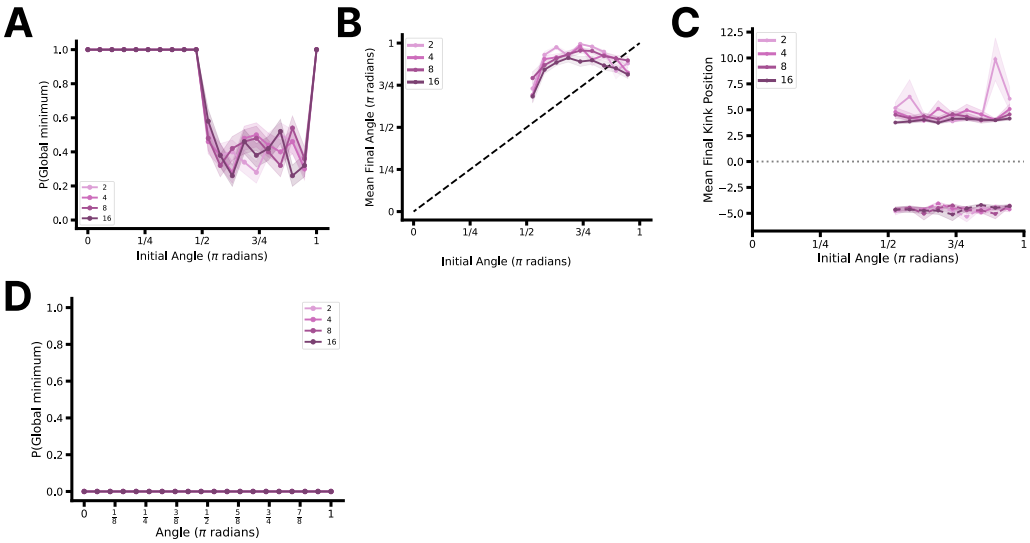

Figure 12: Importance of initial similarity between student and teacher neuron in different number of input dimensions. (**A**) Success rates as a function of different initial angles between student and teacher nodes for different input dimensions (*shade of purple*) simulated as described in Appendix B.1. (**B**) Final angle for neurons that reach suboptimal local minima in single-neuron student-teacher systems starting from the correct readout sign. The unsuccessful neurons are rotated away from the teacher. (**C**) Same as (B), but showing the final kink position. (**D**) Success rates for teachers and students with a single neuron when the student has an opposite sign of the teacher. The shaded region indicates the standard error of the binomial proportion. (**C**) Success rates for a single neuron as a function of input dimensionality. The shaded region indicates the standard error of the binomial proportion.

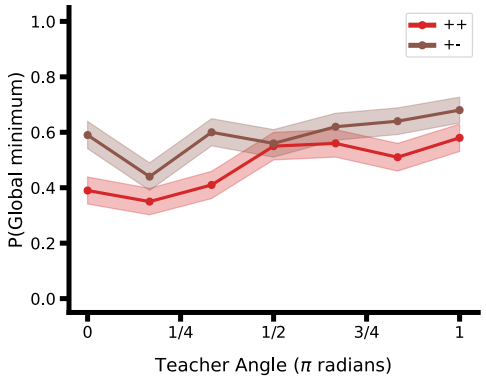

Figure 13: Fraction of students reaching the global minimum as a function of the angle between the two teacher nodes for gradient flow with finite data. The shaded region indicates the standard error of the binomial proportion.

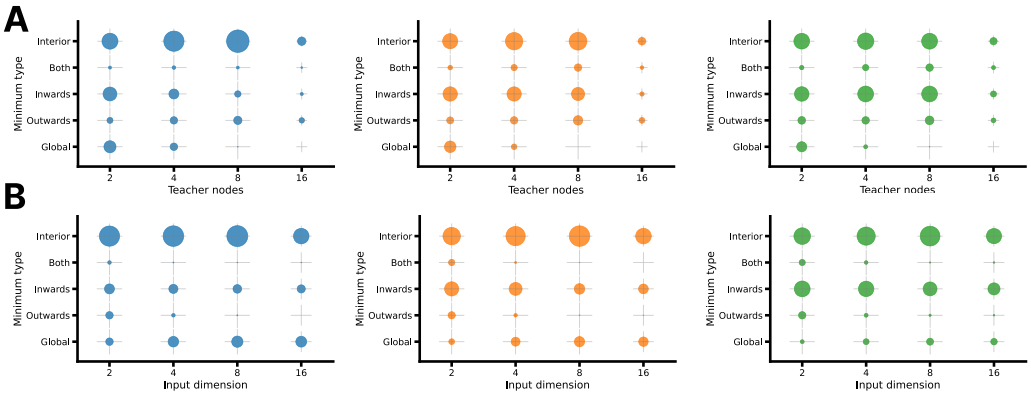

Figure 14: Local minima reached by the system for $\rho = 1$. (**A**) Bubble plots showing how the reached minima change for increasing M for each of the different teacher distributions with $D = 2$ constant. The size of the bubbles indicates how frequently that neuron is reached for each of the different teacher distributions. The crosses indicate the scale that corresponds to 100% of neurons. (**B**) Same as in (A), but for increasing D with $M = 4$ constant.

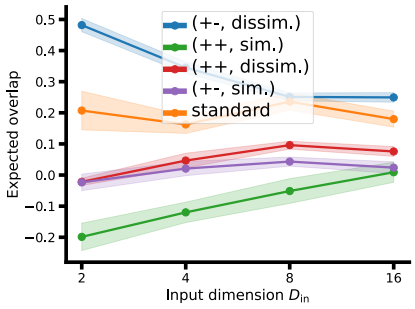

Figure 15: Average starting overlap as a function of increasing dimensionality.

Overlap terms Overlap terms tend towards zero for increasing dimensionality (Figure 15).

AUROC prediction We plot individual overlaps vs. losses for teachers for a single setting in Figure 16.

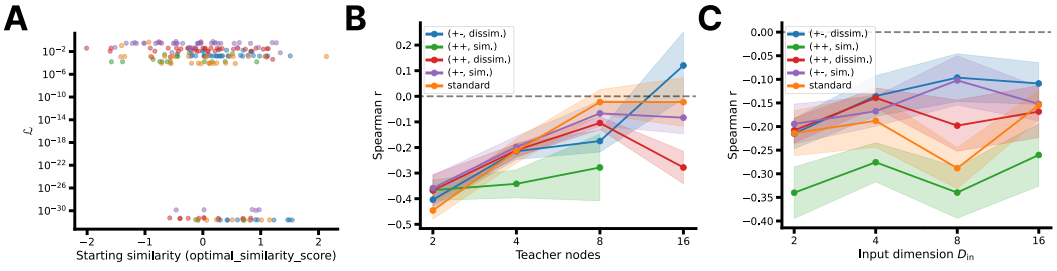

Figure 16: Relationship between starting similarity and final success, as quantified using AUROC.

G Real-world physics dataset

As described in the main text, we compute the function metrics Hessian top share, sign ratio, and linearity on the set of physics equations. After fitting, we then compare these with metrics computed on the trained neural network over units with with $\|w_j\| > 0$. *Directional (unweighted) alignment* is the mean absolute cosine between unit weight directions over distinct pairs, $\langle |\hat{w}_i^\top \hat{w}_j| \rangle_{i<j}$ with $\hat{w}_j = w_j/\|w_j\|$, counting every unit equally; it is sign-blind and asks only whether units share an *axis*. *Same-sign slope agreement* instead asks whether the units' slopes reinforce or cancel. Unit j has incoming weights w_j and an outgoing readout weight v_j; where it is active its contribution to the output gradient is $v_j w_j$, so its slope direction is $s_j = \mathrm{sign}(v_j)\,\hat{w}_j$. Weighting each unit by its contribution $c_j = |v_j|\,\|w_j\|$, normalized to $p_j = c_j / \sum_k c_k$, it is the contribution-weighted mean signed cosine between slope directions over distinct pairs, slope-agreement $= \frac{\sum_{i<j} p_i\, p_j\, (s_i^\top s_j)}{\sum_{i<j} p_i\, p_j} \in [-1, 1]$, equal to $+1$ when every unit's slope points the same way, 0 at chance, and negative when the slopes oppose. Because the readout sums the units, two units that share an axis but carry opposite readout signs cancel, which a target with a large linear part cannot afford; it is thus the network-side counterpart of the coherence of the target's own gradient field. A unit is *out of bounds* when its ReLU kink, at $\kappa_j = -b_j/\|w_j\|$ along $\hat{w}_j$, has left the data: with pre-activations $z_j = w_j \cdot x + b_j$, it is always on (affine, $\mathrm{relu}(z) = z$) when $Q_{0.01}(z_j) > 0$ and dead when $Q_{0.99}(z_j) < 0$, i.e. its kink lies outside the inner 98% of the data on its own axis; we report the fraction of units in either state. *Training MSE* is the final mean squared error on the standardized training set ($R^2 = 1 - \mathrm{MSE}$).

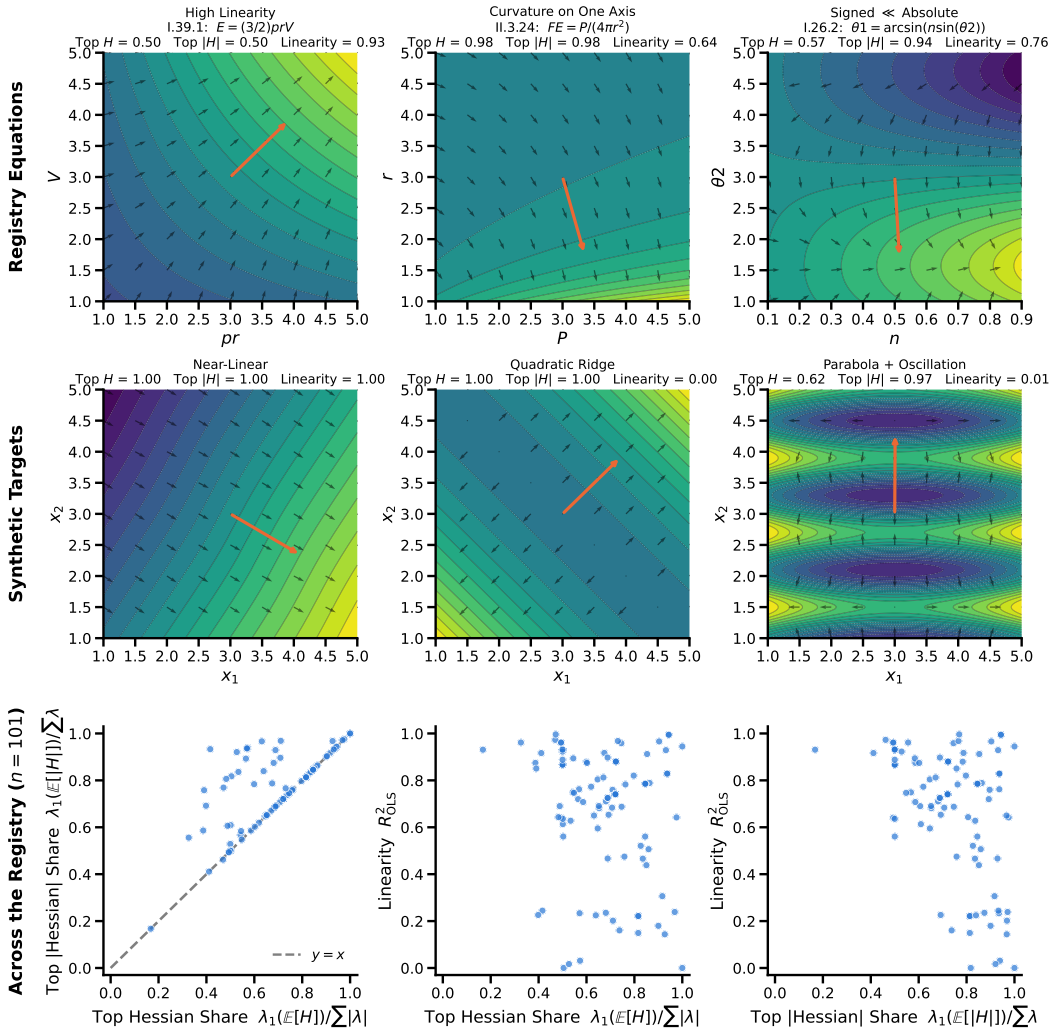

Figure 17: Three descriptors of a target function: top Hessian share, top absolute Hessian share, and linearity. Each map shows a standardized two-input target (colour) with its gradient field (black arrows) and mean gradient (orange), annotated with the top Hessian share $\mathrm{Top}\, H = \max_i |\lambda_i| / \sum_i |\lambda_i|$ of the expected Hessian $\mathbb{E}[H]$, the same share of the expected absolute Hessian $\mathbb{E}[|H|]$ ($\mathrm{Top}\,|H|$), and linearity (the R^2 of the target's best linear fit). (**A**) Top row: three AI Feynman equations spanning the descriptors — a near-linear target, one whose curvature lies on a single axis, and one whose signed curvature reverses so that $\mathrm{Top}\, H \ll \mathrm{Top}\,|H|$. (**B**) Middle row: synthetic two-input targets built to isolate each descriptor; "parabola + oscillation" shows the signed/absolute gap, which opens only when curvature spans more than one axis. (**C**) Bottom row: the descriptors against one another across all $n = 101$ equations; the dashed line is the bound $\mathrm{Top}\,|H| \geq \mathrm{Top}\, H$, since entrywise $|H|$ cannot cancel.

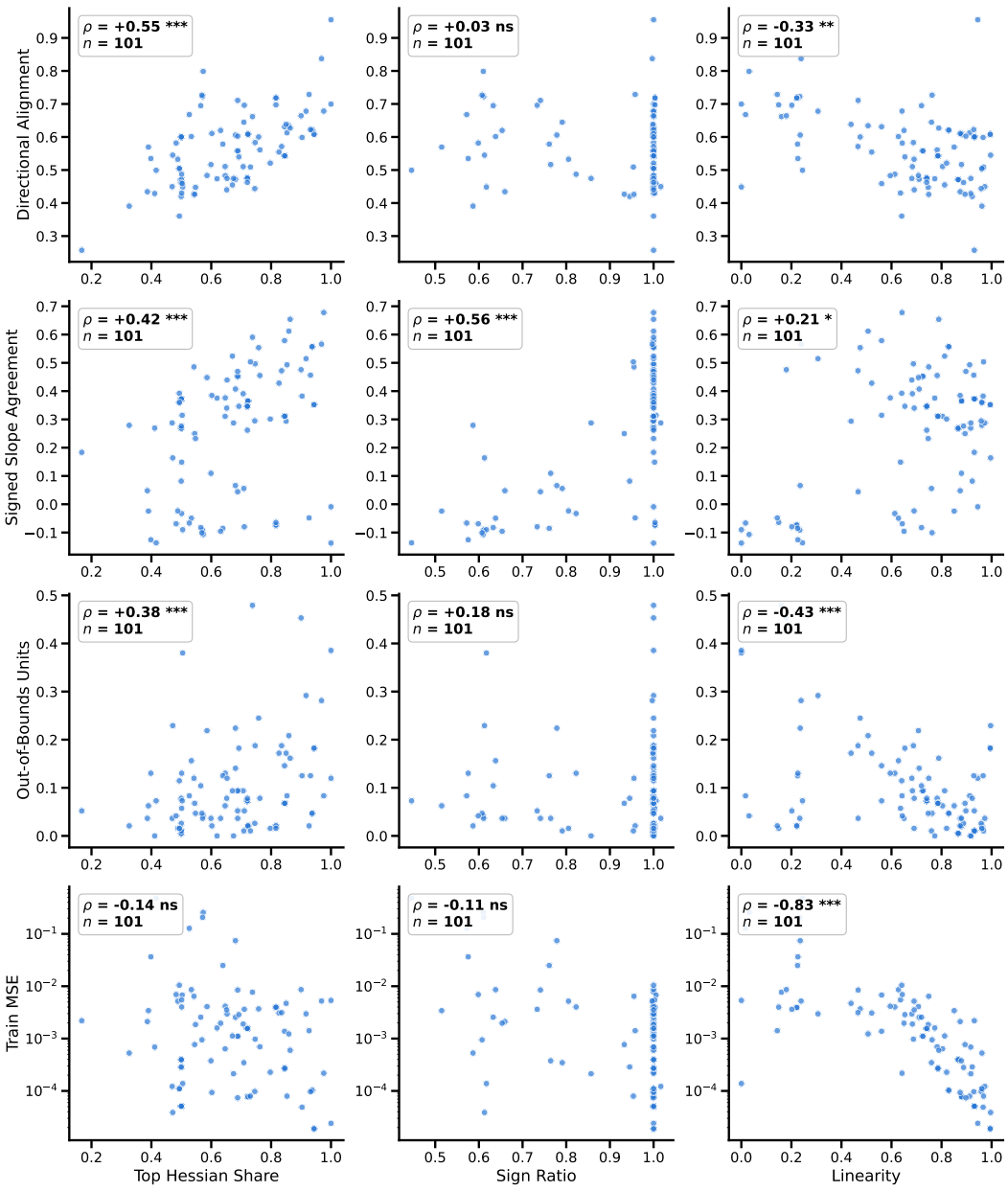

Figure 18: Target curvature predicts what the trained hidden layer does. Three target descriptors (columns) against four properties of the trained layer (rows), one point per equation ($n = 101$; ReLU, 16 hidden units, 960 epochs). Columns: the top Hessian share $\max_i |\lambda_i| / \sum_i |\lambda_i|$ of $\mathbb{E}[H]$; the sign ratio (Top H divided by the same share of $\mathbb{E}[|H|]$), i.e. how much of the leading concentration survives sign cancellation; and linearity, the R^2 of the target's best linear fit. Rows: directional (unweighted) alignment between hidden-unit directions; same-sign slope agreement; the fraction of out-of-bounds units (kink outside the inner 98% of the data, i.e. dead or purely linear); and training MSE (log axis). Each panel gives the Spearman correlation ρ with Benjamini–Hochberg-adjusted significance ($^*q < 0.05$, $^{**}q < 0.01$, $^{***}q < 0.001$) and n.

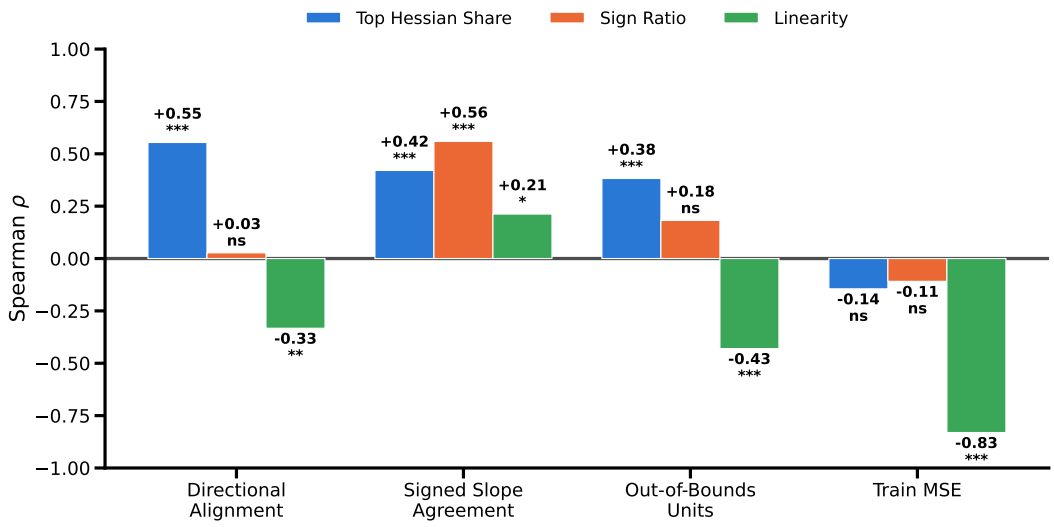

Figure 19: Target curvature predicts hidden-layer alignment and same-sign slope agreement across the AI Feynman equations. Spearman correlation ρ between three target descriptors — top Hessian share (blue), sign ratio (orange), and linearity (green, the R^2 of the target's best linear fit) — and four properties of the trained hidden layer ($n = 101$; ReLU, 16 hidden units, 960 epochs). Bars are signed; labels give ρ and Benjamini–Hochberg-adjusted significance ($^*q < 0.05$, $^{**}q < 0.01$, $^{***}q < 0.001$, ns otherwise). The sign ratio is the strongest predictor of same-sign slope agreement, whereas linearity dominates training MSE.