Gradient Descent on Two ReLU Neurons: Global Landscape and Bifurcation Dynamics

Binghua Li[*]
University of California, Berkeley
binghuali@berkeley.edu

Mengzhe Li[*]
École Normale Supérieure Paris-Saclay
mengzhe.li_mengzhe@ens-paris-saclay.fr

Denny Wu
New York University & Flatiron Institute
dwu@flatironinstitute.org

Tianhao Wang
University of California, San Diego
tianhaowang@ucsd.edu

Abstract

Understanding how gradient descent learns features remains a central challenge in neural network theory. We study this question in a minimal multi-index setting that already exhibits rich multi-phase gradient dynamics: a realizable, matched-architecture two-neuron ReLU teacher-student model under isotropic Gaussian inputs, trained by population gradient flow from small initialization. We show that the dynamics are organized by two directions: the "easy" bisector direction, which carries the leading signal, and the "hard" splitting direction, which governs specialization. To characterize the loss structure and learning behavior, we analyze the population squared-loss landscape and show that every critical point with both student weights nonzero is either a global minimum or a saddle. We then track the gradient-flow trajectory, showing that student neurons first collapse toward a bisector saddle before escaping and specializing to teacher neurons. Our results provide both a landscape and dynamics account of the multi-phase symmetry-breaking behavior that arises in a simple multi-index model and under standard algorithmic and architectural choices.

1 Introduction

A central promise of gradient-based training of neural networks is *feature learning*: rather than fitting the target function on a fixed feature space, neural networks can extract features by aligning their parameters with the latent structure of the data or target function. A recent line of theoretical works has made this intuition precise by analyzing teacher-student models under Gaussian inputs, where the target function is a *multi-index model* (Abbe et al., 2022; Damian et al., 2022; Ba et al., 2022) which depends on a small number of input directions (e.g., a shallow neural network with $O_d(1)$ neurons). In these settings, it has been demonstrated that gradient descent training of a shallow "student" neural network discovers the relevant subspace, and achieves a sample complexity superior to fixed features methods (or neural networks in the "lazy" regime (Chizat et al., 2019)).

However, most existing theory becomes tractable only after imposing simplifying assumptions that deviate from standard algorithmic and architectural choices. Examples include (i) near-orthogonal teacher neurons (Li et al., 2020; Ren and Lee, 2025), (ii) local convergence conditions (Zhou et al., 2021; Zhu et al., 2025), (iii) activation functions with high *information exponent* (Oko et al., 2024; Şimşek et al., 2025; Ren et al., 2025), and (iv) modified algorithms such as loss preprocessing (Damian et al., 2022; Zhang et al., 2026) and spherical/Stiefel gradients (Ben Arous et al., 2021,

[*]Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

2026; Abbe et al., 2023). While these assumptions enable clean analyses, they do not capture the complex *multi-phase* learning dynamics present in standard ReLU networks (Boursier et al., 2022; Wang and Ma, 2023; Min et al., 2025). A notable multi-index example exhibiting such nontrivial dynamics is the realizable setting in which both the student and teacher are two-layer neural networks with matching activation functions and more than one neuron: for commonly used nonlinearities such as ReLU, the optimization trajectory exhibits symmetry-breaking and bifurcation behavior, with neurons initially moving toward an "average" direction before specializing to different teacher directions. While this plateau-specialization phenomenon has appeared in the statistical physics literature (Saad and Solla, 1995; Goldt et al., 2020), it is not captured by recent analyses that rely on the above simplifications. To our knowledge, a rigorous and quantitative understanding of the loss landscape and multi-phase training dynamics remains incomplete and largely open.

1.1 Our contributions

We study arguably the simplest multi-index model setting that exhibits this bifurcation phenomenon: learning a two-neuron ReLU network under isotropic Gaussian inputs via population gradient flow, starting from small initialization. Specifically, the teacher model is

$$f^*(x) = \mathrm{ReLU}(\langle v_1, x\rangle) + \mathrm{ReLU}(\langle v_2, x\rangle), \qquad x \sim \mathcal{N}(0, I_d), \tag{1.1}$$

where $\mathrm{ReLU}(z) = \max\{z, 0\}$, $\|v_1\|_2 = \|v_2\|_2 = 1$ and $\angle(v_1, v_2) = \alpha \in (0, \frac{\pi}{2}]$, and the student has the same architecture with trainable weights $w_1, w_2 \in \mathbb{R}^d$. Unlike the teacher weights, the student weights are not constrained in norm. We call $V := \mathrm{span}\{v_1, v_2\}$ the *teacher plane*. A pair of student weights is *in-plane* if both weights lie in V, and *out-of-plane* if at least one of them has a nonzero component orthogonal to V.

As we will see, the geometry of this model is organized by two directions in the teacher plane: the bisector $(v_1 + v_2)$, which carries the dominant signal, and the difference $(v_1 - v_2)$, which governs specialization. In the language of leap complexity for Gaussian multi-index models (Abbe et al., 2023; Bietti et al., 2025), the target function (1.1) has *leap exponent* 2, where the bisector and difference vectors specify the "easy" leap-1 direction and the "hard" leap-2 direction, respectively. This suggests a two-stage picture for gradient-based learning: the student neurons are first attracted toward the leap-1 bisector, and only later separate toward the individual teacher directions along the leap-2 direction. Our goal is to make this collapse-then-splitting picture quantitative by studying both the population squared loss landscape and the gradient-flow dynamics from small initialization. Specifically,

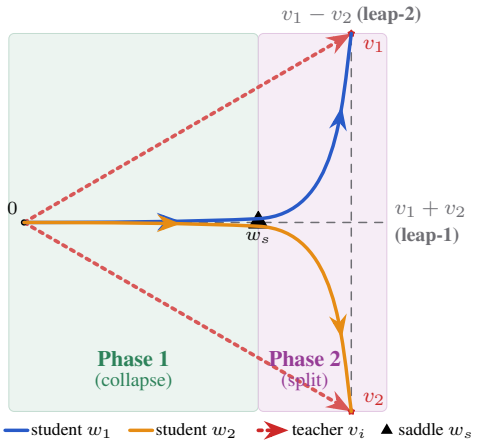


Figure 1: Illustration of bifurcation dynamics.

- **Landscape analysis.** Section 3 studies the loss geometry and the critical-point structure for both the correlation loss and the MSE loss. We reduce the stationarity conditions to a three-dimensional problem and isolate the remaining in-plane and out-of-plane critical families. We establish that every critical point of the population loss with both student weights nonzero is either a global minimum or a saddle, so none of these critical points is a spurious local minimum. This resolves a two-neuron conjecture from Tian (2017) for every teacher angle $\alpha \le \alpha_\star \approx 127.5°$, including obtuse angles.

- **Small-initialization dynamics.** Section 4 studies the gradient-flow trajectory from small initialization and characterizes the symmetry breaking and specialization behavior — see Figure 1. We show that for sufficiently small initialization, both student neurons stay close to the bisector for fixed finite time and approach the symmetric saddle. We then analyze escape from that saddle, including global convergence to a minimum for typical initializations, as well as a quantitative escape-time bound that is logarithmic in the ambient dimension d with a structured initialization that small i.i.d. Gaussian initialization satisfies with constant probability.

1.2 Further related work

Learning multi-index models. Recent works have shown shallow neural networks can efficiently learn low-dimensional latent structure via gradient descent (Abbe et al., 2022; Damian et al., 2022; Ba et al., 2022; Bietti et al., 2022; Glasgow, 2024; Dandi et al., 2024). Specifically, the sample complexity has been shown to be superior to that of kernel methods and to be governed by the *information exponent* (Dudeja and Hsu, 2018; Ben Arous et al., 2021) for single-index models and the *leap exponent* (Abbe et al., 2023; Bietti et al., 2025) for multi-index models. Our target function (1.1) is an example of a leap-2 multi-index model, but, crucially, it also contains a leap-1 subspace, which complicates the SGD dynamics. In fact, most prior works studying gradient-based feature learning either directly assume link functions with information exponent 2 (Li et al., 2020; Martin et al., 2024; Ren and Lee, 2025; Ben Arous et al., 2025) or higher (Oko et al., 2024; Şimşek et al., 2025; Ren et al., 2025), or remove the low-leap components through algorithmic modifications such as deflation / projection (Ge et al., 2021; Abbe et al., 2023) and loss preprocessing (Damian et al., 2022; Zhang et al., 2026). Consequently, these analyses do not fully characterize the learning dynamics of two ReLU neurons with standard gradient-based training.

Training dynamics of ReLU networks. Restricting to shallow neural networks with ReLU activations, prior works have studied landscape properties such as the existence of spurious minima (Safran and Shamir, 2018; Safran et al., 2021); the concurrent work of Huang et al. (2026) describes the local minima of wider networks in high dimension through a few summary statistics, mostly for orthonormal teacher weights, and studies families of solutions with a block-symmetric structure. We instead focus on two neurons with possibly nonorthogonal teacher weights. For this model, we classify every stationary point with both student weights nonzero, including all saddles, and we also analyze the gradient-flow dynamics. For gradient-based training dynamics, while the single-neuron case is relatively well-understood (Tian, 2017; Soltanolkotabi, 2017; Yehudai and Shamir, 2020; Xu and Du, 2023), when the target network contains more than one neuron, existing dynamical analyses either assume local convergence, thereby bypassing the bifurcation behavior (Zhou et al., 2021; Zhu et al., 2025), or require exponential width to invoke mean-field global convergence (Chizat, 2022; Akiyama and Suzuki, 2021).

The two-phase trajectory illustrated in Figure 1 also relates to the "plateau" phenomenon observed in the statistical mechanics literature (Biehl et al., 1996; Fukumizu and Amari, 2000). Specifically, the specialization phenomenon has been investigated in Goldt et al. (2020); Oostwal et al. (2021) using the ODE analysis originated from Saad and Solla (1995); Riegler and Biehl (1995). However, to our knowledge, a complete understanding of the landscape and rigorous characterization of the learning dynamics have not yet been established even in the two-neuron setting (1.1) studied here. Our contribution is therefore not the phenomenon itself but a rigorous and quantitative description of it in this setting, without numerical or heuristic approximations.

2 Preliminary

Notation. For a nonzero vector $w \in \mathbb{R}^d$, we denote its direction by $\bar{w} := w/\|w\|_2$, and write orthogonal projection onto the subspace perpendicular to w as $P_w^{\perp} := I - \bar{w}\bar{w}^\top \in \mathbb{R}^{d\times d}$. The angle between two nonzero vectors $u, v \in \mathbb{R}^d$ is $\angle(u,v) := \arccos(\langle u, v\rangle/(\|u\|_2\,\|v\|_2))$. For vectors $a_1, \dots, a_k$, $\operatorname{span}\{a_1, \dots, a_k\}$ denotes the linear subspace generated by them. We use $\phi(z) := \mathrm{ReLU}(z) = \max\{0, z\}$ for the activation function.

2.1 Problem setup

We use the teacher f^* defined in (1.1). The student has trainable weights $w_1, w_2 \in \mathbb{R}^d$ and output

$$f(x) = \phi(\langle w_1, x\rangle) + \phi(\langle w_2, x\rangle). \tag{2.1}$$

The student is trained by gradient flow on the population squared loss under isotropic Gaussian inputs:

$$L(w_1, w_2) := \tfrac{1}{2}\,\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\big[(f(x) - f^*(x))^2\big], \quad \tfrac{\mathrm{d}w_i}{\mathrm{d}t} = -\nabla_{w_i} L(w_1, w_2), \quad i \in \{1, 2\}, \tag{2.2}$$

starting from an arbitrary nonzero initialization $w(0) = (w_1(0), w_2(0))$ with $w_1(0) \neq 0$ and $w_2(0) \neq 0$. We define its initialization scale by

$$r_{\text{init}} := \tfrac{1}{2}\big(\|w_1(0)\|_2 + \|w_2(0)\|_2\big). \tag{2.3}$$

For independent Gaussian initialization $w_i(0) \sim \mathcal{N}(0, \sigma^2 I_d)$, $i = 1, 2$, both norms are of order $\sigma\sqrt{d}$ with high probability. Our main focus is the small-initialization regime, which for the Gaussian initialization above corresponds to taking σ sufficiently small. Small initialization is a common assumption in theoretical analyses of gradient-based training dynamics (Stöger and Soltanolkotabi, 2021; Boursier et al., 2022; Ren et al., 2025). In our setting, it ensures that the teacher force dominates the interaction betweens student neurons in the early dynamics (Section 4.1 and Appendix I.4).

Within the teacher plane, we highlight two important directions, the *bisector* and the *splitting direction*, which form an orthonormal basis of the plane:

$$\bar{v} := \frac{v_1+v_2}{\|v_1+v_2\|_2}, \qquad v_\Delta := \frac{v_1-v_2}{\|v_1-v_2\|_2}. \tag{2.4}$$

We define the initial planar splitting amplitude by

$$a_{\text{init}} := \big|\langle w_1(0) - w_2(0), v_\Delta\rangle\big|. \tag{2.5}$$

For $C \geq 1$, we say that the initialization is C-balanced if

$$\max_{i=1,2} \|w_i(0)\|_2 \leq C \min_{i=1,2} \|w_i(0)\|_2. \tag{2.6}$$

We use the following shorthand for the angles between student and teacher directions:

$$\theta_{ij} := \angle(w_i, v_j), \quad i, j \in \{1, 2\}, \qquad \gamma := \angle(w_1, w_2). \tag{2.7}$$

The next subsection explains the teacher-side hierarchy behind the two-stage collapse-then-splitting dynamics in the small-initialization regime, as illustrated in Figure 1.

2.2 Leap complexity of the teacher

We use the standard Hermite-support notion of leap complexity for Gaussian multi-index models (Abbe et al., 2023; Bietti et al., 2025); here only the following two-coordinate version is needed. Let $\{h_k\}_{k\geq 0}$ denote the normalized probabilists' Hermite polynomials. Suppose U, W are independent standard Gaussians and $g \in L^2(\mathcal{N}(0, I_2))$. Its Hermite expansion in $L^2(\mathcal{N}(0, I_2))$ is

$$g(U, W) = \sum\nolimits_{m,n\geq 0} \widehat{g}_{m,n} h_m(U) h_n(W). \tag{2.8}$$

We add the two coordinates one at a time, first U and then W. The leap for U is the lowest degree of a nonzero term that involves U alone, $\min\{m > 0 : \widehat{g}_{m,0} \neq 0\}$. Once U has been learned, the leap for W is the lowest degree in W of any nonzero term that involves W, $\min\{n > 0 : \widehat{g}_{m,n} \neq 0 \text{ for some } m \geq 0\}$.

For the two-ReLU teacher network defined in (1.1), let $z_1 := \langle \bar{v}, x\rangle$ and $z_2 := \langle v_\Delta, x\rangle$ be the projections of the input onto the bisector and splitting directions defined in (2.4). Because the teacher weights satisfy $v_1, v_2 \in \text{span}\{\bar{v}, v_\Delta\}$, $f^*(x)$ depends on x only through these two coordinates, and hence we can write $f^*(x) = g(z_1, z_2)$. Also, since $(z_1, z_2) \sim \mathcal{N}(0, I_2)$, we may Hermite expand g with respect to coordinates (z_1, z_2). The following proposition shows that although the nonlinearity ϕ has information/leap exponent 1, the sum of two ReLU neurons is a leap-2 function.

Proposition 2.1 (Leap complexity of teacher)**.** *The two-ReLU teacher network f^* in (1.1) has leap exponent 2, where the bisector coordinate z_1 has leap 1 and the splitting coordinate z_2 has leap 2.*

The proof is in Appendix A.4. This leap structure gives useful intuition for the two-stage dynamics shown in Figure 1. Specifically, the leap at each subspace determines the gradient magnitude and hence the order of learning: the leap-1 bisector $\bar{v}$ gives a linear signal visible at random initialization, while the leap-2 splitting direction v_Δ appears in a smaller quadratic signal. Consequently, near small initialization the student–student interaction is negligible, and each student neuron is driven by its correlation with the teacher. At a random initialization in high dimension, each student direction is nearly orthogonal to the teacher plane, so the degree-one part of this correlation, which points toward $\bar{v}$, dominates. We show in Theorem 3.1 that, over unit-norm student weights, the full correlation term is minimized at the bisector, which results in a collapse toward $\bar{v}$ before the higher-order signal along v_Δ drives splitting and specialization.

3 Landscape analysis

We begin with a landscape analysis, with the goal of classifying all stationary points of the population squared loss with both student weights nonzero as global minima or saddles.

3.1 Warm-up: landscape of the correlation loss

We first work with a more general setting where the teacher is a width-k ReLU network,

$$f^*(x) = \sum_{i=1}^{k} \phi(\langle v_i, x\rangle), \qquad k \geq 2. \tag{3.1}$$

The student has m ReLU neurons, $f(x) = \sum_{j=1}^{m} \phi(\langle w_j, x\rangle)$, with trainable weights $w_1, \ldots, w_m \in \mathbb{R}^d$. The setup in Section 2 corresponds to the special case $k = m = 2$.

By the positive homogeneity of ReLU, the population squared loss expands as

$$\frac{1}{2}\mathbb{E}[(f(x) - f^*(x))^2] = \frac{1}{2}\mathbb{E}[f^*(x)^2] - \sum_{j=1}^{m} \|w_j\|_2 \cdot \mathbb{E}[f^*(x)\phi(\langle \bar{w}_j, x\rangle)] + O\Bigg(\sum_{i,j=1}^{m} \|w_i\|_2 \|w_j\|_2\Bigg).$$

The first term is constant in the student weights, while the last term is second order in their norms. Thus, for small student weights, the first-order decrease of the loss is governed by the correlations between the teacher output $f^*(x)$ and the normalized student activations $\phi(\langle \bar{w}_j, x\rangle)$. We refer to this leading L^2 correlation term as the correlation loss,

$$L_{\mathrm{corr}}(w_1, \ldots, w_m) := -\sum_{j=1}^{m} \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\big[f^*(x)\phi(\langle w_j, x\rangle)\big]. \tag{3.2}$$

For the width-two result below, let $\alpha_\star \in (\pi/2, \pi)$ be the unique solution of $\tan(\alpha_\star/2) = \pi - \alpha_\star/2$. Numerically, $\alpha_\star \approx 127.521476°$. While many existing works on gradient-based feature learning employ a correlation-loss approximation to simplify the SGD dynamics (Abbe et al., 2023; Damian et al., 2023; Lee et al., 2024), the following theorem shows that the collapsed configuration at the normalized average teacher direction is the unique minimizer on the unit sphere, and hence cannot recover the two teacher neurons.

Theorem 3.1. *Consider the teacher model f^* in (3.1). Suppose that $v_1, \ldots, v_k$ are unit vectors satisfying $\angle(v_i, v_j) = \alpha$ for all $i \neq j$, and that either $0 < \alpha \leq \pi/2$, or $k = 2$ and $0 < \alpha \leq \alpha_\star$. Let $\bar{v} = (\sum_{i=1}^{k} v_i)/\|\sum_{i=1}^{k} v_i\|_2$. Then, for every student width m, $(\bar{v}, \ldots, \bar{v})$ is the unique global minimum of L_{corr} among all unit-norm student weights.*

The proof is in Appendix B. For $k = 2$, the threshold $\alpha_\star$ is sharp: above it, the bisector is no longer a local minimizer of L_{corr} on the unit sphere (Lemma B.5). The following two observations connect the theorem to the training dynamics.

- This result aligns with the first stage of the dynamics observed in Figure 1: near small initialization, the squared loss is dominated by the correlation term, which is minimized when all student directions equal the averaged teacher direction $\bar{v}$. In the two-neuron setting of (1.1), this averaged direction is exactly the bisector, explaining why the student neurons initially move toward this direction. This result is close in spirit to Şimşek et al. (2025), which studies correlation-loss gradient in the high-information-exponent setting, whereas our ReLU setting has information exponent 1.
- On the other hand, since the correlation loss L_{corr} ignores the student-student interaction terms when $\|w_j\|_2$ becomes non-negligible, it cannot explain why collapsed students eventually repel and specialize. To capture this effect, we have to characterize the full squared-loss landscape.

3.2 Landscape of the squared loss

Now we turn to the squared loss (2.2). To simplify the analysis, we write the first-order condition in angular variables. Recall the angle shorthand from (2.7): $\theta_{ij} = \angle(w_i, v_j)$ and $\gamma = \angle(w_1, w_2)$.

Lemma 3.2 (Characterization of stationary points)**.** *Define $a_1 := \sin\theta_{11} + \sin\theta_{12} - \pi\|w_1\|_2 - \sin\gamma\|w_2\|_2$ and $a_2 := \sin\theta_{21} + \sin\theta_{22} - \pi\|w_2\|_2 - \sin\gamma\|w_1\|_2$. Then for $w_1, w_2 \neq 0$, (w_1, w_2) is a stationary point of the squared loss L in (2.2) if and only if*

$$\begin{aligned}(\pi - \theta_{11})v_1 + (\pi - \theta_{12})v_2 + a_1\bar{w}_1 - (\pi - \gamma)\|w_2\|_2\bar{w}_2 = 0,\\ (\pi - \theta_{21})v_1 + (\pi - \theta_{22})v_2 + a_2\bar{w}_2 - (\pi - \gamma)\|w_1\|_2\bar{w}_1 = 0.\end{aligned} \tag{3.3}$$

The proof is in Appendix C.1.1. With the above first-order condition in hand, we classify all stationary points with $w_1 \neq 0$ and $w_2 \neq 0$, up to permutation of student neurons.

Theorem 3.3 (Landscape of the squared loss)**.** *Let L be the population squared loss of the two-neuron ReLU model in (2.2), and let (w_1, w_2) be a stationary point of L with $w_1 \neq 0$ and $w_2 \neq 0$. Define*

$$r_+ := \tfrac{2}{\pi}\big((\pi - \tfrac{\alpha}{2})\cos\tfrac{\alpha}{2} + \sin\tfrac{\alpha}{2}\big), \qquad r_- := \tfrac{2}{\pi}\big(\sin\tfrac{\alpha}{2} - \tfrac{\alpha}{2}\cos\tfrac{\alpha}{2}\big). \tag{3.4}$$

Then, after possibly swapping w_1 and w_2, exactly one of the following holds.

1. ***Global minimum:*** *$w_1 = v_1$ and $w_2 = v_2$; this is a global minimum.*
2. ***In-plane saddles:*** *$w_1, w_2 \in \operatorname{span}\{v_1, v_2\}$ satisfy one of the following three forms:*

$$\begin{cases} w_1 = s_1\bar{v}, \quad w_2 = s_2\bar{v}, & \text{for some } s_1, s_2 > 0 \text{ with } s_1 + s_2 = r_+; \\ w_1 = -s_1\bar{v}, \quad w_2 = -s_2\bar{v}, & \text{for some } s_1, s_2 > 0 \text{ with } s_1 + s_2 = r_-; \\ w_1 = r_+\bar{v}, \quad w_2 = -r_-\bar{v}. \end{cases} \tag{3.5}$$

   Every such point has a strictly negative Hessian direction.

3. ***Out-of-plane saddles:*** *necessarily $d \geq 3$, and for some unit vector $\xi \perp \operatorname{span}\{v_1, v_2\}$,*

$$w_1 = r_\alpha(\kappa_\alpha)\left(\cos\kappa_\alpha \cdot \bar{v} + \sin\kappa_\alpha \cdot \xi\right), \qquad w_2 = r_\alpha(\kappa_\alpha)\left(\cos\kappa_\alpha \cdot \bar{v} - \sin\kappa_\alpha \cdot \xi\right), \tag{3.6}$$

   where $\kappa_\alpha \in (0, \pi/2)$ and $r_\alpha(\kappa_\alpha) > 0$ depend only on α (Lemma C.2). Every such point has a strictly negative Hessian direction.

Consequently, the squared loss L has no spurious local minima on $\{(w_1, w_2) : w_1 \neq 0,\ w_2 \neq 0\}$.

To the best of our knowledge, this is the first complete classification of all stationary points with both student weights nonzero (under Euclidean geometry) in a realizable, matched-architecture shallow ReLU teacher–student model with more than one teacher neuron. We make the following remark on the significance of this landscape result.

Remark 3.4. Tian (2017) conjectured that the unconstrained population squared loss for $k = 2$ ReLU teacher-student under isotropic Gaussian inputs has no spurious local minima. Tian's own analysis was conditional on a numerically-verified inequality (Conjecture 1 of Tian, 2017) and resolved only the case where both student vectors lie inside or outside $\operatorname{Cone}(v_1, v_2)$. Wu et al. (2018) made progress by proving the absence of spurious local minima for *orthogonal* teachers on the *unit-norm manifold*. Our Theorem 3.3 settles this open problem for every $\alpha \leq \alpha_\star$: it provides an unconditional, complete classification of every critical point of the unconstrained squared loss with both student weights nonzero, showing that each is either the global minimum or a saddle. The proof uses the angle only through the inequality $\sin(\alpha/2) \leq (\pi - \alpha/2)\cos(\alpha/2)$, which holds exactly when $\alpha \leq \alpha_\star$, a range that includes obtuse angles (Theorem C.8). For $\alpha > \alpha_\star$, the stationary equations have further solution families (Proposition C.7), and a numerical search finds them at every tested angle above $\alpha_\star$ and none at or below it (Appendix I.7).

We prove Theorem 3.3 in three steps. The first is a dimension reduction: any solution (w_1, w_2) of (3.3) is contained in the teacher plane together with at most one orthogonal direction. The next two steps analyze the resulting three-dimensional problem and the planar cases that remain.

3.2.1 Step 1: reduce to three dimensions

The dimension reduction is a simple consequence of the characterization in Lemma 3.2. It is useful to decompose each student weight into the part lying in the teacher plane $\operatorname{span}\{v_1, v_2\}$ and the part orthogonal to that plane. For a solution of (3.3), the two orthogonal parts cannot vary independently: the system (3.3) forces them to be linearly dependent, as summarized in the following proposition.

Proposition 3.5. *Let $w_1, w_2 \neq 0$ satisfy (3.3). For $i = 1, 2$, decompose $w_i = w_i^{\parallel} + w_i^{\perp}$, where $w_i^{\parallel} \in \operatorname{span}\{v_1, v_2\}$ and $w_i^{\perp} \perp \operatorname{span}\{v_1, v_2\}$. Then $w_1^{\perp}$ and $w_2^{\perp}$ are linearly dependent. Consequently, $\dim \operatorname{span}\{v_1, v_2, w_1, w_2\} \leq 3$.*

The proof is in Appendix C.1.3. Consequently, for any stationary point (w_1, w_2), the vectors w_1, w_2 are contained in the teacher plane together with at most one orthogonal direction.

3.2.2 Step 2: reduce the nonplanar case to two dimensions

Given the previous reduction, there are two possible types of stationary points: either both w_1 and w_2 lie in the teacher plane, or at least one of them lies outside the teacher plane. We next show that for the latter case, the weights lie in the plane spanned by the bisector and an orthogonal direction.

Lemma 3.6. *Let $w_1, w_2 \neq 0$ satisfy (3.3). If at least one of w_1, w_2 lies outside the teacher plane, then $w_1, w_2 \in \operatorname{span}\{\bar{v}, \xi\}$ for some unit vector $\xi \perp \operatorname{span}\{v_1, v_2\}$.*

The proof is in Appendix C.2.2. Geometrically, Lemma 3.6 says that for a stationary point, if one weight is not contained in the teacher plane, both weights have zero component along the splitting direction v_Δ. Thus the three-dimensional problem reduces to two planar problems.

3.2.3 Step 3: solve the 2D cases

Out-of-plane solutions. We first consider solutions in $\operatorname{span}\{\bar{v}, \xi\}$ that do not lie in the teacher plane. By Lemma C.5, such a solution is symmetric with respect to the teacher plane: after possibly replacing ξ by $-\xi$, the two student weights have the same component in $\bar{v}$ and opposite components in ξ. This corresponds to the out-of-plane saddle form in (3.6). The stationarity condition therefore reduces to a scalar equation for the angle away from the teacher plane.

Lemma 3.7 (Out-of-plane stationary points)**.** *Let $\alpha \in (0, \pi/2]$, suppose $d \geq 3$, and let ξ be any unit vector perpendicular to $\operatorname{span}\{v_1, v_2\}$. For any nonzero $w_1, w_2 \in \operatorname{span}\{\bar{v}, \xi\}$ such that $\langle w_1, \xi\rangle \neq 0$ or $\langle w_2, \xi\rangle \neq 0$, the pair (w_1, w_2) is a stationary point of the squared loss (2.2) if and only if, after possibly swapping the students or replacing ξ by $-\xi$, it has the form (3.6). Furthermore, every such stationary point has a strictly negative Hessian direction and thus is a saddle.*

The proof is in Appendix C.3. Geometrically, the teacher pulls both students toward the bisector, and when their components orthogonal to the teacher plane are equal in size and opposite in sign, the interaction between the students balances this pull. Since the Gaussian input is rotationally invariant, every unit vector $\xi \perp V$ gives such a point, and each is a saddle because moving the two students apart along $(v_\Delta, -v_\Delta)$ decreases the loss at second order.

In-plane solutions. We now turn to the other planar case where $w_1, w_2 \in \operatorname{span}\{v_1, v_2\}$. This case contains the global minima where the student weights match the teacher weights up to permutation; the only other stationary points are the three bisector-supported families listed below.

Lemma 3.8 (In-plane stationary points)**.** *Let r_+ and r_- be as defined in (3.4). Let $w_1, w_2 \neq 0$ satisfy $w_1, w_2 \in \operatorname{span}\{v_1, v_2\}$. Then (w_1, w_2) is a stationary point of the squared loss (2.2) if and only if, after possibly swapping w_1 and w_2, either (w_1, w_2) is a global minimum with $w_1 = v_1, w_2 = v_2$ or (w_1, w_2) is a saddle satisfying one of the three forms in (3.5).*

The in-plane classification is proved in Appendix C.4. Combining the above results yields the complete classification of stationary points in Theorem 3.3. The symmetric point $w_1 = w_2 = (r_+/2)\bar{v}$ in the positive-bisector saddle family is the saddle shown in Figure 1.

4 Training dynamics

Next we study the gradient-flow trajectory from small initialization. We write $r_s = r_+/2$ with r_+ from Theorem 3.3, $w_s = r_s\bar{v}$, and $w_{\text{saddle}} := (w_s, w_s)$ for the symmetric positive-bisector saddle.

4.1 Phase 1: alignment with the bisector

We first study how the gradient flow approaches the saddle point w_{saddle}. This happens in two subphases: the student directions first align with $\bar{v}$, and then their norms grow to r_s.

Subphase 1: Directional alignment. By direct calculation, the student directions along gradient flow admit the following evolution:

$$\frac{\mathrm{d}\bar{w}_i}{\mathrm{d}t} = \frac{1}{2\pi\|w_i\|_2} P^{\perp}_{\bar{w}_i}\Big(\underbrace{(\pi - \theta_{i1})v_1 + (\pi - \theta_{i2})v_2}_{\text{teacher force}} - \underbrace{(\pi - \gamma)w_{3-i}}_{\text{student interaction}} \Big), \qquad i = 1, 2.$$

The projected teacher force is the directional part of the teacher-student correlation term analyzed in Theorem 3.1. Note that this term dominates the student interaction when $\|w_{3-i}\|_2$ is very small. Thus

the teacher force controls the angular motion while the weights are small. Moreover, the projected teacher force points toward the bisector by Lemma E.3: $\langle \bar{v}, P^{\perp}_{\bar{w}_i}\big((\pi-\theta_{i1})v_1+(\pi-\theta_{i2})v_2\big)\rangle \geq 0$, with equality only at $\bar{w}_i = \bar{v}$ or $\bar{w}_i = -\bar{v}$. Thus, except at these two directions, the teacher force increases the alignment of the student neurons with $\bar{v}$. The factor $1/\|w_i\|_2$ makes this aligning very fast near small initialization, while the norm barely changes.

Subphase 2: Radial growth toward the saddle. The radial growth that follows takes only a bounded time, over which the direction dynamics amplify the angular errors by at most a bounded factor, so for sufficiently small initialization the alignment from Subphase 1 persists. When $w_i = r_i\bar{v}$ exactly, the norm equations reduce to $\dot{r}_i = r_s - \frac{1}{2}(r_1+r_2)$, so the total radius r_1+r_2 is attracted to $r_+ = 2r_s$ while the imbalance $r_1 - r_2$ stays fixed. A small, balanced initialization therefore carries both norms nearly together to r_s, and the full trajectory follows this reduced dynamics up to small errors until it enters a neighborhood of w_{saddle}.

Combining the two subphases gives the following finite-time approach result.

Lemma 4.1 (Phase 1: finite-time approach to the bisector saddle)**.** *Fix $d \geq 2$ and unit teacher weights $v_1, v_2 \in \mathbb{R}^d$ with $\alpha := \angle(v_1, v_2) \in (0, \pi/2]$. Consider the two-student population gradient flow (2.2) under isotropic Gaussian inputs, and write $w(t) = (w_1(t), w_2(t))$. Fix an absolute constant $C \geq 1$ and an angular margin $\eta \in (0, \pi/2)$. For every target radius $\rho > 0$, there are $t_0 = t_0(\alpha, \rho) < \infty$ and $r_* = r_*(\alpha, C, \eta, \rho) > 0$ such that the following holds. Recall the initialization scale r_{init} from (2.3) and the balance condition (2.6). If $0 < r_{\text{init}} \leq r_*$, the initialization is C-balanced, and $\angle(\bar{w}_i(0), -\bar{v}) \geq \eta$ for $i = 1, 2$, then the gradient-flow solution exists on $[0, t_0]$, both student weights remain nonzero on this interval, and there is a time $t_{\text{in}} \leq t_0$ such that $\|w(t_{\text{in}}) - w_{\text{saddle}}\|_2 < \rho$.*

The proof is in Appendix E.3. If $w_1(0), w_2(0)$ are i.i.d. $\mathcal{N}(0, \sigma^2 I_d)$ with $\sigma \leq r_*/(2\sqrt{d})$, these conditions hold with high probability (Corollary E.8).

4.2 Phase 2: escape from the bisector saddle

Before turning to the local escape mechanism, we establish a qualitative global convergence result for Lebesgue-almost every initialization, with no restriction on its scale or structure.

Theorem 4.2 (Global convergence)**.** *Fix $d \geq 2$ and $\alpha = \angle(v_1, v_2) \in (0, \pi/2]$. For Lebesgue-almost every initialization $(w_1(0), w_2(0)) \in (\mathbb{R}^d)^2$, that is, for every initialization outside a set of Lebesgue measure zero in $(\mathbb{R}^d)^2$, the population gradient-flow trajectory (2.2) is defined for every $t \geq 0$, both student weights remain nonzero at every finite time, and $(w_1(t), w_2(t)) \to (v_1, v_2)$ or (v_2, v_1).*

The proof is in Appendix D.2. Note that while Theorem 4.2 establishes global convergence, it does not quantify the convergence rate or the escape behavior. The following subsections characterize how the trajectory escapes the saddle under additional assumptions.

4.2.1 Escape directions near the bisector saddle

Near w_{saddle}, Taylor expansion of gradient flow gives

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\big(w(t) - w_{\text{saddle}}\big) = -\nabla^2 L(w_{\text{saddle}})\big(w(t) - w_{\text{saddle}}\big) + o\big(\|w(t) - w_{\text{saddle}}\|_2\big). \tag{4.1}$$

Thus negative Hessian directions are pushed away from the saddle, positive directions shrink, and the kernel remains flat. Recall the landscape theorem Theorem 3.3 identifies the positive-bisector saddle family; here we focus on $w_{\text{saddle}} = r_s(\bar{v}, \bar{v})$. Complete the orthonormal pair $(\bar{v}, v_\Delta)$ from (2.4) to an orthonormal basis $\bar{v}, v_\Delta, u_3, \ldots, u_d$ of $\mathbb{R}^d$. The Hessian directions at this saddle are as follows.

Lemma 4.3 (Hessian eigenspaces at the bisector saddle)**.** *In the setting of Theorem 3.3, the Hessian of the loss at the bisector saddle w_{saddle} has the following eigenspace structure:*

- *The negative eigenspace of $\nabla^2 L(w_{\text{saddle}})$ is $\operatorname{span}\{(v_\Delta, -v_\Delta), (u_j, -u_j) : 3 \leq j \leq d\}$. Moreover, $(v_\Delta, -v_\Delta)$ has eigenvalue $-\frac{2\sin(\alpha/2)}{\pi r_s}$, while each $(u_j, -u_j)$ has eigenvalue $-\frac{\sin(\alpha/2)}{\pi r_s}$.*
- *The positive eigenspace of $\nabla^2 L(w_{\text{saddle}})$ is $\operatorname{span}\{(\bar{v}, \bar{v}), (v_\Delta, v_\Delta), (u_j, u_j) : 3 \leq j \leq d\}$.*
- *The kernel of $\nabla^2 L(w_{\text{saddle}})$ is $\operatorname{span}\{(\bar{v}, -\bar{v})\}$.*

The proof is in Appendix H.1. We observe that the planar antisymmetric mode $(v_\Delta, -v_\Delta)$ is the most unstable direction: its eigenvalue is twice as negative as the out-of-plane antisymmetric eigenvalues.

Thus, if a trajectory enters a sufficiently small saddle neighborhood with homogeneous coordinates in the Hessian eigenspace, the planar split would grow fastest along $(v_\Delta, -v_\Delta)$ and dominate the escape. This gives the geometric intuition for the observed in-plane splitting in Figure 1.

Write $w_-(t)$ for the projection of $w(t) - w_{\text{saddle}}$ onto the negative eigenspace in the lemma, and let ϵ be the *smallest value of* $\|w_-(t)\|_2$ *over the gradient flow trajectory*. Near the saddle, keeping only the linear term in the Taylor expansion (4.1) gives the approximation

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\|w_-(t)\|_2^2 \approx 2\langle w_-(t), -\nabla^2 L(w_{\text{saddle}})w_-(t)\rangle \asymp \|w_-(t)\|_2^2.$$

Thus, after $\|w_-(t)\|_2$ reaches its minimum, it will grow exponentially. This suggests that the escape time from the saddle is on the order of $\log(1/\epsilon)$ (see a rigorous version in Theorem 4.6). However, this calculation does not fully describe the escape path. In particular, it leaves two questions open:

1. *Which negative direction carries the escape?* The component $w_-(t)$ includes both the planar split $(v_\Delta, -v_\Delta)$ and the out-of-plane splits $(u_j, -u_j)$, $j \geq 3$.
2. *How large is* ϵ*?* Can this quantity be decided by properties of the initialization?

4.2.2 Exclusion of out-of-plane saddles

Recall the orthonormal basis $\{\bar{v}, v_\Delta, u_3, \ldots, u_d\}$. For each student weight $i \in \{1, 2\}$, we decompose $w_i(t) = z_{i,1}(t)\bar{v} + z_{i,2}(t)v_\Delta + \sum_{j=3}^d z_{i,j}(t)u_j$. In words, $z_{i,1}$ is the bisector coordinate, $z_{i,2}$ is the planar splitting coordinate, and $z_{i,j}$, $j \geq 3$, are the out-of-plane coordinates.

Assumption 4.4. Fix $C_\perp > 0$. The planar splitting coordinates have opposite signs, $z_{1,2}(0)z_{2,2}(0) < 0$, and every out-of-plane coordinate is dominated by the corresponding planar splitting coordinate: $|z_{i,j}(0)| \leq C_\perp |z_{i,2}(0)|$ for every $i \in \{1, 2\}$ and $3 \leq j \leq d$.

Notice that only the split inside the teacher plane receives an additive force from the teacher directions. The out-of-plane coordinates have no such source: each is multiplied by a scalar and coupled to the corresponding coordinate of the other student. Hence the conditions of Assumption 4.4 continue to hold along the flow, so the trajectory stays away from saddles with zero planar split and nonzero out-of-plane split. The theorem below makes this comparison precise. Since the loss is differentiable only where both student weights are nonzero, the solution is defined up to the first time a weight reaches zero; we write $[0, T_{\max})$ for this maximal interval, with $T_{\max} = \infty$ if no weight reaches zero. By Theorem 4.2, $T_{\max} = \infty$ for Lebesgue-almost every initialization.

Theorem 4.5 (Separation from out-of-plane saddles)**.** *Under Assumption 4.4,* $|z_{1,j}(t) - z_{2,j}(t)| \leq C_\perp |z_{1,2}(t) - z_{2,2}(t)|$ *for every* $j \geq 3$ *and* $0 \leq t < T_{\max}$*. Let* $\mathcal{S}_\perp$ *denote the out-of-plane saddle family in Theorem 3.3. Then there exists a constant* $c > 0$ *(depending on* α*,* $C_\perp$*, and* d*) such that* $\|w(t) - \widehat{w}\|_2 \geq c$ *for every* $\widehat{w} \in \mathcal{S}_\perp$ *and* $0 \leq t < T_{\max}$*.*

This gives an alternative proof of global convergence for initializations satisfying Assumption 4.4, which Gaussian initialization satisfies for some $C_\perp$ with probability $1/2$ (Appendix F).

4.2.3 Logarithmic escape under structured initialization

The results above are qualitative and do not quantify how long symmetry breaking takes. We now study this time scale under a structured initialization condition, which holds with constant probability at small random initialization. We use the coordinates $z_{i,j}$ of Section 4.2.2 and write $\mu_\alpha := 2\sin(\alpha/2)/(\pi r_s)$ for the planar escape rate from Lemma 4.3. For $\varepsilon > 0$, let

$$T_\varepsilon := \inf\left\{t \geq 0 : \sum\nolimits_{i=1}^{2}\sum\nolimits_{j=2}^{d} z_{i,j}(t)^2 \geq \varepsilon^2\right\}$$

be the first time the total squared distance of the student weights from the bisector axis reaches ε^2. The next theorem shows that T_ε is governed by the initial planar splitting and, up to a bounded error, depends logarithmically on a_{init} and r_{init}. The proof is in Appendix H.2.

Theorem 4.6. *Fix* $\alpha \in (0, \pi/2]$ *and an absolute constant* $C > 1$*. For every sufficiently small* $\varepsilon > 0$*, there are constants* $c \in (0, 1)$ *and* $K < \infty$*, depending only on* α, C, ε*, such that the following holds for every* $d \geq 2$*. Consider the gradient flow in (2.2) from a* C*-balanced initialization (see (2.6)) satisfying* $r_{\text{init}}\sqrt{d} \leq c$*,* $r_{\text{init}}\sqrt{d}\left(1 + \log(\sqrt{d}/r_{\text{init}}) + \log d\right) \leq c$ *and* $a_{\text{init}} \geq r_{\text{init}}/(C\sqrt{d})$*, with* $z_{1,1}(0), z_{2,1}(0) > 0$ *and* $z_{1,2}(0)z_{2,2}(0) < 0$*. Then* T_ε *is finite and satisfies*

$$\left|T_\varepsilon - \tfrac{1}{\mu_\alpha}\log\tfrac{r_{\text{init}}^{\mu_\alpha}}{a_{\text{init}}}\right| \leq K.$$

Remark 4.7 (Dimension dependence and i.i.d. initialization). When the initial planar splitting is of the size of a single coordinate, that is, $a_{\text{init}} \asymp r_{\text{init}}/\sqrt{d}$, the theorem gives $T_\varepsilon = \left(\mu_\alpha^{-1} - 1\right) \log \frac{1}{r_{\text{init}}} + \frac{1}{2\mu_\alpha} \log d + O(1)$. Independent Gaussian initialization $w_i(0) \sim \mathcal{N}(0, \sigma^2 I_d)$, $i = 1, 2$, is a natural example (Appendix H.3). With constant probability, it satisfies the other conditions of Theorem 4.6, namely C-balance, $a_{\text{init}} \geq r_{\text{init}}/(C\sqrt{d})$ and the sign conditions, with $r_{\text{init}} \asymp \sigma\sqrt{d}$ and $a_{\text{init}} \asymp \sigma \asymp r_{\text{init}}/\sqrt{d}$. On this event, for $0 < \sigma \leq 1$, the two upper bounds on r_{init} hold once $\sigma d\big(1 + \log(1/\sigma) + \log d\big)$ is sufficiently small, and the theorem gives $T_\varepsilon = \left(\mu_\alpha^{-1} - 1\right) \log \frac{1}{\sigma} + \frac{1}{2} \log d + O_{\alpha,C,\varepsilon}(1)$. If $\sigma = d^{-\beta}$ for a fixed $\beta > 1$, these bounds hold for all sufficiently large d, and $T_\varepsilon = \Theta(\log d)$. This logarithmic dependence on d is consistent with the "search phase" in learning functions of information or leap exponent 2 (Chen et al., 2019; Ben Arous et al., 2021; Montanari and Wang, 2026).

Empirical validation. Figure 2 supports the $\frac{1}{2} \log d$ growth of T_ε that the Gaussian formula above predicts at fixed σ, and shows that the ambient dimension mainly affects the time needed to break the bisector symmetry. Further experiments in Appendix I show the same collapse and split under SGD on a finite sample, in wider networks, and with GELU in place of ReLU. From larger initialization, the students can skip the collapse, but they still reach a global minimum. Networks trained on MNIST and CIFAR-10 show a similar collapse and recovery of the hidden-layer weight directions.

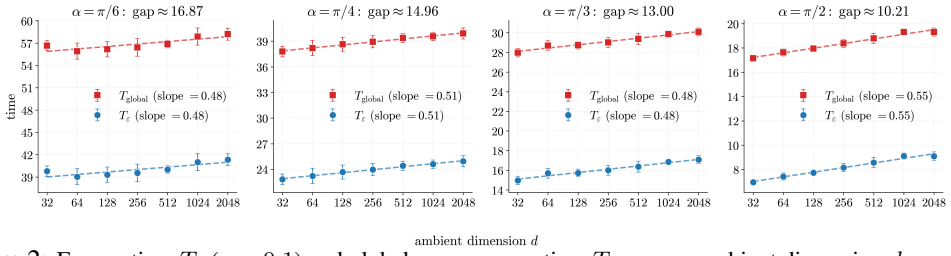


Figure 2: Escape time T_ε ($\varepsilon = 0.1$) and global-convergence time T_{global} vs. ambient dimension d, one panel per α, for random initialization at fixed $\sigma = 10^{-4}$; mean ± 1 s.d. over 5 seeds. T_{global} is when the state first reaches distance 0.05 from the global minimizer set. Dashed lines are log-d fits with slopes in $[0.48, 0.55]$; the gap $T_{\text{global}} - T_\varepsilon$ (panel titles) is almost d-independent.

5 Conclusion

We studied population gradient flow for a two-neuron ReLU teacher-student model. For the loss landscape, we show that the loss has no spurious local minima with both student weights nonzero, but it has a bisector saddle. For the dynamics, we show that, from small initialization, gradient flow first learns the easy bisector direction and approaches this saddle, then escapes along the harder splitting direction in time logarithmic in d. Table 1 in the appendix summarizes our main results.

Scope and limitations. Our proofs assume two neurons, isotropic Gaussian inputs, and population gradient flow, but some results hold more generally. The correlation-loss result (Theorem 3.1) holds for a teacher with k neurons at equal pairwise angles and for a student of any width. Wider ReLU networks can have spurious local minima even when the student and the teacher have the same width, the inputs are Gaussian, and the teacher weights are orthonormal (Safran and Shamir, 2018). The absence of spurious minima and the convergence from almost every initialization in Theorem 4.2 are therefore specific to two neurons, and in our experiments at matched widths 16 and 32, gradient flow does not reach a global minimum (Appendix I.5). With three or more teacher neurons, the dimension reduction of Proposition 3.5 also fails, because the components of the student weights orthogonal to the span of the teacher weights no longer need to point in a common direction.

The bisector saddle and the leap-2 escape time are specific to our model, but a similar collapse followed by a recovery appears in more general settings. In deep matrix factorization, convolutional networks, and transformer pretraining, the effective rank of the weight matrices has been observed to drop early in training and then rise again (Chou et al., 2024; Huh et al., 2023; Shivagunde et al., 2026). We see a similar pattern in networks trained on MNIST and CIFAR-10 (Appendix I.8).

Open directions include (i) obtaining an end-to-end escape-time analysis from random initialization with high probability, and (ii) extending the landscape and dynamical characterization to other teacher-student settings, larger networks, non-Gaussian inputs, and high-leap multi-index functions.

Acknowledgments and Disclosure of Funding

The authors thank Margalit Glasgow, Jason D. Lee, and Theodor Misiakiewicz for discussions and feedback. TW acknowledges support from Google. The authors declare no competing interests.

References

ABBE, E., BOIX-ADSERÀ, E. and MISIAKIEWICZ, T. (2022). The merged-staircase property: a necessary and nearly sufficient condition for SGD learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*. PMLR.

ABBE, E., BOIX-ADSERÀ, E. and MISIAKIEWICZ, T. (2023). SGD learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *The Thirty Sixth Annual Conference on Learning Theory*. PMLR.

AKIYAMA, S. and SUZUKI, T. (2021). On learnability via gradient method for two-layer ReLU neural networks in teacher-student setting. In *International Conference on Machine Learning*.

BA, J., ERDOGDU, M. A., SUZUKI, T., WANG, Z., WU, D. and YANG, G. (2022). High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *Advances in Neural Information Processing Systems* **35** 37932–37946.

BAH, B., RAUHUT, H., TERSTIEGE, U. and WESTDICKENBERG, M. (2022). Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers. *Information and Inference: A Journal of the IMA* **11** 307–353.

BEN AROUS, G., ERDOGDU, M. A., VURAL, N. M. and WU, D. (2025). Learning quadratic neural networks in high dimensions: SGD dynamics and scaling laws. *Advances in Neural Information Processing Systems* **38** 162874–162961.

BEN AROUS, G., GERBELOT, C. and PICCOLO, V. (2026). Stochastic gradient descent in high dimensions for multi-spiked tensor PCA. *Communications on Pure and Applied Mathematics* **79** 2291–2369.

BEN AROUS, G., GHEISSARI, R. and JAGANNATH, A. (2021). Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research* **22** 1–51.

BIEHL, M., RIEGLER, P. and WÖHLER, C. (1996). Transient dynamics of on-line learning in two-layered neural networks. *Journal of Physics A: Mathematical and General* **29** 4769–4780.

BIETTI, A., BRUNA, J. and PILLAUD-VIVIEN, L. (2025). On learning Gaussian multi-index models with gradient flow part I: General properties and two-timescale learning. *Communications on Pure and Applied Mathematics* **78** 2354–2435.

BIETTI, A., BRUNA, J., SANFORD, C. and SONG, M. J. (2022). Learning single-index models with shallow neural networks. *Advances in neural information processing systems* **35** 9768–9783.

BOLTE, J., DANIILIDIS, A. and LEWIS, A. (2007). The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization* **17** 1205–1223.

BOURSIER, E., PILLAUD-VIVIEN, L. and FLAMMARION, N. (2022). Gradient flow dynamics of shallow ReLU networks for square loss and orthogonal inputs. *Advances in Neural Information Processing Systems* **35** 20105–20118.

CHEN, Y., CHI, Y., FAN, J. and MA, C. (2019). Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming* **176** 5–37.

CHIZAT, L. (2022). Sparse optimization on measures with over-parameterized gradient descent. *Mathematical Programming* **194** 487–532.

CHIZAT, L., OYALLON, E. and BACH, F. (2019). On lazy training in differentiable programming. *Advances in neural information processing systems* **32**.

CHOU, H.-H., GIESHOFF, C., MALY, J. and RAUHUT, H. (2024). Gradient descent for deep matrix factorization: Dynamics and implicit bias towards low rank. *Applied and Computational Harmonic Analysis* **68** 101595.

DAMIAN, A., LEE, J. and SOLTANOLKOTABI, M. (2022). Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*. PMLR.

DAMIAN, A., NICHANI, E., GE, R. and LEE, J. D. (2023). Smoothing the landscape boosts the signal for SGD: Optimal sample complexity for learning single index models. *Advances in Neural Information Processing Systems* **36** 752–784.

DANDI, Y., KRZAKALA, F., LOUREIRO, B., PESCE, L. and STEPHAN, L. (2024). How two-layer neural networks learn, one (giant) step at a time. *Journal of Machine Learning Research* **25** 1–65.

DUDEJA, R. and HSU, D. (2018). Learning single-index models in Gaussian space. In *Conference On Learning Theory*. PMLR.

FUKUMIZU, K. and AMARI, S.-I. (2000). Local minima and plateaus in hierarchical structures of multilayer perceptrons. *Neural networks* **13** 317–327.

GALANTI, T., SIEGEL, Z. S., GUPTE, A. and POGGIO, T. A. (2025). SGD with weight decay secretly minimizes the ranks of your neural networks. In *Conference on Parsimony and Learning*, vol. 280 of *Proceedings of Machine Learning Research*.

GE, R., REN, Y., WANG, X. and ZHOU, M. (2021). Understanding deflation process in over-parametrized tensor decomposition. *Advances in Neural Information Processing Systems* **34** 1299–1311.

GLASGOW, M. (2024). SGD finds then tunes features in two-layer neural networks with near-optimal sample complexity: A case study in the XOR problem. In *The Twelfth International Conference on Learning Representations*.

GOLDT, S., ADVANI, M. S., SAXE, A. M., KRZAKALA, F. and ZDEBOROVÁ, L. (2020). Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. *Journal of Statistical Mechanics: Theory and Experiment* 124010.

HE, K., ZHANG, X., REN, S. and SUN, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE International Conference on Computer Vision*.

HENDRYCKS, D. and GIMPEL, K. (2016). Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415* .

HIRSCH, M. W., PUGH, C. C. and SHUB, M. (1977). *Invariant Manifolds*, vol. 583 of *Lecture Notes in Mathematics*. Springer.

HUANG, J., LOUREIRO, B. and SARAO MANNELLI, S. (2026). Sharp description of local minima in the loss landscape of high-dimensional two-layer ReLU neural networks. In *Proceedings of the 43rd International Conference on Machine Learning*.

HUH, M., MOBAHI, H., ZHANG, R., CHEUNG, B., AGRAWAL, P. and ISOLA, P. (2023). The low-rank simplicity bias in deep networks. *Transactions on Machine Learning Research* .

KRIZHEVSKY, A. (2009). Learning multiple layers of features from tiny images. Tech. rep., University of Toronto.

LECUN, Y., BOTTOU, L., BENGIO, Y. and HAFFNER, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86** 2278–2324.

LEE, J. D., OKO, K., SUZUKI, T. and WU, D. (2024). Neural network learns low-dimensional polynomials with SGD near the information-theoretic limit. *Advances in Neural Information Processing Systems* **37** 58716–58756.

LI, Y., MA, T. and ZHANG, H. R. (2020). Learning over-parametrized two-layer neural networks beyond NTK. In *Conference on learning theory*. PMLR.

MARTIN, S., BACH, F. and BIROLI, G. (2024). On the impact of overparameterization on the training of a shallow neural network in high dimensions. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

MIN, H., ZHU, Z. and VIDAL, R. (2025). Neural collapse under gradient flow on shallow ReLU networks for orthogonally separable data. *Advances in Neural Information Processing Systems* **38** 105551–105590.

MONTANARI, A. and WANG, Z. (2026). Phase transitions for feature learning in neural networks. *arXiv preprint arXiv:2602.01434* .

OKO, K., SONG, Y., SUZUKI, T. and WU, D. (2024). Learning sum of diverse features: computational hardness and efficient gradient-based training for ridge combinations. In *The Thirty Seventh Annual Conference on Learning Theory*. PMLR.

OOSTWAL, E., STRAAT, M. and BIEHL, M. (2021). Hidden unit specialization in layered neural networks: ReLU vs. sigmoidal activation. *Physica A: Statistical Mechanics and its Applications* **564** 125517.

REN, Y. and LEE, J. D. (2025). Learning orthogonal multi-index models: A fine-grained information exponent analysis. *Advances in Neural Information Processing Systems* **38** 183843–183894.

REN, Y., NICHANI, E., WU, D. and LEE, J. D. (2025). Emergence and scaling laws in SGD learning of shallow neural networks. *Advances in Neural Information Processing Systems* **38** 42829–42911.

RIEGLER, P. and BIEHL, M. (1995). On-line backpropagation in two-layered neural networks. *Journal of Physics A: Mathematical and General* **28** L507–L513.

SAAD, D. and SOLLA, S. A. (1995). On-line learning in soft committee machines. *Physical Review E* **52** 4225–4243.

SAFRAN, I. and SHAMIR, O. (2018). Spurious local minima are common in two-layer ReLU neural networks. In *International conference on machine learning*. PMLR.

SAFRAN, I. M., YEHUDAI, G. and SHAMIR, O. (2021). The effects of mild over-parameterization on the optimization landscape of shallow ReLU neural networks. In *Conference on Learning Theory*. PMLR.

SHIVAGUNDE, N., DESHPANDE, V., MUCKATIRA, S. and RUMSHISKY, A. (2026). Beyond perplexity: A geometric and spectral study of low-rank pre-training. *arXiv preprint arXiv:2605.13652* .

ŞIMŞEK, B., BENDJEDDOU, A. and HSU, D. (2025). Learning Gaussian multi-index models with gradient flow: Time complexity and directional convergence. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, vol. 258 of *Proceedings of Machine Learning Research*. PMLR.

SOLTANOLKOTABI, M. (2017). Learning ReLUs via gradient descent. *Advances in neural information processing systems* **30**.

STÖGER, D. and SOLTANOLKOTABI, M. (2021). Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. In *Advances in Neural Information Processing Systems*, vol. 34.

SUMNER, J. S., JAGERS, A. A., VOWE, M. and ANGLESIO, J. (1991). Inequalities involving trigonometric functions. *The American Mathematical Monthly* **98** 264–267.

TIAN, Y. (2017). An analytical formula of population gradient for two-layered ReLU network and its applications in convergence and critical point analysis. In *International Conference on Machine Learning*.

WANG, M. and MA, C. (2023). Understanding multi-phase optimization dynamics and rich nonlinear behaviors of ReLU networks. *Advances in Neural Information Processing Systems* **36** 35654–35747.

WU, C., LUO, J. and LEE, J. D. (2018). No spurious local minima in a two hidden unit ReLU network. In *International Conference on Learning Representations, Workshop Track*.

XU, W. and DU, S. (2023). Over-parameterization exponentially slows down gradient descent for learning a single neuron. In *The Thirty Sixth Annual Conference on Learning Theory*. PMLR.

YEHUDAI, G. and SHAMIR, O. (2020). Learning a single neuron with gradient methods. In *Conference on Learning Theory*. PMLR.

YUNIS, D., PATEL, K. K., WHEELER, S., SAVARESE, P., VARDI, G., LIVESCU, K., MAIRE, M. and WALTER, M. R. (2024). Approaching deep learning through the spectral dynamics of weights. *arXiv preprint arXiv:2408.11804* .

ZHANG, B., WANG, Z., FU, H. and LEE, J. D. (2026). Neural networks learn generic multi-index models near information-theoretic limit. In *The Fourteenth International Conference on Learning Representations*.

ZHOU, M., GE, R. and JIN, C. (2021). A local convergence theory for mildly over-parameterized two-layer neural network. In *Conference on Learning Theory*. PMLR.

ZHU, Z., LIU, F. and CEVHER, V. (2025). How gradient descent balances features: A dynamical analysis for two-layer neural networks. In *The Thirteenth International Conference on Learning Representations*.

Contents

1 Introduction **1**

1.1 Our contributions . . . 2

1.2 Further related work . . . 3

2 Preliminary **3**

2.1 Problem setup . . . 3

2.2 Leap complexity of the teacher . . . 4

3 Landscape analysis **4**

3.1 Warm-up: landscape of the correlation loss . . . 5

3.2 Landscape of the squared loss . . . 5

3.2.1 Step 1: reduce to three dimensions . . . 6

3.2.2 Step 2: reduce the nonplanar case to two dimensions . . . 7

3.2.3 Step 3: solve the 2D cases . . . 7

4 Training dynamics **7**

4.1 Phase 1: alignment with the bisector . . . 7

4.2 Phase 2: escape from the bisector saddle . . . 8

4.2.1 Escape directions near the bisector saddle . . . 8

4.2.2 Exclusion of out-of-plane saddles . . . 9

4.2.3 Logarithmic escape under structured initialization . . . 9

5 Conclusion **10**

A Technical lemmas **18**

A.1 Gaussian ReLU kernel . . . 18

A.2 Closed-form loss and gradient flow . . . 20

A.3 Hessian of the population loss . . . 22

A.4 Proof of Proposition 2.1 . . . 22

B Correlation-loss landscape **23**

B.1 Reduction to one student in $\mathrm{span}\{v_1, \dots, v_k\}$ . . . 23

B.2 The case $\alpha \le \pi/2$ . . . 24

B.3 The case $k = 2$ and $\alpha > \pi/2$ . . . 27

B.4 Proof of Theorem 3.1 . . . 28

C Squared-loss landscape **28**

C.1 Stationarity and dimension reduction . . . 28

C.1.1 Stationary equations . . . 29

C.1.2 Uniqueness of the out-of-plane root . . . 29

C.1.3 Reduction to three dimensions . . . 30

C.2 Three-dimensional symmetry reduction . . . 30
C.2.1 The teacher force . . . 30
C.2.2 Elimination of the splitting coordinate . . . 32
C.3 Out-of-plane stationary points . . . 32
C.3.1 A scalar inequality . . . 32
C.3.2 Mirror symmetry . . . 35
C.3.3 Exact scalar characterization . . . 37
C.3.4 Negative curvature . . . 38
C.4 In-plane stationary points . . . 38
C.4.1 Coordinates in the teacher plane . . . 38
C.4.2 Stationary points above the angle threshold . . . 40
C.4.3 Classification up to the angle threshold . . . 40
C.5 Completion of the squared-loss landscape theorem . . . 44

D Global convergence and boundary avoidance **45**
D.1 Almost-everywhere avoidance of zero weights . . . 45
D.1.1 Polar coordinates near $\{w_1 = 0\}$ . . . 46
D.1.2 Equilibria of X with $r_1 = 0$ . . . 47
D.1.3 The equilibria $(0, -v_1, v_2)$ and $(0, -v_2, v_1)$ . . . 49
D.1.4 Avoidance of $\partial\Omega$ . . . 53
D.2 Saddle avoidance and global convergence . . . 55

E Approach to the bisector saddle **55**
E.1 Norm bounds . . . 55
E.2 Alignment with the bisector and balance of the norms . . . 56
E.3 Proofs of Lemma 4.1 and Corollary E.8 . . . 59

F Exclusion of saddles and global convergence **60**

G Approach to the bisector saddle under structured initialization **62**
G.1 Coordinates and estimates near the bisector . . . 62
G.2 Alignment with the bisector . . . 66
G.3 Growth of the bisector coordinates up to $r_s/2$ . . . 67
G.4 Entry into a neighborhood of the bisector saddle . . . 68

H Logarithmic escape from the bisector saddle **70**
H.1 Hessian at the bisector saddle . . . 70
H.2 Proof of Theorem 4.6 . . . 71
H.3 Gaussian initialization . . . 73

I Additional experiments **75**
I.1 Common setup and measured quantities . . . 75

Table 1: Summary of the main results, their assumptions, and their conclusions. Here $\alpha_\star \approx 127.5^\circ > \pi/2$ is the angle threshold defined in Section 3.1. Unless stated otherwise, the teacher has two unit neurons at angle α, inputs are isotropic Gaussian, and training is population gradient flow.

Result	Assumptions	Conclusion
Theorem 3.1 (correlation loss)	Equiangular width-k teacher with $\alpha \le \pi/2$, or $k = 2$ with $\alpha \le \alpha_\star$ (obtuse angles included), and any student width	$(\bar{v}, \dots, \bar{v})$ is the unique minimizer over unit-norm student weights
Theorems 3.3 and C.8 (squared loss)	$\alpha \le \alpha_\star$ (obtuse angles included), both student weights nonzero	Every stationary point is a global minimum or a strict saddle, so there are no spurious local minima
Lemma 4.1 and Corollary E.8 (Phase 1)	$\alpha \le \pi/2$, small C-balanced initialization at angle at least η from $-\bar{v}$ (probability $\ge 1 - 6e^{-cd}$ for Gaussian initialization)	The trajectory enters any prescribed neighborhood of w_{saddle} within a time independent of d
Theorem 4.2 (global convergence)	$\alpha \le \pi/2$, Lebesgue-almost every initialization of any scale	Convergence to (v_1, v_2) or (v_2, v_1)
Theorem 4.5 (saddle exclusion)	Opposite-sign planar splits, out-of-plane coordinates dominated by them (probability $1/2$ for Gaussian initialization)	The trajectory stays a positive distance from the out-of-plane saddles
Theorem 4.6 (escape time)	$\alpha \le \pi/2$, small C-balanced initialization with $a_{\mathrm{init}} \ge r_{\mathrm{init}}/(C\sqrt{d})$ and sign conditions (constant probability for Gaussian initialization)	$T_\varepsilon = \mu_\alpha^{-1} \log(r_{\mathrm{init}}^{\mu_\alpha}/a_{\mathrm{init}}) + O(1)$, which is $\Theta(\log d)$ for $a_{\mathrm{init}} \asymp r_{\mathrm{init}}/\sqrt{d}$ and $\log(1/r_{\mathrm{init}}) = O(\log d)$

I.2 Escape time and dimension 76
I.3 Finite-sample SGD 76
I.4 Initialization scale 77
I.5 Wider networks 77
I.6 GELU activations 78
I.7 Numerical search for stationary points 79
I.8 Networks trained on real data 79

Summary of results. Table 1 collects the main results of the paper together with their assumptions and conclusions.

A Technical lemmas

This appendix collects the lemmas used in the landscape and training-dynamics proofs. Throughout this appendix, unless stated otherwise, $v_1, v_2 \in \mathbb{R}^d$ are unit teacher directions with $\alpha = \angle(v_1, v_2) \in (0, \pi/2]$, and the angle notation θ_{ij} and γ follows the convention in (2.7). Whenever directions or angles of student weights are used, the corresponding weights are assumed to be nonzero.

A.1 Gaussian ReLU kernel

We first recall the Gaussian ReLU kernel identity from Safran and Shamir (2018). For nonzero $u, v \in \mathbb{R}^d$, let $\theta_{u,v} := \angle(u, v)$. Then

$$\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\big[\phi(\langle u, x\rangle)\phi(\langle v, x\rangle)\big] = \|u\|_2\|v\|_2\, l(\theta_{u,v}), \tag{A.1a}$$

$$l(\theta) := \frac{(\pi-\theta)\cos\theta + \sin\theta}{2\pi}, \qquad \theta \in [0, \pi]. \tag{A.1b}$$

Lemma A.1 (Regularity of L and continuation of gradient flow)**.** *The Gaussian ReLU kernel in (A.1a) is jointly C^2 in its two nonzero vector arguments, including at aligned and antipodal directions. Consequently, the population loss L is C^2 on the open set where both student weights are nonzero. On this set, gradient flow has a unique local solution, and any solution that stays in a compact subset of this set on $[0, T)$, with $T < \infty$, extends beyond T.*

Proof of Lemma A.1. Write

$$\mathcal{U} = (\mathbb{R}^d \setminus \{0\})^2, \qquad K(u, v) = \|u\|_2\|v\|_2\, l(\angle(u, v)).$$

For $(u, v) \in \mathcal{U}$, write

$$c = \langle \bar{u}, \bar{v}\rangle, \quad z = \bar{u} - c\bar{v}, \quad \tau = \|z\|_2 = \sqrt{1-c^2}.$$

When $|c| < 1$, $K(u, v) = \|u\|_2\|v\|_2\, l(\arccos c)$ is smooth. Near $c = \pm 1$, the sign of c is fixed and

$$\theta = \angle(u, v) = \begin{cases} \arcsin\tau, & c > 0, \\ \pi - \arcsin\tau, & c < 0, \end{cases}$$

$$l(\theta) = \begin{cases} c/2 + F(\tau), & c > 0, \\ F(\tau), & c < 0, \end{cases} \qquad F(\tau) = \frac{\tau - \sqrt{1-\tau^2}\arcsin\tau}{2\pi}.$$

The function F is odd and analytic near zero, with

$$\arcsin\tau = \tau + \frac{\tau^3}{6} + \frac{3\tau^5}{40} + O(\tau^7),$$

$$\sqrt{1-\tau^2} = 1 - \frac{\tau^2}{2} - \frac{\tau^4}{8} + O(\tau^6),$$

$$F(\tau) = \frac{\tau^3}{6\pi} + \frac{\tau^5}{15\pi} + O(\tau^7) = \tau^3\psi(\tau^2),$$

where ψ is analytic near zero and $\psi(0) = 1/(6\pi)$.

For $H(z) = \|z\|_2^3$ and $z \neq 0$,

$$\nabla H(z) = 3\|z\|_2 z,$$

$$\nabla^2 H(z) = 3\|z\|_2 I_d + \frac{3zz^\top}{\|z\|_2}, \quad \|\nabla^2 H(z)\|_{\mathrm{op}} \le 6\|z\|_2.$$

At zero,

$$\frac{|H(z) - H(0)|}{\|z\|_2} = \|z\|_2^2 \longrightarrow 0, \qquad \frac{\|\nabla H(z)\|_2}{\|z\|_2} = 3\|z\|_2 \longrightarrow 0.$$

Thus $\nabla H(0) = 0$ and $\nabla^2 H(0) = 0$, and $\nabla^2 H(z) \to \nabla^2 H(0)$ as $z \to 0$. Hence $H \in C^2(\mathbb{R}^d)$, and $F(\|z\|_2) = H(z)\psi(\|z\|_2^2)$ is C^2 near 0. Since c, z, $\|u\|_2$, and $\|v\|_2$ depend smoothly on $(u, v) \in \mathcal{U}$, the two endpoint formulas give $K \in C^2(\mathcal{U})$. Since

$$L(w_1, w_2) = \frac{1}{2}\sum_{i,j=1}^{2} K(w_i, w_j) - \sum_{i,j=1}^{2} K(w_i, v_j) + \frac{1}{2}\sum_{i,j=1}^{2} K(v_i, v_j),$$

it follows that $L \in C^2(\mathcal{U})$ and $-\nabla L \in C^1(\mathcal{U})$. Local existence and uniqueness follow from the Picard–Lindelöf theorem. For continuation, suppose that $w([0,T)) \subset \mathcal{Q}$ for some $T < \infty$ and some compact set $\mathcal{Q} \subset \mathcal{U}$, and let $B_{\mathcal{Q}} = \max_{w \in \mathcal{Q}} \|\nabla L(w)\|_2 < \infty$. For $0 \le s < t < T$,

$$\|w(t) - w(s)\|_2 \le \int_s^t \|\nabla L(w(a))\|_2 \, \mathrm{d}a \le B_{\mathcal{Q}}(t-s).$$

Therefore the limit $w_T = \lim_{t \uparrow T} w(t)$ exists and lies in $\mathcal{Q} \subset \mathcal{U}$. The local solution starting at w_T extends w beyond T. ■

Lemma A.2 (Monotonicity and bounds for l)**.** *The function l defined in (A.1b) is strictly decreasing on $[0,\pi]$ and satisfies $0 \le l(\theta) \le 1/2$ for every $\theta \in [0,\pi]$.*

Proof of Lemma A.2. Differentiating gives

$$l'(\theta) = -\frac{(\pi - \theta)\sin\theta}{2\pi}, \tag{A.2}$$

which is negative for $\theta \in (0,\pi)$. Hence l is strictly decreasing on $[0,\pi]$. The endpoint values $l(0) = 1/2$ and $l(\pi) = 0$ give the claimed bounds. ■

Lemma A.3 (Bounds on $l(\theta_1)+l(\theta_2)$)**.** *Let $u \in \mathbb{R}^d$ be nonzero and write $\theta_j := \angle(u, v_j)$ for $j = 1, 2$. Then*

$$2l\left(\pi - \frac{\alpha}{2}\right) \le l(\theta_1) + l(\theta_2) \le 2l\left(\frac{\alpha}{2}\right).$$

Moreover, equality in the upper bound holds if and only if $\theta_1 = \theta_2 = \alpha/2$.

Proof of Lemma A.3. We begin with the lower bound. Set

$$\vartheta_j := \pi - \theta_j = \angle(-\bar{u}, v_j), \qquad j = 1, 2.$$

The triangle inequality for angles gives

$$\vartheta_1 + \vartheta_2 \ge \alpha.$$

Write

$$q(x) := \sin x - x\cos x, \qquad l(\pi - x) = \frac{q(x)}{2\pi}, \qquad x \in [0,\pi].$$

Since $q'(x) = x\sin x \ge 0$, the function q is increasing. Let

$$a := \min\{\vartheta_1, \alpha\}, \qquad b := \alpha - a.$$

Then $a, b \in [0,\alpha]$, $a + b = \alpha$, $a \le \vartheta_1$, and $b \le \vartheta_2$. Indeed, if $\vartheta_1 < \alpha$, then $b = \alpha - \vartheta_1 \le \vartheta_2$; otherwise $b = 0$. Therefore,

$$q(\vartheta_1) + q(\vartheta_2) \ge q(a) + q(b). \tag{A.3}$$

Moreover,

$$q''(x) = \sin x + x\cos x \ge 0, \qquad x \in [0,\alpha],$$

because $\alpha \le \pi/2$. Thus q is convex on $[0,\alpha]$, and Jensen's inequality yields

$$q(a) + q(b) \ge 2q\left(\frac{a+b}{2}\right) = 2q\left(\frac{\alpha}{2}\right). \tag{A.4}$$

Combining (A.3) and (A.4) gives

$$l(\theta_1) + l(\theta_2) = l(\pi - \vartheta_1) + l(\pi - \vartheta_2) \ge 2l\left(\pi - \frac{\alpha}{2}\right). \tag{A.5}$$

We now prove the upper bound and identify its equality case.

Case 1: $\theta_1, \theta_2 \in [0, \alpha]$.

Define l_1 on $[0, \alpha]$ by

$$l_1(x) \coloneqq l(x) + l(\alpha - x).$$

The triangle inequality gives $\theta_2 \geq \alpha - \theta_1$. Since l is decreasing,

$$l(\theta_1) + l(\theta_2) \leq l(\theta_1) + l(\alpha - \theta_1) = l_1(\theta_1). \tag{A.6}$$

Differentiating (A.2) gives

$$\begin{aligned} 2\pi l_1''(x) &= \sin x - (\pi - x)\cos x + \sin(\alpha - x) - (\pi - \alpha + x)\cos(\alpha - x) \\ &= 2\left(\sin\frac{\alpha}{2} - \left(\pi - \frac{\alpha}{2}\right)\cos\frac{\alpha}{2}\right)\cos\left(x - \frac{\alpha}{2}\right) - 2\left(x - \frac{\alpha}{2}\right)\sin\frac{\alpha}{2}\sin\left(x - \frac{\alpha}{2}\right). \end{aligned} \tag{A.7}$$

Because $0 < \alpha/2 \leq \pi/4$,

$$\tan\frac{\alpha}{2} \leq 1 < \frac{3\pi}{4} \leq \pi - \frac{\alpha}{2}.$$

Hence

$$\sin\frac{\alpha}{2} - \left(\pi - \frac{\alpha}{2}\right)\cos\frac{\alpha}{2} < 0.$$

For $x \in [0, \alpha]$, $|x - \alpha/2| \leq \alpha/2 \leq \pi/4$, so

$$\cos\left(x - \frac{\alpha}{2}\right) > 0, \qquad \left(x - \frac{\alpha}{2}\right)\sin\left(x - \frac{\alpha}{2}\right) \geq 0.$$

It follows from (A.7) that $l_1''(x) < 0$ on $[0, \alpha]$.

The symmetry $l_1(x) = l_1(\alpha - x)$ and strict concavity show that l_1 has a unique maximum at $x = \alpha/2$. Therefore,

$$l_1(x) \leq l_1\left(\frac{\alpha}{2}\right) = 2l\left(\frac{\alpha}{2}\right), \qquad x \in [0, \alpha], \tag{A.8}$$

with equality only at $x = \alpha/2$. Combining (A.6) and (A.8) yields

$$l(\theta_1) + l(\theta_2) \leq 2l\left(\frac{\alpha}{2}\right).$$

If equality holds, (A.8) forces $\theta_1 = \alpha/2$. Equality in (A.6), together with the strict monotonicity of l, then forces $\theta_2 = \alpha - \theta_1 = \alpha/2$. Conversely, $\theta_1 = \theta_2 = \alpha/2$ attains equality.

Case 2: Exactly one of θ_1, θ_2 lies in $(\alpha, \pi]$. By symmetry, assume $\theta_1 \in (\alpha, \pi]$ and $\theta_2 \in [0, \alpha]$.

The reverse triangle inequality gives $\theta_1 - \theta_2 \leq \alpha$, so $\theta_2 \geq \theta_1 - \alpha$. Monotonicity of l then gives

$$l(\theta_1) + l(\theta_2) \leq l(\theta_1) + l(\theta_1 - \alpha) \leq l(\alpha) + l(0) \leq 2l\left(\frac{\alpha}{2}\right).$$

The last inequality is (A.8) evaluated at $x = 0$. Since $\alpha/2 > 0$, the equality condition in (A.8) shows that this inequality is strict.

Case 3: $\theta_1, \theta_2 \in (\alpha, \pi]$. Monotonicity gives

$$l(\theta_1) + l(\theta_2) \leq 2l(\alpha) \leq 2l\left(\frac{\alpha}{2}\right).$$

The second inequality is strict because l is strictly decreasing and $\alpha > 0$.

Thus the upper bound holds in all three cases, with equality if and only if $\theta_1 = \theta_2 = \alpha/2$. Together with the lower bound (A.5), this completes the proof. ■

A.2 Closed-form loss and gradient flow

The identities in this subsection hold for the wider range $\alpha = \angle(v_1, v_2) \in (0, \pi)$. We recall the angles θ_{ij} and γ from (2.7), the kernel profile l from (A.1b), and the shorthand $\theta_{u,v} = \angle(u, v)$ from Appendix A.1. For nonzero $u, v \in \mathbb{R}^d$, the derivative of the Gaussian ReLU kernel with respect to its first argument is

$$\nabla_u \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[\phi(\langle u, x\rangle)\phi(\langle v, x\rangle)\right] = \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[\mathbb{1}(u^\top x \geq 0)\phi(\langle v, x\rangle)x\right]$$

$$= \frac{\|v\|_2 \sin\theta_{u,v}}{2\pi}\bar{u} + \frac{\pi - \theta_{u,v}}{2\pi}v. \tag{A.9}$$

The Lipschitz continuity of ReLU and Gaussian integrability justify differentiating under the expectation. The hyperplane $\{x : u^\top x = 0\}$ has Gaussian measure zero, so using either $>$ or $\geq$ in the indicator gives the same expectation.

Expanding the squared residual and applying (A.1a) to each student–student, teacher–teacher, and student–teacher pair gives, for nonzero w_1, w_2,

$$\begin{aligned} L(w_1, w_2) &= \frac{1}{2}\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[\big(\phi(\langle w_1, x\rangle) + \phi(\langle w_2, x\rangle) - \phi(\langle v_1, x\rangle) - \phi(\langle v_2, x\rangle)\big)^2\right] \\ &= \frac{1}{4}\big(\|w_1\|_2^2 + \|w_2\|_2^2\big) + \|w_1\|_2\|w_2\|_2 l(\gamma) - \sum_{i=1}^{2}\|w_i\|_2 \sum_{j=1}^{2} l(\theta_{ij}) + l(\alpha) + \frac{1}{2}. \end{aligned} \tag{A.10}$$

Since l is bounded by Lemma A.2, the right-hand side of (A.10) extends continuously to all $w_1, w_2 \in \mathbb{R}^d$ if every term carrying a factor $\|w_i\|_2$ with $w_i = 0$ is set to zero. We use (A.10) with this convention whenever one or both student weights vanish.

On the domain $w_1, w_2 \neq 0$, applying (A.9) to the student–teacher and student–student terms in the gradient flow (2.2) gives

$$\frac{\mathrm{d}w_1}{\mathrm{d}t} = \frac{1}{2\pi}\big((\pi - \theta_{11})v_1 + (\pi - \theta_{12})v_2 + (\sin\theta_{11} + \sin\theta_{12})\bar{w}_1\big) - \left(\frac{1}{2} + \frac{\sin\gamma}{2\pi}\frac{\|w_2\|_2}{\|w_1\|_2}\right)w_1 - \frac{\pi - \gamma}{2\pi}w_2, \tag{A.11}$$

$$\frac{\mathrm{d}w_2}{\mathrm{d}t} = \frac{1}{2\pi}\big((\pi - \theta_{21})v_1 + (\pi - \theta_{22})v_2 + (\sin\theta_{21} + \sin\theta_{22})\bar{w}_2\big) - \frac{\pi - \gamma}{2\pi}w_1 - \left(\frac{1}{2} + \frac{\sin\gamma}{2\pi}\frac{\|w_1\|_2}{\|w_2\|_2}\right)w_2. \tag{A.12}$$

Lemma A.4 (Direction and norm equations)**.** *Let $w(t) = (w_1(t), w_2(t))$ solve the gradient flow (2.2). At each time t for which $w_1(t), w_2(t) \neq 0$, the following identities hold:*

$$\frac{\mathrm{d}\bar{w}_1}{\mathrm{d}t} = \frac{1}{2\pi\|w_1\|_2}(I - \bar{w}_1\bar{w}_1^\top)\big((\pi - \theta_{11})v_1 + (\pi - \theta_{12})v_2 - (\pi - \gamma)w_2\big), \tag{A.13}$$

$$\frac{\mathrm{d}\bar{w}_2}{\mathrm{d}t} = \frac{1}{2\pi\|w_2\|_2}(I - \bar{w}_2\bar{w}_2^\top)\big((\pi - \theta_{21})v_1 + (\pi - \theta_{22})v_2 - (\pi - \gamma)w_1\big), \tag{A.14}$$

$$\begin{aligned} \frac{\mathrm{d}\|w_1\|_2}{\mathrm{d}t} &= \frac{1}{2\pi}\big(\sin\theta_{11} + \sin\theta_{12} + (\pi - \theta_{11})\cos\theta_{11} + (\pi - \theta_{12})\cos\theta_{12}\big) \\ &\quad - \frac{1}{2}\|w_1\|_2 - \frac{\sin\gamma + (\pi - \gamma)\cos\gamma}{2\pi}\|w_2\|_2, \end{aligned} \tag{A.15}$$

$$\begin{aligned} \frac{\mathrm{d}\|w_2\|_2}{\mathrm{d}t} &= \frac{1}{2\pi}\big(\sin\theta_{21} + \sin\theta_{22} + (\pi - \theta_{21})\cos\theta_{21} + (\pi - \theta_{22})\cos\theta_{22}\big) \\ &\quad - \frac{\sin\gamma + (\pi - \gamma)\cos\gamma}{2\pi}\|w_1\|_2 - \frac{1}{2}\|w_2\|_2. \end{aligned} \tag{A.16}$$

Proof of Lemma A.4. Project (A.11) and (A.12) onto the orthogonal complement of each $\bar{w}_i$ for the direction equations and onto each $\bar{w}_i$ for the norm equations. ■

Lemma A.5 (Bounds on the radial velocity)**.** *Let $w(t) = (w_1(t), w_2(t))$ solve the gradient flow (2.2). At each time t for which $w_1(t), w_2(t) \neq 0$,*

$$\frac{r_-}{2} - \frac{\|w_1\|_2 + \|w_2\|_2}{2} \leq \frac{\mathrm{d}\|w_i\|_2}{\mathrm{d}t} \leq 1 - \frac{\|w_i\|_2}{2}, \qquad i = 1, 2,$$

where r_- is defined in (3.4).

Proof of Lemma A.5. By (A.1b), the norm equation (A.15) reads

$$\frac{\mathrm{d}\|w_1\|_2}{\mathrm{d}t} = l(\theta_{11}) + l(\theta_{12}) - \frac{1}{2}\|w_1\|_2 - l(\gamma)\|w_2\|_2.$$

By Lemma A.3 and (3.4), $l(\theta_{11}) + l(\theta_{12}) \geq 2l(\pi - \alpha/2) = r_-/2$, and Lemma A.2 gives $l(\theta_{11}) + l(\theta_{12}) \leq 1$ and $0 \leq l(\gamma) \leq 1/2$. These bounds give the claim for $i = 1$; the case $i = 2$ follows in the same way from (A.16). ■

A.3 Hessian of the population loss

As in Appendix A.2, we consider $\alpha \in (0, \pi)$. We recall $\theta_{w,v} = \angle(w, v)$ and write

$$n_{v,w} := \bar{v} - \cos\theta_{w,v}\bar{w}, \qquad \bar{n}_{v,w} := \frac{n_{v,w}}{\|n_{v,w}\|_2}$$

for nonzero $w, v \in \mathbb{R}^d$ with $\theta_{w,v} \in (0, \pi)$.

Lemma A.6 (Hessian of the population loss)**.** *For nonzero w_1, w_2, the Hessian of the population loss $L(w_1, w_2)$ has diagonal blocks*

$$\frac{\partial^2 L}{\partial w_i^2} = \frac{1}{2} I_d + Q_1(w_i, w_{3-i}) - Q_1(w_i, v_1) - Q_1(w_i, v_2), \qquad i = 1, 2,$$

and off-diagonal blocks

$$\frac{\partial^2 L}{\partial w_i \partial w_{3-i}} = Q_2(w_i, w_{3-i}), \qquad i = 1, 2,$$

where

$$Q_1(w, v) = \frac{\sin\theta_{w,v}\|v\|_2}{2\pi\|w\|_2}\left(I_d - \bar{w}\bar{w}^\top + \bar{n}_{v,w}\bar{n}_{v,w}^\top\right),$$
$$Q_2(w, v) = \frac{1}{2\pi}\left((\pi - \theta_{w,v})I_d + \bar{n}_{v,w}\bar{w}^\top + \bar{n}_{w,v}\bar{v}^\top\right).$$

At collinear pairs, Q_1 and Q_2 are defined by continuous extension: $Q_1(w, v) = 0$ when $\bar{w} = \pm\bar{v}$, while $Q_2(w, v) = \frac{1}{2} I_d$ when $\bar{w} = \bar{v}$ and $Q_2(w, v) = 0$ when $\bar{w} = -\bar{v}$.

Proof of Lemma A.6. This is the block Hessian formula from Theorem 5 of Safran and Shamir (2018), specialized to the two-student, two-teacher loss (2.2). ∎

A.4 Proof of Proposition 2.1

We use the normalized probabilists' Hermite polynomials from Section 2.2,

$$h_k(x) := \frac{(-1)^k}{\sqrt{k!}} e^{x^2/2} \frac{\mathrm{d}^k}{\mathrm{d}x^k} e^{-x^2/2}, \qquad k \geq 0.$$

Proof of Proposition 2.1. Let $a := \cos(\alpha/2)$ and $b := \sin(\alpha/2)$. Since $\|v_1\|_2 = \|v_2\|_2 = 1$, the directions in (2.4) satisfy

$$v_1 = a\bar{v} + bv_\Delta, \qquad v_2 = a\bar{v} - bv_\Delta.$$

With $z_1 = \langle \bar{v}, x\rangle$ and $z_2 = \langle v_\Delta, x\rangle$ as in Section 2.2,

$$f^*(x) = g(z_1, z_2) := \phi(az_1 + bz_2) + \phi(az_1 - bz_2).$$

We expand g as in (2.8) with $(U, W) = (z_1, z_2)$, and write

$$\widehat{\phi}_k := \mathbb{E}_{G\sim\mathcal{N}(0,1)}[\phi(G)h_k(G)].$$

Since $a^2 + b^2 = 1$, the Hermite addition formula

$$h_k(az_1 \pm bz_2) = \sum_{m+n=k} \sqrt{\frac{k!}{m!\,n!}}\, a^m (\pm b)^n h_m(z_1) h_n(z_2)$$

and the orthonormality of $\{h_m(z_1)h_n(z_2)\}_{m,n\geq 0}$ give

$$\mathbb{E}[\phi(az_1 \pm bz_2)h_m(z_1)h_n(z_2)] = \widehat{\phi}_{m+n} \sqrt{\frac{(m+n)!}{m!\,n!}}\, a^m (\pm b)^n.$$

Therefore

$$\widehat{g}_{m,n} = \widehat{\phi}_{m+n} \sqrt{\frac{(m+n)!}{m!\,n!}}\, a^m b^n \left(1 + (-1)^n\right).$$

In particular, $\widehat{g}_{m,1} = 0$ for every $m \geq 0$.

It remains to compute $\widehat{g}_{1,0}$ and $\widehat{g}_{0,2}$. Since $h_1(x) = x$ and $h_2(x) = (x^2 - 1)/\sqrt{2}$,

$$\widehat{\phi}_1 = \mathbb{E}[G^2 \, \mathbb{1}(G > 0)] = \frac{1}{2}, \qquad \widehat{\phi}_2 = \frac{1}{\sqrt{2}} \mathbb{E}[(G^3 - G) \, \mathbb{1}(G > 0)] = \frac{1}{2\sqrt{\pi}}.$$

Thus

$$\widehat{g}_{1,0} = 2\widehat{\phi}_1 a = a \neq 0, \qquad \widehat{g}_{0,2} = 2\widehat{\phi}_2 b^2 = \frac{b^2}{\sqrt{\pi}} \neq 0,$$

where $a > 0$ and $b > 0$ because $\alpha \in (0, \pi/2]$. Hence, along the ordering z_1 then z_2, the leap to z_1 is 1, and the leap to z_2 is 2 because $\widehat{g}_{m,1} = 0$ for every m while $\widehat{g}_{0,2} \neq 0$. The leap exponent is therefore 2. ■

Remark A.7 (Unequal output weights). The argument above uses that the two teacher neurons have equal output weights. If instead

$$y = c_1 \phi(\langle v_1, x \rangle) + c_2 \phi(\langle v_2, x \rangle), \qquad c_1 \neq c_2,$$

then $y = c_1 \phi(az_1 + bz_2) + c_2 \phi(az_1 - bz_2)$, and Stein's lemma gives

$$\mathbb{E}[z_2 y] = \frac{b}{2}(c_1 - c_2) \neq 0.$$

The splitting coordinate then appears at degree 1, so the leap-2 structure is specific to the symmetric teacher considered here.

B Correlation-loss landscape

B.1 Reduction to one student in $\operatorname{span}\{v_1, \ldots, v_k\}$

The correlation loss (3.2) is separable across student neurons, so it is enough to analyze one student on the unit sphere. For a single student $w \in \mathbb{S}^{d-1}$, (A.1a) gives, with l as in (A.1b),

$$L_{\text{corr}}(w) = -\sum_{i=1}^{k} l(\angle(w, v_i)).$$

We also write $g(t) = l(\arccos t)$ for $t \in [-1, 1]$, so that $L_{\text{corr}}(w) = -\sum_{i=1}^{k} g(\langle w, v_i \rangle)$ and

$$g(t) = \frac{t(\pi - \arccos t) + \sqrt{1 - t^2}}{2\pi}. \tag{B.1}$$

A direct computation gives

$$g'(t) = \frac{\pi - \arccos t}{2\pi}, \qquad g''(t) = \frac{1}{2\pi\sqrt{1 - t^2}} > 0 \quad (t \in (-1, 1)). \tag{B.2}$$

We next compute $\langle \bar{v}, v_i \rangle$, where $\bar{v} = \sum_i v_i / \| \sum_i v_i \|_2$ is the direction in Theorem 3.1. Since $\langle v_i, v_j \rangle = \cos\alpha$ for $i \neq j$,

$$\left\| \sum_{i=1}^{k} v_i \right\|_2^2 = k + k(k-1)\cos\alpha = k(1 + (k-1)\cos\alpha), \qquad \left\langle v_i, \sum_{j=1}^{k} v_j \right\rangle = 1 + (k-1)\cos\alpha.$$

Therefore the angle $\rho := \angle(\bar{v}, v_i)$ is the same for every i, with

$$\cos\rho = \sqrt{\frac{1 + (k-1)\cos\alpha}{k}}. \tag{B.3}$$

Under the setting of Theorem 3.1, we have $1 + (k-1)\cos\alpha > 0$: this is immediate when $\alpha \leq \pi/2$, and for $k = 2$ it follows from $\alpha < \pi$. Since also $\alpha > 0$, we have $\rho \in (0, \pi/2)$. Moreover, $L_{\text{corr}}(\bar{v}) = -kg(\cos\rho)$.

Lemma B.1 (Minimizers lie in $\operatorname{span}\{v_1, \dots, v_k\}$)**.** *Under the setting of Theorem 3.1, every minimizer of L_{corr} on $\mathbb{S}^{d-1}$ lies in* $\operatorname{span}\{v_1, \dots, v_k\}$.

Proof of Lemma B.1. Let w_* be a minimizer of L_{corr} on the unit sphere. At a unit vector w, the gradient of (3.2) is $-\sum_{i=1}^k g'(\langle w, v_i\rangle)v_i$ plus a multiple of w, so the Lagrange multiplier condition gives, for some $\lambda \in \mathbb{R}$,

$$\sum_{i=1}^k g'(\langle w_*, v_i\rangle)v_i = \lambda w_*.$$

The left-hand side lies in $\operatorname{span}\{v_1, \dots, v_k\}$, so it suffices to show that $\lambda \neq 0$. If $\lambda = 0$, taking the inner product with $\sum_{j=1}^k v_j$ gives

$$0 = \sum_{i=1}^k g'(\langle w_*, v_i\rangle)\left\langle v_i, \sum_{j=1}^k v_j \right\rangle = (1 + (k-1)\cos\alpha)\sum_{i=1}^k g'(\langle w_*, v_i\rangle).$$

Because $1 + (k-1)\cos\alpha > 0$ and $g' \geq 0$, this forces $g'(\langle w_*, v_i\rangle) = 0$ for every i. Since $g'(t) = 0$ only at $t = -1$, we get $w_* = -v_i$ for all i, which contradicts $\alpha > 0$. Hence $\lambda \neq 0$ and $w_* \in \operatorname{span}\{v_1, \dots, v_k\}$. ∎

It remains to compare $L_{\mathrm{corr}}(w)$ with $L_{\mathrm{corr}}(\bar{v})$ for unit vectors $w \in \operatorname{span}\{v_1, \dots, v_k\}$. We do this for $0 < \alpha \leq \pi/2$ in Lemmas B.3 and B.4, and for $k = 2$ and $\pi/2 < \alpha < \pi$ in Lemma B.5.

B.2 The case $\alpha \leq \pi/2$

For $0 < \alpha \leq \pi/2$ and general k, we treat $\rho \geq \pi/4$ and $\rho \leq \pi/4$ separately. The first case relies on the following scalar inequality for g.

Lemma B.2. *Let $\rho \in [\pi/4, \pi/2)$, and let g be as in (B.1). Then*

$$g(y) \leq g(\cos\rho) + \frac{\pi - \rho}{4\pi\cos\rho}(y^2 - \cos^2\rho), \qquad y \in [-1, 1],$$

with equality only at $y = \cos\rho$.

Proof of Lemma B.2. Write $y = \cos\theta$ with $\theta \in [0, \pi]$. Since $g(\cos\theta) = l(\theta)$, where l is defined in (A.1b), the inequality is equivalent to $R_\rho(\theta) \geq 0$, where

$$R_\rho(\theta) := l(\rho) + \frac{\pi - \rho}{4\pi\cos\rho}(\cos^2\theta - \cos^2\rho) - l(\theta).$$

Note that $R_\rho(\rho) = 0$. By (A.2),

$$R_\rho'(\theta) = \frac{\sin\theta}{2\pi}\, H_\rho(\theta), \qquad H_\rho(\theta) := (\pi - \theta) - \frac{\pi - \rho}{\cos\rho}\cos\theta.$$

We show that $H_\rho < 0$ on $(0, \rho)$ and $H_\rho > 0$ on (ρ, π).

On $[0, \rho]$, we have $H_\rho(\rho) = 0$ and $H_\rho(0) = \pi - (\pi - \rho)/\cos\rho$, which is negative if and only if $\pi\cos\rho < \pi - \rho$. The function $\rho \mapsto \pi - \rho - \pi\cos\rho$ is positive at $\rho = \pi/4$, where it equals $3\pi/4 - \pi/\sqrt{2}$, and its derivative $-1 + \pi\sin\rho \geq -1 + \pi/\sqrt{2}$ is positive on $[\pi/4, \pi/2)$. Hence $H_\rho(0) < 0$. Moreover,

$$H_\rho'(\theta) = -1 + \frac{\pi - \rho}{\cos\rho}\sin\theta, \qquad H_\rho''(\theta) = \frac{\pi - \rho}{\cos\rho}\cos\theta,$$

so H_ρ is convex on $[0, \pi/2]$. Since $H_\rho(0) < 0 = H_\rho(\rho)$, convexity gives $H_\rho < 0$ on $(0, \rho)$.

On $(\rho, \pi/2]$, we use $H_\rho'(\rho) = -1 + (\pi - \rho)\tan\rho$. The function $\rho \mapsto (\pi - \rho)\tan\rho$ equals $3\pi/4 > 1$ at $\rho = \pi/4$ and is increasing on $[\pi/4, \pi/2)$, since

$$\frac{\mathrm{d}}{\mathrm{d}\rho}\big((\pi - \rho)\tan\rho\big) = \frac{(\pi - \rho) - \sin\rho\cos\rho}{\cos^2\rho} > 0$$

because $\pi - \rho > \pi/2 > \sin\rho\cos\rho$. Hence $H'_\rho(\rho) > 0$. By convexity, H'_ρ is nondecreasing on $[0, \pi/2]$, so $H'_\rho > 0$ on $[\rho, \pi/2]$ and $H_\rho(\theta) > H_\rho(\rho) = 0$ for $\rho < \theta \le \pi/2$.

On $[\pi/2, \pi)$, we have $\cos\theta \le 0$ and $\pi - \theta > 0$, so $H_\rho(\theta) > 0$.

Since $\sin\theta > 0$ on $(0, \pi)$, the function R_ρ is strictly decreasing on $[0, \rho]$ and strictly increasing on $[\rho, \pi]$. Thus $R_\rho(\theta) > R_\rho(\rho) = 0$ for $\theta \neq \rho$. ■

The case $\rho \ge \pi/4$ now follows from Lemma B.2.

Lemma B.3. *Suppose that $0 < \alpha \le \pi/2$ and $\rho \ge \pi/4$, with ρ as in (B.3). Then $L_{\mathrm{corr}}(w) \ge L_{\mathrm{corr}}(\bar{v})$ for every unit vector $w \in \mathrm{span}\{v_1, \dots, v_k\}$, with equality only at $w = \bar{v}$.*

Proof of Lemma B.3. Applying Lemma B.2 to each $\langle w, v_i\rangle$ and summing over i gives

$$L_{\mathrm{corr}}(w) - L_{\mathrm{corr}}(\bar{v}) = kg(\cos\rho) - \sum_i g(\langle w, v_i\rangle) \ge \frac{\pi - \rho}{4\pi\cos\rho}\left(k\cos^2\rho - \sum_i \langle w, v_i\rangle^2\right),$$

with equality only if $\langle w, v_i\rangle = \cos\rho$ for every i. The coefficient $(\pi - \rho)/(4\pi\cos\rho)$ is positive. Moreover, $\sum_i \langle w, v_i\rangle^2 = w^\top(\sum_i v_i v_i^\top)w$. The matrix $\sum_i v_i v_i^\top$ has eigenvalue $1 + (k-1)\cos\alpha = k\cos^2\rho$ in the direction $\bar{v}$ and eigenvalue $1 - \cos\alpha$ on the orthogonal complement of $\bar{v}$ inside $\mathrm{span}\{v_1, \dots, v_k\}$. Since $\cos\alpha \ge 0$, we have $1 - \cos\alpha \le k\cos^2\rho$, and hence $\sum_i \langle w, v_i\rangle^2 \le k\cos^2\rho$. Therefore $L_{\mathrm{corr}}(w) \ge L_{\mathrm{corr}}(\bar{v})$.

If $L_{\mathrm{corr}}(w) = L_{\mathrm{corr}}(\bar{v})$, the inequality in the display is an equality, so $\langle w, v_i\rangle = \cos\rho = \langle \bar{v}, v_i\rangle$ for all i. The Gram matrix

$$(\langle v_i, v_j\rangle)_{i,j} = (1 - \cos\alpha)I + \cos\alpha\,\mathbf{1}\mathbf{1}^\top$$

is positive definite because $0 \le \cos\alpha < 1$, and both w and $\bar{v}$ lie in $\mathrm{span}\{v_1, \dots, v_k\}$. Hence $w = \bar{v}$. ■

When $\rho \le \pi/4$, we use a different argument: we write w in terms of its angle φ from $\bar{v}$ and bound $L_{\mathrm{corr}}(w)$ separately for $\varphi \ge \rho$ and $\varphi \le \rho$.

Lemma B.4. *Suppose that $0 < \alpha \le \pi/2$ and $\rho \le \pi/4$, with ρ as in (B.3). Then $L_{\mathrm{corr}}(w) \ge L_{\mathrm{corr}}(\bar{v})$ for every unit vector $w \in \mathrm{span}\{v_1, \dots, v_k\}$, with equality only at $w = \bar{v}$.*

Proof of Lemma B.4. Let $\tilde{v}_i = (v_i - \cos\rho\,\bar{v})/\sin\rho$ be the unit vector in the direction of the component of v_i orthogonal to $\bar{v}$. These vectors satisfy

$$v_i = \cos\rho\,\bar{v} + \sin\rho\,\tilde{v}_i, \qquad \tilde{v}_i \perp \bar{v}, \qquad \|\tilde{v}_i\|_2 = 1, \qquad \sum_i \tilde{v}_i = 0, \qquad \langle \tilde{v}_i, \tilde{v}_j\rangle = -\frac{1}{k-1} \quad (i \neq j). \tag{B.4}$$

We write

$$w = \cos\varphi\,\bar{v} + \sin\varphi\,u,$$

where $\varphi \in [0, \pi]$, $u \perp \bar{v}$, $\|u\|_2 = 1$, and $u \in \mathrm{span}\{\tilde{v}_1, \dots, \tilde{v}_k\}$. When $\sin\varphi = 0$, the vector u is arbitrary, and we fix any such u. Let $t_i = \langle u, \tilde{v}_i\rangle$ and

$$z_i = \langle w, v_i\rangle = \cos\rho\cos\varphi + t_i\sin\rho\sin\varphi.$$

With a slight abuse of notation, we write

$$L_{\mathrm{corr}}(\varphi, u) \coloneqq L_{\mathrm{corr}}(\cos\varphi\,\bar{v} + \sin\varphi\,u) = -\sum_{i=1}^k g(z_i),$$

where g is defined in (B.1). In particular, $L_{\mathrm{corr}}(0, u) = L_{\mathrm{corr}}(\bar{v})$.

We note that the numbers t_i satisfy

$$\sum_i t_i = 0, \qquad \sum_i t_i^2 = \frac{k}{k-1}, \qquad -1 \le t_i \le 1.$$

Indeed, $\sum_i t_i = \langle u, \sum_i \tilde{v}_i \rangle = 0$, and $|t_i| \le 1$ by Cauchy–Schwarz. For every j, (B.4) gives

$$\left\langle \sum_i t_i \tilde{v}_i, \tilde{v}_j \right\rangle = t_j - \frac{1}{k-1} \sum_{i \neq j} t_i = \frac{k}{k-1} t_j = \left\langle \frac{k}{k-1} u, \tilde{v}_j \right\rangle.$$

Both $\sum_i t_i \tilde{v}_i$ and u lie in $\operatorname{span}\{\tilde{v}_1, \dots, \tilde{v}_k\}$, so $\sum_i t_i \tilde{v}_i = \frac{k}{k-1} u$. Taking the inner product with u gives $\sum_i t_i^2 = k/(k-1)$.

We first show that $L_{\mathrm{corr}}(\varphi, u)$ is nondecreasing in φ on $[\rho, \pi]$. We have

$$\frac{\partial}{\partial \varphi} L_{\mathrm{corr}}(\varphi, u) = \cos\rho \sin\varphi \sum_i g'(z_i) - \sin\rho \cos\varphi \sum_i t_i g'(z_i). \tag{B.5}$$

Since $\sum_i t_i = 0$, the identity $\sum_{i<j} (t_i - t_j)\big(g'(z_i) - g'(z_j)\big) = k \sum_i t_i g'(z_i) - \big(\sum_i t_i\big)\big(\sum_i g'(z_i)\big)$ gives

$$\sum_i t_i g'(z_i) = \frac{1}{k} \sum_{i<j} (t_i - t_j)\big(g'(z_i) - g'(z_j)\big) \ge 0, \tag{B.6}$$

because g' is increasing by (B.2) and z_i is nondecreasing in t_i (recall $\sin\rho > 0$ and $\sin\varphi \ge 0$).

If $\varphi \ge \pi/2$, then $\cos\varphi \le 0$, so both terms in (B.5) are nonnegative, by $g' \ge 0$ and (B.6). If $\rho \le \varphi < \pi/2$, then $\cos\varphi > 0$ and $\tan\varphi \ge \tan\rho$, so $\cot\rho \tan\varphi \ge 1 \ge t_i$ and

$$\frac{\partial}{\partial \varphi} L_{\mathrm{corr}}(\varphi, u) = \sin\rho \cos\varphi \sum_i (\cot\rho \tan\varphi - t_i)\, g'(z_i) \ge 0.$$

Hence $L_{\mathrm{corr}}(\varphi, u)$ is nondecreasing in φ on $[\rho, \pi]$, and

$$L_{\mathrm{corr}}(\varphi, u) \ge L_{\mathrm{corr}}(\rho, u), \qquad \varphi \ge \rho.$$

It remains to consider $0 \le \varphi \le \rho$. We first take $0 \le \varphi < \rho$; the bound (B.7) obtained below then extends to $\varphi = \rho$ by continuity. Since $\rho \le \pi/4$, we have $\rho + \varphi < \pi/2$, and for every $t \in [-1, 1]$,

$$0 \le \cos(\rho + \varphi) \le \cos\rho \cos\varphi + t \sin\rho \sin\varphi \le \cos(\rho - \varphi) < 1.$$

Hence the function $h_\varphi(t) = g(\cos\rho\cos\varphi + t \sin\rho \sin\varphi)$ is C^3 on $[-1, 1]$. Differentiating (B.2) once more gives $g'''(s) = s/(2\pi(1 - s^2)^{3/2})$, which is nonnegative for $s \in [0, 1)$. By the chain rule,

$$h_\varphi'''(t) = (\sin\rho \sin\varphi)^3\, g'''(\cos\rho\cos\varphi + t \sin\rho\sin\varphi) \ge 0, \qquad t \in [-1, 1].$$

Let $s_0 = -1/(k-1) \in [-1, 0)$, and let Q_φ be the quadratic polynomial

$$Q_\varphi(t) = h_\varphi(s_0) + h_\varphi'(s_0)(t - s_0) + \frac{h_\varphi(1) - h_\varphi(s_0) - h_\varphi'(s_0)(1 - s_0)}{(1 - s_0)^2} (t - s_0)^2,$$

which satisfies $Q_\varphi(1) = h_\varphi(1)$, $Q_\varphi(s_0) = h_\varphi(s_0)$, and $Q_\varphi'(s_0) = h_\varphi'(s_0)$. For $t \in [-1, 1]$, the Hermite interpolation remainder with a double node at s_0 and a simple node at 1 gives

$$h_\varphi(t) - Q_\varphi(t) = \frac{h_\varphi'''(\xi)}{6} (t - s_0)^2 (t - 1)$$

for some $\xi \in [-1, 1]$. Since $h_\varphi'''(\xi) \ge 0$ and $t - 1 \le 0$, we have $h_\varphi \le Q_\varphi$ on $[-1, 1]$, and hence $\sum_i h_\varphi(t_i) \le \sum_i Q_\varphi(t_i)$.

Since Q_φ is quadratic, $\sum_i Q_\varphi(t_i)$ depends on the t_i only through $\sum_i t_i = 0$ and $\sum_i t_i^2 = k/(k-1)$. The k numbers $1, s_0, \dots, s_0$ have the same two sums, because $1 + (k-1)s_0 = 0$ and $1 + (k-1)s_0^2 = k/(k-1)$. Therefore

$$-L_{\mathrm{corr}}(\varphi, u) = \sum_i h_\varphi(t_i) \le Q_\varphi(1) + (k-1) Q_\varphi(s_0) = h_\varphi(1) + (k-1) h_\varphi(s_0) = -L_{\mathrm{corr}}(\varphi, \tilde{v}_1), \tag{B.7}$$

where the last equality holds because $u = \tilde{v}_1$ gives $t_1 = 1$ and $t_j = s_0$ for $j \neq 1$, by (B.4). Thus, for each $\varphi \in [0, \rho]$, $L_{\mathrm{corr}}(\varphi, \cdot)$ is minimized at $u = \tilde{v}_1$.

For $u = \tilde{v}_1$, we have $z_1 = \cos(\rho - \varphi)$, and $z_j = z_2$ for every $j \geq 2$, where

$$z_2 = \cos\rho\cos\varphi - \frac{\sin\rho\sin\varphi}{k-1}, \qquad L_{\mathrm{corr}}(\varphi, \tilde{v}_1) = -g\big(\cos(\rho-\varphi)\big) - (k-1)g(z_2). \tag{B.8}$$

We now show that $L_{\mathrm{corr}}(\varphi, \tilde{v}_1) > L_{\mathrm{corr}}(\bar{v})$ for $0 < \varphi \leq \rho$. Since $0 \leq \varphi \leq \rho \leq \pi/4$,

$$0 \leq \cos(\rho + \varphi) \leq z_2 \leq \cos\rho. \tag{B.9}$$

Since the arccosine is decreasing, (B.9) gives $\rho \leq \arccos z_2 \leq \rho + \varphi$. We write $\delta(\varphi) = \arccos z_2 - \rho$, so that $0 \leq \delta(\varphi) \leq \varphi$. By (B.2),

$$g'\big(\cos(\rho-\varphi)\big) = \frac{\pi - \rho + \varphi}{2\pi}, \qquad g'(z_2) = \frac{\pi - \rho - \delta(\varphi)}{2\pi}.$$

Differentiating (B.8) therefore gives

$$\begin{aligned}\frac{\partial}{\partial\varphi} L_{\mathrm{corr}}(\varphi, \tilde{v}_1) &= \frac{1}{2\pi}\Big[-(\pi - \rho + \varphi)\sin(\rho - \varphi) \\ &\qquad + (k-1)\big(\pi - \rho - \delta(\varphi)\big)\left(\cos\rho\sin\varphi + \frac{\sin\rho\cos\varphi}{k-1}\right)\Big] \\ &= \frac{1}{2\pi}\Big[\cos\rho\sin\varphi\,[k(\pi-\rho) + \varphi - (k-1)\delta(\varphi)] - \sin\rho\cos\varphi\big(\varphi + \delta(\varphi)\big)\Big].\end{aligned} \tag{B.10}$$

The bracket in (B.10) is an affine function of $\delta(\varphi)$ with slope $-(k-1)\cos\rho\sin\varphi - \sin\rho\cos\varphi < 0$. Since $\delta(\varphi) \leq \varphi$, replacing $\delta(\varphi)$ by φ gives

$$\frac{\partial}{\partial\varphi} L_{\mathrm{corr}}(\varphi, \tilde{v}_1) \geq \frac{1}{2\pi}\left[\cos\rho\sin\varphi\,[k(\pi-\rho) - (k-2)\varphi] - 2\varphi\sin\rho\cos\varphi\right]. \tag{B.11}$$

For $0 < \varphi \leq \rho$, dividing by $\cos\rho\sin\varphi > 0$ shows that the right-hand side of (B.11) is positive whenever

$$2\tan\rho \cdot \varphi\cot\varphi < k(\pi - \rho) - (k-2)\varphi.$$

This inequality holds: since $\varphi\cot\varphi \leq 1$ and $\rho \leq \pi/4$, the left-hand side is at most $2\tan\rho \leq 2$, while the right-hand side is at least $k(\pi-\rho) - (k-2)\rho = k\pi - 2(k-1)\rho \geq (k+1)\pi/2 \geq 3\pi/2 > 2$. Hence $L_{\mathrm{corr}}(\varphi, \tilde{v}_1)$ is strictly increasing on $[0, \rho]$, and

$$L_{\mathrm{corr}}(\varphi, \tilde{v}_1) > L_{\mathrm{corr}}(0, \tilde{v}_1) = L_{\mathrm{corr}}(\bar{v}), \qquad 0 < \varphi \leq \rho.$$

Combining the two ranges with (B.7), $L_{\mathrm{corr}}(\varphi, u) \geq L_{\mathrm{corr}}(\varphi, \tilde{v}_1) > L_{\mathrm{corr}}(\bar{v})$ for $0 < \varphi \leq \rho$, and $L_{\mathrm{corr}}(\varphi, u) \geq L_{\mathrm{corr}}(\rho, u) \geq L_{\mathrm{corr}}(\rho, \tilde{v}_1) > L_{\mathrm{corr}}(\bar{v})$ for $\varphi \geq \rho$. Thus $L_{\mathrm{corr}}(w) \geq L_{\mathrm{corr}}(\bar{v})$, with equality only when $\varphi = 0$, that is, $w = \bar{v}$. ■

B.3 The case $k = 2$ and $\alpha > \pi/2$

For $k = 2$, we have $\rho = \alpha/2$. An obtuse α therefore gives $\cos\alpha < 0$, which is outside the range of Lemma B.3. In this case $\mathrm{span}\{v_1, v_2\} = \mathrm{span}\{\bar{v}, v_\Delta\}$, and we compare L_{corr} along its unit circle directly.

Lemma B.5. *Let $k = 2$ and $\pi/2 < \alpha < \pi$, and let $\alpha_\star$ be as in Section 3.1.*

(a) If $\alpha \leq \alpha_\star$, then $\bar{v}$ is the unique minimizer of L_{corr} on the unit sphere.

(b) If $\alpha > \alpha_\star$, then $\bar{v}$ is not a local minimizer of L_{corr} on the unit sphere.

Proof of Lemma B.5. Recall from (2.4) that $\bar{v}$ and v_Δ form an orthonormal basis of the teacher plane, with

$$v_1 = \cos(\alpha/2)\,\bar{v} + \sin(\alpha/2)\,v_\Delta, \qquad v_2 = \cos(\alpha/2)\,\bar{v} - \sin(\alpha/2)\,v_\Delta.$$

As in the proof of Lemma B.4, we write $L_{\mathrm{corr}}(\varphi, v_\Delta) = L_{\mathrm{corr}}(\cos\varphi\,\bar{v} + \sin\varphi\,v_\Delta)$. For $0 \leq \varphi \leq \pi/2$, with l as in (A.1b),

$$L_{\mathrm{corr}}(\varphi, v_\Delta) = -l(|\alpha/2 - \varphi|) - l(\alpha/2 + \varphi).$$

For $0 < \varphi < \alpha/2$, direct differentiation gives

$$\frac{\partial}{\partial \varphi} L_{\mathrm{corr}}(\varphi, v_\Delta) = \frac{\cos(\alpha/2)\,\sin\varphi}{\pi}\left[(\pi - \alpha/2) - \tan(\alpha/2)\cdot\varphi\cot\varphi\right]. \tag{B.12}$$

We first prove (a). Since $\arccos(-t) = \pi - \arccos t$, the formula (B.1) gives $g(t) - g(-t) = t/2$ for $t \in [-1, 1]$. Consequently,

$$L_{\mathrm{corr}}(-w) - L_{\mathrm{corr}}(w) = \frac{1}{2}\langle w, v_1 + v_2\rangle = \cos(\alpha/2)\,\langle w, \bar{v}\rangle.$$

By Lemma B.1, every minimizer lies in $\mathrm{span}\{\bar{v}, v_\Delta\}$. By the identity above, a minimizer cannot have $\langle w, \bar{v}\rangle < 0$, and reflection across $\bar{v}$ leaves L_{corr} unchanged while changing the sign of $\langle w, v_\Delta\rangle$. It therefore suffices to show that $L_{\mathrm{corr}}(\varphi, v_\Delta) > L_{\mathrm{corr}}(\bar{v})$ for $0 < \varphi \le \pi/2$. The definition $\tan(\alpha_\star/2) = \pi - \alpha_\star/2$ of $\alpha_\star$, together with the monotonicity of $x \mapsto \tan x + x$, gives $\tan(\alpha/2) \le \pi - \alpha/2$. Since $\varphi\cot\varphi < 1$, the derivative (B.12) is positive for $0 < \varphi < \alpha/2$. For $\alpha/2 < \varphi \le \pi/2$,

$$\frac{\partial}{\partial \varphi} L_{\mathrm{corr}}(\varphi, v_\Delta) = -l'(\varphi - \alpha/2) - l'(\varphi + \alpha/2) > 0$$

by (A.2), since $0 < \varphi + \alpha/2 < \pi$. The one-sided derivatives agree at $\varphi = \alpha/2$ because $l'(0) = 0$. Hence $L_{\mathrm{corr}}(\varphi, v_\Delta) > L_{\mathrm{corr}}(\bar{v})$ for every $\varphi \in (0, \pi/2]$.

We now prove (b). Since $\alpha > \alpha_\star$, we have $\tan(\alpha/2) > \pi - \alpha/2$. Differentiating (B.12) at $\varphi = 0$ gives

$$\frac{\partial^2}{\partial \varphi^2} L_{\mathrm{corr}}(\varphi, v_\Delta)\Big|_{\varphi=0} = \frac{(\pi - \alpha/2)\cos(\alpha/2) - \sin(\alpha/2)}{\pi} < 0,$$

while the first derivative vanishes at $\varphi = 0$. Hence $L_{\mathrm{corr}}(\varphi, v_\Delta) < L_{\mathrm{corr}}(\bar{v})$ for small $\varphi > 0$, so $\bar{v}$ is not a local minimizer of L_{corr}. ■

B.4 Proof of Theorem 3.1

Proof of Theorem 3.1. Since L_{corr} is continuous on the compact unit sphere, it has a minimizer, and by Lemma B.1 every minimizer lies in $\mathrm{span}\{v_1, \ldots, v_k\}$. If $0 < \alpha \le \pi/2$, then Lemma B.3 (when $\rho \ge \pi/4$) or Lemma B.4 (when $\rho \le \pi/4$) shows that $\bar{v}$ is the only unit vector in $\mathrm{span}\{v_1, \ldots, v_k\}$ attaining the minimum of L_{corr}. If $k = 2$ and $\pi/2 < \alpha \le \alpha_\star$, the same conclusion is part (a) of Lemma B.5. In either case, $\bar{v}$ is the unique minimizer of L_{corr} on the unit sphere. For unit-norm $w_1, \ldots, w_m$,

$$L_{\mathrm{corr}}(w_1, \ldots, w_m) = \sum_{j=1}^{m} L_{\mathrm{corr}}(w_j) \ge m L_{\mathrm{corr}}(\bar{v}),$$

with equality if and only if $w_j = \bar{v}$ for every j. Thus $(\bar{v}, \ldots, \bar{v})$ is the unique global minimum. ■

Remark B.6 (Obtuse angles for $k > 2$). For fixed $k > 2$ and $d \ge k$, $(\bar{v}, \ldots, \bar{v})$ remains the unique global minimizer when $\cos\alpha$ is negative and sufficiently close to zero. This follows from a perturbation argument at $\alpha = \pi/2$. There, for a single student, $\bar{v}$ is a strict minimizer of L_{corr} on the unit sphere, with positive definite Hessian on the sphere, and L_{corr} is bounded away from its minimum outside any neighborhood of $\bar{v}$. Both properties persist under small changes of α. The size of this range of α may depend on k and d, and we do not determine how far it extends.

C Squared-loss landscape

C.1 Stationarity and dimension reduction

The proofs below use the condition $\alpha \le \alpha_\star$ in the form $\sin(\alpha/2) \le (\pi - \alpha/2)\cos(\alpha/2)$. The following lemma shows that the two are equivalent.

Lemma C.1 (Equivalent angle threshold)**.** *The equation* $\tan(\alpha/2) = \pi - \alpha/2$ *has exactly one solution in* $(\pi/2, \pi)$*, namely the threshold* $\alpha_\star$ *of Section 3.1. Moreover, for every* $\alpha \in (0, \pi)$*, the following conditions are equivalent: (a)* $\alpha \le \alpha_\star$*; (b)* $\tan(\alpha/2) \le \pi - \alpha/2$*; (c)* $\sin(\alpha/2) \le (\pi - \alpha/2)\cos(\alpha/2)$.

Proof of Lemma C.1. The function $\alpha \mapsto \tan(\alpha/2) + \alpha/2 - \pi$ is strictly increasing on $(0, \pi)$, since its derivative is $\big(\sec^2(\alpha/2) + 1\big)/2 > 0$. It equals $1 - 3\pi/4 < 0$ at $\alpha = \pi/2$ and tends to $+\infty$ as $\alpha \uparrow \pi$. Thus it has a unique root $\alpha_\star \in (\pi/2, \pi)$ and is nonpositive exactly on $(0, \alpha_\star]$, so (a) and (b) are equivalent. Since $\cos(\alpha/2) > 0$, multiplying (b) by $\cos(\alpha/2)$ shows that (b) and (c) are equivalent. ■

C.1.1 Stationary equations

We first derive the stationarity equations (3.3) from the gradient-flow equations (A.11)–(A.12).

Proof of Lemma 3.2. Since $\frac{\mathrm{d}w_i}{\mathrm{d}t} = -\nabla_{w_i} L$, the pair (w_1, w_2) is stationary if and only if $\frac{\mathrm{d}w_1}{\mathrm{d}t} = \frac{\mathrm{d}w_2}{\mathrm{d}t} = 0$ in (A.11)–(A.12). Setting $\frac{\mathrm{d}w_1}{\mathrm{d}t} = 0$ in (A.11), multiplying by 2π, and using $w_i = \|w_i\|_2 \bar{w}_i$ gives

$$(\pi - \theta_{11})v_1 + (\pi - \theta_{12})v_2 + a_1\bar{w}_1 - (\pi - \gamma)\|w_2\|_2\bar{w}_2 = 0,$$

which is the first identity in (3.3). Setting $\frac{\mathrm{d}w_2}{\mathrm{d}t} = 0$ in (A.12) similarly yields the second identity. These steps can be reversed, so (3.3) holds if and only if (w_1, w_2) is stationary. ■

C.1.2 Uniqueness of the out-of-plane root

The out-of-plane saddles in Theorem 3.3 are described by the following functions. For $0 < \kappa < \pi/2$, define

$$\theta_\alpha(\kappa) = \arccos\left(\cos\kappa \cdot \cos\frac{\alpha}{2}\right), \tag{C.1}$$

$$r_\alpha(\kappa) = \frac{2\sin\theta_\alpha(\kappa)}{\sin(2\kappa) + 2\kappa}, \tag{C.2}$$

$$F_\alpha(\kappa) = (\pi - \theta_\alpha(\kappa))\cos\frac{\alpha}{2} - (\pi - 2\kappa)r_\alpha(\kappa)\cos\kappa. \tag{C.3}$$

The saddles are determined by a root of F_α in $(0, \pi/2)$, and we next show that this root exists and is unique.

Lemma C.2 (Unique out-of-plane root)**.** *For every $\alpha \in (0, \pi)$, the equation $F_\alpha(\kappa) = 0$ has exactly one solution in $(0, \pi/2)$, which we denote by κ_α.*

Proof of Lemma C.2. Fix $\alpha \in (0, \pi)$. By (C.1), $\theta_\alpha(\kappa) = \arccos\big(\cos(\alpha/2)\cos\kappa\big)$, so for $0 < \kappa < \pi/2$ we have $\theta_\alpha(\kappa) < \pi/2$ and $0 < \theta'_\alpha(\kappa) = \cos(\alpha/2)\sin\kappa/\sin\theta_\alpha(\kappa) < 1$, where the last inequality follows from $\sin^2\theta_\alpha(\kappa) - \cos^2(\alpha/2)\sin^2\kappa = \sin^2(\alpha/2) > 0$. Using $\sin(2\kappa) = 2\sin\kappa\cos\kappa$, we first rewrite (C.2) as

$$r_\alpha(\kappa) = \frac{2\sin\theta_\alpha(\kappa)}{\sin(2\kappa) + 2\kappa} = \frac{\sin\theta_\alpha(\kappa)}{\kappa + \sin\kappa\cos\kappa}.$$

Substituting this expression into (C.3) yields

$$F_\alpha(\kappa) = \big(\pi - \theta_\alpha(\kappa)\big)\cos(\alpha/2) - \frac{(\pi - 2\kappa)\sin\theta_\alpha(\kappa)\cos\kappa}{\kappa + \sin\kappa\cos\kappa}.$$

The fraction in the second term is strictly positive because $0 < \kappa < \pi/2$ and $0 < \theta_\alpha(\kappa) < \pi/2$. Dividing by this fraction shows that $F_\alpha(\kappa) = 0$ if and only if $Q_\alpha(\kappa) = 1$, where

$$\begin{aligned} Q_\alpha(\kappa) &= \frac{\cos(\alpha/2)\,\big(\pi - \theta_\alpha(\kappa)\big)(\kappa + \sin\kappa\cos\kappa)}{(\pi - 2\kappa)\sin\theta_\alpha(\kappa)\cos\kappa} \\ &= \cos(\alpha/2)\,\frac{\pi - \theta_\alpha(\kappa)}{\pi - 2\kappa}\,\frac{\kappa + \sin\kappa\cos\kappa}{\sin\kappa\cos\kappa}\,\frac{\sin\kappa}{\sin\theta_\alpha(\kappa)} \\ &= \cos(\alpha/2)\,\frac{\pi - \theta_\alpha(\kappa)}{\pi - 2\kappa}\left(1 + \frac{2\kappa}{\sin(2\kappa)}\right)\frac{\sin\kappa}{\sin\theta_\alpha(\kappa)}. \end{aligned} \tag{C.4}$$

We can show that each of the three positive factors after $\cos(\alpha/2)$ in the last line is strictly increasing in κ. For the first factor,

$$\frac{\mathrm{d}}{\mathrm{d}\kappa}\frac{\pi - \theta_\alpha(\kappa)}{\pi - 2\kappa} = \frac{2\big(\pi - \theta_\alpha(\kappa)\big) - (\pi - 2\kappa)\theta'_\alpha(\kappa)}{(\pi - 2\kappa)^2} > \frac{2\kappa}{(\pi - 2\kappa)^2} > 0.$$

Here $\pi - 2\kappa > 0$, $\theta'_\alpha(\kappa) < 1$, and $\theta_\alpha(\kappa) < \pi/2$ give

$$2\big(\pi - \theta_\alpha(\kappa)\big) - (\pi - 2\kappa)\theta'_\alpha(\kappa) > 2\big(\pi - \theta_\alpha(\kappa)\big) - (\pi - 2\kappa) = \pi - 2\theta_\alpha(\kappa) + 2\kappa > 2\kappa.$$

For the second factor, $0 < 2\kappa < \pi$ implies $t \sin t > 0$ for $0 < t < 2\kappa$, so integration by parts gives

$$0 < \int_0^{2\kappa} t \sin t \, \mathrm{d}t = (-t\cos t + \sin t)\big|_0^{2\kappa} = \sin(2\kappa) - 2\kappa\cos(2\kappa).$$

Therefore

$$\frac{\mathrm{d}}{\mathrm{d}\kappa}\left(1 + \frac{2\kappa}{\sin(2\kappa)}\right) = \frac{2\big(\sin(2\kappa) - 2\kappa\cos(2\kappa)\big)}{\sin^2(2\kappa)} > 0.$$

For the third factor, the identity $\sin\kappa / \sin\theta_\alpha(\kappa) = \big(1 + \sin^2(\alpha/2)\cot^2\kappa\big)^{-1/2}$ and $\sin^2(\alpha/2) > 0$ give strict monotonicity. Hence Q_α in (C.4) is continuous and strictly increasing on $(0, \pi/2)$.

Finally, as $\kappa \downarrow 0$, we have $\theta_\alpha(\kappa) \to \alpha/2$, so the third factor tends to 0 while the first two stay bounded. As $\kappa \uparrow \pi/2$, we have $\theta_\alpha(\kappa) \to \pi/2$, so the first two factors tend to $+\infty$ while the third tends to 1. Hence

$$\lim_{\kappa\downarrow 0} Q_\alpha(\kappa) = 0, \qquad \lim_{\kappa\uparrow\pi/2} Q_\alpha(\kappa) = +\infty.$$

The intermediate value theorem and strict monotonicity therefore give exactly one solution of $Q_\alpha(\kappa) = 1$, and thus exactly one root of F_α in $(0, \pi/2)$. ■

C.1.3 Reduction to three dimensions

The teacher terms in (3.3) lie in the teacher plane, so the components of (3.3) orthogonal to this plane involve only the students. These components give the dimension reduction in Proposition 3.5.

Proof of Proposition 3.5. Projecting the two equations in (3.3) onto the orthogonal complement of $\operatorname{span}\{v_1, v_2\}$ gives $a_1\bar{w}_1^\perp - (\pi - \gamma)\|w_2\|_2\bar{w}_2^\perp = 0$ and $a_2\bar{w}_2^\perp - (\pi - \gamma)\|w_1\|_2\bar{w}_1^\perp = 0$. If $\gamma < \pi$, then $(\pi - \gamma)\|w_2\|_2 > 0$, so the first identity expresses $\bar{w}_2^\perp$ as a multiple of $\bar{w}_1^\perp$, possibly zero. If $\gamma = \pi$, then $\bar{w}_2 = -\bar{w}_1$, so $\bar{w}_2^\perp = -\bar{w}_1^\perp$. In either case $\bar{w}_1^\perp$ and $\bar{w}_2^\perp$ are linearly dependent, and hence so are $w_1^\perp$ and $w_2^\perp$. Thus w_1 and w_2 lie in the span of the teacher plane and at most one further direction, which gives $\dim\operatorname{span}\{v_1, v_2, w_1, w_2\} \le 3$. ■

C.2 Three-dimensional symmetry reduction

C.2.1 The teacher force

The angle between $\bar{v}$ and each teacher is $\alpha/2$, so

$$v_1 = \cos(\alpha/2)\,\bar{v} + \sin(\alpha/2)\,v_\Delta, \qquad v_2 = \cos(\alpha/2)\,\bar{v} - \sin(\alpha/2)\,v_\Delta.$$

For a unit vector x, we define its teacher force by

$$T(x) = \big(\pi - \angle(x, v_1)\big)v_1 + \big(\pi - \angle(x, v_2)\big)v_2.$$

The teacher terms in the ith equation of (3.3) are $T(\bar{w}_i)$. The proof of Lemma 3.6 relies on the following comparison between the directions of $T(x)$ and x.

Lemma C.3 (Direction of the teacher force)**.** *Assume $0 < \alpha \le \alpha_\star$, and let $x \in \mathbb{R}^d$ satisfy $\|x\|_2 = 1$. Write $p = \langle x, \bar{v}\rangle$, $q = \langle x, v_\Delta\rangle$, and $x_P = p\bar{v} + qv_\Delta$. Then $\langle T(x), \bar{v}\rangle > 0$ and $\operatorname{sgn}\langle T(x), v_\Delta\rangle = \operatorname{sgn} q$. If $p > 0$ and $q \ne 0$, the teacher force lies strictly closer to the bisector than x_P:*

$$0 < \frac{\langle T(x), v_\Delta\rangle}{\langle T(x), \bar{v}\rangle} < \frac{q}{p} \quad \text{if } q > 0, \qquad \frac{q}{p} < \frac{\langle T(x), v_\Delta\rangle}{\langle T(x), \bar{v}\rangle} < 0 \quad \text{if } q < 0.$$

Proof of Lemma C.3. We write $H(t) = \pi - \arccos t$ for $-1 \le t \le 1$. This function is strictly increasing and nonnegative, and it vanishes only at $t = -1$. Since x and v_j are unit vectors, $\pi - \angle(x, v_j) = H(\langle x, v_j\rangle)$ for $j = 1, 2$, and the decomposition of v_1, v_2 in the basis $\bar{v}, v_\Delta$ gives

$$\langle x, v_1\rangle = p\cos(\alpha/2) + q\sin(\alpha/2), \qquad \langle x, v_2\rangle = p\cos(\alpha/2) - q\sin(\alpha/2).$$

Substituting into the definition of $T(x)$ and taking inner products with $\bar{v}$ and v_Δ yields

$$\langle T(x), \bar{v}\rangle = \cos(\alpha/2)\,\big(H(p\cos(\alpha/2)+q\sin(\alpha/2)) + H(p\cos(\alpha/2)-q\sin(\alpha/2))\big),$$
$$\langle T(x), v_\Delta\rangle = \sin(\alpha/2)\,\big(H(p\cos(\alpha/2)+q\sin(\alpha/2)) - H(p\cos(\alpha/2)-q\sin(\alpha/2))\big).$$

Since $\cos(\alpha/2) > 0$ and $H \geq 0$, the first quantity is nonnegative. It vanishes only if $\langle x, v_1\rangle = \langle x, v_2\rangle = -1$, that is, only if $x = -v_1 = -v_2$, which is impossible because $v_1 \neq v_2$. Hence $\langle T(x), \bar{v}\rangle > 0$. In the second quantity, the two arguments of H differ by $2q\sin(\alpha/2)$, and $\sin(\alpha/2) > 0$, so the strict monotonicity of H shows that $\langle T(x), v_\Delta\rangle$ has the sign of q.

It remains to prove the bounds on the ratio when $p > 0$ and $q \neq 0$. By reflection across the $\bar{v}$-axis, we may assume $q > 0$. The lower bound then follows from the sign of $\langle T(x), v_\Delta\rangle$, and since $\langle T(x), \bar{v}\rangle > 0$ and $p > 0$, the upper bound is equivalent to $q\langle T(x), \bar{v}\rangle - p\langle T(x), v_\Delta\rangle > 0$. We write x_P in polar form, $x_P = \rho(\cos\varphi\,\bar{v} + \sin\varphi\, v_\Delta)$ with $0 < \rho \leq 1$ and $0 < \varphi < \pi/2$, so that $p = \rho\cos\varphi$ and $q = \rho\sin\varphi$. Then

$$p\cos(\alpha/2) + q\sin(\alpha/2) = \rho\cos(\varphi - \alpha/2), \qquad p\cos(\alpha/2) - q\sin(\alpha/2) = \rho\cos(\varphi + \alpha/2).$$

Substituting these identities into the formulas for $\langle T(x), \bar{v}\rangle$ and $\langle T(x), v_\Delta\rangle$ and using the angle-addition formulas, we obtain

$$\begin{aligned}
\frac{q\langle T(x), \bar{v}\rangle - p\langle T(x), v_\Delta\rangle}{\rho} &= \cos(\alpha/2)\Big[H\big(\rho\cos(\varphi-\alpha/2)\big) + H\big(\rho\cos(\varphi+\alpha/2)\big)\Big]\sin\varphi \\
&\quad - \sin(\alpha/2)\Big[H\big(\rho\cos(\varphi-\alpha/2)\big) - H\big(\rho\cos(\varphi+\alpha/2)\big)\Big]\cos\varphi \\
&= H\big(\rho\cos(\varphi-\alpha/2)\big)\big(\cos(\alpha/2)\sin\varphi - \sin(\alpha/2)\cos\varphi\big) \\
&\quad + H\big(\rho\cos(\varphi+\alpha/2)\big)\big(\cos(\alpha/2)\sin\varphi + \sin(\alpha/2)\cos\varphi\big) \\
&= H\big(\rho\cos(\varphi-\alpha/2)\big)\sin(\varphi-\alpha/2) + H\big(\rho\cos(\varphi+\alpha/2)\big)\sin(\varphi+\alpha/2).
\end{aligned}$$

If $\varphi \geq \alpha/2$, both terms on the right are nonnegative and the second is positive, so the expression is positive.

Suppose now that $0 < \varphi < \alpha/2$, and set $x_0 = \alpha/2 - \varphi$ and $y_0 = \alpha/2 + \varphi$, so that $0 < x_0 < y_0 < \alpha < \pi$. Since $\sin(\varphi - \alpha/2) = -\sin x_0$ and $\sin(\varphi + \alpha/2) = \sin y_0$, the expression is positive exactly when

$$H(\rho\cos x_0)\sin x_0 < H(\rho\cos y_0)\sin y_0. \tag{C.5}$$

To compare the two sides, we write $R(\rho) = H(\rho\cos x_0)/H(\rho\cos y_0)$ and $K(u) = u/\big(\sqrt{1-u^2}\,H(u)\big)$. For $-1 < u < 1$,

$$K'(u) = \frac{H(u) - u\sqrt{1-u^2}}{H(u)^2(1-u^2)^{3/2}}.$$

The numerator is positive: for $u < 0$ both of its terms are positive, and for $u \geq 0$ we have $H(u) \geq \pi/2 > 1/2 \geq u\sqrt{1-u^2}$. Thus K is strictly increasing. Since $\cos x_0 > \cos y_0$, it follows that

$$\frac{\mathrm{d}}{\mathrm{d}\rho}\log R(\rho) = \frac{K(\rho\cos x_0) - K(\rho\cos y_0)}{\rho} > 0,$$

and therefore

$$R(\rho) \leq R(1) = \frac{\pi - x_0}{\pi - y_0}. \tag{C.6}$$

It therefore suffices to show that $(\pi - x_0)\sin x_0 < (\pi - y_0)\sin y_0$. Using $x_0 = \alpha/2 - \varphi$ and $y_0 = \alpha/2 + \varphi$, we have

$$(\pi - y_0)\sin y_0 - (\pi - x_0)\sin x_0 = 2\big((\pi - \alpha/2)\cos(\alpha/2)\sin\varphi - \varphi\sin(\alpha/2)\cos\varphi\big).$$

By Lemma C.1, $(\pi - \alpha/2)\cos(\alpha/2) \geq \sin(\alpha/2)$, so the right-hand side is at least $2\sin(\alpha/2)(\sin\varphi - \varphi\cos\varphi)$, which is positive because $\tan\varphi > \varphi$. Combined with (C.6), this gives (C.5), and hence the upper bound for $q > 0$. The case $q < 0$ follows by reflection. ∎

C.2.2 Elimination of the splitting coordinate

Proof of Lemma 3.6. By Proposition 3.5, there exists a unit vector $\xi \perp \operatorname{span}\{\bar{v}, v_\Delta\}$ such that $\bar{w}_1, \bar{w}_2 \in \operatorname{span}\{\bar{v}, v_\Delta, \xi\}$. We write $\bar{w}_i = p_i\bar{v} + q_i v_\Delta + \eta_i \xi$ and $T_i = T(\bar{w}_i)$ for $i = 1, 2$. Since one of the students lies outside the teacher plane, at least one η_i is nonzero. By Lemma C.3,

$$\langle T_i, \bar{v}\rangle > 0, \qquad \operatorname{sgn}\langle T_i, v_\Delta\rangle = \operatorname{sgn} q_i. \tag{C.7}$$

Each T_i lies in the teacher plane, so it has no ξ-component.

We first show that $0 < \gamma < \pi$. If $\gamma = 0$, then $\bar{w}_1 = \bar{w}_2$, and the first equation in (3.3) becomes $T_1 + (a_1 - \pi\|w_2\|_2)\bar{w}_1 = 0$. Its ξ-coordinate shows that the coefficient of $\bar{w}_1$ vanishes, so $T_1 = 0$, which contradicts (C.7). If $\gamma = \pi$, then $\bar{w}_2 = -\bar{w}_1$, both η_i are nonzero, and the equations become $T_i + a_i\bar{w}_i = 0$. Their ξ-coordinates give $a_i = 0$, and hence again $T_i = 0$. Thus

$$0 < \gamma < \pi. \tag{C.8}$$

We next show that $\eta_1\eta_2 \neq 0$. If, for instance, $\eta_1 = 0$, then $\eta_2 \neq 0$, and the ξ-coordinate of the first equation in (3.3) gives $(\pi - \gamma)\|w_2\|_2\eta_2 = 0$, which contradicts (C.8). The case $\eta_2 = 0$ is the same.

Let $t = \eta_2/\eta_1$. The ξ-coordinates of the two equations in (3.3) give $a_1 = (\pi - \gamma)\|w_2\|_2 t$ and $a_2 = (\pi - \gamma)\|w_1\|_2/t$, and substituting these values back into (3.3) yields

$$\bar{w}_2 = t\bar{w}_1 + \frac{1}{(\pi - \gamma)\|w_2\|_2}T_1, \qquad T_2 = -\frac{\|w_1\|_2}{t\|w_2\|_2}T_1. \tag{C.9}$$

Since T_1 and T_2 both have positive $\bar{v}$-coordinate, the coefficient in the second identity is positive. Hence $t < 0$, and T_2 is a positive multiple of T_1. By (C.7), the numbers q_1 and q_2 are then either both zero or both nonzero with the same sign.

It remains to rule out the second possibility. Suppose that q_1 and q_2 are nonzero with the same sign. By reflection across the $\bar{v}$-axis, we may assume $q_1, q_2 > 0$, so that $\langle T_1, v_\Delta\rangle > 0$. Taking the $\bar{v}$- and v_Δ-coordinates of the first identity in (C.9) gives

$$p_2 = tp_1 + \frac{\langle T_1, \bar{v}\rangle}{(\pi - \gamma)\|w_2\|_2}, \quad q_2 = tq_1 + \frac{\langle T_1, v_\Delta\rangle}{(\pi - \gamma)\|w_2\|_2}. \tag{C.10}$$

If $p_1 \leq 0$, the first identity gives $p_2 > 0$, because $t < 0$ and $\langle T_1, \bar{v}\rangle > 0$. After exchanging the students if necessary, we may therefore assume $p_1 > 0$. Then Lemma C.3, applied to $\bar{w}_1$, gives $\langle T_1, v_\Delta\rangle/\langle T_1, \bar{v}\rangle < q_1/p_1$, that is, $q_1\langle T_1, \bar{v}\rangle - p_1\langle T_1, v_\Delta\rangle > 0$. Substituting (C.10), we obtain

$$\begin{aligned} q_2\langle T_1, \bar{v}\rangle - p_2\langle T_1, v_\Delta\rangle &= \langle T_1, \bar{v}\rangle\left(tq_1 + \frac{\langle T_1, v_\Delta\rangle}{(\pi - \gamma)\|w_2\|_2}\right) - \langle T_1, v_\Delta\rangle\left(tp_1 + \frac{\langle T_1, \bar{v}\rangle}{(\pi - \gamma)\|w_2\|_2}\right) \\ &= t(q_1\langle T_1, \bar{v}\rangle - p_1\langle T_1, v_\Delta\rangle) < 0, \end{aligned} \tag{C.11}$$

where the last inequality uses $t < 0$. Since $\langle T_1, \bar{v}\rangle$, $\langle T_1, v_\Delta\rangle$, and q_2 are positive, this inequality forces $p_2 > 0$. We may then apply Lemma C.3 to $\bar{w}_2$. Since T_2 is a positive multiple of T_1, it gives

$$\frac{\langle T_1, v_\Delta\rangle}{\langle T_1, \bar{v}\rangle} = \frac{\langle T_2, v_\Delta\rangle}{\langle T_2, \bar{v}\rangle} < \frac{q_2}{p_2},$$

that is, $q_2\langle T_1, \bar{v}\rangle - p_2\langle T_1, v_\Delta\rangle > 0$. This contradicts (C.11). Hence $q_1 = q_2 = 0$, and $w_1, w_2 \in \operatorname{span}\{\bar{v}, \xi\}$. ∎

C.3 Out-of-plane stationary points

By Lemma 3.6, a stationary point with a student outside the teacher plane lies in $\operatorname{span}\{\bar{v}, \xi\}$. For such points, we first show that the two students are mirror images across the teacher plane, then reduce stationarity to the equation $F_\alpha(\kappa) = 0$, and finally find a direction of negative curvature. Together these steps prove Lemma 3.7.

C.3.1 A scalar inequality

In the proof of Lemma C.5 below, an out-of-plane stationary point that is not symmetric across the teacher plane leads to two scalar identities. The following inequality shows that these identities are incompatible.

Lemma C.4 (A scalar inequality)**.** *Let* $0 < t < m < \pi/2$*, and put*

$$X = t + \frac{2(\pi - 2m)}{2m - \sin 2m}\big(t - \cos^2 m \tan t\big), \qquad Y = t + \frac{2(\pi - 2m)}{2m + \sin 2m}\big(t + \sin^2 m \cot t\big).$$

If $\lambda > 0$ *satisfies* $\lambda X = \pi/2 + \arctan(\lambda \tan t / \tan^2 m)$*, then* $\lambda X > Y/\lambda$*.*

Proof of Lemma C.4. We fix m and write $a = m - \sin m \cos m$, $b = m + \sin m \cos m$, $P(t) = t - \cos^2 m \tan t$, and $Q(t) = t + \sin^2 m \cot t$. Then $a, b > 0$, $X = t + (\pi - 2m)P(t)/a$, and $Y = t + (\pi - 2m)Q(t)/b$. We have $X > 0$ and $Y > 0$, since $P(t) > t - \cos^2 t \tan t = t - \sin t \cos t > 0$ and $Q(t) > 0$.

We first collect three bounds on P and Q. Since $P'(t) = 1 - \cos^2 m \sec^2 t > 0$ and $Q'(t) = 1 - \sin^2 m \csc^2 t < 0$ on $0 < t < m$, and since $P(m) = a$ and $Q(m) = b$, we have $0 < P(t) < a$ and $Q(t) > b$. The product $P(t)Q(t)$ is also increasing on $(0, m)$, because

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(P(t)Q(t)\big) = \frac{\tan^2 m - \tan^2 t}{1 + \tan^2 m}\left[Q(t) - \frac{P(t)}{\tan^2 t}\right] = \frac{\tan^2 m - \tan^2 t}{1 + \tan^2 m}\,\frac{\tan t - t + t\tan^2 t}{\tan^2 t} > 0.$$

Hence $P(t)Q(t) < P(m)Q(m) = ab$. Finally, the function $t(P(t)/a + Q(t)/b)$ equals $2m$ at $t = m$, and we claim that it is increasing on $(0, m)$, so that $t(P(t)/a + Q(t)/b) < 2m$. Its derivative is

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[t\left(\frac{P(t)}{a} + \frac{Q(t)}{b}\right)\right] = \frac{2t - \cos^2 m(\tan t + t\sec^2 t)}{a} + \frac{2t + \sin^2 m(\cot t - t\csc^2 t)}{b}. \tag{C.12}$$

For the first numerator in (C.12), the inequalities $0 < t < m < \pi/2$ give $\cos^2 m < \cos^2 t$. Since $\tan t + t\sec^2 t > 0$, it follows that

$$2t - \cos^2 m(\tan t + t\sec^2 t) > 2t - \cos^2 t(\tan t + t\sec^2 t) = t - \sin t \cos t > 0.$$

For the second numerator, the inequality $t - \sin t \cos t > 0$ gives $\cot t - t\csc^2 t = (\sin t \cos t - t)/\sin^2 t < 0$. Since $0 < \sin^2 m < 1$, multiplying this negative quantity by $\sin^2 m$ makes it strictly larger. Therefore

$$2t + \sin^2 m(\cot t - t\csc^2 t) > 2t + \cot t - t\csc^2 t = \frac{\sin t \cos t - t\cos 2t}{\sin^2 t} > 0.$$

For the last inequality, the numerator vanishes at $t = 0$ and its derivative is $2t \sin 2t > 0$ on $(0, \pi/2)$. Both numerators in (C.12) are therefore positive, which proves the claim.

Expanding XY and using $P(t)Q(t) < ab$ and $t(P(t)/a + Q(t)/b) < 2m$, we obtain

$$\begin{aligned}(\pi - m)^2 - XY &= (\pi - m)^2 - \left(t + (\pi - 2m)\frac{P(t)}{a}\right)\left(t + (\pi - 2m)\frac{Q(t)}{b}\right) \\ &= m^2 - t^2 + (\pi - 2m)\left[2m - t\left(\frac{P(t)}{a} + \frac{Q(t)}{b}\right)\right] + (\pi - 2m)^2\left(1 - \frac{P(t)Q(t)}{ab}\right) \\ &> 0.\end{aligned}$$

Therefore $\sqrt{XY} < \pi - m$.

We also need the bound

$$\frac{t(aQ(t) - bP(t))}{ab - P(t)Q(t)} < \frac{\pi}{2}, \tag{C.13}$$

in which the denominator is positive and so is the numerator, since $aQ(t) - bP(t) > ab - bP(t) = b(a - P(t)) > 0$. Because $b = m + \frac{1}{2}\sin 2m < \pi/2$, it is enough to show that the left-hand side is less than b, that is,

$$b(ab - P(t)Q(t)) - t(aQ(t) - bP(t)) > 0.$$

A direct expansion gives

$$b(ab - P(t)Q(t)) - t(aQ(t) - bP(t)) = bm^2 - at^2 - 2m\sin^2 m\, t\cot t. \tag{C.14}$$

At $t = m$, the right-hand side of (C.14) is zero. Its derivative with respect to t is

$$2t\left[m\sin^2 m\,\frac{t - \sin t\cos t}{t\sin^2 t} - a\right]. \tag{C.15}$$

For $0 < u < \pi/2$, the function $u \mapsto (u - \sin u \cos u)/(u \sin^2 u)$ is strictly increasing, since

$$\frac{\mathrm{d}}{\mathrm{d}u}\left[\frac{u - \sin u \cos u}{u \sin^2 u}\right] = \frac{\cos u}{\sin^3 u}\left[\frac{\tan u}{u} + \left(\frac{\sin u}{u}\right)^2 - 2\right] > 0$$

by Wilker's inequality (Sumner et al., 1991). Since $t < m$, the bracketed factor $m \sin^2 m\,(t - \sin t \cos t)/(t \sin^2 t) - a$ in (C.15) is strictly less than $m \sin^2 m\,\frac{m - \sin m \cos m}{m \sin^2 m} - a = 0$. Thus the derivative in (C.15) is negative on $(0, m)$. Since the right-hand side of (C.14) vanishes at $t = m$, it is positive for $0 < t < m$, which proves (C.13).

The conclusion of the lemma will follow from the inequality

$$\sqrt{XY} + \arctan\left(\frac{\tan^2 m}{\tan t}\sqrt{\frac{X}{Y}}\right) < \pi. \tag{C.16}$$

If $\sqrt{XY} \le \pi/2$, this holds because the arctangent is less than $\pi/2$. We may therefore assume $\sqrt{XY} > \pi/2$. Since $\sqrt{XY} < \pi - m$, we then have $m < \pi - \sqrt{XY} < \pi/2$. The tangent function is strictly convex on $(0, \pi/2)$, so $\tan u > \tan m + \sec^2 m\,(u - m)$ for every $u \in (0, \pi/2)$ with $u \ne m$. Taking $u = \pi - \sqrt{XY}$ and using $\sec^2 m = 1 + \tan^2 m$ gives $\tan(\pi - \sqrt{XY}) > \tan m + (1 + \tan^2 m)(\pi - \sqrt{XY} - m)$. Hence (C.16) follows once we show

$$\frac{\tan^2 m}{\tan t}\sqrt{\frac{X}{Y}} \le \tan m + (1 + \tan^2 m)(\pi - \sqrt{XY} - m). \tag{C.17}$$

Multiplying this inequality by $\cos^2 m$ and using $\sqrt{XY} = Y\sqrt{X/Y}$, we can rewrite it as $\sqrt{X/Y}\,(Y + \sin^2 m \cot t) \le \pi - m + \sin m \cos m$. Since $Y + \sin^2 m \cot t$ equals $t + \sin^2 m \cot t$ times $(\pi - m + \sin m \cos m)/(m + \sin m \cos m)$, this inequality is equivalent to

$$X\big(t + \sin^2 m \cot t\big)^2 \le Y(m + \sin m \cos m)^2. \tag{C.18}$$

To prove (C.18) when $\sqrt{XY} > \pi/2$, we fix t and m and set, for $\mu \ge \pi - 2m$, $X_\mu = t + \mu P(t)/a$ and $Y_\mu = t + \mu Q(t)/b$. At $\mu = \pi - 2m$, these are the original quantities X and Y, and (C.18) reads $Y_{\pi-2m} b^2 - X_{\pi-2m} Q(t)^2 \ge 0$. The expression $Y_\mu b^2 - X_\mu Q(t)^2$ is affine in μ, with slope

$$\frac{\mathrm{d}}{\mathrm{d}\mu}\left(Y_\mu b^2 - X_\mu Q(t)^2\right) = \frac{Q(t)(ab - P(t)Q(t))}{a} > 0.$$

Suppose that (C.18) fails. Then this affine expression is negative at $\mu = \pi - 2m$, so it vanishes at exactly one $\mu_* > \pi - 2m$, where $X_{\mu_*} Q(t)^2 = Y_{\mu_*} b^2$. Solving this equation gives $\mu_* = ta\big(Q(t)^2 - b^2\big)/\big(Q(t)(ab - P(t)Q(t))\big)$, and therefore

$$\sqrt{X_{\mu_*} Y_{\mu_*}} = \frac{b}{Q(t)}\,Y_{\mu_*} = \frac{tb}{Q(t)} + \mu_* = t\,\frac{aQ(t) - bP(t)}{ab - P(t)Q(t)} < \frac{\pi}{2}$$

by (C.13). Since X_μ and Y_μ increase with μ, this gives $\sqrt{XY} < \sqrt{X_{\mu_*} Y_{\mu_*}} < \pi/2$, contrary to our assumption. Hence (C.18) holds, and so does (C.17). Together with the tangent-line bound, it gives $(\tan^2 m/\tan t)\sqrt{X/Y} < \tan(\pi - \sqrt{XY})$, and since $0 < \pi - \sqrt{XY} < \pi/2$, taking arctangents proves (C.16).

We now derive the lemma from (C.16). Since $\arctan r + \arctan(1/r) = \pi/2$ for $r > 0$, the inequality (C.16) can be written as

$$\sqrt{XY} < \frac{\pi}{2} + \arctan\left(\frac{\tan t}{\tan^2 m}\sqrt{\frac{Y}{X}}\right),$$

and dividing by $\sqrt{Y/X}$ gives

$$\frac{\frac{\pi}{2} + \arctan\left(\frac{\tan t}{\tan^2 m}\sqrt{\frac{Y}{X}}\right)}{\sqrt{Y/X}} > X. \tag{C.19}$$

For fixed t and m, the function $r \mapsto \big(\pi/2 + \arctan(r \tan t / \tan^2 m)\big)/r$ on $r > 0$ is strictly decreasing, because its derivative is

$$\frac{\frac{r \tan t/\tan^2 m}{1+(r \tan t/\tan^2 m)^2} - \big(\frac{\pi}{2} + \arctan(r \tan t/\tan^2 m)\big)}{r^2} < 0.$$

By the hypothesis of the lemma, this function takes the value X at $r = \lambda$, while by (C.19) its value at $r = \sqrt{Y/X}$ is larger than X. Hence $\lambda > \sqrt{Y/X}$, and consequently $\lambda X > \sqrt{XY} > Y/\lambda$.

■

C.3.2 Mirror symmetry

We now use Lemma C.4 to show that the two students are mirror images across the teacher plane.

Lemma C.5 (Out-of-plane symmetry)**.** *Let ξ be any unit vector orthogonal to $\operatorname{span}\{v_1, v_2\}$. Assume $w_1, w_2 \neq 0$ satisfy (3.3), lie in $\operatorname{span}\{\bar{v}, \xi\}$, and do not both lie in the teacher plane. Then, after possibly replacing ξ by $-\xi$, there exist $r > 0$ and $\kappa \in (0, \pi/2)$ such that*

$$w_1 = r(\cos\kappa\, \bar{v} + \sin\kappa\, \xi), \qquad w_2 = r(\cos\kappa\, \bar{v} - \sin\kappa\, \xi).$$

Proof of Lemma C.5. We use coordinates with respect to the orthonormal basis $\bar{v}, v_\Delta, \xi$, so that $v_1 = (\cos(\alpha/2), \sin(\alpha/2), 0)$ and $v_2 = (\cos(\alpha/2), -\sin(\alpha/2), 0)$. Since $w_1, w_2 \in \operatorname{span}\{\bar{v}, \xi\}$, we can write $\bar{w}_i = (x_i, 0, z_i)$. Then $\langle \bar{w}_i, v_1 \rangle = \langle \bar{w}_i, v_2 \rangle$, so each student makes the same angle with both teachers; we denote this angle by θ_i. The system (3.3) becomes

$$(\pi - \theta_i)(v_1 + v_2) + \big(2\sin\theta_i - \pi\|w_i\|_2 - \sin\gamma\, \|w_j\|_2\big)\bar{w}_i - (\pi - \gamma)\|w_j\|_2 \bar{w}_j = 0, \qquad \{i, j\} = \{1, 2\}. \tag{C.20}$$

We first show that $z_1 z_2 \neq 0$. By hypothesis, z_1 and z_2 are not both zero. If $z_1 = 0$ and $z_2 \neq 0$, then $\bar{w}_1$ and $\bar{w}_2$ are neither parallel nor antiparallel, so $0 < \gamma < \pi$, and the third coordinate of (C.20) for $i = 1$ gives $(\pi - \gamma)\|w_2\|_2 z_2 = 0$, a contradiction. The case $z_2 = 0$ is the same.

We next show that $0 < \gamma < \pi$. If γ were 0 or π, then $\bar{w}_2 = \pm\bar{w}_1$, and (C.20) for $i = 1$ would express $(\pi - \theta_1)(v_1 + v_2)$, a nonzero multiple of $\bar{v}$, as a multiple of $\bar{w}_1$. This is impossible because $z_1 \neq 0$.

To remove the term in $\bar{w}_i$ from the ith equation, we take its inner product with $(z_i, 0, -x_i)$, which is orthogonal to $\bar{w}_i$. This gives

$$\begin{aligned} 2\cos(\alpha/2)(\pi - \theta_1) z_1 + (\pi - \gamma)\|w_2\|_2 (x_1 z_2 - z_1 x_2) &= 0, \\ 2\cos(\alpha/2)(\pi - \theta_2) z_2 - (\pi - \gamma)\|w_1\|_2 (x_1 z_2 - z_1 x_2) &= 0. \end{aligned} \tag{C.21}$$

The coefficients $2\cos(\alpha/2)(\pi - \theta_i)$ and $(\pi - \gamma)\|w_j\|_2$ are positive. The first identity therefore gives $\operatorname{sgn}(x_1 z_2 - z_1 x_2) = -\operatorname{sgn} z_1$, and the second gives $\operatorname{sgn}(x_1 z_2 - z_1 x_2) = \operatorname{sgn} z_2$, so z_1 and z_2 have opposite signs. Replacing ξ by $-\xi$ if necessary, we may assume $z_1 > 0 > z_2$ and write $\bar{w}_1 = (\cos p, 0, \sin p)$ and $\bar{w}_2 = (\cos q, 0, -\sin q)$ with $p, q \in (0, \pi)$. Then $x_1 z_2 - z_1 x_2 = -\sin(p + q)$, which is negative by the first identity in (C.21). Hence $p + q < \pi$ and $\gamma = p + q$, and (C.21) becomes

$$\begin{aligned} (\pi - p - q)\sin(p + q)\|w_2\|_2 &= 2\cos(\alpha/2)(\pi - \theta_1)\sin p, \\ (\pi - p - q)\sin(p + q)\|w_1\|_2 &= 2\cos(\alpha/2)(\pi - \theta_2)\sin q. \end{aligned} \tag{C.22}$$

The teacher angles satisfy $\cos\theta_1 = \cos(\alpha/2)\cos p$ and $\cos\theta_2 = \cos(\alpha/2)\cos q$. Taking the inner product of the ith equation in (C.20) with $\bar{w}_i$ gives

$$\begin{aligned} \pi\|w_1\|_2 + \big(\sin(p + q) + (\pi - p - q)\cos(p + q)\big)\|w_2\|_2 &= 2\big(\sin\theta_1 + (\pi - \theta_1)\cos\theta_1\big), \\ \big(\sin(p + q) + (\pi - p - q)\cos(p + q)\big)\|w_1\|_2 + \pi\|w_2\|_2 &= 2\big(\sin\theta_2 + (\pi - \theta_2)\cos\theta_2\big). \end{aligned} \tag{C.23}$$

Eliminating $\|w_1\|_2, \|w_2\|_2$ from (C.23) and (C.22) gives

$$\begin{aligned} &(\pi - p - q)\sin(p + q)\big(\sin\theta_1 + (\pi - \theta_1)\cos\theta_1\big) \\ &\quad = \big(\sin(p + q) + (\pi - p - q)\cos(p + q)\big)\cos(\alpha/2)(\pi - \theta_1)\sin p + \pi\cos(\alpha/2)(\pi - \theta_2)\sin q, \end{aligned} \tag{C.24}$$

$$(\pi - p - q)\sin(p + q)\big(\sin\theta_2 + (\pi - \theta_2)\cos\theta_2\big)$$

$$= \pi\cos(\alpha/2)(\pi-\theta_1)\sin p + \big(\sin(p+q) + (\pi-p-q)\cos(p+q)\big)\cos(\alpha/2)(\pi-\theta_2)\sin q. \tag{C.25}$$

We now show that (C.24) and (C.25) force $p = q$. Suppose instead that $p \neq q$; by symmetry we may assume $p > q$, so that

$$0 < \frac{p-q}{2} < \frac{p+q}{2} < \frac{\pi}{2}. \tag{C.26}$$

The formula (C.1) for $\theta_\alpha(\kappa)$ makes sense for every $\kappa \in (0,\pi)$, and $\theta_1 = \theta_\alpha(p)$, $\theta_2 = \theta_\alpha(q)$. Direct differentiation gives

$$\theta'_\alpha(\kappa) = \frac{\cos(\alpha/2)\sin\kappa}{\sqrt{1-\cos^2(\alpha/2)\cos^2\kappa}} > 0, \qquad 1-\theta'_\alpha(\kappa)^2 = \frac{\sin^2(\alpha/2)}{1-\cos^2(\alpha/2)\cos^2\kappa} > 0,$$

so $0 < \theta'_\alpha(\kappa) < 1$ on $(0,\pi)$. By the mean value theorem, there exists $\zeta \in (q,p)$ such that $\theta_1 - \theta_2 = \theta_\alpha(p) - \theta_\alpha(q) = \theta'_\alpha(\zeta)(p-q)$. Since $0 < \theta'_\alpha(\zeta) < 1$ and $p > q$, it follows that

$$0 < \frac{\theta_1-\theta_2}{2} < \frac{p-q}{2}. \tag{C.27}$$

Moreover, $\cos\theta_1 + \cos\theta_2 = \cos(\alpha/2)(\cos p + \cos q) > 0$ because $p+q < \pi$, so $(\theta_1+\theta_2)/2 < \pi/2$.

Adding and subtracting the identities $\cos\theta_1 = \cos(\alpha/2)\cos p$ and $\cos\theta_2 = \cos(\alpha/2)\cos q$, we obtain

$$\begin{aligned}
\cos\frac{\theta_1+\theta_2}{2}\cos\frac{\theta_1-\theta_2}{2} &= \cos(\alpha/2)\cos\frac{p+q}{2}\cos\frac{p-q}{2},\\
\sin\frac{\theta_1+\theta_2}{2}\sin\frac{\theta_1-\theta_2}{2} &= \cos(\alpha/2)\sin\frac{p+q}{2}\sin\frac{p-q}{2}.
\end{aligned} \tag{C.28}$$

Hence

$$\tan\frac{\theta_1+\theta_2}{2}\tan\frac{\theta_1-\theta_2}{2} = \tan\frac{p+q}{2}\tan\frac{p-q}{2}.$$

Since $0 < (\theta_1+\theta_2)/2 < \pi/2$, this gives

$$\pi - \frac{\theta_1+\theta_2}{2} = \frac{\pi}{2} + \arctan\left(\frac{\tan\frac{\theta_1-\theta_2}{2}}{\tan\frac{p+q}{2}\tan\frac{p-q}{2}}\right). \tag{C.29}$$

We now subtract (C.25) from (C.24) and write p, q, θ_1, θ_2 in terms of the half-sums $(p+q)/2$, $(\theta_1+\theta_2)/2$ and the half-differences $(p-q)/2$, $(\theta_1-\theta_2)/2$. This gives

$$\begin{aligned}
&(\pi-p-q)\sin(p+q)\Big[2\cos\frac{\theta_1+\theta_2}{2}\sin\frac{\theta_1-\theta_2}{2} - 2\left(\pi-\frac{\theta_1+\theta_2}{2}\right)\sin\frac{\theta_1+\theta_2}{2}\sin\frac{\theta_1-\theta_2}{2}\\
&\quad - (\theta_1-\theta_2)\cos\frac{\theta_1+\theta_2}{2}\cos\frac{\theta_1-\theta_2}{2}\Big]\\
&= \cos(\alpha/2)\big(\sin(p+q) + (\pi-p-q)\cos(p+q) - \pi\big)\Big[2\left(\pi-\frac{\theta_1+\theta_2}{2}\right)\cos\frac{p+q}{2}\sin\frac{p-q}{2}\\
&\quad - (\theta_1-\theta_2)\sin\frac{p+q}{2}\cos\frac{p-q}{2}\Big].
\end{aligned}$$

Using (C.28) in this identity and collecting the terms containing $\pi - (\theta_1+\theta_2)/2$, we get

$$\begin{aligned}
&\left(\pi-\frac{\theta_1+\theta_2}{2}\right)\big(p+q-\sin(p+q)\big)\cos\frac{p+q}{2}\tan\frac{p-q}{2}\\
&\quad = \sin\frac{p+q}{2}\left[\big(p+q-\sin(p+q)\big)\frac{\theta_1-\theta_2}{2} + 2(\pi-p-q)\left(\frac{\theta_1-\theta_2}{2} - \cos^2\frac{p+q}{2}\tan\frac{\theta_1-\theta_2}{2}\right)\right].
\end{aligned}$$

Therefore

$$\pi - \frac{\theta_1+\theta_2}{2} = \tan\frac{p+q}{2}\cot\frac{p-q}{2}\left[\frac{\theta_1-\theta_2}{2} + \frac{2(\pi-p-q)}{p+q-\sin(p+q)}\left(\frac{\theta_1-\theta_2}{2} - \cos^2\frac{p+q}{2}\tan\frac{\theta_1-\theta_2}{2}\right)\right]. \tag{C.30}$$

Adding (C.24) and (C.25) instead, we obtain in the same way

$$
\begin{aligned}
&(\pi - p - q)\sin(p+q)\Big[2\sin\frac{\theta_1+\theta_2}{2}\cos\frac{\theta_1-\theta_2}{2} + 2\left(\pi - \frac{\theta_1+\theta_2}{2}\right)\cos\frac{\theta_1+\theta_2}{2}\cos\frac{\theta_1-\theta_2}{2} \\
&\quad + (\theta_1-\theta_2)\sin\frac{\theta_1+\theta_2}{2}\sin\frac{\theta_1-\theta_2}{2}\Big] \\
&= \cos(\alpha/2)\big(\sin(p+q) + (\pi - p - q)\cos(p+q) + \pi\big)\Big[2\left(\pi - \frac{\theta_1+\theta_2}{2}\right)\sin\frac{p+q}{2}\cos\frac{p-q}{2} \\
&\quad - (\theta_1-\theta_2)\cos\frac{p+q}{2}\sin\frac{p-q}{2}\Big].
\end{aligned}
$$

Using (C.28) again and collecting $\pi - (\theta_1+\theta_2)/2$, we obtain

$$
\begin{aligned}
&\left(\pi - \frac{\theta_1+\theta_2}{2}\right)\big(p+q+\sin(p+q)\big)\sin\frac{p+q}{2}\cot\frac{p-q}{2} \\
&= \cos\frac{p+q}{2}\left[\big(p+q+\sin(p+q)\big)\frac{\theta_1-\theta_2}{2} + 2(\pi-p-q)\left(\frac{\theta_1-\theta_2}{2} + \sin^2\frac{p+q}{2}\cot\frac{\theta_1-\theta_2}{2}\right)\right].
\end{aligned}
$$

Thus

$$
\begin{aligned}
\pi - \frac{\theta_1+\theta_2}{2} = \cot\frac{p+q}{2}\tan\frac{p-q}{2}\Bigg[&\frac{\theta_1-\theta_2}{2} + \frac{2(\pi-p-q)}{p+q+\sin(p+q)} \\
&\times\left(\frac{\theta_1-\theta_2}{2} + \sin^2\frac{p+q}{2}\cot\frac{\theta_1-\theta_2}{2}\right)\Bigg].
\end{aligned}
\tag{C.31}
$$

We now apply Lemma C.4 with $t = (\theta_1-\theta_2)/2$, $m = (p+q)/2$, and $\lambda = \tan\frac{p+q}{2}\big/\tan\frac{p-q}{2}$. By (C.26) and (C.27), we have $0 < t < m < \pi/2$. In the notation of the lemma, (C.30) says that $\pi - (\theta_1+\theta_2)/2 = \lambda X$, and (C.31) says that $\pi - (\theta_1+\theta_2)/2 = Y/\lambda$. Combined with (C.29), the first identity shows that λ satisfies the hypothesis of the lemma, which then gives $\lambda X > Y/\lambda$. This contradicts $\lambda X = \pi - (\theta_1+\theta_2)/2 = Y/\lambda$. Hence $p = q$.

Thus $\bar{w}_1 = (\cos p, 0, \sin p)$ and $\bar{w}_2 = (\cos p, 0, -\sin p)$ with $0 < p < \pi/2$, so that $\theta_1 = \theta_2$ and $\gamma = 2p$. It remains to show that $\|w_1\|_2 = \|w_2\|_2$. Dividing the third coordinates of (C.20) by $\sin p$ gives

$$
\begin{aligned}
0 &= 2\sin\theta_1 - \pi\|w_1\|_2 + \big((\pi-\gamma) - \sin\gamma\big)\|w_2\|_2, \\
0 &= 2\sin\theta_1 - \pi\|w_2\|_2 + \big((\pi-\gamma) - \sin\gamma\big)\|w_1\|_2.
\end{aligned}
$$

Subtracting these two equations yields $(2\pi - \gamma - \sin\gamma)(\|w_1\|_2 - \|w_2\|_2) = 0$, and the coefficient is positive because $0 < \gamma < \pi$. Hence $\|w_1\|_2 = \|w_2\|_2$, and the lemma follows with $r = \|w_1\|_2$ and $\kappa = p$.

■

C.3.3 Exact scalar characterization

By Lemma C.5, an out-of-plane stationary point is determined by a common norm r and an angle κ from $\bar{v}$. It remains to find the pairs (r, κ) that give stationary points. We recall θ_α, r_α, and F_α from (C.1)–(C.3), and that κ_α is the unique root of F_α in $(0, \pi/2)$ by Lemma C.2.

Proof of the characterization in Lemma 3.7. By Lemma C.5, after possibly swapping the students or replacing ξ by $-\xi$, every pair in question has the form

$$
w_1 = r(\cos\kappa\,\bar{v} + \sin\kappa\,\xi), \qquad w_2 = r(\cos\kappa\,\bar{v} - \sin\kappa\,\xi) \tag{C.32}
$$

for some $r > 0$ and $0 < \kappa < \pi/2$, and every such pair lies in $\operatorname{span}\{\bar{v}, \xi\}$ outside the teacher plane. For the pair (C.32), all four teacher–student angles equal $\theta_\alpha(\kappa)$ by (C.1), and the angle between the students is $\gamma = 2\kappa$. We write

$$
\begin{aligned}
R(r,\kappa) &= 2\sin\theta_\alpha(\kappa) - r\big(\sin(2\kappa) + 2\kappa\big), \\
G_\alpha(r,\kappa) &= \big(\pi - \theta_\alpha(\kappa)\big)\cos(\alpha/2) - (\pi - 2\kappa)r\cos\kappa.
\end{aligned}
$$

Substituting (C.32) into (3.3) reduces its two vector equations to

$$\big(2G_\alpha(r,\kappa)+\cos\kappa\, R(r,\kappa)\big)\bar v+\sin\kappa\, R(r,\kappa)\xi=0,$$
$$\big(2G_\alpha(r,\kappa)+\cos\kappa\, R(r,\kappa)\big)\bar v-\sin\kappa\, R(r,\kappa)\xi=0.$$

Since $\sin\kappa>0$, the pair is stationary if and only if

$$R(r,\kappa)=0,\qquad G_\alpha(r,\kappa)=0.$$

Since $\sin(2\kappa)+2\kappa>0$, the first equation means $r=2\sin\theta_\alpha(\kappa)/(\sin(2\kappa)+2\kappa)=r_\alpha(\kappa)$. With this value of r, the second equation reads

$$0=G_\alpha(r_\alpha(\kappa),\kappa)=\big(\pi-\theta_\alpha(\kappa)\big)\cos(\alpha/2)-(\pi-2\kappa)r_\alpha(\kappa)\cos\kappa=F_\alpha(\kappa).$$

By Lemma C.2, the pair is therefore stationary if and only if $\kappa=\kappa_\alpha$ and $r=r_\alpha(\kappa_\alpha)$. ■

C.3.4 Negative curvature

To finish the proof of Lemma 3.7, we move the two students toward different teachers, along the direction $(v_\Delta,-v_\Delta)$, and show that the loss has negative curvature in this direction.

Proof of the strict-saddle claim in Lemma 3.7. Fix a stationary point (C.32) with $\kappa=\kappa_\alpha$ and $r=r_\alpha(\kappa_\alpha)$, and consider the perturbation $(v_\Delta,-v_\Delta)$. Since v_Δ is orthogonal to $\bar w_1$ and $\bar w_2$, the blocks in Lemma A.6 give, for $i=1,2$,

$$v_\Delta^\top Q_1(w_i,w_{3-i})v_\Delta=\frac{\sin(2\kappa)}{2\pi},$$
$$v_\Delta^\top Q_1(w_i,v_j)v_\Delta=\frac{\sin\theta_\alpha(\kappa)}{2\pi r}\left(1+\frac{\sin^2(\alpha/2)}{\sin^2\theta_\alpha(\kappa)}\right),\qquad j\in\{1,2\},$$
$$v_\Delta^\top Q_2(w_1,w_2)v_\Delta=\frac{\pi-2\kappa}{2\pi}.$$

Consequently,

$$(v_\Delta,-v_\Delta)^\top\nabla^2L(w_1,w_2)(v_\Delta,-v_\Delta)$$
$$=1+\frac{\sin(2\kappa)}{\pi}-\frac{2\sin\theta_\alpha(\kappa)}{\pi r}\left(1+\frac{\sin^2(\alpha/2)}{\sin^2\theta_\alpha(\kappa)}\right)-\frac{\pi-2\kappa}{\pi}.$$

Since $r=r_\alpha(\kappa)$, (C.2) gives $2\sin\theta_\alpha(\kappa)/r=\sin(2\kappa)+2\kappa$, and therefore

$$(v_\Delta,-v_\Delta)^\top\nabla^2L(w_1,w_2)(v_\Delta,-v_\Delta)=-\frac{\sin(2\kappa)+2\kappa}{\pi}\frac{\sin^2(\alpha/2)}{\sin^2\theta_\alpha(\kappa)}<0.$$

Thus every point of the out-of-plane family has a strictly negative Hessian direction. ■

C.4 In-plane stationary points

In the teacher plane, each student is described by its norm and its angle from $\bar v$, and stationarity reduces to four scalar equations. Solving them for $0<\alpha\le\alpha_\star$ proves Lemma 3.8.

C.4.1 Coordinates in the teacher plane

For $\varphi\in[-\pi,\pi)$, write $e(\varphi)=\cos\varphi\,\bar v+\sin\varphi\,v_\Delta$. Its derivative is the positively oriented unit tangent $e'(\varphi)=-\sin\varphi\,\bar v+\cos\varphi\,v_\Delta$. For $\varphi,\vartheta\in\mathbb{R}$, define their circular distance by

$$d_{\mathbb{S}^1}(\varphi,\vartheta):=\min_{k\in\mathbb{Z}}|\varphi-\vartheta+2\pi k|=\arccos\big(\cos(\varphi-\vartheta)\big)\in[0,\pi].$$

With l as in (A.1b), define

$$\mathcal{P}(\varphi)=2\pi\Big(l\big(d_{\mathbb{S}^1}(\varphi,\alpha/2)\big)+l\big(d_{\mathbb{S}^1}(\varphi,-\alpha/2)\big)\Big),$$
$$\mathcal{P}'(\varphi)=-\big(\pi-d_{\mathbb{S}^1}(\varphi,\alpha/2)\big)\sin(\varphi-\alpha/2)-\big(\pi-d_{\mathbb{S}^1}(\varphi,-\alpha/2)\big)\sin(\varphi+\alpha/2).$$

The function $\mathcal{P}$ is 2π-periodic, even, and continuously differentiable. In these coordinates, (3.3) splits into a radial and a tangential equation for each student.

Lemma C.6 (Planar stationarity equations)**.** *Let $0 < \alpha < \pi$, let $w_1, w_2 \in \operatorname{span}\{v_1, v_2\}$ be nonzero, and write*

$$w_i = \|w_i\|_2\, e(\varphi_i), \qquad \varphi_i \in [-\pi, \pi), \quad i = 1, 2,$$

and let $\gamma = d_{\mathbb{S}^1}(\varphi_1, \varphi_2)$. Then (w_1, w_2) is stationary if and only if

$$\pi\|w_1\|_2 + 2\pi l(\gamma)\|w_2\|_2 = \mathcal{P}(\varphi_1), \tag{C.33}$$

$$\pi\|w_2\|_2 + 2\pi l(\gamma)\|w_1\|_2 = \mathcal{P}(\varphi_2), \tag{C.34}$$

$$\mathcal{P}'(\varphi_1) = (\pi - \gamma)\|w_2\|_2 \sin(\varphi_2 - \varphi_1), \tag{C.35}$$

$$\mathcal{P}'(\varphi_2) = (\pi - \gamma)\|w_1\|_2 \sin(\varphi_1 - \varphi_2). \tag{C.36}$$

If in addition $0 < \alpha \le \alpha_\star$, then $\mathcal{P}'(\varphi) < 0$ for $0 < \varphi < \pi$ and $\mathcal{P}'(\varphi) > 0$ for $-\pi < \varphi < 0$. In particular, the zeros of $\mathcal{P}'$ are exactly the integer multiples of π.

Proof of Lemma C.6. The closed-form loss (A.10) gives

$$2\pi L = \frac{\pi}{2}(\|w_1\|_2^2 + \|w_2\|_2^2) + 2\pi\|w_1\|_2\|w_2\|_2\, l(\gamma) - \|w_1\|_2\mathcal{P}(\varphi_1) - \|w_2\|_2\mathcal{P}(\varphi_2) + \text{constant}.$$

Differentiating with respect to $\|w_1\|_2$ and $\|w_2\|_2$ gives (C.33) and (C.34).

For the tangential equations, we write $B_i = (\pi - \theta_{i1})v_1 + (\pi - \theta_{i2})v_2$. The direction equations (A.13)–(A.14) state that $(I - e(\varphi_i)e(\varphi_i)^\top)\big(B_i - (\pi - \gamma)w_{3-i}\big) = 0$. Since $e'(\varphi_i) \perp e(\varphi_i)$, taking the inner product with $e'(\varphi_i)$ removes the projection and gives $\langle e'(\varphi_i), B_i\rangle = (\pi - \gamma)\langle e'(\varphi_i), w_{3-i}\rangle$. To identify the left-hand side with $\mathcal{P}'(\varphi_i)$, let $a = \pm\alpha/2$. If $\varphi \not\equiv a$ and $\varphi \not\equiv a + \pi$ modulo 2π, then $\cos d_{\mathbb{S}^1}(\varphi, a) = \cos(\varphi - a)$, and differentiating gives $\sin d_{\mathbb{S}^1}(\varphi, a)\, \partial_\varphi d_{\mathbb{S}^1}(\varphi, a) = \sin(\varphi - a)$. Since $2\pi l'(t) = -(\pi - t)\sin t$ by (A.2) and $\langle e'(\varphi), e(a)\rangle = -\sin(\varphi - a)$, the chain rule gives

$$\begin{aligned}\frac{\partial}{\partial\varphi} 2\pi l\big(d_{\mathbb{S}^1}(\varphi, a)\big) &= 2\pi l'\big(d_{\mathbb{S}^1}(\varphi, a)\big)\frac{\partial}{\partial\varphi} d_{\mathbb{S}^1}(\varphi, a) = -\big(\pi - d_{\mathbb{S}^1}(\varphi, a)\big)\sin(\varphi - a) \\ &= \big(\pi - d_{\mathbb{S}^1}(\varphi, a)\big)\langle e'(\varphi), e(a)\rangle.\end{aligned}$$

With $v_1 = e(\alpha/2)$, $v_2 = e(-\alpha/2)$, $\theta_{i1} = d_{\mathbb{S}^1}(\varphi_i, \alpha/2)$, and $\theta_{i2} = d_{\mathbb{S}^1}(\varphi_i, -\alpha/2)$, it follows that

$$\mathcal{P}'(\varphi_i) = (\pi - \theta_{i1})\langle e'(\varphi_i), v_1\rangle + (\pi - \theta_{i2})\langle e'(\varphi_i), v_2\rangle = \langle e'(\varphi_i), B_i\rangle.$$

Since $\langle e'(\varphi_i), e(\varphi_{3-i})\rangle = \sin(\varphi_{3-i} - \varphi_i)$, this gives (C.35)–(C.36).

The computation above assumes that $e(\varphi_i) \neq \pm v_1, \pm v_2$, because $\varphi \mapsto d_{\mathbb{S}^1}(\varphi, a)$ is not differentiable at $\varphi \equiv a$ and $\varphi \equiv a + \pi$. The composition $\varphi \mapsto l(d_{\mathbb{S}^1}(\varphi, a))$ is nevertheless differentiable at these points. Near $\varphi \equiv a$ the circular distance is $|\varphi - a|$, and near $\varphi \equiv a + \pi$ it is $\pi - |\varphi - a - \pi|$. Since the formula (A.1b) for l extends smoothly to $\mathbb{R}$ and $l'(0) = l'(\pi) = 0$ by (A.2), both one-sided derivatives vanish. Hence the formula for $\mathcal{P}'(\varphi_i)$ holds for every φ_i, and the tangential equations remain valid also when $\gamma = 0$ or $\gamma = \pi$. Conversely, the gradient with respect to each student lies in the teacher plane, and the four equations say that its radial and tangential components vanish. Hence the four equations are equivalent to stationarity.

It remains to determine the sign of $\mathcal{P}'$ when $0 < \alpha \le \alpha_\star$. Since $\mathcal{P}$ is even, $\mathcal{P}'$ is odd, so it suffices to consider $0 < \varphi < \pi$. For $0 \le \varphi \le \alpha/2$, direct expansion gives

$$\mathcal{P}'(\varphi) = 2\sin\varphi\big(\varphi\cot\varphi\sin(\alpha/2) - (\pi - \alpha/2)\cos(\alpha/2)\big).$$

For $0 < \varphi \le \alpha/2$, this is negative, because $\varphi\cot\varphi < 1$ and $\sin(\alpha/2) \le (\pi - \alpha/2)\cos(\alpha/2)$ by Lemma C.1. For $\alpha/2 \le \varphi \le \pi - \alpha/2$,

$$\mathcal{P}'(\varphi) = 2\big((\alpha/2)\sin(\alpha/2)\cos\varphi - (\pi - \varphi)\cos(\alpha/2)\sin\varphi\big).$$

If $\varphi \ge \pi/2$, both terms are nonpositive and the second is negative. If $\alpha/2 \le \varphi < \pi/2$, then $(\pi - \varphi)\tan\varphi > (\alpha/2)\tan(\alpha/2)$ because $\pi - \varphi > \alpha/2$ and $\tan\varphi \ge \tan(\alpha/2)$; multiplying by $\cos(\alpha/2)\cos\varphi > 0$ shows again that $\mathcal{P}'(\varphi) < 0$. Finally, for $\pi - \alpha/2 \le \varphi \le \pi$, we put $x = \pi - \varphi \in [0, \alpha/2]$ and find

$$\mathcal{P}'(\varphi) = -2\big((\alpha/2)\cos(\alpha/2)\sin x + x\sin(\alpha/2)\cos x\big),$$

which is negative for $x > 0$ and vanishes at $x = 0$, that is, at $\varphi = \pi$. Together with $\mathcal{P}'(0) = 0$ and the oddness of $\mathcal{P}'$, this proves the claims about the sign and the zeros of $\mathcal{P}'$. ∎

C.4.2 Stationary points above the angle threshold

The assumption $\alpha \le \alpha_\star$ in the sign condition cannot be dropped. Above the threshold, $\mathcal{P}'$ vanishes inside $(0, \alpha/2)$, which gives the following stationary points.

Proposition C.7 (Additional stationary families above the threshold)**.** *Suppose $\alpha \in (\alpha_\star, \pi)$. There is a unique $\varphi_0 \in (0, \alpha/2)$ satisfying*

$$\varphi_0 \cot \varphi_0 = (\pi - \alpha/2)\cot(\alpha/2). \tag{C.37}$$

For $\sigma = \pm 1$ and every $s_1, s_2 > 0$ with $s_1 + s_2 = \mathcal{P}(\varphi_0)/\pi$, the pair $(w_1, w_2) = \big(s_1 e(\sigma\varphi_0), s_2 e(\sigma\varphi_0)\big)$ is stationary. These families are not among the configurations listed in Lemma 3.8.

Proof of Proposition C.7. The function $\varphi \mapsto \varphi\cot\varphi$ is strictly decreasing on $(0, \alpha/2)$, since its derivative $(\sin\varphi\cos\varphi - \varphi)/\sin^2\varphi$ is negative. Moreover, $\varphi\cot\varphi \to 1$ as $\varphi \downarrow 0$, and $(\alpha/2)\cot(\alpha/2) < (\pi - \alpha/2)\cot(\alpha/2) < 1$. Here the first inequality uses $\alpha < \pi$, and the second is equivalent to $\tan(\alpha/2) > \pi - \alpha/2$, which holds by Lemma C.1 because $\alpha > \alpha_\star$. Thus (C.37) has a unique root $\varphi_0 \in (0, \alpha/2)$.

The distances from φ_0 to the two teachers, $\alpha/2 - \varphi_0$ and $\alpha/2 + \varphi_0$, lie in $(0, \pi)$. Since l is strictly decreasing on $[0, \pi]$ by (A.2) and $l(\pi) = 0$, it follows that $\mathcal{P}(\varphi_0) > 0$, so there are positive s_1, s_2 with the stated sum.

By the formula for $\mathcal{P}'$ on $[0, \alpha/2]$ in the proof of Lemma C.6,

$$\mathcal{P}'(\varphi) = 2\sin\varphi\sin(\alpha/2)\,(\varphi\cot\varphi - (\pi - \alpha/2)\cot(\alpha/2))$$

for $0 < \varphi < \alpha/2$, so $\mathcal{P}'(\varphi_0) = 0$. For the pair in the proposition, $\gamma = 0$, so the tangential equations reduce to $\mathcal{P}'(\varphi_0) = 0$ and both radial equations reduce to $\pi(s_1 + s_2) = \mathcal{P}(\varphi_0)$. By Lemma C.6, the pair is stationary. Since $0 < \varphi_0 < \alpha/2$, these points are not parallel to $\pm\bar{v}$ and differ from (v_1, v_2), so they are not among the configurations in Lemma 3.8. ∎

C.4.3 Classification up to the angle threshold

We now return to $0 < \alpha \le \alpha_\star$, which we assume for the rest of this subsection. The points 0, $\pm\alpha/2$, and $\pm(\pi - \alpha/2)$ split $(-\pi, \pi)$ into the open intervals

$$\begin{aligned} &J_2 = (-\pi, -\pi + \alpha/2), \quad J_1 = (-\pi + \alpha/2, -\alpha/2), \quad J_0 = (-\alpha/2, 0),\\ &I_0 = (0, \alpha/2), \qquad\quad I_1 = (\alpha/2, \pi - \alpha/2), \qquad I_2 = (\pi - \alpha/2, \pi). \end{aligned}$$

The only symmetries of the loss that we use are the exchange of the two students and the reflection $\varphi \mapsto -\varphi$ across $\bar{v}$; the map $\varphi \mapsto \varphi + \pi$ is not a symmetry of the loss.

Proof of Lemma 3.8. As in Lemma C.6, we write $w_i = \|w_i\|_2\, e(\varphi_i)$ with $\varphi_i \in [-\pi, \pi)$. Both r_+ and r_- in (3.4) are positive; for r_-, this uses $\tan(\alpha/2) > \alpha/2$.

We first treat collinear directions. If $\gamma = 0$, the two directions coincide, and the tangential equations (C.35)–(C.36) give $\mathcal{P}'(\varphi_1) = 0$, so by Lemma C.6 the common direction is $\bar{v}$ or $-\bar{v}$. Since $2\pi l(0) = \pi$ and

$$\begin{aligned} \mathcal{P}(0) &= 2\big((\pi - \alpha/2)\cos(\alpha/2) + \sin(\alpha/2)\big) = \pi r_+,\\ \mathcal{P}(\pi) &= 2\big(\sin(\alpha/2) - (\alpha/2)\cos(\alpha/2)\big) = \pi r_-, \end{aligned}$$

the radial equations (C.33)–(C.34) give $\|w_1\|_2 + \|w_2\|_2 = r_+$ in the first case and $\|w_1\|_2 + \|w_2\|_2 = r_-$ in the second. If $\gamma = \pi$, the right-hand sides of the tangential equations vanish, so $\mathcal{P}'(\varphi_1) = \mathcal{P}'(\varphi_2) = 0$ and the two directions are $\bar{v}$ and $-\bar{v}$. After exchanging the students, $(\varphi_1, \varphi_2) = (0, -\pi)$, and since $l(\pi) = 0$, the radial equations give $\|w_1\|_2 = r_+$ and $\|w_2\|_2 = r_-$. This determines all stationary points with collinear directions.

Noncollinear directions. Assume now that $0 < \gamma < \pi$. Multiplying the two tangential equations gives

$$\mathcal{P}'(\varphi_1)\mathcal{P}'(\varphi_2) = -(\pi - \gamma)^2\|w_1\|_2\|w_2\|_2\sin^2(\varphi_2 - \varphi_1) < 0.$$

By Lemma C.6, one direction therefore lies in $(-\pi, 0)$ and the other in $(0, \pi)$. After relabeling, we write $\varphi_- = \varphi_1 \in (-\pi, 0)$ and $\varphi_+ = \varphi_2 \in (0, \pi)$. Since $\mathcal{P}'(\varphi_-) > 0$, the tangential equation (C.35)

gives $(\pi-\gamma)\|w_2\|_2\sin(\varphi_+-\varphi_-)>0$, so $\sin(\varphi_+-\varphi_-)>0$. As $0<\varphi_+-\varphi_-<2\pi$, it follows that

$$0<\varphi_+-\varphi_-<\pi. \tag{C.38}$$

Up to the exchange of the students and the reflection $\varphi\mapsto-\varphi$, (C.38) leaves four cases for the intervals containing φ_- and φ_+:

case	(φ_-,φ_+)
1	$J_0\times I_0$,
2	$J_0\times I_1$ or, by reflection, $J_1\times I_0$,
3	$J_2\times I_0$ or, by reflection, $J_0\times I_2$,
4	$J_1\times I_1$.

The products $J_1\times I_2$, $J_2\times I_1$, and $J_2\times I_2$ are excluded by (C.38), since on them $\varphi_+-\varphi_-\geq\pi$. Directions at the endpoints $\pm\alpha/2$ and $\pm(\pi-\alpha/2)$ are included in the adjacent cases. We now show that these cases contain no stationary point other than $(w_1,w_2)=(v_1,v_2)$, which lies on the boundary of Case 1.

Case 1: $\varphi_-\in J_0$ and $\varphi_+\in I_0$. We write $\varphi_+=x$ and $\varphi_-=-y$ with $0<x,y\leq\alpha/2$, so that $\gamma=x+y$, and set $p=x-y$. Then

$$|p|\leq\gamma\leq\alpha<\pi. \tag{C.39}$$

On $[0,\alpha/2]$,

$$\mathcal{P}(\varphi)=2\Big(\big((\pi-\alpha/2)\cos(\alpha/2)+\sin(\alpha/2)\big)\cos\varphi+\varphi\sin(\alpha/2)\sin\varphi\Big),$$

and $\mathcal{P}'(\varphi)$ is given by the first formula in the proof of Lemma C.6. Here $(\varphi_1,\varphi_2)=(-y,x)$. Since $\mathcal{P}$ is even and $(\pi-\gamma)\sin\gamma>0$, the tangential equations give

$$\|w_1\|_2=-\frac{\mathcal{P}'(x)}{(\pi-\gamma)\sin\gamma},\qquad\|w_2\|_2=-\frac{\mathcal{P}'(y)}{(\pi-\gamma)\sin\gamma},$$

and the radial equations read $\pi\|w_2\|_2+2\pi l(\gamma)\|w_1\|_2=\mathcal{P}(x)$ and $\pi\|w_1\|_2+2\pi l(\gamma)\|w_2\|_2=\mathcal{P}(y)$. Substituting the norms and multiplying by $(\pi-\gamma)\sin\gamma$, we obtain

$$\begin{aligned}-\pi\mathcal{P}'(y)-2\pi l(\gamma)\mathcal{P}'(x)&=(\pi-\gamma)\sin\gamma\,\mathcal{P}(x),\\-\pi\mathcal{P}'(x)-2\pi l(\gamma)\mathcal{P}'(y)&=(\pi-\gamma)\sin\gamma\,\mathcal{P}(y).\end{aligned}$$

After division by 2, both equations are linear and homogeneous in the nonzero vector $\big((\pi-\alpha/2)\cos(\alpha/2),\sin(\alpha/2)\big)$, with coefficient matrix

$$M=\begin{pmatrix}\gamma\sin y+\sin\gamma\sin x & -\gamma(\pi-x)\cos y-\sin\gamma(\pi-y)\cos x\\ \gamma\sin x+\sin\gamma\sin y & -\gamma(\pi-y)\cos x-\sin\gamma(\pi-x)\cos y\end{pmatrix}.$$

Hence $\det M=0$. Expanding the determinant and using $x+y=\gamma$ and $x-y=p$, we find

$$\begin{aligned}\det M&=(\gamma^2-\sin^2\gamma)\big((\pi-x)\sin x\cos y-(\pi-y)\sin y\cos x\big)\\&=\frac{\gamma^2-\sin^2\gamma}{2}\big((2\pi-\gamma)\sin p-p\sin\gamma\big).\end{aligned}$$

Since $\gamma>\sin\gamma>0$, it follows that $(2\pi-\gamma)\sin p=p\sin\gamma$. If $p\neq0$, this says that $\sin|p|/|p|=\sin\gamma/(2\pi-\gamma)$. But $|p|\leq\gamma<\pi$ by (C.39), and $\sin t/t$ is strictly decreasing on $(0,\pi)$, so $\sin|p|/|p|\geq\sin\gamma/\gamma>\sin\gamma/(2\pi-\gamma)$, a contradiction. Hence $p=0$.

Thus $x=y=\gamma/2$, and either of the two equations above becomes

$$(\gamma+\sin\gamma)\left((\pi-\alpha/2)\cos(\alpha/2)\sin\frac{\gamma}{2}-\sin(\alpha/2)\left(\pi-\frac{\gamma}{2}\right)\cos\frac{\gamma}{2}\right)=0.$$

Since $\gamma+\sin\gamma>0$, it follows that $(\pi-\gamma/2)\cot(\gamma/2)=(\pi-\alpha/2)\cot(\alpha/2)$. The function $\gamma\mapsto(\pi-\gamma/2)\cot(\gamma/2)$ is strictly decreasing on $(0,\pi)$, because its derivative is

$$-\frac{\left(\pi-\frac{\gamma}{2}\right)+\sin\frac{\gamma}{2}\cos\frac{\gamma}{2}}{2\sin^2\frac{\gamma}{2}}<0.$$

Hence $\gamma=\alpha$ and $(\varphi_-,\varphi_+)=(-\alpha/2,\alpha/2)$. Since $-\mathcal{P}'(\alpha/2)=(\pi-\alpha)\sin\alpha$, the formulas for the norms give $\|w_1\|_2=\|w_2\|_2=1$. Thus the only stationary point in Case 1, including its boundary, is (v_1,v_2), up to the exchange of the students.

Case 2: $\varphi_- \in J_0$ and $\varphi_+ \in I_1$. By the reflection $\varphi \mapsto -\varphi$, we may assume that $\varphi_- \in J_0$ and $\varphi_+ \in I_1$, including the endpoints except $\varphi_+ = \pi - \alpha/2$, which is treated in Case 3. We write $\varphi_1 = \varphi_- = \alpha/2 - b_0$ and $\varphi_2 = \varphi_+ = \alpha/2 + a_0$, and set $s = a_0 + b_0$ and $\rho = \alpha - b_0$. Then $a_0 \geq 0$, $\alpha/2 < b_0 \leq \alpha$, $s < \pi$ by (C.38), and $a_0 + \alpha < \pi$ because $\varphi_+ < \pi - \alpha/2$.

First suppose that $a_0 > 0$ and $\rho > 0$. The teacher–student angles are $\theta_{11} = b_0$, $\theta_{12} = \rho$, $\theta_{21} = a_0$, and $\theta_{22} = a_0 + \alpha$. Substituting them into (C.33)–(C.36), using the tangential equations to eliminate the radii, and collecting the coefficients of $\sin b_0$, $\sin a_0$, $\sin\rho$, and $\sin(a_0+\alpha)$, we obtain

$$0 = -a_0(2\pi - s)\sin b_0 + s(2\pi - a_0 - \alpha)\sin\rho + \sin s\big(\rho\sin(a_0+\alpha) - b_0\sin a_0\big),$$
$$0 = -b_0(2\pi - s)\sin a_0 + s\rho\sin(a_0+\alpha) + \sin s\big((2\pi - a_0 - \alpha)\sin\rho - a_0\sin b_0\big).$$

We write $U = a_0\sin b_0$, $V = b_0\sin a_0$, $M = (2\pi - a_0 - \alpha)\sin\rho$, and $N = \rho\sin(a_0+\alpha)$, which are all positive. Since $2\pi - a_0 - \alpha = 2\pi - s - \rho$ and $a_0 + \alpha = s + \rho$, these equations can be written as

$$(2\pi - s)U + \sin s\, V = sM + \sin s\, N, \tag{C.40}$$
$$\sin s\, U + (2\pi - s)V = \sin s\, M + sN. \tag{C.41}$$

and adding and subtracting them gives

$$(2\pi - s + \sin s)(U + V) = (s + \sin s)(M + N), \tag{C.42}$$
$$(2\pi - s - \sin s)(U - V) = (s - \sin s)(M - N). \tag{C.43}$$

Since $0 < \rho < s + \rho < \pi$, we have $2\pi - s - \rho > s + \rho$, and since $\sin t/t$ is strictly decreasing on $(0,\pi)$, $M > (s+\rho)\sin\rho > \rho\sin(s+\rho) = N$. The coefficients in (C.43) are positive, so $U > V$, that is, $\sin b_0/b_0 > \sin a_0/a_0$, and hence $a_0 > b_0$. Together with $b_0 > \alpha/2$ and $s + \rho = a_0 + \alpha < \pi$, this gives

$$0 < \rho = \alpha - b_0 < b_0 < a_0, \qquad b_0 < \frac{s}{2}, \qquad \rho < \min\left\{\frac{s}{2}, \pi - s\right\}. \tag{C.44}$$

To compare $U - V$ with $M - N$, we fix $s \in (0,\pi)$ and introduce

$$R_s(x) = \frac{(s-x)\sin x - x\sin(s-x)}{(s-x)\sin x + x\sin(s-x)}, \qquad Q_s(x) = \frac{(2\pi - s - x)\sin x - x\sin(s+x)}{(2\pi - s - x)\sin x + x\sin(s+x)},$$

for $0 < x < s/2$ and $0 < x < \min\{s/2, \pi - s\}$, respectively. Both denominators are positive, the second because $0 < s + x < \pi$. By (C.44), $(U-V)/(U+V) = R_s(b_0)$ and $(M-N)/(M+N) = Q_s(\rho)$, so dividing (C.43) by (C.42) gives

$$R_s(b_0) = \Lambda_s Q_s(\rho), \qquad \Lambda_s = \frac{s - \sin s}{s + \sin s}\frac{2\pi - s + \sin s}{2\pi - s - \sin s}. \tag{C.45}$$

As $x \downarrow 0$, $R_s(x)$ and $Q_s(x)$ tend to $R_s(0) = (s - \sin s)/(s + \sin s)$ and $Q_s(0) = (2\pi - s - \sin s)/(2\pi - s + \sin s)$, so $\Lambda_s = R_s(0)/Q_s(0)$.

We next show that

$$R_s(x) < \frac{s-x}{s}R_s(0), \qquad 0 < x < s/2, \tag{C.46}$$
$$Q_s(x) \geq \frac{s-x}{s}Q_s(0), \qquad 0 < x < \min\{s/2, \pi - s\}. \tag{C.47}$$

For the first bound, we write $A(t) = \log(t/\sin t)$, which is positive for $0 < t < \pi$, and $\psi(t) = (t - \sin t)/\big(t(t + \sin t)\big)$. Dividing the numerator and the denominator of $R_s(x)$ by $\sin x\sin(s-x)$ gives

$$R_s(x) = \tanh\left(\frac{A(s-x) - A(x)}{2}\right) < \tanh\left(\frac{A(s-x)}{2}\right) = (s-x)\psi(s-x).$$

To see that ψ is nondecreasing on $(0,\pi)$, we compute

$$\psi'(t) = \frac{\sin^2 t + 2t\sin t - t^2(1 + 2\cos t)}{t^2(t + \sin t)^2}.$$

The numerator vanishes at $t = 0$ and has derivative $2\big(t^2 \sin t - (1+\cos t)(t - \sin t)\big) \geq 0$, because $t - \sin t = \int_0^t (1 - \cos u)\,\mathrm{d}u \leq t(1-\cos t)$ implies $(1+\cos t)(t-\sin t) \leq t \sin^2 t \leq t^2 \sin t$. Hence $R_s(x) < (s-x)\psi(s) = \frac{s-x}{s} R_s(0)$, which is (C.46).

For the second bound, we write $B(t) = \log\big((2\pi - t)/\sin t\big)$ for $0 < t < \pi$. In the same way, $Q_s(x) = \tanh\big((B(s+x) - A(x))/2\big)$, and we first show that

$$B(s+x) - A(x) \geq \frac{s-x}{s} B(s). \tag{C.48}$$

With $A(0) = 0$, the function $H_s(x) = B(s+x) - A(x) - \frac{s-x}{s} B(s)$ vanishes at $x = 0$, and

$$H_s'(x) = \frac{B(s)}{s} - \left(\frac{1}{x} - \cot x\right) - \left(\cot(s+x) + \frac{1}{2\pi - s - x}\right).$$

The function $t \mapsto 1/t - \cot t$ is increasing on $(0, \pi)$, since its derivative $1/\sin^2 t - 1/t^2$ is positive, and $t \mapsto \cot t + 1/(2\pi - t)$ is decreasing, since its derivative $-1/\sin^2 t + 1/(2\pi - t)^2$ is negative. Because $x < s/2$ and $s + x > s$,

$$H_s'(x) \geq \frac{B(s)}{s} - \left(\frac{2}{s} - \cot\frac{s}{2}\right) - \left(\cot s + \frac{1}{2\pi - s}\right).$$

Since $\cot(s/2) = 1/\sin s + \cot s$, the right-hand side is nonnegative as soon as

$$B(s) \geq 2 - \frac{s}{\sin s} + \frac{s}{2\pi - s}. \tag{C.49}$$

Applying $\log z \geq 2(z-1)/(z+1)$ for $z > 1$ with $z = (2\pi - s)/\sin s$, we see that (C.49) follows from

$$s\big((2\pi - s)^2 - \sin^2 s\big) \geq 4(2\pi - s)\sin^2 s. \tag{C.50}$$

If $0 < s \leq 1$, then $\sin s \leq s$ and $(2\pi - s)/s \geq 2\pi - 1 > 2 + \sqrt{5}$, which implies $(2\pi - s)^2 - s^2 \geq 4(2\pi - s)s$ and hence (C.50). If $1 \leq s < \pi$, then $\sin^2 s \leq 1$, and it is enough to show that $\chi(s) = s\big((2\pi - s)^2 - 1\big) - 4(2\pi - s)$ is positive. Since $\chi''(s) = 6s - 8\pi < 0$ on $[1, \pi]$, this follows from $\chi(1) = 4(\pi^2 - 3\pi + 1) > 0$ and $\chi(\pi) = \pi(\pi^2 - 5) > 0$. Thus (C.49) holds, so $H_s' \geq 0$, and (C.48) follows from $H_s(0) = 0$. Since $\tanh$ is increasing, and concave on $[0, \infty)$ with $\tanh 0 = 0$,

$$Q_s(x) \geq \tanh\left(\frac{s-x}{s}\frac{B(s)}{2}\right) \geq \frac{s-x}{s}\tanh\left(\frac{B(s)}{2}\right) = \frac{s-x}{s} Q_s(0),$$

which is (C.47).

Since $0 < \rho < b_0$ by (C.44), the two bounds give

$$R_s(b_0) < \frac{s - b_0}{s} R_s(0) < \frac{s-\rho}{s} R_s(0) = \Lambda_s \frac{s-\rho}{s} Q_s(0) \leq \Lambda_s Q_s(\rho),$$

which contradicts (C.45). Hence there is no stationary point with $a_0 > 0$ and $\rho > 0$.

It remains to treat the endpoints. If $\rho = 0$ and $a_0 > 0$, then $M = N = 0$, and (C.40)–(C.41) say that (U, V) lies in the kernel of the matrix with diagonal entries $2\pi - s$ and off-diagonal entries $\sin s$. This matrix is invertible because $2\pi - s > \sin s$, so $U = V = 0$, which contradicts $a_0 > 0$. If $a_0 = 0$, then $\varphi_+ = \alpha/2$ lies in the closure of Case 1, where the only stationary point is (v_1, v_2). Thus Case 2 contributes no further stationary point.

Case 3: $\varphi_- \in J_2$ and $\varphi_+ \in I_0$. By reflection, we may assume that $\varphi_- \in J_2$ and $\varphi_+ \in I_0$, and we write $\varphi_+ = p$ and $\varphi_- = -\pi + q$ with $0 \leq p, q \leq \alpha/2$. Then $\varphi_+ - \varphi_- = \pi - (q - p)$, so (C.38) gives $p < q$, and $\gamma = \pi - \delta$ with $\delta = q - p \in (0, \alpha/2]$. Substituting these directions into (C.33)–(C.36) and eliminating the two radii, we obtain

$$0 = \sin\delta\, A_p + \delta B_{p,q} - \pi C_q, \tag{C.51}$$

where

$$\begin{aligned} A_p &= 2p\sin(\alpha/2)\cos p - 2(\pi - \alpha/2)\cos(\alpha/2)\sin p, \\ B_{p,q} &= 2(\pi - \alpha/2)\cos(\alpha/2)\sin q - 2p\sin(\alpha/2)\cos q + 2\sin(\alpha/2)\cos p\sin\delta, \\ C_q &= 2q\sin(\alpha/2)\cos q + \alpha\cos(\alpha/2)\sin q. \end{aligned}$$

By Lemma C.1, $\sin(\alpha/2) \le (\pi - \alpha/2)\cos(\alpha/2)$. Since $\tan p > p$ for $0 < p < \pi/2$, it follows that $A_p = 2p\cos p\big(\sin(\alpha/2) - (\pi - \alpha/2)\cos(\alpha/2)\tan p/p\big) < 0$ for $p > 0$, while $A_0 = 0$. Thus $A_p \le 0$. Next, the middle term of $B_{p,q}$ is nonpositive. Dropping it and using $\delta \le q \le \alpha/2$, $\cos p \le 1$, and $\sin\delta \le \sin q < \pi\cos q$, where the last inequality holds because $\tan q \le \tan(\alpha/2) \le \pi - \alpha/2 < \pi$, we obtain

$$\begin{aligned}\delta B_{p,q} &\le 2\delta(\pi - \alpha/2)\cos(\alpha/2)\sin q + 2\delta\sin(\alpha/2)\cos p\sin\delta \\ &< \pi\alpha\cos(\alpha/2)\sin q + 2\pi q\sin(\alpha/2)\cos q = \pi C_q.\end{aligned}$$

Hence the right-hand side of (C.51) is negative, a contradiction. Thus Case 3 contains no stationary point.

Case 4: $\varphi_- \in J_1$ and $\varphi_+ \in I_1$. We write $\varphi_+ = \alpha/2 + a_0$ and $\varphi_- = -\alpha/2 - b_0$ with $a_0, b_0 \ge 0$, so that $\gamma = \alpha + a_0 + b_0 < \pi$ and the teacher–student angles are $\theta_{11} = b_0 + \alpha$, $\theta_{12} = b_0$, $\theta_{21} = a_0$, and $\theta_{22} = a_0 + \alpha$. Substituting them into the equations of Lemma C.6 and eliminating the two radii, we obtain

$$\begin{aligned}0 = &-\sin\gamma\big((\gamma - a_0)\sin a_0 + (\gamma - a_0 - \alpha)\sin(a_0 + \alpha)\big) \\ &- (2\pi - \gamma)\big(a_0\sin(b_0 + \alpha) + (a_0 + \alpha)\sin b_0\big).\end{aligned}$$

Both terms on the right are nonpositive, and the second is negative when $a_0, b_0 > 0$. Hence there is no stationary point with $a_0, b_0 > 0$. If exactly one of a_0 and b_0 vanishes, the point lies in the closure of Case 2, and if both vanish, it is (v_1, v_2).

Together with the collinear directions, Cases 1–4 show that every stationary point is one of the configurations listed in Lemma 3.8.

Stationarity and curvature of the listed points. Conversely, every listed configuration is stationary. The pair (v_1, v_2) reproduces the teacher, so the loss, which is nonnegative, vanishes there; it is therefore a global minimum. For two students along $\bar{v}$, we have $\gamma = 0$ and $\mathcal{P}'(0) = 0$, and since $2\pi l(0) = \pi$ and $\mathcal{P}(0) = \pi r_+$, the equations of Lemma C.6 hold exactly when $\|w_1\|_2 + \|w_2\|_2 = r_+$. The same computation for two students along $-\bar{v}$ gives $\|w_1\|_2 + \|w_2\|_2 = r_-$. For the pair $(r_+\bar{v}, -r_-\bar{v})$, we have $\mathcal{P}'(0) = \mathcal{P}'(\pi) = 0$ and $l(\pi) = 0$, and the radial equations hold with the norms r_+ and r_-.

It remains to find a direction of negative curvature at each of these saddles. If both students are parallel to $\bar{v}$, or both are parallel to $-\bar{v}$, the Hessian in Lemma A.6, restricted to perturbations $(t_1 v_\Delta, t_2 v_\Delta)$, is

$$H_{v_\Delta} = \begin{pmatrix} \frac{1}{2} - \frac{2\sin(\alpha/2)}{\pi\|w_1\|_2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} - \frac{2\sin(\alpha/2)}{\pi\|w_2\|_2} \end{pmatrix},$$

The direction $(1, -1)$, which moves the two students apart along v_Δ, gives

$$(1,-1)H_{v_\Delta}(1,-1)^\top = -\frac{2\sin(\alpha/2)}{\pi}\left(\frac{1}{\|w_1\|_2} + \frac{1}{\|w_2\|_2}\right) < 0.$$

At $(r_+\bar{v}, -r_-\bar{v})$, the off-diagonal Hessian block vanishes by the continuous extension in Lemma A.6, and the perturbation $(0, v_\Delta)$ of the second student gives

$$\frac{1}{2} - \frac{2\sin(\alpha/2)}{\pi r_-} = -\frac{\sin(\alpha/2) + (\alpha/2)\cos(\alpha/2)}{2\big(\sin(\alpha/2) - (\alpha/2)\cos(\alpha/2)\big)} < 0.$$

Thus every listed configuration other than (v_1, v_2) has a strictly negative Hessian direction. ■

C.5 Completion of the squared-loss landscape theorem

The statements of Proposition 3.5 and Lemmas 3.6 to 3.8 assume $0 < \alpha \le \pi/2$, but their proofs in the preceding subsections use only $0 < \alpha \le \alpha_\star$, with $\alpha_\star$ as in Section 3.1. These results therefore hold for every $\alpha \in (0, \alpha_\star]$, and they give the following extension of Theorem 3.3.

Theorem C.8 (Squared-loss landscape up to the angle threshold)**.** *Let $0 < \alpha \le \alpha_\star$, and let (w_1, w_2) with $w_1, w_2 \ne 0$ be a stationary point of the squared loss (2.2). After possibly swapping the students, exactly one of the following holds:*

(1) $w_1 = v_1$ *and* $w_2 = v_2$*. This is a global minimum.*

(2) The pair belongs to one of the three in-plane families

$$\begin{aligned} w_1 = s_1\bar{v}, \quad w_2 = s_2\bar{v}, \quad s_1, s_2 > 0, \quad s_1 + s_2 = r_+, \\ w_1 = -s_1\bar{v}, \quad w_2 = -s_2\bar{v}, \quad s_1, s_2 > 0, \quad s_1 + s_2 = r_-, \\ w_1 = r_+\bar{v}, \quad w_2 = -r_-\bar{v}. \end{aligned}$$

Every such point has a strictly negative Hessian direction.

(3) Necessarily $d \geq 3$*, and there is a unit vector* $\xi \perp \operatorname{span}\{v_1, v_2\}$ *such that*

$$\begin{aligned} w_1 &= r_\alpha(\kappa_\alpha)(\cos\kappa_\alpha\,\bar{v} + \sin\kappa_\alpha\,\xi), \\ w_2 &= r_\alpha(\kappa_\alpha)(\cos\kappa_\alpha\,\bar{v} - \sin\kappa_\alpha\,\xi). \end{aligned}$$

Every such point has a strictly negative Hessian direction.

Consequently, every local minimum of the squared loss with $w_1 \neq 0$ *and* $w_2 \neq 0$ *is a global minimum.*

Proof of Theorem C.8. By Lemma 3.2, the pair satisfies (3.3), so by Proposition 3.5 the vectors v_1, v_2, w_1, w_2 span a space of dimension at most three. If $w_1, w_2 \in \operatorname{span}\{v_1, v_2\}$, Lemma 3.8 shows that the pair is (v_1, v_2) or belongs to one of the three families in (2), and that each point of these families has a strictly negative Hessian direction.

Otherwise, $d \geq 3$, and there is a unit vector $\xi \perp \operatorname{span}\{v_1, v_2\}$ such that $w_1, w_2 \in \operatorname{span}\{\bar{v}, v_\Delta, \xi\}$. By Lemma 3.6, $w_1, w_2 \in \operatorname{span}\{\bar{v}, \xi\}$, and Lemma 3.7 shows that the pair has the form in (3) and has a strictly negative Hessian direction. Since this pair does not lie in $\operatorname{span}\{v_1, v_2\}$, the three cases are disjoint.

Conversely, the same lemmas show that every configuration in (1)–(3) is stationary. The pair (v_1, v_2) has zero loss, and every other stationary point has a strictly negative Hessian direction. Since a local minimum with $w_1 \neq 0$ and $w_2 \neq 0$ is stationary, it is a global minimum. ■

Proof of Theorem 3.3. Since $\pi/2 < \alpha_\star$, the theorem is the restriction of Theorem C.8 to $0 < \alpha \leq \pi/2$. ■

D Global convergence and boundary avoidance

For the proof of Theorem 4.2, we fix $d \geq 2$ and unit teachers with $\alpha = \angle(v_1, v_2) \in (0, \pi/2]$, and we write $\Omega = \{(w_1, w_2) \in (\mathbb{R}^d)^2 : w_1 \neq 0,\ w_2 \neq 0\}$, its boundary $\partial\Omega = \{(w_1, w_2) : w_1 = 0 \text{ or } w_2 = 0\}$, $\mathcal{G} = \{(v_1, v_2), (v_2, v_1)\}$ and $r_i = \|w_i\|_2$. A set is null if it has Lebesgue measure zero; on a manifold, this is understood in local coordinates. For the student output f in (2.1), the cross term $2\mathbb{E}[\phi(\langle w_1, x\rangle)\phi(\langle w_2, x\rangle)]$ in $\mathbb{E}[f(x)^2]$ is nonnegative, and $\mathbb{E}[\phi(\langle w_i, x\rangle)^2] = r_i^2/2$ by (A.1a). Since $f^2 \leq 2(f^*)^2 + 2(f - f^*)^2$, we obtain, for all $w_1, w_2 \in \mathbb{R}^d$,

$$\frac{1}{2}\big(r_1^2 + r_2^2\big) \leq \mathbb{E}[f(x)^2] \leq 2\mathbb{E}[f^*(x)^2] + 4L(w_1, w_2). \tag{D.1}$$

Since L does not increase along gradient flow, every trajectory stays in a bounded set determined by its initial loss.

D.1 Almost-everywhere avoidance of zero weights

For $w(0) \in \Omega$, let $[0, T)$ be the maximal interval on which the solution of gradient flow exists in Ω; it is well defined by Lemma A.1, and $T \in (0, \infty]$ depends on $w(0)$. A trajectory may leave Ω in finite time, and it may approach $\partial\Omega$ as $t \to \infty$. We show that both happen only for a null set of initializations. The first lemma shows that the two weights cannot tend to zero together.

Lemma D.1 (Lower bound on $\|w_1\|_2 + \|w_2\|_2$)**.** *For every* $w(0) \in \Omega$ *and* $0 \leq t < T$*,*

$$r_1(t) + r_2(t) \geq r_- + \big(r_1(0) + r_2(0) - r_-\big)e^{-t} \geq \min\{r_1(0) + r_2(0),\, r_-\},$$

where r_- *is defined in* (3.4)*.*

Proof of Lemma D.1. By Lemma A.5, $\dot r_i \ge (r_- - r_1 - r_2)/2$ for $i = 1, 2$. The sum therefore satisfies $\frac{\mathrm{d}}{\mathrm{d}t}(r_1 + r_2) \ge r_- - (r_1 + r_2)$, and Grönwall's inequality gives the first inequality. The second holds because the middle expression is a convex combination of $r_1(0) + r_2(0)$ and r_-. ■

D.1.1 Polar coordinates near $\{w_1 = 0\}$

By symmetry it suffices to study trajectories along which w_1 tends to zero. We use the coordinates $(r_1, \bar w_1, w_2)$, where $w_1 = r_1 \bar w_1$, and set

$$\Psi_{w_2}(\bar w_1) = l(\theta_{11}) + l(\theta_{12}) - r_2\, l(\gamma),$$

with $\theta_{11}, \theta_{12}, \gamma$ as in (2.7). Substituting $w_1 = r_1 \bar w_1$ into (A.10), with its convention at $w_1 = 0$, gives

$$L(r_1 \bar w_1, w_2) = L(0, w_2) - r_1 \Psi_{w_2}(\bar w_1) + \frac{r_1^2}{4}. \tag{D.2}$$

For a nonzero vector b, (A.2) shows that the spherical gradient of $u \mapsto l(\angle(u, b))$ at $\bar w_1$ is

$$\nabla_{\mathbb{S}^{d-1}}\big[l(\angle(\cdot, b))\big](\bar w_1) = p\big(\angle(\bar w_1, b)\big) P^{\perp}_{\bar w_1} \bar b, \qquad p(\theta) = \frac{\pi - \theta}{2\pi}. \tag{D.3}$$

With this formula, (A.15) and (A.13) become the first two equations below, and differentiating (D.2) in w_2 gives the third:

$$\begin{aligned} \dot r_1 &= \Psi_{w_2}(\bar w_1) - \frac{r_1}{2}, \\ \dot{\bar w}_1 &= \frac{1}{r_1} \nabla_{\mathbb{S}^{d-1}} \Psi_{w_2}(\bar w_1), \\ \dot w_2 &= F_b(w_2) + r_1 \nabla_{w_2} \Psi_{w_2}(\bar w_1), \end{aligned} \tag{D.4}$$

where $F_b(w_2) = -\nabla_{w_2} L(0, w_2)$; thus $\dot w_2 = F_b(w_2)$ is the gradient flow of L restricted to $\{w_1 = 0\}$.

The factor $1/r_1$ in the second equation is removed by the rescaled time $\tau(t) = \int_0^t \mathrm{d}t'/r_1(t')$. With $' = \mathrm{d}/\mathrm{d}\tau$, the triple $(r_1, \bar w_1, w_2)$ solves $(r_1, \bar w_1, w_2)' = X(r_1, \bar w_1, w_2)$, where the vector field X is r_1 times the right-hand side of (D.4); its components are $r_1(\Psi_{w_2}(\bar w_1) - r_1/2)$, $\nabla_{\mathbb{S}^{d-1}} \Psi_{w_2}(\bar w_1)$ and $r_1(F_b(w_2) + r_1 \nabla_{w_2} \Psi_{w_2}(\bar w_1))$. These formulas make sense for every $r_1 \in \mathbb{R}$ and every $\bar w_1 \in \mathbb{S}^{d-1}$. By Lemma A.1, the kernel in (A.1a) is C^2 in its two nonzero arguments, so $\Psi_{w_2}(\bar w_1)$ and $L(0, w_2)$ are C^2 in $(\bar w_1, w_2)$ for $w_2 \neq 0$. Hence X is a C^1 vector field on $\mathbb{R} \times \mathbb{S}^{d-1} \times (\mathbb{R}^d \setminus \{0\})$, and $X(0, \bar w_1, w_2) = (0, \nabla_{\mathbb{S}^{d-1}} \Psi_{w_2}(\bar w_1), 0)$.

The next lemma describes a trajectory whose first weight tends to zero.

Lemma D.2 (Behavior as w_1 tends to zero)**.** *Let $w(0) \in \Omega$, and suppose that $r_1(t) \to 0$ as $t \uparrow T$. If $T = \infty$, suppose also that $\int_0^\infty \|\dot w(t)\|_2 \,\mathrm{d}t < \infty$. Then $w_2(t)$ converges to some $w_{2,*} \neq 0$, $\tau(t) \to \infty$, and $\Psi_{w_2(t)}(\bar w_1(t))$ converges to some $\psi_* \le 0$, with $\psi_* = 0$ if $T = \infty$. Every limit point u_* of $\bar w_1(t)$ as $t \uparrow T$ satisfies $\nabla_{\mathbb{S}^{d-1}} \Psi_{w_{2,*}}(u_*) = 0$ and $\Psi_{w_{2,*}}(u_*) = \psi_*$.*

Proof of Lemma D.2. By (D.1) and Lemma D.1, there are $0 < c_0 \le C_0$ and $t_0 < T$ such that $r_1(t) \le C_0$ and $c_0 \le r_2(t) \le C_0$ for $t_0 \le t < T$. On the compact set $\{|r_1| \le C_0,\ c_0 \le \|w_2\|_2 \le C_0\}$, the field X, its derivative, and the right-hand sides of the first and third equations in (D.4) are bounded.

If $T < \infty$, then $\dot w_2$ and $\dot r_1$ are bounded on $[t_0, T)$. Hence $w_2(t)$ converges, and $r_1(t) \le C(T - t)$ for some C, so that $\tau(t) - \tau(t_0) \ge C^{-1} \log \frac{T - t_0}{T - t} \to \infty$. If $T = \infty$, then $w_2(t)$ converges because its path has finite length, and $\tau(t) - \tau(t_0) \ge (t - t_0)/C_0 \to \infty$. In both cases $\|w_{2,*}\|_2 \ge c_0$, and $\int_{\tau(t_0)}^\infty \|w_2'(\tau)\|_2 \,\mathrm{d}\tau = \int_{t_0}^T \|\dot w_2(t)\|_2 \,\mathrm{d}t < \infty$.

Let $\psi(\tau) = \Psi_{w_2(\tau)}(\bar w_1(\tau))$. Since $\bar w_1' = \nabla_{\mathbb{S}^{d-1}} \Psi_{w_2}(\bar w_1)$, we have $\psi' = \|\nabla_{\mathbb{S}^{d-1}} \Psi_{w_2}(\bar w_1)\|_2^2 + D_{w_2} \Psi_{w_2}(\bar w_1)[w_2']$. The last term is integrable, because $D_{w_2} \Psi$ is bounded on the compact set above and w_2' is integrable. Since ψ is bounded, ψ minus the integral of this term is bounded and nondecreasing. Hence ψ has a limit ψ_*, and $\int^\infty \|\nabla_{\mathbb{S}^{d-1}} \Psi_{w_2}(\bar w_1)\|_2^2 \,\mathrm{d}\tau < \infty$. The derivative of $\tau \mapsto \|\nabla_{\mathbb{S}^{d-1}} \Psi_{w_2}(\bar w_1)\|_2^2$ is bounded, because $\bar w_1'$ and w_2' are bounded and $\nabla_{\mathbb{S}^{d-1}} \Psi$ is C^1. Barbalat's lemma therefore gives $\nabla_{\mathbb{S}^{d-1}} \Psi_{w_2}(\bar w_1) \to 0$, and continuity gives the two conditions at every limit point u_*.

Finally, $\dot{r}_1 = \Psi_{w_2}(\bar{w}_1) - r_1/2$ by (D.4). If $\psi_* > 0$, then $\dot{r}_1 \geq \psi_*/4$ for t close to T, which contradicts $r_1(t) \to 0$. If $T = \infty$ and $\psi_* < 0$, then $\dot{r}_1 \leq \psi_*/2$ for all large t, so r_1 reaches zero in finite time, which contradicts $T = \infty$. ■

D.1.2 Equilibria of X with $r_1 = 0$

Let $\mathcal{Z}_0$ be the set of points $(0, \bar{w}_1, w_2)$ with $w_2 \neq 0$, $\nabla_{\mathbb{S}^{d-1}} \Psi_{w_2}(\bar{w}_1) = 0$ and $\Psi_{w_2}(\bar{w}_1) \leq 0$. These points are equilibria of X, and by Lemma D.2, for a trajectory as in that lemma, the limit points of $(r_1, \bar{w}_1, w_2)$ lie in $\mathcal{Z}_0$. We show that the divergence of X is positive on $\mathcal{Z}_0$ except at two points, so that trajectories staying near the other points form a null set.

Lemma D.3 (A sine inequality at equilibria of X)**.** *Let $(0, \bar{w}_1, w_2) \in \mathcal{Z}_0$. Then $r_2 \sin\gamma \leq \sin\theta_{11} + \sin\theta_{12}$, with equality if and only if $\Psi_{w_2}(\bar{w}_1) = 0$ and*

$$(\bar{w}_1, w_2) = (-v_1, v_2) \quad \text{or} \quad (\bar{w}_1, w_2) = (-v_2, v_1). \tag{D.5}$$

Proof of Lemma D.3. For $\theta \in (0, \pi)$, define Q by $Q(p(\theta)) = l(\theta)/\sin\theta$, with p as in (D.3). By (A.1b), $Q(p(\theta))/p(\theta) = \cot\theta + 1/(\pi - \theta)$. The function $l(\theta)/\sin\theta$ has derivative $-((\pi - \theta) + \sin\theta\cos\theta)/(2\pi\sin^2\theta)$, which is negative because $\frac{1}{2}|\sin 2\theta| = \frac{1}{2}|\sin 2(\pi - \theta)| < \pi - \theta$, and second derivative $(1 + (\pi - \theta)\cot\theta)/(\pi\sin^2\theta) = 2l(\theta)/\sin^3\theta$, which is positive because $l > 0$ on $[0, \pi)$ by Lemma A.2. The derivative of $\cot\theta + 1/(\pi - \theta)$ is $1/(\pi - \theta)^2 - 1/\sin^2\theta$, which is negative because $\sin\theta = \sin(\pi - \theta) < \pi - \theta$. Since p is affine and decreasing, Q is strictly increasing and strictly convex on $(0, 1/2)$, and $Q(p)/p$ is strictly increasing there.

By (D.3), the condition $\nabla_{\mathbb{S}^{d-1}} \Psi_{w_2}(\bar{w}_1) = 0$ reads

$$p(\theta_{11}) P^{\perp}_{\bar{w}_1} v_1 + p(\theta_{12}) P^{\perp}_{\bar{w}_1} v_2 = r_2 p(\gamma) P^{\perp}_{\bar{w}_1} \bar{w}_2, \tag{D.6}$$

and the norm of the right-hand side is $r_2 p(\gamma)\sin\gamma$.

Since Q is defined only on $(0, 1/2)$, we first treat the case in which $\bar{w}_1$ is aligned with a teacher, say $\theta_{11} = 0$, so that $\theta_{12} = \alpha$. Taking norms in (D.6) gives $p(\alpha)\sin\alpha = r_2 p(\gamma)\sin\gamma$, so $\gamma \in (0, \pi)$. Since $\Psi_{w_2}(\bar{w}_1) \leq 0$ and $l(0) = 1/2$,

$$\frac{1}{2} + l(\alpha) \leq r_2 l(\gamma) = p(\alpha)\sin\alpha \, \frac{Q(p(\gamma))}{p(\gamma)}, \qquad l(\alpha) = p(\alpha)\sin\alpha \, \frac{Q(p(\alpha))}{p(\alpha)}.$$

Hence $Q(p(\gamma))/p(\gamma) > Q(p(\alpha))/p(\alpha)$, so $p(\gamma) > p(\alpha)$ and $r_2 \sin\gamma = p(\alpha)\sin\alpha/p(\gamma) < \sin\alpha = \sin\theta_{11} + \sin\theta_{12}$.

We may therefore assume that $\theta_{11}, \theta_{12} \neq 0$. Let $S = \sin\theta_{11} + \sin\theta_{12}$. Then $S > 0$, since $S = 0$ would force $\theta_{11} = \theta_{12} = \pi$ and hence $v_1 = v_2$. A teacher with $\theta_{1j} = \pi$ contributes nothing to $\Psi_{w_2}(\bar{w}_1)$ or to (D.6), because $l(\pi) = p(\pi) = 0$, and we omit it from the sums below.

If the left-hand side of (D.6) vanishes, so does the right-hand side, and $\gamma \in \{0, \pi\}$. The case $\gamma = \pi$ is impossible, because then $\Psi_{w_2}(\bar{w}_1) = l(\theta_{11}) + l(\theta_{12}) > 0$. Hence $\gamma = 0$ and $r_2 \sin\gamma = 0 < S$.

Otherwise, let $\hat{p}$ be the norm of the left-hand side of (D.6) divided by S. The triangle inequality gives $0 < \hat{p} \leq (p(\theta_{11})\sin\theta_{11} + p(\theta_{12})\sin\theta_{12})/S < 1/2$, where the last inequality holds because $p(\theta_{1j}) < 1/2$ for $\theta_{1j} \neq 0$. Taking norms in (D.6) gives $r_2 p(\gamma)\sin\gamma = S\hat{p} > 0$, so $\gamma \in (0, \pi)$. By Jensen's inequality for the convex function Q with weights $\sin\theta_{1j}/S$, and then by the monotonicity of Q,

$$l(\theta_{11}) + l(\theta_{12}) = S \sum_{j=1}^{2} \frac{\sin\theta_{1j}}{S} \, Q(p(\theta_{1j})) \geq SQ\Big(\sum_{j=1}^{2} \frac{\sin\theta_{1j}}{S} \, p(\theta_{1j})\Big) \geq SQ(\hat{p}). \tag{D.7}$$

On the other hand, $\Psi_{w_2}(\bar{w}_1) \leq 0$ gives

$$l(\theta_{11}) + l(\theta_{12}) \leq r_2 l(\gamma) = r_2 p(\gamma)\sin\gamma \, \frac{Q(p(\gamma))}{p(\gamma)} = S\hat{p} \, \frac{Q(p(\gamma))}{p(\gamma)}.$$

Combining the two bounds, $Q(\hat{p})/\hat{p} \leq Q(p(\gamma))/p(\gamma)$. Hence $\hat{p} \leq p(\gamma)$, and $r_2 \sin\gamma = S\hat{p}/p(\gamma) \leq S$, which proves the inequality.

It remains to identify the case of equality. If $r_2 \sin\gamma = S$, then $\hat{p} = p(\gamma)$, so equality holds in (D.7), in the triangle inequality, and in $\Psi_{w_2}(\bar{w}_1) \le 0$. If both $\sin\theta_{11}$ and $\sin\theta_{12}$ were positive, strict convexity of Q would give $\theta_{11} = \theta_{12}$, and equality in the triangle inequality would make $P^{\perp}_{\bar{w}_1} v_1$ and $P^{\perp}_{\bar{w}_1} v_2$ positive multiples of each other. These projections have the same norm, and $\langle \bar{w}_1, v_1\rangle = \langle \bar{w}_1, v_2\rangle$, so $v_1 = v_2$, a contradiction. Hence one sine vanishes, say $\sin\theta_{11} = 0$; since $\theta_{11} \neq 0$, we have $\theta_{11} = \pi$ and $\bar{w}_1 = -v_1$. Then $S = \sin\theta_{12} > 0$ and $\hat{p} = p(\theta_{12})$, so $p(\gamma) = p(\theta_{12})$ gives $\gamma = \theta_{12}$, and $r_2 \sin\gamma = S$ gives $r_2 = 1$. Now (D.6) gives $P^{\perp}_{\bar{w}_1} v_2 = P^{\perp}_{\bar{w}_1} \bar{w}_2$, and $\langle \bar{w}_1, v_2\rangle = \cos\theta_{12} = \cos\gamma = \langle \bar{w}_1, \bar{w}_2\rangle$, so $w_2 = v_2$. The case $\sin\theta_{12} = 0$ gives $(\bar{w}_1, w_2) = (-v_2, v_1)$.

Conversely, $\Psi_{v_2}(u) = l(\angle(u, v_1))$ for every $u \in \mathbb{S}^{d-1}$, because the two terms involving v_2 cancel. This function is nonnegative and vanishes at $u = -v_1$, so both $\Psi_{v_2}(-v_1)$ and its spherical gradient vanish there. At $(\bar{w}_1, w_2) = (-v_1, v_2)$, we have $r_2 \sin\gamma = \sin(\pi - \alpha) = \sin\theta_{11} + \sin\theta_{12}$. The point $(-v_2, v_1)$ is symmetric. ■

Lemma D.4 (Positive divergence of X)**.** *Let $\mathcal{Z}$ be the set $\mathcal{Z}_0$ without the two points with $(\bar{w}_1, w_2)$ as in (D.5). Then $\operatorname{div} X > 0$ on $\mathcal{Z}$, where the divergence and the volume on $\mathbb{R} \times \mathbb{S}^{d-1} \times (\mathbb{R}^d \setminus \{0\})$ are those of the product metric. Consequently, every compact $\mathcal{K} \subset \mathcal{Z}$ has a relatively compact open neighborhood U such that the set $\mathcal{T}(U)$ of points of U whose forward orbit under X stays in U is null.*

Proof of Lemma D.4. At $r_1 = 0$, the derivative of the first component of X with respect to r_1 is $\Psi_{w_2}(\bar{w}_1)$, the divergence of the second component on the sphere is $\operatorname{tr} \operatorname{Hess}_{\mathbb{S}^{d-1}} \Psi_{w_2}(\bar{w}_1)$, and the third component has the factor r_1, so its divergence in w_2 vanishes. To compute the trace, let b be a unit vector, $u \in \mathbb{S}^{d-1}$ and $\theta = \angle(u, b)$. By (A.2), the first and second derivatives of $l(\arccos(\cdot))$ at $\cos\theta$ are $(\pi - \theta)/(2\pi)$ and $1/(2\pi \sin\theta)$. Hence the spherical Hessian of $u \mapsto l(\angle(u, b))$ is $\frac{1}{2\pi\sin\theta} P^{\perp}_u b b^{\top} P^{\perp}_u - \frac{(\pi-\theta)\cos\theta}{2\pi} P^{\perp}_u$, and its trace is $\frac{\sin\theta}{2\pi} - (d-1)\frac{(\pi-\theta)\cos\theta}{2\pi} = \frac{d\sin\theta}{2\pi} - (d-1)l(\theta)$. This identity extends to $\theta \in \{0, \pi\}$ by continuity, since Ψ is C^2. Applying it to the three terms of Ψ_{w_2}, we obtain

$$\operatorname{div} X(0, \bar{w}_1, w_2) = \frac{d}{2\pi}\big(\sin\theta_{11} + \sin\theta_{12} - r_2 \sin\gamma\big) - (d-2)\Psi_{w_2}(\bar{w}_1). \tag{D.8}$$

On $\mathcal{Z}$, the last term is nonnegative because $d \ge 2$ and $\Psi_{w_2}(\bar{w}_1) \le 0$, and the first term is positive by Lemma D.3, since the points of equality are excluded. Hence $\operatorname{div} X > 0$ on $\mathcal{Z}$.

Let $\mathcal{K} \subset \mathcal{Z}$ be compact. By continuity, there are a relatively compact open neighborhood U of $\mathcal{K}$ and $c > 0$ such that $\operatorname{div} X \ge c$ on U. Orbits that stay in U exist for all $\tau \ge 0$, because U is relatively compact. For each integer $N \ge 1$, the set of points of U whose orbit stays in U on $[0, N]$ is open, by continuous dependence on initial conditions, and $\mathcal{T}(U)$ is the intersection of these sets. Hence $\mathcal{T}(U)$ is Borel. Let Φ_τ be the flow of X. For $\tau > 0$, Φ_τ maps $\mathcal{T}(U)$ injectively into U, and by Liouville's formula its Jacobian determinant on $\mathcal{T}(U)$ is $\exp\int_0^\tau (\operatorname{div} X) \circ \Phi_{\tau'} \, \mathrm{d}\tau' \ge e^{c\tau}$. Hence $\operatorname{vol}(U) \ge e^{c\tau} \operatorname{vol}(\mathcal{T}(U))$, and letting $\tau \to \infty$ gives $\operatorname{vol}(\mathcal{T}(U)) = 0$. ■

We now turn this lemma into a statement about the original variables.

Lemma D.5. *There are null sets $\mathcal{B}_1, \mathcal{B}_2, \ldots \subset \Omega$ with the following property. Let w be as in Lemma D.2, and suppose that no limit point of $(\bar{w}_1(t), w_2(t))$ as $t \uparrow T$ is one of the two points in (D.5). Then there is n such that $w(t) \in \mathcal{B}_n$ for all t sufficiently close to T.*

Proof of Lemma D.5. For $n \ge 1$, let $\mathcal{K}_n$ be the set of points $(0, \bar{w}_1, w_2) \in \mathcal{Z}$ with $1/n \le \|w_2\|_2 \le n$ whose distance from the two excluded points is at least $1/n$. Each $\mathcal{K}_n$ is closed and bounded, hence compact. Let U_n be the neighborhood of $\mathcal{K}_n$ given by Lemma D.4. The map $(r_1, \bar{w}_1, w_2) \mapsto (r_1 \bar{w}_1, w_2)$ is smooth between manifolds of dimension $2d$, so it maps the null set $\mathcal{T}(U_n)$ onto a null set; let $\mathcal{B}_n$ be its intersection with Ω.

Let w be as in the lemma. By Lemma D.2, the set of limit points of $(r_1, \bar{w}_1, w_2)$ as $\tau \to \infty$ is a compact subset of $\mathcal{Z}$. It has positive distance from the two excluded points, so it lies in some $\mathcal{K}_n$. Since $(r_1, \bar{w}_1, w_2)$ stays bounded, it approaches this limit set as $\tau \to \infty$, and hence lies in U_n for all large τ. It therefore lies in $\mathcal{T}(U_n)$ for all large τ, and $w(t) \in \mathcal{B}_n$ for all t close to T. ■

D.1.3 The equilibria $(0, -v_1, v_2)$ and $(0, -v_2, v_1)$

At the two points in (D.5), both terms in (D.8) vanish, so Lemma D.4 does not apply. We treat $(\bar{w}_1, w_2) = (-v_1, v_2)$; exchanging v_1 and v_2 gives the other point. We show that a trajectory reaching this point in finite time does so at explicit rates, and that, after a blow-up of the variables by these rates, it lies on a manifold of dimension less than $2d$.

We write $\bar{w}_1 = \exp_{-v_1} z$ with $z \in v_1^\perp$, where $\exp_{-v_1}$ is the exponential map of the sphere at $-v_1$, whose tangent space there is $v_1^\perp$, and we let $\mathsf{G}(z)$ be the matrix of the round metric in these normal coordinates. Then $\mathsf{G}(0) = I$, the first derivatives of G vanish at $z = 0$, and $\mathsf{G}(z)z = z$ by the Gauss lemma. In rescaled time, the angular equation reads $z' = \mathsf{G}(z)^{-1}\nabla_z \Psi_{w_2}(\exp_{-v_1} z)$, and therefore $\langle z, z'\rangle = \langle z, \nabla_z \Psi_{w_2}(\exp_{-v_1} z)\rangle$. We write $K(u, v) = \|u\|_2\|v\|_2 l(\angle(u, v))$ for the kernel in (A.1a), so that $\Psi_{w_2}(\bar{w}_1) = K(\bar{w}_1, v_1) + K(\bar{w}_1, v_2) - K(\bar{w}_1, w_2)$. Since $\angle(\exp_{-v_1} z, v_1) = \pi - \|z\|_2$, the first term equals $l(\pi - \|z\|_2) = (\sin\|z\|_2 - \|z\|_2\cos\|z\|_2)/(2\pi) = \|z\|_2^3 a_0(\|z\|_2^2)$, and its gradient is $\|z\|_2 z\, a_1(\|z\|_2^2)$, where a_0 and a_1 are real analytic with $a_0(0) = 1/(6\pi)$ and $a_1(0) = 1/(2\pi)$. By (A.9),

$$G_0 := \nabla_{w_2} K(-v_1, w_2)\big|_{w_2 = v_2} = \frac{\sin\alpha\, v_2 - \alpha v_1}{2\pi}, \qquad F_b(v_2) = \frac{\sin\alpha\, v_2 + (\pi - \alpha)v_1}{2\pi},$$

and a direct computation gives $G_0^\top F_b(v_2) = -\nu_\alpha$, where

$$\nu_\alpha = \frac{\alpha(\pi - \alpha) - \sin^2\alpha - (\pi - 2\alpha)\sin\alpha\cos\alpha}{4\pi^2}.$$

The numerator vanishes at $\alpha = 0$ and has derivative $2(\pi - 2\alpha)\sin^2\alpha$, which is positive on $(0, \pi/2)$; hence $\nu_\alpha > 0$ for $\alpha \in (0, \pi/2]$. Since $F_b(v_2) \neq 0$, the flow $\dot{w}_2 = F_b(w_2)$ passes through v_2 with nonzero velocity, and we first straighten its orbit near v_2.

Lemma D.6 (Coordinates near v_2)**.** *There are a cubic polynomial curve Γ with $\Gamma(0) = v_2$, $\Gamma'(0) = -F_b(v_2)$, $\Gamma''(0) = DF_b(v_2)F_b(v_2)$ and $\Gamma'(s) = -F_b(\Gamma(s)) + O(s^3)$ as $s \downarrow 0$, and a linear isometry $J : \mathbb{R}^{d-1} \to F_b(v_2)^\perp$, such that $\chi(s, y) = \Gamma(s) + Jy$ is a real-analytic diffeomorphism from a neighborhood of $(0, 0)$ onto a neighborhood of v_2. Writing $w_2 = \chi(s, y)$ along the trajectory, the third equation in (D.4) reads*

$$\dot{s} = -1 + R_s(s, y) + r_1 B_s(s, y, z), \qquad \dot{y} = R_y(s, y) + r_1 B_y(s, y, z), \tag{D.9}$$

where $B = (B_s, B_y)$ is real analytic near the origin, $R = (R_s, R_y)$ is C^1, and, for some $C, \delta > 0$, $\|R(s, y)\|_2 \le C(s^3 + \|y\|_2)$ whenever $0 \le s \le \delta$ and $\|y\|_2 \le \delta$. Moreover, for every $\Lambda < \infty$, after decreasing δ, there is a real-analytic function $\widehat{R}$ on a neighborhood of $\{0\} \times \{\hat{y} : \|\hat{y}\|_2 \le \Lambda\}$ such that $R(s, s^3\hat{y}) = s^3\widehat{R}(s, \hat{y})$ for $0 \le s \le \delta$ and $\|\hat{y}\|_2 \le \Lambda$.

Proof of Lemma D.6. For $w_2 = v_2 + s\omega$ with $s > 0$ and $P_{v_2}^\perp \omega \neq 0$, (A.9) gives

$$\nabla_{w_2} K(w_2, v_2) = \frac{v_2}{2} + \frac{1}{2\pi}\left(\frac{s\|P_{v_2}^\perp\omega\|_2}{\|w_2\|_2^2}\, w_2 - \arctan\left(\frac{s\|P_{v_2}^\perp\omega\|_2}{1 + s\langle v_2, \omega\rangle}\right) v_2\right).$$

The right-hand side is real analytic in (s, ω) near $(0, -F_b(v_2))$, because $P_{v_2}^\perp F_b(v_2) = \frac{\pi - \alpha}{2\pi} P_{v_2}^\perp v_1 \neq 0$. Replacing $\nabla_{w_2} K(w_2, v_2)$ by it in $F_b(w_2) = -w_2/2 + \nabla_{w_2} K(w_2, v_1) + \nabla_{w_2} K(w_2, v_2)$ defines a real-analytic function $\widetilde{F}(s, \omega)$ near $(0, -F_b(v_2))$ with $\widetilde{F}(s, \omega) = F_b(v_2 + s\omega)$ for $s > 0$ and $\widetilde{F}(0, \omega) = F_b(v_2)$ for all ω. In particular, $D_\omega \widetilde{F}(0, \cdot)$ and $D_\omega^2 \widetilde{F}(0, \cdot)$ vanish, and $\partial_s \widetilde{F}(0, \omega) = DF_b(v_2)\omega$, since F_b is C^1 by Lemma A.1.

Define

$$\omega_1 = -\frac{1}{2}\partial_s \widetilde{F}(0, -F_b(v_2)), \qquad \omega_2 = -\frac{1}{3}\Big(\frac{1}{2}\partial_s^2 \widetilde{F}(0, -F_b(v_2)) + D_\omega \partial_s \widetilde{F}(0, -F_b(v_2))[\omega_1]\Big),$$

and $\Gamma(s) = v_2 + s\,\omega(s)$ with $\omega(s) = -F_b(v_2) + s\omega_1 + s^2\omega_2$. Expanding $\widetilde{F}(s, \omega(s))$ in s, and using that $D_\omega \widetilde{F}(0, \cdot)$ and $D_\omega^2 \widetilde{F}(0, \cdot)$ vanish, we find that the coefficients of 1, s and s^2 in $\widetilde{F}(s, \omega(s)) + \Gamma'(s)$ vanish. This proves $\Gamma'(s) = -F_b(\Gamma(s)) + O(s^3)$, and $\Gamma''(0) = 2\omega_1 = DF_b(v_2)F_b(v_2)$.

Fix a linear isometry $J : \mathbb{R}^{d-1} \to F_b(v_2)^\perp$. The matrix $[\Gamma'(s)\ J]$ is invertible for small s, because $\Gamma'(0)$ is orthogonal to the range of J. By the inverse function theorem, χ is a real-analytic diffeomorphism near $(0,0)$. Define

$$R(s,y) = [\Gamma'(s)\ J]^{-1}\big(F_b(\chi(s,y)) + \Gamma'(s)\big),$$
$$B(s,y,z) = [\Gamma'(s)\ J]^{-1}\nabla_{w_2}\Psi_{w_2}(\exp_{-v_1} z)\big|_{w_2=\chi(s,y)}.$$

Since $\dot w_2 = \Gamma'(s)\dot s + J\dot y$, the third equation in (D.4) becomes (D.9). The function R is C^1 because F_b is. The function B is real analytic, because $\nabla_{w_2}\Psi_{w_2}(\bar w_1) = -\nabla_{w_2}K(\bar w_1, w_2)$ and the angle between $\exp_{-v_1} z$ and $\chi(s,y)$ stays near $\pi - \alpha \in (0,\pi)$, where K is analytic. By the property of Γ, $R(s,0) = O(s^3)$, and R is Lipschitz in y; this gives the bound on R.

Finally, for bounded $\hat y$ we have $\chi(s, s^3\hat y) = v_2 + s\big(\omega(s) + s^2 J\hat y\big)$. Since $D_\omega\widetilde F(0,\cdot)$ and $D^2_\omega\widetilde F(0,\cdot)$ vanish, the coefficients of 1, s and s^2 in $\widetilde F(s, \omega(s) + s^2 J\hat y) + \Gamma'(s)$ do not depend on $\hat y$, and so they vanish. Dividing this analytic function by s^3 and multiplying by $[\Gamma'(s)\ J]^{-1}$ defines $\widehat R(s,\hat y)$, which satisfies $R(s, s^3\hat y) = s^3\widehat R(s,\hat y)$ for $s > 0$. ∎

Along a trajectory that reaches $(0, -v_1, v_2)$ at time T, we will see in Lemma D.8 that $\Psi_{w_2}(\bar w_1)$ behaves like $-\nu_\alpha(T-t)$ and r_1 like $\nu_\alpha(T-t)^2/2$, that s is close to $T-t$, and that $\|z\|_2$ and $\|y\|_2$ are at most of order $(T-t)^{1/2}$ and $(T-t)^3$. We therefore blow up the variables by these powers. When $\alpha = \pi/2$, the constant term in the angular equation of the blow-up vanishes and $\|z\|_2$ is at most of order $T-t$, so this case uses different powers. The next lemma gives the expansions and the equations in the blow-up variables.

Lemma D.7 (Blow-up near $(0,-v_1,v_2)$)**.** *Let $h(z) = D_{w_2}K(\exp_{-v_1} z, w_2)[F_b(v_2)]\big|_{w_2=v_2}$ and $\beta_0 = B_y(0,0,0)$. Then $\nabla h(0) = -\frac{(\pi-2\alpha)\sin\alpha}{4\pi^2}(v_2 - \cos\alpha\, v_1)$, and if $\alpha = \pi/2$, then $\nabla^2 h(0) = \frac{1}{16}I + \frac{1}{4\pi^2}v_2 v_2^\top$, where I is the identity on $v_1^\perp$. For bounded $\hat y$ and (z,s) near $(0,0)$,*

$$\Psi_{\chi(s,s^3\hat y)}(\exp_{-v_1} z) = l(\pi - \|z\|_2) + s\, g(z,s,\hat y), \tag{D.10}$$

where g is real analytic, $g(z,0,\hat y) = h(z)$, and, if $\alpha = \pi/2$, $\partial_s\nabla_z g(0,0,\hat y) = v_2/(16\pi^2)$.

Fix $\Lambda < \infty$, and let $\|\hat z\|_2, |\hat r|, \|\hat y\|_2 \le \Lambda$. Suppose first that $0 < \alpha < \pi/2$. If we set $s = \zeta^2$, $z = \zeta\hat z$, $r_1 = \zeta^4\hat r$ and $y = \zeta^6\hat y$ with $\zeta > 0$, and change time by $\mathrm{d}\sigma = \zeta\,\mathrm{d}\tau$, the rescaled-time equations for (s, z, r_1, y) become

$$\begin{aligned} \frac{\mathrm{d}\zeta}{\mathrm{d}\sigma} &= -\tfrac12\zeta^2\hat r + \zeta^6 E_1, & \frac{\mathrm{d}\hat z}{\mathrm{d}\sigma} &= q_1(\hat z) + \zeta E_2, \\ \frac{\mathrm{d}\hat r}{\mathrm{d}\sigma} &= \zeta\hat r(2\hat r - \nu_\alpha) + \zeta^2 E_3, & \frac{\mathrm{d}\hat y}{\mathrm{d}\sigma} &= \zeta(\hat r^2\beta_0 + 3\hat r\hat y) + \zeta^2 E_4, \end{aligned} \tag{D.11}$$

where $q_1(\hat z) = \|\hat z\|_2\hat z/(2\pi) + \nabla h(0)$. If $\alpha = \pi/2$, we set instead $z = s\hat z$, $r_1 = s^2\hat r$ and $y = s^3\hat y$ with $s > 0$, and $\mathrm{d}\sigma = s\,\mathrm{d}\tau$; this gives

$$\begin{aligned} \frac{\mathrm{d}s}{\mathrm{d}\sigma} &= -s\hat r + s^3 E_1, & \frac{\mathrm{d}\hat z}{\mathrm{d}\sigma} &= q_2(\hat z, \hat r) + sE_2, \\ \frac{\mathrm{d}\hat r}{\mathrm{d}\sigma} &= \hat r(2\hat r - \nu_\alpha) + sE_3, & \frac{\mathrm{d}\hat y}{\mathrm{d}\sigma} &= \hat r^2\beta_0 + 3\hat r\hat y + sE_4, \end{aligned} \tag{D.12}$$

where $q_2(\hat z, \hat r) = \|\hat z\|_2\hat z/(2\pi) + (\nabla^2 h(0) + \hat r I)\hat z + v_2/(16\pi^2)$. In the first case, the remainders $E_1, \dots, E_4$ extend to C^1 functions on a neighborhood of $\{\zeta = 0\}$, including negative values of ζ; in the second case, the same holds with s in place of ζ.

Proof of Lemma D.7. By (A.9), $h(z) = \big(\sin\theta_{12}\langle v_2, F_b(v_2)\rangle + (\pi - \theta_{12})\langle \bar w_1, F_b(v_2)\rangle\big)/(2\pi)$, where $\bar w_1 = \exp_{-v_1} z$. At $z = 0$, $\bar w_1 = -v_1$ and $\theta_{12} = \pi - \alpha$, so $h(0) = G_0^\top F_b(v_2) = -\nu_\alpha$. Since $\bar w_1 = -v_1 + z + O(\|z\|_2^2)$ and $\cos\theta_{12} = -\cos\alpha + \langle z, v_2\rangle + O(\|z\|_2^2)$, the differential of θ_{12} at $z = 0$ is $-\langle v_2, \mathrm{d}z\rangle/\sin\alpha$. Differentiating h at $z = 0$, with $\langle v_2, F_b(v_2)\rangle = (\sin\alpha + (\pi-\alpha)\cos\alpha)/(2\pi)$ and $\langle v_1, F_b(v_2)\rangle = ((\pi-\alpha) + \sin\alpha\cos\alpha)/(2\pi)$, gives the formula for $\nabla h(0)$. When $\alpha = \pi/2$, expanding h to second order in the same way gives $h(z) = -\nu_\alpha + \|z\|_2^2/32 + \langle v_2, z\rangle^2/(8\pi^2) + O(\|z\|_2^3)$, and hence the formula for $\nabla^2 h(0)$.

The terms $K(\exp_{-v_1} z, v_2) - K(\exp_{-v_1} z, \chi(s, s^3\hat{y}))$ of $\Psi_{\chi(s,s^3\hat{y})}(\exp_{-v_1} z)$ are real analytic near $(z, s) = (0, 0)$, because both angles stay near $\pi - \alpha$, and they vanish at $s = 0$. Hence they equal $s\, g(z, s, \hat{y})$ with g real analytic, and $g(z, 0, \hat{y}) = -D_{w_2} K(\exp_{-v_1} z, v_2)[\Gamma'(0)] = h(z)$. Since $\hat{y}$ enters $\chi(s, s^3\hat{y})$ only at order s^3, a second-order expansion in s gives

$$\partial_s g(z, 0, \hat{y}) = -\frac{1}{2}\Big(D_{w_2} K(\exp_{-v_1} z, v_2)[\Gamma''(0)] + D^2_{w_2} K(\exp_{-v_1} z, v_2)[F_b(v_2), F_b(v_2)]\Big).$$

When $\alpha = \pi/2$, we have $\Gamma''(0) = DF_b(v_2)F_b(v_2) = (\frac{1}{4\pi} - \frac{1}{8})v_1 - \frac{1}{4\pi}v_2$, and at $z = 0$ the gradients in z of the two terms in the bracket are $-v_2/(8\pi^2)$ and 0; hence $\partial_s \nabla_z g(0, 0, \hat{y}) = v_2/(16\pi^2)$.

By the expansion of $l(\pi - \|z\|_2)$ given at the beginning of this subsection and $\mathsf{G}(z)^{-1} z = z$, the angular equation is $z' = \|z\|_2 z\, a_1(\|z\|_2^2) + s\, \mathsf{G}(z)^{-1} \nabla_z g(z, s, \hat{y})$, with $y = s^3\hat{y}$. Because $\mathsf{G}(0) = I$ and $D\mathsf{G}(0) = 0$, the function $\mathsf{G}^{-1}\nabla_z g$ equals $\nabla h(0)$ at $(0, 0, \hat{y})$, and its derivatives in z and s there are $\nabla^2 h(0)$ and $\partial_s \nabla_z g(0, 0, \hat{y})$.

Suppose that $0 < \alpha < \pi/2$, and substitute $z = \zeta\hat{z}$ and $s = \zeta^2$. The functions $g(\zeta\hat{z}, \zeta^2, \hat{y}) - g(0, 0, \hat{y})$ and $\mathsf{G}^{-1}\nabla_z g(\zeta\hat{z}, \zeta^2, \hat{y}) - \nabla h(0)$ are analytic and vanish at $\zeta = 0$, so they are ζ times analytic functions; also, $a_1(\zeta^2\|\hat{z}\|_2^2) - 1/(2\pi)$ is ζ^2 times an analytic function, and $\hat{z} \mapsto \|\hat{z}\|_2^3$ and $\hat{z} \mapsto \|\hat{z}\|_2 \hat{z}$ are C^1. Hence $\Psi_{w_2}(\bar{w}_1) = \zeta^2(-\nu_\alpha + \zeta k_1)$ and $z' = \zeta^2(q_1(\hat{z}) + \zeta k_2)$, where k_1 and k_2 are C^1 functions of $(\zeta, \hat{z}, \hat{y})$ on a neighborhood of $\zeta = 0$. Suppose now that $\alpha = \pi/2$, and substitute $z = s\hat{z}$. Then $\mathsf{G}^{-1}\nabla_z g(s\hat{z}, s, \hat{y})$ vanishes at $s = 0$, and its derivative in s there is $\nabla^2 h(0)\hat{z} + v_2/(16\pi^2)$. Taylor's formula with integral remainder gives, in the same way, $\Psi_{w_2}(\bar{w}_1) = s(-\nu_\alpha + s k_1)$ and $z' = s^2(q_2(\hat{z}, 0) + s k_2)$, with k_1 and k_2 C^1 on a neighborhood of $s = 0$.

By (D.4), (D.9) and Lemma D.6, the remaining rescaled-time equations are

$$s' = r_1\big(-1 + s^3\widehat{R}_s + r_1 B_s\big), \qquad y' = r_1\big(s^3\widehat{R}_y + r_1 B_y\big), \qquad r_1' = r_1\big(\Psi_{w_2}(\bar{w}_1) - r_1/2\big).$$

We substitute the blow-up variables and divide by ζ to pass to the time σ (by s when $\alpha = \pi/2$), and we write $B_y = \beta_0 + \zeta \cdot (\text{analytic})$ (or $B_y = \beta_0 + s \cdot (\text{analytic})$ when $\alpha = \pi/2$). For example, $s = \zeta^2$ gives $2\zeta\zeta' = -\zeta^4\hat{r}(1 - \zeta^6\widehat{R}_s - \zeta^4\hat{r}B_s)$, so $\mathrm{d}\zeta/\mathrm{d}\sigma = \zeta'/\zeta = -\frac{1}{2}\zeta^2\hat{r} + \zeta^6 E_1$. The other equations follow in the same way from $\hat{z}' = (z' - \zeta'\hat{z})/\zeta$, $\hat{r}' = (r_1' - 4\zeta^3\zeta'\hat{r})/\zeta^4$ and $\hat{y}' = (y' - 6\zeta^5\zeta'\hat{y})/\zeta^6$, and their analogues for $\alpha = \pi/2$; the term $\hat{r}\hat{z}$ in q_2 comes from $-s'\hat{z}/s$. Each E_i is a sum of products of analytic functions, k_1, k_2, and polynomials in the blow-up variables. These expressions remain defined and C^1 for $\zeta \le 0$ (for $s \le 0$ when $\alpha = \pi/2$), which gives the stated extension. ∎

Lemma D.8 (Rates of approach to $(0, -v_1, v_2)$)**.** *Let $w(0) \in \Omega$ with $T < \infty$, and suppose that $r_1(t) \to 0$, $\bar{w}_1(t) \to -v_1$ and $w_2(t) \to v_2$ as $t \uparrow T$. Then, as $t \uparrow T$,*

$$\frac{s(t)}{T - t} \to 1, \qquad \frac{r_1(t)}{(T - t)^2} \to \frac{\nu_\alpha}{2}, \qquad \frac{y(t)}{(T - t)^3} \to -\frac{\nu_\alpha}{6}\beta_0,$$

and there are $M < \infty$ and $t_0 < T$ such that $\|z(t)\|_2 \le M\sqrt{T - t}$ for $t_0 \le t < T$. If $\alpha = \pi/2$, then also $\|z(t)\|_2 \le M(T - t)$ for $t_0 \le t < T$.

Proof of Lemma D.8. Throughout the proof, t is close to T, and C denotes a constant that may change from line to line. By the third equation in (D.4), $\dot{w}_2 \to F_b(v_2)$, because $r_1 \to 0$ and $\nabla_{w_2}\Psi$ is bounded near $(-v_1, v_2)$; so $(w_2(t) - v_2)/(T - t) \to -F_b(v_2)$ and $\|w_2(t) - v_2\|_2 \le C(T - t)$.

The difference $K(\exp_{-v_1} z, v_2) - K(\exp_{-v_1} z, w_2)$ is real analytic near $(z, w_2) = (0, v_2)$, vanishes at $w_2 = v_2$, and has gradient $-G_0$ in w_2 at $(0, v_2)$. Together with the expansion of $l(\pi - \|z\|_2)$ given at the beginning of this subsection, this gives $\delta > 0$ and $C_1 < \infty$ such that, for $\|z\|_2 \le \delta$ and $\|w_2 - v_2\|_2 \le \delta$,

$$\begin{aligned}\Big|\Psi_{w_2}(\exp_{-v_1} z) - \frac{\|z\|_2^3}{6\pi} + G_0^\top(w_2 - v_2)\Big| &\le C_1\big(\|z\|_2^5 + \|z\|_2\|w_2 - v_2\|_2 + \|w_2 - v_2\|_2^2\big),\\ \Big\|\nabla_z \Psi_{w_2}(\exp_{-v_1} z) - \frac{\|z\|_2 z}{2\pi}\Big\|_2 &\le C_1\big(\|w_2 - v_2\|_2 + \|z\|_2^3\big).\end{aligned} \tag{D.13}$$

By the first inequality, $\Psi_{w_2(t)}(\bar{w}_1(t)) \ge -C(T - t)$, since $\|z\|_2^3/(6\pi) - C_1\|z\|_2^5 \ge 0$ for small $\|z\|_2$. Solving the first equation in (D.4) backward from $r_1(T) = 0$ gives

$$r_1(t) = -\int_t^T e^{(t' - t)/2}\Psi_{w_2(t')}(\bar{w}_1(t'))\,\mathrm{d}t', \tag{D.14}$$

and hence $0 < r_1(t) \le C(T-t)^2$. By (D.9), $\dot{s} \to -1$, and $s(t) \to 0$ because $w_2 \to v_2 = \chi(0,0)$. Integrating (D.9) from t to T, and using the bound on R in Lemma D.6 and the boundedness of B, we obtain

$$|s(t) - (T-t)| + \|y(t)\|_2 \le C \int_t^T \big(r_1(t') + (T-t')^3 + \|y(t')\|_2\big)\,\mathrm{d}t'.$$

Since $r_1(t') \le C(T-t')^2$, Grönwall's inequality gives $|s(t) - (T-t)| + \|y(t)\|_2 \le C(T-t)^3$. In particular, $s(t)/(T-t) \to 1$, and $(T-t)/2 \le s(t) \le 2(T-t)$.

Next we bound $\|z\|_2$. By the second inequality in (D.13) and the identity for $\langle z, z'\rangle$ given at the beginning of this subsection, the upper right Dini derivative of $\|z\|_2$ in rescaled time satisfies $D_\tau^+\|z\|_2 \ge c_2\|z\|_2^2 - C_2 s$ for some $c_2, C_2 > 0$. Let $m_1 = \|z\|_2/\sqrt{s}$. Since $s' = -r_1(1 - R_s - r_1 B_s) \le 0$, the factor $s^{-1/2}$ does not decrease, so $D_\tau^+ m_1 \ge \sqrt{s}\,(c_2 m_1^2 - C_2)$. Choose M_1 with $c_2 M_1^2 \ge 2C_2$. If $m_1 \ge M_1$ at some time t_1, then m_1 stays above M_1, and $D_\tau^+ m_1 \ge (c_2/2)\sqrt{s}\,m_1^2$ gives

$$\frac{1}{m_1(t)} \le \frac{1}{m_1(t_1)} - \frac{c_2}{2}\int_{t_1}^t \frac{\sqrt{s(t')}}{r_1(t')}\,\mathrm{d}t'.$$

Since $\sqrt{s}/r_1 \ge C^{-1}(T-t)^{-3/2}$, the integral tends to infinity as $t \uparrow T$, and the right-hand side becomes negative, which is impossible. Hence $m_1 < M_1$, and $\|z\|_2 \le M_1\sqrt{s} \le M_1\sqrt{2(T-t)}$.

Dividing the first inequality in (D.13) by $T-t$, and using $\|z\|_2^3/(T-t) \le C\sqrt{T-t} \to 0$ and $(w_2 - v_2)/(T-t) \to -F_b(v_2)$, we obtain $\Psi_{w_2}(\bar{w}_1)/(T-t) \to G_0^\top F_b(v_2) = -\nu_\alpha$. Substituting $t' = t + \lambda(T-t)$ in (D.14) gives

$$\frac{r_1(t)}{(T-t)^2} = -\int_0^1 e^{\lambda(T-t)/2}\,\frac{\Psi_{w_2(t')}(\bar{w}_1(t'))}{T-t}\,\mathrm{d}\lambda.$$

For fixed $\lambda \in [0,1)$, the quotient in the integrand equals $(1-\lambda)\Psi_{w_2(t')}(\bar{w}_1(t'))/(T-t')$, which tends to $-(1-\lambda)\nu_\alpha$, and these quotients are bounded. Dominated convergence gives $r_1(t)/(T-t)^2 \to \nu_\alpha \int_0^1 (1-\lambda)\,\mathrm{d}\lambda = \nu_\alpha/2$. Finally, $y(t) = -\int_t^T \big(R_y(s,y) + r_1 B_y(s,y,z)\big)\,\mathrm{d}t'$. The integral of R_y is $O((T-t)^4)$ by the bound on R. In the integral of $r_1 B_y$, the substitution $t' = t + \lambda(T-t)$, the limit of $r_1/(T-t)^2$, the limit $B_y \to \beta_0$, and dominated convergence give $y(t)/(T-t)^3 \to -\frac{\nu_\alpha}{2}\beta_0 \int_0^1 (1-\lambda)^2\,\mathrm{d}\lambda = -\frac{\nu_\alpha}{6}\beta_0$.

It remains to treat $\alpha = \pi/2$, where $\nabla h(0) = 0$. Write $y = s^3\hat{y}$; by the bounds above, $\hat{y}$ is bounded. By (D.10) and Taylor's formula for $\nabla_z g$ at $(0,0,\hat{y})$,

$$\nabla_z \Psi_{w_2}(\exp_{-v_1} z) = \|z\|_2 z\, a_1(\|z\|_2^2) + s\nabla^2 h(0) z + \frac{s^2}{16\pi^2} v_2 + O\big(s\|z\|_2^2 + s^2\|z\|_2 + s^3\big),$$

uniformly in $\hat{y}$. Since $\nabla^2 h(0)$ is positive definite and $\|z\|_2 \le M_1\sqrt{s}$, this gives $D_\tau^+\|z\|_2 \ge c_3\|z\|_2^2 - C_3 s^2$ for some $c_3, C_3 > 0$. Let $m_2 = \|z\|_2/s$. As before, $D_\tau^+ m_2 \ge s(c_3 m_2^2 - C_3)$, and $\int_{t_1}^t s(t')/r_1(t')\,\mathrm{d}t' \to \infty$ because $s/r_1 \ge C^{-1}(T-t)^{-1}$. The argument used for m_1 gives $\|z\|_2 \le M_2 s \le 2M_2(T-t)$. The lemma holds with M the largest of the constants obtained. ∎

Proposition D.9 (Trajectories reaching $(0, -v_1, v_2)$)**.** *There is a null set $\mathcal{B} \subset \Omega$ with the following property. If $w(0) \in \Omega$, $T < \infty$, and $r_1(t) \to 0$, $\bar{w}_1(t) \to -v_1$ and $w_2(t) \to v_2$ as $t \uparrow T$, then $w(t) \in \mathcal{B}$ for all t sufficiently close to T.*

Proof of Proposition D.9. Suppose first that $0 < \alpha < \pi/2$, and consider the blow-up variables of (D.11). By Lemma D.8, as $t \uparrow T$ we have $\zeta = \sqrt{s} \to 0$, $\hat{r} = r_1/s^2 \to \nu_\alpha/2$, $\hat{y} = y/s^3 \to -\nu_\alpha\beta_0/6$, and $\hat{z} = z/\sqrt{s}$ stays bounded. Moreover, $\mathrm{d}\sigma/\mathrm{d}t = \sqrt{s}/r_1$ behaves like $(2/\nu_\alpha)(T-t)^{-3/2}$, so $\sigma \to \infty$. We apply Lemma D.7 with Λ larger than the bounds on $\|\hat{z}\|_2$, $|\hat{r}|$ and $\|\hat{y}\|_2$ along the trajectory. Then $Y = (\zeta, \hat{z}, \hat{r}, \hat{y})$ is a bounded solution of (D.11) for all large σ, with $\zeta(\sigma) \to 0$.

Since $(\pi - 2\alpha)\sin\alpha > 0$, we have $\nabla h(0) \ne 0$, and $q_1 = \nabla V_1$ for the strictly convex function $V_1(\hat{z}) = \|\hat{z}\|_2^3/(6\pi) + \langle \nabla h(0), \hat{z}\rangle$, whose only critical point is $\hat{z}_1^* = -(2\pi/\|\nabla h(0)\|_2)^{1/2}\,\nabla h(0)$. Along a solution of $\mathrm{d}\hat{z}/\mathrm{d}\sigma = q_1(\hat{z})$, the derivative of $\frac{1}{2}\|\hat{z} - \hat{z}_1^*\|_2^2$ is $\langle \hat{z} - \hat{z}_1^*, \nabla V_1(\hat{z}) - \nabla V_1(\hat{z}_1^*)\rangle$. By the monotonicity of ∇V_1 along rays from $\hat{z}_1^*$ and the compactness of the unit sphere, this derivative

has a positive lower bound on $\{\|\hat{z} - \hat{z}_1^*\|_2 \geq a\}$ for each $a > 0$. Hence no solution with $\hat{z}(\sigma') \neq \hat{z}_1^*$ at some σ' is defined and bounded on $[\sigma', \infty)$, so a solution that is defined and bounded on $\mathbb{R}$ is constant, equal to $\hat{z}_1^*$.

Let $\sigma_n \to \infty$. The translates $Y(\sigma_n + \cdot)$ are bounded and have bounded derivatives, so a subsequence converges uniformly on compact intervals to a solution of (D.11) on $\mathbb{R}$ with $\zeta \equiv 0$. Its $\hat{z}$-component is a bounded solution of $\mathrm{d}\hat{z}/\mathrm{d}\sigma = q_1(\hat{z})$ on $\mathbb{R}$, so it equals $\hat{z}_1^*$. Therefore $\hat{z}(\sigma) \to \hat{z}_1^*$, and $Y(\sigma)$ converges to $Y_1^* = (0, \hat{z}_1^*, \nu_\alpha/2, -\nu_\alpha\beta_0/6)$. At Y_1^*, the partial derivatives of the right-hand sides of the ζ-, $\hat{r}$- and $\hat{y}$-equations vanish: the first is $O(\zeta^2)$, and the other two are ζ times functions that vanish at Y_1^*, plus $O(\zeta^2)$. In the order $(\zeta, \hat{r}, \hat{y}, \hat{z})$, the Jacobian at Y_1^* is therefore block lower triangular, with a zero block of size $d+1$ and the block $Dq_1(\hat{z}_1^*) = (2\pi)^{-1}(\|\hat{z}_1^*\|_2 I + \hat{z}_1^*(\hat{z}_1^*)^\top/\|\hat{z}_1^*\|_2)$, which is positive definite. By the C^1 center-stable manifold theorem (Hirsch et al., 1977), there is a C^1 manifold $\mathcal{M}_1$ of dimension $d+1$ through Y_1^* that contains every forward orbit staying in a small neighborhood of Y_1^*. Hence $Y(\sigma) \in \mathcal{M}_1$ for all large σ.

Now let $\alpha = \pi/2$, so that $\nabla h(0) = 0$, and consider the blow-up variables of (D.12). By Lemma D.8, $s \to 0$, $\hat{r} = r_1/s^2 \to \nu_\alpha/2$, $\hat{y} = y/s^3 \to -\nu_\alpha\beta_0/6$, and $\hat{z} = z/s$ stays bounded. Moreover, $\mathrm{d}\sigma/\mathrm{d}t = s/r_1$ behaves like $(2/\nu_\alpha)(T-t)^{-1}$, so $\sigma \to \infty$. At $s = 0$ and $\hat{r} = \nu_\alpha/2$, the $\hat{z}$-equation is $\mathrm{d}\hat{z}/\mathrm{d}\sigma = \nabla V_2(\hat{z})$ for the strongly convex function

$$V_2(\hat{z}) = \frac{\|\hat{z}\|_2^3}{6\pi} + \frac{1}{2}\Big\langle \hat{z}, \Big(\nabla^2 h(0) + \frac{\nu_\alpha}{2} I\Big)\hat{z}\Big\rangle + \frac{\langle v_2, \hat{z}\rangle}{16\pi^2},$$

whose only critical point $\hat{z}_2^*$ is nonzero because $\nabla V_2(0) = v_2/(16\pi^2)$. The argument of the acute case shows that $Y = (s, \hat{z}, \hat{r}, \hat{y})$ converges to $Y_2^* = (0, \hat{z}_2^*, \nu_\alpha/2, -\nu_\alpha\beta_0/6)$. In the order $(s, \hat{r}, \hat{z}, \hat{y})$, the Jacobian at Y_2^* is block lower triangular with diagonal blocks $-\nu_\alpha/2$, ν_α, $D_{\hat{z}} q_2(\hat{z}_2^*, \nu_\alpha/2)$ and $(3\nu_\alpha/2)I$, where

$$D_{\hat{z}} q_2(\hat{z}_2^*, \nu_\alpha/2) = \frac{1}{2\pi}\Big(\|\hat{z}_2^*\|_2 I + \frac{\hat{z}_2^*(\hat{z}_2^*)^\top}{\|\hat{z}_2^*\|_2}\Big) + \nabla^2 h(0) + \frac{\nu_\alpha}{2} I$$

is positive definite. Hence Y_2^* has one stable direction and $2d-1$ unstable directions, and its stable manifold $\mathcal{M}_2$, a C^1 curve (Hirsch et al., 1977), contains $Y(\sigma)$ for all large σ.

Finally, we return to w. In the two cases, the maps

$$(\zeta, \hat{z}, \hat{r}, \hat{y}) \mapsto \big(\zeta^4 \hat{r} \exp_{-v_1}(\zeta\hat{z}), \chi(\zeta^2, \zeta^6\hat{y})\big), \qquad (s, \hat{z}, \hat{r}, \hat{y}) \mapsto \big(s^2 \hat{r} \exp_{-v_1}(s\hat{z}), \chi(s, s^3\hat{y})\big),$$

defined for $\zeta > 0$ and $s > 0$ respectively, are C^1 and send $Y(\sigma)$ to $w(t)$. The parts of $\mathcal{M}_1$ with $\zeta > 0$ and of $\mathcal{M}_2$ with $s > 0$ are C^1 manifolds of dimensions $d+1$ and 1, both less than $2d$ since $d \geq 2$. Their images under these maps are therefore null in $(\mathbb{R}^d)^2$. Let $\mathcal{B}$ be the intersection of the union of the two images with Ω. By the above, $w(t) \in \mathcal{B}$ for all t sufficiently close to T. ■

D.1.4 Avoidance of $\partial\Omega$

The results above give null sets that contain $w(t)$ for all t close to T. The next lemma turns them into null sets of initial points.

Lemma D.10 (Preimages of null sets under the flow)**.** *Let $\mathcal{A} \subset \Omega$ be a null set. The set of $w(0) \in \Omega$ for which $w(t) \in \mathcal{A}$ for all t in some nonempty open subinterval of $[0, T)$ is null.*

Proof of Lemma D.10. Such a trajectory lies in $\mathcal{A}$ at some rational time $t_* \geq 0$. Since $L \in C^2(\Omega)$ by Lemma A.1, the time-t_* map of gradient flow is a C^1 diffeomorphism from the open set of initial points with $T > t_*$ onto an open subset of Ω, with the backward flow as its inverse. Hence the set of $w(0)$ with $w(t_*) \in \mathcal{A}$ is null, and the union over rational t_* is null. ■

Proposition D.11 (Avoidance of $\partial\Omega$ in finite time)**.** *The set of $w(0) \in \Omega$ with $T < \infty$ is null.*

Proof of Proposition D.11. Let $w(0) \in \Omega$ with $T < \infty$. For $i = 1, 2$,

$$\|\nabla_{w_i} L(w)\|_2 = \big\|\mathbb{E}\big[(f(x) - f^*(x))\,\mathbb{1}(\langle w_i, x\rangle > 0)\,x\big]\big\|_2 \leq \big(\mathbb{E}(f-f^*)^2\big)^{1/2}\big(\mathbb{E}\|x\|_2^2\big)^{1/2} \leq \sqrt{2dL(w(0))}$$

along the trajectory, so $w(t)$ is Lipschitz on $[0, T)$ and has a limit $w(T)$. By Lemma A.1, $w(T) \notin \Omega$, since otherwise the solution would extend beyond T; and $w(T) \neq (0, 0)$ by Lemma D.1. Exchanging the two students if necessary, we may assume that $w_1(T) = 0 \neq w_2(T)$. Then Lemma D.2 applies.

If no limit point of $(\bar{w}_1(t), w_2(t))$ is one of the two points in (D.5), then by Lemma D.5, $w(t) \in \mathcal{B}_n$ for some n and all t close to T. Otherwise, say, $(-v_1, v_2)$ is a limit point. Then $w_2(T) = v_2$, and the limit ψ_* in Lemma D.2 equals $\Psi_{v_2}(-v_1) = 0$. Since $\Psi_{v_2}(u) = l(\angle(u, v_1))$ vanishes only at $u = -v_1$ by Lemma A.2, every limit point of $\bar{w}_1(t)$ equals $-v_1$. Hence $\bar{w}_1(t) \to -v_1$, and by Proposition D.9, $w(t)$ lies in its null set $\mathcal{B}$ for all t close to T. The point $(-v_2, v_1)$ is handled in the same way, with the teachers exchanged.

In all cases, $w(t)$ lies, for all t close to T, in one of countably many null sets: the sets $\mathcal{B}_n$, the set $\mathcal{B}$ of Proposition D.9 and its analogue with the teachers exchanged, and the images of all these sets under the exchange of the two students. Lemma D.10 completes the proof. ■

A trajectory that stays in Ω for all time can still approach $\partial\Omega$ as $t \to \infty$. We first show that every such trajectory converges, and then that its limit lies on $\partial\Omega$ only for a null set of initializations.

Proposition D.12 (Convergence of global solutions)**.** *Let $w(0) \in \Omega$ with $T = \infty$. Then the trajectory has finite length, it converges to some $w_\infty \in (\mathbb{R}^d)^2$, and $0 \in \partial L(w_\infty)$, where ∂L is the limiting (Mordukhovich) subdifferential. In particular, $\nabla L(w_\infty) = 0$ if $w_\infty \in \Omega$.*

Proof of Proposition D.12. Since $L(w(t))$ is nonincreasing and nonnegative, it has a limit L_∞, and since

$$\frac{\mathrm{d}}{\mathrm{d}t} L(w(t)) = -\|\nabla L(w(t))\|_2^2, \tag{D.15}$$

we have $\int_0^\infty \|\nabla L(w(t))\|_2^2 \,\mathrm{d}t < \infty$. By (D.1), the trajectory is bounded. Hence there are $t_n \to \infty$ with $\nabla L(w(t_n)) \to 0$ and $w(t_n) \to w_\infty$ for some w_∞.

For nonzero arguments, the kernel K is built from norms, inner products, algebraic operations and $\arccos$ on $[-1, 1]$, whose graph is the inverse graph of $\cos$ on $[0, \pi]$. Extending it by zero where an argument vanishes preserves subanalyticity, and the bound $0 \le K(u, v) \le \|u\|_2\|v\|_2/2$, which follows from Lemma A.2, gives continuity there. Hence L is continuous and globally subanalytic on $(\mathbb{R}^d)^2$. On Ω, $\partial L(w) = \{\nabla L(w)\}$, because L is C^1 there. By continuity, $L(w_\infty) = L_\infty$, and since $\nabla L(w(t_n)) \in \partial L(w(t_n))$ tends to zero while $L(w(t_n)) \to L(w_\infty)$, the closedness of the limiting subdifferential gives $0 \in \partial L(w_\infty)$.

Because w_∞ may lie on $\partial\Omega$, where L is not differentiable, we use the nonsmooth Łojasiewicz inequality for continuous subanalytic functions (Bolte et al., 2007, Theorem 3.1). It gives $r_0, C, \delta_0 > 0$ and $\mu \in [0, 1)$ such that

$$|L(w) - L_\infty|^\mu \le C \operatorname{dist}\big(0, \partial L(w)\big) \qquad \text{whenever } \|w - w_\infty\|_2 < r_0 \text{ and } 0 < |L(w) - L_\infty| < \delta_0.$$

If $L(w(t)) = L_\infty$ at some finite time, then by (D.15) the trajectory is constant from that time on. Otherwise, whenever $\|w(t) - w_\infty\|_2 < r_0$ and $L(w(t)) - L_\infty < \delta_0$,

$$-\frac{\mathrm{d}}{\mathrm{d}t}\big(L(w(t)) - L_\infty\big)^{1-\mu} = (1-\mu)\frac{\|\nabla L(w(t))\|_2^2}{(L(w(t)) - L_\infty)^\mu} \ge \frac{1-\mu}{C}\|\dot{w}(t)\|_2.$$

Choose n so large that $\|w(t_n) - w_\infty\|_2 < r_0/2$, $L(w(t_n)) - L_\infty < \delta_0$, and $\frac{C}{1-\mu}(L(w(t_n)) - L_\infty)^{1-\mu} < r_0/2$. If the trajectory left the ball $\{\|w - w_\infty\|_2 < r_0\}$ after t_n, integrating the last inequality up to the first exit time would show that its length after t_n is less than $r_0/2$, which is impossible. Hence it stays in the ball, and the same inequality gives $\int_{t_n}^\infty \|\dot{w}(t)\|_2 \,\mathrm{d}t < \infty$. The trajectory therefore converges, and its limit is w_∞. ■

Proposition D.13 (Avoidance of $\partial\Omega$ as $t \to \infty$)**.** *The set of $w(0) \in \Omega$ with $T = \infty$ whose trajectory converges to a point of $\partial\Omega$ is null.*

Proof of Proposition D.13. Let w be such a trajectory, with limit $w_\infty \in \partial\Omega$. By Lemma D.1, $w_\infty \ne (0, 0)$; exchanging the two students if necessary, we may assume that $w_\infty = (0, w_{2,*})$ with $w_{2,*} \ne 0$. By Proposition D.12, the trajectory has finite length, so Lemma D.2 applies with $T = \infty$. Moreover, by the third equation in (D.4), $\dot{w}_2(t) \to F_b(w_{2,*})$; since $w_2(t)$ converges, this limit is zero. On the other hand, $F_b(v_2) \ne 0$ by the formula in Appendix D.1.3, and $F_b(v_1) \ne 0$ by symmetry. Hence $w_{2,*} \notin \{v_1, v_2\}$, no limit point of $(\bar{w}_1(t), w_2(t))$ is one of the two points in (D.5), and by Lemma D.5, $w(t) \in \mathcal{B}_n$ for some n and all large t. Together with the images of the sets $\mathcal{B}_n$ under the exchange of the two students, Lemma D.10 completes the proof. ■

D.2 Saddle avoidance and global convergence

It remains to show that trajectories converge to strict saddles in Ω only for a null set of initializations.

Proposition D.14 (Avoidance of strict saddles)**.** *The set of $w(0) \in \Omega$ with $T = \infty$ whose trajectory converges to a strict saddle of L in Ω, that is, a stationary point in Ω at which $\nabla^2 L$ has a negative eigenvalue, is null.*

Proof of Proposition D.14. For $n \geq 1$, choose a function in $C_c^\infty(\Omega)$ equal to one on a neighborhood of $\{w : \|w\|_2 \leq n,\ \min(r_1, r_2) \geq 1/n\}$, and let L_n be its product with L, extended by zero outside Ω. Then L_n is C^2 on $(\mathbb{R}^d)^2$ by Lemma A.1, and its gradient has compact support, so its gradient flow is defined for all time. By Bah et al. (2022, Theorem 28), the set $\mathcal{S}_n$ of initial points whose L_n-gradient flow converges to a strict saddle of L_n is null.

Suppose that a trajectory converges to a strict saddle $w_\infty \in \Omega$ of L. Then there are n and t_1 such that w_∞ and $w(t)$ for all $t \geq t_1$ lie in the open set where $L_n = L$. Hence w_∞ is a strict saddle of L_n, and for every $t \geq t_1$ the L_n-gradient flow starting at $w(t)$ follows the original trajectory, so $w(t) \in \mathcal{S}_n$. Lemma D.10, applied to $\bigcup_n \mathcal{S}_n \cap \Omega$, completes the proof. ∎

Proof of Theorem 4.2. The initializations with $w_1(0) = 0$ or $w_2(0) = 0$ form a null set, so we consider $w(0) \in \Omega$. Outside the null set of Proposition D.11, $T = \infty$, that is, the solution exists for all $t \geq 0$ and both weights stay nonzero. By Proposition D.12, it converges to some w_∞ with $0 \in \partial L(w_\infty)$. Outside the null set of Proposition D.13, $w_\infty \in \Omega$, and then $\nabla L(w_\infty) = 0$. By Theorem 3.3, every stationary point in Ω outside $\mathcal{G}$ has a strictly negative Hessian direction, that is, it is a strict saddle, and by Proposition D.14 convergence to a strict saddle happens only on a null set. Hence, for almost every initialization, the solution exists for all $t \geq 0$, both weights stay nonzero, and $w(t)$ converges to (v_1, v_2) or (v_2, v_1). ∎

E Approach to the bisector saddle

As in Lemma 4.1, we consider $\alpha = \angle(v_1, v_2) \in (0, \pi/2]$. We recall the set Ω of pairs (w_1, w_2) with $w_1, w_2 \neq 0$ and the notation $r_i = \|w_i\|_2$ from Appendix D. For $w(0) \in \Omega$, we write $[0, T_{\max})$ for the maximal interval on which the gradient-flow solution exists in Ω; it is well defined by Lemma A.1. By (3.4) and (A.1b), $r_- = 4l(\pi - \alpha/2)$ and $r_+ = 4l(\alpha/2)$, so Lemma A.2 gives $0 < r_- < 2$ and $r_+ \leq 2$. For unit vectors $p, q \in \mathbb{R}^d$, we also use

$$\angle(p, q) = 2 \arcsin \frac{\|p - q\|_2}{2} \leq \frac{\pi}{2} \|p - q\|_2, \tag{E.1}$$

which holds because $\arcsin$ is convex on $[0, 1]$, with $\arcsin 0 = 0$ and $\arcsin 1 = \pi/2$.

E.1 Norm bounds

Lemma E.1. *Let $w(0) \in \Omega$. For $0 \leq t < T_{\max}$ and $i = 1, 2$, we have $r_i(t) \leq r_i(0) + t$. If, in addition, $r_i(0) \leq 2$, then $r_i(t) \leq 2$ for $0 \leq t < T_{\max}$.*

Proof of Lemma E.1. By Lemma A.5, $\dot{r}_i \leq 1 - r_i/2 \leq 1$, and integration gives the first bound. The same inequality reads $\frac{\mathrm{d}}{\mathrm{d}t}(r_i - 2) \leq -(r_i - 2)/2$, so Grönwall's inequality gives $r_i(t) - 2 \leq (r_i(0) - 2)e^{-t/2} \leq 0$. ∎

Lemma E.2. *Let $w(0) \in \Omega$ and $T > 0$ satisfy $\max_{i=1,2} r_i(0) + T \leq r_-/4$. Then $T_{\max} > T$, and*

$$r_i(0) + \frac{r_-}{4} t \leq r_i(t) \leq r_i(0) + t, \qquad 0 \leq t \leq T, \quad i = 1, 2.$$

Proof of Lemma E.2. Let $0 \leq t < \min\{T, T_{\max}\}$. By Lemma E.1, $r_i(t) \leq r_i(0) + t \leq r_-/4$ for $i = 1, 2$, so Lemma A.5 gives $\dot{r}_i(t) \geq r_-/2 - (r_1(t) + r_2(t))/2 \geq r_-/4$. Integration proves the two bounds for these t. If $T_{\max} \leq T$, the bounds keep $w(t)$, $0 \leq t < T_{\max}$, in the compact subset of Ω on which $r_i(0) \leq \|w_i\|_2 \leq r_-/4$ for $i = 1, 2$. By Lemma A.1, the solution then extends beyond $T_{\max}$, which is impossible. Hence $T_{\max} > T$, and the bounds hold at $t = T$ by continuity. ∎

E.2 Alignment with the bisector and balance of the norms

The alignment of $\bar{w}_i$ with $\bar{v}$ is driven by the component along $\bar{v}$ of the projected teacher force in (A.13).

Lemma E.3. *For a unit vector $q \in \mathbb{R}^d$, let*

$$\psi(q) := \Big\langle \bar{v}, P_q^{\perp}\big((\pi - \angle(q, v_1))v_1 + (\pi - \angle(q, v_2))v_2\big)\Big\rangle$$

be the component along $\bar{v}$ of the projected teacher force at q. Then $\psi(q) \geq 0$, with equality if and only if $q = \bar{v}$ or $q = -\bar{v}$.

Proof of Lemma E.3. We write $\theta_j = \angle(q, v_j)$ for $j = 1, 2$. Since $\bar{v} = (v_1 + v_2)/(2\cos(\alpha/2))$, we have $\langle \bar{v}, v_j\rangle = \cos(\alpha/2)$ and $\langle \bar{v}, q\rangle = (\cos\theta_1 + \cos\theta_2)/(2\cos(\alpha/2))$. Expanding $P_q^{\perp} = I - qq^{\top}$ therefore gives

$$2\cos\frac{\alpha}{2}\,\psi(q) = 2\cos^2\frac{\alpha}{2}\,(2\pi - \theta_1 - \theta_2) - (\cos\theta_1 + \cos\theta_2)\big((\pi - \theta_1)\cos\theta_1 + (\pi - \theta_2)\cos\theta_2\big).$$

By symmetry we may assume $\theta_1 \geq \theta_2$. We set $x = \theta_1 + \theta_2$ and $y = \theta_1 - \theta_2$. The triangle inequality for angles, applied to q, v_1, v_2 and to $-q, v_1, v_2$, gives $\alpha \leq x \leq 2\pi - \alpha$ and $0 \leq y \leq \alpha \leq \pi/2$. Since $\cos\theta_1 + \cos\theta_2 = 2\cos(x/2)\cos(y/2)$ and

$$(\pi - \theta_1)\cos\theta_1 + (\pi - \theta_2)\cos\theta_2 = (2\pi - x)\cos\frac{x}{2}\cos\frac{y}{2} + y\sin\frac{x}{2}\sin\frac{y}{2},$$

the identity above becomes

$$2\cos\frac{\alpha}{2}\,\psi(q) = 2(2\pi - x)\Big(\cos^2\frac{\alpha}{2} - \cos^2\frac{x}{2}\cos^2\frac{y}{2}\Big) - \frac{y}{2}\sin x\sin y. \tag{E.2}$$

As $x/2 \in [\alpha/2, \pi - \alpha/2]$, we have $\cos^2(x/2) \leq \cos^2(\alpha/2)$, so the right-hand side is at least

$$2(2\pi - x)\cos^2\frac{\alpha}{2}\sin^2\frac{y}{2} - \frac{y}{2}\sin x\sin y = 2\sin\frac{y}{2}\Big((2\pi - x)\cos^2\frac{\alpha}{2}\sin\frac{y}{2} - \frac{y}{2}\cos\frac{y}{2}\sin x\Big).$$

Since $0 \leq y/2 \leq \pi/4$ and $\tan u \geq u$ for $0 \leq u < \pi/2$, we have $\sin(y/2) \geq (y/2)\cos(y/2)$, and hence

$$2\cos\frac{\alpha}{2}\,\psi(q) \geq \frac{y}{2}\sin y\,\Big((2\pi - x)\cos^2\frac{\alpha}{2} - \sin x\Big). \tag{E.3}$$

For $x \leq \pi$, the last factor is at least $\pi\cos^2(\alpha/2) - 1 \geq \pi/2 - 1 > 0$, because $\cos^2(\alpha/2) \geq 1/2$. For $\pi < x \leq 2\pi - \alpha$, it is positive because $\sin x < 0$ and $2\pi - x > 0$. Since $\sin y \geq 0$, it follows that $\psi(q) \geq 0$.

If $q = \pm\bar{v}$, then $\theta_1 = \theta_2$, so the teacher force is a multiple of $v_1 + v_2$ and $P_q^{\perp}$ annihilates it; thus $\psi(q) = 0$. Conversely, suppose that $\psi(q) = 0$. Then (E.3) forces $y\sin y = 0$, that is, $y = 0$, and (E.2) reduces to $2(2\pi - x)(\cos^2(\alpha/2) - \cos^2(x/2)) = 0$. Since $x < 2\pi$, this gives $x = \alpha$ or $x = 2\pi - \alpha$. With $y = 0$, we have $\langle q, \bar{v}\rangle = \cos(x/2)/\cos(\alpha/2)$, which equals 1 in the first case and -1 in the second. Hence $q = \bar{v}$ or $q = -\bar{v}$. ∎

Lemma E.4. *For every $\eta \in (0, \pi/2)$ and $\varepsilon \in (0, 1]$, there is $\delta = \delta(\alpha, \eta, \varepsilon) > 0$, independent of d, such that $\psi(q) \geq \delta$ for every unit vector $q \in \mathbb{R}^d$ with $\|q - \bar{v}\|_2 \geq \varepsilon$ and $\angle(q, \bar{v}) \leq \pi - \eta$.*

Proof of Lemma E.4. Let $a = \langle q, \bar{v}\rangle$ and $b = \langle q, v_{\Delta}\rangle$, so that $a^2 + b^2 \leq 1$. By (2.4), $v_1 = \cos(\alpha/2)\bar{v} + \sin(\alpha/2)v_{\Delta}$ and $v_2 = \cos(\alpha/2)\bar{v} - \sin(\alpha/2)v_{\Delta}$, so $\cos\angle(q, v_1) = a\cos(\alpha/2) + b\sin(\alpha/2)$ and $\cos\angle(q, v_2) = a\cos(\alpha/2) - b\sin(\alpha/2)$. Expanding $P_q^{\perp} = I - qq^{\top}$ in the definition of ψ shows that $\psi(q) = \tilde{\psi}(a, b)$ for a continuous function $\tilde{\psi}$ on the closed unit disk $\mathcal{D} \subset \mathbb{R}^2$ that depends only on α. The two conditions on q read $a \leq 1 - \varepsilon^2/2$ and $a \geq -\cos\eta$. They define a compact set $\mathcal{K} \subset \mathcal{D}$, which is nonempty because $\varepsilon \leq 1$. Every point of $\mathcal{D}$ is the pair (a, b) of some unit vector in $\mathbb{R}^3$, so Lemma E.3, applied in $\mathbb{R}^3$ to two teachers at angle α, shows that $\tilde{\psi}$ is nonnegative on $\mathcal{D}$ and vanishes only at $(1, 0)$ and $(-1, 0)$. Neither point lies in $\mathcal{K}$. Hence $\delta := \min_{\mathcal{K}}\tilde{\psi}$ is positive, and it depends only on α, η and ε. ∎

Lemma E.5. *Let $w(0) \in \Omega$, and let $E_i := \frac{1}{2}\|\bar{w}_i - \bar{v}\|_2^2$ for $i = 1, 2$. For $0 \le t < T_{\max}$,*

$$\dot{E}_i \le -\frac{\psi(\bar{w}_i)}{2\pi r_i} + \frac{r_{3-i}}{2r_i} \le \frac{r_{3-i}}{2r_i}, \qquad i = 1, 2, \tag{E.4}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}(E_1 + E_2) \le 2 \max_{j=1,2} \frac{r_{3-j}}{r_j} (E_1 + E_2). \tag{E.5}$$

Proof of Lemma E.5. Since $E_i = 1 - \langle \bar{w}_i, \bar{v} \rangle$, (A.13) and (A.14) give

$$\dot{E}_i = -\frac{\psi(\bar{w}_i)}{2\pi r_i} + \frac{\pi - \gamma}{2\pi} \frac{r_{3-i}}{r_i} \langle \bar{v}, P_{\bar{w}_i}^{\perp} \bar{w}_{3-i} \rangle.$$

The first inequality in (E.4) follows from $0 \le \pi - \gamma \le \pi$ and $|\langle \bar{v}, P_{\bar{w}_i}^{\perp} \bar{w}_{3-i} \rangle| \le 1$, and the second from Lemma E.3. For (E.5), let $e_i = \bar{w}_i - \bar{v}$. Since $P_{\bar{w}_i}^{\perp} \bar{w}_i = 0$, we have $P_{\bar{w}_i}^{\perp} \bar{v} = -P_{\bar{w}_i}^{\perp} e_i$ and $P_{\bar{w}_i}^{\perp} \bar{w}_{3-i} = P_{\bar{w}_i}^{\perp}(e_{3-i} - e_i)$. Since $P_{\bar{w}_i}^{\perp}$ is an orthogonal projection, the Cauchy–Schwarz inequality gives

$$\left|\left\langle \bar{v}, P_{\bar{w}_i}^{\perp} \bar{w}_{3-i} \right\rangle\right| = \left|\left\langle P_{\bar{w}_i}^{\perp} e_i, P_{\bar{w}_i}^{\perp}(e_{3-i} - e_i) \right\rangle\right| \le \|e_i\|_2 \big(\|e_1\|_2 + \|e_2\|_2\big).$$

Dropping the nonpositive first term and summing over i, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}(E_1 + E_2) \le \frac{1}{2} \max_{j=1,2} \frac{r_{3-j}}{r_j} \big(\|e_1\|_2 + \|e_2\|_2\big)^2 \le \max_{j=1,2} \frac{r_{3-j}}{r_j} \big(\|e_1\|_2^2 + \|e_2\|_2^2\big),$$

which is (E.5). ∎

While the weights are small, the factor $1/r_i$ in the first term of (E.4) makes the alignment fast. The next lemma shows that it takes place on a time scale proportional to r_{init}.

Lemma E.6. *Let $\eta \in (0, \pi/2)$, $0 < m \le M$ and $\varepsilon \in (0, 1]$. There are constants $\Lambda, r_0 > 0$, depending only on $\alpha, \eta, m, M, \varepsilon$ and not on d, with the following property. If $w(0) \in \Omega$ satisfies*

$$r_{\mathrm{init}} \le r_0, \qquad m r_{\mathrm{init}} \le r_i(0) \le M r_{\mathrm{init}}, \qquad \angle(\bar{w}_i(0), \bar{v}) \le \pi - \eta, \qquad i = 1, 2,$$

then $T_{\max} > \Lambda r_{\mathrm{init}}$ and $\|\bar{w}_i(\Lambda r_{\mathrm{init}}) - \bar{v}\|_2 < \varepsilon$ for $i = 1, 2$.

Proof of Lemma E.6. Let $\delta = \delta(\alpha, \eta, \varepsilon/2)$ be the constant of Lemma E.4, and set

$$\Lambda := M\big(e^{6\pi/\delta} - 1\big), \qquad B := \frac{M + \Lambda}{2m}, \qquad r_0 := \min\left\{ \frac{r_-}{4(M + \Lambda)}, \frac{\delta}{4\pi B(M + \Lambda)}, \frac{\varepsilon^2}{8B\Lambda} \right\}.$$

Since $\max_i r_i(0) + \Lambda r_{\mathrm{init}} \le (M + \Lambda) r_{\mathrm{init}} \le r_-/4$, Lemma E.2 gives $T_{\max} > \Lambda r_{\mathrm{init}}$ and

$$m r_{\mathrm{init}} \le r_i(t) \le M r_{\mathrm{init}} + t \le (M + \Lambda) r_{\mathrm{init}}, \qquad 0 \le t \le \Lambda r_{\mathrm{init}}, \quad i = 1, 2.$$

In particular, $r_{3-i}/r_i \le 2B$ on this interval, and Lemma E.5 gives

$$\dot{E}_i \le -\frac{\psi(\bar{w}_i)}{2\pi r_i} + B \le B, \qquad 0 \le t \le \Lambda r_{\mathrm{init}}, \quad i = 1, 2. \tag{E.6}$$

Fix i, and suppose that $\|\bar{w}_i(t) - \bar{v}\|_2 \ge \varepsilon/2$ for all $t \in [0, \Lambda r_{\mathrm{init}}]$. At every such t with $\angle(\bar{w}_i(t), \bar{v}) \le \pi - \eta$, Lemma E.4, the bound $r_i \le (M + \Lambda) r_{\mathrm{init}}$ and the choice of r_0 give

$$\dot{E}_i \le -\frac{\delta}{2\pi(M + \Lambda) r_{\mathrm{init}}} + B \le -B < 0.$$

Let t_* be the supremum of the times $t \le \Lambda r_{\mathrm{init}}$ such that $\angle(\bar{w}_i, \bar{v}) \le \pi - \eta$ on $[0, t]$. By continuity, the angle condition holds on $[0, t_*]$, so E_i is decreasing there and $\dot{E}_i(t_*) < 0$. The angle $\angle(\bar{w}_i, \bar{v}) = \arccos(1 - E_i)$ increases with E_i. If $t_* < \Lambda r_{\mathrm{init}}$, then $E_i(t) < E_i(t_*) \le E_i(0)$ for t slightly larger than t_*, and the angle condition holds beyond t_*, contradicting the definition of t_*. Hence $t_* = \Lambda r_{\mathrm{init}}$, and $\psi(\bar{w}_i) \ge \delta$ on the whole interval. The first bound in (E.6), together with $r_i(t) \le M r_{\mathrm{init}} + t$, now gives

$$E_i(0) - E_i(\Lambda r_{\mathrm{init}}) \ge \int_0^{\Lambda r_{\mathrm{init}}} \left(\frac{\delta}{2\pi(M r_{\mathrm{init}} + t)} - B \right) \mathrm{d}t = \frac{\delta}{2\pi} \log\left(1 + \frac{\Lambda}{M}\right) - B\Lambda r_{\mathrm{init}} \ge 3 - \frac{\varepsilon^2}{8} > 2.$$

This is impossible, since $E_i \ge 0$ and $E_i(0) = 1 - \cos\angle(\bar{w}_i(0), \bar{v}) \le 1 + \cos\eta < 2$. Hence there is $t_i \in [0, \Lambda r_{\mathrm{init}}]$ with $\|\bar{w}_i(t_i) - \bar{v}\|_2 < \varepsilon/2$, that is, $E_i(t_i) < \varepsilon^2/8$. The second bound in (E.6) and the choice of r_0 then give $E_i(\Lambda r_{\mathrm{init}}) \le E_i(t_i) + B\Lambda r_{\mathrm{init}} < \varepsilon^2/4$, so $\|\bar{w}_i(\Lambda r_{\mathrm{init}}) - \bar{v}\|_2 = \sqrt{2E_i(\Lambda r_{\mathrm{init}})} < \varepsilon$. ∎

After this interval, the norms need a time of order one to grow. As long as the ratio r_{3-i}/r_i stays bounded, (E.5) allows the direction errors to grow only by a bounded factor over such a time.

Lemma E.7. *Let $C \geq 1$, $\eta \in (0, \pi/2)$ and $T_\alpha := r_-/8$. For every $\varepsilon \in (0,1]$ and $T > T_\alpha$, there is $r_0 = r_0(\alpha, C, \eta, \varepsilon, T) > 0$, independent of d, with the following property. If the initialization is C-balanced, $0 < r_{\mathrm{init}} \leq r_0$, and $\angle(\bar{w}_i(0), -\bar{v}) \geq \eta$ for $i = 1, 2$, then $T_{\max} > T$ and*

$$\|\bar{w}_i(t) - \bar{v}\|_2 < \varepsilon, \qquad |r_1(t) - r_2(t)| < \varepsilon, \qquad T_\alpha \leq t \leq T, \quad i = 1, 2.$$

Proof of Lemma E.7. Let

$$K := \exp\left(C + \frac{256\,T}{r_-^2}\right), \qquad \varepsilon_* := \frac{\min\{\varepsilon, r_-/2\}}{1 + 2\pi T}.$$

Let Λ and r_{a} be the constants of Lemma E.6 for $m = 2/(C+1)$, $M = 2$, the margin η and the tolerance $\varepsilon_*/(2\sqrt{K})$, and set

$$r_0 := \min\left\{r_{\mathrm{a}}, \frac{r_-}{32}, \frac{T_\alpha}{\Lambda}, \frac{\min\{\varepsilon, r_-/2\}}{2(2+\Lambda)}\right\}.$$

Let $0 < r_{\mathrm{init}} \leq r_0$. Since $r_1(0) + r_2(0) = 2r_{\mathrm{init}}$ and $\max_i r_i(0) \leq C \min_i r_i(0)$, we have $2r_{\mathrm{init}}/(C+1) \leq r_i(0) \leq 2r_{\mathrm{init}}$. Since $r_{\mathrm{init}} \leq r_-/32$, we get $\max_i r_i(0) + T_\alpha \leq r_-/16 + r_-/8 < r_-/4$, so Lemma E.2 gives $T_{\max} > T_\alpha$ and $r_i(0) + r_- t/4 \leq r_i(t) \leq r_i(0) + t$ for $0 \leq t \leq T_\alpha$. Hence $r_i(T_\alpha) \geq r_-^2/32$. Moreover, $r_{3-i}(t)/r_i(t) \leq (r_{3-i}(0) + t)/(r_i(0) + r_- t/4)$. The right-hand side is a monotone function of t that equals $r_{3-i}(0)/r_i(0) \leq C$ at $t = 0$ and tends to $4/r_-$ as $t \to \infty$. Hence

$$\frac{r_{3-i}(t)}{r_i(t)} \leq \max\left\{C, \frac{4}{r_-}\right\}, \qquad 0 \leq t \leq T_\alpha, \quad i = 1, 2. \tag{E.7}$$

The angular hypothesis is equivalent to $\angle(\bar{w}_i(0), \bar{v}) \leq \pi - \eta$, so Lemma E.6 applies. At $\tau_{\mathrm{a}} := \Lambda r_{\mathrm{init}} \leq T_\alpha$, it gives $\|\bar{w}_i(\tau_{\mathrm{a}}) - \bar{v}\|_2 < \varepsilon_*/(2\sqrt{K})$, and hence $E_1(\tau_{\mathrm{a}}) + E_2(\tau_{\mathrm{a}}) < \varepsilon_*^2/(4K)$, with E_i as in Lemma E.5.

Let $T' := \min\{T, T_{\max}\}$, and let τ be the supremum of the times $t \in [T_\alpha, T')$ such that $r_1, r_2 > r_-^2/64$ on $[T_\alpha, t]$. Since $r_i(T_\alpha) \geq r_-^2/32$, we have $\tau > T_\alpha$. Since $r_i(0) \leq 2r_{\mathrm{init}} \leq 2$, Lemma E.1 gives $r_i \leq 2$ on $[0, T_{\max})$, so $r_{3-i}/r_i \leq 128/r_-^2$ on $[T_\alpha, \tau)$. Since $r_- < 2$ and $C \geq 1$, we have $2T_\alpha \max\{C, 4/r_-\} = \max\{Cr_-/4, 1\} \leq C$, and therefore

$$\int_{\tau_{\mathrm{a}}}^{t} 2 \max_{j=1,2} \frac{r_{3-j}}{r_j} \mathrm{d}u \leq 2T_\alpha \max\left\{C, \frac{4}{r_-}\right\} + \frac{256\,T}{r_-^2} \leq \log K, \qquad \tau_{\mathrm{a}} \leq t < \tau.$$

Hence (E.5), (E.7) and Grönwall's inequality give $E_1(t) + E_2(t) \leq K(E_1(\tau_{\mathrm{a}}) + E_2(\tau_{\mathrm{a}})) < \varepsilon_*^2/4$ for $\tau_{\mathrm{a}} \leq t < \tau$, and therefore

$$\|\bar{w}_i(t) - \bar{v}\|_2 < \varepsilon_*, \qquad \tau_{\mathrm{a}} \leq t < \tau, \quad i = 1, 2. \tag{E.8}$$

We next bound $r_1 - r_2$. By (A.15), (A.16) and (A.1b),

$$\frac{\mathrm{d}}{\mathrm{d}t}(r_1 - r_2) = \sum_{j=1}^{2}\big(l(\theta_{1j}) - l(\theta_{2j})\big) - \Big(\frac{1}{2} - l(\gamma)\Big)(r_1 - r_2).$$

Since $l(\gamma) \leq 1/2$ by Lemma A.2, solving this linear equation gives

$$|r_1(t) - r_2(t)| \leq |r_1(0) - r_2(0)| + \int_0^t \Big|\sum_{j=1}^{2}\big(l(\theta_{1j}) - l(\theta_{2j})\big)\Big| \mathrm{d}u.$$

The integrand is at most 1, because $0 \leq l \leq 1/2$. It is also at most γ, because $|l'| \leq 1/2$ by (A.2) and $|\theta_{1j} - \theta_{2j}| \leq \gamma$ by the triangle inequality for angles. On $[\tau_{\mathrm{a}}, \tau)$, (E.1) and (E.8) give $\gamma \leq \angle(\bar{w}_1, \bar{v}) + \angle(\bar{w}_2, \bar{v}) < \pi\varepsilon_*$. Using the bound 1 on $[0, \tau_{\mathrm{a}}]$, $|r_1(0) - r_2(0)| \leq 2r_{\mathrm{init}}$ and the choice of ε_* and r_0, we obtain

$$|r_1(t) - r_2(t)| \leq (2 + \Lambda) r_{\mathrm{init}} + \pi\varepsilon_* T < \min\left\{\varepsilon, \frac{r_-}{2}\right\}, \qquad 0 \leq t < \tau. \tag{E.9}$$

Suppose that $\tau < T'$. Then $r_i(\tau) = r_-^2/64$ for some i, while $r_i > r_-^2/64$ on $[T_\alpha, \tau)$. By (E.9) and continuity, $r_{3-i}(\tau) \le r_-^2/64 + r_-/2$, and Lemma A.5 gives $\dot r_i(\tau) \ge r_-/2 - r_-^2/64 - r_-/4 > 0$, because $r_- < 2$. This contradicts $r_i > r_i(\tau)$ on $[T_\alpha, \tau)$, so $\tau = T'$. If $T_{\max} \le T$, then $\min\{r_i(0), r_-^2/64\} \le r_i \le 2$ on $[0, T_{\max})$ for $i = 1, 2$, so the trajectory stays in a compact subset of Ω; Lemma A.1 then extends the solution beyond $T_{\max}$, which is impossible. Hence $T_{\max} > T$ and $\tau = T$. By continuity, (E.9) also holds at $t = T$, and $\|\bar w_i(T) - \bar v\|_2 \le \varepsilon_*$. Since $\varepsilon_* < \varepsilon$ and $\tau_{\mathrm{a}} \le T_\alpha$, this proves the lemma. ∎

E.3 Proofs of Lemma 4.1 and Corollary E.8

Proof of Lemma 4.1. If the conclusion holds for some ρ, it holds with the same t_0 and r_* for every larger ρ, so we may assume $\rho \le 1$. Let $T_\alpha = r_-/8$ be as in Lemma E.7, and set

$$t_0 := T_\alpha + \log\frac{16}{\rho}, \qquad \varepsilon := \frac{\rho}{20}, \qquad r_* := \min\{r_0(\alpha, C, \eta, \varepsilon, t_0), 1\},$$

where r_0 is given by Lemma E.7. Then t_0 depends only on α and ρ, and r_* only on α, C, η, ρ. Suppose that $0 < r_{\mathrm{init}} \le r_*$. By Lemma E.7, the solution exists on $[0, t_0]$, both weights are nonzero there, and the bounds of that lemma hold on $[T_\alpha, t_0]$. Moreover, $r_i(0) \le 2r_{\mathrm{init}} \le 2$, so $r_i \le 2$ on $[0, t_0]$ by Lemma E.1.

Adding (A.15) and (A.16), and using $r_+ = 4l(\alpha/2)$ and (A.1b), we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}(r_1 + r_2) + (r_1 + r_2) - r_+ = \sum_{i,j=1}^{2}\Big(l(\theta_{ij}) - l\Big(\frac{\alpha}{2}\Big)\Big) + \Big(\frac12 - l(\gamma)\Big)(r_1 + r_2).$$

We bound the right-hand side on $[T_\alpha, t_0]$. Since $|l'| \le 1/2$ by (A.2) and $|\theta_{ij} - \alpha/2| \le \angle(\bar w_i, \bar v)$ by the triangle inequality for angles, (E.1) gives $|l(\theta_{ij}) - l(\alpha/2)| < \pi\varepsilon/4$. In the same way, $\gamma < \pi\varepsilon$, and hence $0 \le \frac12 - l(\gamma) = l(0) - l(\gamma) < \pi\varepsilon/2$. Since $r_1 + r_2 \le 4$, the right-hand side is less than $3\pi\varepsilon$ in absolute value. Solving the linear equation for $r_1 + r_2 - r_+$ on $[T_\alpha, t_0]$, and using $|r_1(T_\alpha) + r_2(T_\alpha) - r_+| \le 4$, we get $|r_1(t_0) + r_2(t_0) - r_+| \le 4e^{-(t_0 - T_\alpha)} + 3\pi\varepsilon = \rho/4 + 3\pi\varepsilon$. Recall from Section 4 that $r_s = r_+/2$ and $w_s = r_s\bar v$. Hence $r_i - r_s = \frac12(r_1 + r_2 - r_+) \pm \frac12(r_1 - r_2)$, and therefore $|r_i(t_0) - r_s| < \rho/8 + 3\pi\varepsilon/2 + \varepsilon/2$. Since $w_i - w_s = (r_i - r_s)\bar w_i + r_s(\bar w_i - \bar v)$ and $r_s \le 1$, it follows that

$$\|w(t_0) - w_{\mathrm{saddle}}\|_2 \le \sum_{i=1}^{2}\|w_i(t_0) - w_s\|_2 < \frac{\rho}{4} + (3\pi + 3)\varepsilon < \rho.$$

Thus $t_{\mathrm{in}} = t_0$ has the required property. ∎

Corollary E.8 (Phase 1 under Gaussian initialization)**.** *In the setting of Lemma 4.1, fix $C > 1$, $\eta \in (0, \pi/2)$, $\rho > 0$, and let t_0, r_* be the constants in that lemma. Suppose $w_1(0), w_2(0) \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, \sigma^2 I_d)$, where $0 < \sigma \le r_*/(2\sqrt{d})$. There exists $c = c(C, \eta) > 0$ such that, with probability at least $1 - 6e^{-cd}$, the gradient-flow solution exists on $[0, t_0]$, both student weights remain nonzero on this interval, and*

$$\|w(t_{\mathrm{in}}) - w_{\mathrm{saddle}}\|_2 < \rho \qquad \textit{for some } t_{\mathrm{in}} \in [0, t_0].$$

Proof of Corollary E.8. Let $\lambda := \min\{1/2, (C^2 - 1)/(C^2 + 1)\} > 0$. For each i, $\|w_i(0)\|_2^2/\sigma^2$ has the χ^2 distribution with d degrees of freedom, whose moment generating function is $(1 - 2s)^{-d/2}$ for $s < 1/2$. Markov's inequality applied to $\exp(s\|w_i(0)\|_2^2/\sigma^2)$, with $s = \lambda/(2(1 + \lambda))$ and with $s = -\lambda/(2(1 - \lambda))$, gives

$$\mathbb{P}\{\|w_i(0)\|_2^2 \ge (1 + \lambda)\sigma^2 d\} \le \exp\Big(-\frac{d}{2}\big(\lambda - \log(1 + \lambda)\big)\Big) \le e^{-d\lambda^2/8},$$
$$\mathbb{P}\{\|w_i(0)\|_2^2 \le (1 - \lambda)\sigma^2 d\} \le \exp\Big(-\frac{d}{2}\big(-\lambda - \log(1 - \lambda)\big)\Big) \le e^{-d\lambda^2/4}.$$

The last inequalities hold because, for $0 < \lambda \le 1/2$,

$$\lambda - \log(1 + \lambda) = \int_0^\lambda \frac{u}{1 + u}\mathrm{d}u \ge \frac{\lambda^2}{4}, \qquad -\lambda - \log(1 - \lambda) = \int_0^\lambda \frac{u}{1 - u}\mathrm{d}u \ge \frac{\lambda^2}{2}.$$

Let $\mathcal{E}$ be the event that $(1-\lambda)\sigma^2 d < \|w_i(0)\|_2^2 < (1+\lambda)\sigma^2 d$ and $\angle(\bar{w}_i(0), -\bar{v}) \geq \eta$ for $i = 1, 2$. If the norm condition holds but the angular condition fails for some i, then $\langle w_i(0), \bar{v}\rangle < -\|w_i(0)\|_2 \cos\eta < -\sigma\sqrt{(1-\lambda)d}\cos\eta$. Here $\langle w_i(0), \bar{v}\rangle/\sigma$ is standard normal, and a standard normal Z satisfies $\mathbb{P}\{Z \leq -a\} \leq e^{-a^2/2}$ for $a \geq 0$. A union bound therefore gives

$$\mathbb{P}(\mathcal{E}^c) \leq 4e^{-d\lambda^2/8} + 2e^{-(1-\lambda)d\cos^2\eta/2} \leq 6e^{-cd}, \qquad c := \min\left\{\frac{\lambda^2}{8}, \frac{(1-\lambda)\cos^2\eta}{2}\right\}.$$

On $\mathcal{E}$, both initial weights are nonzero, and the choice of λ gives

$$\max_i \|w_i(0)\|_2 \leq \sqrt{\frac{1+\lambda}{1-\lambda}} \min_i \|w_i(0)\|_2 \leq C \min_i \|w_i(0)\|_2, \qquad r_{\text{init}} < \sigma\sqrt{(1+\lambda)d} \leq 2\sigma\sqrt{d} \leq r_*.$$

Hence Lemma 4.1 applies on $\mathcal{E}$ and gives the conclusion. ■

F Exclusion of saddles and global convergence

As in Theorem 3.3, we consider $\alpha = \angle(v_1, v_2) \in (0, \pi/2]$. We recall the set Ω of pairs (w_1, w_2) with $w_1, w_2 \neq 0$ and the notation $r_i = \|w_i\|_2$ from Appendix D. For $w(0) \in \Omega$, we write $[0, T_{\max})$ for the maximal interval on which the gradient-flow solution exists in Ω; it is well defined by Lemma A.1. We use the coordinates $z_{i,j}$ of Section 4.2.2 with respect to the orthonormal basis $\{\bar{v}, v_\Delta, u_3, \ldots, u_d\}$.

Since the teachers are unit vectors at angle α, (2.4) gives $v_1 = \cos(\alpha/2)\,\bar{v} + \sin(\alpha/2)\,v_\Delta$ and $v_2 = \cos(\alpha/2)\,\bar{v} - \sin(\alpha/2)\,v_\Delta$. Hence

$$\cos\theta_{i1} - \cos\theta_{i2} = \langle \bar{w}_i, v_1 - v_2\rangle = \frac{2\sin(\alpha/2)}{r_i}\, z_{i,2}, \qquad i = 1, 2. \tag{F.1}$$

Since $\arccos$ is strictly decreasing, $\theta_{i2} - \theta_{i1}$ has the sign of $z_{i,2}$ and vanishes with it. Taking inner products of (A.11) and (A.12) with $\bar{v}$, v_Δ and u_j, and noting that the terms in $\bar{w}_i$ and w_i there combine to $a_i w_i/(2\pi r_i)$, with a_i as in Lemma 3.2, we obtain, on Ω,

$$\begin{aligned}
\dot{z}_{1,1} &= \frac{1}{2\pi}\left[(2\pi - \theta_{11} - \theta_{12})\cos\frac{\alpha}{2} + \frac{a_1}{r_1} z_{1,1} - (\pi - \gamma) z_{2,1}\right], \\
\dot{z}_{2,1} &= \frac{1}{2\pi}\left[(2\pi - \theta_{21} - \theta_{22})\cos\frac{\alpha}{2} + \frac{a_2}{r_2} z_{2,1} - (\pi - \gamma) z_{1,1}\right], \\
\dot{z}_{1,2} &= \frac{1}{2\pi}\left[(\theta_{12} - \theta_{11})\sin\frac{\alpha}{2} + \frac{a_1}{r_1} z_{1,2} - (\pi - \gamma) z_{2,2}\right], \\
\dot{z}_{2,2} &= \frac{1}{2\pi}\left[(\theta_{22} - \theta_{21})\sin\frac{\alpha}{2} + \frac{a_2}{r_2} z_{2,2} - (\pi - \gamma) z_{1,2}\right], \\
\dot{z}_{1,j} &= \frac{1}{2\pi}\left[\frac{a_1}{r_1} z_{1,j} - (\pi - \gamma) z_{2,j}\right], \qquad 3 \leq j \leq d, \\
\dot{z}_{2,j} &= \frac{1}{2\pi}\left[\frac{a_2}{r_2} z_{2,j} - (\pi - \gamma) z_{1,j}\right], \qquad 3 \leq j \leq d.
\end{aligned} \tag{F.2}$$

By (F.1), the first term in the equation for $z_{i,2}$, which comes from the teachers, vanishes together with $z_{i,2}$. At a zero of one planar splitting coordinate, the sign of its derivative is therefore determined by the planar splitting coordinate of the other student.

Lemma F.1 (Signs of the planar splitting coordinates)**.** *Let $w(0) \in \Omega$ with $z_{1,2}(0) z_{2,2}(0) < 0$. Then $z_{1,2}(t) z_{2,2}(t) < 0$ for $0 \leq t < T_{\max}$.*

Proof of Lemma F.1. Exchanging the two students if necessary, we may assume that $z_{1,2}(0) > 0 > z_{2,2}(0)$. By (F.1), the right-hand sides of the third and fourth lines of (F.2) vanish on the subspace where $z_{1,2} = z_{2,2} = 0$, so the vector field of the flow is tangent to this subspace. The field is locally Lipschitz on Ω by Lemma A.1, and uniqueness of solutions in both time directions shows that a trajectory that meets the subspace lies in it at all times. Since $z_{1,2}(0) \neq 0$, the trajectory never meets it.

Suppose now that $z_{1,2}(t)z_{2,2}(t) = 0$ for some $t < T_{\max}$, and let t_1 be the first such time. By the previous paragraph, only one of the two coordinates vanishes at t_1. We treat the case $z_{1,2}(t_1) = 0 > z_{2,2}(t_1)$; the other is symmetric. Then $\gamma(t_1) < \pi$, since otherwise $w_2(t_1)$ would be a negative multiple of $w_1(t_1)$ and $z_{2,2}(t_1)$ would vanish as well. As $\theta_{11} = \theta_{12}$ at t_1, the third line of (F.2) gives $\dot z_{1,2}(t_1) = -(\pi - \gamma(t_1))z_{2,2}(t_1)/(2\pi) > 0$. This is impossible: $z_{1,2}$ is positive on $[0, t_1)$ and vanishes at t_1, so $\dot z_{1,2}(t_1) \le 0$. ∎

Under the hypothesis of Lemma F.1, $z_{1,2}$ and $z_{2,2}$ keep opposite signs, and $\theta_{i2} - \theta_{i1}$ has the sign of $z_{i,2}$. The third and fourth lines of (F.2) then give, for $0 \le t < T_{\max}$,

$$\frac{\mathrm{d}}{\mathrm{d}t}|z_{i,2}| = \frac{1}{2\pi}\left[|\theta_{i2} - \theta_{i1}| \sin\frac{\alpha}{2} + \frac{a_i}{r_i}|z_{i,2}| + (\pi - \gamma)|z_{3-i,2}|\right], \qquad i = 1, 2. \tag{F.3}$$

The equations for the out-of-plane coordinates in (F.2) involve the same factors a_i/r_i and $\pi - \gamma$ as (F.3), but no term coming from the teachers.

Lemma F.2 (Comparison of out-of-plane and planar coordinates)**.** *Let $w(0)$ satisfy Assumption 4.4. Then $|z_{i,j}(t)| \le C_\perp |z_{i,2}(t)|$ and $|z_{1,j}(t) - z_{2,j}(t)| \le C_\perp |z_{1,2}(t) - z_{2,2}(t)|$ for $0 \le t < T_{\max}$, $i \in \{1, 2\}$ and $3 \le j \le d$.*

Proof of Lemma F.2. By Lemma F.1, (F.3) holds on $[0, T_{\max})$. Fix j with $3 \le j \le d$. For $i = 1, 2$, we set $p_i^{\pm} := C_\perp |z_{i,2}| \mp z_{i,j}$. The first claim says that all four functions $p_i^{\pm}$ are nonnegative, and by Assumption 4.4 this holds at $t = 0$. By (F.3) and the last two lines of (F.2),

$$\dot p_i^{\pm} = \frac{1}{2\pi}\left[C_\perp |\theta_{i2} - \theta_{i1}| \sin\frac{\alpha}{2} + \frac{a_i}{r_i} p_i^{\pm} + (\pi - \gamma) p_{3-i}^{\mp}\right].$$

By Lemma F.1 and (F.1), $\theta_{i2} \ne \theta_{i1}$. Hence, if all four functions are nonnegative and one of them vanishes, its derivative is positive. Let t_2 be the supremum of the times $t < T_{\max}$ such that all four functions are nonnegative on $[0, t]$, and suppose that $t_2 < T_{\max}$. Then one of them is negative at times arbitrarily close to t_2 from above. By continuity, it vanishes at t_2 while the others are nonnegative there, so its derivative at t_2 is positive and it is positive just after t_2, which is a contradiction. This proves the first claim. Since $z_{1,2}$ and $z_{2,2}$ have opposite signs, $|z_{1,2}| + |z_{2,2}| = |z_{1,2} - z_{2,2}|$, and the second claim follows from $|z_{1,j} - z_{2,j}| \le |z_{1,j}| + |z_{2,j}|$. ∎

Proof of Theorem 4.5. The first claim is the second estimate of Lemma F.2. If $d = 2$, the family $\mathcal{S}_\perp$ is empty by Theorem 3.3, so we may assume that $d \ge 3$. As in Proposition 3.5, we write $w_i^\perp = \sum_{j=3}^{d} z_{i,j} u_j$ for the component of w_i orthogonal to the teacher plane. Since $u_3, \dots, u_d$ are orthonormal, Lemma F.2 gives

$$\|w_1^\perp(t) - w_2^\perp(t)\|_2 \le C_\perp \sqrt{d-2}\, |z_{1,2}(t) - z_{2,2}(t)|. \tag{F.4}$$

Let $\widehat w \in \mathcal{S}_\perp$. By (3.6), $\widehat w_1$ and $\widehat w_2$ are orthogonal to v_Δ, and $\|\widehat w_1^\perp - \widehat w_2^\perp\|_2 = 2 r_\alpha(\kappa_\alpha) \sin\kappa_\alpha > 0$. By the Cauchy–Schwarz inequality,

$$\begin{aligned} |z_{1,2}(t) - z_{2,2}(t)| &= \big|\big\langle (w_1(t) - \widehat w_1) - (w_2(t) - \widehat w_2), v_\Delta \big\rangle\big| \\ &\le \|w_1(t) - \widehat w_1\|_2 + \|w_2(t) - \widehat w_2\|_2 \le \sqrt{2}\, \|w(t) - \widehat w\|_2, \end{aligned}$$

and since orthogonal projection does not increase norms, the same bound holds with $w_i(t) - \widehat w_i$ replaced by $w_i^\perp(t) - \widehat w_i^\perp$. The triangle inequality and (F.4) now give

$$2 r_\alpha(\kappa_\alpha) \sin\kappa_\alpha \le \|w_1^\perp(t) - w_2^\perp(t)\|_2 + \sqrt{2}\, \|w(t) - \widehat w\|_2 \le \sqrt{2}\big(1 + C_\perp\sqrt{d-2}\big)\|w(t) - \widehat w\|_2.$$

Hence the theorem holds with $c = \sqrt{2}\, r_\alpha(\kappa_\alpha) \sin\kappa_\alpha / (1 + C_\perp\sqrt{d-2})$. ∎

Every in-plane saddle in (3.5) has zero planar splitting coordinates, while near each of them (F.3) makes at least one of these coordinates grow.

Lemma F.3 (Avoidance of in-plane saddles)**.** *Let $w(0) \in \Omega$ with $z_{1,2}(0)z_{2,2}(0) < 0$ and $T_{\max} = \infty$. Then $w(t)$ does not converge to an in-plane saddle in (3.5).*

Proof of Lemma F.3. Suppose that $w(t)$ converges to such a saddle w_∞. Both components of w_∞ are multiples of $\bar{v}$, so $z_{1,2}(t)$ and $z_{2,2}(t)$ tend to zero. By Lemma F.1, (F.3) holds for all $t \geq 0$.

First suppose that w_∞ has one of the first two forms in (3.5), that is, $w_\infty = (s_1\bar{v}, s_2\bar{v})$ or $w_\infty = (-s_1\bar{v}, -s_2\bar{v})$ with $s_1, s_2 > 0$. At w_∞, we have $\gamma = 0$ and $\sin\theta_{ij} = \sin(\alpha/2)$ for all i, j, so $a_i/r_i + \pi - \gamma = 2\sin(\alpha/2)/s_i$. By continuity, $a_i/r_i + \pi - \gamma \geq \sin(\alpha/2)/s_i$ for all sufficiently large t. Adding the two equations in (F.3) and dropping the nonnegative terms that contain $|\theta_{i2} - \theta_{i1}|$, we obtain, for such t,

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(|z_{1,2}| + |z_{2,2}|\big) \geq \frac{1}{2\pi}\sum_{i=1}^{2}\left(\frac{a_i}{r_i} + \pi - \gamma\right)|z_{i,2}| \geq \frac{\sin(\alpha/2)}{2\pi\max\{s_1, s_2\}}\big(|z_{1,2}| + |z_{2,2}|\big).$$

Thus $|z_{1,2}| + |z_{2,2}|$ grows exponentially for large t, contradicting its convergence to zero.

Otherwise, w_∞ has the third form in (3.5). Let i be the index of the student whose limit is $-r_-\bar{v}$. At w_∞, we have $\gamma = \pi$ and $\sin\theta_{i1} = \sin\theta_{i2} = \sin(\alpha/2)$, so $a_i(w_\infty) = 2\sin(\alpha/2) - \pi r_-$, which is positive because $r_- < (2/\pi)\sin(\alpha/2)$ by (3.4). By continuity, $a_i/r_i \geq a_i(w_\infty)/(2r_-)$ for all sufficiently large t, and (F.3) gives $\frac{\mathrm{d}}{\mathrm{d}t}|z_{i,2}| \geq a_i(w_\infty)|z_{i,2}|/(4\pi r_-)$. Again $|z_{i,2}|$ grows exponentially, contradicting its convergence to zero. ■

The conclusion of the next corollary is contained in Theorem 4.2. What is new is the proof, in which Theorem 4.5 and Lemma F.3 take the place of the stable-manifold argument of Proposition D.14.

Corollary F.4 (Global convergence under Assumption 4.4)**.** *For Lebesgue-almost every initialization satisfying Assumption 4.4, the solution exists for all $t \geq 0$, both student weights remain nonzero, and $w(t)$ converges to (v_1, v_2) or (v_2, v_1).*

Proof of Corollary F.4. As in the proof of Theorem 4.2, outside the null set of Proposition D.11 we have $T_{\max} = \infty$, and by Proposition D.12 the solution converges to some w_∞. Outside the null set of Proposition D.13, $w_\infty \in \Omega$, and then $\nabla L(w_\infty) = 0$. Letting $t \to \infty$ in Theorem 4.5 shows that $w_\infty \notin \mathcal{S}_\perp$, and Lemma F.3 shows that w_∞ is not an in-plane saddle. By Theorem 3.3, w_∞ is (v_1, v_2) or (v_2, v_1). ■

Lemma F.5 (Gaussian initialization)**.** *Let $d \geq 2$, $\sigma > 0$ and $w_1(0), w_2(0) \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I_d)$. Then the coordinates $z_{i,j}(0)$, $i \in \{1,2\}$, $1 \leq j \leq d$, are independent $\mathcal{N}(0,\sigma^2)$ random variables. With probability $1/2$, Assumption 4.4 holds for some $C_\perp > 0$ depending on the initialization.*

Proof of Lemma F.5. The coordinates $z_{i,j}(0)$ are the inner products of $w_1(0)$ and $w_2(0)$ with the vectors of the orthonormal basis $\{\bar{v}, v_\Delta, u_3, \dots, u_d\}$. Since $w_1(0)$ and $w_2(0)$ are independent and $\mathcal{N}(0, \sigma^2 I_d)$ is invariant under orthogonal transformations, this gives the first claim. In particular, $z_{1,2}(0)$ and $z_{2,2}(0)$ are independent, and each is positive with probability $1/2$ and negative with probability $1/2$, so $z_{1,2}(0)z_{2,2}(0) < 0$ with probability $1/2$. On this event, Assumption 4.4 holds with $C_\perp$ the largest of 1 and the ratios $|z_{i,j}(0)|/|z_{i,2}(0)|$, $i \in \{1,2\}$, $3 \leq j \leq d$. Off this event, it fails for every $C_\perp$. ■

G Approach to the bisector saddle under structured initialization

We prove Theorem 4.6 in this appendix and the next. Here we follow the trajectory from the initialization until it comes close to the bisector saddle; Appendix H.2 studies the escape from the saddle and completes the proof.

G.1 Coordinates and estimates near the bisector

As in Theorem 4.6, we consider $\alpha \in (0, \pi/2]$. We use the set Ω, the norms $r_i = \|w_i\|_2$, the maximal time $T_{\max}$ and the coordinates $z_{i,j}$ as in Appendix F, and, as in Proposition 3.5, we write $w_i^\perp = \sum_{j=3}^{d} z_{i,j}u_j$ for the component of w_i orthogonal to the teacher plane. We describe the pair (w_1, w_2) by

$$\begin{aligned} R &\coloneqq \frac{z_{1,1} + z_{2,1}}{2}, & \delta &\coloneqq \frac{z_{1,1} - z_{2,1}}{2}, & a &\coloneqq z_{1,2} - z_{2,2}, \\ Y &\coloneqq \big(\|w_1^\perp\|_2^2 + \|w_2^\perp\|_2^2\big)^{1/2}, & N_\perp &\coloneqq \big(z_{1,2}^2 + z_{2,2}^2 + Y^2\big)^{1/2}. \end{aligned} \tag{G.1}$$

Here R and δ are the mean and half the difference of the two bisector coordinates, $|a| = |\langle w_1 - w_2, v_\Delta\rangle|$ is the planar splitting amplitude, and Y is the norm of the out-of-plane part of (w_1, w_2). Since $N_\perp^2 = \sum_{i=1}^2 \sum_{j=2}^d z_{i,j}^2$, the time T_ε defined before Theorem 4.6 is the first time at which $N_\perp$ reaches ε.

If $z_{1,2}(0) > 0 > z_{2,2}(0)$, then Lemma F.1 gives

$$z_{1,2}(t) > 0 > z_{2,2}(t), \qquad a(t) = |z_{1,2}(t)| + |z_{2,2}(t)| > 0, \qquad 0 \le t < T_{\max}. \tag{G.2}$$

In this case $a(0) = a_{\text{init}}$ by (2.5), and, since $a = |z_{1,2}| + |z_{2,2}|$, we have $|z_{i,2}| \le a$, $\|w_i^\perp\|_2 \le Y$ and $N_\perp \le a + Y$.

We recall the rate $\mu_\alpha = 2\sin(\alpha/2)/(\pi r_s)$ defined before Theorem 4.6. By (3.4) and $r_s = r_+/2$, we have $\pi r_s = (\pi - \alpha/2)\cos(\alpha/2) + \sin(\alpha/2)$. Hence $\mu_\alpha < 1$ is equivalent to $\tan(\alpha/2) < \pi - \alpha/2$, which holds because $\tan(\alpha/2) \le 1 < 3\pi/4 \le \pi - \alpha/2$ for $\alpha \le \pi/2$. Thus

$$0 < \mu_\alpha < 1. \tag{G.3}$$

Three identities follow from (F.2). First, by the definition of a_i in Lemma 3.2,

$$\frac{a_i}{r_i}\, z_{i,1} = (\sin\theta_{i1} + \sin\theta_{i2})\frac{z_{i,1}}{r_i} - \pi z_{i,1} - \sin\gamma\, \frac{r_{3-i}}{r_i}\, z_{i,1}.$$

We substitute this into the first two lines of (F.2) and use $(2\pi - \alpha)\cos(\alpha/2) + 2\sin(\alpha/2) = 2\pi r_s$, which is (3.4). Since $z_{1,1} + z_{2,1} = 2R$, we obtain, for $i = 1, 2$,

$$\begin{aligned} 2\pi\big(\dot z_{i,1} - (r_s - R)\big) = (\alpha - \theta_{i1} - \theta_{i2})\cos\frac{\alpha}{2} + \Big((\sin\theta_{i1} + \sin\theta_{i2})\frac{z_{i,1}}{r_i} - 2\sin\frac{\alpha}{2}\Big) \\ + \gamma z_{3-i,1} - \sin\gamma\, \frac{r_{3-i}}{r_i}\, z_{i,1}. \end{aligned} \tag{G.4}$$

Second, at each time at which $z_{1,2} > 0 > z_{2,2}$, the two equations (F.3) hold, and their sum gives

$$2\pi\dot a = \sum_{i=1}^2 \Big[|\theta_{i2} - \theta_{i1}|\sin\frac{\alpha}{2} + \Big(\frac{a_i}{r_i} + \pi - \gamma\Big)|z_{i,2}|\Big]. \tag{G.5}$$

Third, the last two lines of (F.2) read $2\pi\dot w_1^\perp = (a_1/r_1)w_1^\perp - (\pi - \gamma)w_2^\perp$ and $2\pi\dot w_2^\perp = (a_2/r_2)w_2^\perp - (\pi - \gamma)w_1^\perp$. They give

$$\begin{aligned} \pi\frac{\mathrm{d}}{\mathrm{d}t}Y^2 &= \frac{a_1}{r_1}\|w_1^\perp\|_2^2 + \frac{a_2}{r_2}\|w_2^\perp\|_2^2 - 2(\pi - \gamma)\langle w_1^\perp, w_2^\perp\rangle \\ &\le \Big(\max_{i=1,2}\frac{a_i}{r_i} + \pi - \gamma\Big)Y^2; \end{aligned} \tag{G.6}$$

the inequality in the second line follows from $0 \le \pi - \gamma \le \pi$ and $-2\langle w_1^\perp, w_2^\perp\rangle \le Y^2$. The equations for $w_1^\perp$ and $w_2^\perp$ are linear with scalar coefficients, so $w_1^\perp(t)$ and $w_2^\perp(t)$ remain in $\operatorname{span}\{w_1^\perp(0), w_2^\perp(0)\}$. In particular, $Y \equiv 0$ if $Y(0) = 0$. The estimates below involve the out-of-plane components only through Y and the norms $\|w_i^\perp\|_2$, which is why their constants do not depend on d.

The bisector makes the angle $\alpha/2$ with both teachers, so a student with $z_{i,1} > 0$ that is close to the bisector axis $\mathbb{R}\bar v$ has both teacher angles close to $\alpha/2$. We write $\rho_i := (z_{i,2}^2 + \|w_i^\perp\|_2^2)^{1/2}$ for the distance of w_i from $\mathbb{R}\bar v$. The next lemma quantifies this in terms of the ratio $\rho_i/z_{i,1}$.

Lemma G.1 (Angles near the bisector)**.** *Fix $\alpha \in (0, \pi/2]$ and set $C_{\text{angle}} := 6/\sin^5(3\alpha/8)$. Let $d \ge 2$ and $i \in \{1, 2\}$, and suppose that $z_{i,1} > 0$ and $\rho_i \le (\alpha/8)\, z_{i,1}$, with ρ_i as above. Then $\angle(\bar w_i, \bar v) \le \rho_i/z_{i,1}$, $\theta_{i1}, \theta_{i2} \in [3\alpha/8, 5\alpha/8]$, $0 \le r_i - z_{i,1} \le \rho_i^2/(2z_{i,1})$, and*

$$\Big|\theta_{i2} - \theta_{i1} - \frac{2z_{i,2}}{z_{i,1}}\Big| \le C_{\text{angle}}\, \frac{|z_{i,2}|\,\rho_i^2}{z_{i,1}^3}, \tag{G.7}$$

$$|\theta_{i1} + \theta_{i2} - \alpha| + \Big|(\sin\theta_{i1} + \sin\theta_{i2})\frac{z_{i,1}}{r_i} - 2\sin\frac{\alpha}{2}\Big| \le C_{\text{angle}}\, \frac{\rho_i^2}{z_{i,1}^2}. \tag{G.8}$$

Proof of Lemma G.1. Since $z_{i,1} > 0$, the angle $\angle(\bar{w}_i, \bar{v})$ is acute. Its tangent is $\rho_i / z_{i,1}$, which is at most $\alpha/8$. As $\arctan y \le y$ for $y \ge 0$, this gives the first claim. Since $\angle(\bar{v}, v_1) = \angle(\bar{v}, v_2) = \alpha/2$, the triangle inequality for angles gives $|\theta_{ij} - \alpha/2| \le \angle(\bar{w}_i, \bar{v}) \le \alpha/8$, which is the second claim. Since $r_i \ge z_{i,1}$, the identity $r_i - z_{i,1} = \rho_i^2/(r_i + z_{i,1})$ gives the third claim and

$$0 \le 1 - \frac{z_{i,1}}{r_i} \le \frac{\rho_i^2}{2 z_{i,1}^2}.$$

As recalled at the start of Appendix F, $v_1 = \cos(\alpha/2)\,\bar{v} + \sin(\alpha/2)\,v_\Delta$ and $v_2 = \cos(\alpha/2)\,\bar{v} - \sin(\alpha/2)\,v_\Delta$. Hence $\cos\theta_{i1} = m_i + \Delta_i$ and $\cos\theta_{i2} = m_i - \Delta_i$, where

$$m_i := \cos\frac{\alpha}{2}\,\frac{z_{i,1}}{r_i}, \qquad \Delta_i := \sin\frac{\alpha}{2}\,\frac{z_{i,2}}{r_i}.$$

By the second claim, the numbers $m_i \pm \Delta_i$, their midpoint m_i, and $\cos(\alpha/2)$ all lie in the interval $[\cos(5\alpha/8), \cos(3\alpha/8)] \subset (0,1)$. For y in this interval, $1 - y^2 \ge \sin^2(3\alpha/8)$, and therefore

$$|\arccos'(y)| = \frac{1}{(1-y^2)^{1/2}} \le \frac{1}{\sin(3\alpha/8)}, \qquad |\arccos''(y)| = \frac{y}{(1-y^2)^{3/2}} \le \frac{1}{\sin^3(3\alpha/8)},$$

and $|\arccos'''(y)| = (1+2y^2)(1-y^2)^{-5/2} \le 3/\sin^5(3\alpha/8)$. We also note that $|\Delta_i| \le |z_{i,2}|/z_{i,1}$, since $r_i \ge z_{i,1}$.

For the difference, $\theta_{i2} - \theta_{i1} = \arccos(m_i - \Delta_i) - \arccos(m_i + \Delta_i) = \int_{m_i-\Delta_i}^{m_i+\Delta_i} (-\arccos'(y))\mathrm{d}y$. The error of the midpoint rule on an interval of length $2|\Delta_i|$ is at most $(2|\Delta_i|)^3/24$ times the maximum of the second derivative of the integrand, so

$$\bigl|\theta_{i2} - \theta_{i1} + 2\Delta_i \arccos'(m_i)\bigr| \le \frac{|\Delta_i|^3}{\sin^5(3\alpha/8)} \le \frac{|z_{i,2}|^3}{z_{i,1}^3 \sin^5(3\alpha/8)}.$$

Since $r_i^2(1 - m_i^2) = z_{i,1}^2 \sin^2(\alpha/2) + \rho_i^2$, we have

$$-2\Delta_i \arccos'(m_i) = \frac{2 z_{i,2}}{z_{i,1}} \Bigl(1 + \frac{\rho_i^2}{z_{i,1}^2 \sin^2(\alpha/2)}\Bigr)^{-1/2}.$$

As $0 \le 1 - (1+y)^{-1/2} \le y/2$ for $y \ge 0$ and $\sin(\alpha/2) \ge \sin(3\alpha/8)$, it follows that

$$\Bigl|\theta_{i2} - \theta_{i1} - \frac{2 z_{i,2}}{z_{i,1}}\Bigr| \le \Bigl(\frac{1}{\sin^2(3\alpha/8)} + \frac{1}{\sin^5(3\alpha/8)}\Bigr)\frac{|z_{i,2}|\,\rho_i^2}{z_{i,1}^3},$$

which implies (G.7).

For the sum, Taylor's formula with the Lagrange remainder gives

$$\bigl|\arccos(m_i + \Delta_i) + \arccos(m_i - \Delta_i) - 2\arccos m_i\bigr| \le \frac{\Delta_i^2}{\sin^3(3\alpha/8)} \le \frac{z_{i,2}^2}{z_{i,1}^2 \sin^3(3\alpha/8)}.$$

Moreover, $0 \le \cos(\alpha/2) - m_i \le 1 - z_{i,1}/r_i$, so the mean value theorem gives

$$|2\arccos m_i - \alpha| = 2\Bigl|\arccos m_i - \arccos\cos\frac{\alpha}{2}\Bigr| \le \frac{2(1 - z_{i,1}/r_i)}{\sin(3\alpha/8)} \le \frac{\rho_i^2}{z_{i,1}^2 \sin(3\alpha/8)}.$$

Together, the last two bounds give $|\theta_{i1} + \theta_{i2} - \alpha| \le 2\rho_i^2/(z_{i,1}^2 \sin^3(3\alpha/8))$.

Finally, $\sin\theta_{i1} + \sin\theta_{i2} = 2\sin\bigl((\theta_{i1}+\theta_{i2})/2\bigr)\cos\bigl((\theta_{i2}-\theta_{i1})/2\bigr)$. Since the sine is 1-Lipschitz and $0 \le 1 - \cos y \le y^2/2$,

$$|\sin\theta_{i1} + \sin\theta_{i2} - 2\sin(\alpha/2)| \le |\theta_{i1} + \theta_{i2} - \alpha| + \frac{(\theta_{i2} - \theta_{i1})^2}{4}.$$

By the mean value theorem, $|\theta_{i2} - \theta_{i1}| \le 2|\Delta_i|/\sin(3\alpha/8) \le 2|z_{i,2}|/(z_{i,1}\sin(3\alpha/8))$. The right side above is therefore at most $3\rho_i^2/(z_{i,1}^2 \sin^3(3\alpha/8))$. We multiply by $z_{i,1}/r_i \le 1$ and add $2\sin(\alpha/2)(1 - z_{i,1}/r_i) \le \rho_i^2/z_{i,1}^2$. This bounds the second term on the left of (G.8) by $4\rho_i^2/(z_{i,1}^2 \sin^3(3\alpha/8))$, and (G.8) follows. ∎

On the bisector axis, where $N_\perp = 0$ and $z_{1,1}, z_{2,1} > 0$, both teacher angles equal $\alpha/2$, $\gamma = 0$ and $r_i = z_{i,1}$, so (G.4) gives $\dot z_{1,1} = \dot z_{2,1} = r_s - R$: so R relaxes to r_s and δ stays constant. The next lemma shows that near the axis this remains true up to quadratic errors, while a grows at the rate $\mu_\alpha r_s/R$ and Y at most at the rate $\mu_\alpha r_s/(2R)$. At $R = r_s$ these are the two unstable rates of Lemma 4.3.

Lemma G.2 (Equations of motion near the bisector)**.** *Fix $\alpha \in (0, \pi/2]$ and set $\nu_0 := \min\{1/8, 7\alpha/64\}$. There is $C_{\mathrm{bis}} \geq 1$, depending only on α, such that the following holds for every $d \geq 2$. Let $w(0) \in \Omega$, and let $t \in [0, T_{\max})$ be a time at which $z_{1,2} > 0 > z_{2,2}$, $0 < R \leq 2r_s$ and $a + Y + |\delta| \leq \nu_0 R$. Then $7R/8 \leq z_{i,1} \leq r_i \leq 9R/8$ for $i = 1, 2$, and*

$$|\dot R - (r_s - R)| + |\dot\delta| \leq C_{\mathrm{bis}} \frac{a^2 + Y^2}{R^2},$$

$$\Big|\dot a - \frac{2\sin(\alpha/2)}{\pi R}\, a\Big| \leq C_{\mathrm{bis}}\Big(\frac{a+Y}{R} + \frac{|\delta|}{R^2} + \frac{a^2+Y^2}{R^3}\Big) a,$$

$$\frac{\mathrm{d}}{\mathrm{d}t} Y^2 \leq \Big[\frac{2\sin(\alpha/2)}{\pi R} + C_{\mathrm{bis}}\Big(\frac{a+Y}{R} + \frac{|\delta|}{R^2} + \frac{a^2+Y^2}{R^3}\Big)\Big] Y^2.$$

Proof of Lemma G.2. In this proof, C_α denotes a constant depending only on α, whose value may change from line to line. Since $|\delta| + a + Y \leq \nu_0 R \leq R/8$, $|z_{i,2}| \leq a$ and $\|w_i^\perp\|_2 \leq Y$,

$$\frac{7}{8} R \leq R - |\delta| \leq z_{i,1} \leq r_i \leq z_{i,1} + |z_{i,2}| + \|w_i^\perp\|_2 \leq R + |\delta| + a + Y \leq \frac{9}{8} R,$$

which is the first claim. Moreover, $\rho_i \leq a + Y \leq \nu_0 R \leq (8\nu_0/7) z_{i,1} \leq (\alpha/8) z_{i,1}$, so Lemma G.1 applies to both students, with $\rho_i^2 \leq a^2 + Y^2$. By that lemma and the triangle inequality for angles,

$$\gamma \leq \angle(\bar w_1, \bar v) + \angle(\bar w_2, \bar v) \leq \frac{2(a+Y)}{7R/8} = \frac{16(a+Y)}{7R} \leq \frac{16\nu_0}{7}.$$

We first prove the bound for $\dot R$ and $\dot\delta$. Let $i \in \{1, 2\}$ and $j = 3 - i$. By Lemma G.1, the first two terms on the right of (G.4) are at most $C_\alpha(a^2 + Y^2)/R^2$ in absolute value. We write the remaining term as

$$\gamma z_{j,1} - \sin\gamma\, \frac{r_j}{r_i}\, z_{i,1} = \gamma\Big(z_{j,1} - \frac{r_j}{r_i}\, z_{i,1}\Big) + (\gamma - \sin\gamma)\frac{r_j}{r_i}\, z_{i,1}.$$

Since $z_{j,1} r_i - r_j z_{i,1} = z_{j,1}(r_i - z_{i,1}) - z_{i,1}(r_j - z_{j,1})$, Lemma G.1 and the first claim give $|z_{j,1} - r_j z_{i,1}/r_i| \leq C_\alpha(a^2 + Y^2)/R$. Also $0 \leq \gamma - \sin\gamma \leq \gamma^3/6$, $r_j z_{i,1}/r_i \leq r_j \leq 9R/8$ and $(a+Y)^2 \leq 2(a^2 + Y^2)$. With the bound for γ, we obtain

$$\Big|\gamma z_{j,1} - \sin\gamma\, \frac{r_j}{r_i}\, z_{i,1}\Big| \leq C_\alpha \frac{(a+Y)(a^2+Y^2)}{R^2} \leq C_\alpha \frac{a^2+Y^2}{R^2},$$

where the last step uses $a + Y \leq \nu_0 R \leq 2\nu_0 r_s$. Hence $|\dot z_{i,1} - (r_s - R)| \leq C_\alpha(a^2 + Y^2)/R^2$ for $i = 1, 2$, and the first bound follows from $\dot R = (\dot z_{1,1} + \dot z_{2,1})/2$ and $\dot\delta = (\dot z_{1,1} - \dot z_{2,1})/2$.

For a, we estimate the two terms in (G.5). By (G.7),

$$\Big||\theta_{i2} - \theta_{i1}| - \frac{2|z_{i,2}|}{z_{i,1}}\Big| \leq C_\alpha \frac{|z_{i,2}|\,(a^2+Y^2)}{R^3}.$$

By the definition of a_i in Lemma 3.2,

$$\frac{a_i}{r_i} + \pi - \gamma = \frac{\sin\theta_{i1} + \sin\theta_{i2}}{r_i} - \sin\gamma\, \frac{r_j}{r_i} - \gamma.$$

Dividing (G.8) by $z_{i,1}$ shows that the first term on the right differs from $2\sin(\alpha/2)/z_{i,1}$ by at most $C_\alpha(a^2+Y^2)/R^3$, and the other two terms are at most $\gamma(r_j/r_i + 1) \leq (16(a+Y)/(7R))(9/7+1) \leq 6(a+Y)/R$ in absolute value. Finally, $z_{i,1} = R \pm \delta$ gives $|1/z_{i,1} - 1/R| \leq 8|\delta|/(7R^2)$. Therefore

$$\Big|\frac{a_i}{r_i} + \pi - \gamma - \frac{2\sin(\alpha/2)}{R}\Big| \leq C_\alpha\Big(\frac{a+Y}{R} + \frac{|\delta|}{R^2} + \frac{a^2+Y^2}{R^3}\Big),$$

$$\Big||\theta_{i2} - \theta_{i1}| - \frac{2|z_{i,2}|}{R}\Big| \leq C_\alpha\Big(\frac{|\delta|}{R^2} + \frac{a^2+Y^2}{R^3}\Big)|z_{i,2}|.$$

Substituting these bounds into (G.5) and using $|z_{1,2}| + |z_{2,2}| = a$ proves the second bound.

The bound for $a_i/r_i + \pi - \gamma$ holds for both i. Together with (G.6), it proves the third bound. ∎

G.2 Alignment with the bisector

At a small initialization, the ratios a/R, Y/R and $|\delta|/R$ need not be small, so Lemma G.2 does not yet apply. By Lemma E.6, the directions $\bar{w}_i$ first align with $\bar{v}$ in a time of order r_{init}. Afterwards R grows at a rate close to r_s, while a, Y and δ change little. The next lemma follows the trajectory until R is a large multiple of r_{init}, when all three ratios are small.

Lemma G.3 (Alignment with the bisector)**.** *Fix $\alpha \in (0, \pi/2]$, $C \geq 1$ and $\nu > 0$. There are $r_0 \in (0,1]$, $B > 1$ and $C_{\mathrm{align}} \geq 1$, depending only on α, C and ν, with $Br_0 < r_s/4$, such that the following holds for every $d \geq 2$. Suppose that the initialization is C-balanced, that $0 < r_{\mathrm{init}} \leq r_0$, and that $z_{1,1}(0), z_{2,1}(0) > 0$ and $z_{1,2}(0) > 0 > z_{2,2}(0)$. Then there is a first time $\tau_0 \in [0, T_{\max})$ at which $R(\tau_0) = Br_{\mathrm{init}}$; it satisfies $\tau_0 \leq C_{\mathrm{align}} r_{\mathrm{init}}$,*

$$C_{\mathrm{align}}^{-1} a_{\mathrm{init}} \leq a(t) \leq C_{\mathrm{align}} a_{\mathrm{init}}, \qquad Y(t) + |\delta(t)| \leq C_{\mathrm{align}} r_{\mathrm{init}}, \qquad 0 \leq t \leq \tau_0,$$

and $a(\tau_0) + Y(\tau_0) + |\delta(\tau_0)| \leq \nu R(\tau_0)$.

Proof of Lemma G.3. We first collect bounds that hold on a time interval of length proportional to r_{init}. Since $r_1(0) + r_2(0) = 2r_{\mathrm{init}}$ and the initialization is C-balanced, $2r_{\mathrm{init}} \leq (1+C)\min_i r_i(0)$, so $r_{\mathrm{init}}/C \leq r_i(0) \leq 2r_{\mathrm{init}}$. Also $a_{\mathrm{init}} \leq r_1(0) + r_2(0) = 2r_{\mathrm{init}}$, $Y(0) \leq r_1(0) + r_2(0) = 2r_{\mathrm{init}}$ and $|\delta(0)| \leq r_{\mathrm{init}}$. Let $\lambda > 0$, and suppose that $r_{\mathrm{init}} \leq r_-/(4(2+\lambda))$, with r_- as in (3.4). Since $r_- = 4l(\pi - \alpha/2) < 2$ by (A.1b) and Lemma A.2, this gives $r_{\mathrm{init}} < 1/4$. Then $\max_i r_i(0) + \lambda r_{\mathrm{init}} \leq r_-/4$, and Lemma E.2 gives $T_{\max} > \lambda r_{\mathrm{init}}$ and $r_{\mathrm{init}}/C \leq r_i(t) \leq (2+\lambda) r_{\mathrm{init}}$ for $0 \leq t \leq \lambda r_{\mathrm{init}}$. On this interval, $|a_i| \leq 2 + \pi r_i + r_{3-i}$ and $r_{\mathrm{init}} < 1$ give $|a_i|/r_i \leq (2C + \pi + C(2+\lambda))/r_{\mathrm{init}}$. Let J be the reflection that fixes $\bar{v}, u_3, \dots, u_d$ and maps v_Δ to $-v_\Delta$, so that $Jv_2 = v_1$. The triangle inequality for angles and (E.1) give

$$|\theta_{i2} - \theta_{i1}| = |\angle(J\bar{w}_i, v_1) - \angle(\bar{w}_i, v_1)| \leq \angle(J\bar{w}_i, \bar{w}_i) \leq \frac{\pi}{2} \|J\bar{w}_i - \bar{w}_i\|_2 = \frac{\pi |z_{i,2}|}{r_i}.$$

By (G.5) and (G.6), these bounds give $|\dot{a}| \leq 3C(1+\lambda) a/r_{\mathrm{init}}$ and $\frac{\mathrm{d}}{\mathrm{d}t} Y^2 \leq 4C(1+\lambda) Y^2/r_{\mathrm{init}}$. Since $r_i \leq (2+\lambda) r_{\mathrm{init}} \leq r_-/4 < 1/2$, we have $|a_i| \leq 2 + (\pi+1)/2$, and the first two lines of (F.2) give $2\pi |\dot{z}_{i,1}| \leq 2\pi + |a_i| + \pi r_{3-i} < 4\pi$. Hence $|\dot{\delta}| \leq 2$. Integrating, we find that, for $0 \leq t \leq \lambda r_{\mathrm{init}}$,

$$e^{-3C(1+\lambda)\lambda} a_{\mathrm{init}} \leq a(t) \leq e^{3C(1+\lambda)\lambda} a_{\mathrm{init}}, \quad Y(t) \leq 2e^{2C(1+\lambda)\lambda} r_{\mathrm{init}}, \quad |\delta(t)| \leq r_{\mathrm{init}} + 2t. \tag{G.9}$$

Next we consider times at which both directions are close to $\bar{v}$. Let

$$\varepsilon_{\mathrm{a}} := \min\Big\{\nu_0, \frac{r_s}{4C_{\mathrm{angle}} + 4r_s}, \frac{\nu r_s}{4C_{\mathrm{angle}} + 12 r_s}\Big\},$$

with ν_0 as in Lemma G.2, and suppose that at some time $\|\bar{w}_i - \bar{v}\|_2 \leq \varepsilon_{\mathrm{a}}$ for $i = 1, 2$ and $R \leq r_s/4$. Then $z_{i,1}/r_i = \langle \bar{w}_i, \bar{v}\rangle = 1 - \|\bar{w}_i - \bar{v}\|_2^2/2 \geq 7/8$ and $\rho_i/r_i = \|\bar{w}_i - \langle \bar{w}_i, \bar{v}\rangle \bar{v}\|_2 \leq \varepsilon_{\mathrm{a}}$, so $\rho_i \leq (8\varepsilon_{\mathrm{a}}/7) z_{i,1} \leq (\alpha/8) z_{i,1}$. By Lemma G.1, the first two terms on the right of (G.4) are at most $2C_{\mathrm{angle}} \varepsilon_{\mathrm{a}}^2$ in absolute value, and the triangle inequality for angles gives $\gamma \leq 16\varepsilon_{\mathrm{a}}/7$. Since $z_{j,1} \leq 2R$ and $r_j \leq (8/7) z_{j,1} \leq 16R/7$ for $j = 1, 2$, the last two terms are at most $\gamma(z_{3-i,1} + r_{3-i}) \leq 11\varepsilon_{\mathrm{a}} R$. We use $R \leq r_s/4$ and the choice of ε_{a} for the first bound below, and we sum $|z_{i,2}| + \|w_i^\perp\|_2 \leq \sqrt{2}\,\varepsilon_{\mathrm{a}} r_i$ over i for the second:

$$|\dot{z}_{i,1} - (r_s - R)| \leq (C_{\mathrm{angle}} + r_s)\varepsilon_{\mathrm{a}} \leq \frac{r_s}{4}, \qquad a + Y \leq \sqrt{2}\,\varepsilon_{\mathrm{a}}(r_1 + r_2) \leq 4\varepsilon_{\mathrm{a}} R. \tag{G.10}$$

Consequently $\dot{R} \geq r_s/2$ and $|\dot{\delta}| \leq (C_{\mathrm{angle}} + r_s)\varepsilon_{\mathrm{a}}$ at such times.

Let Λ and r_{a} be the constants of Lemma E.6 for $\eta = \pi/3$, $m = 1/C$, $M = 2$ and $\varepsilon = \varepsilon_{\mathrm{a}}/\sqrt{2}$. We set

$$B := 3 + \Lambda + \frac{2(1+2\Lambda)}{\nu}, \qquad \lambda_1 := \Lambda + \frac{4B}{r_s}, \qquad r_0 := \min\Big\{r_{\mathrm{a}}, \frac{r_-}{4(2+\lambda_1)}, \frac{\varepsilon_{\mathrm{a}}^2}{4C(2+\lambda_1)\lambda_1}\Big\},$$

and $C_{\mathrm{align}} := (3 + 2\lambda_1) e^{3C(1+\lambda_1)\lambda_1}$. These depend only on α, C and ν, and $C_{\mathrm{align}} \geq \lambda_1$.

We note that $Br_0 < Br_-/(4\lambda_1) \leq r_- r_s/16 < r_s/4$, because $\lambda_1 \geq 4B/r_s$ and $r_- < 2$. Suppose now that $r_{\mathrm{init}} \leq r_0$. With $T_1 := \lambda_1 r_{\mathrm{init}}$, (G.9) holds with $\lambda = \lambda_1$ on $[0, T_1]$. Since $\angle(\bar{w}_i(0), \bar{v}) <$

$\pi/2$, Lemma E.6 applies, and at $t_{\mathrm{a}} := \Lambda r_{\mathrm{init}}$ we have $E_i(t_{\mathrm{a}}) < \varepsilon_{\mathrm{a}}^2/4$, where $E_i = \frac{1}{2}\|\bar{w}_i - \bar{v}\|_2^2$ as in Lemma E.5. By (E.4), $\dot{E}_i \le r_{3-i}/(2r_i) \le C(2+\lambda_1)/2$ on $[0, T_1]$, so the choice of r_0 gives $E_i < \varepsilon_{\mathrm{a}}^2/4 + \varepsilon_{\mathrm{a}}^2/8 < \varepsilon_{\mathrm{a}}^2/2$ on $[t_{\mathrm{a}}, T_1]$. Thus $\|\bar{w}_i - \bar{v}\|_2 < \varepsilon_{\mathrm{a}}$ on $[t_{\mathrm{a}}, T_1]$.

For $0 \le t \le t_{\mathrm{a}}$, the bound $r_i(t) \le r_i(0) + t$ of Lemma E.2 gives $R \le (r_1 + r_2)/2 \le (2+\Lambda) r_{\mathrm{init}} < B r_{\mathrm{init}}$, so R does not reach $B r_{\mathrm{init}}$ before t_{a}. As long as $t_{\mathrm{a}} \le t \le T_1$ and $R \le B r_{\mathrm{init}}$, we have $\|\bar{w}_i - \bar{v}\|_2 < \varepsilon_{\mathrm{a}}$ and $R \le r_s/4$, so $\dot{R} \ge r_s/2$. If R stayed below $B r_{\mathrm{init}}$ on $[0, T_1]$, we would obtain $R(T_1) \ge R(t_{\mathrm{a}}) + (r_s/2)(T_1 - t_{\mathrm{a}}) = R(t_{\mathrm{a}}) + 2B r_{\mathrm{init}}$, where $R(t_{\mathrm{a}}) > 0$ because $z_{i,1} \ge (7/8) r_i$ there. This is impossible, so τ_0 exists, $t_{\mathrm{a}} < \tau_0 \le T_1 \le C_{\mathrm{align}} r_{\mathrm{init}}$, and $\tau_0 - t_{\mathrm{a}} \le 2(B r_{\mathrm{init}} - R(t_{\mathrm{a}}))/r_s \le 2B r_{\mathrm{init}}/r_s$. The bounds for a, Y and δ on $[0, \tau_0]$ follow from (G.9) with $\lambda = \lambda_1$.

At τ_0, (G.10) gives $a + Y \le 4\varepsilon_{\mathrm{a}} R(\tau_0)$. Integrating the bound for $\dot{\delta}$ from t_{a}, and using $|\delta(t_{\mathrm{a}})| \le (1 + 2\Lambda) r_{\mathrm{init}}$, we obtain

$$\frac{a(\tau_0) + Y(\tau_0) + |\delta(\tau_0)|}{R(\tau_0)} \le 4\varepsilon_{\mathrm{a}} + \frac{1 + 2\Lambda}{B} + \frac{2(C_{\mathrm{angle}} + r_s)\varepsilon_{\mathrm{a}}}{r_s} \le \frac{\nu}{2} + \frac{\nu}{2},$$

by the choices of ε_{a} and B. ∎

G.3 Growth of the bisector coordinates up to $r_s/2$

Once the ratios are small, Lemma G.2 applies. Then R grows at a rate close to $r_s - R$, δ changes little, and, since $\mu_\alpha r_s \int \mathrm{d}t/R \approx \mu_\alpha \log R$, a grows like R^{μ_α} while Y grows at most like $R^{\mu_\alpha/2}$. Since $\mu_\alpha < 1$ by (G.3), the ratios a/R and Y/R decrease, and the trajectory stays near the bisector until R reaches $r_s/2$.

Lemma G.4 (Growth of the bisector coordinates)**.** *Fix $\alpha \in (0, \pi/2]$, and let ν_0 be as in Lemma G.2. There are $\nu \in (0, \nu_0]$ and $C_{\mathrm{scale}} \ge 1$, depending only on α, such that the following holds for every $d \ge 2$. Let $w(0) \in \Omega$ satisfy $z_{1,2}(0) > 0 > z_{2,2}(0)$, $0 < R_0 < r_s/2$ and $a_0 + Y_0 + |\delta_0| \le \nu R_0$, where R_0, a_0, Y_0 and δ_0 denote the values of R, a, Y and δ at $t = 0$. Then there is a first time $t_* \in [0, T_{\max})$ at which $R(t_*) = r_s/2$; it satisfies $0 < t_* < 2$,*

$$C_{\mathrm{scale}}^{-1} a_0 \Big(\frac{R(t)}{R_0}\Big)^{\mu_\alpha} \le a(t) \le C_{\mathrm{scale}} a_0 \Big(\frac{R(t)}{R_0}\Big)^{\mu_\alpha}, \qquad Y(t) \le C_{\mathrm{scale}} Y_0 \Big(\frac{R(t)}{R_0}\Big)^{\mu_\alpha/2}, \qquad 0 \le t \le t_*,$$

and

$$|\delta(t_*)| \le C_{\mathrm{scale}} \Big(|\delta_0| + \frac{a_0^2 + Y_0^2}{R_0} + a(t_*)^2 \log \frac{1}{R_0} \Big).$$

Proof of Lemma G.4. Let C_{bis} be the constant of Lemma G.2, and let $\nu_1 := \min\{\nu_0, (r_s/(4C_{\mathrm{bis}}))^{1/2}\}$. We first derive bounds that hold for every $\nu \le \nu_1/4$; the value of ν, together with a constant $Q \ge 2$, is fixed afterwards.

Let τ be the supremum of the times $s \in [0, T_{\max})$ such that $R < r_s/2$ and $a + Y + |\delta| < \nu_1 R$ on $[0, s]$. Both conditions hold at $t = 0$, so $\tau > 0$. On $[0, \tau)$, we have $R > 0$, because $a > 0$ and $a + Y + |\delta| < \nu_1 R$, and the signs (G.2) hold, so Lemma G.2 applies. Its first bound and $(a^2 + Y^2)/R^2 \le \nu_1^2$ give $\dot{R} \ge r_s - R - r_s/4 \ge r_s/4$. Hence R is increasing, $R \ge R_0$, and $t < 4(r_s/2 - R_0)/r_s < 2$ for $t < \tau$. Moreover $7R_0/8 \le r_i \le 9r_s/16$ on $[0, \tau)$. If $\tau = T_{\max}$, the solution would stay in a compact subset of Ω on a bounded interval, which Lemma A.1 excludes. Thus $\tau < \min\{2, T_{\max}\}$, the estimates above extend to $[0, \tau]$ by continuity, and at τ either $R = r_s/2$ or $a + Y + |\delta| = \nu_1 R$.

Since $\dot{R} \ge r_s/4$, we have $\mathrm{d}s \le (4/r_s)\mathrm{d}R$, so an integral in time of a nonnegative quantity over $[0, t] \subset [0, \tau]$ is at most $4/r_s$ times the corresponding integral in R over $[R_0, R(t)]$. Moreover, by the first bound of Lemma G.2,

$$\Big| \frac{1}{\dot{R}} - \frac{1}{r_s - R} \Big| = \frac{|\dot{R} - (r_s - R)|}{\dot{R}\,(r_s - R)} \le \frac{8C_{\mathrm{bis}}}{r_s^2} \frac{a^2 + Y^2}{R^2}.$$

Since $1/(R(r_s - R)) = (1/R + 1/(r_s - R))/r_s$ and $0 \le \log((r_s - R_0)/(r_s - R)) \le \log 2$ for $R \le r_s/2$, it follows that

$$\Big| \mu_\alpha r_s \int_0^t \frac{\mathrm{d}s}{R} - \mu_\alpha \log \frac{R(t)}{R_0} \Big| \le \mu_\alpha \log 2 + \frac{8C_{\mathrm{bis}}}{r_s} \int_{R_0}^{R(t)} \frac{a^2 + Y^2}{R^3}\,\mathrm{d}R. \tag{G.11}$$

Let τ_Q be the supremum of the times $s \in [0,\tau]$ such that, on $[0,s]$,

$$a \le Q a_0 \Big(\frac{R}{R_0}\Big)^{\mu_\alpha}, \qquad Y \le Q Y_0 \Big(\frac{R}{R_0}\Big)^{\mu_\alpha/2}. \tag{G.12}$$

If $Y_0 > 0$, both inequalities are strict at $t = 0$. If $Y_0 = 0$, then $Y \equiv 0$ and the second inequality holds throughout, while the first is strict at $t = 0$. In either case $\tau_Q > 0$, and if $\tau_Q < \tau$, then, by continuity, the first inequality, or the second with $Y_0 > 0$, is an equality at τ_Q. On $[0,\tau_Q]$, since $R \ge R_0$, $\mu_\alpha < 1$ and $a_0, Y_0 \le \nu R_0$, (G.12) gives

$$\int_{R_0}^{R(t)} \frac{a^2+Y^2}{R^3}\,\mathrm{d}R \le Q^2\Big(\frac{a_0^2}{R_0^{2\mu_\alpha}}\int_{R_0}^{\infty} R^{2\mu_\alpha-3}\mathrm{d}R + \frac{Y_0^2}{R_0^{\mu_\alpha}}\int_{R_0}^{\infty} R^{\mu_\alpha-3}\mathrm{d}R\Big) \le \frac{Q^2\nu^2}{1-\mu_\alpha}.$$

Similarly, $R^{2\mu_\alpha-2} \le R_0^{2\mu_\alpha-2}$ and $R^{\mu_\alpha-2} \le R_0^{\mu_\alpha-2}$ for $R \ge R_0$ give $\int_{R_0}^{R(t)}(a^2+Y^2)R^{-2}\mathrm{d}R \le 2Q^2\nu^2 R(t)$, and the first bound of Lemma G.2 yields

$$|\delta(t)| \le |\delta_0| + \frac{4C_{\mathrm{bis}}}{r_s}\int_{R_0}^{R(t)} \frac{a^2+Y^2}{R^2}\,\mathrm{d}R \le \Big(\nu + \frac{8C_{\mathrm{bis}}Q^2\nu^2}{r_s}\Big)R(t). \tag{G.13}$$

For the term with δ in Lemma G.2, we use $|\delta(s)| \le |\delta_0| + \int_0^s |\dot\delta|$ and exchange the order of integration. Since $\int_{s'}^t \mathrm{d}s/R(s)^2 \le (4/r_s)\int_{R(s')}^{\infty}\mathrm{d}R/R^2 = 4/(r_s R(s'))$, we find

$$\int_0^t \frac{|\delta|}{R^2}\,\mathrm{d}s \le \frac{4}{r_s}\Big(\frac{|\delta_0|}{R_0} + \int_0^t \frac{|\dot\delta|}{R}\,\mathrm{d}s\Big) \le \frac{4}{r_s}\Big(\nu + C_{\mathrm{bis}}\int_0^t \frac{a^2+Y^2}{R^3}\,\mathrm{d}s\Big).$$

Combining these bounds with (G.11), the second and third bounds of Lemma G.2, $(a+Y)/R < \nu_1 \le 1$ and $t < 2$, we obtain, for $0 \le t \le \tau_Q$,

$$\Big|\log\frac{a(t)}{a_0} - \mu_\alpha \log\frac{R(t)}{R_0}\Big| \le \Gamma_1 + \Gamma_2(\nu + Q^2\nu^2), \qquad Y(t)^2 \le Y_0^2\Big(\frac{R(t)}{R_0}\Big)^{\mu_\alpha} e^{\Gamma_1 + \Gamma_2(\nu+Q^2\nu^2)},$$

where $\Gamma_1 := \mu_\alpha \log 2 + 2C_{\mathrm{bis}}$ and Γ_2 depends only on α; the second bound follows from Grönwall's inequality for Y^2. We now choose $Q := 2e^{\Gamma_1+1}$ and then $\nu \in (0, \nu_1/4]$ so small that

$$\Gamma_2(\nu+Q^2\nu^2) \le 1, \qquad (2Q+1)\nu + \frac{8C_{\mathrm{bis}}Q^2\nu^2}{r_s} \le \frac{\nu_1}{2}.$$

Then $a \le (Q/2)a_0(R/R_0)^{\mu_\alpha}$ and $Y \le (Q/2)Y_0(R/R_0)^{\mu_\alpha/2}$ on $[0,\tau_Q]$, so at τ_Q the first inequality in (G.12) is strict, and so is the second when $Y_0 > 0$. Hence $\tau_Q = \tau$. Moreover, $a_0(R/R_0)^{\mu_\alpha} \le \nu R$ and $Y_0(R/R_0)^{\mu_\alpha/2} \le \nu R$, so (G.13) gives $a + Y + |\delta| \le \nu_1 R/2$ on $[0,\tau]$. At τ, therefore, $R = r_s/2$, and $t_* = \tau \in (0,2)$.

The bounds for a and Y hold with $C_{\mathrm{scale}} \ge Q$. For $\delta(t_*)$, we use the first inequality in (G.13) with (G.12). The contribution of Y^2 to the integral is at most $Q^2Y_0^2/((1-\mu_\alpha)R_0)$. For the contribution of a^2, if $\mu_\alpha < 1/2$, then $a_0^2R_0^{-2\mu_\alpha}\int_{R_0}^{\infty} R^{2\mu_\alpha-2}\mathrm{d}R = a_0^2/((1-2\mu_\alpha)R_0)$. If $\mu_\alpha \ge 1/2$, then $R^{2\mu_\alpha-2} \le 1/R$ for $R \le r_s/2$, because $r_s = 2l(\alpha/2) \le 1$ by (3.4), (A.1b) and Lemma A.2. Hence $a_0^2R_0^{-2\mu_\alpha}\int_{R_0}^{r_s/2} R^{2\mu_\alpha-2}\mathrm{d}R \le a_0^2R_0^{-2\mu_\alpha}\log(1/R_0)$. In this case, the lower bound for a at t_* gives $a_0R_0^{-\mu_\alpha} \le (Q/2)(2/r_s)^{\mu_\alpha}a(t_*) \le (Q/r_s)a(t_*)$. Hence the bound for $\delta(t_*)$ holds with C_{scale} large enough, depending only on α. ∎

G.4 Entry into a neighborhood of the bisector saddle

We fix $c_0 := r_s/64$ and consider the segment

$$\mathcal{C}_{c_0} := \{((r_s+s)\bar v, (r_s-s)\bar v) : |s| \le c_0\}$$

of the first family in (3.5); it contains w_{saddle}. For $|s| \le c_0$, $\|w_1 - (r_s+s)\bar v\|_2^2 + \|w_2 - (r_s-s)\bar v\|_2^2 = (R+\delta-r_s-s)^2 + (R-\delta-r_s+s)^2 + N_\perp^2$. Hence

$$\mathrm{dist}\big((w_1,w_2),\mathcal{C}_{c_0}\big)^2 = 2(R-r_s)^2 + 2\min_{|s|\le c_0}(\delta-s)^2 + N_\perp^2. \tag{G.14}$$

The analysis in Appendix H.2 starts once the trajectory is within a fixed distance $\varepsilon_{\mathrm{entry}}$ of $\mathcal{C}_{c_0}$. By Lemma G.4, the quantities a, Y and δ are small when R reaches $r_s/2$. Afterwards R relaxes exponentially to r_s, and over the bounded time this takes, a and Y change at most by bounded factors and δ changes by $O(a^2+Y^2)$. The next lemma combines this with Lemmas G.3 and G.4.

Lemma G.5 (Entry into a neighborhood of the bisector saddle)**.** *Fix $\alpha \in (0, \pi/2]$, $C \geq 1$ and $\varepsilon_{\mathrm{entry}} \in (0, c_0]$. There are $c_{\mathrm{entry}} \in (0,1]$ and $C_{\mathrm{entry}} \geq 1$, depending only on α, C and $\varepsilon_{\mathrm{entry}}$, such that the following holds for every $d \geq 2$. Suppose that the initialization is C-balanced, that $z_{1,1}(0), z_{2,1}(0) > 0$, $z_{1,2}(0) > 0 > z_{2,2}(0)$ and $a_{\mathrm{init}} \geq r_{\mathrm{init}}/(C\sqrt{d})$, and that*

$$r_{\mathrm{init}}\sqrt{d} \leq c_{\mathrm{entry}}, \qquad r_{\mathrm{init}}\sqrt{d}\left(1 + \log\frac{\sqrt{d}}{r_{\mathrm{init}}} + \log d\right) \leq c_{\mathrm{entry}}. \tag{G.15}$$

We write $\varepsilon_{\mathrm{init}}$ for the left side of the second inequality. Then there is a first time $t_{\mathrm{entry}} \in [0, T_{\max})$ at which $\mathrm{dist}((w_1, w_2), \mathcal{C}_{c_0}) \leq \varepsilon_{\mathrm{entry}}$. Moreover $t_{\mathrm{entry}} \leq C_{\mathrm{entry}}$, and $a_{\mathrm{e}} := a(t_{\mathrm{entry}})$ and $q_{\mathrm{e}} := r_{\mathrm{init}}\sqrt{d} + a_{\mathrm{e}}^2(1 + \log(1/r_{\mathrm{init}}))$ satisfy

$$\begin{aligned} &C_{\mathrm{entry}}^{-1} a_{\mathrm{init}} r_{\mathrm{init}}^{-\mu_\alpha} \leq a_{\mathrm{e}} \leq C_{\mathrm{entry}} a_{\mathrm{init}} r_{\mathrm{init}}^{-\mu_\alpha}, \qquad Y(t_{\mathrm{entry}})^2 \leq C_{\mathrm{entry}} r_{\mathrm{init}}\sqrt{d}\, a_{\mathrm{e}}, \\ &|\delta(t_{\mathrm{entry}})| \leq C_{\mathrm{entry}} q_{\mathrm{e}}, \qquad \sup_{0 \leq t \leq t_{\mathrm{entry}}} N_\perp(t) + q_{\mathrm{e}} \log\frac{1}{a_{\mathrm{e}}} \leq C_{\mathrm{entry}} \varepsilon_{\mathrm{init}}^{1-\mu_\alpha}. \end{aligned} \tag{G.16}$$

Proof of Lemma G.5. Let ν and C_{scale} be the constants of Lemma G.4, let r_0, B and C_{align} be the constants of Lemma G.3 for α, C and this ν, and let C_{bis} be the constant of Lemma G.2. The constants $C_1, \dots, C_6$ below depend only on α, C and $\varepsilon_{\mathrm{entry}}$. We require $c_{\mathrm{entry}} \leq r_0$; a further condition is imposed below. Since $c_{\mathrm{entry}} \leq 1$, the first inequality in (G.15) gives $r_{\mathrm{init}} \leq r_{\mathrm{init}}\sqrt{d} \leq 1$; hence $\log(\sqrt{d}/r_{\mathrm{init}}) \geq 0$ and $r_{\mathrm{init}}\sqrt{d} \leq \varepsilon_{\mathrm{init}} \leq c_{\mathrm{entry}}$.

We first follow the trajectory until R reaches $r_s/2$. By Lemma G.3, the time τ_0 exists, and $R(\tau_0) = Br_{\mathrm{init}} \leq Br_0 < r_s/4$. We apply Lemma G.4 to the solution $t \mapsto w(\tau_0 + t)$; its hypotheses hold by (G.2) and Lemma G.3. Let $t_1 := \tau_0 + t_*$ be the first time after τ_0 at which $R = r_s/2$, and let $R_0 = Br_{\mathrm{init}}$, a_0, Y_0, δ_0 be the values of R, a, Y, δ at τ_0. Then $C_{\mathrm{align}}^{-1} a_{\mathrm{init}} \leq a_0 \leq C_{\mathrm{align}} a_{\mathrm{init}}$ and $Y_0 + |\delta_0| \leq C_{\mathrm{align}} r_{\mathrm{init}}$. We also use $a_{\mathrm{init}} \leq r_1(0) + r_2(0) = 2r_{\mathrm{init}}$, $a_{\mathrm{init}} \geq r_{\mathrm{init}}/(C\sqrt{d})$, $r_s/(2R_0) = r_s/(2Br_{\mathrm{init}})$ and $r_s/(2B) < 1$. On $[0, \tau_0]$ we bound a and Y by Lemma G.3, and on $[\tau_0, t_1]$ we use $(R/R_0)^{\mu_\alpha} \leq (r_s/(2Br_{\mathrm{init}}))^{\mu_\alpha} \leq r_{\mathrm{init}}^{-\mu_\alpha}$. For $\delta(t_1)$, we use $(a_0^2 + Y_0^2)/R_0 \leq 5C_{\mathrm{align}}^2 r_{\mathrm{init}}/B$ and $\log(1/R_0) \leq \log(1/r_{\mathrm{init}})$. With these facts, Lemma G.4 gives

$$\begin{aligned} &C_1^{-1} a_{\mathrm{init}} r_{\mathrm{init}}^{-\mu_\alpha} \leq a(t_1) \leq C_1 a_{\mathrm{init}} r_{\mathrm{init}}^{-\mu_\alpha}, \qquad \frac{Y(t_1)^2}{a(t_1)} \leq C_{\mathrm{scale}}^3 \frac{Y_0^2}{a_0} \leq C_1 r_{\mathrm{init}}\sqrt{d}, \\ &a(t) + Y(t) \leq C_1 r_{\mathrm{init}}^{1-\mu_\alpha} \quad (0 \leq t \leq t_1), \qquad |\delta(t_1)| \leq C_1\Big(r_{\mathrm{init}} + a(t_1)^2\Big(1 + \log\frac{1}{r_{\mathrm{init}}}\Big)\Big). \end{aligned} \tag{G.17}$$

Since $\sup_{0<r\leq 1} r^{1-\mu_\alpha}(1 + \log(1/r)) \leq 1/(1-\mu_\alpha)$, these bounds give

$$a(t_1) + Y(t_1) + |\delta(t_1)| \leq C_2 r_{\mathrm{init}}^{1-\mu_\alpha}. \tag{G.18}$$

Next we follow the trajectory from t_1 to t_{entry}. We consider the region where $r_s/4 < R < 2r_s$ and $a + Y + |\delta| < \nu_0 r_s/4$. There Lemma G.2 applies, and it gives

$$|\dot{R} - (r_s - R)| + |\dot{\delta}| \leq \frac{16 C_{\mathrm{bis}}}{r_s^2}(a^2 + Y^2), \qquad |\dot{a}| \leq C_3 a, \qquad \frac{\mathrm{d}}{\mathrm{d}t} Y^2 \leq C_3 Y^2,$$

with $C_3 := (3 + C_{\mathrm{bis}}(1 + r_s))/r_s$, because $2\sin(\alpha/2)/(\pi R) \leq 3/r_s$, $(a+Y)/R \leq 1$, $|\delta|/R^2 \leq 4\nu_0/r_s$ and $(a^2 + Y^2)/R^3 \leq 4\nu_0^2/r_s$ in this region. Let $T_2 := \log(2r_s/\varepsilon_{\mathrm{entry}}) > 0$, $C_4 := e^{C_3 T_2}(1 + 16 C_{\mathrm{bis}} r_s^{-2}(1 + T_2) e^{C_3 T_2})$ and $S := a(t_1) + Y(t_1) + |\delta(t_1)|$. We now impose the remaining condition on c_{entry}: $C_2 c_{\mathrm{entry}}^{1-\mu_\alpha} \leq \min\{1, \nu_0 r_s/(8C_4), \varepsilon_{\mathrm{entry}}/(8C_4)\}$. By (G.18) and $r_{\mathrm{init}} \leq c_{\mathrm{entry}}$, S satisfies the same bound, so $w(t_1)$ lies in the region.

Let τ_1 be the supremum of the times $t \in [t_1, T_{\max})$ with $t \leq t_1 + T_2$ such that the trajectory lies in the region on $[t_1, t]$. For $t_1 \leq t < \tau_1$, Grönwall's inequality gives $a(t) + Y(t) \leq e^{C_3 T_2}(a(t_1) + Y(t_1))$ and $e^{-C_3 T_2} a(t_1) \leq a(t) \leq e^{C_3 T_2} a(t_1)$. Integrating the bound for $\dot{\delta}$, and applying the variation-of-constants formula to $\frac{\mathrm{d}}{\mathrm{d}t}(R - r_s) = -(R - r_s) + (\dot{R} - (r_s - R))$ with $R(t_1) - r_s = -r_s/2$, we obtain

$$|\delta(t) - \delta(t_1)| + \Big|R(t) - r_s + \frac{r_s}{2} e^{-(t-t_1)}\Big| \leq \frac{16 C_{\mathrm{bis}}}{r_s^2}(1 + T_2) e^{2C_3 T_2}\big(a(t_1) + Y(t_1)\big)^2.$$

Since $S \le 1$, the last two bounds give, for $t_1 \le t < \tau_1$,

$$a(t) + Y(t) + |\delta(t)| \le C_4 S, \qquad \Big| R(t) - r_s + \frac{r_s}{2} e^{-(t-t_1)} \Big| \le C_4 S^2. \tag{G.19}$$

Hence, on $[t_1, \tau_1)$, we have $a + Y + |\delta| \le \nu_0 r_s/8$ and $3r_s/8 \le R \le 9r_s/8$, since $C_4 S^2 \le C_4 S \le r_s/8$. By Lemma G.2, these bounds also give $21r_s/64 \le r_i \le 81r_s/64$. If $\tau_1 = T_{\max}$, the solution would therefore stay in a compact subset of Ω on the bounded interval $[t_1, T_{\max})$, which Lemma A.1 excludes. Hence $\tau_1 < T_{\max}$, the bounds extend to τ_1 by continuity, and they keep the trajectory inside the region at τ_1. Thus $\tau_1 = t_2 \coloneqq t_1 + T_2$, and (G.19) holds on $[t_1, t_2]$.

At t_2, $\frac{r_s}{2} e^{-T_2} = \varepsilon_{\text{entry}}/4$, so $|R - r_s| \le \varepsilon_{\text{entry}}/4 + \varepsilon_{\text{entry}}/8$. By (G.14) with $s = 0$ and $N_\perp \le a + Y$,

$$\operatorname{dist}\big((w_1, w_2), \mathcal{C}_{c_0}\big)^2 \le 2(R - r_s)^2 + 2\delta^2 + (a + Y)^2 \le 2\Big(\frac{3\varepsilon_{\text{entry}}}{8}\Big)^2 + 3\Big(\frac{\varepsilon_{\text{entry}}}{8}\Big)^2 < \varepsilon_{\text{entry}}^2.$$

Before t_1, we have $R < r_s/2$, by the definitions of τ_0 and t_*, and (G.14) gives $\operatorname{dist}((w_1, w_2), \mathcal{C}_{c_0}) > r_s/\sqrt{2} > \varepsilon_{\text{entry}}$. Hence t_{entry} exists and $t_1 \le t_{\text{entry}} \le t_2 \le C_{\text{align}} + 2 + T_2$.

It remains to prove (G.16). Let $C_5 \coloneqq C_1 C_4^2$. On $[t_1, t_{\text{entry}}]$, a changes at most by the factor $e^{C_3 T_2} \le C_4$, Y^2 increases at most by this factor, and $a + Y \le C_4(a(t_1) + Y(t_1))$. With (G.17), this gives the bounds for a_{e} and $Y(t_{\text{entry}})^2$ with C_5 in place of C_{entry}, and the bound $N_\perp \le a + Y \le C_5 r_{\text{init}}^{1-\mu_\alpha}$ on $[0, t_{\text{entry}}]$. By the bound for $\dot\delta$ above, $|\delta(t_{\text{entry}}) - \delta(t_1)| \le C_4(a(t_1)^2 + Y(t_1)^2)$, and $Y(t_1)^2 \le C_1 r_{\text{init}} \sqrt{d}\, a(t_1) \le C_1 r_{\text{init}} \sqrt{d}$ because $a(t_1) \le S \le 1$. Together with (G.17) and $a(t_1) \le e^{C_3 T_2} a_{\text{e}} \le C_4 a_{\text{e}}$, this gives the bound for $\delta(t_{\text{entry}})$.

For the last bound in (G.16), we note that $a_{\text{e}} \le C_4 S \le 1$ and that $r_{\text{init}}^{-\mu_\alpha} \ge 1$, so $a_{\text{e}} \ge C_5^{-1} a_{\text{init}} \ge r_{\text{init}}/(C C_5 \sqrt{d})$. Hence, with $C_6 \coloneqq 1 + \log(C C_5)$, $\log(1/a_{\text{e}}) \le \log(C C_5) + \log(\sqrt{d}/r_{\text{init}}) \le C_6(1 + \log(\sqrt{d}/r_{\text{init}}) + \log d)$, and $r_{\text{init}} \sqrt{d} \log(1/a_{\text{e}}) \le C_6 \varepsilon_{\text{init}}$. Since $r_{\text{init}} \sqrt{d} \le 1$, we have $d \le r_{\text{init}}^{-2}$ and $1 + \log(\sqrt{d}/r_{\text{init}}) + \log d \le 4(1 + \log(1/r_{\text{init}}))$. With $a_{\text{e}} \le 2 C_5 r_{\text{init}}^{1-\mu_\alpha}$, it follows that

$$a_{\text{e}}^2 \Big(1 + \log \frac{1}{r_{\text{init}}}\Big) \log \frac{1}{a_{\text{e}}} \le 16 C_5^2 C_6\, r_{\text{init}}^{2-2\mu_\alpha} \Big(1 + \log \frac{1}{r_{\text{init}}}\Big)^2 \le \frac{64 C_5^2 C_6}{(1-\mu_\alpha)^2}\, r_{\text{init}}^{1-\mu_\alpha},$$

where we used $\sup_{0<r\le 1} r^{1-\mu_\alpha}(1 + \log(1/r))^2 \le 4/(1-\mu_\alpha)^2$. Since $r_{\text{init}} \le \varepsilon_{\text{init}} \le 1$, the terms $C_5 r_{\text{init}}^{1-\mu_\alpha}$, $C_6 \varepsilon_{\text{init}}$ and the last bound are all at most a constant times $\varepsilon_{\text{init}}^{1-\mu_\alpha}$. This proves (G.16) with C_{entry} large enough. ∎

H Logarithmic escape from the bisector saddle

H.1 Hessian at the bisector saddle

Proof of Lemma 4.3. At w_{saddle}, both students equal $w_s = r_s \bar v$. By the continuous extension in Lemma A.6, the off-diagonal blocks $Q_2(w_s, w_s)$ of the Hessian equal $\frac{1}{2} I_d$, and the term $Q_1(w_s, w_s)$ in the diagonal blocks vanishes. As recalled at the start of Appendix F, $v_1 = \cos(\alpha/2)\bar v + \sin(\alpha/2) v_\Delta$ and $v_2 = \cos(\alpha/2)\bar v - \sin(\alpha/2) v_\Delta$. Hence the angle between w_s and each teacher is $\alpha/2$, $\bar n_{v_1, w_s} = v_\Delta$ and $\bar n_{v_2, w_s} = -v_\Delta$, so $Q_1(w_s, v_1) = Q_1(w_s, v_2) = (\mu_\alpha/4)(I_d - \bar v \bar v^\top + v_\Delta v_\Delta^\top)$, where $\mu_\alpha = 2\sin(\alpha/2)/(\pi r_s)$ is the rate defined before Theorem 4.6. Both diagonal blocks therefore equal

$$A \coloneqq \frac{1}{2} I_d - \frac{\mu_\alpha}{2}\big(I_d - \bar v \bar v^\top + v_\Delta v_\Delta^\top\big).$$

The vectors $\bar v$, v_Δ and $u_3, \dots, u_d$ are eigenvectors of A, with eigenvalues $\frac{1}{2}$, $\frac{1}{2} - \mu_\alpha$ and $\frac{1}{2} - \mu_\alpha/2$, respectively, because $I_d - \bar v \bar v^\top + v_\Delta v_\Delta^\top$ maps them to 0, $2v_\Delta$ and u_j. The Hessian maps (x, x) to $((A + \frac{1}{2} I_d)x, (A + \frac{1}{2} I_d)x)$ and $(x, -x)$ to $((A - \frac{1}{2} I_d)x, -(A - \frac{1}{2} I_d)x)$. Hence the vectors $(x, \pm x)$ with $x \in \{\bar v, v_\Delta, u_3, \dots, u_d\}$ are eigenvectors of the Hessian, with the following eigenvalues:

x	eigenvalue of (x, x)	eigenvalue of $(x, -x)$
$\bar v$	1	0
v_Δ	$1 - \mu_\alpha$	$-\mu_\alpha$
u_j, $3 \le j \le d$	$1 - \mu_\alpha/2$	$-\mu_\alpha/2$

By (G.3), $\mu_\alpha < 1$, so the eigenvalues in the middle column are positive. The $2d$ vectors $(x, \pm x)$ are orthogonal, so they form a basis of $(\mathbb{R}^d)^2$, and the table gives the negative, positive and zero eigenspaces stated in Lemma 4.3. ∎

H.2 Proof of Theorem 4.6

By Lemma G.5, the trajectory comes within a small fixed distance of the segment $\mathcal{C}_{c_0}$ after a bounded time, with a small planar splitting amplitude a_{e}. Afterwards a grows at a rate close to μ_α until $N_\perp$ reaches ε, which takes a time of order $\mu_\alpha^{-1}\log(1/a_{\mathrm{e}})$. Over so long a time, a constant error in the growth rate would change the escape time by an amount that grows with $\log(1/a_{\mathrm{e}})$, so we need errors that vanish at the saddle. Once the rate $2\sin(\alpha/2)/(\pi R)$ in Lemma G.2 is compared with μ_α, the bounds of that lemma take this form near w_{saddle}.

Lemma H.1 (Equations of motion near the bisector saddle)**.** *Fix $\alpha \in (0,\pi/2]$, and let ν_0 be as in Lemma G.2. There is $C_{\mathrm{local}} \geq 1$, depending only on α, such that the following holds for every $d \geq 2$. Let $w(0) \in \Omega$, and let $t \in [0, T_{\max})$ be a time at which $z_{1,2} > 0 > z_{2,2}$ and*

$$|R - r_s| + |\delta| + a + Y \leq \nu_0 r_s. \tag{H.1}$$

Then $r_s/2 \leq r_i \leq 2r_s$ for $i = 1, 2$, and

$$|\dot a - \mu_\alpha a| \leq C_{\mathrm{local}}\big(|R - r_s| + |\delta| + a + Y\big)a, \tag{H.2}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}Y^2 \leq \big(\mu_\alpha + C_{\mathrm{local}}(|R - r_s| + |\delta| + a + Y)\big)Y^2, \tag{H.3}$$

$$|\dot R - (r_s - R)| + |\dot\delta| \leq C_{\mathrm{local}}(a^2 + Y^2). \tag{H.4}$$

Proof of Lemma H.1. Let C_{bis} be the constant of Lemma G.2. We first check the hypotheses of that lemma. Since $\nu_0 \leq 1/8$, (H.1) gives $7r_s/8 \leq R \leq 9r_s/8$. If $R \geq r_s$, then $a+Y+|\delta| \leq \nu_0 r_s \leq \nu_0 R$; if $R < r_s$, then $a + Y + |\delta| \leq \nu_0 r_s - (r_s - R) \leq \nu_0 R$, because $\nu_0 \leq 1$. Hence the lemma applies. Its first claim gives $r_s/2 \leq 7R/8 \leq r_i \leq 9R/8 \leq 2r_s$, and its first bound, together with $R \geq 7r_s/8$, gives (H.4) whenever $C_{\mathrm{local}} \geq 2C_{\mathrm{bis}}/r_s^2$.

For the other two bounds, we compare the error terms of Lemma G.2 with $|R - r_s| + |\delta| + a + Y$. Since $R \geq 7r_s/8$ and $a + Y \leq r_s/8$,

$$\frac{a+Y}{R} + \frac{|\delta|}{R^2} + \frac{a^2+Y^2}{R^3} \leq \frac{2(a+Y)}{r_s} + \frac{2|\delta|}{r_s^2} + \frac{a+Y}{4r_s^2} \leq 3\big(1 + r_s^{-2}\big)\big(|\delta| + a + Y\big),$$

where the last step uses $1/r_s \leq 1 + r_s^{-2}$. Moreover, $2\sin(\alpha/2)/(\pi R) = \mu_\alpha r_s/R$, and $\mu_\alpha < 1$ by (G.3), so

$$\left|\frac{2\sin(\alpha/2)}{\pi R} - \mu_\alpha\right| = \frac{\mu_\alpha|R - r_s|}{R} \leq \frac{2|R - r_s|}{r_s} \leq 2\big(1 + r_s^{-2}\big)|R - r_s|.$$

Substituting these two bounds into the second and third bounds of Lemma G.2 gives (H.2) and (H.3) with $C_{\mathrm{local}} := 4C_{\mathrm{bis}}(1 + r_s^{-2})$, since $C_{\mathrm{bis}} \geq 1$. This constant also satisfies the requirement above. ∎

For each $\alpha \in (0,\pi/2]$, with ν_0 and C_{local} as in Lemma G.2 and Lemma H.1, we set $\zeta := \min\{\nu_0 r_s/8, \mu_\alpha/(16C_{\mathrm{local}})\}$ and $\varepsilon_{\mathrm{entry}} := \sqrt{2}\,\zeta/100$, which depend only on α. This choice is made so that, whenever $|R - r_s| + |\delta| + a + Y \leq 4\zeta$, the inequality (H.1) holds and (H.2) gives $\dot a \geq \mu_\alpha a/2$. Since $\nu_0 \leq 1/8$, we have $\varepsilon_{\mathrm{entry}} < \zeta \leq r_s/64 = c_0$, so Lemma G.5 applies with this $\varepsilon_{\mathrm{entry}}$. Below, t_{entry}, a_{e} and q_{e} are as in that lemma.

Lemma H.2 (Escape from a neighborhood of the bisector saddle)**.** *Fix $\alpha \in (0,\pi/2]$ and $C \geq 1$, and let c_{entry} and C_{entry} be the constants of Lemma G.5 for α, C and $\varepsilon_{\mathrm{entry}}$. There is $\varepsilon_0 > 0$, depending only on α and C, such that for every $\varepsilon \in (0, \varepsilon_0]$ there is $c_{\mathrm{pass}} \in (0, c_{\mathrm{entry}}]$, depending only on α, C and ε, with the following property. For every $d \geq 2$ and every initialization that satisfies the hypotheses of Lemma G.5 with c_{pass} in place of c_{entry}, the time T_ε is finite and*

$$\left|T_\varepsilon - t_{\mathrm{entry}} - \frac{1}{\mu_\alpha}\log\frac{1}{a_{\mathrm{e}}}\right| \leq \frac{C_{\mathrm{local}} + \log(4/\varepsilon)}{\mu_\alpha}. \tag{H.5}$$

Proof of Lemma H.2. We set $C_Y := 4C_{\mathrm{entry}}e^{4C_{\mathrm{local}}}$, $C_\delta := 8(C_{\mathrm{entry}} + C_{\mathrm{local}}(1 + 2C_Y)/\mu_\alpha)$ and $\varepsilon_0 := \zeta/(1 + 2C_\delta\zeta)$. Then $C_\delta \geq 16C_{\mathrm{local}}$, since $C_Y \geq 4$ and $\mu_\alpha < 1$, and, for $0 < \varepsilon \leq \varepsilon_0$,

$$2C_\delta\varepsilon^2 + \varepsilon \leq \zeta. \tag{H.6}$$

We fix $\varepsilon \in (0, \varepsilon_0]$. Since $\mu_\alpha < 1$, we have $C_{\mathrm{local}}\zeta \le \mu_\alpha/16 < 1/16$; as $C_{\mathrm{local}} \ge 1$, also $\varepsilon \le \zeta < 1/16$. We choose $c_{\mathrm{pass}} \in (0, c_{\mathrm{entry}}]$ so small that

$$\beta := 2C_{\mathrm{entry}} c_{\mathrm{pass}}^{1-\mu_\alpha} \le \min\Big\{\frac{\varepsilon}{8C_Y}, \frac{\zeta}{2C_\delta}, \frac{\zeta^2}{8C_Y}\Big\}.$$

Now let $d \ge 2$, and consider an initialization that satisfies the hypotheses of Lemma G.5 with c_{pass} in place of c_{entry}. Since $c_{\mathrm{pass}} \le c_{\mathrm{entry}}$, that lemma applies, and $\varepsilon_{\mathrm{init}} \le c_{\mathrm{pass}}$. The first inequality in (G.15) gives $r_{\mathrm{init}} \le r_{\mathrm{init}}\sqrt{d} \le c_{\mathrm{pass}} \le 1$, and $c_{\mathrm{pass}} \le c_{\mathrm{pass}}^{1-\mu_\alpha} \le \beta/2$. Since $a_{\mathrm{init}} \le r_1(0) + r_2(0) = 2r_{\mathrm{init}}$, (G.16) gives $a_{\mathrm{e}} \le 2C_{\mathrm{entry}} r_{\mathrm{init}}^{1-\mu_\alpha} \le \beta$ and $\sup_{0\le t\le t_{\mathrm{entry}}} N_\perp(t) + q_{\mathrm{e}}\log(1/a_{\mathrm{e}}) \le C_{\mathrm{entry}}\varepsilon_{\mathrm{init}}^{1-\mu_\alpha} \le \beta/2$. Since $\beta \le \varepsilon/(8C_Y) < 1/e$, we have $\log(1/a_{\mathrm{e}}) \ge 1$. Hence

$$a_{\mathrm{e}} \le \beta, \qquad r_{\mathrm{init}}\sqrt{d} \le \frac{\beta}{2}, \qquad q_{\mathrm{e}} \le q_{\mathrm{e}}\log\frac{1}{a_{\mathrm{e}}} \le \frac{\beta}{2}, \qquad \sup_{0\le t\le t_{\mathrm{entry}}} N_\perp(t) \le \frac{\beta}{2}. \tag{H.7}$$

By (G.16), $Y(t_{\mathrm{entry}})^2 \le C_{\mathrm{entry}} r_{\mathrm{init}}\sqrt{d}\, a_{\mathrm{e}}$ and $|\delta(t_{\mathrm{entry}})| \le C_{\mathrm{entry}} q_{\mathrm{e}}$. Since the distance from (w_1, w_2) to $\mathcal{C}_{c_0}$ is at most $\varepsilon_{\mathrm{entry}}$ at t_{entry}, (G.14) gives $|R(t_{\mathrm{entry}}) - r_s| \le \varepsilon_{\mathrm{entry}}/\sqrt{2} = \zeta/100$.

By (G.2), $z_{1,2} > 0 > z_{2,2}$ on $[0, T_{\max})$, so Lemma H.1 applies wherever (H.1) holds. Let τ be the supremum of the times $t \in [t_{\mathrm{entry}}, T_{\max})$ such that, on $[t_{\mathrm{entry}}, t]$, we have $a < 2\varepsilon$, the inequality (H.1) holds strictly, and

$$Y^2 < C_Y r_{\mathrm{init}}\sqrt{d}\, a, \qquad |\delta| < C_\delta(q_{\mathrm{e}} + a^2), \qquad |R - r_s| < \zeta, \qquad \int_{t_{\mathrm{entry}}}^{t} |R(u) - r_s|\,\mathrm{d}u < 1. \tag{H.8}$$

These conditions hold at t_{entry}: $a_{\mathrm{e}} \le \beta < 2\varepsilon$; the first two inequalities follow from the bounds at t_{entry} above, because $C_Y, C_\delta > C_{\mathrm{entry}}$; and $|R(t_{\mathrm{entry}}) - r_s| \le \zeta/100$. Moreover, by (H.7) and the choice of β, we have $C_{\mathrm{entry}} q_{\mathrm{e}} \le C_\delta\beta/16 \le \zeta/32$, $a_{\mathrm{e}} \le \zeta/32$ and $Y(t_{\mathrm{entry}})^2 \le C_{\mathrm{entry}}\beta^2/2 \le C_Y\beta/8 \le \zeta^2/64$, so the left side of (H.1) is less than $\zeta < \nu_0 r_s$ at t_{entry}. Hence $\tau > t_{\mathrm{entry}}$.

We now show that, on $[t_{\mathrm{entry}}, \tau)$, each of these conditions other than $a < 2\varepsilon$ holds with room to spare. On this interval, $a < 2\varepsilon \le 1$, so (H.7) and (H.8) give $Y^2 \le C_Y\beta \le \zeta^2/8$ and $|\delta| \le C_\delta\beta/2 + 4C_\delta\varepsilon^2$. Since $C_\delta\beta/2 \le \zeta/4$ and, by (H.6), $4C_\delta\varepsilon^2 + 2\varepsilon \le 2\zeta$, it follows that $|R - r_s| + |\delta| + a + Y \le \zeta + (\zeta/4 + 2\zeta) + \zeta/2 \le 4\zeta$. Since $16C_{\mathrm{local}}\zeta \le \mu_\alpha$, (H.2) gives

$$\dot a \ge \frac{\mu_\alpha}{2} a \qquad \text{on } [t_{\mathrm{entry}}, \tau). \tag{H.9}$$

Hence a is strictly increasing, and $\mathrm{d}t \le 2\mathrm{d}a/(\mu_\alpha a)$. Changing the variable of integration from t to a, and using $a \le 1$ and $Y^2 \le C_Y r_{\mathrm{init}}\sqrt{d}\, a$, we obtain, for $t_{\mathrm{entry}} \le t < \tau$,

$$\begin{aligned} &t - t_{\mathrm{entry}} \le \frac{2}{\mu_\alpha}\log\frac{1}{a_{\mathrm{e}}}, \qquad \int_{t_{\mathrm{entry}}}^{t} a\,\mathrm{d}u \le \frac{2a(t)}{\mu_\alpha}, \qquad \int_{t_{\mathrm{entry}}}^{t} a^2\,\mathrm{d}u \le \frac{a(t)^2}{\mu_\alpha}, \\ &\int_{t_{\mathrm{entry}}}^{t} Y^2\,\mathrm{d}u \le \frac{2C_Y}{\mu_\alpha} r_{\mathrm{init}}\sqrt{d}\, a(t), \qquad \int_{t_{\mathrm{entry}}}^{t} Y\,\mathrm{d}u \le \frac{4}{\mu_\alpha}\big(C_Y r_{\mathrm{init}}\sqrt{d}\, a(t)\big)^{1/2}. \end{aligned} \tag{H.10}$$

Integrating (H.4) and using (H.10) together with $r_{\mathrm{init}}\sqrt{d}\, a \le r_{\mathrm{init}}\sqrt{d} \le q_{\mathrm{e}}$, we obtain

$$|\delta(t)| \le C_{\mathrm{entry}} q_{\mathrm{e}} + \frac{C_{\mathrm{local}}}{\mu_\alpha}\big(a(t)^2 + 2C_Y r_{\mathrm{init}}\sqrt{d}\, a(t)\big) \le \frac{C_\delta}{8}\big(q_{\mathrm{e}} + a(t)^2\big).$$

By (H.8), (H.10) and (H.7),

$$\int_{t_{\mathrm{entry}}}^{t} (|\delta| + a + Y)\,\mathrm{d}u \le \frac{1}{\mu_\alpha}\Big(2C_\delta q_{\mathrm{e}}\log\frac{1}{a_{\mathrm{e}}} + C_\delta a(t)^2 + 2a(t) + 4\big(C_Y\beta\big)^{1/2}\Big) \le \frac{13\zeta}{2\mu_\alpha} < \frac{1}{2},$$

where we used $C_\delta\beta \le \zeta/2$, $4C_\delta\varepsilon^2 + 4\varepsilon \le 4\zeta$, $(C_Y\beta)^{1/2} \le \zeta/2$ and $16\zeta \le \mu_\alpha$. Since $a > 0$, Y^2/a is continuously differentiable, and (H.2) and (H.3) give

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{Y^2}{a} \le 2C_{\mathrm{local}}\big(|R - r_s| + |\delta| + a + Y\big)\frac{Y^2}{a}.$$

By Grönwall's inequality, the last integral bound and the bound $\int_{t_{\text{entry}}}^{t} |R - r_s|\,\mathrm{d}u < 1$, we obtain

$$\frac{Y(t)^2}{a(t)} \le \frac{Y(t_{\text{entry}})^2}{a_{\text{e}}} e^{3C_{\text{local}}} \le C_{\text{entry}} e^{3C_{\text{local}}} r_{\text{init}}\sqrt{d} < \frac{C_Y}{4} r_{\text{init}}\sqrt{d}.$$

Applying the variation-of-constants formula to $\frac{\mathrm{d}}{\mathrm{d}t}(R - r_s) = -(R - r_s) + (\dot{R} - (r_s - R))$ and using (H.4), we obtain

$$|R(t) - r_s| \le e^{-(t-t_{\text{entry}})}\frac{\zeta}{100} + C_{\text{local}} \int_{t_{\text{entry}}}^{t} e^{-(t-u)}\big(a(u)^2 + Y(u)^2\big)\,\mathrm{d}u \le \frac{\zeta}{100} + C_{\text{local}}\big(4\varepsilon^2 + C_Y\beta\big) < \frac{\zeta}{2},$$

because $32C_{\text{local}}\varepsilon^2 \le 2C_\delta\varepsilon^2 \le \zeta$ and $C_{\text{local}}C_Y\beta \le C_{\text{local}}\zeta^2/8 \le \zeta/128$. Integrating the middle expression over $[t_{\text{entry}}, t]$ and exchanging the order of integration, we find in the same way

$$\int_{t_{\text{entry}}}^{t} |R(u) - r_s|\,\mathrm{d}u \le \frac{\zeta}{100} + \frac{C_{\text{local}}}{\mu_\alpha}\big(a(t)^2 + 2C_Y r_{\text{init}}\sqrt{d}\,a(t)\big) \le \frac{\zeta}{100} + \frac{C_{\text{local}}}{\mu_\alpha}\big(4\varepsilon^2 + 2C_Y\beta\big) < \frac{1}{2},$$

by the same two bounds and $16\zeta \le \mu_\alpha$. Finally, with the bounds just proved,

$$|R - r_s| + |\delta| + a + Y \le \frac{\zeta}{2} + \Big(\frac{C_\delta\beta}{16} + \frac{C_\delta}{2}\varepsilon^2 + 2\varepsilon\Big) + \frac{\zeta}{2} \le \frac{\zeta}{2} + \Big(\frac{\zeta}{32} + 2\zeta\Big) + \frac{\zeta}{2} < \frac{\nu_0 r_s}{2}.$$

These bounds are stronger than the conditions defining τ, other than $a < 2\varepsilon$. By Lemma H.1, also $r_s/2 \le r_i \le 2r_s$ on $[t_{\text{entry}}, \tau)$, and (H.9) gives $\tau - t_{\text{entry}} \le (2/\mu_\alpha)\log(2\varepsilon/a_{\text{e}})$. If $\tau = T_{\max}$, the solution would stay in a compact subset of Ω on a bounded interval, which Lemma A.1 excludes. Hence $\tau < T_{\max}$. By continuity the stronger bounds hold at τ, so $a(\tau) = 2\varepsilon$.

Since $a_{\text{e}} \le \beta < \varepsilon/4$, there is a first time $t_- \in (t_{\text{entry}}, \tau)$ at which $a = \varepsilon/4$. On $[t_{\text{entry}}, t_-]$, the bound for Y^2/a and $C_Y\beta \le \varepsilon/8$ give $Y^2 \le C_Y\beta\varepsilon/32 \le (\varepsilon/16)^2$, so $N_\perp \le a + Y \le 5\varepsilon/16$. Together with (H.7), this shows that $N_\perp < \varepsilon$ on $[0, t_-]$, so $T_\varepsilon \ge t_-$. At τ, we have $N_\perp^2 \ge z_{1,2}^2 + z_{2,2}^2 \ge a^2/2 = 2\varepsilon^2$, so $T_\varepsilon \le \tau < \infty$.

It remains to estimate t_- and τ. By (H.2) and the integral bounds above, $\int_{t_{\text{entry}}}^{\tau} |\dot{a}/a - \mu_\alpha|\,\mathrm{d}u \le C_{\text{local}}$. Since $a(t_-) = \varepsilon/4$ and $a(\tau) = 2\varepsilon$, it follows that

$$\frac{1}{\mu_\alpha}\Big(\log\frac{\varepsilon}{4} - C_{\text{local}}\Big) \le t_- - t_{\text{entry}} - \frac{1}{\mu_\alpha}\log\frac{1}{a_{\text{e}}} \le \tau - t_{\text{entry}} - \frac{1}{\mu_\alpha}\log\frac{1}{a_{\text{e}}} \le \frac{1}{\mu_\alpha}\big(\log(2\varepsilon) + C_{\text{local}}\big).$$

As $t_- \le T_\varepsilon \le \tau$ and $\log(2\varepsilon) < 0$, this proves (H.5). ∎

Proof of Theorem 4.6. Let ε_0 be as in Lemma H.2 for the given α and C, fix $\varepsilon \in (0, \varepsilon_0]$, and let c_{pass} be as in that lemma. We show that the theorem holds with $c := c_{\text{pass}}/2$ (so that $c < 1$) and

$$K := C_{\text{entry}} + \frac{C_{\text{local}} + \log(4/\varepsilon) + \log C_{\text{entry}}}{\mu_\alpha},$$

which depend only on α, C and ε. Let $d \ge 2$, and consider an initialization as in the theorem. Exchanging the two students if necessary, we may assume that $z_{1,2}(0) > 0 > z_{2,2}(0)$; this changes none of T_ε, r_{init} and a_{init}. Since $c \le c_{\text{pass}}$, the initialization then satisfies the hypotheses of Lemma G.5 with c_{pass} in place of c_{entry}. By Lemma H.2, T_ε is finite and satisfies (H.5). By Lemma G.5, $0 \le t_{\text{entry}} \le C_{\text{entry}}$, and (G.16) gives $|\log(1/a_{\text{e}}) - \log(r_{\text{init}}^{\mu_\alpha}/a_{\text{init}})| \le \log C_{\text{entry}}$. These three bounds give $|T_\varepsilon - \mu_\alpha^{-1}\log(r_{\text{init}}^{\mu_\alpha}/a_{\text{init}})| \le K$. ∎

H.3 Gaussian initialization

Lemma H.3 (Gaussian initialization)**.** *For every $C > 1$, there is $p_C > 0$, depending only on C, such that the following holds for every $d \ge 2$ and $\sigma > 0$. If $w_1(0), w_2(0) \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I_d)$, then, with probability at least p_C, the initialization is C-balanced, $z_{1,1}(0), z_{2,1}(0) > 0$, $z_{1,2}(0)z_{2,2}(0) < 0$, and*

$$\frac{\sigma\sqrt{d}}{2} < r_{\text{init}} < 2\sigma\sqrt{d}, \qquad a_{\text{init}} > \frac{r_{\text{init}}}{C\sqrt{d}}, \qquad \frac{\sigma}{C} < a_{\text{init}} < 4\sigma.$$

Proof of Lemma H.3. As in the proof of Corollary E.8, let $\lambda := \min\{1/2, (C^2-1)/(C^2+1)\}$, so that $0 < \lambda \le 1/2$ and $1+\lambda \le C^2(1-\lambda) < C^2$. Let $\mathcal{E}$ be the event that

$$z_{1,1}(0) > 0, \quad z_{2,1}(0) > 0, \quad z_{1,2}(0)z_{2,2}(0) < 0, \quad \frac{\sigma\sqrt{1+\lambda}}{C} < |z_{1,2}(0)| < 2\sigma, \quad |z_{2,2}(0)| < 2\sigma,$$
$$(1-\lambda)\sigma^2 d < \|w_i(0)\|_2^2 < (1+\lambda)\sigma^2 d \quad \text{for } i = 1, 2.$$

We first show that the conclusions hold on $\mathcal{E}$. The norm bounds give $\max_i \|w_i(0)\|_2 < \sqrt{(1+\lambda)/(1-\lambda)}\, \min_i \|w_i(0)\|_2 \le C \min_i \|w_i(0)\|_2$, so the initialization is C-balanced. They also give $\sigma\sqrt{(1-\lambda)d} < r_{\text{init}} < \sigma\sqrt{(1+\lambda)d}$, which implies the bounds on r_{init} because $\lambda \le 1/2$. Since $z_{1,2}(0)$ and $z_{2,2}(0)$ have opposite signs, $a_{\text{init}} = |z_{1,2}(0)| + |z_{2,2}(0)|$. Hence $a_{\text{init}} < 4\sigma$ and $a_{\text{init}} > \sigma\sqrt{1+\lambda}/C$. The last quantity is at least σ/C, and it is greater than $r_{\text{init}}/(C\sqrt{d})$ by the upper bound $r_{\text{init}} < \sigma\sqrt{(1+\lambda)d}$.

It remains to bound $\mathbb{P}(\mathcal{E})$ from below. By Lemma F.5, the rescaled coordinates $z_{i,j}(0)/\sigma$ are independent standard normal random variables. Dividing each condition defining $\mathcal{E}$ by a power of σ shows that $\mathcal{E}$ is the event that the vector $(z_{i,j}(0)/\sigma)_{i,j}$ lies in a set $U \subset \mathbb{R}^{2d}$ that depends only on C and d. Hence $\mathbb{P}(\mathcal{E})$ depends only on C and d.

Let $\mathcal{E}_0$ be the event defined by the first line of conditions, and let Z be a standard normal random variable. The event $\mathcal{E}_0$ involves only $z_{1,1}(0)$, $z_{2,1}(0)$, $z_{1,2}(0)$ and $z_{2,2}(0)$. The cases $z_{1,2}(0) > 0$ and $z_{1,2}(0) < 0$ have the same probability by symmetry, so independence gives

$$p_0 := \mathbb{P}(\mathcal{E}_0) = \frac{1}{2}\,\mathbb{P}\Big\{\frac{\sqrt{1+\lambda}}{C} < Z < 2\Big\}\,\mathbb{P}\{0 < Z < 2\} > 0,$$

because $\sqrt{1+\lambda}/C < 1$. This number depends only on C. The random variable $\|w_i(0)\|_2^2/\sigma^2$ has mean d and variance $2d$, so by Chebyshev's inequality the norm condition fails for some i with probability at most $4/(\lambda^2 d)$. Hence $\mathbb{P}(\mathcal{E}) \ge p_0 - 4/(\lambda^2 d) \ge p_0/2$ for $d \ge 8/(p_0\lambda^2)$.

The set U is open, because it is defined by strict inequalities between continuous functions. It is also nonempty: it contains the vector with entry -1 at $(i,j) = (2,2)$ and all other entries equal to 1, because $\sqrt{1+\lambda}/C < 1 < 2$ and, for each i, the squares of the entries (i,j), $1 \le j \le d$, sum to d. The standard Gaussian distribution on $\mathbb{R}^{2d}$ gives positive mass to every nonempty open set, so $\mathbb{P}(\mathcal{E}) > 0$ for every $d \ge 2$. The lemma therefore holds with p_C the minimum of $p_0/2$ and the finitely many probabilities $\mathbb{P}(\mathcal{E})$ with $2 \le d < 8/(p_0\lambda^2)$. ∎

Combining this lemma with Theorem 4.6 gives the escape time from small Gaussian initialization.

Corollary H.4 (Escape time under Gaussian initialization)**.** *Fix $\alpha \in (0, \pi/2]$, $C > 1$ and $\varepsilon > 0$ small enough for Theorem 4.6, and let c and K be the constants of that theorem and p_C the constant of Lemma H.3. Let $d \ge 2$ and $0 < \sigma \le 1$ satisfy $4\sigma d\big(1 + \log(1/\sigma) + \log d\big) \le c$, and let $w_1(0), w_2(0) \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I_d)$. Then, with probability at least p_C, T_ε is finite and*

$$\Big|T_\varepsilon - \Big(\frac{1}{\mu_\alpha} - 1\Big)\log\frac{1}{\sigma} - \frac{1}{2}\log d\Big| \le K + \log 2 + \frac{\max\{\log 4, \log C\}}{\mu_\alpha}.$$

In particular, if $\sigma = d^{-\beta}$ for a fixed $\beta > 1$, the condition on σ holds for all sufficiently large d, and the bound gives $T_\varepsilon = \big((\mu_\alpha^{-1} - 1)\beta + \frac{1}{2}\big)\log d + O(1)$, where the coefficient of $\log d$ is at least $\frac{1}{2}$.

Proof of Corollary H.4. Consider the event of Lemma H.3, which has probability at least p_C. On it, the initialization is C-balanced, satisfies the sign conditions and $a_{\text{init}} \ge r_{\text{init}}/(C\sqrt{d})$, and $r_{\text{init}}\sqrt{d} < 2\sigma d \le c/2$, where the last step uses $4\sigma d \le c$, which follows from the hypothesis because $1 + \log(1/\sigma) + \log d \ge 1$. Since $\sqrt{d}/r_{\text{init}} < 2/\sigma$, $\sigma \le 1$ and $\log 2 \le \log d$, we have

$$1 + \log\frac{\sqrt{d}}{r_{\text{init}}} + \log d < 1 + \log 2 + \log\frac{1}{\sigma} + \log d \le 2\Big(1 + \log\frac{1}{\sigma} + \log d\Big),$$

and the right side is positive. Together with $0 < r_{\text{init}}\sqrt{d} < 2\sigma d$, this gives $r_{\text{init}}\sqrt{d}\,\big(1 + \log(\sqrt{d}/r_{\text{init}}) + \log d\big) < 4\sigma d\big(1 + \log(1/\sigma) + \log d\big) \le c$. Thus Theorem 4.6 applies, so T_ε

is finite and $|T_\varepsilon - \mu_\alpha^{-1}\log(r_{\text{init}}^{\mu_\alpha}/a_{\text{init}})| \le K$. Moreover,

$$\frac{1}{\mu_\alpha}\log\frac{r_{\text{init}}^{\mu_\alpha}}{a_{\text{init}}} - \Big(\frac{1}{\mu_\alpha} - 1\Big)\log\frac{1}{\sigma} - \frac{1}{2}\log d = \log\frac{r_{\text{init}}}{\sigma\sqrt{d}} - \frac{1}{\mu_\alpha}\log\frac{a_{\text{init}}}{\sigma}.$$

By Lemma H.3, the first term on the right has absolute value less than $\log 2$, and $\log(a_{\text{init}}/\sigma)$ lies between $-\log C$ and $\log 4$. This proves the bound.

If $\sigma = d^{-\beta}$ with $\beta > 1$, then $\sigma \le 1$ and $4\sigma d\big(1 + \log(1/\sigma) + \log d\big) = 4d^{1-\beta}\big(1 + (1+\beta)\log d\big)$, which tends to zero as $d \to \infty$. Substituting $\log(1/\sigma) = \beta \log d$ into the bound gives the expansion of T_ε, and its coefficient of $\log d$ is at least $\frac{1}{2}$ because $\mu_\alpha < 1$ by (G.3). ■

I Additional experiments

This appendix collects numerical experiments that complement the theory. The first is the experiment behind Figure 2. The next four test whether the collapse-then-split behavior survives changes to the training algorithm, the initialization scale, the width, and the activation. The sixth is a numerical search for stationary points across teacher angles, and the last looks for an analogue of the behavior in networks trained on real data.

I.1 Common setup and measured quantities

Model and data. Except in Appendices I.5 and I.8, we use the two-neuron model of Section 2, with teacher $f^*(x) = \phi(\langle v_1, x\rangle) + \phi(\langle v_2, x\rangle)$ and student $f_W(x) = \phi(\langle w_1, x\rangle) + \phi(\langle w_2, x\rangle)$, where $W = (w_1, w_2) \in (\mathbb{R}^d)^2$ and ϕ is the ReLU unless stated otherwise. The teacher weights are unit vectors at angle $\alpha = \pi/3$. We choose coordinates in which $v_{1,2} = (\cos(\alpha/2), \pm\sin(\alpha/2), 0, \dots, 0)$, so that the bisector is $\bar v = e_1$ and the splitting direction is $v_\Delta = e_2$. The input dimension is $d = 128$ unless stated otherwise. Inputs are drawn as $x \sim \mathcal{N}(0, I_d)$, and each label is the noiseless teacher output $y = f^*(x)$.

Losses and time. The population loss is $L(W) = \frac{1}{2}\mathbb{E}_x[(f_W(x) - f^*(x))^2]$ as in (2.2), and we evaluate it and its gradient in closed form through the Gaussian ReLU kernel (A.1a). In the finite-sample experiments, an update on a minibatch B of samples (x_l, y_l) uses the gradient of the empirical loss $\widehat{L}_B(W) = \frac{1}{2|B|}\sum_{l\in B}(f_W(x_l) - y_l)^2$. For gradient descent or SGD with step size η, we report time as $t = \eta \times$ (number of updates), which makes t comparable to the time of the gradient flow (2.2). Unless stated otherwise, we record the state at checkpoints spaced 0.05 time units apart.

Measured quantities. Write $\bar w_i = w_i/\|w_i\|_2$ for the student directions. For $\alpha = \pi/3$, the symmetric saddle $w_{\text{saddle}} = (r_s\bar v, r_s\bar v)$ has per-neuron norm $r_s = \big((\pi - \alpha/2)\cos(\alpha/2) + \sin(\alpha/2)\big)/\pi \approx 0.881$. We track the following quantities.

- The *bisector alignment* $A(W) = \min_{i=1,2}\langle \bar w_i, \bar v\rangle$ is the smaller of the two cosine similarities between a student weight and the bisector. It equals 1 exactly when both students point along $\bar v$.
- The *mean neuron norm* is $\bar r(W) = (\|w_1\|_2 + \|w_2\|_2)/2$, which equals r_s at the symmetric saddle.
- The *separation*
$$S(W) = \frac{|\langle w_1 - w_2, v_\Delta\rangle|}{\|v_1 - v_2\|_2} = \frac{|\langle w_1 - w_2, v_\Delta\rangle|}{2\sin(\alpha/2)}$$
measures how far the two students have split along the splitting direction. It is 0 when the two students have the same coordinate along v_Δ, as at the symmetric saddle, and it is 1 when (w_1, w_2) equals (v_1, v_2) or (v_2, v_1). The *directional separation* $S_{\text{dir}}(W)$ is defined in the same way with $\bar w_1, \bar w_2$ in place of w_1, w_2, so it ignores the norms.
- The *normalized distance to the saddle*
$$D(W) = \Big(\frac{\|w_1 - r_s\bar v\|_2^2 + \|w_2 - r_s\bar v\|_2^2}{2r_s^2}\Big)^{1/2}$$
is the root-mean-square distance of the two students from $r_s\bar v$, relative to r_s. It is 0 exactly at w_{saddle}.

Events. We locate the phases of a trajectory through *events*, each defined by a condition on the quantities above. An event occurs at time t if its condition holds at five consecutive checkpoints, the first of which is at time t. Requiring five checkpoints prevents a single noisy checkpoint from triggering an event. Each event is searched for only after the preceding event in the sequence has occurred. The conditions differ between experiments and are stated in each subsection.

Runs. Every reported number is the mean over five independent runs. Runs differ in the random seeds for the initialization and, where applicable, for the dataset and the order of the minibatches.

I.2 Escape time and dimension

This experiment produces Figure 2. We take the teacher angles $\alpha \in \{\pi/6, \pi/4, \pi/3, \pi/2\}$ and the dimensions $d \in \{32, 64, 128, 256, 512, 1024, 2048\}$. The initial weights $w_1(0), w_2(0) \sim \mathcal{N}(0, \sigma^2 I_d)$ are independent, with $\sigma = 10^{-4}$, and the students follow the population gradient flow (2.2), computed with the closed-form gradient of Appendix I.1. For each run, we record two times: the escape time T_ε with $\varepsilon = 0.1$, defined before Theorem 4.6, and the global-convergence time T_{global}, the first time at which the distance from W to the set $\{(v_1, v_2), (v_2, v_1)\}$ of global minimizers is at most 0.05. Each marker in the figure shows the mean and one standard deviation over five runs. The dashed lines are fits of the form $a + b \log d$, with slopes b between 0.48 and 0.55, and the panel titles give the gap $T_{\text{global}} - T_\varepsilon$.

I.3 Finite-sample SGD

Setup. For each run, we draw one training set of $n = 100d$ inputs $x_l \sim \mathcal{N}(0, I_d)$ with labels $y_l = f^*(x_l)$ and reuse it throughout training. We train with minibatch SGD with step size $\eta = 2.5 \times 10^{-4}$ and batch size 128, without momentum or weight decay. At the start of each epoch, the training set is shuffled and split into $n/128$ disjoint minibatches, and each update is $w_i \leftarrow w_i - \eta \nabla_{w_i} \widehat{L}_B(W)$ for $i = 1, 2$. The initial weights $w_1(0), w_2(0) \sim \mathcal{N}(0, 10^{-12} I_d)$ are independent, so that $\|w_i(0)\|_2 \approx 10^{-5}$. Each run continues until the students reach a teacher configuration or until $t = 160$. We record A, $\bar{r}$, S, and the population loss L at each checkpoint, and Table 2 reports their means over five runs at four fixed times.

Table 2: Finite-sample SGD at $d = 128$, averaged over five runs.

Statistic	$t = 0$	$t = 0.5$	$t = 10$	$t = 25$
Bisector alignment A	0.00941	0.9907	0.9985	0.9353
Mean neuron norm $\bar{r}$	1.19×10^{-5}	0.3460	0.8816	0.9338
Separation S	1.15×10^{-6}	5.61×10^{-4}	0.0319	0.508
Population loss L	0.8045	0.3192	0.02863	0.01298

At $t = 10$, both students are aligned with the bisector, their mean norm 0.8816 matches the collapsed radius $r_s \approx 0.881$, and their separation is still small, so the trajectory is near the symmetric saddle. By $t = 25$, the separation has reached 0.508 and the loss has dropped further, so the students have split. All five runs subsequently reached a neighborhood of one of the two global minima, at different times.

To check the time scales, we repeat the experiment for $d \in \{32, 128, 512, 2048\}$, with the same step size, batch size, and initialization, and with $n = 100d$ samples for each d. We use three events, in this order. The students are *aligned* when $A(W) \geq 0.95$, *collapsed* when $D(W) \leq 0.05$, and *split* when $S(W) \geq 0.50$. For each run, we record the time from alignment to collapse and the time from collapse to split, and Table 3 reports their means over five runs.

Table 3: Mean elapsed times under finite-sample SGD, averaged over five runs.

Elapsed time	$d = 32$	$d = 128$	$d = 512$	$d = 2048$
Alignment to collapse	1.42	1.47	1.47	1.36
Collapse to split	20.33	24.59	30.45	35.07

Collapse is reached much faster than splitting, and the time from alignment to collapse is nearly independent of d. A least-squares fit of the mean collapse-to-split time against $\log d$, with an intercept and the four values of d, has $R^2 = 0.997$, in line with the logarithmic dependence on d in Theorem 4.6.

I.4 Initialization scale

Setup. We run gradient descent on the population loss, $W \leftarrow W - \eta \nabla L(W)$, with $\eta = 10^{-3}$ and the gradient computed in closed form, up to $t = 150$, that is, for 150,000 steps. No data are sampled. The initial weights are $w_i(0) = \sigma g_i$ with independent $g_1, g_2 \sim \mathcal{N}(0, I_d)$, for the per-coordinate scales $\sigma \in \{10^{-3}, 10^{-2}, 0.05, 0.10, 0.15\}$. Within each run, the same pair (g_1, g_2) is used for every σ, so that only the scale changes. The mean initial neuron norm in Table 4 is the average of $\|w_i(0)\|_2$ over both neurons and all runs, and it is close to $\sigma\sqrt{d}$. A *collapse* occurs when $D(W) \le 0.25$ and $S(W) \le 0.20$, that is, when both students are close to w_{saddle} and have not yet split. A subsequent *split* occurs when $S(W) \ge 0.50$. We say that the collapse is *bypassed* when no collapse occurs before $t = 150$. The final population loss is $L(W)$ at $t = 150$.

Table 4: Dependence on the initialization scale, population gradient descent at $d = 128$, averaged over five runs.

Per-coordinate σ	Mean initial neuron norm	Observed trajectory	Mean final population loss
10^{-3}	0.0112	collapse, then split	1.615×10^{-10}
10^{-2}	0.1118	collapse, then split	1.440×10^{-8}
0.05	0.5589	collapse bypassed	1.791×10^{-7}
0.10	1.1178	collapse bypassed	3.724×10^{-7}
0.15	1.6767	collapse bypassed	5.133×10^{-7}

Collapse occurs at the two smallest scales and is bypassed at the three larger ones, while every run ends near a global minimum (Table 4). This matches the mechanism behind Phase 1. In the directional dynamics of Section 4.1, the teacher force is of order one while the student–student interaction is of order $\|w_{3-i}\|_2$, so the teacher force dominates only while the weights are small, and it aligns both students with the bisector before their norms grow. Once the initial norms are comparable to the saddle radius r_s, the interaction is already significant at the start, and the trajectory can specialize without passing near a collapsed configuration.

The landscape results explain why missing the collapse costs nothing. The collapsed stationary points sit at a definite radius, so an initialization at or above that scale need not come near them. By Theorem 3.3, every stationary point with both student weights nonzero is a global minimum or a saddle, and by Theorem 4.2 gradient flow reaches a global minimum from Lebesgue-almost every initialization.

I.5 Wider networks

We train m students against k teachers with orthonormal weights $v_a = e_a$, using mean-normalized networks

$$f_W(x) = \frac{1}{m} \sum_{i=1}^{m} \phi(\langle w_i, x \rangle), \qquad f_*(x) = \frac{1}{k} \sum_{a=1}^{k} \phi(\langle v_a, x \rangle),$$

and population gradient flow on $\frac{1}{2}\mathbb{E}_x[(f_W(x) - f_*(x))^2]$, with $x \sim \mathcal{N}(0, I_d)$ and $d = 128$, where the loss and its gradient are computed in closed form. We integrate the flow with an adaptive Dormand–Prince 5(4) scheme in double precision, with relative tolerance 10^{-8}, absolute tolerance 10^{-13}, and maximal step 0.1. Because of the factor $1/m$ in f_W, each student moves at a speed of order $1/m$, so we report the rescaled time $\tau = t/m$ to compare different widths. The initial weights $w_i(0) \sim \mathcal{N}(0, r_0^2 I_d/d)$ are independent, with $r_0 = 10^{-5}$. In each run, we draw one $32 \times d$ Gaussian matrix and give the model with m students its first m rows. We record the state every 0.1 units of τ.

To define the measured quantities, let $\bar{w}_i = w_i/\|w_i\|_2$, let $\bar{v} = k^{-1/2} \sum_{a=1}^{k} e_a$ be the normalized teacher mean, let $T = \text{span}\{e_1, \ldots, e_k\}$ be the teacher span, and write P_T and $P_{T^\perp}$ for the orthogonal projections onto T and its orthogonal complement. For each student, $a_i = \langle \bar{w}_i, \bar{v} \rangle$ is its

alignment with $\bar{v}$, and $c_i = P_T \bar{w}_i - a_i \bar{v}$ is its *teacher-contrast component*, the part of its direction that lies in T but is orthogonal to $\bar{v}$. Angle brackets denote averages over the m students, for example $\langle c \rangle = \frac{1}{m} \sum_i c_i$ and $\langle \bar{w} \rangle = \frac{1}{m} \sum_i \bar{w}_i$. We track the following quantities.

- The *directional dispersion* $\frac{1}{m} \sum_i \|\bar{w}_i - \langle \bar{w} \rangle\|_2^2$ is 0 when all students point in the same direction.
- The *contrast spread* $S_k = \left(\frac{1}{m} \sum_i \|c_i - \langle c \rangle\|_2^2 / (1 - 1/k) \right)^{1/2}$ is 0 when all students have the same teacher-contrast component, and it is 1 when each teacher direction e_a is matched by exactly m/k students.
- The *contrast fraction* $C/(C + C_\perp)$, with $C = \frac{1}{m} \sum_i \|c_i - \langle c \rangle\|_2^2$ and $C_\perp = \frac{1}{m} \sum_i \|P_{T^\perp}(\bar{w}_i - \langle \bar{w} \rangle)\|_2^2$, is the share of the dispersion orthogonal to $\bar{v}$ that lies in T.
- The *teacher-span occupancy* $\frac{1}{m} \sum_i \|P_T \bar{w}_i\|_2^2$ is the average squared length of the part of each student direction that lies in T.
- The *effective participation* $(\sum_i E_i)^2 / (m \sum_i E_i^2)$, with $E_i = \|c_i - \langle c \rangle\|_2^2$, lies in $[1/m, 1]$ and is small when only a few students carry the spread.

A *collapse* occurs when $\min_i a_i \geq 0.95$, the directional dispersion is at most 0.075, and $S_k \leq 0.20$. A subsequent *split* occurs when $S_k \geq 0.50$, the contrast fraction is at least 0.8, the teacher-span occupancy is at least 0.8, and the effective participation is at least 0.25. As in Appendix I.1, each event requires five consecutive checkpoints.

Table 5: Wider networks, population gradient flow at $d = 128$, averaged over five runs. Times are in units of $\tau = t/m$.

Students m	Teachers k	Mean collapse time	Mean split time
16	16	0.10	144.02
32	32	0.22	252.84
32	16	0.10	180.22
16	32	0.22	213.94

Collapse followed by a split occurs in all four configurations (Table 5), covering matched, overparameterized, and underparameterized students. This experiment concerns the transient behavior only. In separate long-time runs from the same initializations, integrated to $\tau = 5 \times 10^4$ for $(16, 16)$ and $(32, 32)$, to $\tau = 10^5$ for $(32, 16)$, and to $\tau = 2 \times 10^5$ for $(16, 32)$, we examine the last $1{,}000$ units of τ. The overparameterized configuration $(m, k) = (32, 16)$ reaches numerically zero loss, in the sense that the normalized loss $\mathbb{E}[(f_W - f_*)^2]/\mathbb{E}[f_*^2]$ stays below 10^{-10}. The matched configurations $(16, 16)$ and $(32, 32)$ and the underparameterized configuration $(16, 32)$ instead end numerically stationary with positive loss. Over the last $1{,}000$ units of τ, their root-mean-square velocity $(\frac{1}{m} \sum_i \|\mathrm{d}w_i/\mathrm{d}\tau\|_2^2)^{1/2}$ stays below 10^{-7}, their weights move by at most 10^{-4} in root-mean-square distance, and their normalized loss varies by at most 10^{-9}. Thus, unlike the two-neuron case, global convergence does not hold uniformly at larger width, consistent with the spurious local minima of wider ReLU networks found by Safran and Shamir (2018).

I.6 GELU activations

We replace the ReLU by the exact GELU $\phi(z) = z\Phi(z)$ (Hendrycks and Gimpel, 2016), where Φ is the standard Gaussian distribution function, either in both networks (a realizable model) or only in the student, with a ReLU teacher (a misspecified model).

In each run, we draw one training set of $n = 100d = 12{,}800$ inputs $x_l \sim \mathcal{N}(0, I_d)$ and an independent validation set of the same size, with noiseless labels $y_l = f^*(x_l)$ given by the teacher of that activation pair. We train with minibatch SGD on $\widehat{L}_B$ with step size $\eta = 10^{-3}$ and batch size 128, without momentum or weight decay, reshuffling the training set into disjoint minibatches at each epoch, up to $t = 100$. The initial weights $w_1(0), w_2(0) \sim \mathcal{N}(0, 0.01^2 I_d)$ are independent. Within a run, the three activation pairs share the initialization, the inputs, and the order of the minibatches. For each activation pair, we replace r_s by the radius $r_\phi > 0$ of the symmetric stationary point $(r_\phi \bar{v}, r_\phi \bar{v})$, which solves the radial stationarity equation $\mathbb{E}\big[(f^*(x) - 2\phi(r_\phi \langle \bar{v}, x \rangle))\, \phi'(r_\phi \langle \bar{v}, x \rangle) \langle \bar{v}, x \rangle\big] = 0$. We

solve this equation numerically with Gauss–Hermite quadrature, and $r_\phi = r_s$ for the ReLU pair. A *collapse* occurs when $A(W) \geq 0.95$, $S_{\mathrm{dir}}(W) \leq 0.20$, and $D(W) \leq 0.25$ with r_ϕ in place of r_s. A subsequent *split* occurs when $S_{\mathrm{dir}}(W) \geq 0.50$. The final risk is the mean squared error of the student on the validation set at $t = 100$, divided by the mean of $f^*(x)^2$ on the validation set.

Table 6: Activation pairs, minibatch SGD at $d = 128$, averaged over five runs.

Teacher / student	Mean collapse time	Mean split time	Mean final normalized risk
ReLU / ReLU	2.23	18.87	1.377×10^{-7}
GELU / GELU	1.92	23.27	2.295×10^{-7}
ReLU / GELU	2.16	21.40	1.922×10^{-2}

All three pairs show a collapse followed by a split (Table 6). The nonzero final risk of the ReLU/GELU pair is expected under misspecification. The times should not be compared across activations, since the activation also changes the gradient scale and the radius of the symmetric stationary state.

I.7 Numerical search for stationary points

We search numerically for zeros of the exact population gradient of the two-neuron model at teacher angles $\alpha \in \{90^\circ, 120^\circ, \alpha_\star, 128^\circ, 140^\circ, 150^\circ, 165^\circ\}$, where $\alpha_\star \approx 127.521476^\circ$.

We look for points $W = (w_1, w_2)$ with $w_1, w_2 \neq 0$ and $\nabla L(W) = 0$, where L is the population loss of the two-neuron ReLU model and the gradient is computed in closed form. By Proposition 3.5, every such point can be rotated so that both weights lie in the span of the teacher plane and one transverse direction. We therefore represent each student weight by three coordinates, two in the teacher plane and one transverse, which gives six unknowns. For each angle, we solve $\nabla L(W) = 0$ by nonlinear least squares, using SciPy's trust-region reflective method with all tolerances set to 10^{-12}, from 750 starting points. These are 300 random starts, with directions uniform on the sphere and norms log-uniform in $[0.04, 2.5]$, together with 300 starts in the teacher plane whose norms solve the two radial stationarity equations and 150 starts in the teacher plane with nearly collinear students. We also use the closed-form families of Theorem 3.3 as starting points. A solution is kept if every component of $\nabla L(W)$ is at most 5×10^{-14} in absolute value and $5 \times 10^{-9} < \|w_i\|_2 \leq 20$ for both students. In the recorded run, every retained point has every component of the full gradient at most 5.6×10^{-17} in absolute value. We count retained points up to the symmetries of the problem, namely exchanging the two students and reflecting the transverse coordinate or the v_Δ-coordinate. A point is *listed* if its dimensionless distance to one of the families of Theorem 3.3 is at most 10^{-10}, where distances are measured relative to $\max\{1, \|w_1\|_2, \|w_2\|_2\}$.

Table 7: Numerically found stationary families, counted up to symmetry. "Listed" families are those of Theorem 3.3, and "other" families are not.

α	Listed	Other	Types of the other families
90°	5	0	–
120°	5	0	–
$127.521476^\circ = \alpha_\star$	5	0	–
128°	5	2	same-direction tilted, in-plane noncollinear
140°	5	2	same-direction tilted, in-plane noncollinear
150°	5	3	same-direction tilted, in-plane noncollinear, nonplanar ($d \geq 3$)
165°	5	3	same-direction tilted, in-plane noncollinear, nonplanar ($d \geq 3$)

Up to $\alpha_\star$, every stationary point found belongs to a family listed in Theorem 3.3, in agreement with Theorem C.8. Above $\alpha_\star$, the search finds additional families (Table 7), whose distance to every listed family is at least 7×10^{-2} in the same dimensionless distance, far above the classification tolerance 10^{-10}. This is consistent with Proposition C.7. A finite multistart search is numerical evidence and does not prove completeness.

I.8 Networks trained on real data

Finally, we ask whether an analogue of collapse-then-split appears in networks trained on real data.

We use the standard splits of MNIST (LeCun et al., 1998), with 60,000 training and 10,000 test images, standard normalization, and no augmentation. For CIFAR-10 (Krizhevsky, 2009), we use 50,000 training and 10,000 test images with channelwise normalization, random crops, and horizontal flips. The MNIST model is a $784 \to 128 \to 10$ MLP with a ReLU hidden layer. The CIFAR-10 model has two convolution–ReLU–max-pool blocks with 32 and 64 channels, followed by a 128-dimensional ReLU hidden layer and a linear classifier, with no normalization layers. Both models are trained by minibatch SGD with momentum 0.9 and batch size 128 on the cross-entropy loss. For MNIST, we train for 20 epochs with learning rate 0.05, without weight decay or a learning-rate schedule. For CIFAR-10, we train for 160 epochs with learning rate 0.1, weight decay 5×10^{-4}, one epoch of linear warmup, and cosine decay. The weights of the feature extractor, that is, of all layers before the classifier, are drawn from the Kaiming initialization (He et al., 2015) and multiplied by 0.03 for MNIST and 0.1 for CIFAR-10, while the classifier keeps its Kaiming scale. Test accuracy is measured on the test split at the end of training.

We track the same object as the theory, the incoming weight direction of each hidden neuron. For a layer with incoming weight matrix $W \in \mathbb{R}^{m \times d}$ with rows $w_i^\top$, let $q_i = w_i / \|w_i\|_2$ and let $Q \in \mathbb{R}^{m \times d}$ have rows $q_i^\top$. For a convolutional layer, each filter is flattened into one row. With λ_j the eigenvalues of the Gram matrix $G = QQ^\top$, whose entries are the cosine similarities $\langle q_i, q_j \rangle$, and $p_j = \lambda_j / \sum_l \lambda_l$, we report the effective rank

$$\operatorname{erank}(G) = \exp\Big(-\sum_j p_j \log p_j\Big) \in [1, \min(m, d)].$$

The effective rank is the exponential of the Shannon entropy of the normalized spectrum. It equals 1 when all rows of Q are parallel and equals the number of rows when the rows are orthonormal. Biases are excluded, and all rows remain nonzero during training. A decrease means that the incoming weight directions concentrate in a lower-dimensional subspace, and a later increase means that they spread out again.

Table 8: Effective rank of the hidden-layer weight directions, averaged over five runs.

Dataset and model	Tracked weights	Initial → early minimum → final	Test accuracy
MNIST, MLP	hidden, 128×784	$117.94 \to 15.86 \to 54.83$	98.17%
CIFAR-10, CNN	final hidden, 128×4096	$126.04 \to 53.93 \to 104.54$	85.22%

Table 9: Effective rank along one representative run of each network. Step N is the state after N SGD updates.

MNIST	Step 0	Step 1	Step 4	Step 16	Step 64	Step 256	Epoch 1	Epoch 20
Hidden layer	117.92	110.70	27.36	**15.48**	20.18	28.35	32.59	55.28
CIFAR-10	Step 0	Step 64	Step 128	Step 256	Epoch 1	Epoch 10	Epoch 40	Epoch 160
Final hidden layer	126.02	123.55	62.25	**53.68**	66.85	94.82	98.98	105.06

In both networks, the effective rank of the hidden-layer weight directions falls sharply early in training and then partially recovers, while the networks reach ordinary test accuracy (Tables 8 and 9). In the representative CIFAR-10 run, the two convolutional layers show the same ordering more weakly, with effective ranks $18.16 \to 8.28 \to 11.93$ in the first layer and $57.55 \to 21.84 \to 39.83$ in the second. As in the synthetic experiment of Appendix I.4, the effect weakens as the initialization scale grows.

The same low-rank-then-recovery pattern has been reported in other settings, with different weight matrices and rank surrogates. In deep matrix factorization trained by gradient descent, the effective rank of the end-to-end matrix first falls to about one, stays on a rank-one plateau, and then rises through successive plateaus toward the rank of the target (Chou et al., 2024). For a convolutional network trained on CIFAR-100 with momentum SGD, the entropy effective rank of a convolutional layer falls rapidly, reaches a minimum at an intermediate epoch, and then increases (Huh et al., 2023). In LLaMA-style pretraining, some full-rank baselines show an early decrease of the effective and stable ranks followed by a partial recovery (Shivagunde et al., 2026). The pattern depends on the

layer, architecture, optimizer, and rank measure, and it is not universal (Yunis et al., 2024; Galanti et al., 2025).

These experiments show an analogue of collapse-then-split in weight space, but they do not show neuron-level specialization. In the two-neuron model, the collapsed state is the symmetric saddle of Theorem 3.3, and the plateau is the escape time from it, governed by the leap-2 structure of the target. These identifications and the resulting time scale are specific to our model, and we do not claim that the networks above pass near the same saddle or follow the same time scale.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately describe the paper's setting, assumptions, and two main contributions: the landscape classification and the small-initialization gradient-flow analysis. The formal theorems and appendices support the same claims, so the stated scope matches the results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in the conclusion, in a paragraph on scope and limitations, which notes that the landscape and convergence results hold only for two neurons, that wider ReLU networks can have spurious local minima, that the dimension reduction does not extend to three or more teacher neurons, and that an end-to-end escape-time analysis from random initialization with high probability remains open. It also makes clear that the analysis is for a population gradient-flow model with isotropic Gaussian inputs and a two-neuron teacher-student architecture.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper is primarily theoretical, and the assumptions are stated in the setup and in the individual theorem or lemma statements. The main results are cross-referenced to complete proofs in the appendices, with supporting lemmas and formulas numbered where needed.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The numerical experiments are supporting evidence for the theoretical results, and the paper provides sufficient details to reproduce them, including the ODE, parameter choices, initialization scheme, number of seeds, integrator, and stopping rules. The additional experiments in Appendix I state their models, optimizers, hyperparameters, initialization scales, event criteria, and numbers of runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper does not provide open-access code. The synthetic experiments use data generated from the stated models, the real-data experiments use the public MNIST and CIFAR-10 benchmarks, and the paper provides the mathematical and experimental details needed to reproduce the reported numerical results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The main-text experiment is a population gradient-flow simulation with synthetic initializations, so it has no data splits or test sets. The finite-sample experiments in Appendix I draw a Gaussian dataset for each run, and the MNIST and CIFAR-10 experiments use the standard training and test splits. The paper specifies the relevant parameters, initialization scheme, and stopping criteria needed to understand the reported results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The reported figure shows means over five random seeds with error bars equal to one standard deviation. The source of variability is the random initialization. The tables in Appendix I report means over five independent runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The paper does not report detailed compute resources such as worker type, memory use, or wall-clock time. The experiments are lightweight population ODE simulations, but precise compute-resource information is not included.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

Justification: The research is theoretical and uses synthetic Gaussian-input models and simulations, together with the public MNIST and CIFAR-10 benchmarks. It does not involve human subjects, private data, scraped data, or deployment of a high-risk model.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: The work is foundational mathematical analysis of a simplified teacher-student model and does not introduce a deployable system, dataset, or model. We do not identify a direct path to societal harms or benefits beyond broad progress in understanding optimization dynamics.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release pretrained models, generative models, scraped datasets, or other assets with a high risk of misuse. The numerical results are based on synthetic simulations of the stated mathematical model.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The only external assets are the MNIST and CIFAR-10 datasets, used in Appendix I.8, whose creators are cited (LeCun et al., 1998; Krizhevsky, 2009). No pretrained models are used. Existing scientific works are cited where relevant.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release a new dataset, model, benchmark, or code package. The numerical results are generated from synthetic simulations described in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing, user studies, annotation work, or other research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human subjects, human-derived data, or crowdsourced labor. IRB or equivalent human-subjects review is therefore not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs were used to brainstorm possible proof ideas and to assist with writing and editing. All theorem statements, proofs, calculations, and numerical results were checked and validated by the human authors; LLM outputs were not used as an automated component of the method or as evidence.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.