Gradient Flow Dynamics of Teacher-Student Distillation with the Squared Loss

Berkan Ottlik[1,2]

Flatiron Institute[1], Columbia University[2]

We study a teacher-student learning setup, where a "student" one layer neural network tries to approximate a fixed "teacher" one layer neural network. We analyze the population gradient flow dynamics in the previously unstudied setting with exactly and under-parameterization, even Hermite polynomial activation functions, and squared loss. In the toy model with 2 teacher neurons and 2 student neurons, we fully characterize all critical points. We identify "tight-balance" critical points which are frequently encountered in simulation and greatly slow down training. We prove that with favorable initialization, we avoid tight-balance critical points and converge to the global optimum. We extend tight-balance critical points and favorable initializations to the multi-neuron exact and under-parameterized regimes. Additionally, we compare dynamics under the squared loss to the simpler correlation loss and describe the loss landscape in the multi-neuron exact and under-parameterized regimes. Finally, we discuss potential implications our work could have for training neural networks with even activation functions.

This is a manuscript of my work during my summer internship at the Flatiron Institute. I was mainly advised by Berfin Şimşek and helped by Denny Wu. This writing reflects my own thoughts and not necessarily those of Berfin and Denny.

Contents

1 Introduction

We have two main motivations for pursing this work: (1) We are fascinated by neural network (NN) dynamics. This fascination stems from an intrinsic curiosity about complex systems and similarities to physical models such as the Ising model[1]. NNs are particularly well-suited for study since, unlike many complex systems, we have full information about individual neurons and how they update. (2) We are optimistic that mathematically analyzing gradient-based training dynamics of NNs in key understudied regimes can make progress towards key NN research themes and ultimately make NNs more controllable and efficient [Belkin (2023)]. Our work is most directly related to the research themes of loss and activation function choices, optimization, distillation, and superposition of features.

In this paragraph, I informally compare and contrast our research methodology to other sciences. Please note this informal treatment does not do justice to the rich field of metascience. Natural science research (and some deep learning theory) is typically theory driven. Researchers run experiments to find phenomena existing theories cannot explain and use experimental results to develop new theories. Machine learning research is typically methods driven. Researchers have hypotheses about what could provably or empirically improve model performance, that they test by proving theorems or testing on a validation dataset [Wolpert (1995); Haussler and Pitt (1989)]. Our research is phenomena driven. We aim to understand the phenomena of a simple model and generalize the phenomena as much as possible. We hope our results are useful for theory and methods driven research in NNs.

We specifically focus on the under-parameterized teacher-student learning setup, where a "student" one layer NN tries to approximate a fixed "teacher" one layer NN. This setting allows us to directly study distillation and feature superposition. Additionally, it allows us to study learning any function that can be parameterized with a neural network.

[1] Our loss and the Ising model Hamiltonian are similar in that they both have a repulsive and an attractive force. Perhaps the biggest difference is the Ising model has binary valued spins while our weight vectors are real valued.

1.1 Related works

Many works have studied gradient flow trajectories in the teacher-student setup. Saad and Solla (1995) derive the gradient flow differential equations, yet rely on numerical integration to understand dynamics. Du and Lee (2018) gives convergence rates in the over-parameterized regime with the squared loss but only considers one student neuron. Martin et al. (2024) considers over exact and under-parameterized regimes with the squared loss and multiple student and teacher neurons, but Martin et al. (2024) only considers the quadratic activation function. Simsek et al. (2023, 2024) studies the under-parameterized setting with multiple student and teacher neurons, but they relies on the correlation loss to decouple student neuron dynamics, potentially concealing important phenomena.

The under-parameterized teacher-student setup is directly related to distillation. In fact, it is equivalent to knowledge distillation in binary classification NNs [Hinton et al. (2015)]. Existing theories of NN distillation require the linear representation hypotheses [Boix-Adsera (2024)]. Additionally, prior works on the superposition of features study empirically study and interpret different under-parameterized models [Elhage et al. (2022)].

1.2 Our model

We study the teacher-student setup and focus on the under-parameterized regime, since it is understudied and relevant to distillation, superposition, and large language model training. To be as realistic as possible, we consider the squared loss and multiple teacher and student neurons. To keep our analysis tractable, we consider the population spherical gradient flow dynamics with even Hermite polynomial activation functions for both the teacher and student and orthonormal teacher vectors. Even with these assumptions, it is difficult to get a precise understanding of the critical points, let alone the dynamics. Thus, we began by studying a toy model with 2 teacher neurons and 2 student neurons. Then, we generalized the results to the exact and under-parameterized settings.

We now formally introduce our model. We consider the following (typically non-convex) optimization problem

$$L^{n,k}(\{w_i\}_{i=1}^{n}) = \frac{1}{2} \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \Big[\Big(\sum_{i=1}^{n} H_\rho(w_i^\top x) - \sum_{j=1}^{k} H_\rho(v_j^\top x) \Big)^2 \Big],$$

where $\mathcal{D} = \mathcal{N}(0, I_d)$, $w_i, v_j \in \mathbb{S}^{d-1}$, and H_ρ is the ρth normalized Hermite polynomial. We can expand out the square and use the linearity of expectation to rewrite our objective as

$$= \frac{1}{2} \sum_{i=1}^{n} \sum_{l\neq i}^{n} g_\rho(w_i^\top x, w_l^\top x) - \sum_{i=1}^{n} \sum_{j=1}^{k} g_\rho(w_i^\top x, v_j^\top x) + \frac{1}{2} \sum_{j=1}^{k} \sum_{j'=1}^{k} g_\rho(v_j^\top x, v_{j'}^\top x),$$

where $g_\rho(a, b) = \mathbb{E}[H_\rho(a) H_\rho(b)]$. Since w_i and v_j are unit vectors, the correlation between $w_i^\top x$ and $v_j^\top x$ is $w_i^\top v_j$. By Proposition 11.31 in O'Donnell (2021), $g_\rho(w_i^\top x, v_j^\top x) = (w_i^\top v_j)^\rho$. The first and last terms are constants in terms of $w_i \in \mathbb{S}^{d-1}$, because of the spherical constraint and independence of w_i respectively. We can rewrite our objective as

$$= \frac{1}{2} \sum_{i=1}^{n} \sum_{q=1}^{n} (w_i^\top w_q)^\rho - \sum_{i=1}^{n} \sum_{j=1}^{k} (w_i^\top v_j)^\rho + C = \sum_{i=1}^{n} \sum_{i'>i}^{n} (w_i^\top w_{i'})^\rho - \sum_{i=1}^{n} \sum_{j=1}^{k} (w_i^\top v_j)^\rho + C.$$

Setting

$$W = \begin{bmatrix} — & w_1^\top & — \\ & \vdots & \\ — & w_n^\top & — \end{bmatrix} \in \mathbb{R}^{n \times d} \text{ and } V = \begin{bmatrix} — & v_1^\top & — \\ & \vdots & \\ — & v_k^\top & — \end{bmatrix} \in \mathbb{R}^{k \times d}$$

and denoting $(\cdot)^{\circ \rho}$ as the Hadamard power that raises each element of its input to the ρth power, we can rewrite the objective as

$$= \frac{1}{2}\, \mathbb{1}_n^\top (WW^\top)^{\circ \rho}\, \mathbb{1}_n - \mathbb{1}_n^\top (WV^\top)^{\circ \rho}\, \mathbb{1}_k + C,$$

where $\mathbb{1}_n \in \{1\}^n$. We refer to the first term as the student repulsive term, as it encourages student vectors to become orthogonal. We refer to the second term as the teacher attractive term, as it encourages student vectors to align with teacher vectors. This model is identical, up to constant factors, to that of Martin et al. (2024) when the activation function is the second normalized Hermite polynomial. See appendix A for more details.

1.3 Contributions

In section 2, we conduct a full critical point analysis of the toy model with two student and two teacher neurons. We also briefly describe the critical points in more general cases. In section 3, we discuss dynamics, and prove that we avoid "tight-balance" critical points from favorable initializations. We generalize these initializations to the multi teacher and student neuron exact and under-parameterized setting. In section 4, we connect our results to machine learning practice, in particular to activation function choices.

Full proofs are in the appendix.

2 Critical point analysis

2.1 Derivative

We first compute the derivative of our objective with respect to $w_{\alpha,\beta}$,

$$\frac{\partial L^{n,k}(W)}{\partial w_{\alpha,\beta}} = \rho w_{:,\beta}^\top (W w_\alpha)^{\circ \rho - 1} - \rho v_{:,\beta}^\top (V w_\alpha)^{\circ \rho - 1}.$$

Therefore,

$$\frac{\partial L^{n,k}(W)}{\partial W} = \rho (WW^\top)^{\circ \rho - 1} W - \rho (WV^\top)^{\circ \rho - 1} V.$$

We now consider the spherical derivative. Specifically, we restrict the rows of the derivative to be on the sphere by projecting each row onto the tangent space $T_{w_\alpha}(\mathbb{S}^{d-1})$ of the sphere

$$\begin{aligned} \overline{\frac{\partial L^{n,k}(W)}{\partial W}} &= \begin{bmatrix} — & P_{w_1}\left(\frac{\partial L^{n,k}(W)}{\partial w_1}\right) & — \\ & \vdots & \\ — & P_{w_n}\left(\frac{\partial L^{n,k}(W)}{\partial w_n}\right) & — \end{bmatrix} \\ &= \rho (WW^\top)^{\circ \rho - 1} W - \rho (WV^\top)^{\circ \rho - 1} V \\ &\quad - \left(\left[\rho (WW^\top)^{\circ \rho - 1} W - \rho (WV^\top)^{\circ \rho - 1} V \right] W^\top \circ I_n \right) W. \end{aligned}$$

Let $A = \rho(WW^\top)^{\circ\rho-1}W - \rho(WV^\top)^{\circ\rho-1}V$, then

$$\overline{\frac{\partial L^{n,k}(W)}{\partial W}} = A - (AW^\top \circ I_n)W.$$

2.2 Finding general critical points

We set the derivative to 0. Thus, every W that satisfies the following equation is a critical point,

$$B - (BW^\top \circ I_n)W = 0,$$

where $B = A/\rho$. We first identify general critical points, and then identify critical points in the toy model.

2.2.1 Orthogonal copy critical points

Theorem 2.1 (Orthogonal copy critical points)**.** *Every arrangement of student vectors that satisfies the following conditions is a critical point:*

- *for each $i \in [n]$, $w_i = v_j$ for some v_j,*
- *for all $i \in [n]$ and $j \in [k]$, $w_i^\top v_j \in \{-1, 0, 1\}$.*

Proof Sketch. Consider the case where ρ is odd. Then the ith row of B is

$$B_i = w_i \sum_{j=1}^{n} w_j^\top w_i - w_i \sum_{j=1}^{k} v_j^\top w_i.$$

Since $\|w_i\|_2^2 = 1$, the ith row of the spherical gradient simplifies to 0. This completes the proof for the odd case.

Consider the case where ρ is even. Then the ith row of B is

$$B_i = w_i \sum_{j=1}^{n} |w_j^\top w_i| - w_i \sum_{j=1}^{k} |v_j^\top w_i|.$$

Since $\|w_i\|_2^2 = 1$, the ith row of the spherical gradient simplifies to 0. This completes the proof for the even case. This completes the entire proof. □

2.2.2 Euclidean gradient can never be 0

We want to show that the Euclidean gradient can never be 0, i.e. $B \neq 0$.

Simple case $\rho = 2$. In the simple case $\Phi(W) = W$. A necessary condition for $W^\top W = V^\top V$, is

$$\text{Tr}(W^\top W) = \text{Tr}(V^\top V).$$

Observe

$$\text{Tr}(W^\top W) = \text{Tr}(WW^\top) = \|w_1\|^2 + \ldots + \|w_n\|^2 = n,$$

and similarly $\text{Tr}(V^\top V) = k$. Therefore it is not true that $W^\top W = V^\top V$.

Even ρ case.

Lemma 2.2 (Euclidean gradient cannot be 0)**.** *If the teacher vectors are orthogonal, ρ is even, $\rho > 0$, and $n < k$, there exist no configuration of student neurons such that $B = 0$.*

Proof Sketch. We aim to show $B \neq 0$. Observe a sufficient condition is that there exists a $i \in [n]$ such that the squared ℓ^2 norm of the ith row of $(WW^\top)^{\circ\rho-1}W$ is not equal to the squared ℓ^2 norm of the ith row of $(WV^\top)^{\circ\rho-1}V$.

We compute the squared ℓ^2 norm of the ith row of $(WV^\top)^{\circ\rho-1}V$,

$$\|((WV^\top)^{\circ\rho-1}V)_i\|_2^2 = \|\sum_{j=1}^{k}(v_j^\top w_i)^{\rho-1}v_j\|_2^2 = \sum_{j=1}^{k}(v_j^\top w_i)^{2\rho-2}.$$

By lemma B.1, we know this quantity is at most 1, with equality only when $w_i = v_j$ for some $j \in [k]$.

We compute the squared ℓ^2 norm of the ith row of $(WW^\top)^{\circ\rho-1}W$,

$$\|((WW^\top)^{\circ\rho-1}W)_i\|_2^2 = 2\sum_{j\neq i}^{n}(w_j^\top w_i)^{\rho} + 2\sum_{\substack{j'>j \\ j',j\neq i}}^{n}(w_j^\top w_i)^{\rho-1}(w_{j'}^\top w_i)^{\rho-1}w_{j'}^\top w_j + 1 + \sum_{j\neq i}^{n}(w_j^\top w_i)^{2\rho-2},$$

and show this quantity is at least 1, therefore showing that the norms can never be equal and proving the original statement. This completes the proof. □

2.2.3 Scaled rows critical points

Recall our critical point condition is

$$B - (BW^\top \circ I_n)W = 0.$$

This is equivalent to the condition for all $i \in [n]$, $B_i = (B_i^\top w_i)w_i$. Observe it suffices to say B_i is some scalar multiple of w_i since if there exists some constant $\beta_i \in \mathbb{R}$ such that $B_i = \beta_i w_i$,

$$\begin{aligned}\beta_i w_i &= (\beta_i w_i^\top w_i)w_i \\ \beta_i w_i &= \beta_i w_i,\end{aligned}$$

implying $\beta_i = B_i^\top w_i$.

Expanding out B_i, we have

$$\begin{aligned}B_i &= W^\top(Ww_i)^{\circ\rho-1} - V^\top(Vw_i)^{\circ\rho-1} \\ &= w_i + \sum_{j\neq i}(w_i^\top w_j)^{\rho-1}w_j - \sum_{j=1}^{k}(w_i^\top v_j)^{\rho-1}v_j.\end{aligned}$$

Student in span of teachers. We first consider the case where $w_i \in \operatorname{span}(\{v_1, \dots, v_k\})$. Since the teacher vectors are unit norm and orthogonal, there exists some $\alpha_i \in \mathbb{S}^{k-1}$ such that $w_i = \alpha_{i,1}v_1 + \cdots + \alpha_{i,k}v_k$. We must now show that there exists some $\beta_i \in \mathbb{R}$ such that

$$\sum_{j\neq i}^{n}(\alpha_i^\top \alpha_j)^{\rho-1}(\alpha_{j,1}v_1 + \cdots + \alpha_{j,k}v_k) - \sum_{j=1}^{k}\alpha_{i,j}^{\rho-1}v_j = (\beta_i - 1)(\alpha_{i,1}v_1 + \cdots + \alpha_{i,k}v_k).$$

Since the teacher vectors are orthogonal, an equivalent condition is for all $i \in [n]$ and for all $l \in [k]$,[2]

$$\sum_{j \neq i}^{n} (\alpha_i^\top \alpha_j)^{\rho-1} \alpha_{j,l} - \alpha_{i,l}^{\rho-1} = (\beta_i - 1)\alpha_{i,l}.$$

We can multiply each side of the expression above by $\alpha_{i,l}$ for all $l \in [k]$,

$$\sum_{j \neq i}^{n} (\alpha_i^\top \alpha_j)^{\rho-1} \alpha_{i,l}\alpha_{j,l} - \alpha_{i,l}^{\rho} = (\beta_i - 1)\alpha_{i,l}^2.$$

Summing these equations over all $l \in [k]$ for a fixed i,

$$\begin{aligned} \sum_{j \neq i}^{n} (\alpha_i^\top \alpha_j)^{\rho-1} \sum_{l=1}^{k} \alpha_{i,l}\alpha_{j,l} - \sum_{l=1}^{k} \alpha_{i,l}^{\rho} = (\beta_i - 1) \sum_{l=1}^{k} \alpha_{i,l}^2 \\ \sum_{j \neq i}^{n} (\alpha_i^\top \alpha_j)^{\rho} - \|\alpha_i\|_\rho^\rho = \beta_i - 1. \end{aligned}$$

Solving for β_i,

$$\beta_i = 1 + \sum_{j \neq i}^{n} (\alpha_i^\top \alpha_j)^{\rho} - \|\alpha_i\|_\rho^\rho.$$

Substituting β_i to our prior equation,

$$\sum_{j \neq i}^{n} (\alpha_i^\top \alpha_j)^{\rho-1} \alpha_{j,l} - \alpha_{i,l}^{\rho-1} = (\sum_{j \neq i}^{n} (\alpha_i^\top \alpha_j)^{\rho} - \|\alpha_i\|_\rho^\rho)\alpha_{i,l}.$$

Writing the equalities in a vectorized form, we have an equivalent critical points condition that is more interpretable,

$$\begin{aligned} \sum_{j \neq i}^{n} (\alpha_i^\top \alpha_j)^{\rho-1} \alpha_j - \alpha_i^{\circ \rho-1} &= \sum_{j \neq i}^{n} (\alpha_i^\top \alpha_j)^{\rho} \alpha_i - \|\alpha_i\|_\rho^\rho \alpha_i \\ \sum_{j \neq i}^{n} (\alpha_i^\top \alpha_j)^{\rho-1} \Big(\alpha_j - (\alpha_i^\top \alpha_j)\alpha_i\Big) &= \Big(\alpha_i^{\circ \rho-1} - \|\alpha_i\|_\rho^\rho \alpha_i\Big) \\ \sum_{j \neq i}^{n} (\alpha_i^\top \alpha_j)^{\rho-1} \Big(\alpha_j - (\alpha_i^\top \alpha_j)\alpha_i\Big) &= \Big(\alpha_i^{\circ \rho-1} - (\alpha_i^{\circ \rho-1 \top} \alpha_i)\alpha_i\Big). \end{aligned}$$

Observe for any vectors $u, v \in \mathbb{R}^d$, $u - (u^\top v)v$ is the orthogonal projection of u onto the orthogonal complement of v.

2.3 Toy model critical points

In general, it is still difficult to find critical points from the simplified form in the previous section. Therefore we consider a simplified case where $n = k = d = 2$ and $\rho > 2$ and ρ is even. In this case,

[2] If the student vectors are orthogonal, the condition simplifies to $-\alpha_{i,l}^{\rho-1} = (\beta_i - 1)\alpha_{i,l}$, which clearly holds if $\alpha_{i,l} \in \{-1, 0, 1\}$.

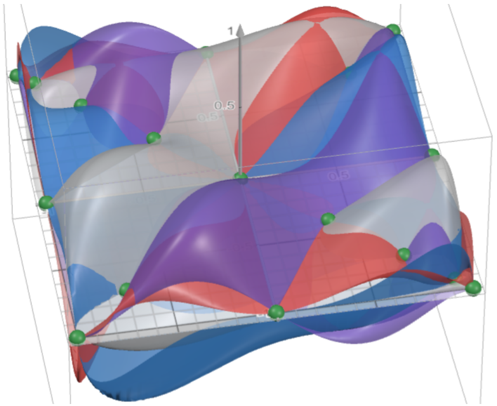


Figure 1: Critical points condition functions when $n = k = d = 2$, $\rho = 4$. All critical points are labeled with green dots.

Name	**Condition**
Copy	$w_1 = v_1$ and $w_2 = v_2$
Tight-balance	$w_1^\top v_1 = w_2^\top v_1 = \cos^{-1}(\pi/8)$ and $w_1^\top v_2 = -w_2^\top v_2 = \sin^{-1}(\pi/8)$
Balance	$w_1^\top v_1 = w_2^\top v_1 = 2^{-1/2}$ and $w_1^\top v_1 = -w_2^\top v_1 = 2^{-1/2}$
Same copy	$w_1 = w_2 = v_1$
Same balance	$w_1 = w_2$, $w_1^\top v_1 = 2^{-1/2}$, and $w_1^\top v_2 = 2^{-1/2}$

Table 1: Full description of the critical points found from figure 1 up to sign and teacher symmetries.

W is a critical point iff it satisfies

$$(\alpha_1^\top \alpha_2)^{\rho-1}\Big(\alpha_2 - (\alpha_1^\top \alpha_2)\alpha_1\Big) = \alpha_1^{\circ\rho-1} - (\alpha_1^{\circ\rho-1\top}\alpha_1)\alpha_1$$
$$(\alpha_2^\top \alpha_1)^{\rho-1}\Big(\alpha_1 - (\alpha_2^\top \alpha_1)\alpha_2\Big) = \alpha_2^{\circ\rho-1} - (\alpha_2^{\circ\rho-1\top}\alpha_2)\alpha_2,$$

where $\alpha_i = (w_i^\top v_1, w_i^\top v_2)$. We describe all ways this is possible,

Recall $\|\alpha_i\|_2 = 1$. Therefore, we can rewrite the conditions above in terms of α_{11} and α_{21} in non-vectorized form. For simplicity we denote $a = \alpha_{11}$, $b = \alpha_{21}$, and $c = ab + \sqrt{(1-a^2)(1-b^2)}$,

$$c^{\rho-1}\Big(b - ca\Big) - a^{\rho-1} + (a^\rho + \sqrt{1-a^2}^{\rho})a = 0$$
$$c^{\rho-1}\Big(\sqrt{1-b^2} - c\sqrt{1-a^2}\Big) - \sqrt{1-a^2}^{\rho-1} + (a^\rho + \sqrt{1-a^2}^{\rho})\sqrt{1-a^2} = 0$$
$$c^{\rho-1}\Big(a - cb\Big) - b^{\rho-1} + (b^\rho + \sqrt{1-b^2}^{\rho})b = 0$$
$$c^{\rho-1}\Big(\sqrt{1-a^2} - c\sqrt{1-b^2}\Big) - \sqrt{1-b^2}^{\rho-1} + (b^\rho + \sqrt{1-b^2}^{\rho})\sqrt{1-b^2} = 0.$$

Note it is difficult to find the zeros of these functions. Therefore, we plot these functions for $\rho = 4$ in figure 1 and by inspection find a full description of the critical points (all green dots in figure 1) in table 1.

When we vary ρ, each critical point stays the same except the tight-balance critical point. Specifically, the angle of the student vectors to the closer teacher decreases as ρ increases. We show how to find this angle θ_ρ.

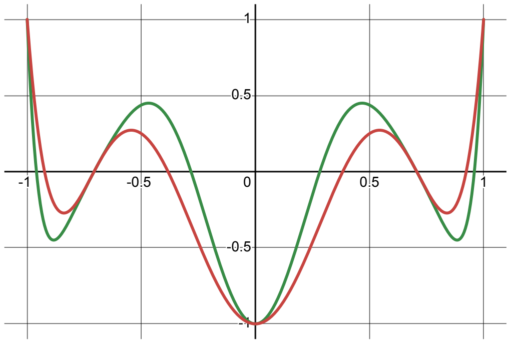


Figure 2: Tight-balance critical point condition when $n = k = d = 2$. The red line corresponds to $\rho = 4$ and the green line corresponds to $\rho = 6$.

Observe in the tight-balance critical point, $a = -b$ and consequentially $\sqrt{1-a^2} = \sqrt{1-b^2}$. In this case, $c = -a^2 + \sqrt{(1-a^2)^2} = 1 - 2a^2$ and we can simplify our critical point conditions to

$$(1-2a^2)^{\rho-1}\Big(-a-(1-2a^2)a\Big) - a^{\rho-1} + (a^\rho + \sqrt{1-a^2}^{\rho})a = 0$$

$$(1-2a^2)^{\rho-1}\Big(\sqrt{1-a^2}-(1-2a^2)\sqrt{1-a^2}\Big) - \sqrt{1-a^2}^{\rho-1} + (a^\rho + \sqrt{1-a^2}^{\rho})\sqrt{1-a^2} = 0.$$

We can further simplify to

$$(1-2a^2)^{\rho-1}(-2a+2a^3) - a^{\rho-1} + a^{\rho+1} - a\sqrt{1-a^2}^{\rho} = 0$$

$$2a^2\sqrt{1-a^2}(1-2a^2)^{\rho-1} - \sqrt{1-a^2}^{\rho-1} + a^\rho\sqrt{1-a^2} - \sqrt{1-a^2}^{\rho+1} = 0.$$

Recall that at a tight-balance critical point, $a \neq 0$, therefore we can simplify the conditions to

$$-2(1-2a^2)^{\rho-1} - a^{\rho-2} + 2a^2(1-2a^2)^{\rho-1} + a^\rho - \sqrt{1-a^2}^{\rho} = 0$$

$$-\sqrt{1-a^2}^{\rho-2} + 2a^2(1-2a^2)^{\rho-1} + a^\rho - \sqrt{1-a^2}^{\rho} = 0$$

By subtracting the second equation from the first, we get a more general condition

$$-2(1-2a^2)^{\rho-1} - a^{\rho-2} + \sqrt{1-a^2}^{\rho-2} = 0.$$

Unfortunately, this condition is also difficult to solve analytically. We visualize the function in figure 2. As ρ varies between different even values larger than 2, the condition maintains the 0 at $2^{-1/2}$, but the other zeroes vary. We empirically solve for these zeros and plot them as a function of ρ in figure 3. For $\rho = 4$, $\theta_\rho = \pi/8$, and plugging this back into the conditions above verifies that it is a critical point. For larger ρ, we did not find a simple analytic form, yet empirically they satisfy the conditions above.

2.4 Generalizing tight-balance critical points

We generalize tight-balance critical points to the exact and under-parameterized setting while keeping all other assumptions. In the simplest case, a pair of student vectors is orthogonal to all other student vectors and lies in the span of two teacher vectors. This decouples the dynamics of the pair of student vectors from all other student vectors and reduces back to the toy model where $n = k = d = 2$. Thus, the pair can form a tight-balance critical point. If the pair of student vectors is not orthogonal to all other student vectors or the pair of student vectors does not lie in the span of two teacher vectors, the dynamics will no longer decouple to the toy model where $n = k = d = 2$.

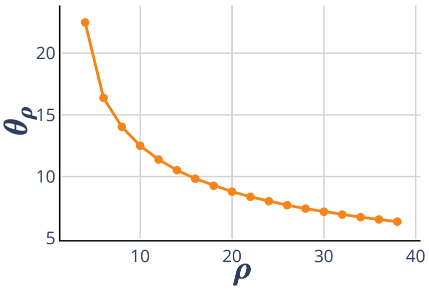


Figure 3: Tight-balance critical point angle as a function of ρ.

2.5 Minima

We can write polynomial programs for finding the optimal configuration of neurons. However, these programs are in general not convex and difficult to solve. Fortunately, identifying the global minimum with orthogonal teacher vectors is simple.

Theorem 2.3 (Global minimum)**.** *When the teacher vectors are pair-wise orthogonal and the student network is under or exactly parameterized and ρ is even, a student configuration where each student vector copies a different teacher vector (up to sign symmetries) achieves the lowest loss. When $\rho = 2$, any configuration where the students vector are orthogonal and lie in the span of the teacher vectors is also a global minimum.*

Proof. Recall the loss has two terms dependent on W, $\frac{1}{2}\mathbb{1}_n^\top(WW^\top)^{\circ\rho}\mathbb{1}_n$ and $-\mathbb{1}_n^\top(WV^\top)^{\circ\rho}\mathbb{1}_k$.

Consider the first term. Observe $(WW^\top)^{\circ\rho}$ always has ones on the main diagonal and non-negative off diagonal entries. Therefore, this term is minimized when the off diagonals are all 0, which is only the case when w_is are pairwise orthogonal.

Consider the second term. Observe $-\mathbb{1}_n^\top(WV^\top)^{\circ\rho}\mathbb{1}_k$ is minimized when $\sum_{j=1}^k(w_i^\top v_j)^\rho$ is maximized for each $i \in [n]$. By lemma B.1, when $\rho > 2$ this is only the case if $w_i = v_j$ or -v_j. When $\rho = 2$, this is the case if w_i lies in the span of the teacher vectors.

Put together, this implies a configuration of student vectors is a global minimum iff it meets the minimum condition for both the first and the second term. The stated configurations are the only ones that achieve this. This completes the proof. □

3 Dynamics

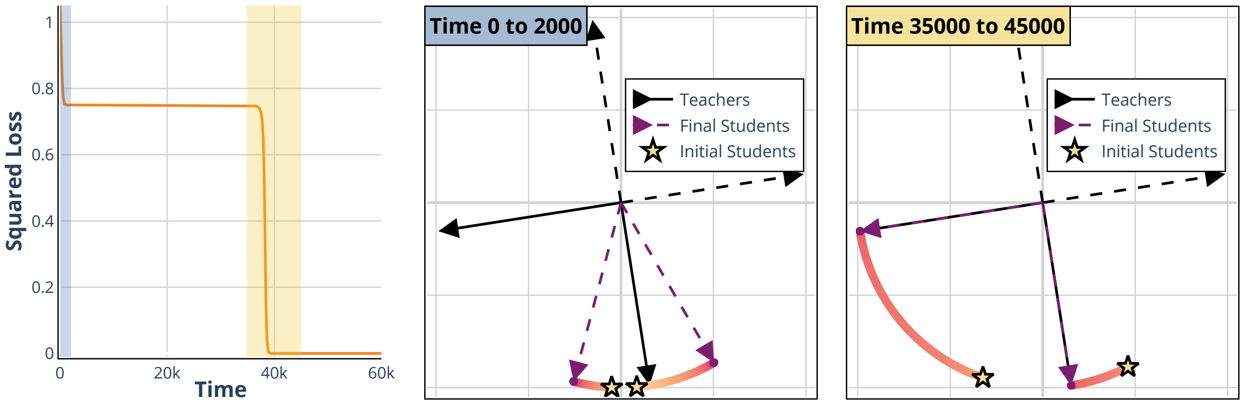
We study the training dynamics under the spherical population gradient flow,

$$\frac{dW^{(t)}}{dt} = -\nabla^{\mathbb{S}^{d-1}} L^{n,k}(W) = -\frac{\overline{\partial L^{n,k}(W)}}{\partial w_\alpha}.$$

We first describe the dynamics of the tight-balance critical points. Then we show that from favorable "split-cone initialization", we can avoid the tight-balance critical points and converge straight to the global minima. Finally, we compare the dynamics with the squared loss to the correlation loss.

3.1 Tight-balance critical point dynamics with 2 students and 2 teachers

We first consider $n = k = 2$. We simulate spherical gradient flow and analyze the dynamics empirically. When both students are initialized close to the same teacher, they often first get stuck at the tight-balance critical point, and eventually escape to the global minimum. See figure 4.



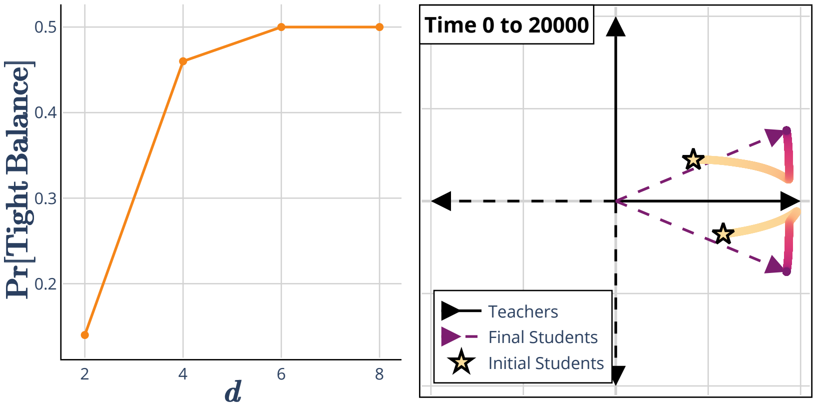
Loss curve

Close initialization to tight-balance critical point.

Tight-balance critical point to global minimum.

Figure 4: Dynamics when $n = k = d = 2$ and $\rho = 4$.

Interestingly, the probability of encountering a tight-balance critical point from uniform random initialization on the sphere increases with dimension d 5. Qualitatively, when converging to the tight-balance critical points in higher dimensions, the student vectors first appear to go into the span of the teachers, entering near one of the teachers. Then they split off into the tight-balance configuration just like in the 2 dimensional case before.



Estimated from 100 trials

$n = k = 2,\ d = 4$

Figure 5: Left: probability of encountering tight-balance critical point where $n = k = 2$, $\rho = 4$, and d varies. Right: Dynamics when $n = k = 2$, $\rho = 4$, and $d = 4$

When $n \le k \le d$, tight-balance critical points still frequently occur. However, only a few pairs of student neurons find tight-balance critical points, and the rest converge directly to a teacher vector.

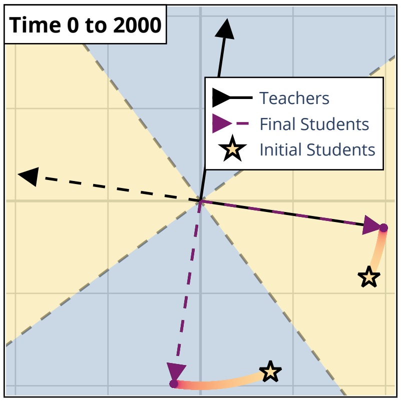


Figure 6: Split cone initialization when $n = k = d = 2$.

3.2 Toy model split-cone initialization convergence

When student vectors are initialized closest to different teachers (up to sign symmetries), student vectors avoid tight-balance critical points and quickly converge to the global minimum. The visualization of this initialization, see figure 6, looks similar to light cones in special relativity. Hence, we named this initialization split-cone initialization.

Lemma 3.1. *Suppose $n = k = d = 2$, ρ is even, and $\rho > 2$. If student vectors are initialized closest to different teacher vectors (up to sign symmetries), then the student vectors will converge to their closest teacher vectors.*

Proof Sketch. We first show if the student vectors are initialized closest to different teachers, they will remain closest to their respective teacher throughout all of training. We show this by studying the behavior of the spherical gradient flow when a student vector is nearly perfectly in between the two teacher vectors.

The only critical point attainable following split teacher initialization is the copy global minimum. Therefore gradient flow will converge to this critical point. □

Unfortunately, this exact split-cone behavior is only exhibited when $d = 2$. The simplest counter example can be constructed when $n = 2$ and $k = d = 3$, where regardless scaling on the student repulsive term or the teacher attractive force, we are not guaranteed to stay in the cone we initialize in. However, with the right scaling in our objective, we can describe an slightly more restricted initialization from which we will converge to the globally optimal copy configuration.

Lemma 3.2. *Suppose $n \leq k \leq d < \infty$, ρ is even, $\rho > 2$, and we have $1/n$ scaling on the student repulsive term. If all student vectors are initialized such that $w_i^\top v_i > 2^{-1/2}+\epsilon$ for $\epsilon = (40(\rho-3.9))^{-1}$, then the student vectors will converge to their closest teacher vectors.*

Proof Sketch. Observe with the new scaling our loss is

$$L^{n,k}(\{w_i\}_{i=1}^n) = \frac{1}{2n}\mathbb{1}_n^\top(WW^\top)^{\circ\rho}\,\mathbb{1}_n - \mathbb{1}_n^\top(WV^\top)^{\circ\rho}\,\mathbb{1}_k + C,$$

We use the same proof technique as before. We first write out the spherical gradient on the ith student vector as

$$\overline{\frac{\partial L^{n,k}(W)}{\partial w_i}} = \rho\Bigg(\sum_{j\neq i}(w_i^\top w_j)^{\rho-1}(w_j - (w_i^\top w_j)w_i) - (w_i^{\circ\rho-1} - w_i\sum_{j=1}^{k} w_{i,j}^{\rho})\Bigg).$$

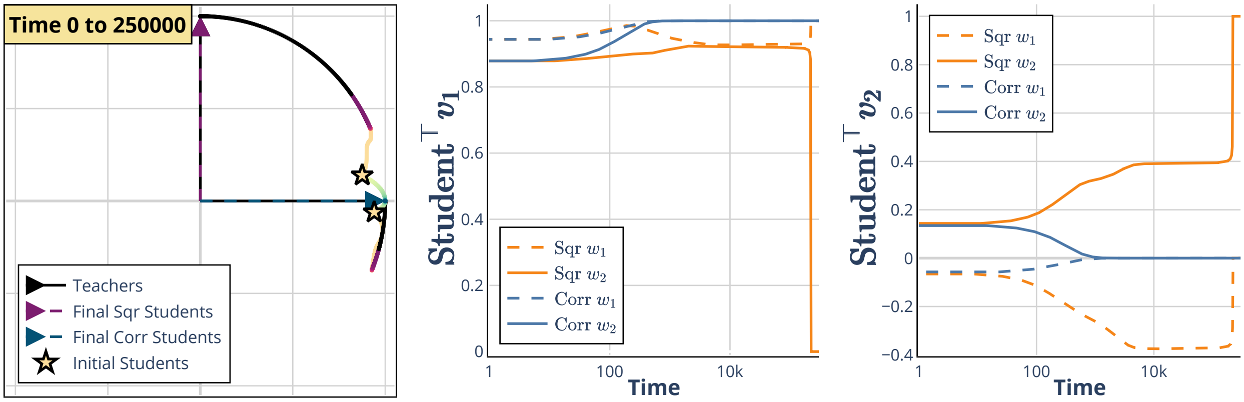
Suppose $w_{i,i} = 2^{-1/2} + \epsilon$, then our spherical gradient is

$$\begin{aligned} f(\epsilon) &= ((2^{-1/2}+\epsilon)^{\circ\rho-1} - (2^{-1/2}+\epsilon)\sum_{j=1}^{k} w_{i,j}^{\rho}) - \sum_{j\neq i}(w_i^\top w_j)^{\rho-1}(w_{j,i} - (w_i^\top w_j)(2^{-1/2}+\epsilon)) \\ &= (2^{-1/2}+\epsilon)^{\circ\rho-1} - (2^{-1/2}+\epsilon)^{\rho+1} - (2^{-1/2}+\epsilon)\sum_{j\neq i}^{k} w_{i,j}^{\rho} \\ &\quad - \sum_{j\neq i}^{n}(w_i^\top w_j)^{\rho-1}w_{j,i} + (2^{-1/2}+\epsilon)\sum_{j\neq i}^{n}(w_i^\top w_j)^{\rho}. \end{aligned}$$

The first three terms represent the "attractive" student to teacher force, and the last two terms represent the "repulsive" student to student force. We bound all the terms and achieve the desired result. □

3.3 Squared versus correlation loss

Under the correlation loss, it has been shown by Simsek et al. (2023) that student vectors monotonically approach the closest teacher vector. Under the squared loss, the student repulsive force result in more complicated dynamics when student vectors are initialized close to the same teacher. While student vectors first appear to monotonically approach the same teacher, they instead reach the tight-balance critical point in 2 dimensions and eventually converge to different teachers. See figure 7.



$n = k = 2$, $d = 3$ Inner product with v_1. Inner product with v_2.

Figure 7: Comparisons of square and correlation loss for $n = k = 2$, $d = 3$, and $\rho = 4$.

4 Broader connections

4.1 Activation function choices

Why don't people use even activation functions? One hypothesis is that even activation functions with small derivative near 0 might lead to regions of very small gradients around initialization or when the magnitudes of weights are small or something.

We ran very simple experiments to test this hypothesis and found that even activation functions with small derivatives near 0 tend to do poorly. Some of these activation functions such as quartic activations were also unstable during training, perhaps also due to large derivatives far from the origin. Even activation functions with larger derivatives near 0 such as the absolute value activation function and quadratic activation performed more comparably to ReLU activation functions. See figure 8.

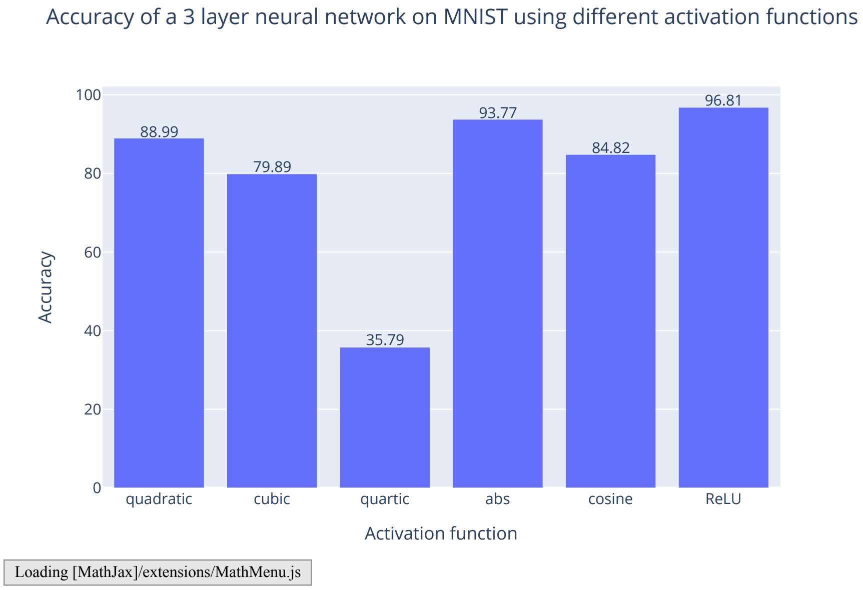


Figure 8: Performance of various even activation functions and ReLU on MNIST. Trained a 3 layer neural network with 100 neurons per layer. Trained with Adam.

5 Discussion and future directions

We made our analysis tractable by studying a toy model. In future works, we also want to take appropriate limits. Additionally, we want to expand to more activation functions, specifically odd normalized Hermite polynomials and ReLUs. We also want to prove convergence rates from various initializations to more carefully describe the effect of tight-balance critical points. We also want to further generalize tight-balance critical points and see if we can empirically observe them when looseining other assumptions such as the orthogonality of the teacher vectors. Additionally, we empirically observe global convergence to the global minimum in the toy model, but we want to prove it.

We want to explore further connections with distillation, superposition, and activation functions choices. Specifically, we wonder if our results can suggest favorable initializations for self-distillation. We also wonder what studying different activation functions will reveal for feature superposition. Superposition could be especially interesting with unequally weighted features, i.e. loosening the standard Gaussian or the unit length teacher assumptions, and correlated teacher vectors. Regarding activation function choices, we wonder if tight-balance critical points are encountered when training real models with even activation functions on real data. We also wonder if absolute value activation functions can perform competitively to ReLU or GeLU activations.

References

Misha Belkin. The necessity of machine learning theory in mitigating ai risk, July 2023. URL https://mishabelkin.substack.com/p/the-necessity-of-machine-learning.

Enric Boix-Adsera. Towards a theory of model distillation, 2024. URL https://arxiv.org/abs/2403.09053.

Simon Du and Jason Lee. On the power of over-parametrization in neural networks with quadratic activation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1329–1338. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/du18a.html.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

David Haussler and Leonard Pitt. *Proceedings of the 1988 Workshop on Computational Learning Theory: MIT, August 3-5, 1988.* M. Kaufmann, San Mateo, CA, 1989. Print.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

Simon Martin, Francis Bach, and Giulio Biroli. On the impact of overparameterization on the training of a shallow neural network in high dimensions. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 3655–3663. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/martin24a.html.

Ryan O'Donnell. Analysis of boolean functions, 2021.

David Saad and Sara A. Solla. On-line learning in soft committee machines. *Phys. Rev. E*, 52: 4225–4243, Oct 1995. doi: 10.1103/PhysRevE.52.4225. URL https://link.aps.org/doi/10.1103/PhysRevE.52.4225.

Berfin Simsek, Amire Bendjeddou, Wulfram Gerstner, and Johanni Brea. Should underparameterized student networks copy or average teacher weights? In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 78028–78068. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/f5ccb3ab757131a93586ef61ec701533-Paper-Conference.pdf.

Berfin Simsek, Amire Bendjeddou, and Daniel Hsu. Learning gaussian multi-index models with gradient flow: Time complexity and directional convergence, 2024. URL https://arxiv.org/abs/2411.08798.

David Wolpert. *The mathematics of generalization: the proceedings of the SFI/CNLS Workshop on Formal Approaches to Supervised Learning.* Santa Fe Institute studies in the sciences of complexity. Proceedings. Addison-Wesley Pub. Co, Reading, Mass, 1995. ISBN 9780201409857.

A Relation to previous works with quadratic activation function

We consider the similarity of our work to Proposition 4.1 in Martin et al. (2024). We now assume $w_i, v_j \in \mathbb{R}^d$. Recall

$$H_2(x) = \frac{1}{\sqrt{2}}(x^2 - 1).$$

By Isserlis' Theorem,

$$\begin{aligned}
g_2(w_i, v_j) &= \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[H_2(w_i^\top x)H_2(v_j^\top x)] = \mathop{\mathbb{E}}_{a,b\sim\mathcal{N}(0,\Sigma)}[H_2(a)H_2(b)] \\
&= \frac{1}{2}\mathop{\mathbb{E}}_{a,b\sim\mathcal{N}(0,\Sigma)}[(a^2-1)(b^2-1)] \\
&= \frac{1}{2}\mathop{\mathbb{E}}_{a,b\sim\mathcal{N}(0,\Sigma)}[a^2b^2 - a^2 - b^2 + 1] \\
&= \frac{1}{2}\mathop{\mathbb{E}}_{a,b\sim\mathcal{N}(0,\Sigma)}[a^2b^2] - \frac{1}{2}\|w_i\|^2 - \frac{1}{2}\|v_j\|^2 + \frac{1}{2} \\
&= \frac{1}{2}\|w_i\|^2\|v_j\|^2 + (w_i^\top v_j)^2 - \frac{1}{2}\|w_i\|^2 - \frac{1}{2}\|v_j\|^2 + \frac{1}{2},
\end{aligned}$$

where

$$\Sigma = \begin{bmatrix} \|w_i\|^2 & w_i^\top v_j \\ w_i^\top v_j & \|v_j\|^2 \end{bmatrix}.$$

Clearly, if we keep the original assumption that $w_i, v_j \in \mathbb{S}^{d-1}$, we are left with $(w_i^\top v_j)^2$. On the other hand, Martin et al. (2024) considers the squared loss, where,

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[(w_i^\top x)^2(v_j^\top x)^2] = \mathop{\mathbb{E}}_{a,b\sim\mathcal{N}(0,\Sigma)}[a^2b^2] = \|w_i\|^2\|v_j\|^2 + 2(w_i^\top v_j)^2.$$

Rewriting the loss for our setting with different scaling factors,

$$\begin{aligned}
L^{n,k}(\{w_i\}_{i=1}^n) &= \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\Big[\Big(\frac{1}{n}\sum_{i=1}^n H_2(w_i^\top x) - \frac{1}{k}\sum_{j=1}^k H_2(v_j^\top x)\Big)^2\Big] \\
&= \frac{1}{n^2}\sum_{i=1}^n\sum_{q=1}^n g_\rho(w_i^\top x, w_q^\top x) - 2\frac{1}{nk}\sum_{i=1}^n\sum_{j=1}^k g_\rho(w_i^\top x, v_j^\top x) + \frac{1}{k^2}\sum_{j=1}^k\sum_{q=1}^k g_\rho(v_j^\top x, v_q^\top x) \\
&= \frac{1}{n^2}\sum_{i=1}^n\sum_{q=1}^n \frac{1}{2}\|w_i\|^2\|w_q\|^2 + (w_i^\top w_q)^2 - \frac{1}{2}\|w_i\|^2 - \frac{1}{2}\|w_q\|^2 + \frac{1}{2} \\
&\quad - 2\frac{1}{nk}\sum_{i=1}^n\sum_{j=1}^k \frac{1}{2}\|w_i\|^2\|v_j\|^2 + (w_i^\top v_j)^2 - \frac{1}{2}\|w_i\|^2 - \frac{1}{2}\|v_j\|^2 + \frac{1}{2} \\
&\quad + \frac{1}{k^2}\sum_{j=1}^k\sum_{r=1}^k \frac{1}{2}\|v_j\|^2\|v_r\|^2 + (v_j^\top v_r)^2 - \frac{1}{2}\|v_j\|^2 - \frac{1}{2}\|v_r\|^2 + \frac{1}{2}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{q=1}^{n} -\frac{1}{2}\|w_i\|^2 - \frac{1}{2}\|w_q\|^2 + \frac{1}{2} - 2\frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k} -\frac{1}{2}\|w_i\|^2 - \frac{1}{2}\|v_j\|^2 + \frac{1}{2} \\
&\quad + \frac{1}{k^2}\sum_{j=1}^{k}\sum_{r=1}^{k} -\frac{1}{2}\|v_j\|^2 - \frac{1}{2}\|v_r\|^2 + \frac{1}{2} \\
&\quad + \frac{1}{2}\Bigg(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{q=1}^{n} \|w_i\|^2\|w_q\|^2 + 2(w_i^\top w_q)^2 - 2\frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k} \|w_i\|^2\|v_j\|^2 + 2(w_i^\top v_j)^2 \\
&\quad + \frac{1}{k^2}\sum_{j=1}^{k}\sum_{r=1}^{k} \|v_j\|^2\|v_r\|^2 + 2(v_j^\top v_r)^2\Bigg).
\end{aligned}
$$

Observe the term with the parentheses can be rewritten in terms of an expectation of the quadratic loss and the other terms can be simplified,

$$
\begin{aligned}
&= \frac{1}{2} + 1 + 1 - \frac{1}{n}\sum_{i=1}^{n}\|w_i\|^2 + \frac{1}{n}\sum_{i=1}^{n}\|w_i\|^2 + \frac{1}{k}\sum_{j=1}^{k}\|v_j\|^2 - \frac{1}{k}\sum_{j=1}^{k}\|v_j\|^2 \\
&\quad + \frac{1}{2}\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(\frac{1}{n}\sum_{i=1}^{n}(w_i^\top x)^2 - \frac{1}{k}\sum_{j=1}^{k}(v_j^\top x)^2\right)^2\right] \\
&= \frac{5}{2} + \frac{1}{2}\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(\frac{1}{n}\sum_{i=1}^{n}(w_i^\top x)^2 - \frac{1}{k}\sum_{j=1}^{k}(v_j^\top x)^2\right)^2\right]
\end{aligned}
$$

The derivative of our loss is the same as Martin et al. (2024) up to scaling, additive factors, and multiplicative factors.

B Critical point analysis proofs

B.1 Derivative

We first compute the derivative of our objective with respect to $w_{\alpha,\beta}$,

$$
\begin{aligned}
\frac{\partial L^{n,k}(W)}{\partial w_{\alpha,\beta}} &= \frac{1}{2}\frac{\partial}{\partial w_{\alpha,\beta}}\sum_{i=1}^{n}\sum_{q=1}^{n}(w_i^\top w_q)^\rho - \frac{\partial}{\partial w_{\alpha,\beta}}\sum_{i=1}^{n}\sum_{j=1}^{k}(w_i^\top v_j)^\rho \\
&= \sum_{i\in[n],i\neq\alpha}\frac{\partial}{\partial w_{\alpha,\beta}}(w_\alpha^\top w_i)^\rho + \frac{1}{2}\frac{\partial}{\partial w_{\alpha,\beta}}(w_\alpha^\top w_\alpha)^\rho - \sum_{j=1}^{k}\frac{\partial}{\partial w_{\alpha,\beta}}(w_\alpha^\top v_j)^\rho \\
&= \sum_{i\in[n],i\neq\alpha}\rho w_{i,\beta}(w_\alpha^\top w_i)^{\rho-1} + \rho w_{\alpha,\beta}(w_\alpha^\top w_\alpha)^{\rho-1} - \sum_{j=1}^{k}\rho v_{j,\beta}(w_\alpha^\top v_j)^{\rho-1} \\
&= \rho w_{:,\beta}^\top(Ww_\alpha)^{\circ\rho-1} - \rho v_{:,\beta}^\top(Vw_\alpha)^{\circ\rho-1}.
\end{aligned}
$$

Therefore,

$$
\frac{\partial L^{n,k}(W)}{\partial w_\alpha} = \rho W^\top(Ww_\alpha)^{\circ\rho-1} - \rho V^\top(Vw_\alpha)^{\circ\rho-1},
$$

and

$$\begin{aligned}\frac{\partial L^{n,k}(W)}{\partial W} &= \begin{bmatrix} \frac{\partial L^{n,k}(W)}{\partial w_{1,1}} & \\ & \ddots \end{bmatrix} \\ &= \rho(WW^\top)^{\circ\rho-1}W - \rho(WV^\top)^{\circ\rho-1}V.\end{aligned}$$

B.2 Derivative on the sphere

We restrict the rows of the derivative to be on the sphere by projecting each row onto the tangent space $T_{w_\alpha}(\mathbb{S}^{d-1})$ of the sphere

$$\begin{aligned}\overline{\frac{\partial L^{n,k}(W)}{\partial W}} &= \begin{bmatrix} - & P_{w_1}\left(\frac{\partial L^{n,k}(W)}{\partial w_1}\right) & - \\ & \vdots & \\ - & P_{w_n}\left(\frac{\partial L^{n,k}(W)}{\partial w_n}\right) & - \end{bmatrix} \\ &= \begin{bmatrix} - & (I - w_1 w_1^\top)\left(\frac{\partial L^{n,k}(W)}{\partial w_1}\right) & - \\ & \vdots & \\ - & (I - w_n w_n^\top)\left(\frac{\partial L^{n,k}(W)}{\partial w_n}\right) & - \end{bmatrix}.\end{aligned}$$

Let $\operatorname{diag}(u)$ denote the matrix with the entries of the vector u as its diagonal. We consider the αth row of the projected derivative matrix

$$\begin{aligned}\overline{\frac{\partial L^{n,k}(W)}{\partial w_\alpha}} &= (I - w_\alpha w_\alpha^\top)\Big(\rho W^\top (W w_\alpha)^{\circ\rho-1} - \rho V^\top (V w_\alpha)^{\circ\rho-1}\Big) \\ &= \rho W^\top (W w_\alpha)^{\circ\rho-1} - \rho w_\alpha w_\alpha^\top W^\top (W w_\alpha)^{\circ\rho-1} \\ &\quad - \rho V^\top (V w_\alpha)^{\circ\rho-1} + \rho w_\alpha w_\alpha^\top V^\top (V w_\alpha)^{\circ\rho-1} \\ &= \rho W^\top (W w_\alpha)^{\circ\rho-1} - \rho w_\alpha w_\alpha^\top W^\top (W w_\alpha)^{\circ\rho-1} \\ &\quad - \rho V^\top (V w_\alpha)^{\circ\rho-1} + \rho w_\alpha w_\alpha^\top V^\top (V w_\alpha)^{\circ\rho-1} \\ &= \rho W^\top (W w_\alpha)^{\circ\rho-1} - \rho w_\alpha w_\alpha^\top W^\top (W w_\alpha)^{\circ\rho-1} \\ &\quad - \rho V^\top (V w_\alpha)^{\circ\rho-1} + \rho w_\alpha w_\alpha^\top V^\top (V w_\alpha)^{\circ\rho-1}.\end{aligned}$$

Putting it all together, our spherical gradient is

$$\begin{aligned}\overline{\frac{\partial L^{n,k}(W)}{\partial W}} &= \rho(WW^\top)^{\circ\rho-1}W - \rho\big((WW^\top)^{\circ\rho-1}WW^\top \circ I_n\big)W \\ &\quad - \rho(WV^\top)^{\circ\rho-1}V + \rho\big((WV^\top)^{\circ\rho-1}VW^\top \circ I_n\big)W \\ &= \rho(WW^\top)^{\circ\rho-1}W - \rho(WV^\top)^{\circ\rho-1}V \\ &\quad - \Big(\Big[\rho(WW^\top)^{\circ\rho-1}W - \rho(WV^\top)^{\circ\rho-1}V\Big]W^\top \circ I_n\Big)W.\end{aligned}$$

Let $A = \rho(WW^\top)^{\circ\rho-1}W - \rho(WV^\top)^{\circ\rho-1}V$, then

$$\overline{\frac{\partial L^{n,k}(W)}{\partial W}} = A - (AW^\top \circ I_n)W.$$

I have experimentally verified that this is the correct expression for the spherical gradient.

B.3 Maximum of sum of dot products

Lemma B.1 (Maximum of sum of dot products)**.** *Suppose $v_1, \dots, v_k \in \mathbb{R}^d$ are pairwise orthogonal and unit length. Consider the optimization problem*

$$\max_{w \in \mathbb{S}^{d-1}} \sum_{j=1}^{k} (w^\top v_j)^\rho.$$

The solution is 1. When $\rho = 2$, span$(v_1, \dots, v_k) \cup \mathbb{S}^{d-1}$ is the set of all w that attain 1. When $\rho > 2$, $\{v_1, \dots, v_k\}$ is the set of all w that attain 1.

Proof. When $\rho = 2$, observe

$$\sum_{j=1}^{k} (w^\top v_j)^2 = (Vw)^\top (Vw) = \|Vw\|_2^2.$$

Consider the orthonormal matrix $V' \in \mathbb{R}^{d \times d}$ which has $v_1, \dots, v_k$ as its first k rows and other unit vectors for the last $d-k$ rows that satisfy orthogonality. Since $w \in \text{span}(v_1, \dots, v_k)$, the inner product of w and any of the last $d-k$ rows of V' is 0. Therefore $\|Vw\|_2^2 = \|V'w\|_2^2$. Observe V' is a rotation matrix and recall rotation matrices preserve distances. Therefore $\|V'w\|_2^2 = \|w\|_2^2 = 1$, proving the first case.

Consider $\rho > 2$. Let $u = Vw$. By the orthogonality of the teacher vectors, any $u \in \text{span}(v_1, \dots, v_k) \cup \mathbb{S}^{d-1}$ corresponds to a unique $w \in \text{span}(v_1, \dots, v_k)$. Therefore we can write the Lagrangian function as

$$\mathcal{L}(u) = \sum_{j=1}^{k} u_j^\rho + \lambda \Big(\sum_{j=1}^{k} u_j^2 - 1 \Big).$$

Recall,

$$\frac{d\mathcal{L}}{du_j} = \rho u_j^{\rho-1} + 2\lambda u_j = 0$$

is a condition for the maximum, and thus

$$\rho u_j^{\rho-2} = -2\lambda.$$

The left hand side is a strictly increasing function of u_j for all $\rho > 2$, therefore there is a unique u_j^* that satisfies this equation. This implies that all non-zero coordinates of u must be equal and that candidates for the optimal solution have the form

$$u^*(l) = \Big(\underbrace{0, \dots, 0}_{k-l}, \underbrace{\frac{1}{\sqrt{l}}, \dots, \frac{1}{\sqrt{l}}}_{l} \Big), \quad l \in [k].$$

Evaluating the objective for all these candidate solutions, we have

$$\sum_{j=1}^{k} u_j^{*\rho} = l^{1-\rho/2},$$

which is maximized to 1 at $l = 1$ for every $\rho > 2$, implying the optimal solution is $u^* = (1, 0, \dots, 0)$. This corresponds to some v_j (or $-v_j$ in the case of even ρ), proving the second case. □

B.4 Identifying critical points

Theorem B.2 (Orthogonal copy critical points)**.** *Every arrangement of student vectors that satisfies the following conditions is a critical point:*

- *for each $i \in [n]$, $w_i = v_j$ for some v_j,*
- *for all $i \in [n]$ and $j \in [k]$, $w_i^\top v_j \in \{-1, 0, 1\}$.*

Proof. Consider the case where ρ is odd. Then the ith row of B is

$$\begin{aligned} B_i &= \sum_{j=1}^{n} (w_j^\top w_i)^{\rho-1} w_j - \sum_{j=1}^{k} (v_j^\top w_i)^{\rho-1} v_j \\ &= w_i \sum_{j=1}^{n} w_j^\top w_i - w_i \sum_{j=1}^{k} v_j^\top w_i. \end{aligned}$$

Since $\|w_i\|_2^2 = 1$, the ith row of the spherical gradient simplifies to

$$\begin{aligned} \overline{\frac{\partial L^{n,k}(W)}{\partial w_i}} &= w_i \sum_{j=1}^{n} w_j^\top w_i - w_i \sum_{j=1}^{k} v_j^\top w_i - \Big(w_i \sum_{j=1}^{n} w_j^\top w_i - w_i \sum_{j=1}^{k} v_j^\top w_i \Big)^\top w_i w_i \\ &= w_i \sum_{j=1}^{n} w_j^\top w_i - w_i \sum_{j=1}^{k} v_j^\top w_i - w_i \sum_{j=1}^{n} w_j^\top w_i + w_i \sum_{j=1}^{k} v_j^\top w_i = 0. \end{aligned}$$

This completes the proof for the odd case.

Consider the case where ρ is even. Then the ith row of B is

$$\begin{aligned} B_i &= \sum_{j=1}^{n} (w_j^\top w_i)^{\rho-1} w_j - \sum_{j=1}^{k} (v_j^\top w_i)^{\rho-1} v_j \\ &= w_i \sum_{j=1}^{n} |w_j^\top w_i| - w_i \sum_{j=1}^{k} |v_j^\top w_i|. \end{aligned}$$

Since $\|w_i\|_2^2 = 1$, the ith row of the spherical gradient simplifies to

$$\begin{aligned} \overline{\frac{\partial L^{n,k}(W)}{\partial w_i}} &= w_i \sum_{j=1}^{n} |w_j^\top w_i| - w_i \sum_{j=1}^{k} |v_j^\top w_i| - \Big(w_i \sum_{j=1}^{n} |w_j^\top w_i| - w_i \sum_{j=1}^{k} |v_j^\top w_i| \Big)^\top w_i w_i \\ &= w_i \sum_{j=1}^{n} |w_j^\top w_i| - w_i \sum_{j=1}^{k} |v_j^\top w_i| - w_i \sum_{j=1}^{n} |w_j^\top w_i| + w_i \sum_{j=1}^{k} |v_j^\top w_i| = 0. \end{aligned}$$

This completes the proof for the even case. This completes the entire proof. □

Lemma B.3 (Euclidean gradient cannot be 0)**.** *If the teacher vectors are orthogonal, ρ is even, $\rho > 0$, and $n < k$, there exist no configuration of student neurons such that $B = 0$.*

Proof. We aim to show $B \neq 0$. Observe a sufficient condition is that there exists a $i \in [n]$ such that the squared ℓ^2 norm of the ith row of $(WW^\top)^{\circ\rho-1}W$ is not equal to the squared ℓ^2 norm of the ith row of $(WV^\top)^{\circ\rho-1}V$.

We compute the squared ℓ^2 norm of the ith row of $(WV^\top)^{\circ\rho-1}V$,

$$\|((WV^\top)^{\circ\rho-1}V)_i\|_2^2 = \|\sum_{j=1}^{k}(v_j^\top w_i)^{\rho-1}v_j\|_2^2 = \sum_{j=1}^{k}(v_j^\top w_i)^{2\rho-2}.$$

By lemma B.1, we know this quantity is at most 1, with equality only when $w_i = v_j$ for some $j \in [k]$.

We compute the squared ℓ^2 norm of the ith row of $(WW^\top)^{\circ\rho-1}W$,

$$\begin{aligned}\|((WW^\top)^{\circ\rho-1}W)_i\|_2^2 &= \|\sum_{j=1}^{k}(w_j^\top w_i)^{\rho-1}w_j\|_2^2 \\ &= 2\sum_{j'>j}^{n}(w_j^\top w_i)^{\rho-1}(w_{j'}^\top w_i)^{\rho-1}w_{j'}^\top w_j + \sum_{j=1}^{n}(w_j^\top w_i)^{2\rho-2} \\ &= 2\sum_{j\neq i}^{n}(w_j^\top w_i)^{\rho} + 2\sum_{\substack{j'>j\\ j',j\neq i}}^{n}(w_j^\top w_i)^{\rho-1}(w_{j'}^\top w_i)^{\rho-1}w_{j'}^\top w_j + 1 + \sum_{j\neq i}^{n}(w_j^\top w_i)^{2\rho-2}.\end{aligned}$$

We wish to show this quantity is at least 1, therefore showing that the norms can never be equal and proving the original statement. It is sufficient to show

$$2\sum_{\substack{j'>j\\ j',j\neq i}}^{n}(w_j^\top w_i)^{\rho-1}(w_{j'}^\top w_i)^{\rho-1}w_{j'}^\top w_j + \sum_{j\neq i}^{n}(w_j^\top w_i)^{2\rho-2} \geq 0.$$

Observe

$$2\sum_{\substack{j'>j\\ j',j\neq i}}^{n}(w_j^\top w_i)^{\rho-1}(w_{j'}^\top w_i)^{\rho-1}w_{j'}^\top w_j + \sum_{j\neq i}^{n}(w_j^\top w_i)^{2\rho-2} = \left\|\sum_{j\neq i}(w_j^\top w_i)^{\rho-1}w_j\right\|_2^2 \geq 0.$$

This completes the proof. □

C Dynamics proofs

C.1 Toy model split-cone initialization convergence proof

Lemma C.1. *Suppose $n = k = d = 2$ and ρ is even and $\rho > 2$. If student vectors are initialized closest to different teacher vectors (up to sign symmetries), then the student vectors will converge to their closest teacher vectors.*

Proof. We first show if the student vectors are initialized closest to different teachers, they will remain closest to their respective teacher throughout all of training. We show this by studying the behavior of the spherical gradient flow when a student vector is nearly perfectly in between the two teacher vectors.

Recall the spherical gradient on the ith student vector is

$$\begin{aligned}\overline{\frac{\partial L^{n,k}(W)}{\partial w_i}} &= \rho W^\top (Ww_i)^{\circ\rho-1} - \rho V^\top (Vw_i)^{\circ\rho-1} \\ &\quad - \rho w_i w_i^\top W^\top (Ww_i)^{\circ\rho-1} + \rho w_i w_i^\top V^\top (Vw_i)^{\circ\rho-1} \\ &= \rho\bigg(w_i + \sum_{j\neq i}(w_i^\top w_j)^{\rho-1} w_j - \sum_{j=1}^{k}(w_i^\top v_j)^{\rho-1} v_j \\ &\quad - w_i w_i^\top w_i - w_i w_i^\top \sum_{j\neq i}(w_i^\top w_j)^{\rho-1} w_j + w_i w_i^\top \sum_{j=1}^{k}(w_i^\top v_j)^{\rho-1} v_j \bigg) \\ &= \rho\bigg(\sum_{j\neq i}(w_i^\top w_j)^{\rho-1} w_j - \sum_{j=1}^{k}(w_i^\top v_j)^{\rho-1} v_j - w_i \sum_{j\neq i}(w_i^\top w_j)^{\rho} + w_i \sum_{j=1}^{k}(w_i^\top v_j)^{\rho} \bigg) \\ &= \rho\bigg(\sum_{j\neq i}(w_i^\top w_j)^{\rho-1}(w_j - (w_i^\top w_j)w_i) - \sum_{j=1}^{k}(w_i^\top v_j)^{\rho-1}(v_j - (w_i^\top v_j)w_i) \bigg).\end{aligned}$$

We can split the gradient flow update into the sum of three terms, each of which lie in the tangent space of w_i,

$$-\underbrace{\rho(w_i^\top w_j)^{\rho-1}(w_j - (w_i^\top w_j)w_i)}_{1} + \underbrace{\rho(w_i^\top v_1)^{\rho-1}(v_1 - (w_i^\top v_1)w_i)}_{2} + \underbrace{\rho(w_i^\top v_2)^{\rho-1}(v_2 - (w_i^\top v_2)w_i)}_{3}.$$

We assume $w_i^\top v_i = 2^{-1/2} + \epsilon$ for small $\epsilon > 0$ and symmetrically $i \in [2]$. Observe that the update is only one dimensional, so as long as it is pointing in the direction of v_1, we will never escape the cone. Therefore we must only compare the magnitudes. Observe

$$\begin{aligned}\big\|(w_i^\top v_1)^{\rho-1}(v_1 - (w_i^\top v_1)w_i)\big\|_2 &= (w_i^\top v_1)^{\rho-1}\big\|(v_1 - (w_i^\top v_1)w_i)\big\|_2 \\ &= (w_i^\top v_1)^{\rho-1}\sqrt{\|v_1\|^2 - 2(w_i^\top v_1)^2 + (w_i^\top v_1)^2\|w_1\|} \\ &= (w_i^\top v_1)^{\rho-1}\sqrt{1 - (w_i^\top v_1)^2} = (w_i^\top v_1)^{\rho-1}(w_i^\top v_2).\end{aligned}$$

We consider the case where $2^{-1/2} < w_2^\top v_2 < 2^{-1/2} + \epsilon$ and $w_2^\top v_1 < 0$, since otherwise, term 1 pushes w_1 in the direction of v_1.

We must show

$$-(w_1^\top w_2)^{\rho-1}\sqrt{1 - (w_1^\top w_2)^2} + (w_1^\top v_1)^{\rho-1}(w_1^\top v_2) - (w_1^\top v_2)^{\rho-1}(w_1^\top v_1) > 0.$$

Since $w_1^\top v_1 = 2^{-1/2} + \epsilon$,

$$w_1^\top v_2 = \sqrt{1 - (2^{-1/2} + \epsilon)^2},$$

and

$$-2^{-1/2}(2^{-1/2} + \epsilon) + 2^{-1/2}\sqrt{1 - (2^{-1/2} + \epsilon)^2} < w_1^\top w_2 < 0.$$

Therefore,

$$\begin{aligned}
(w_1^\top w_2)^{\rho-1}&\sqrt{1-(w_1^\top w_2)^2} + (w_1^\top v_1)^{\rho-1}(w_1^\top v_2) - (w_1^\top v_2)^{\rho-1}(w_1^\top v_1)\\
&> (w_1^\top w_2)^{\rho-1} + (w_1^\top v_1)^{\rho-1}(w_1^\top v_2) - (w_1^\top v_2)^{\rho-1}(w_1^\top v_1)\\
&> \Big(-2^{-1/2}(2^{-1/2}+\epsilon) + 2^{-1/2}\sqrt{1-(2^{-1/2}+\epsilon)^2}\Big)^{\rho-1}\\
&\quad + (2^{-1/2}+\epsilon)^{\rho-1}\sqrt{1-(2^{-1/2}+\epsilon)^2} - \sqrt{1-(2^{-1/2}+\epsilon)^2}^{\rho-1}(2^{-1/2}+\epsilon) = f(\epsilon).
\end{aligned}$$

When $\epsilon = 0$, $\sqrt{1-(2^{-1/2}+\epsilon)^2} = 2^{-1/2}$, and

$$\begin{aligned}
f(0) &= \Big(-2^{-1/2}(2^{-1/2}+\epsilon) + 2^{-1/2}\sqrt{1-(2^{-1/2}+\epsilon)^2}\Big)^{\rho-1}\\
&\quad + (2^{-1/2}+\epsilon)^{\rho-1}\sqrt{1-(2^{-1/2}+\epsilon)^2} - \sqrt{1-(2^{-1/2}+\epsilon)^2}^{\rho-1}(2^{-1/2}+\epsilon)\\
&= \Big(-2^{-1/2}2^{-1/2} + 2^{-1/2}2^{-1/2}\Big)^{\rho-1} + (2^{-1/2})^{\rho-1}2^{-1/2} - (2^{-1/2})^{\rho-1}2^{-1/2} = 0.
\end{aligned}$$

Additionally, the derivative with respect to ϵ is

$$\begin{aligned}
f'(\epsilon) = \Big(&-2^{-1/2}\epsilon - \frac{2^{-1/2}+\epsilon}{2^{-1/2}\sqrt{1-(2^{-1/2}+\epsilon)^2}}\Big)\Big(-2^{-1/2}(2^{-1/2}+\epsilon) + 2^{-1/2}\sqrt{1-(2^{-1/2}+\epsilon)^2}\Big)^{\rho-2}\\
&+ (\rho-1)(2^{-1/2}+\epsilon)^{\rho-2}\sqrt{1-(2^{-1/2}+\epsilon)^2} + \frac{(2^{-1/2}+\epsilon)^\rho}{\sqrt{1-(2^{-1/2}+\epsilon)^2}}\\
&- \sqrt{1-(2^{-1/2}+\epsilon)^2}^{\rho-1} + (\rho-1)(2^{-1/2}+\epsilon)\sqrt{1-(2^{-1/2}+\epsilon)^2}^{\rho-3}(2^{-1/2}+\epsilon).
\end{aligned}$$

When $\epsilon = 0$, the derivative is equal to

$$\begin{aligned}
f'(0) &= \Big(-\frac{2^{-1/2}}{2^{-1/2}2^{-1/2}}\Big)\Big(-2^{-1/2}(2^{-1/2}) + 2^{-1/2}2^{-1/2}\Big)^{\rho-2} + (\rho-1)(2^{-1/2})^{\rho-2}2^{-1/2}\\
&\quad + \frac{(2^{-1/2})^\rho}{2^{-1/2}} - \frac{1}{\sqrt{2}}^{\rho-1} + (\rho-1)(2^{-1/2})\frac{1}{\sqrt{2}}^{\rho-3}2^{-1/2}\\
&= 2(\rho-1)(2^{-1/2})^{\rho-1}.
\end{aligned}$$

Observe f is continuous for $-1-\sqrt{2}/2 < \epsilon < 1-\sqrt{2}/2$ Therefore, there exist a $0 < \delta < 1-\sqrt{2}/2$ such that for all $0 < \epsilon < \delta$, $f(\epsilon) > 0$. This proves throughout all of training, $w_1^\top v_1 > w_1^\top v_2$ and similarly for w_2. In words, if the student vectors are initialized closest to different teachers, they will remain closest to their respective teacher throughout all of training.

The only critical point attainable following split teacher initialization is the copy global minimum. Therefore gradient flow will converge to this critical point.

Trivially,

$$(2^{-1/2}+\epsilon) - \sqrt{\frac{1}{2} - \sqrt{2}\epsilon - \epsilon^2} > 0.$$

This completes the proof. □

Lemma C.2. *Suppose $n \le k \le d < \infty$, ρ is even, $\rho > 2$, and we have $1/n$ scaling on the student repulsive term. If all student vectors are initialized such that $w_i^\top v_i > 2^{-1/2}+\epsilon$ for $\epsilon = (40(\rho-3.9))^{-1}$, then the student vectors will converge to their closest teacher vectors.*

Proof. We first show if the student vectors are initialized closest to different teachers, they will remain closest to their respective teacher throughout all of training. We show this by showing spherical gradient flow takes a student vector towards the closest teacher if the student vector is within a certain distance of the teacher.

WLOG we assume $v_i = e_i$.

Then the spherical gradient on the ith student vector is

$$\begin{aligned}\overline{\frac{\partial L^{n,k}(W)}{\partial w_i}} &= \rho\Bigg(\sum_{j\neq i}(w_i^\top w_j)^{\rho-1}(w_j - (w_i^\top w_j)w_i) - \sum_{j=1}^{k}(w_{i,j})^{\rho-1}(e_j - w_{i,j}w_i)\Bigg) \\ &= \rho\Bigg(\sum_{j\neq i}(w_i^\top w_j)^{\rho-1}(w_j - (w_i^\top w_j)w_i) - (w_i^{\circ\rho-1} - w_i\sum_{j=1}^{k} w_{i,j}^{\rho})\Bigg).\end{aligned}$$

For convenience, we omit the multiplicative ρ factor and consider how much the spherical gradient flow points in the direction of the closest teacher vector $v_1 - (w_i^\top v_1)w_i = e_i - w_{i,i}w_i$,

$$\begin{aligned}&-\overline{\frac{\partial L^{n,k}(W)}{\partial w_i}}^\top (e_i - w_{i,i}w_i) \\ &\qquad = -\Bigg(\sum_{j\neq i}(w_i^\top w_j)^{\rho-1}(w_j - (w_i^\top w_j)w_i) - (w_i^{\circ\rho-1} - w_i\sum_{j=1}^{k} w_{i,j}^{\rho}w_i)\Bigg)^\top (e_i - w_{i,i}w_i) \\ &\qquad = -\Bigg(\sum_{j\neq i}(w_i^\top w_j)^{\rho-1}(w_j - (w_i^\top w_j)w_i) - (w_i^{\circ\rho-1} - w_i\sum_{j=1}^{k} w_{i,j}^{\rho})\Bigg)^\top (e_i) \\ &\qquad = (w_{i,i}^{\circ\rho-1} - w_{i,i}\sum_{j=1}^{k} w_{i,j}^{\rho}) - \sum_{j\neq i}(w_i^\top w_j)^{\rho-1}(w_{j,i} - (w_i^\top w_j)w_{i,i}).\end{aligned}$$

Suppose $w_{i,i} = 2^{-1/2} + \epsilon$,

$$\begin{aligned}f(\epsilon) &= ((2^{-1/2}+\epsilon)^{\circ\rho-1} - (2^{-1/2}+\epsilon)\sum_{j=1}^{k} w_{i,j}^{\rho}) - \sum_{j\neq i}(w_i^\top w_j)^{\rho-1}(w_{j,i} - (w_i^\top w_j)(2^{-1/2}+\epsilon)) \\ &= (2^{-1/2}+\epsilon)^{\circ\rho-1} - (2^{-1/2}+\epsilon)^{\rho+1} - (2^{-1/2}+\epsilon)\sum_{j\neq i}^{k} w_{i,j}^{\rho} \\ &\quad - \sum_{j\neq i}^{n}(w_i^\top w_j)^{\rho-1}w_{j,i} + (2^{-1/2}+\epsilon)\sum_{j\neq i}^{n}(w_i^\top w_j)^{\rho}.\end{aligned}$$

The first three terms represent the "attractive" student to teacher force, and the last two terms represent the "repulsive" student to student force. By lemma B.1, the attractive terms are minimized when there exists some $l \in [k] \setminus i$ such that $w_{i,l} = \sqrt{1-(2^{-1/2}+\epsilon)^2}$ and all other entries of w_i are 0. We write this formally as

$$a(\epsilon) \geq (2^{-1/2}+\epsilon)^{\circ\rho-1} - (2^{-1/2}+\epsilon)^{\rho+1} - (2^{-1/2}+\epsilon)\sqrt{1-(2^{-1/2}+\epsilon)^2}^{\rho}$$

The repulsive terms can be lower bounded as

$$
\begin{aligned}
r(\epsilon) &\geq -(2^{-1/2}+\epsilon)\Big(\sum_{j\neq i}^{n}(w_i^\top w_j)^{\rho-1} - \sum_{j\neq i}^{n}(w_i^\top w_j)^{\rho}\Big) \\
&\geq \min_{\theta\in[0,1-(2^{-1/2}+\epsilon)^2]} -(2^{-1/2}+\epsilon)(n-1)(\theta^{\rho-1}-\theta^{\rho}) \\
&= -(2^{-1/2}+\epsilon)(n-1)\Big((1-(2^{-1/2}+\epsilon)^2)^{\rho-1} - (1-(2^{-1/2}+\epsilon)^2)^{\rho}\Big)
\end{aligned}
$$

since $w_{j,i} \leq \sqrt{1-(2^{-1/2}+\epsilon)^2} \leq (2^{-1/2}+\epsilon)$.

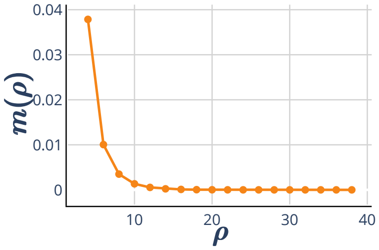


Figure 9: Lower bound on ϵ with nice properties.

We find for $\epsilon \geq m(\rho)$, see figure 9, with appropriate scaling, e.g. $1/n$ on the repulsive force,

$$f(\epsilon) = \frac{1}{n}r(\epsilon) + a(\epsilon) \geq 0.$$

Observe it is sufficient for

$$\epsilon = \frac{1}{40(\rho-3.9)}.$$

This completes the proof. □