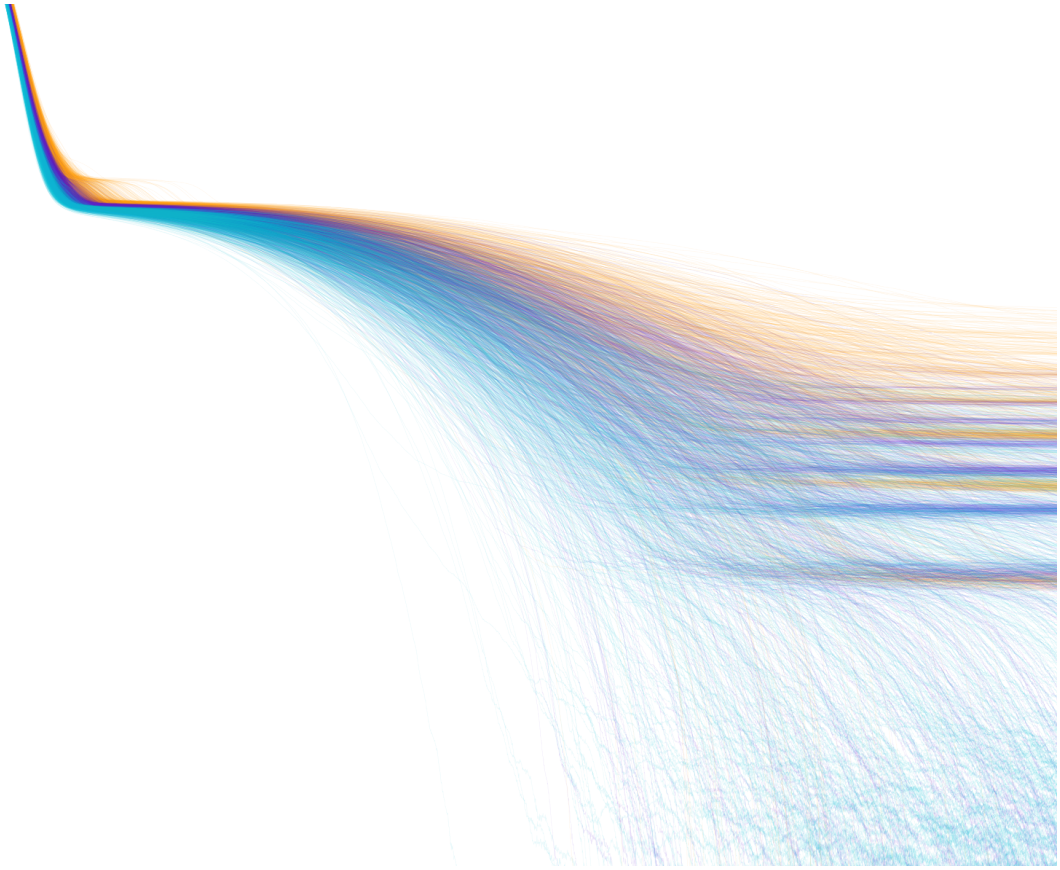


Mean-Field Analysis of Loss Landscapes in Overparameterized Two-Layer ReLU Neural Networks

Master's thesis in Computer science and engineering

JIE HUANG

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2026

Master's thesis 2026

Mean-Field Analysis of Loss Landscapes in Overparameterized Two-Layer ReLU Neural Networks

JIE HUANG

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2026

Mean-Field Analysis of Loss Landscapes in Overparameterized Two-Layer ReLU Neural Networks
JIE HUANG

Supervisor: Stefano Sarao Mannelli , DSAI, Computer Science and Engineering
Examiner: Ashkan Panahi, DSAI, Computer Science and Engineering

Master's Thesis 2026
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Evolution of population loss under gradient flow for $M = 20$ teacher units, with 1,000 random initializations per condition in two-layer ReLU neural networks with fixed second-layer weights. The well-specified model ($K = 20$ student units, orange) often gets trapped in discrete high-loss plateaus. Mild over-parameterization ($K = 21$, purple; $K = 22$, blue) progressively destabilizes intermediate plateaus, allowing more trajectories to escape towards the global minimum.

Typeset in LaTeX
Gothenburg, Sweden 2026

Mean-Field Analysis of Loss Landscapes in Overparameterized Two-Layer ReLU Neural Networks
JIE HUANG Department of Physics and Astronomy
Chalmers University of Technology and University of Gothenburg

Abstract

We study the loss landscape of high-dimensional two-layer ReLU neural networks in the teacher-student setting through a macroscopic description in terms of summary statistics, where the training dynamics reduce to a closed system of ordinary differential equations for the order parameters. When the student and teacher networks have the same size, we identify several structured families of spurious local minima organized into discrete loss levels. These families exhibit block-structured order parameters, characterized by aligned and anti-aligned groups of student units. Motivated by this structure, we introduce a reduced ansatz and compute theoretical fixed points whose loss values and order-parameter patterns agree with large-scale ODE simulations. When the student network is slightly overparameterized by adding a single extra neuron, the loss landscape changes qualitatively: the lowest-order spurious minima family is destabilized, the basin of attraction of the global minimum expands, and the remaining high-loss families become less frequent. Using the string method as an exploratory diagnostic, we find that representative fixed points separated by visible barriers in the equal-size setting become less isolated under overparameterization, with paths often attracted toward the global-minimum basin. These results suggest a mechanism by which overparameterization improves optimization in this model.

Keywords: Loss landscape, overparameterization, spurious local minima, teacher-student networks, mean-field dynamics.

Statement on Prior Dissemination

Parts of the material presented in this thesis have previously been disseminated in the arXiv preprint *Sharp description of local minima in the loss landscape of high-dimensional two-layer ReLU neural networks* (arXiv:2604.09412) by Jie Huang, Bruno Loureiro, and Stefano Sarao Mannelli.

Some mathematical formulations, figures, numerical results, and textual explanations in this thesis overlap with that preprint. The present thesis provides a more comprehensive presentation of the work in the format required for the Master's thesis project, including additional background, methodological details, discussion, and supporting material.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Stefano Sarao Mannelli, for his guidance, support, and patience throughout this thesis project. His help was particularly valuable when I encountered difficult technical questions, especially in the construction of the induced metric used in the string method analysis and in the derivation of the related mathematical formulas.

I would like to thank Bruno Loureiro for his valuable advice on the broader research direction and on the presentation of the results. His comments helped me clarify the structure of the work and better communicate the main findings.

I would like to thank my examiner, Ashkan Panahi, for his valuable feedback and for examining this thesis.

The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at Tetralith, partially funded by the Swedish Research Council through grant agreement no. 2022-06725, under project NAISS 2024/22-1082.

Finally, I would like to thank my parents for their continuous support and encouragement throughout my studies.

Jie Huang, Gothenburg, 2026-06-05

Contents

List of Figures

List of Tables

1

Introduction

Modern neural networks are trained by gradient-based methods on loss functions which are highly non-convex and whose structure depends on both the model architecture and the data distribution. However, large neural networks are often trained successfully in practice despite the difficulty posed by non-convexity, which indicates that overparameterization may significantly affect the geometry of the loss landscape. This motivates the study of how overparameterization affects the geometry of loss landscapes, including the structure of critical points and the organization of minima and saddles. Developing a principled understanding of these landscapes remains a central challenge in building a mathematical theory of deep learning.

Recent years have seen significant progress in this direction, especially in work on infinitely wide ReLU neural networks within the mean-field limit, where global convergence guarantees can be established [1, 2, 3, 4]. Nevertheless, these results do not clearly explain how the geometry of the loss landscape changes at finite width as overparameterization is increased. In particular, they provide limited guidance on how the structure of the loss landscape evolves with model size. To study this question, we consider the population loss landscape of teacher-student two-layer ReLU networks with Gaussian inputs. In this setting, Safran and Shamir [5] established the existence of spurious local minima and showed that increasing the width of the student network, i.e., overparameterization, makes global minima easier to reach. Related finite-width loss landscape phenomena have also been studied in [6, 7]. However, these works did not provide a systematic geometric description of the critical points and their structure.

This setting is closely related to the statistical-physics literature on teacher-student learning dynamics. Early work on soft committee machines introduced order-parameter descriptions of two-layer networks [8, 9], and more recent studies have developed rigorous and unified high-dimensional limits for such dynamics [10, 11, 12, 13, 14, 15]. These approaches provide a natural macroscopic description of learning, but they have mostly been used to analyze the time evolution of the loss rather than the detailed fixed-point structure in the loss landscape. Complementary geometric studies have emphasized the role of symmetry-breaking transitions, connectivity, and saddle-mediated dynamics in complex loss landscapes [16, 17, 18, 19, 20, 21]. Together, these lines of work motivate a geometric analysis of the loss landscape of finite-width teacher-student ReLU neural networks.

In this work, we develop a description of the population loss landscape based on a

summary-statistics representation in terms of order parameters, leading to a set of self-consistent ordinary differential equations governing the dynamics. This formulation allows us to study the high-dimensional gradient flow dynamics directly in order-parameter space. We use it to identify structured fixed-point families of the population loss and to investigate how these families change under overparameterization.

We first identify these families empirically through large-scale experiments, finding that the population loss landscape is organized into several discrete families of spurious local minima. Analysis of their order-parameter structure reveals a block structure in which student units split into aligned and anti-aligned groups relative to the teacher. Motivated by this observation, we introduce a block-structured ansatz and reduce the full fixed-point equations to a finite system of nonlinear equations. The numerical solutions of these equations give theoretical fixed points whose loss values and order-parameter patterns reproduce those observed in the ODE simulations.

Based on these theoretical fixed points, we compare two model size regimes: the well-specified case with equal-size student and teacher networks, and the overparameterized case in which the student network contains more neurons than the teacher. In the well-specified case, multiple spurious minima families coexist with the global minimum. Under overparameterization, the basin of attraction of the global minimum expands, the lowest-order spurious minima family disappears, and the remaining higher-order spurious minima families become substantially less prevalent. To probe the loss landscape between representative fixed-point configurations, we then use the string method [22, 23] as an exploratory tool. In the well-specified regime, the resulting paths show visible barriers between fixed points of the same spurious minima family. In contrast, in the overparameterized regime, the paths initialized between some high-order representative fixed points are often attracted toward the global minimum basin.

Together, these results provide a characterization of the population loss landscape of two-layer ReLU networks and suggest a mechanism by which overparameterization improves optimization in this setting.

2

Problem Formulation

2.1 Teacher-Student Neural Networks

We consider two-layer neural networks in the teacher-student framework, defined as follows. Each training example (x^μ, y^μ) consists of an input vector $x^\mu \in \mathbb{R}^d$ and an output scalar y^μ, where the components of x^μ are drawn i.i.d. from the standard normal distribution $\mathcal{N}(0,1)$. The target output is generated by a fixed *teacher* network with M hidden units as

$$y^\mu = \phi(x^\mu, \theta^*), \tag{2.1}$$

where $\theta^* = (v^* \in \mathbb{R}^M, w^* \in \mathbb{R}^{M\times d})$, and

$$\phi(x, \theta^*) = \sum_{m=1}^{M} v_m^*\, g\left(\frac{{w_m^*}^\top x}{\sqrt{d}}\right). \tag{2.2}$$

Here, the mth row of w^* is $(w_m^*)^\top$, where $w_m^* \in \mathbb{R}^d$ is regarded as a column vector, and $g(\cdot)$ denotes the activation function (e.g., sigmoid or ReLU). The *student* network, with K hidden units, produces the prediction

$$\phi(x, \theta) = \sum_{k=1}^{K} v_k\, g\left(\frac{{w_k}^\top x}{\sqrt{d}}\right), \tag{2.3}$$

where $\theta = (v \in \mathbb{R}^K, w \in \mathbb{R}^{K\times d})$. The kth row of w is $w_k^\top$, where each $w_k \in \mathbb{R}^d$ is regarded as a column vector.

2.2 Population loss and learning dynamics

The objective of the student network is to approximate the teacher function by minimizing the difference between their outputs. We define the *population loss* (or generalization error) as

$$\mathcal{L}(\theta; \theta^*) = \frac{1}{2}\,\mathbb{E}_x\left[\Big(\varphi(x, \theta) - \varphi(x, \theta^*)\Big)^2\right], \tag{2.4}$$

where the expectation is taken over the input distribution.

We consider gradient-flow learning on this population loss, where both layers of the student network are trained. Writing

$$\Delta(x) = \varphi(x, \theta) - \varphi(x, \theta^*), \tag{2.5}$$

the dynamics of the student parameters are given by

$$\dot{w}_k = -\eta_w \, \mathbb{E}_x \left[\Delta(x) \, v_k \, g' \left(\frac{w_k^\top x}{\sqrt{d}} \right) \frac{x}{\sqrt{d}} \right], \tag{2.6}$$

and

$$\dot{v}_k = -\eta_v \, \mathbb{E}_x \left[\Delta(x) \, g \left(\frac{w_k^\top x}{\sqrt{d}} \right) \right]. \tag{2.7}$$

Here, η_w and η_v respectively denote the learning rates for the first and second layer.

For the sake of completeness, we discussed the general two-layer network case, were both first layer (w) and second layer (v) are trained. In the special case where the second-layer weights are fixed, the model reduces to the soft committee machine (SCM), which will be considered later in our analysis.

2.3 Order Parameters

Following the notation used by Goldt et al.[10], we use indices $i, j, k = 1, \ldots, K$ to denote student hidden units and $m, n = 1, \ldots, M$ to denote teacher hidden units in this thesis.

To analyze the learning dynamics, we introduce a set of macroscopic variables (order parameters) that capture the alignments of the student and teacher weights. Following the statistical physics approach to neural network learning [10, 8], we define the overlap matrices

$$Q_{ij} = \frac{1}{d} w_i^\top w_j, \quad R_{im} = \frac{1}{d} w_i^\top w_m^*, \quad T_{mn} = \frac{1}{d} {w_m^*}^\top w_n^*. \tag{2.8}$$

Here, $Q \in \mathbb{R}^{K \times K}$ measures correlations between student units, $R \in \mathbb{R}^{K \times M}$ quantifies the alignment between student and teacher units, and $T \in \mathbb{R}^{M \times M}$ encodes the fixed correlations among teacher weights.

In addition, the second-layer weights v_k (student) and v_m^* (teacher) form part of the macroscopic description, as they directly modulate the network output and enter the learning dynamics. Together, the set of variables

$$m = (Q, R, T, v, v^*) \tag{2.9}$$

constitutes a complete set of order parameters for the system.

In the limit of infinite input dimension ($d \to \infty$), averages over the Gaussian input distribution can be expressed entirely in terms of these quantities. As a result, the high-dimensional learning dynamics of the weights reduce to a closed system

of deterministic ordinary differential equations governing the evolution of the order parameters (see Chap. 3).

For analytical simplicity, we consider orthonormal teacher weights in some theoretical analyses. In this case, $T = I_M$. This assumption simplifies the expressions without affecting the qualitative behavior of the system; extensions to general teacher configurations are discussed in Appendix A.

2.4 Parameterization regimes

We define parameterization regimes based on the relative width of the student and teacher networks. Let M and K denote the number of hidden units in the teacher and student networks, respectively:

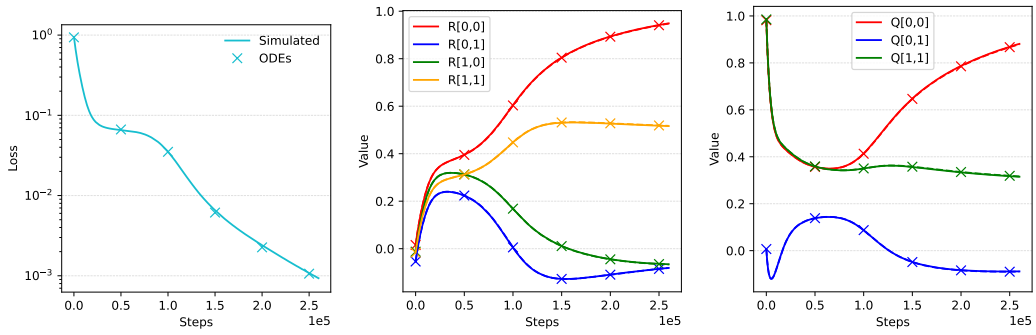
- **Well-specified regime:** $K = M$. In this case, the function class represented by the student network matches that of the teacher network.
- **Overparameterized regime:** $K > M$. In this case, the student network represents a strictly larger class of functions than the teacher network.

In the teacherstudent setting considered here, the complexity of the target function is explicitly controlled, making it natural to define overparameterization relative to the teacher.

3

Mean-field Dynamics

As shown in Fig. 3.1, the mean-field ODEs provide an accurate description of the loss dynamics observed in simulations of finite-width neural networks trained by gradient descent. We now provide the derivation of these equations.



(a) Training loss **(b)** Entries of R **(c)** Entries of Q

Figure 3.1: Comparison between mean-field ODEs predictions and network simulations. Panel (a) shows the training loss, while panels (b) and (c) display representative entries of the order parameters R and Q as functions of training steps. Solid lines denote simulation results and markers indicate ODE predictions.

3.1 Population Gradient

We consider training the network using Gradient Descent (GD), Normalized GD (nGD), or Orthonormalized GD (onGD) on the mean-squared error (MSE) loss function. In this section, we focus on the derivation for standard GD with ReLU activation functions. The derivations for nGD and onGD are detailed in Appendix C, and integrals for other activations (e.g., Leaky ReLU, erf) are provided in Appendix B.

For notational convenience, we denote expectations over the input distribution by

$$\langle f(x) \rangle \equiv \mathbb{E}_x[f(x)], \tag{3.1}$$

where the expectation is taken over the input distribution. So the population loss between the student and teacher networks can be written as:

$$\mathcal{L} = \frac{1}{2} \left\langle \left[\sum_{k=1}^{K} v_k g(\lambda_k) - \sum_{m=1}^{M} v_m^* g(\rho_m) \right]^2 \right\rangle, \tag{3.2}$$

where $\lambda_k = \frac{w_k^\top x}{\sqrt{d}}$ and $\rho_m = \frac{w_m^{*\top} x}{\sqrt{d}}$ are the pre-activations (also referred to as local fields). The continuous-time dynamics of the student weights w_i under gradient descent are given by:

$$\frac{dw_i}{dt} = -\eta \, \nabla_{w_i} \mathcal{L} = -\eta \, v_i \left\langle \Delta \, g'(\lambda_i) \frac{x}{\sqrt{d}} \right\rangle, \tag{3.3}$$

$$\frac{dv_i}{dt} = -\eta \nabla_{v_i} \mathcal{L} = -\eta \left\langle \Delta \, g(\lambda_i) \right\rangle, \tag{3.4}$$

where $\Delta = \phi(x) - \phi^*(x) = \sum_j v_j g(\lambda_j) - \sum_m v_m^* g(\rho_m)$ is the error signal, and η is the learning rate for weights.

3.2 Derivation of Dynamics

The pre-activations λ and ρ are jointly Gaussian random variables with zero mean. The covariance matrix Σ is fully determined by the order parameters defined in Eqs. 2.8:

$$Q_{ik} = \frac{w_i^\top w_k}{d} = \langle \lambda_i \lambda_k \rangle, \quad R_{in} = \frac{w_i^\top w_n^*}{d} = \langle \lambda_i \rho_n \rangle, \quad T_{nm} = \frac{w_n^{*\top} w_m^*}{d} = \langle \rho_n \rho_m \rangle.$$

To derive the mean-field ODEs, we apply the chain rule to the definition of the order parameters. For the student-student overlap $Q_{ik} = \frac{1}{d} w_i^\top w_k$:

$$\frac{dQ_{ik}}{dt} = \frac{1}{d} \left(\frac{dw_i}{dt}^\top w_k + w_i^\top \frac{dw_k}{dt} \right). \tag{3.5}$$

Substituting the weight dynamics from Eq. (3.3) into the first term, we obtain:

$$\frac{1}{d} \frac{dw_i}{dt}^\top w_k = -\frac{\eta}{d} v_i \left\langle \Delta \cdot g'(\lambda_i) \frac{x^\top w_k}{\sqrt{d}} \right\rangle = -\frac{\eta}{d} v_i \left\langle \Delta \cdot g'(\lambda_i) \lambda_k \right\rangle. \tag{3.6}$$

In the large-d limit, the dynamics evolve on a slow time scale of order $1/d$. By absorbing this factor $1/d$ into the continuous time variable (i.e., defining the macroscopic time step $dt \sim \frac{1}{d}$), we obtain the $O(1)$ dynamical equations:

$$\frac{dQ_{ik}}{dt} = -\eta v_i \langle \Delta g'(\lambda_i) \lambda_k \rangle - \eta v_k \langle \Delta g'(\lambda_k) \lambda_i \rangle, \tag{3.7}$$

where η denotes the effective learning rate. Equivalently, we can also absorb the full prefactor η/d into the time variable; this only changes the normalization of time and does not affect the resulting mean-field dynamics. A related discussion can be found in Arnaboldi et al. [15]. The second term follows by symmetry ($i \leftrightarrow k$). Expanding the error signal Δ, we arrive at terms of the form $\langle g'(\lambda_i) \lambda_k g(\lambda_j) \rangle$, which motivates the definition of the integral I_3. This formulation applies generally to two-layer

networks with activation function $g(\cdot)$ and hidden layer width not larger than $\mathcal{O}(d)$. See [15] for a detailed discussion.

Similarly, for the student-teacher overlap R_{in}:

$$\frac{dR_{in}}{dt} = \frac{1}{d}\frac{dw_i}{dt}^{\top} w_n^* = -\frac{\eta}{d} v_i \langle \Delta g'(\lambda_i)\rho_n \rangle. \tag{3.8}$$

After the same rescaling of time as above, this becomes

$$\frac{dR_{in}}{dt} = -\eta v_i \langle \Delta g'(\lambda_i)\rho_n \rangle. \tag{3.9}$$

Finally, for the second-layer weights v_i, the gradient descent dynamics are given by:

$$\frac{dv_i}{dt} = -\eta \left\langle \Delta \cdot g(\lambda_i) \right\rangle. \tag{3.10}$$

Substituting the explicit form of the error signal $\Delta = \sum_{j=1}^{K} v_j g(\lambda_j) - \sum_{m=1}^{M} v_m^* g(\rho_m)$, we obtain:

$$\frac{dv_i}{dt} = -\eta \left[\sum_{j=1}^{K} v_j \langle g(\lambda_j) g(\lambda_i) \rangle - \sum_{m=1}^{M} v_m^* \langle g(\rho_m) g(\lambda_i) \rangle \right]. \tag{3.11}$$

Here, the expectations are taken over the joint Gaussian distribution of the relevant local fields. This naturally leads to these integrals:

$$I_2(a,b) \equiv \langle g(a) g(b) \rangle, \tag{3.12}$$
$$I_3(a,b,c) \equiv \langle g'(a)\, b\, g(c) \rangle, \tag{3.13}$$

where a, b, c denote local fields chosen from the student fields $\{\lambda_i\}$ and teacher fields $\{\rho_m\}$.

3.3 Full Equations

In the equations below, we use a compact index notation to indicate which local fields enter each Gaussian expectation. For example,

$$I_2(i,j) = \langle g(\lambda_i) g(\lambda_j) \rangle, \qquad I_2(i,m) = \langle g(\lambda_i) g(\rho_m) \rangle, \tag{3.14}$$
$$I_3(i,n,m) = \langle g'(\lambda_i) \rho_n g(\rho_m) \rangle, \qquad I_3(i,k,j) = \langle g'(\lambda_i) \lambda_k g(\lambda_j) \rangle. \tag{3.15}$$

The corresponding covariance entries are determined by the order parameters Q, R, and T, which we will introduce in the next subsection. The mean-field ODEs

training for two-layer network are:

$$\frac{dR_{in}}{dt} = \eta v_i \left[\sum_{m=1}^{M} v_m^* I_3(i,n,m) - \sum_{j=1}^{K} v_j I_3(i,n,j) \right], \tag{3.16}$$

$$\begin{aligned} \frac{dQ_{ik}}{dt} &= \eta v_i \left[\sum_{m=1}^{M} v_m^* I_3(i,k,m) - \sum_{j=1}^{K} v_j I_3(i,k,j) \right] \\ &\quad + \eta v_k \left[\sum_{m=1}^{M} v_m^* I_3(k,i,m) - \sum_{j=1}^{K} v_j I_3(k,i,j) \right], \end{aligned} \tag{3.17}$$

$$\frac{dv_i}{dt} = \eta \left[\sum_{m=1}^{M} v_m^* I_2(i,m) - \sum_{j=1}^{K} v_j I_2(i,j) \right]. \tag{3.18}$$

3.4 ReLU Activation Function Case

For ReLU, the Gaussian integrals have closed-form expressions. We collect here the explicit formulas used in the dynamics; corresponding expressions for other activations are given in Appendix B.

3.4.1 Two-Variable Integral

This integral appears in the dynamics of the second-layer weights v.

$$I_2(x_1, x_2) = \langle g(x_1) g(x_2) \rangle. \tag{3.19}$$

For ReLU, when the variables (x_1, x_2) are jointly Gaussian with zero mean and covariance Σ_2, the explicit forms are given by:

$$I_2(x_1, x_2) = \frac{\sqrt{C_{11}C_{22} - C_{12}^2}}{2\pi} + \frac{C_{12}}{2\pi} \arccos\left(-\frac{C_{12}}{\sqrt{C_{11}C_{22}}} \right). \tag{3.20}$$

where $C_{ij} = \langle x_i x_j \rangle$ are the elements of the covariance matrix

$$\Sigma_2 = \begin{pmatrix} C_{11} & C_{12} \\ C_{12} & C_{22} \end{pmatrix},$$

In the mean-field ODEs of Sec.3.3, the notation $I_2(i,m)$ has the covariance matrix

$$\Sigma_2 = \begin{pmatrix} Q_{ii} & R_{im} \\ R_{mi} & T_{mm} \end{pmatrix}.$$

3.4.2 Three-Variable Integral

This integral appears in the dynamics of Q and R. We define I_3 specifically as the expectation involving a derivative, a multiplier, and an activation input:

$$I_3(x_0, x_1, x_2) = \langle g'(x_0)\, x_1\, g(x_2) \rangle. \tag{3.21}$$

For ReLU, we use $g'(x) = H(x) = \mathbf{1}_{x>0}$ (Heaviside step function) almost everywhere. When the variables (x_0, x_1, x_2) are jointly Gaussian with zero mean and covariance Σ_3, the analytic solution is:

$$I_3(x_0, x_1, x_2) = \frac{1}{2\pi}\left[C_{12}\arccos\left(-\frac{C_{02}}{\sqrt{C_{22}C_{00}}}\right) + \frac{C_{01}}{C_{00}}\sqrt{C_{22}C_{00} - C_{02}^2}\right], \tag{3.22}$$

where $C_{ij} = \langle x_i x_j \rangle$ are the elements of the covariance matrix

$$\Sigma_3 = \begin{pmatrix} C_{00} & C_{01} & C_{02} \\ C_{01} & C_{11} & C_{12} \\ C_{02} & C_{12} & C_{22} \end{pmatrix}.$$

In the mean-field ODEs of Sec. 3.3, the notation $I_3(i, n, m)$ has the covariance matrix

$$\Sigma_3 = \begin{pmatrix} Q_{ii} & R_{in} & R_{im} \\ R_{ni} & T_{nn} & T_{nm} \\ R_{mi} & T_{mn} & T_{mm} \end{pmatrix}.$$

These closed-form expressions allow the mean-field ODEs to be evaluated numerically without sampling, resulting in a deterministic low-dimensional dynamical system.

3.5 Loss Function in Terms of Order Parameters

Finally, utilizing the definition of the two-variable integral I_2, we can express the population loss $\mathcal{L}$ in a compact closed form. Expanding the squared error term and exchanging the summation and expectation operations, we obtain:

$$\begin{aligned}\mathcal{L} &= \frac{1}{2}\left\langle \left(\sum_{k=1}^{K} v_k g(\lambda_k) - \sum_{m=1}^{M} v_m^* g(\rho_m)\right)^2 \right\rangle \\ &= \frac{1}{2}\sum_{i,j=1}^{K} v_i v_j \langle g(\lambda_i) g(\lambda_j)\rangle + \frac{1}{2}\sum_{n,m=1}^{M} v_n^* v_m^* \langle g(\rho_n) g(\rho_m)\rangle - \sum_{k=1}^{K}\sum_{m=1}^{M} v_k v_m^* \langle g(\lambda_k) g(\rho_m)\rangle.\end{aligned} \tag{3.23}$$

Using the definition $I_2(a, b) = \langle g(a)g(b)\rangle$, and adopting the compact index notation, the loss function becomes:

$$\mathcal{L} = \frac{1}{2}\sum_{i,j=1}^{K} v_i v_j I_2(i, j) + \frac{1}{2}\sum_{n,m=1}^{M} v_n^* v_m^* I_2(n, m) - \sum_{k=1}^{K}\sum_{m=1}^{M} v_k v_m^* I_2(k, m). \tag{3.24}$$

3.6 Stochastic Gradient Descent

In the case of Stochastic Gradient Descent [9, 8], the ODE for Q contains an additional term resulting from the stochasticity of the update. The corresponding

mean-field equations can be written as:

$$\frac{dR_{in}}{dt} = \eta v_i \left[\sum_{m=1}^{M} v_m^* I_3(i,n,m) - \sum_{j=1}^{K} v_j I_3(i,n,j) \right], \tag{3.25}$$

$$\begin{aligned} \frac{dQ_{ik}}{dt} = {} & \eta v_i \left[\sum_{m=1}^{M} v_m^* I_3(i,k,m) - \sum_{j=1}^{K} v_j I_3(i,k,j) \right] \\ & + \eta v_k \left[\sum_{m=1}^{M} v_m^* I_3(k,i,m) - \sum_{j=1}^{K} v_j I_3(k,i,j) \right] \\ & + \eta^2 v_i v_k \Bigg[\sum_{m,n=1}^{M} v_m^* v_n^* I_4(i,k,n,m) - 2 \sum_{j=1}^{K} \sum_{m=1}^{M} v_j v_m^* I_4(i,k,m,j) \\ & + \sum_{j,l=1}^{K} v_j v_l I_4(i,k,j,l) \Bigg], \end{aligned} \tag{3.26}$$

$$\frac{dv_i}{dt} = \eta \left[\sum_{n=1}^{M} v_n^* I_2(i,n) - \sum_{j=1}^{K} v_j I_2(i,j) \right], \tag{3.27}$$

where the new term I_4 is defined as $I_4(a,b,c,d) \equiv \langle g'(a)g'(b)g(c)g(d) \rangle$ with a, b, c, d denoting local fields drawn from the student fields $\{\lambda_i\}$ and teacher fields $\{\rho_m\}$.

3.7 Equations of Fixed Points

We consider the case where the second-layer weights are uniform and not trained (soft committee machine). The stationary points of the population risk correspond to the zeros of the gradient flow. These satisfy a system of coupled non-linear equations governing the order parameters Q and R:

$$\mathcal{F}_R(Q,R) = 0, \qquad \mathcal{F}_Q(Q,R) = 0, \tag{3.28}$$

where the explicit forms are given by:

$$0 = \mathcal{F}_R(Q,R) = \frac{dR_{in}}{dt} = \sum_{m=1}^{M} I_3(i,n,m) - \sum_{j=1}^{K} I_3(i,n,j), \tag{3.29}$$

$$0 = \mathcal{F}_Q(Q,R) = \frac{dQ_{ik}}{dt} = \sum_{m=1}^{M} I_3(i,k,m) - \sum_{j=1}^{K} I_3(i,k,j) + \sum_{m=1}^{M} I_3(k,i,m) - \sum_{j=1}^{K} I_3(k,i,j). \tag{3.30}$$

These equations follow from the mean-field dynamics derived from the population loss in Eq. 3.24. As discussed in previous subsections, the dynamics depend on the weights only through the order parameters Q, R defined in Eqs. 2.8. Consequently, any stationary point of the population loss in Eq. 3.24 corresponds to a stationary point of the order parameter dynamics $\mathcal{F}_R(Q,R) = 0,\ \mathcal{F}_Q(Q,R) = 0$ for Gaussian inputs.

4

Landscape Characterization

In this chapter, we numerically integrate the mean-field ODEs for the order parameters Q and R in order to explore the distribution of loss values and the structure of the order parameters at convergence. These simulations provide a quantitative picture of the loss landscape across different student-teacher width pairs.

4.1 Simulation Setup

We consider multiple pairs of (K, M), corresponding to different numbers of student and teacher hidden units, respectively. For each width pair (K, M), we perform **10,000** independent simulations with different random initializations, evolving the mean-field ODEs for (Q, R) until convergence using sufficiently integration time. Each run produces a set of order parameters (Q, R, T) together with the corresponding population loss $\mathcal{L}$, which are recorded for further analysis.

To classify the resulting stationary states, we adopt the following criteria based on fixed-point conditions in Eqs. 3.28: If $\mathcal{L} < 10^{-3}$, the state is labeled as a **Global** minima. Otherwise, we compute the residuals (dQ, dR) from the final (Q, R, T). If $\max\{\|dQ\|_\infty, \|dR\|_\infty\} < 10^{-3}$, the state is labeled as a **Local** minima; otherwise, it is labeled as **None** minima. For each (K, M), we aggregate the outcomes across all 10,000 seeds.

4.2 Summary of minima statistics

We first verify that the mean-field ODE simulations reproduce known results obtained from Safran and Shamir [5]. Table 4.1a reports the proportions of global minima, local minima, and none minima in the well-specified setting $K = M$, while Table 4.1b presents the corresponding statistics in the overparameterized setting $K = M + 1$. The observed proportions are in close agreement with results reported by Safran and Shamir [5], supporting the validity of the mean-field description.

Consistent with their findings, we observe that in the well-specified setting the gradient flow frequently becomes trapped in local minima with non-vanishing generalization error. In contrast, in the overparameterized regime, the proportion of such local minima is significantly reduced, indicating an substantially improved optimization landscape.

(a) Results for case $K = M$.

K	M	Local (%)	Global (%)	None (%)
6	6	0.91	98.82	0.27
7	7	5.30	94.37	0.33
8	8	13.16	86.09	0.75
9	9	21.15	78.53	0.32
10	10	33.27	66.38	0.35
11	11	43.74	55.92	0.34
12	12	54.62	45.12	0.26
13	13	63.87	35.93	0.20
14	14	71.69	28.08	0.23
15	15	77.99	21.81	0.20
16	16	82.48	17.44	0.08
17	17	86.79	13.18	0.03
18	18	90.42	9.51	0.07
19	19	92.54	7.38	0.08
20	20	94.76	5.21	0.03

(b) Results for the overparameterized case $K = M + 1$.

K	M	Local (%)	Global (%)	None (%)
9	8	0.07	99.74	0.19
10	9	0.75	98.95	0.30
11	10	1.93	97.26	0.81
12	11	3.83	94.25	1.92
13	12	5.74	91.42	2.84
14	13	8.37	87.46	4.17
15	14	13.50	81.16	5.34
16	15	18.12	75.36	6.52
17	16	28.58	65.47	5.95
18	17	36.88	57.50	5.62
19	18	42.25	52.62	5.13
20	19	51.46	44.24	4.30

Table 4.1: Statistics of minima under mean-field dynamics. Proportions of local, global, and non-minima found by integrating the mean-field ODEs across 10,000 random seeds for each (K, M) pair.

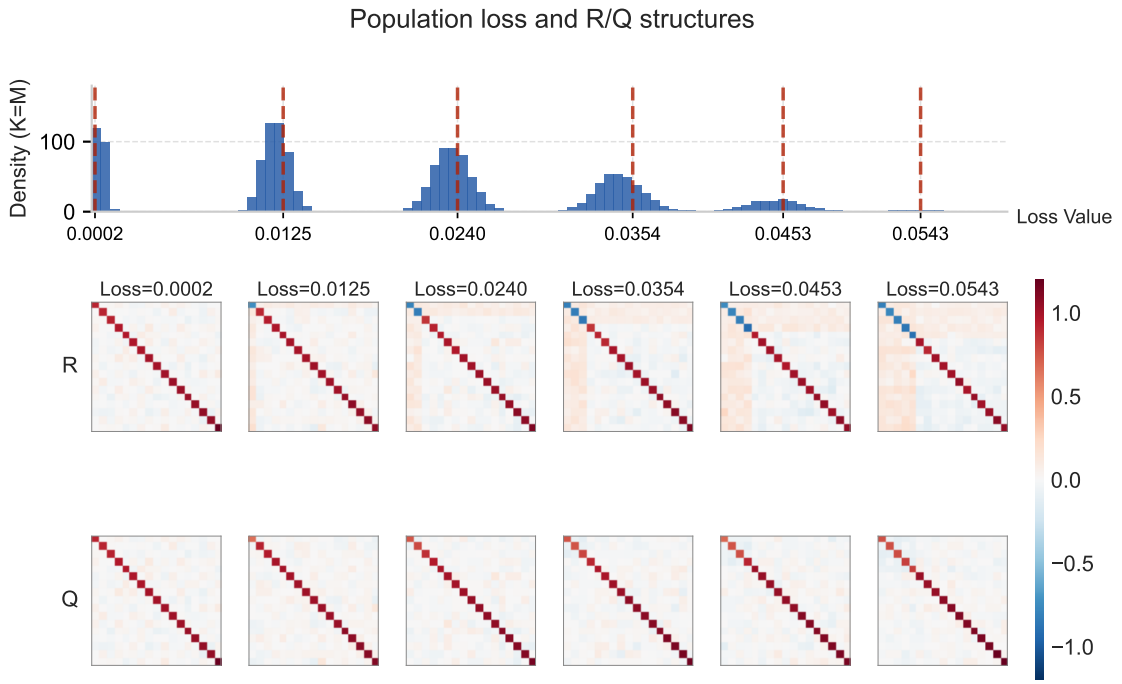


Figure 4.1: Loss Distribution and Corresponding Order Parameters. Final population loss histogram (top) for $(K, M) = (17, 17)$ with vertical lines at selected modes, and the corresponding R (middle) and Q (bottom) heatmaps from representative seeds at those losses within the dominant family. This histogram shows the distribution of population loss reached by gradient flow in the high-dimensional standard normal teacher initializations ($d = 784$ as the content in Appendix A).

4.3 Loss Distribution and Order Parameters

We next examine the distribution of loss values together with the corresponding structure of the order parameters. As shown in Figure 4.1, the distribution of final

loss values exhibits a quantized structure: rather than forming a continuum, the stationary states concentrate tightly around a discrete set of loss levels. This indicates that the loss landscape is organized into distinct families of spurious minima in our setting.

Each non-zero peak in the histogram corresponds to one such family, which is characterized by a specific pattern in the order parameters (Q, R). Near-zero loss is associated with an almost perfect teacher-student alignment, reflected by a strongly diagonal structure in R. At higher loss levels, a subset of diagonal entries of R becomes negative, indicating anti-alignment between certain student and teacher units. This is accompanied by a corresponding structure in the off-diagonal entries, suggesting a coordinated interaction between aligned and anti-aligned components.

These observations reveal that the spurious minima are not arbitrary, but instead exhibit a highly structured organization, which we exploit in the following section to construct a suitable ansatz.

5

Structured Ansatz of Order Parameters

In this chapter, we introduce a block-symmetric ansatz for the order parameters, motivated by the structured patterns observed in the simulations.

5.1 GMM Analysis

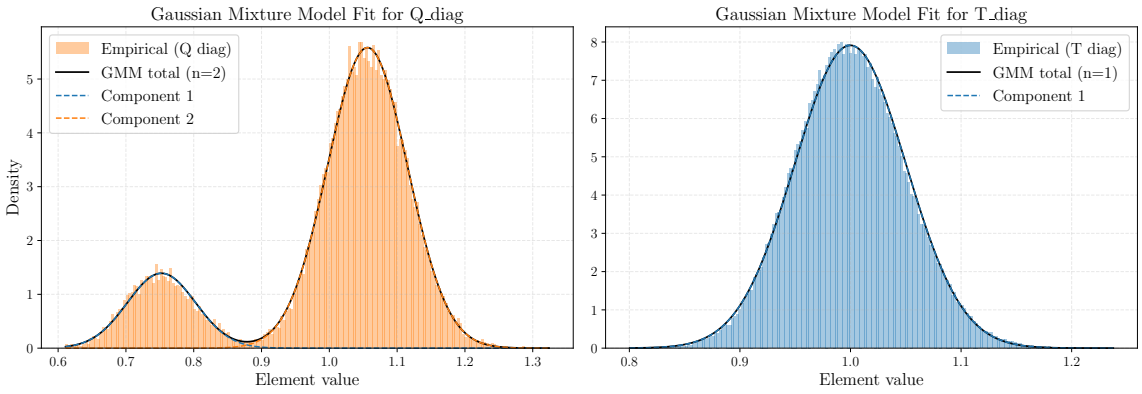


Figure 5.1: Gaussian mixture analysis of diagonal order parameters. The left panel shows the distribution of diagonal elements of the matrix Q, which is well described by a two-component Gaussian mixture model. The right panel shows the distribution of diagonal elements of the matrix T, which follows a single Gaussian distribution centered at 1. The data are obtained by grouping order parameters within the same minima family, re-arranging the corresponding (Q, R) matrices to maximize absolute values of diagonal entries of Q and extracting the diagonal entries.

To characterize the structure of the order parameters, we employ a Gaussian Mixture Model (GMM) analysis for the well-specified case ($K = M$) and the overparameterized case ($K = M + 1$), which represents a distribution as a weighted sum of Gaussian components and provides a simple tool for identifying clustered structure in empirical data.

We focus on a representative family of minima observed in the large-scale simulations (e.g., the same structure shown in Fig.4.1). Due to the permutation invariance of the hidden units, the ordering of neurons is not fixed. To obtain a consistent representation, we align the student units by reordering them to maximize the absolute values of the diagonal entries in the student-teacher overlap matrix R. After

alignment, we analyze the distributions of the diagonal elements of the matrices Q and T across the family of minima. We fit these distributions using a GMM to identify distinct populations of hidden units. The results shown in this subsection are obtained in the high-dimensional setting ($d = 784$) described in Appendix A, without imposing orthonormality on the teacher weights.

In Fig. 5.1, we show an example of this procedure for elements in Q and T. In Fig. 5.1 (Left), the diagonal elements of Q exhibit a clear bimodal distribution. A two-component GMM reveals two well-separated Gaussian components, whose relative weights closely match the ratio k_1/k_2 (with k_1 denoting the size of anti-aligned group and k_2 denoting the size of the aligned group). As a comparison, in Fig. 5.1 (Right) we apply the same GMM analysis to the diagonal elements of the teacher-teacher overlap matrix T. In contrast to Q, the distribution of T is well described by a single Gaussian component ($n = 1$) and centered around 1, consistent with the self-alignment in the teacher representation.

These results show that the student units separate into two distinct populations during training, providing strong empirical support for a block-structured organization of the order parameters and directly motivating the ansatz introduced in the following subsection.

5.2 Ansatz

We formalize this observation by introducing a block-symmetric ansatz for the order parameters (R, Q).

For the well-specified regime ($K = M$), We consider a partition of the hidden units into two groups of sizes k_1 and k_2 (with $k_1 + k_2 = K = M$), where k_1 denotes the number of student units that are anti-aligned with the teacher, and k_2 denotes the number of aligned units. Accordingly, the order parameters R and Q exhibit a block structure defined by:

$$R = \begin{pmatrix} \mathbf{D}_{k_1}(\varsigma, s) & \mathbf{B}(\varphi) \\ \mathbf{B}(f) & \mathbf{D}_{k_2}(b, \tau) \end{pmatrix}, \; Q = \begin{pmatrix} \mathbf{D}_{k_1}(q, e) & \mathbf{B}(u) \\ \mathbf{B}(u) & \mathbf{D}_{k_2}(p, \mu) \end{pmatrix} \tag{5.1}$$

and $T = I_K$, corresponding to the orthonormal teacher setting. The block operators correspond to the group partition dimensions: $\mathbf{D}_n(x, y)$ represents an $n \times n$ matrix with diagonal elements x and off-diagonal elements y, while $\mathbf{B}(z)$ represents a dense matrix of value z with dimensions conforming to the block position (e.g., $k_1 \times k_2$ for the off-diagonal blocks). Formally:

$$\mathbf{D}_n(x, y) \equiv (x - y)I_n + yJ_n, \quad \mathbf{B}(z) \equiv zJ, \tag{5.2}$$

where I_n is the identity matrix and J denotes the all-ones matrix of the requisite size.

For overparameterized case ($K = M + 1$), we treat the additional weights as unknown variables. Motivated by the empirical statistics in the previous subsection,

we propose an ansatz in $K = M + 1$ based on the structure of Q, R for $K = M$ before:

$$R = \begin{pmatrix} \mathbf{r}_0 \\ R_{\mathrm{K=M}} \end{pmatrix}, \quad Q = \begin{pmatrix} \Omega & \mathbf{q}_0 \\ \mathbf{q}_0^\top & Q_{\mathrm{K=M}} \end{pmatrix}, \tag{5.3}$$

where $R_{\mathrm{K=M}}$ and $Q_{\mathrm{K=M}}$ denote the $M \times M$ block matrices defined in Eq. 5.1. The scalar Ω represents the self-overlap of the extra unit. Crucially, the vectors $\mathbf{r}_0$ and $\mathbf{q}_0$ are block vectors partitioned to align with the group sizes k_1 and k_2 of the matched blocks. Formally:

$$\mathbf{q}_0 = (\mathbf{B}(v), \mathbf{B}(\gamma)), \quad \mathbf{r}_0 = (\mathbf{B}(\iota), \mathbf{B}(\kappa)). \tag{5.4}$$

Examples of the Ansatz. We illustrate the ansatz with explicit examples. For the case $k_1 = 2$ and $k_2 = 3$ in the well-specified regime ($K = M$), the matrices R and Q take the form:

$$R = \begin{pmatrix} \varsigma & s & \varphi & \varphi & \varphi \\ s & \varsigma & \varphi & \varphi & \varphi \\ f & f & b & \tau & \tau \\ f & f & \tau & b & \tau \\ f & f & \tau & \tau & b \end{pmatrix}, \quad Q = \begin{pmatrix} q & e & u & u & u \\ e & q & u & u & u \\ u & u & p & \mu & \mu \\ u & u & \mu & p & \mu \\ u & u & \mu & \mu & p \end{pmatrix}. \tag{5.5}$$

And for the case $k_1 = 2$ and $k_2 = 3$ in the overparameterized case ($K = M + 1$), the matrices R and Q take the form:

$$R = \begin{pmatrix} \iota & \iota & \kappa & \kappa & \kappa \\ \varsigma & s & \varphi & \varphi & \varphi \\ s & \varsigma & \varphi & \varphi & \varphi \\ f & f & b & \tau & \tau \\ f & f & \tau & b & \tau \\ f & f & \tau & \tau & b \end{pmatrix}, \quad Q = \begin{pmatrix} \Omega & v & v & \gamma & \gamma & \gamma \\ v & q & e & u & u & u \\ v & e & q & u & u & u \\ \gamma & u & u & p & \mu & \mu \\ \gamma & u & u & \mu & p & \mu \\ \gamma & u & u & \mu & \mu & p \end{pmatrix}. \tag{5.6}$$

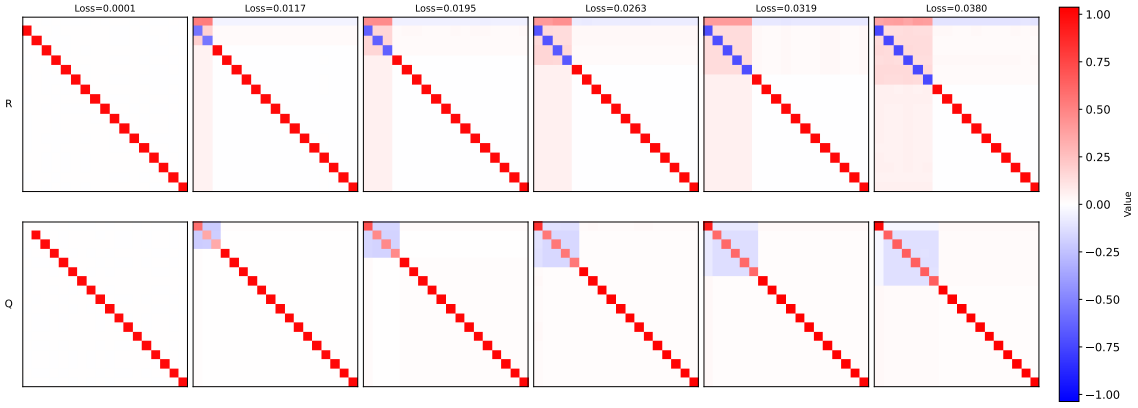


Figure 5.2: Different order parameters at $K = 18$ and $M = 17$. The plots show the values of R (first row) and Q (second row) for minima representative of the different families. From left to right, we see results for $k_1 = 0, 2, 3, 4, 5, 6$.

To provide a direct visual validation of the ansatz, we show a representative example of the order parameters (R^*, Q^*) obtained from simulations with $K = 18$ and $M = 17$ in Fig. 5.2. In this case, we adopt the orthonormal teacher setting described in Appendix A to improve the clarity of the block structure in Q and R.

6

Theoretical Fixed Points

In this chapter, we test the structured ansatz and compute theoretical fixed points. We substitute the ansatz into the fixed-point conditions, generate the resulting reduced equations symbolically, solve them numerically, and compare the theoretical loss values and order-parameter structures with empirical results obtained from ODE simulations.

6.1 Reduction to fixed-point equations

We now use the block-structured ansatz introduced in Chap. 5 to derive a reduced system of fixed-point equations. The goal is to test whether the structured order parameters observed in the ODE simulations correspond to genuine fixed points of the mean-field dynamics, rather than only empirical patterns.

6.1.1 Well-specified case

For the well-specified case $K = M$, we substitute the block ansatz in Eq. 5.1 into the fixed-point equations of the mean-field dynamics. The full stationary conditions are

$$\frac{dR_{in}}{dt} = 0, \qquad \frac{dQ_{ik}}{dt} = 0, \tag{6.1}$$

with the right-hand sides given by Eqs. 3.29 and 3.30. Before imposing the ansatz, these conditions are written for all entries of R and Q. Concretely, the ansatz contains five parameters for Q,

$$q,\ e,\ u,\ p,\ \mu,$$

and six parameters for R,

$$\varsigma,\ s,\ \varphi,\ f,\ b,\ \tau.$$

We therefore extract fixed-point equations for these parameters, denoted by

$$E_q = \frac{dq}{dt} = 0, \quad E_e = \frac{de}{dt} = 0, \quad E_u = \frac{du}{dt} = 0, \quad E_p = \frac{dp}{dt} = 0, \quad E_\mu = \frac{d\mu}{dt} = 0,$$

and

$$E_\varsigma = \frac{d\varsigma}{dt} = 0, \quad E_s = \frac{ds}{dt} = 0, \quad E_\varphi = \frac{d\varphi}{dt} = 0,$$
$$E_f = \frac{df}{dt} = 0, \quad E_b = \frac{db}{dt} = 0, \quad E_\tau = \frac{d\tau}{dt} = 0.$$

This gives a system of 11 coupled nonlinear algebraic equations for the 11 ansatz parameters. For completeness, the explicit equations for the well-specified case are reported in Appendix D.

6.1.2 Overparameterized case

The same reduction principle applies to the overparameterized case $K = M + 1$. In this setting, the ansatz contains the original $K = M$ block structure together with additional parameters $(\iota, \kappa, \Omega, v, \gamma)$ describing the extra student unit and its overlaps. Substituting this enlarged ansatz into the mean-field fixed-point conditions again produces a finite system of nonlinear algebraic equations. However, because the additional unit introduces several new covariance classes, the resulting expressions are substantially longer than in the well-specified case. We therefore do not report these equations explicitly.

6.2 Mathematica-assisted symbolic equation generation

The reduced fixed-point equations are lengthy, because each entry of the mean-field vector field contains sums of Gaussian expectations whose covariance entries depend on the ansatz parameters. To avoid deriving every equation manually, we used Wolfram Mathematica as a symbolic equation generator. Here we describe the procedure for the well-specified case $K = M$; the overparameterized case follows the same logic, with additional parameters for the extra student unit.

The main idea is to encode the ansatz directly at the level of indexed matrix entries. Instead of constructing full matrices Q and R for a fixed value of K, we first define symbolic functions $Q(i, j)$ and $R(i, n)$ whose values are determined by the positions of the indices in the two blocks. For example, diagonal entries, within-block off-diagonal entries, and cross-block entries are assigned to the corresponding ansatz parameters. This makes the subsequent derivation independent of the specific matrix size, except through the block sizes k_1 and k_2.

The second step is to construct each Gaussian expectation through its covariance entries. Each term in the fixed-point equations is of the form $I_3(x_0, x_1, x_2)$, where the three variables may be student or teacher local fields. We therefore define separate covariance maps for the different cases appearing in the equations, such as student-student-teacher terms and student-student-student terms in the equation for Q, and student-teacher-teacher and student-teacher-student terms in the equation for R. These covariance entries are then passed to the closed-form expression for I_3.

Finally, the summations over student and teacher indices are evaluated symbolically. The main technical point is that the summed index may coincide with one of the fixed indices, producing different covariance patterns. The Mathematica implementation handles these cases separately and multiplies each representative term by its multiplicity, such as $k_1 - 1$, $k_1 - 2$, or $k_2 - 1$. This yields symbolic expressions for the right-hand sides of dR_{in}/dt and dQ_{ik}/dt.

After the symbolic expressions for dR_{in}/dt and dQ_{ik}/dt have been constructed, we obtain the reduced equations by evaluating them at entries corresponding to each ansatz parameter. For example, in the first block, the parameter ς appears on the diagonal entries of R, so we use the stationary condition for $R_{1,1}$ to define $E_\varsigma = 0$. Similarly, s corresponds to the off-diagonal entries inside the first block of R, represented by $R_{1,2}$, while φ corresponds to cross-block entries such as R_{1,k_1+1}. For Q, the parameters q, e, and u are obtained analogously from $Q_{1,1}$, $Q_{1,2}$, and Q_{1,k_1+1}. The parameters in the second block are extracted in the same way, using $R_{k_1+1,1}$, R_{k_1+1,k_1+1}, R_{k_1+1,k_1+2}, Q_{k_1+1,k_1+1}, and Q_{k_1+1,k_1+2}. This gives the eleven reduced equations reported in Appendix D.

6.3 Numerical solution of the reduced equations

We solve the reduced fixed point equations using `FindRoot` in Wolfram Mathematica. Since the equations are nonlinear and may admit multiple solutions, the initialization is important. For each fixed value of (k_1, k_2), we initialize the solver with order parameters obtained from ODE simulations in the orthonormal teacher setting. These empirical solutions are already close to the corresponding fixed points and therefore provide stable initial guesses for the solver.

The solutions are first computed in Mathematica using high numerical precision. After convergence, we verify the result by substituting the solution back into the reduced equations and computing the maximum absolute residual,

$$\max_a |E_a|, \tag{6.2}$$

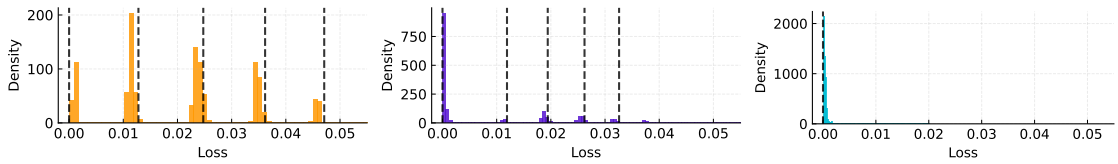
where E_a runs over the reduced fixed-point equations. Within Mathematica, the residuals are at the level expected from the chosen numerical precision.

For the subsequent simulations and visualizations, the fixed-point solutions are exported to JAX data structures in Python. Since JAX computations are performed in floating-point arithmetic, at most in float64 precision, this export introduces a finite-precision truncation. As a result, when the same fixed-point residuals are re-evaluated inside the JAX pipeline, the maximum residual may increase to around 10^{-6} in some cases. This discrepancy is a numerical precision effect rather than a failure of the fixed-point equations.

6.4 Comparison with empirical loss families

For each solution (R^*, Q^*) of the reduced equations, we evaluate the corresponding population loss. As shown in Fig. 6.1, the predicted loss values agree closely with the peaks of the empirical histograms for both the well-specified case $K = M$ and the overparameterized case $K = M + 1$. Minor discrepancies may arise due to differences in numerical precision and finite solver tolerances.

For the overparameterized case, Fig. 6.2 shows the order parameters (R^*, Q^*) obtained from the numerical solver for $K = 18$ and $M = 17$ across different values of



(a) $K = 17, M = 17$. **(b)** $K = 18, M = 17$. **(c)** $K = 19, M = 17$.

Figure 6.1: Loss families and theoretical values. Histograms of final loss values obtained from ODE dynamics at $M = 17$ for the well-specified case $K = M$ and the overparameterized case $K = M + 1$ and $K = M + 2$, starting from 10^4 random initializations in the orthonormal teacher setting. In (a) and (b), dashed vertical lines indicate the theoretical loss values obtained from the reduced fixed-point equations solved in Wolfram Mathematica.

k_1. The resulting structures and loss values closely match those observed directly from ODE simulations in Fig. 5.2.

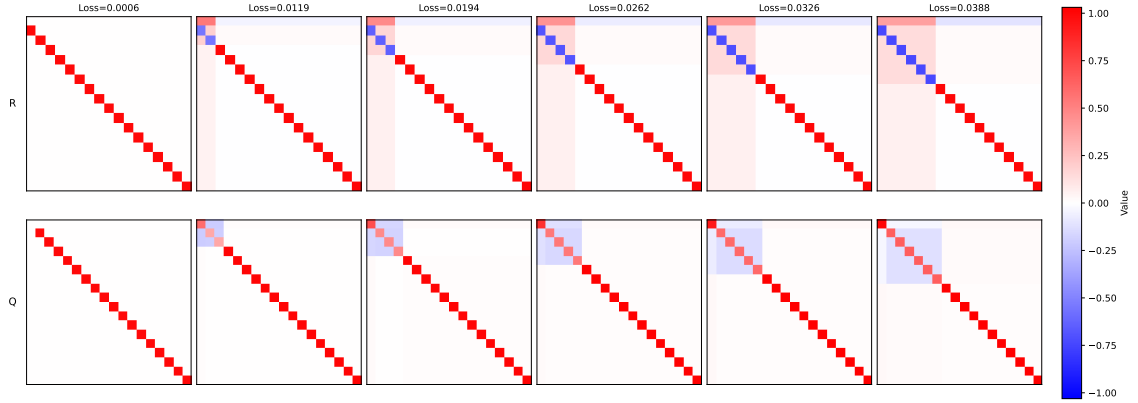


Figure 6.2: Order parameters obtained from solvers at $K = 18$ and $M = 17$. The plots show the values of R (first row) and Q (second row) for fixed points computed by the solver. Columns are ordered from left to right by the number of anti-aligned units, $k_1 = 0, 2, 3, 4, 5, 6$.

7

Exploratory Landscape Analysis with the String Method

7.1 String Method

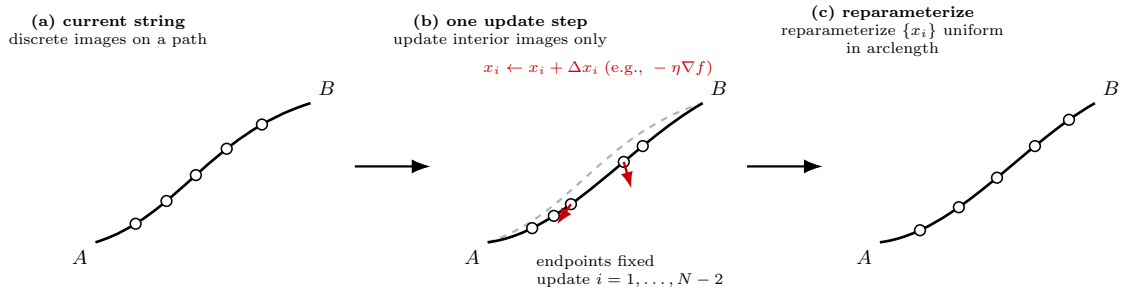


Figure 7.1: Illustration of the string method.

In this section, we introduce the string method [22, 23, 24], which we use as an exploratory numerical tool to probe loss landscape between distinct stationary configurations. This algorithm seeks the Minimum Energy Path (MEP) connecting two fixed configurations in the order parameter space. Conceptually similar to the Nudged Elastic Band method [25], this technique enables the study of low-loss trajectories between distinct points. We illustrate the string method in Fig. 7.1. Given two fixed endpoints, the string method approximates a continuous path in parameter space by a discrete set of images. These images evolve under gradient descent, and are periodically reparameterized to ensure approximately uniform spacing with respect to the chosen metric.

7.1.1 Induced Metric and Interpolation

A crucial requirement for the string method is a proper definition of distance to discretise the path. Since the gradient descent dynamics are defined on the weights $W \in \mathbb{R}^{K \times d}$, the geometry of the landscape is intrinsically Euclidean in the weight space. Therefore, the string evolution in the reduced order parameter space must rely on an induced metric that preserves this geometry.

Given the usage of the mean-field description based on the order parameters (sufficient statistics) $\Theta = (R, Q)$, we lose the information of the individual weights.

We define the distance between two macroscopic states Θ_1 and Θ_2 as the expected squared Euclidean distance between their typical microscopic realizations W_1 and W_2:

$$\mathcal{D}^2(\Theta_1, \Theta_2) = \mathbb{E}_{W_1,W_2}\left[\frac{1}{d}\|W_1 - W_2\|_F^2\right], \tag{7.1}$$

where the expectation is taken over the independent ensembles of weights consistent with the respective order parameters.

Expanding the Frobenius norm yields:

$$\frac{1}{d}\|W_1 - W_2\|_F^2 = \mathrm{Tr}\left(\frac{W_1W_1^\top}{d} + \frac{W_2W_2^\top}{d} - \frac{W_1W_2^\top}{d} - \frac{W_2W_1^\top}{d}\right). \tag{7.2}$$

By definition of the order parameters, the self-overlap terms converge to their expectations:

$$\mathbb{E}\left[\frac{1}{d}W_iW_i^\top\right] = Q_i. \tag{7.3}$$

To evaluate the cross-term $W_1W_2^\top$, we decompose the weights into a signal component (parallel to the teacher subspace) and a noise component (orthogonal to the teacher):

$$W = W_\parallel + W_\perp = RW^* + \Xi, \tag{7.4}$$

where $W^* \in \mathbb{R}^{M\times d}$ are the teacher weights and Ξ represents the orthogonal noise. Substituting this decomposition into the cross-term:

$$\frac{1}{d}W_1W_2^\top = \frac{1}{d}(R_1W^* + \Xi_1)(R_2W^* + \Xi_2)^\top = R_1\left(\frac{W^*W^{*\top}}{d}\right)R_2^\top + \text{noise cross-terms}. \tag{7.5}$$

We assume that W_1 and W_2 are drawn as independent samples from their respective macroscopic ensembles. Under this assumption, the noise components Ξ_1 and Ξ_2 are statistically uncorrelated. In the limit $d \to \infty$, the inner product of independent random vectors vanishes: $\frac{1}{d}\Xi_1\Xi_2^\top \to 0$. Furthermore, assuming orthonormal teacher weights ($T = \frac{1}{d}W^*W^{*\top} = I$), the expectation simplifies to:

$$\mathbb{E}\left[\frac{1}{d}W_1W_2^\top\right] \approx R_1R_2^\top. \tag{7.6}$$

Substituting these expectations back into the distance equation, we obtain the metric employed in our analysis:

$$\mathrm{dist}(\Theta_1, \Theta_2) = \sqrt{\mathrm{Tr}\left(Q_1 + Q_2 - R_1R_2^\top - R_2R_1^\top\right)}. \tag{7.7}$$

We emphasise that the independence assumption used to derive Eq. 7.7 implies that the microscopic noise components Ξ are uncorrelated. While valid for random initialisations in high dimensions, this assumption does not strictly hold for consecutive points along a continuous trajectory, where the noise components are necessarily correlated. However, the Minimum Energy Path is a geometric object that is invariant under reparameterisation of the curve. The specific choice of metric serves only to distribute the discretisation points (images) along the string. Therefore, Eq. 7.7 constitutes a valid parameterisation choice that respects the global Euclidean scaling of the problem, even if it overestimates local distances in the small-step limit.

7.1.2 Algorithm

The string is discretized into $P+1$ images $\{\Theta_0, \Theta_1, \ldots, \Theta_P\}$, where $\Theta_i = (Q_i, R_i)$. The endpoints Θ_0 and Θ_P are fixed at the known minima in our case. The algorithm proceeds iteratively:

1. **Evolution:** Each interior image Θ_i $(i = 1, \ldots, P-1)$ evolves independently for a fictitious time step $\Delta\tau$ according to the mean-field dynamics derived in Sec. 3.3:
$$\Theta_i \leftarrow \Theta_i + \Delta\tau \cdot \mathcal{F}(\Theta_i), \tag{7.8}$$
where $\mathcal{F}(\Theta) = (\dot{Q}, \dot{R})$ denotes the vector field of the order parameters (defined by Eqs. 3.16 and Eqs. 3.17). This step drives the string towards the valley floor (local stationarity).

2. **Reparameterization:** To prevent the images from collapsing into the minima, we redistribute them to maintain uniform spacing. We first compute the total arc-length L of the current string using the metric in Eq. 7.7. Then, we construct a cubic spline interpolation of the path and resample $P+1$ new images at equidistant arc-length coordinates $s_k = \frac{k}{P}L$ (for $k = 0, \ldots, P$).

3. **Termination:** The algorithm is iterated for a sufficient number of steps to ensure the string relaxes to a stationary state. Upon convergence, the transverse forces vanish, and the resulting curve represents the Minimum Energy Path.

7.2 Numerical Sensitivity of String Method

The string method is a useful numerical tool for approximating minimum energy paths between stationary configurations, but its implementation in the reduced order-parameter space depends sensitively on numerical details. In our experiments, the two most relevant issues are the precision truncation of the theoretical endpoints after exporting them from Mathematica to JAX, and the discretization of the path into a chosen number of images.

7.2.1 Precision and Endpoint Relaxation

The endpoints used in the string method are theoretical fixed points obtained from the reduced fixed-point equations in Chap. 6. By definition, a fixed point of the order-parameter dynamics is a zero of the vector field, i.e., of the right-hand side of the order-parameter ODEs,

$$F_Q(Q, R) = 0, \qquad F_R(Q, R) = 0.$$

The solutions computed in Mathematica satisfy these equations in high numerical precision. In practice, we also monitor the fixed-point residual by

$$\mathrm{Res}(Q, R) = \max\left\{\|F_Q(Q, R)\|_\infty, \|F_R(Q, R)\|_\infty\right\}.$$

After exporting the solutions to the JAX implementation, their numerical representation is truncated to the precision used by the Python pipeline. When the same residual is evaluated again in JAX using `float64`, it is therefore not exactly zero and can be of order 10^{-6} in some cases.

This residual measures the size of the gradient flow, not just a small error in the loss value. Therefore, even if the population loss of the exported endpoint is almost unchanged, a nonzero residual can still induce a small drift under the string evolution. This effect is particularly relevant near shallow basins, where a weak but nonzero flow may move the string away from the intended local neighborhood.

To reduce this effect, we apply an endpoint relaxation step before running the string method. Starting from the exported theoretical fixed point, we evolve the endpoint in the JAX `float64` implementation using the same ODEs as in the mean-field dynamics. The relaxation is stopped once both the fixed-point residual and the loss variation reach a stable numerical plateau. In a typical run, after relaxation we obtain values such as

$$L = 1.1697708328661349 \times 10^{-2}, \qquad \mathrm{Res}(Q, R) = 1.704 \times 10^{-9},$$

with

$$\|F_Q(Q, R)\|_\infty = 1.704 \times 10^{-9}, \qquad \|F_R(Q, R)\|_\infty = 1.405 \times 10^{-9},$$

and a loss change of approximately 7.951×10^{-10} over the monitored iteration interval. We regard this as numerically stable for the purpose of fixing the string endpoints.

The relaxed configurations are then used as the fixed endpoints of the string. This procedure does not change the definition of the theoretical fixed points; it only ensures that the endpoints used by the numerical string solver are better adapted to the implemented JAX dynamics.

7.2.2 Image Number and Reparameterization

The second numerical issue is the discretization of the path. The string method represents a continuous curve by a sequence of images, and the number of images determines the resolution of this discretized curve.

If too few images are used, the string may fail to show curved valleys or narrow transition regions between the endpoints. In this case, the discretized path can effectively cut across the landscape and miss the relevant minimum energy path. On the other hand, increasing the number of images is not always beneficial. When neighboring images become very close, small errors introduced by the evolution step and by the reparameterization step can accumulate. In flat regions, shallow basins, or narrow transition channels, these small perturbations may change the qualitative behavior of the string.

The reparameterization step is necessary to maintain a reasonable distribution of images along the path. However, it also introduces interpolated order-parameter configurations. These interpolated configurations are not obtained by following an actual weight-space trajectory, and their behavior depends on the induced metric

and interpolation rule used in the reduced space. In our implementation, the string is reparameterized very frequently, typically after each evolution step. This choice prevents image collapse, but it also means that interpolation errors are introduced repeatedly during the computation. As a result, the reparameterization frequency itself can become a source of numerical sensitivity.

In the following section, we therefore treat the string-method outputs as exploratory diagnostics rather than rigorous evidence for the exact loss landscape.

7.3 Landscape Geometry Observations

As illustrated in Fig. 6.1a, in the well-specified setting a significant fraction of trajectories remain trapped in the high-loss families. In contrast, the addition of even a single extra neuron ($K = M + 1$), Fig. 6.1b, dramatically expands the basin of attraction of the global minimum, although nonzero-loss fixed-point families can still be observed. These empirical results indicate that overparameterization changes the stability and reachability of the spurious families identified in the previous chapters.

In this section, we use the string method as an exploratory tool to probe the geometry of the loss landscape between representative fixed point configurations. In light of the numerical sensitivity discussed above, we do not interpret the resulting paths as rigorous evidence for the exact loss landscape. Instead, we use them to compare the qualitative path behavior in the well-specified and overparameterized regimes.

7.3.1 Well-specified Regime

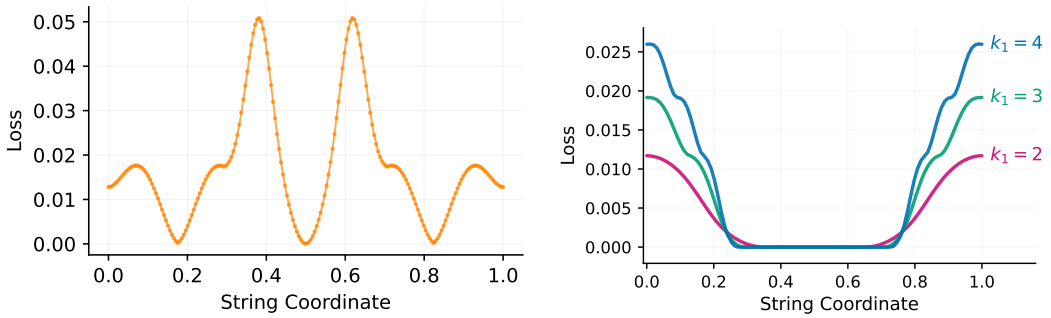
We first consider the well-specified case ($K = M$). Due to permutation symmetry, there exist multiple discrete global minima. We construct a path between two theoretical fixed-point configurations with the same loss value. The first endpoint is the fixed point (R^*, Q^*) obtained from the reduced equations in Chap. 6. The second endpoint is generated by applying a permutation of the student hidden units to this fixed point (R^*, Q^*). If P denotes the corresponding permutation matrix, the permuted endpoint is given by

$$R_{\text{perm}} = PR^*, \qquad Q_{\text{perm}} = PQ^*P^\top.$$

Since the population loss is invariant under permutations of the student hidden units, the two endpoints are distinct representations of the same fixed-point family with same loss value. As shown by the orange curve in Fig. 7.2a, the computed string path exhibits clear barriers between the two endpoints. This indicates that moving between these two fixed points requires passing through regions of higher population loss. We observe the same qualitative behaviour for all permutations examined.

7.3.2 Overparameterized Regime

We now extend the analysis to the overparameterized setting $K = M + 1$, following the same procedure as in the well-specified case. Before applying the string method,



(a) Loss along the string path in $K = M$. **(b) Loss along the string path in $K = M + 1$.**

Figure 7.2: String-method path diagnostics. Left: In the well-specified regime $K = M$, the string connecting two permuted representatives of the same fixed-point family crosses visible loss barriers. **Right:** In the overparameterized regime $K = M+1$, strings between representative fixed points are attracted toward the global-minimum basin. Colors indicate families indexed by k_1, the number of anti-aligned units.

Minima / *Order*	$K = 17$ $M = 17$	$K = 18$ $M = 17$	$K = 19$ $M = 17$
$k_1 = 0$	13.09%	59.29%	99.63%
$k_1 = 1$	27.52%	0.00%	0.00%
$k_1 = 2$	29.05%	2.10%	0.05%
$k_1 = 3$	18.94%	10.83%	0.31%
$k_1 = 4$	7.55%	8.99%	0%

Table 7.1: Observed fixed-point families percentage. Percentage of 10^4 random initializations that converge to minima with k_1 anti-aligned units in different (K, M); $k_1 = 0$ corresponds to the global minimum.

we first summarize which fixed-point families are observed in this regime from the reduced fixed-point equations and the ODE simulations.

The lowest-order spurious family, characterized by a single anti-aligned neuron($k_1 = 1$), disappears in the overparameterized setting considered here. We do not find stable solutions of the fixed-point equations for this family, and the ODE simulations support this observation: across all initializations in Table 7.1 and Fig. 6.1b, no trajectory converges to a state with a single anti-aligned direction. In contrast, higher-order spurious families with $k_1 \geq 2$ still appear in the simulations, although their frequencies are reduced. These surviving families provide the endpoints for the string-method diagnostics below.

For the surviving families with $k_1 \geq 2$, we next examine the string-method paths between representative configurations within the same k_1 family. In the well-specified case, the computed strings between permuted representatives crossed visible loss barriers. We therefore ask whether a similar barrier structure is also observed in the overparameterized regime.

The exploratory results, shown in Fig. 7.2b, indicate a different qualitative behavior. For several pairs of configurations, the interior images of the string are attracted toward the global minimum basin ($\mathcal{L} \approx 0$) rather than crossing barriers between the

endpoint losses.

This collapse of the interior images should not be interpreted as contradicting the existence of the endpoint fixed points. The endpoints are theoretical fixed points of the reduced equations and are further relaxed in the JAX implementation before being used in the string method. Rather, it suggests that these basins are narrow or fragile: small transverse departures from the endpoint neighborhood can lead the path toward the global minimum. This is consistent with the $k_1 = 4$ profile in Fig. 7.2b, where the images closest to the endpoints show small local loss peaks before the string descends toward zero loss. These peaks are hard to see at the scale of the figure but are present in the numerical values, suggesting that the endpoints retain only very narrow local basins. In this sense, the barriers that isolate the corresponding states appear weakened, with the surrounding landscape containing directions toward the global optimum. Although deterministic gradient flow can still converge to these spurious minima, their reduced isolation suggests an increased sensitivity to perturbations. This is qualitatively consistent with the idea that stochastic optimization methods such as SGD may be more likely than deterministic population gradient flow to leave these narrow basins, although this interpretation remains qualitative in the present work.

Thus, the string-method diagnostics suggest that overparameterization changes the landscape geometry around the fixed-point configurations. The high-loss fixed points can still be reached by gradient flow, but the diagnostics reveal a type of path behavior not observed in the well-specified regime: for several representative pairs, the interior images are attracted toward the global-minimum basin. Because of the numerical sensitivity discussed in Sec. 7.2, we interpret this as qualitative evidence for a reshaping of the landscape rather than as a rigorous statement about the basin structure of the loss landscape.

8

Conclusion

We analyzed the population-loss landscape of high-dimensional two-layer ReLU neural networks in the teacher-student setting through a closed set of macroscopic equations for the order parameters. This formulation reduces the high-dimensional gradient-flow dynamics to a deterministic system for the student-student and student-teacher overlaps, allowing us to study the learning dynamics directly in order-parameter space and to relate them to the geometry of the loss landscape using the string method.

In the well-specified regime ($K = M$), we found that gradient flow converges to several structured families of spurious local minima. These families are organized into discrete loss levels and exhibit block-structured order-parameter patterns, characterized by aligned and anti-aligned groups of student units. Based on these observations, we introduced a structured ansatz and used it to derive reduced fixed-point equations. The theoretical fixed points obtained from these equations reproduce the empirical loss families and their corresponding order-parameter structures. String-method diagnostics further indicate that representative configurations in this regime are separated by visible loss barriers.

Under mild overparameterization, the landscape changes qualitatively. The lowest-order spurious loss family is destabilized, the basin of attraction of the global minimum expands, and the frequencies of the remaining high-loss families are strongly reduced. Although some higher-order fixed-point families still survive, the string-method diagnostics suggest that they are less isolated than in the well-specified case: paths initialized between representative configurations are often attracted toward the global-minimum basin rather than crossing stable high-loss barriers.

These results indicate a mechanism through which overparameterization improves optimization in this model. Its effect is not only to remove certain spurious minima, but also to weaken the isolation of the remaining ones and create directions through which the dynamics can more easily approach the global solution. More generally, this work shows that combining mean-field order-parameter dynamics, structured fixed-point ansatzes, and exploratory geometric tools such as the string method can give a sharp and interpretable description of neural network loss landscapes.

Bibliography

[1] L. Chizat and F. Bach, "On the global convergence of gradient descent for over-parameterized models using optimal transport," *Advances in neural information processing systems*, vol. 31, 2018.

[2] S. Mei, A. Montanari, and P.-M. Nguyen, "A mean field view of the landscape of two-layer neural networks," *Proceedings of the National Academy of Sciences*, vol. 115, no. 33, E7665–E7671, 2018.

[3] G. Rotskoff and E. Vanden-Eijnden, "Trainability and accuracy of artificial neural networks: An interacting particle system approach," *Communications on Pure and Applied Mathematics*, vol. 75, no. 9, pp. 1889–1935, 2022.

[4] J. Sirignano and K. Spiliopoulos, "Mean field analysis of neural networks: A law of large numbers," *SIAM Journal on Applied Mathematics*, vol. 80, no. 2, pp. 725–752, 2020.

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

[6] I. M. Safran, G. Yehudai, and O. Shamir, "The effects of mild over-parameterization on the optimization landscape of shallow relu neural networks," in *Conference on Learning Theory*, PMLR, 2021, pp. 3889–3934.

[7] L. Venturi, A. S. Bandeira, and J. Bruna, "Spurious valleys in two-layer neural network optimization landscapes," *arXiv preprint arXiv:1802.06384*, 2018.

[8] M. Biehl and H. Schwarze, "Learning by on-line gradient descent," *Journal of Physics A: Mathematical and general*, vol. 28, no. 3, p. 643, 1995.

[9] D. Saad and S. Solla, "Dynamics of on-line gradient descent learning for multilayer neural networks," *Advances in neural information processing systems*, vol. 8, 1995.

[10] S. Goldt, M. Advani, A. M. Saxe, F. Krzakala, and L. Zdeborová, "Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup," *Advances in neural information processing systems*, vol. 32, 2019.

[11] S. Goldt, M. Mézard, F. Krzakala, and L. Zdeborová, "Modeling the influence of data structure on learning in neural networks: The hidden manifold model," *Physical Review X*, vol. 10, no. 4, p. 041 044, 2020.

[12] G. Ben Arous, R. Gheissari, and A. Jagannath, "High-dimensional limit theorems for sgd: Effective dynamics and critical scaling," *Advances in neural information processing systems*, vol. 35, pp. 25 349–25 362, 2022.

[13] R. Veiga, L. Stephan, B. Loureiro, F. Krzakala, and L. Zdeborová, "Phase diagram of stochastic gradient descent in high-dimensional two-layer neural networks," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 244–23 255, 2022.

[14] L. Arnaboldi, Y. Dandi, F. Krzakala, B. Loureiro, L. Pesce, and L. Stephan, "Online learning and information exponents: The importance of batch size amp; Time/Complexity tradeoffs," in *Proceedings of the 41st International Conference on Machine Learning*, R. Salakhutdinov et al., Eds., ser. Proceedings of Machine Learning Research, vol. 235, PMLR, 21–27 Jul 2024, pp. 1730–1762.

[15] L. Arnaboldi, L. Stephan, F. Krzakala, and B. Loureiro, "From high-dimensional & mean-field dynamics to dimensionless odes: A unifying approach to sgd in two-layers networks," in *The Thirty Sixth Annual Conference on Learning Theory*, PMLR, 2023, pp. 1199–1227.

[16] K. Fukumizu and S.-i. Amari, "Local minima and plateaus in hierarchical structures of multilayer perceptrons," *Neural networks*, vol. 13, no. 3, pp. 317–327, 2000.

[17] Y. Arjevani and M. Field, "On the principle of least symmetry breaking in shallow relu models," *arXiv preprint arXiv:1912.11939*, 2019.

[18] J. Brea, B. Simsek, B. Illing, and W. Gerstner, "Weight-space symmetry in deep networks gives rise to permutation saddles, connected by equal-loss valleys across the loss landscape," *arXiv preprint arXiv:1907.02911*, 2019.

[19] B. Simsek et al., "Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances," in *International Conference on Machine Learning*, PMLR, 2021, pp. 9722–9732.

[20] L. F. Cugliandolo and J. Kurchan, "Analytical solution of the off-equilibrium dynamics of a long-range spin-glass model," *Physical Review Letters*, vol. 71, no. 1, p. 173, 1993.

[21] L. F. Cugliandolo and J. Kurchan, "On the out-of-equilibrium relaxation of the sherrington-kirkpatrick model," *Journal of Physics A: Mathematical and General*, vol. 27, no. 17, p. 5749, 1994.

[22] E. Weinan, W. Ren, and E. Vanden-Eijnden, "String method for the study of rare events," *Physical Review B*, vol. 66, no. 5, p. 052 301, 2002.

[23] W. Ren, E. Vanden-Eijnden, et al., "Simplified and improved string method for computing the minimum energy paths in barrier-crossing events," *The Journal of chemical physics*, vol. 126, no. 16, 2007.

[24] A. Samanta and E. Weinan, "Optimization-based string method for finding minimum energy path," *Communications in Computational Physics*, vol. 14, no. 2, pp. 265–275, 2013.

[25] F. Draxler, K. Veschgini, M. Salmhofer, and F. Hamprecht, "Essentially no barriers in neural network energy landscape," in *International conference on machine learning*, PMLR, 2018, pp. 1309–1318.

A

Robustness to Finite Input Dimensionality

In the main text, our analysis of the fixed points we assume orthonormal teacher configuration ($T = I_M$) for simplifying the equations. While this is exact in the thermodynamic limit ($d \to \infty$), in this section, we investigate the robustness of our predictions to finite input dimensions d. We simulate the dynamics starting from actual neural network weight initialisations at dimensions $d \in \{196, 392, 784\}$.

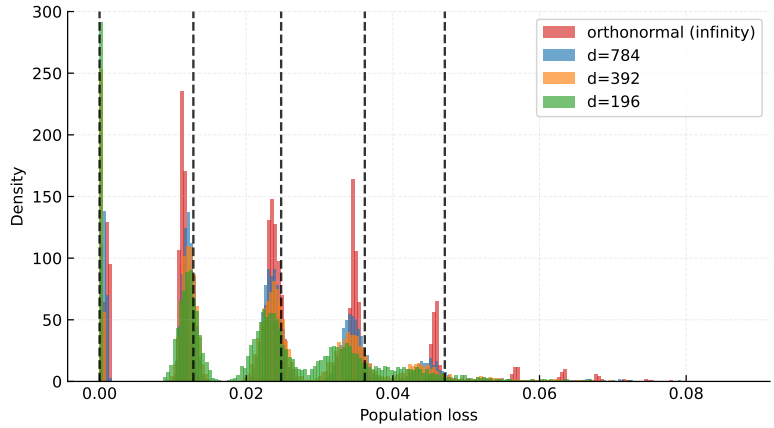


Figure A.1: Impact of finite dimensionality on the loss distribution. Histograms of the final population loss obtained by simulating the ODEs starting from actual neural network weight initialisations at dimensions $d \in \{196, 392, 784\}$, compared against the infinite-dimensional limit (orthonormal teacher, $T = \mathbb{I}$). Dashed vertical lines indicate the corresponding theoretically predicted loss values While smaller dimensions introduce a broadening effect due to variance, the discrete families of minima remain robustly centered around the predicted macroscopic levels.

Fig. A.1 illustrates the impact of finite dimensionality on the final population loss distribution. We compare the empirical histograms obtained from finite d simulations against the idealised infinite-dimensional case. While the discrete, quantised hierarchy of the local minima is strictly preserved, finite dimensions introduce variance. This results in a progressive broadening of the loss distribution peaks around the theoretically predicted values. As expected, the empirical distributions tightly converge toward the analytical infinite-dimensional limit as d increases.

To further validate that the structural symmetries discussed in our ansatz persist beyond the idealised orthonormal teacher scenario, Fig.A.2 visualises the student-

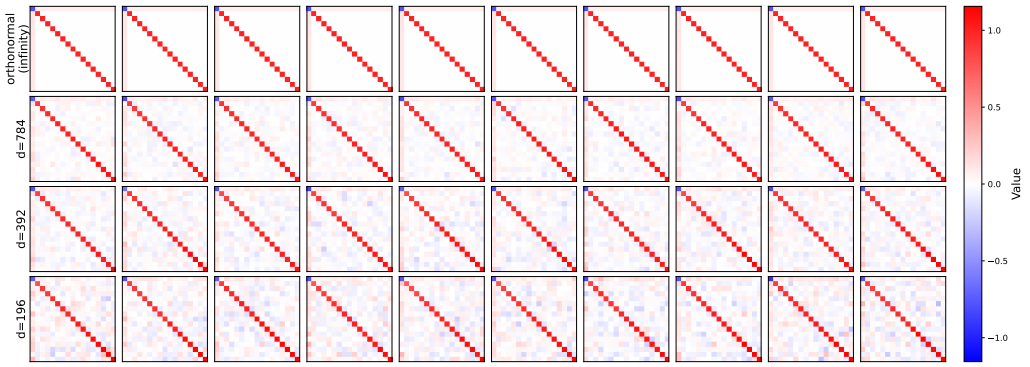


Figure A.2: Student-teacher overlap matrix (R) across different dimensions. Heatmaps of the R matrix for 10 randomly sampled configurations converging to the first-order local minimum ($k_1 = 1$). The top row represents the idealised $d \to \infty$ limit, while subsequent rows correspond to finite dimensions $d = 784, 392,$ and 196. The block-symmetric structure predicted by our theoretical ansatz clearly emerges across all regimes, with finite-size fluctuations increasing at lower dimensions.

teacher overlap matrix R. We uniformly sample 10 configurations at the end of the dynamics from the first-order local minimum (i.e., the second quantised group corresponding to $k_1 = 1$, as observed in Fig.A.1). Despite the noise introduced by the finite-dimensional teacher, the distinct macroscopic block structure, including the aligned and anti-aligned unit clusters, remains clearly identifiable. As d decreases, the off-diagonal fluctuations become more pronounced, yet the fundamental geometric organization in blocks remains robust. This confirms that our ansatz accurately reflects the empirical behaviour of finite-dimensional networks.

B

Alternative Activation Functions

In this section, we provide the analytical expressions for the Gaussian integrals I_2 and I_3 for Leaky ReLU and Error Function (erf) activations. Based on these derivations, we also present numerical observations of the learning landscape and spurious minima structures for these activations in this section. The notation follows the definitions in Chap. 3:

$$I_2(x_1, x_2) = \langle g(x_1)g(x_2)\rangle, \tag{B.1}$$
$$I_3(x_0, x_1, x_2) = \langle g'(x_0)x_1 g(x_2)\rangle, \tag{B.2}$$

where the expectations are taken over a multivariate Gaussian distribution with covariance matrix Σ.

B.1 Leaky ReLU

The Leaky ReLU activation function with a leakage parameter $\alpha \in [0, 1]$ is defined as:

$$g(x) = \mathrm{LReLU}(x; \alpha) = \begin{cases} x & \text{if } x \geq 0, \\ \alpha x & \text{if } x < 0. \end{cases} \tag{B.3}$$

Its derivative is the generalized step function $g'(x) = \mathbf{1}_{x\geq 0} + \alpha \mathbf{1}_{x<0}$ where the derivative at $x = 0$ is defined by convention. Note that setting $\alpha = 0$ recovers the standard ReLU results, while $\alpha = 1$ corresponds to a linear network.

B.1.1 Two-Variable Case

The explicit form for the correlation between two Leaky ReLU units is:

$$I_2^{\mathrm{LReLU}}(\Sigma) = \alpha C_{12} + \frac{(1-\alpha)^2}{2\pi}\left[\sqrt{C_{11}C_{22} - C_{12}^2} + C_{12}\arccos\left(-\frac{C_{12}}{\sqrt{C_{11}C_{22}}}\right)\right], \tag{B.4}$$

where C_{11}, C_{22} are the variances and C_{12} is the covariance of (x_1, x_2).

B.1.2 Three-Variable Case

The integral involving the derivative, a multiplier, and the activation is given by:

$$I_3^{\mathrm{LReLU}}(\Sigma) = \alpha C_{12} + \frac{(1-\alpha)^2}{2\pi}\left[C_{12}\arccos\left(-\frac{C_{02}}{\sqrt{C_{22}C_{00}}}\right) + \frac{C_{01}}{C_{00}}\sqrt{C_{22}C_{00} - C_{02}^2}\right]. \tag{B.5}$$

Here, the covariance matrix for the joint variables (x_0, x_1, x_2) is:

$$\Sigma = \begin{pmatrix} C_{00} & C_{01} & C_{02} \\ C_{01} & C_{11} & C_{12} \\ C_{02} & C_{12} & C_{22} \end{pmatrix}.$$

B.2 Results for Leaky ReLU

Based on the derived integrals, we numerically investigate the stationary points of networks with leaky ReLU activation. With the Leaky ReLU parameter set to $\alpha = 0.01$, the Fig. B.1 shows the results for $K = M = 12$ under nGD. It can be clearly observed that anti-aligned units appear, similar to the ReLU case.

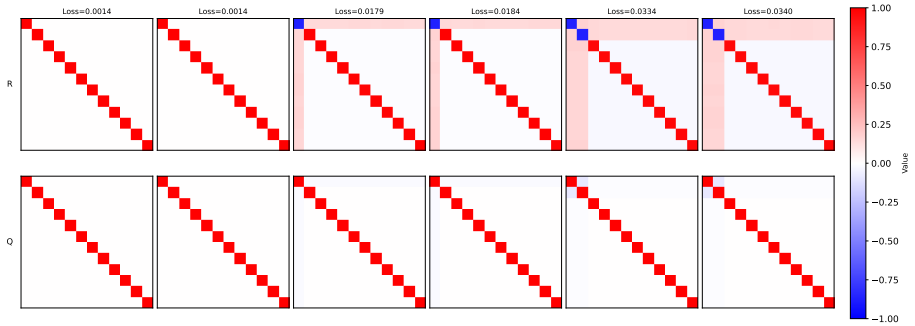


Figure B.1: Order Parameters of Leaky ReLU in $K = M = 12$.

B.3 Sigmoidal (erf)

Consider the sigmoidal activation function defined by the error function:

$$g(x) = \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right). \tag{B.6}$$

The derivative is proportional to a Gaussian: $g'(x) = \sqrt{\frac{2}{\pi}}e^{-x^2/2}$.

B.3.1 Two-Variable Case

Based on the arcsine law for Gaussian integrals, the result is:

$$I_2^{\mathrm{erf}}(\Sigma) = \frac{2}{\pi}\arcsin\left(\frac{C_{12}}{\sqrt{(1+C_{11})(1+C_{22})}}\right). \tag{B.7}$$

B.3.2 Three-Variable Case

The integral involving the derivative of the error function allows for an analytical solution:

$$I_3^{\text{erf}}(\Sigma) = \frac{2}{\pi\sqrt{(1+C_{00})(1+C_{22}) - C_{02}^2}} \left(C_{12} - \frac{C_{01}C_{02}}{1+C_{00}} \right). \tag{B.8}$$

B.4 Results for Sigmoidal (erf) Activation

Based on the derived integrals, we numerically investigate the stationary points of networks with erf activation by running the ODEs for a few seeds. Fig. B.2 illustrates the structure of the order parameters for $K = M = 12$ across different local minima found by nGD. Notice the absence of anti-aligned neurons in sharp contrast with what is observed in the Main Text for ReLU and in Fig.B.1 for Leaky ReLU.

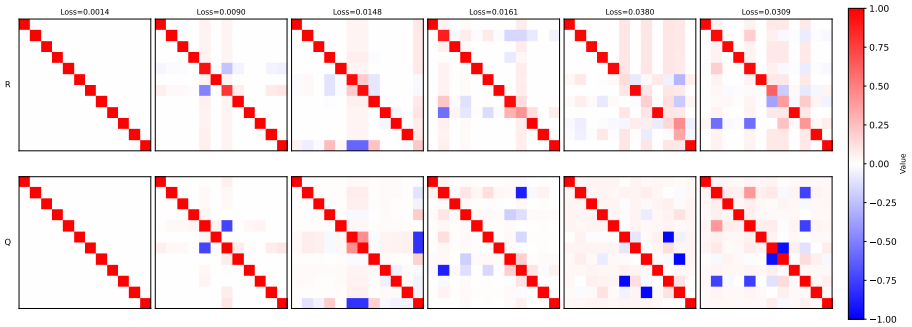


Figure B.2: Order Parameters of erf in $K = M = 12$.

C

Constraint Dynamics

In this section, we derive the mean-field ODEs for Normalised Gradient Descent (nGD) and Orthonormalized Gradient Descent (onGD), which impose distinct geometric constraints on the student weights during training. We also show the resulting learning dynamics for both methods and a brief discussion of stability, convergence properties, and the learnability of nGD and onGD. Consistent with previous sections, we denote the input dimension by d.

C.1 Normalised Student Setting (nGD)

In Normalised Gradient Descent, we enforce the constraint that the norm of each student weight vector remains fixed at initialisation, typically $\|w_i\|^2 = d$ (implying $Q_{ii} = 1$). The continuous-time update rule subtracts the radial component of the gradient:

$$\frac{dw_i}{dt} = -\eta \nabla_{w_i} \mathcal{L} + \frac{\eta}{d}(\nabla_{w_i} \mathcal{L} \cdot w_i^\top) w_i. \tag{C.1}$$

This projection modifies the order parameter dynamics by introducing a decay term proportional to the self-overlap. The resulting ODEs are:

$$\begin{aligned}
\frac{dR_{in}}{dt} &= \eta v_i \left[\sum_{m=1}^{M} v_m^* I_3(i,n,m) - \sum_{j=1}^{K} v_j I_3(i,n,j) \right] \\
&\quad - \eta R_{in} v_i \left[\sum_{m=1}^{M} v_m^* I_3(i,i,m) - \sum_{j=1}^{K} v_j I_3(i,i,j) \right], \\
\frac{dQ_{ik}}{dt} &= \eta v_i \left[\sum_{m=1}^{M} v_m^* I_3(i,k,m) - \sum_{j=1}^{K} v_j I_3(i,k,j) \right] \\
&\quad + \eta v_k \left[\sum_{m=1}^{M} v_m^* I_3(k,i,m) - \sum_{j=1}^{K} v_j I_3(k,i,j) \right] \\
&\quad - \eta Q_{ik} v_i \left[\sum_{m=1}^{M} v_m^* I_3(i,i,m) - \sum_{j=1}^{K} v_j I_3(i,i,j) \right] \\
&\quad - \eta Q_{ik} v_k \left[\sum_{m=1}^{M} v_m^* I_3(k,k,m) - \sum_{j=1}^{K} v_j I_3(k,k,j) \right].
\end{aligned}$$

Note that for the diagonal elements, substituting $k = i$ and $Q_{ii} = 1$ confirms that $\frac{dQ_{ii}}{dt} = 0$, satisfying the constraint.

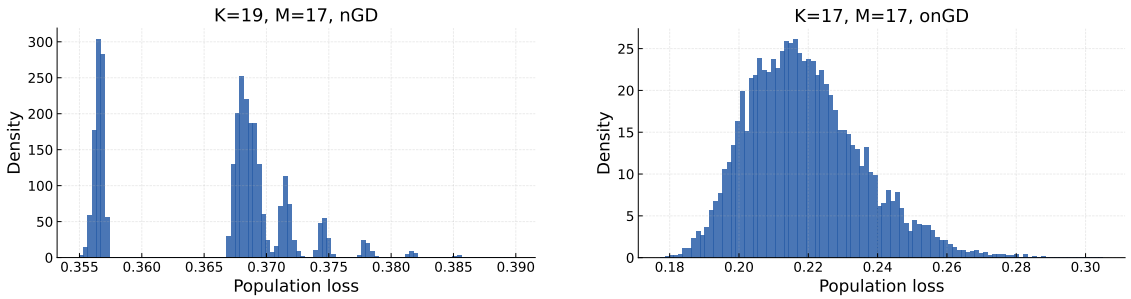
C.2 Empirical Observations for Normalized Gradient Descent

We investigate the loss distribution of Normalized Gradient Descent (nGD) in the overparameterised regime. Specifically, we consider a student network with $K = 19$ hidden units learning from a teacher with $M = 17$ units ($K = M + 2$). Fig. C.1a displays the histogram of the final population loss.

Contrary to the expectation that overparameterisation allows the network to achieve zero loss, nGD exhibits a fundamental barrier. As shown in the histogram, the system fails to reach the rigorous global minima (Loss ≈ 0). Even in the best-found minima (Loss ≈ 0.355), a significant error persists.

This is a direct consequence of the constraint $Q_{ii} = 1$. In unconstrained GD, the $K - M$ excess students would decay to zero norm to eliminate interference, as shown in Fig. 5.2. In nGD, however, the order parameters shown in Fig. C.2 (left columns) indicate that these excess units are constrained to maintain unit norm. They cannot be silenced and effectively act as intrinsic noise sources, creating an irreducible error even when the M target features are perfectly retrieved.

The high density of solutions in the suboptimal interval Loss $\in [0.365, 0.370]$ (approx. 48%) arises from the combinatorial diversity of "blocking" defects. As visualised in the right four columns of Fig. C.2, these states exhibit student groupings of varying sizes (e.g., $k_1 = 3, 4, 5$). Unlike the unique diagonal alignment required for the best solution, there are combinatorially many ways to form these suboptimal clumps. The normalisation constraint stabilises these high-loss configurations, effectively trapping the dynamics in a high-energy part.



(a) nGD Loss Distribution

(b) onGD Loss Distribution

Figure C.1: Loss distribution in different constraints These histograms show the empirical density of the final population loss after 20,000,000 running steps. **(a)** Normalized Gradient Descent (nGD) with $K = 19, M = 17$. Despite overparameterisation, the loss distribution is strictly multimodal, exhibiting discrete peaks at high loss values. **(b)** Orthonormalised Gradient Descent (onGD) with $K = 17, M = 17$. The distribution is unimodal and broad, concentrated entirely in a high-error regime (centered at 0.22).

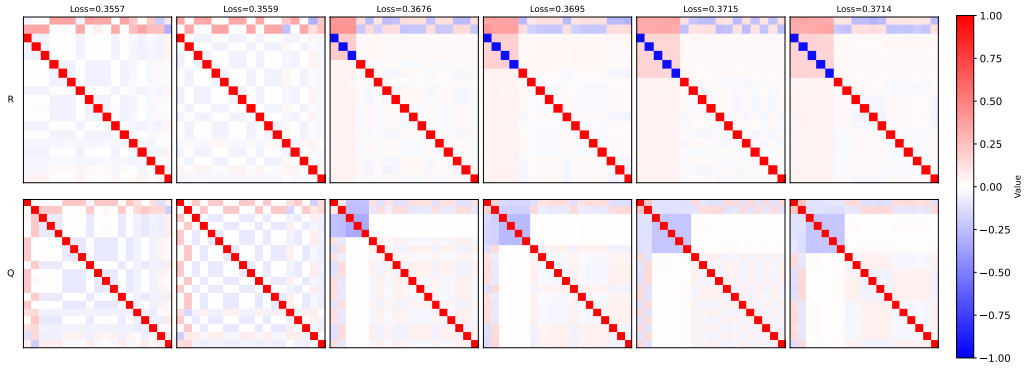


Figure C.2: Order parameters of local minima in nGD ($K = 19, M = 17$). The columns correspond to different final losses. **Left Two Columns (Best Minima):** The R matrices exhibit a near-perfect diagonal structure, indicating successful retrieval of the 17 teacher features. However, the irreducible error persists because the 2 excess students (constrained to $Q_{ii} = 1$) cannot decay to zero. **Right Four Columns (Suboptimal Minima):** These states represent the dominant local minima cluster. They exhibit "blocking" defects (e.g., 2-to-1 assignments seen as red blocks in R), where extra units are misallocated, further elevating the loss.

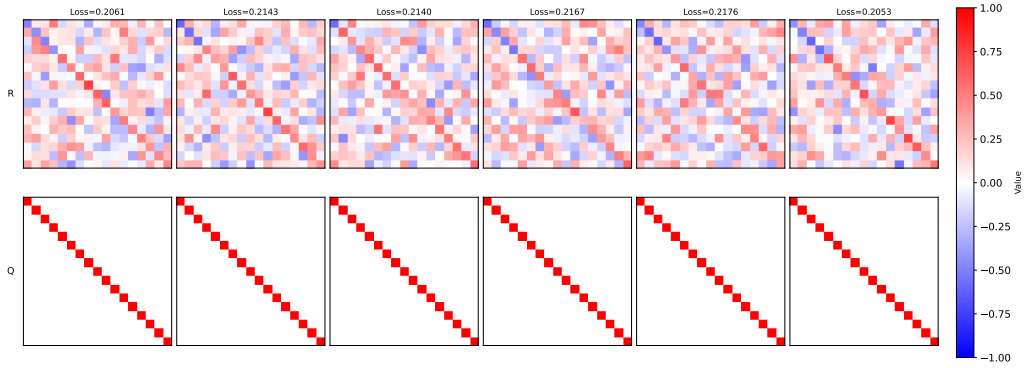


Figure C.3: Final order parameters for onGD ($K = M = 17$). The columns correspond to different final losses. **Bottom Row:** The student-student overlap matrices Q retain a strict identity structure ($Q = I$), satisfying the orthogonality constraint. **Top Row:** The student-teacher overlap matrices R exhibit a disordered, noise-like pattern. Unlike successful learning scenarios shown in Fig. 5.2, the entries remain diffuse and small. This lack of structural alignment confirms that the student units fail to retrieve the teacher vectors in this constrained setting.

C.3 Orthonormalized Student Setting (onGD)

In Orthonormalized Gradient Descent, we enforce the stronger constraint that the student weights remain orthonormal throughout training, i.e., $WW^\top = dI_K$ (implying $Q = I_K$). The update rule projects the gradient onto the tangent space of the Stiefel manifold:

$$\frac{dW}{dt} = -\eta \nabla_W \mathcal{L} + \frac{\eta}{d}(\nabla_W \mathcal{L}\, W^\top)W, \tag{C.2}$$

where $W \in \mathbb{R}^{K\times d}$. Under this setting, the student-student covariance matrix is fixed, so:

$$\frac{dQ_{ik}}{dt} = 0. \tag{C.3}$$

The dynamics of the student-teacher overlap R are modified by a mixing term involving all other student units:

$$\frac{dR_{in}}{dt} = \eta v_i \left[\sum_{m=1}^{M} v_m^* I_3(i,n,m) - \sum_{j=1}^{K} v_j I_3(i,n,j) \right] - \eta v_i \sum_{k=1}^{K} R_{kn} \left[\sum_{m=1}^{M} v_m^* I_3(i,k,m) - \sum_{j=1}^{K} v_j I_3(i,k,j) \right].$$

C.4 Empirical Observations for Orthonormalized Gradient Descent

We next examine the learning dynamics of Orthonormalized Gradient Descent (onGD). We focus on the student-teacher setting with $K = M = 17$. Fig. C.1b presents the distribution of the final population loss.

As shown in Fig. C.1b, onGD does not achieve meaningful learning. The final loss distribution is concentrated in the high-error regime of $[0.18, 0.30]$, with a peak around 0.22. The structural reason for this failure is evident in Fig. C.3: the entries of the student-teacher overlap matrix R remain disordered, confirming that the network is unable to retrieve the teacher configuration.

The failure of onGD appears to be closely related to the excessive rigidity imposed by the orthogonality constraint. In a typical successful learning trajectory (as seen in unconstrained GD), student units often pass through intermediate phases where they become correlated ($Q_{ik} \neq 0$) to "sense" the teacher's structure. By enforcing $Q_{ik} = 0$ at $i \neq k$, onGD effectively forbids these cooperative intermediate states. Our results suggest that for $K = 17$, the gradient flow on this constrained manifold lacks accessible paths connecting the random initialization to the teacher configuration, effectively locking the system in a high-loss state.

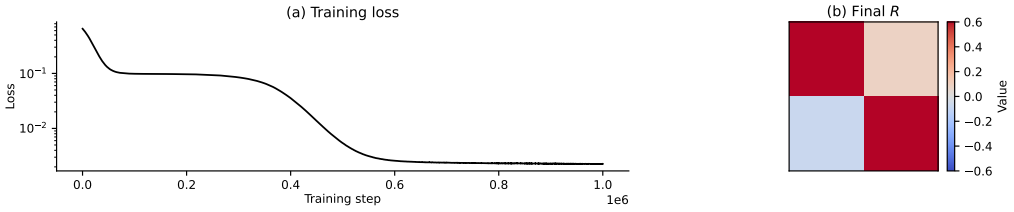


Figure C.4: Dynamics of onGD in a setting with small hidden layers ($K = M = 2$). (a) The training loss (log scale) decreases monotonically and converges to a near-zero value ($\sim 10^{-3}$), indicating successful optimisation. **(b)** The final student-teacher overlap matrix R exhibits a clear diagonal structure (red blocks indicate high positive correlation). This confirms that, unlike the case with larger hidden layers ($K = M = 17$), the orthogonality constraint perfectly retrieves the teacher weights in low dimensions.

To comparison, we present representative learning trajectories in a setting with few hidden units ($K = M = 2$). As shown in Fig. C.4, successful learning is possible in this regime, indicating that the obstruction induced by the orthogonality constraint

is dimension-dependent: systems with only a few hidden units can still navigate the Stiefel manifold to find the solution.

D

Fixed Points Equations

$$E_\varsigma = -\frac{1}{2\pi q}[k_1\,\varsigma\sqrt{q^2-e^2} - \varsigma\sqrt{q^2-e^2} + (k_1-1)\,q\,s\cos^{-1}\left(-\frac{e}{\sqrt{q^2}}\right) + f\,k_2\,q\cos^{-1}\left(-\frac{u}{\sqrt{pq}}\right) - k_1\,\varsigma\sqrt{q-s^2}$$
$$+ k_2\,\varsigma\sqrt{pq-u^2} - k_2\,\varsigma\sqrt{q-\phi^2} + q\,\varsigma\cos^{-1}\left(-\frac{q}{\sqrt{q^2}}\right) + \varsigma\sqrt{q-s^2} - \varsigma\sqrt{q-\varsigma^2} - q\cos^{-1}\left(-\frac{\varsigma}{\sqrt{q}}\right)\Big] = 0, \tag{D.1}$$

$$E_s = -\frac{1}{2\pi q}[k_1\,s\sqrt{q^2-e^2} - s\sqrt{q^2-e^2} + q\cos^{-1}\left(-\frac{e}{\sqrt{q^2}}\right)\big((k_1-2)\,s+\varsigma\big) + f\,k_2\,q\cos^{-1}\left(-\frac{u}{\sqrt{pq}}\right) - k_1\,s\sqrt{q-s^2}$$
$$+ k_2\,s\sqrt{pq-u^2} - k_2\,s\sqrt{q-\varphi^2} + q\,s\cos^{-1}\left(-\frac{q}{\sqrt{q^2}}\right) + s\sqrt{q-s^2} - s\sqrt{q-\varsigma^2} - q\cos^{-1}\left(-\frac{s}{\sqrt{q}}\right)\Big] = 0, \tag{D.2}$$

$$E_\varphi = \frac{1}{2\pi q}[-q\big(b+(k_2-1)\tau\big)\cos^{-1}\left(-\frac{u}{\sqrt{pq}}\right)$$
$$+ \varphi\left(-(k_1-1)\left(\sqrt{q^2-e^2} + q\cos^{-1}\left(-\frac{e}{\sqrt{q^2}}\right)\right) - q\cos^{-1}\left(-\frac{q}{\sqrt{q^2}}\right)\right)$$
$$+ (k_1-1)\,\varphi\sqrt{q-s^2} - k_2\,\varphi\sqrt{pq-u^2} + k_2\,\varphi\sqrt{q-\varphi^2} + \varphi\sqrt{q-\varsigma^2} + q\cos^{-1}\left(-\frac{\varphi}{\sqrt{q}}\right)\Big] = 0, \tag{D.3}$$

$$E_e = \frac{1}{\pi q}\Big[k_1\left(e\left(\sqrt{q-s^2}-\sqrt{q^2-e^2}\right) + qs\cos^{-1}\left(-\frac{s}{\sqrt{q}}\right)\right) + e\sqrt{q^2-e^2} - q\big(e(k_1-2)+q\big)\cos^{-1}\left(-\frac{e}{\sqrt{q^2}}\right)$$
$$+ k_2\left(e\left(\sqrt{q-\varphi^2}-\sqrt{pq-u^2}\right) - qu\cos^{-1}\left(-\frac{u}{\sqrt{pq}}\right)\right) - eq\cos^{-1}\left(-\frac{q}{\sqrt{q^2}}\right) - e\sqrt{q-s^2} + e\sqrt{q-\varsigma^2}$$
$$+ k_2\,\varphi\,q\cos^{-1}\left(-\frac{\varphi}{\sqrt{q}}\right) + q\,\varsigma\cos^{-1}\left(-\frac{s}{\sqrt{q}}\right) + qs\cos^{-1}\left(-\frac{\varsigma}{\sqrt{q}}\right) - 2qs\cos^{-1}\left(-\frac{s}{\sqrt{q}}\right)\Big] = 0, \tag{D.4}$$

$$
\begin{aligned}
E_u =& \frac{1}{2\pi}\Big[\frac{u\sqrt{p-b^2}}{p} + \varphi\cos^{-1}\left(-\frac{b}{\sqrt{p}}\right) + b\cos^{-1}\left(-\frac{\varphi}{\sqrt{q}}\right) \\
&- u\left(\frac{(k_1-1)\left(\sqrt{q^2-e^2} + q\cos^{-1}\left(-\frac{e}{\sqrt{q^2}}\right)\right)}{q} + \cos^{-1}\left(-\frac{q}{\sqrt{q^2}}\right)\right) \\
&- \frac{1}{p}\Big(p\big(e(k_1-1)+q\big)\cos^{-1}\left(-\frac{u}{\sqrt{pq}}\right) + k_1 u\sqrt{pq-u^2}\Big) + (k_1-1)\left(\frac{u\sqrt{p-f^2}}{p} + s\cos^{-1}\left(-\frac{f}{\sqrt{p}}\right)\right) \\
&+ \frac{u\sqrt{p-f^2}}{p} + (k_1-1)\left(f\cos^{-1}\left(-\frac{s}{\sqrt{q}}\right) + \frac{u\sqrt{q-s^2}}{q}\right) + \varsigma\cos^{-1}\left(-\frac{f}{\sqrt{p}}\right) + f\cos^{-1}\left(-\frac{\varsigma}{\sqrt{q}}\right) \\
&- u\left(\frac{(k_2-1)\left(\sqrt{p^2-\mu^2} + p\cos^{-1}\left(-\frac{\mu}{\sqrt{p^2}}\right)\right)}{p} + \cos^{-1}\left(-\frac{p}{\sqrt{p^2}}\right)\right) \\
&- \frac{q\big((k_2-1)\mu+p\big)\cos^{-1}\left(-\frac{u}{\sqrt{pq}}\right) + k_2 u\sqrt{pq-u^2}}{q} \\
&+ (k_2-1)\left(\varphi\cos^{-1}\left(-\frac{\tau}{\sqrt{p}}\right) + \frac{u\sqrt{p-\tau^2}}{p}\right) + (k_2-1)\left(\frac{u\sqrt{q-\varphi^2}}{q} + \tau\cos^{-1}\left(-\frac{\varphi}{\sqrt{q}}\right)\right) \\
&+ \frac{u\sqrt{q-\varphi^2}}{q} + \frac{u\sqrt{q-\varsigma^2}}{q}\Big] = 0,
\end{aligned}
\tag{D.5}
$$

$$
\begin{aligned}
E_f =& -\frac{1}{2\pi p} - f\sqrt{p-b^2} - fk_1\sqrt{p-f^2} + fk_1\sqrt{pq-u^2} + fk_2\sqrt{p^2-\mu^2} + f(k_2-1)\,p\cos^{-1}\left(-\frac{\mu}{\sqrt{p^2}}\right) \\
&- fk_2\sqrt{p-\tau^2} - f\sqrt{p^2-\mu^2} + fp\cos^{-1}\left(-\frac{p}{\sqrt{p^2}}\right) + f\sqrt{p-\tau^2} - p\cos^{-1}\left(-\frac{f}{\sqrt{p}}\right) \\
&+ k_1 ps\cos^{-1}\left(-\frac{u}{\sqrt{pq}}\right) - ps\cos^{-1}\left(-\frac{u}{\sqrt{pq}}\right) + p\varsigma\cos^{-1}\left(-\frac{u}{\sqrt{pq}}\right) = 0,
\end{aligned}
\tag{D.6}
$$

$$
\begin{aligned}
E_b =& -\frac{1}{2\pi p}[-b\sqrt{p-b^2} - bk_1\sqrt{p-f^2} + bk_1\sqrt{pq-u^2} + bk_2\sqrt{p^2-\mu^2} - bk_2\sqrt{p-\tau^2} - b\sqrt{p^2-\mu^2} \\
&+ bp\cos^{-1}\left(-\frac{p}{\sqrt{p^2}}\right) + b\sqrt{p-\tau^2} - p\cos^{-1}\left(-\frac{b}{\sqrt{p}}\right) + k_1 p\varphi\cos^{-1}\left(-\frac{u}{\sqrt{pq}}\right) + (k_2-1)p\tau\cos^{-1}\left(-\frac{\mu}{\sqrt{p^2}}\right)\Big] = 0,
\end{aligned}
\tag{D.7}
$$

$$
\begin{aligned}
E_\tau =& -\frac{1}{2\pi p}[-\tau\sqrt{p-b^2} + p\big(b+(k_2-2)\tau\big)\cos^{-1}\left(-\frac{\mu}{\sqrt{p^2}}\right) - k_1\tau\sqrt{p-f^2} + k_1 p\varphi\cos^{-1}\left(-\frac{u}{\sqrt{pq}}\right) + k_1\tau\sqrt{pq-u^2} \\
&+ k_2\tau\sqrt{p^2-\mu^2} - k_2\tau\sqrt{p-\tau^2} - \tau\sqrt{p^2-\mu^2} + p\tau\cos^{-1}\left(-\frac{p}{\sqrt{p^2}}\right) + \tau\sqrt{p-\tau^2} - p\cos^{-1}\left(-\frac{\tau}{\sqrt{p}}\right)] = 0,
\end{aligned}
\tag{D.8}
$$

$$E_q = \frac{1}{\pi}[-(k_1-1)\sqrt{q^2-e^2} - e(k_1-1)\cos^{-1}\left(-\frac{e}{\sqrt{q^2}}\right) + (k_1-1)\left(\sqrt{q-s^2} + s\cos^{-1}\left(-\frac{s}{\sqrt{q}}\right)\right)$$
$$- k_2\left(\sqrt{pq-u^2} + u\cos^{-1}\left(-\frac{u}{\sqrt{pq}}\right)\right) + k_2\left(\sqrt{q-\varphi^2} + \varphi\cos^{-1}\left(-\frac{\varphi}{\sqrt{q}}\right)\right) - q\cos^{-1}\left(-\frac{q}{\sqrt{q^2}}\right) + \sqrt{q-\varsigma^2}$$
$$+ \varsigma\cos^{-1}\left(-\frac{\varsigma}{\sqrt{q}}\right)\Big] = 0, \tag{D.9}$$

$$E_p = \frac{1}{\pi}[\sqrt{p-b^2} + b\cos^{-1}\left(-\frac{b}{\sqrt{p}}\right) + k_1\left(\sqrt{p-f^2} + f\cos^{-1}\left(-\frac{f}{\sqrt{p}}\right)\right) - k_1\left(\sqrt{pq-u^2} + u\cos^{-1}\left(-\frac{u}{\sqrt{pq}}\right)\right)$$
$$- (k_2-1)\sqrt{p^2-\mu^2} + (k_2-1)(-\mu)\cos^{-1}\left(-\frac{\mu}{\sqrt{p^2}}\right) + k_2\sqrt{p-\tau^2} + (k_2-1)\tau\cos^{-1}\left(-\frac{\tau}{\sqrt{p}}\right)$$
$$- p\cos^{-1}\left(-\frac{p}{\sqrt{p^2}}\right) - \sqrt{p-\tau^2}] = 0, \tag{D.10}$$

$$E_\mu = \frac{1}{\pi p}\Big[\mu\sqrt{p-b^2} + p\tau\cos^{-1}\left(-\frac{b}{\sqrt{p}}\right) + bp\cos^{-1}\left(-\frac{\tau}{\sqrt{p}}\right) + k_1\left(\mu(\sqrt{p-f^2} - \sqrt{pq-u^2}) + fp\cos^{-1}\left(-\frac{f}{\sqrt{p}}\right)\right.$$
$$\left. -pu\cos^{-1}\left(-\frac{u}{\sqrt{pq}}\right)\right) - k_2\mu\left(\sqrt{p^2-\mu^2} + p\cos^{-1}\left(-\frac{\mu}{\sqrt{p^2}}\right)\right) + k_2\mu\sqrt{p-\tau^2} + k_2p\tau\cos^{-1}\left(-\frac{\tau}{\sqrt{p}}\right)$$
$$+ \mu\sqrt{p^2-\mu^2} - p^2\cos^{-1}\left(-\frac{\mu}{\sqrt{p^2}}\right) + 2\mu p\cos^{-1}\left(-\frac{\mu}{\sqrt{p^2}}\right) - \mu p\cos^{-1}\left(-\frac{p}{\sqrt{p^2}}\right)$$
$$- \mu\sqrt{p-\tau^2} - 2p\tau\cos^{-1}\left(-\frac{\tau}{\sqrt{p}}\right)\Big] = 0. \tag{D.11}$$

E

Numerical and Computational Details

The large-scale simulations reported in this work were run on the CPU cluster Tetralith. The CPU nodes used in our experiments are equipped with two Intel Xeon Gold 6130 processors, providing 32 CPU cores per node and 96 GiB of RAM per standard node. No GPU acceleration was used.

For the main ODE simulations, loss histograms, alternative-activation experiments, and constrained-dynamics experiments, we used Slurm job arrays to run independent random initializations in parallel. For settings with both $K < 15$ and $M < 15$, each trajectory was run with one CPU core. For larger settings, where at least one of K or M is at least 15, each trajectory was run with 2 CPU cores. Each array job was assigned a wall-clock limit of 24 hours. The largest runs required up to 10^7 training steps in the well-specified case $K = M$, and up to 1.2×10^7 training steps in the overparameterized case $K = M + 1$. Other simulations and experiments are substantially smaller and can be reproduced on a standard workstation on a laptop/desktop CPU comparable to an Intel i7-13650HX with 32 GB RAM.

Memory usage is modest for all experiments because the simulations evolve the order parameters Q, R and T, rather than full high-dimensional network weights, except in the finite-dimensional robustness checks. The dominant cost is therefore wall-clock time over many independent initializations rather than memory.