An Analysis of On-line Momentum Learning in Soft Committee Machines

Varun Varanasi
Harvard Graduate Program in Biophysics, Harvard University, Cambridge MA 02138, USA.
(Dated: May 16, 2026)

We derive an exact solution to on-line momentum learning in two-layer networks in a noisy-teacher setting. We then investigate the convergence properties of this learning algorithm and study the structure of the symmetric phase characterized by incomplete student-teacher alignment and a plateau in generalization error. Using the derived dynamical equations, we present self-consistency equations to analytically characterize the symmetric points. In simulation, we find that the strong symmetry assumption is only maintained by a small fraction of total observed plateau points. We then study the clustering structure of various plateau points in our numerical simulations and argue that a true analytic theory must account for multi-step symmetry breaking to fully explain the dynamics of momentum based learning rules.

I. INTRODUCTION

There has been recent re-excitement in the machine learning community studying on-line learning. As opposed to traditional learning paradigms that learn a suggestive mapping over batches of a finite data set, on-line learning samples from an infinite distribution and trains the model based on each observation. This framework is thought to be more representative of active, real-world learning. The field is interested in understanding the learning dynamics and generalization properties of on-line learning systems subject to variations in learning algorithms, model choice, and data distribution.

Student-Teacher models present a compelling substrate for studying these questions. By defining the optimal solution as an oracle *teacher* model, the state of the learning *student* system can be well-understood in relation to the known solution. This framework has been studied extensively in the physics of learning community as it can be readily framed in the language of statistical physics where the state of the system can be abstracted to the overlaps between the student and teacher networks, and performance of the student model can be exactly quantified.

Inspired by seminal work by Saad & Solla [1] in the 1990's we continue this framework by investigating exact solutions for momentum-based learning dynamics in two-layer networks. The outline of the paper is as follows. In Section II we introduce the model and task set up, and in Section III we derive an exact solution for on-line learning in mean-field limit. In sections IV and V we discuss the symmetric phase and characterize its glassy landscape through numerics and analytic self-consistency equations.

II. ON-LINE MOMENTUM LEARNING IN A NOISY STUDENT-TEACHER PARADIGM

Consider an on-line student teacher learning task where data is drawn from a zero mean isotropic Gaussian distribution.

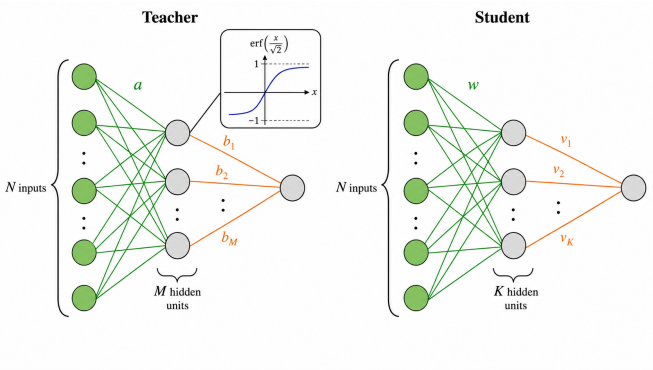

FIG. 1. **Diagram of Student-Teacher Networks.** The input dimension, N is shared across both networks. There are K hidden neurons in the Student Network and M hidden neurons in the Teacher network each imbued with a erf activation function. There is a single readout neuron with readout vectors $v = 1$ and $b = 1$ in the student and teacher networks, respectively.

$$x \sim N(0, \Delta^2 I_N) \tag{2.1}$$

The student and teacher networks are defined as two layer networks with N inputs, K (student)/M (teacher) hidden units, and a single readout neuron. Each neuron is defined with an erf activation function. We label the teacher outputs y and student output as $\hat{y}$ accordingly to the following definitions:

$$\hat{y} = \sum_j^K g(w_j \cdot x) \qquad y = \sum_n^M g(a_n \cdot x) \tag{2.2}$$

$$g = erf(x/\sqrt{2}) \tag{2.3}$$

Note that we implicitly set the readout weights to 1, defining what is known in literature as a soft-committee machine [1]. We make this assumption since the bulk of interesting dynamics are in the learned hidden layer weights and this assumption simplifies later calculations.

The student networks is trained on noisy teacher outputs with MSE loss.

$$MSE = \frac{1}{2}(\hat{y} - y - \sigma\zeta)^2 \tag{2.4}$$

$$= \frac{1}{2}\left(\sum_{i=1}^{K} v_i g(w_i \cdot x) - \sum_{i=1}^{M} b_i g(a_i \cdot x) - \sigma\zeta\right)^2 \tag{2.5}$$

We define σ as the label noise and $\zeta \sim N(0, 1)$.

Momentum Learning Updates

Traditional on-line learning algorithms rely on Stochastic Gradient Descent to navigate the loss landscape defined by the problem. Inspired by heavy-ball optimization, momentum-based learning rules [ADD CITATION] have gained traction as powerful alternatives. Rather than updating the network weights according to the computed gradient at a given time, in momentum-based learning methods, network updates are a weighted function of previous updates. In practice, this is accomplished by introducing an additional memory state variable that introduces second-order dynamics to the system. These methods are thought to be beneficial for navigating sharp loss landscapes in which first-order dynamics can get trapped in local minima.

For our system, we define the weights for a given hidden neuron i as the weight vector w_i. We introduce a corresponding memory parameter Γ_i which controls the direction of movement along the loss landscape. We define momentum gradient descent updates with the parameters β, controlling the memory of the system and η, the learning rate. At a given timestep μ, we receive an observation x^μ and update our weights according to the following equations:

$$\Gamma_i^{\mu+1} = \beta_i \Gamma_i^\mu - (1 - \beta_i)\eta_i \frac{\partial \mathcal{L}}{\partial w_i} \tag{2.6}$$

$$m_i^\mu = N\Gamma_i^\mu \tag{2.7}$$

$$w_i^{\mu+1} = w_i^\mu + \frac{1}{N} m_i^{u+1} \tag{2.8}$$

$$\frac{\partial \mathcal{L}}{\partial w_i} = \left(\sum_{i=1}^{K} g(w_i \cdot x^\mu) - \right. \tag{2.9}$$

$$\left. \sum_{i=1}^{M} g(a_i \cdot x^\mu) - \sigma\zeta\right)(g'(w_i \cdot x^\mu)x^\mu) \tag{2.10}$$

The $1/N$ scaling of memory parameters is preemptively included for our continue time derivation. At this stage, we introduce the following notation:

$$\delta^\mu = \left(\sum_{i=1}^{K} g(w_i \cdot x^\mu) - \sum_{i=1}^{M} g(a_i \cdot x^\mu)\right) \tag{2.11}$$

III. MEAN-FIELD THEORY DYNAMICAL EQUATIONS

Using the statistical physics framework developed in [1, 2], we can apply mean-field techniques to understand learning dynamics of the student network.

Order Parameters

We begin by defining order parameters as the overlaps across the momentum vectors m_i, student weights w_i, and teacher weights a_i.This defines a total of 6 parameters.

$$Q_{ik} = w_i \cdot w_k \qquad R_{ik} = w_i \cdot a_k \qquad T_{ik} = a_i \cdot a_k \tag{3.1}$$

$$S_{ik} = m_i \cdot m_k \qquad U_{ik} = m_i \cdot w_k \qquad P_{ik} = m_i \cdot a_k \tag{3.2}$$

Using the momentum update rules defined above, we can compute step-wise updates to the order parameters themselves. Consider $U_{ik}^{\mu+1}$ as an illustrative example.

$$U_{ik}^{\mu+1} = m_i^{\mu+1} \cdot w_k^{\mu+1} \tag{3.3}$$

$$= \left(N\beta_i\Gamma_i^\mu - (1 - \beta_i)N\eta_i \frac{\partial L}{\partial w_i}\right) \times \tag{3.4}$$

$$\left(w_k^\mu + \beta_k\Gamma_k^\mu - (1 - \beta_k)\eta_k \frac{\partial L}{\partial w_k}\right) \tag{3.5}$$

Local Fields

In the thermodynamic limit, we make the observation that the overlap between a representative data sample x and any parameter of our system, must be a gaussian random variable. We thus define the following quantities as local fields of the system.

$$\lambda_i^\mu = w_i^\mu \cdot x^\mu \qquad \rho_i^\mu = a_i^\mu \cdot x^\mu \qquad \theta_i^\mu = m_i^\mu \cdot x \tag{3.6}$$

which are drawn from the jointly gaussian distribution

$$(\lambda, \rho, \theta) \sim N\left(0, \Delta^2 \begin{bmatrix} Q & R & U^T \\ R^T & T & P \\ U & P^T & S \end{bmatrix}\right). \tag{3.7}$$

We can consequently simplify the order parameter update equations by substituting in local fields and order

parameters where appropriate. In the mean-field limit, we also replace relevant quantities with their expectations under the defined multivariate gaussian. For our example U_{ik}, this results in update equations of the form

$$\begin{aligned} U_{ik}^{\mu+1} = \beta_i U_{ik}^{\mu} + \frac{\beta_i \beta_k}{N} S_{ik}^{\mu} \\ - (1-\beta_k)\eta_k \beta_i \mathbb{E}[\delta^{\mu} g'(\lambda_k)\theta_i] \\ - N(1-\beta_i)\eta_i \mathbb{E}[\delta^{\mu} g'(\lambda_i)\lambda_k] \\ - (1-\beta_i)\eta_i \beta_k \mathbb{E}[\delta^{\mu} g'(\lambda_i)\theta_k] \\ + N^2 \Delta^2 (1-\beta_i)(1-\beta_k)\eta_i \eta_k \mathbb{E}\left[\delta^2 g'(\lambda_i) g'(\lambda_k)\right]. \end{aligned} \tag{3.8}$$

Equations of Motion

These equations can be converted into continuous ODEs by introducing the rescaled time $\Delta\alpha = 1/N$ and the following scaling relations:

$$\beta_i = 1 - \frac{\gamma_i}{N} \qquad \bar{\eta}_i = \eta/N \tag{3.9}$$

The result of this process is a low-dimensional set of coupled ODEs that describe the dynamics of the system. In the new continuous time parameterization γ controls the momentum term. In continuous time, the memory term thus scales approximately with $\exp \gamma\alpha$. In the infinite limit $\gamma \to \infty$ we recover traditional SGD. A more concrete derivation is provided in the appendix. The dynamics can thus be written

$$\frac{dQ_{ik}}{d\alpha} = U_{ki} + U_{ik} \tag{3.10}$$

$$\frac{dR_{ik}}{d\alpha} = P_{ik} \tag{3.11}$$

$$\frac{dS_{ik}}{d\alpha} = (-\gamma_i - \gamma_k) S_{ik} - \gamma_i \bar{\eta}_i v_i F1_{ik} - \gamma_k \bar{\eta}_k v_k F1_{ki} \tag{3.12}$$

$$+ \Delta^2 \gamma_i \gamma_k \bar{\eta}_i \bar{\eta}_k v_i v_k \left(G_{ik} + \sigma^2 J2_{ik}\right) \tag{3.13}$$

$$\frac{dU_{ik}}{d\alpha} = -\gamma_i U_{ik} + S_{ik} - \gamma_i \bar{\eta}_i v_i F2_{ik} \tag{3.14}$$

$$\frac{dP_{ik}}{d\alpha} = -\gamma_i P_{ik} - \gamma_i \bar{\eta}_i v_i F3_{ik}. \tag{3.15}$$

where we introduce the following primitives:

$$F1_{ik} = \mathbb{E}[\delta g'(\lambda_i)\theta_k] \qquad F2_{ik} = \mathbb{E}[\delta g'(\lambda_i)\lambda_k] \tag{3.16}$$

$$F3_{ik} = \mathbb{E}[\delta g'(\lambda_i)\rho_k] \qquad G_{ik} = \mathbb{E}[\delta^2 g'(\lambda_i) g'(\lambda_k)] \tag{3.17}$$

$$H_i = \mathbb{E}[\delta g(\lambda_i)] \qquad J2_{ik} = \mathbb{E}[g'(\lambda_i) g'(\lambda_k)] \tag{3.18}$$

Each primitive admits a closed form expression as gaussian integrals of erf functions. These expansions derived in previous work [1, 2] can be found in the appendix.

Generalization Error

The primary quantity we are concerned about in our learning system is how well the student replicates the teacher output. We measure this via the generalization error which is calculated as the expectation of the squared error of the system over the total distribution of inputs.

$$\varepsilon_g = \mathbb{E}_{x \sim N(0, \Delta^2 I_N)}[(\hat{y} - y)^2] \tag{3.19}$$

Using the same techniques to derived closed forms expressions of the primitive expressions in the dynamics, we can compute the generalization error as a function of the local fields and order parameters.

$$\begin{aligned} \varepsilon_g = \frac{2}{\pi} \Bigg[\sum_{j,l=1}^{K} \arcsin\left(\frac{\Delta^2 Q_{jl}}{\sqrt{(1+\Delta^2 Q_{jj})(1+\Delta^2 Q_{ll})}} \right) \\ + \sum_{n,m=1}^{M} \arcsin\left(\frac{\Delta^2 T_{nm}}{\sqrt{(1+\Delta^2 T_{nn})(1+\Delta^2 T_{mm})}} \right) \\ - 2 \sum_{j=1}^{K} \sum_{n=1}^{M} \arcsin\left(\frac{\Delta^2 R_{jn}}{\sqrt{(1+\Delta^2 Q_{jj})(1+\Delta^2 T_{nn})}} \right) \Bigg]. \end{aligned} \tag{3.20}$$

We've thus derived equations of motion for our momentum-based learning algorithm in the thermodynamic limit. The equations concentrate well for relatively large values of N and can approximate numerical simulations. In Fig 2 we look at alignment between the derived equations of motion and numerical simulations across order parameters, predicted generalization loss, and student-teacher cosine similarity. We consider simple learnable system parameterized by $K = M = 2$ with a constant learning rate $\eta = 2$ and continuous time memory parameter γ.

IV. SYMMETRIC PHASE

It is well-known in that on-line student teacher settings such as the one explored here, there exist two distinct phases of learning characterized by the symmetric and convergent phases [1]. In the initial symmetric phase, generalization error plateaus as the student network navigates a subspace characterized by undifferentiated student-teacher alignment. The network eventually breaks free of the symmetric phase and perfectly aligns with the teacher network. In this new learning regime, we observe the same phenomena. In Fig 2. 2, the student network enters the symmetric phase around $\alpha = 10$ characterized by a plateau in the generalization error and identical alignment of student to teacher neurons.

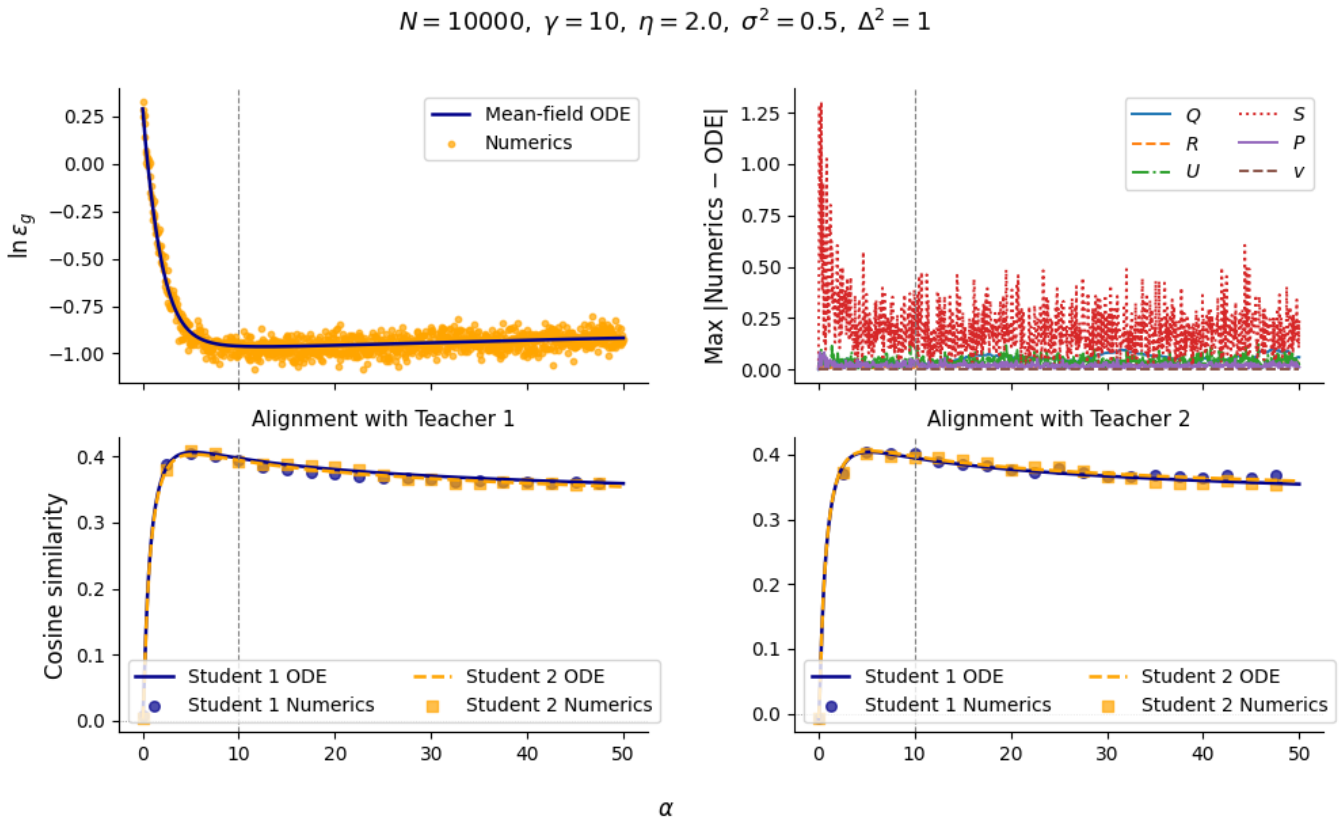

FIG. 2. **Comparison between mean-field dynamics and finite-N simulations.** Simulations were conducted using input size $N = 10,000$ and data drawn from a distribution with variance $\Delta^2 = 1$. Learning dynamics were controlled by a constant memory of $\gamma = 10$, learning rate set to $\eta = 2.0$, and label noise defined of $\sigma^2 = 0.5$. Teacher network was randomly selected with orthonormal weight overlaps such that $T_{nm} = \delta_{nm}$. **(a)** Generalization error over online time α. Derived ODE dynamics (blue) are shown overlaid with numerical simulation results (orange). **(b)** Entry-wise Maximum deviation between numerical simulations and ODE order parameters over training time. **(c,d)** Cosine-Similarity of Student-Teacher alignment across each teacher neurons.

Momentum updates are typically thought to improve learning efficiency in complex landscapes. However, counterintuitively, we observe that student networks are unable to breakout of the symmetric phase over long training times $\alpha \approx 15,000$. This is in stark contrast to observed behavior in Saad and Solla [1], where traditional SGD is able to escape the symmetric phase in $\alpha = 1000$.

To study these transient dynamics, we define the plateau by static order parameters Q and R.

$$\frac{dQ_{ik}}{d\alpha} = 0 \qquad \frac{dR_{ik}}{d\alpha} = 0 \tag{4.1}$$

Immediately from our equations of motion we get the following equalities:

$$U_{ii} = 0 \quad U_{ik} = -U_{ki} \qquad P_{ik} = 0 \tag{4.2}$$

To satisfy stationarity of our plateau fixed point, $\frac{dR}{d\alpha} = 0$ is insufficient since P_{ik} can be transiently 0. We must therefore additionally constrain $\frac{dP_{ik}}{d\alpha} = 0$ so the static generalization error persists. The same stationarity argument can be made for $\frac{dU_{ii}}{d\alpha} = 0$ and $\frac{dU_{ik}}{d\alpha} + \frac{dU_{ki}}{d\alpha} = 0$. Under these constraints, the plateau point is defined by the following structure:

$$\frac{dQ_{ik}}{d\alpha} = 0, \tag{4.3a}$$

$$\frac{dR_{ik}}{d\alpha} = 0, \tag{4.3b}$$

$$\frac{dS_{ik}}{d\alpha} = -(\gamma_i + \gamma_k)S_{ik} - \gamma_i\bar{\eta}_i v_i F1_{ik} - \gamma_k\bar{\eta}_k v_k F1_{ki} + \Delta^2\gamma_i\gamma_k\bar{\eta}_i\bar{\eta}_k v_i v_k \left(G_{ik} + \sigma^2 J2_{ik}^2\right),$$

$$U_{ik}: \quad \begin{cases} S_{ii} = \gamma_i\bar{\eta}_i F2_{ii}, \\ 0 = -\gamma_i U_{ik} - \gamma_k U_{ki} + S_{ik} + S_{ki} - (\gamma_i\bar{\eta}_i + \gamma_k\bar{\eta}_k)\, F2_{ki}, \qquad i \neq k, \end{cases} \tag{4.3c}$$

$$0 = \frac{dP_{ik}}{d\alpha} = F_{ik}^{(3)}. \tag{4.3d}$$

Notice that S must be symmetric by construction as an overlap matrix.

Strong Symmetry Ansatz

The symmetric plateau point as discussed in Saad and Solla [1] is characterized by undifferentiated student-teacher overlaps. We can enforce this structure on our plateau point via the strong symmetry ansatz:

$$Q_{\alpha\beta} = \begin{cases} C & \alpha = \beta \\ q & \alpha \neq \beta \end{cases} \qquad R_{\alpha\beta} = R \tag{4.4}$$

We also assume uniform site dependent parameters $\gamma_i = \gamma$ and $\eta_i = \eta$ Structurally, this assumption corresponds to the single alignment plateau fixed points observed in Fig 2. Leveraging the closed form expressions for each primitive, we can apply the strong symmetry ansatz to produce self-consistent fixed point relations between q, C, and R.

Stationarity of P

The stationarity condition of P implies that $0 = F3_{ik}$. We can expand the definition of $F3_{ik}$ to produce our first equation.

$$q = \frac{KR + (K-1)\Delta^2 RC - 1 - \Delta^2 C + M\Delta^2 R^2}{(K-1)\Delta^2 R} \tag{4.5}$$

A full derivation is included in the appendix.

Anti-Symmetry of U

Recall that S is always symmetric by construction (defined as an overlap). Under the strong symmetry ansatz, we assume that Q and R are symmetric so we it follows that $F2$ must also be symmetric. Any asymmetric evolution of U must arise from asymmetry in U itself. We've previously shown that at the plateau point $U_{ii} = 0$. Consider

$$U_{ik} = \frac{U_{ik} + U_{ki}}{2} + \frac{U_{ik} - U_{ki}}{2} \tag{4.6}$$

Since U is anti-symmetric, at the fixed point we get that

$$\frac{dU_{ik}}{d\alpha} = \frac{1}{2}\left(\frac{dU_{ik}}{d\alpha} - \frac{dU_{ki}}{d\alpha}\right) = -\frac{\gamma}{2}\left(U_{ik} - U_{ki}\right). \tag{4.7}$$

This implies that an off-diagonal element of U should decay corresponding to the difference between the symmetric pairs. For a non-transient fixed point, this collapses to $U_{ik} = U_{ki} = 0$. Therefore, we argue that $U = 0$ at the symmetric fixed point under strong symmetry assumptions. Note that if the learning rate and memory parameters were not identical across sites this would introduce a driving parameter that would allow for an anti-symmetric U.

With the constraint that $U = 0$, we can show that $F1_{ik} = 0$. We now have that for any indices i and k

$$S_{ik} = \gamma\eta F2_{ik} \tag{4.8}$$

$$\frac{dS_{ik}}{d\alpha} = -2\gamma S_{ik} + \Delta^2\gamma^2\eta^2\left(G_{ik} + \sigma^2 J2_{ik}\right) \tag{4.9}$$

Notice that $F2$ is a function of C, R, q which are all static at the plateau points. Therefore, $F2$ and consequently S_{ik} are also static. We can therefore recover two fixed point relations for the diagonal and off-diagonal entries in the above expressions.

Fixed Point Self-Consistency Equations

Under the strong symmetry ansatz, we get the following fixed point self-consistency equations:

$$q = \frac{KR + (K-1)\Delta^2 RC - 1 - \Delta^2 C + M\Delta^2 R^2}{(K-1)\Delta^2 R}, \tag{4.10a}$$

$$\frac{4\Delta^2}{\pi(1+\Delta^2 C)D_C}\left[(K-1)q + C - MR\right] = \Delta^2 \eta G_{ii} + \frac{2\Delta^2 \eta \sigma^2}{\pi D_C}, \tag{4.10b}$$

$$\frac{4}{\pi(1+\Delta^2 C)D_q}\begin{bmatrix}(K-1)q(1+\Delta^2 C) + C(1+\Delta^2 C) - (K-1)\Delta^4 q^2 - \Delta^4 qC \\ -MR(1+\Delta^2 C) + M\Delta^4 qR\end{bmatrix} = \Delta^2 \eta G_{ik} + \frac{2\Delta^2 \eta \sigma^2}{\pi D_q}, \qquad i \neq k. \tag{4.10c}$$

where we define the following quantities.

$$D_C \equiv \sqrt{1 + 2\Delta^2 C} \qquad D_q \equiv \sqrt{1 + 2\Delta^2 C + \Delta^4 C^2 - \Delta^4 q^2}, \tag{4.11}$$

Notice that we preserve η dependence in our fixed point equations. The primitive G_{ik} can similarly be expanded; however, in the interest of legibility, we left the fixed point equations as is. An expansion can be found in the relevant appendix. It is important to note that the fixed point relations defined under the strong symmetry ansatz involve transient functions (arcsin hidden in G). They therefore do not admit a closed form solution unlike in Saad and Solla's original work. We can numerically solve for our q, C, R and validate against our simulated plateau points.

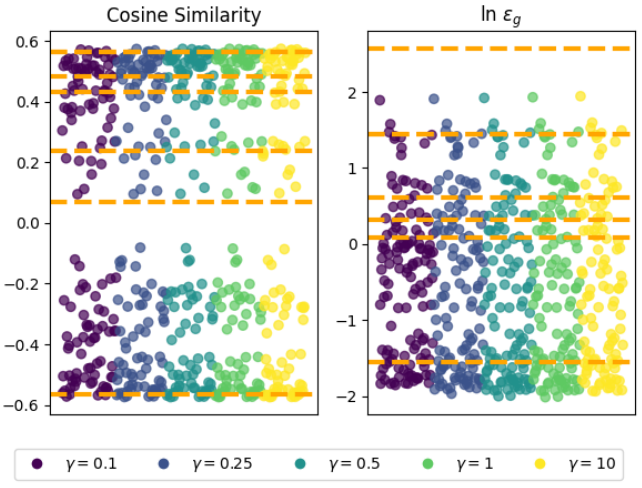

FIG. 3. **Predicted Symmetric Fixed Points vs Simulated ODE Dynamics.** For various γ, we integrate ODE dynamics until $\alpha = 75$ with label noise $\sigma = 0.1$, learning rate $\eta = 3.0$, and data variance $\Delta^2 = 1$. For each run, we record the order parameters of the final state and compute the corresponding generalization error and cosine similarity. Presented data ($n = 581$) is restricted to simulations that exhibit strong symmetry. Deviations from this assumption are discussed in section V.

In Fig 3, we solve the fixed point conditions using numerical solvers to predict predicted symmetric plateau points. To simulate imperfect fixed points, we propose the solutions to the self-consistency equation within a tolerance threshold. We find multiple fixed point solutions as recovered by Biehl et. al in SGD dynamics [3]. The fixed points characterized by q, R, and C can be converted to corresponding generalization errors and cos similarities. The discovered fixed points align well with simulated ODE dynamics for various γ parameterizations. There does however, appear to be a predicted fixed point with generalization loss higher than observed in simulation. This fixed point may be more transient than the others and thus be underrepresented in our data. Also, note that our fixed point predictions and simulations appear to be independent of γ. In other simulations, we observe η dependence as predicted by our expression.

V. REPLICA SYMMETRY BREAKING

One of the most striking observations of our simulated plateau points is the diversity of generalization error, student-teacher alignment, and degree of symmetry. In fact, we observe that only approximately $\sim 6\%$ of our observed plateau points exhibited the strong symmetry discussed in the previous section. A plurality of observed points exhibit complete asymmetry or partial symmetry (Fig 4). This structure is conserved across all γ. We also observed a diversity in generalization error at the plateau phase.

Since seeds are shared across runs, any diversity in the readouts is purely a function of γ. As discussed above, the symmetric phase fixed points are independent of γ; however, as shown above, there exists a continuum of symmetric alignment reminiscent of replica symmetry breaking. Inspired by this observation, we define an overlap quantity for each replica to study the substructure of plateau solutions. We argue that the cosine similarity is the correct order parameter for this function since it is a bounded quantity that directly captures student-teacher alignment from Q and R. With an orthogonal teacher

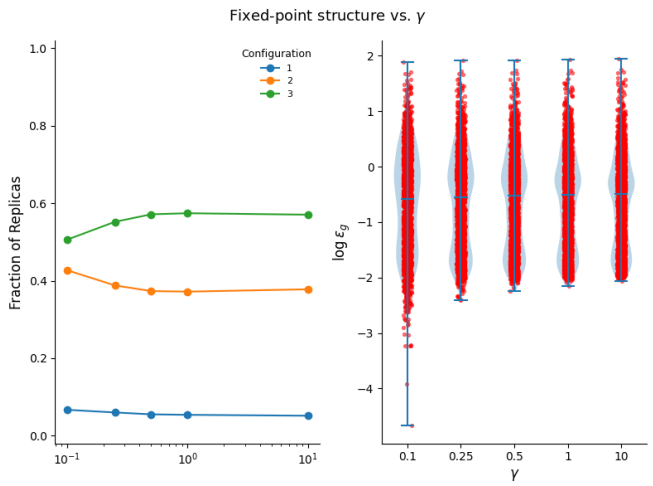

FIG. 4. **Symmetry Breaking in Plateau Points** For various γ, we integrate ODE dynamics until $\alpha = 75$ with label noise $\sigma = 0.1$, learning rate $\eta = 3.0$, and data variance $\Delta^2 = 1$ for 2000 shared seeds. For each run, we record the order parameters of the final state and compute the corresponding generalization error and cosine similarity. We then compute the number of aligned groups in the cos similarity matrix to classify the degree of symmetry (max deviation < 0.1). Generalization error is computed as a function of the final order parameters themselves.

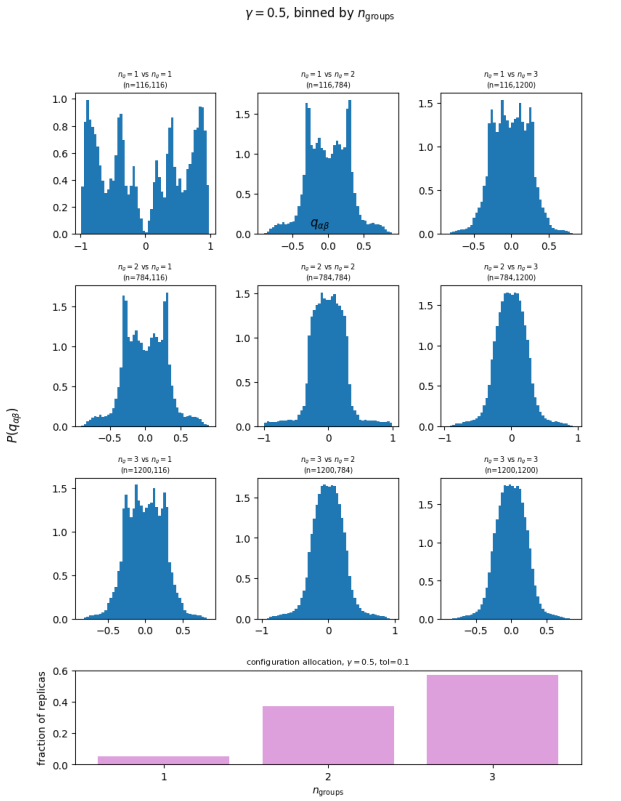

FIG. 5. **Cos-Sim Overlap by Generalization Error** Using the computed overlap quantity for all runs of $\gamma = 0.5$, we differentiate overlaps between replicas by the number of groups (degree of symmetric alignment present). The distribution of each group relative to the total replica population is shown in pink. Overlaps between groups are shown in the off-diagonal elements while within the group is shown on the diagonal.

overlap matrix, we can write the cos-similarity as

$$\mathcal{C}_{ik} = R_{ik}/\sqrt{Q_{ii}} \tag{5.1}$$

To address permutation symmetry, we align the student neurons such that the trace of $\mathcal{C}$ is maximized. We then define an overlap parameter between two replicas as follows:

$$q_{\alpha,\beta} = \frac{1}{K}\sum_{i,k} \mathcal{C}^{\alpha}_{ik}\mathcal{C}^{\beta}_{ki} \tag{5.2}$$

Conditioning our replicas based on qualities of interest, we can plot empirical distributions of $P(q_{\alpha,\beta})$. The structure of these distributions is indicative of the underlying structure of the system.

The overlap parameter analysis reveals strong heterogeneity between different plateau solutions on the basis of configuration. We also find that cluster structures are near identical across each γ run; however, in aggregate, we find that γ runs tend to co-cluster together. We also find that the solutions with the same number of configurations cluster together and more tightly than by γ. We performed a similar clustering analysis based on the euclidean distances between replicas of $\mathcal{C}$ and recover the same results.

VI. CONCLUSION

The initial simulation results from this work indicate that the plateau points in momentum based learning may be more stable than their stochastic gradient descent counterparts. Further analysis into the structure of these fixed points shows that they diverge from their symmetric assumptions. The overlap parameter analysis reveals strong heterogeneity between different plateau solutions on the basis of configuration. This indicates that a true analytic understanding of the plateau phenomena will require reconciling the heterogeneity observed in the simulations. Future work should examine replica symmetry breaking structures to recover the full spectrum of observed behavior. Characterizing these points under various symmetry assumptions is the first step into understanding their stability properties. Future work, should analyze the escape times from these fixed points based on deviations from strong symmetry.

VII. ACKNOWLEDGMENTS

I would like to thank Francesca Mignacco, Francesco Mori, Derya Cansever and Cengiz Pehlevan for the helpful discussions, guidance, and the ongoing collaboration. I also acknowledge ChatGPT for schematic generation.

[1] D. Saad and S. A. Solla, On-line learning in soft committee machines, Phys. Rev. E **52**, 4225 (1995).
[2] F. Mignacco and F. Mori, A statistical physics framework for optimal learning (2025), arXiv:2507.07907 [cond-mat.dis-nn].
[3] M. Biehl, P. Riegler, and C. Wöhler, Transient dynamics of on-line learning in two-layered neural networks, Journal of Physics A: Mathematical and General **29**, 4769 (1996).

Appendix A: Momentum Gradient Updates

Given the teacher-student set up defined, consider the following overlap order parameters:

$$Q_{ik} = w_i \cdot w_k \qquad R_{ik} = w_i \cdot a_k \qquad T_{ik} = a_i \cdot a_k \tag{A1}$$

$$S_{ik} = m_i \cdot m_k \qquad U_{ik} = m_i \cdot w_k \qquad P_{ik} = m_i \cdot a_k \tag{A2}$$

Now we define the following local fields:

$$\lambda_i^\mu = w_i^\mu \cdot x^\mu \qquad \rho_i^\mu = a_i^\mu \cdot x^\mu \qquad \theta_i^\mu = m_i^\mu \cdot x \tag{A3}$$

In the large N limit, we argue that these local fields are drawn from the jointly gaussian distribution:

$$(\lambda, \rho, \theta) \sim N\left(0, \Delta^2 \begin{bmatrix} Q & R & U^T \\ R^T & T & P \\ U & P^T & S \end{bmatrix}\right) \tag{A4}$$

With these definitions, we can expand each on-line update rules to derive mean-field updates by replacing quantities with their corresponding expectations, local fields, and order parameters.

$$\begin{aligned} Q_{ik}^{\mu+1} &= Q_{ik}^\mu + \frac{\beta_i}{N} U_{ik}^\mu + \frac{\beta_k}{N} U_{ki}^\mu + \frac{\beta_i \beta_k}{N^2} S_{ik}^\mu \\ &\quad - (1-\beta_i)\eta_i v_i \mathbb{E}[\delta^\mu g'(\lambda_i)\lambda_k] - (1-\beta_k)\eta_k v_k \mathbb{E}[\delta^\mu g'(\lambda_k)\lambda_i] \\ &\quad - \frac{\beta_k(1-\beta_i)}{N}\eta_i v_i \mathbb{E}[\delta^\mu g'(\lambda_i)\theta_k] - \frac{\beta_i(1-\beta_k)}{N}\eta_k v_k \mathbb{E}[\delta^\mu g'(\lambda_k)\theta_i] \\ &\quad + N\Delta^2(1-\beta_i)(1-\beta_k)\eta_i\eta_k v_i v_k \left(\mathbb{E}\left[\delta^2 g'(\lambda_i)g'(\lambda_k)\right] + \sigma^2 \mathbb{E}[g'(\lambda_i)g'(\lambda_k)]\right), \end{aligned} \tag{A5}$$

$$\begin{aligned} R_{ik}^{\mu+1} &= w_i^{\mu+1} \cdot a_k \\ &= \left(w_i^\mu + \beta_i \Gamma_i^\mu - (1-\beta_i)\eta_i \frac{\partial L}{\partial w_i}\right) \cdot a_k \\ &= R_{ik}^\mu + \frac{\beta_i}{N} P_{ik}^\mu - (1-\beta_i)\eta_i v_i \mathbb{E}[\delta^\mu g'(\lambda_i)\rho_k], \end{aligned} \tag{A6}$$

$$\begin{aligned} S_{ik}^{\mu+1} &= m_i^{\mu+1} \cdot m_k^{\mu+1} \\ &= \left(N\beta_i \Gamma_i^\mu - N(1-\beta_i)\eta_i \frac{\partial L}{\partial w_i}\right) \cdot \left(N\beta_k \Gamma_k^\mu - N(1-\beta_k)\eta_k \frac{\partial L}{\partial w_k}\right) \\ &= \beta_i\beta_k S_{ik}^\mu - N\beta_k(1-\beta_i)\eta_i v_i \mathbb{E}[\delta^\mu g'(\lambda_i)\theta_k] - N\beta_i(1-\beta_k)\eta_k v_k \mathbb{E}[\delta^\mu g'(\lambda_k)\theta_i] \\ &\quad + \Delta^2 N^3 (1-\beta_i)(1-\beta_k)\eta_i\eta_k v_i v_k \left(\mathbb{E}\left[\delta^2 g'(\lambda_i)g'(\lambda_k)\right] + \sigma^2 \mathbb{E}[g'(\lambda_i)g'(\lambda_k)]\right), \end{aligned} \tag{A7}$$

$$\begin{aligned} U_{ik}^{\mu+1} &= m_i^{\mu+1} \cdot w_k^{\mu+1} \\ &= \left(N\beta_i \Gamma_i^\mu - N(1-\beta_i)\eta_i \frac{\partial L}{\partial w_i}\right) \cdot \left(w_k^\mu + \beta_k \Gamma_k^\mu - (1-\beta_k)\eta_k \frac{\partial L}{\partial w_k}\right) \\ &= \beta_i U_{ik}^\mu + \frac{\beta_i\beta_k}{N} S_{ik}^\mu - (1-\beta_k)\eta_k \beta_i v_k \mathbb{E}[\delta^\mu g'(\lambda_k)\theta_i] - N(1-\beta_i)\eta_i v_i \mathbb{E}[\delta^\mu g'(\lambda_i)\lambda_k] \\ &\quad - (1-\beta_i)\eta_i \beta_k v_i \mathbb{E}[\delta^\mu g'(\lambda_i)\theta_k] + N^2\Delta^2(1-\beta_i)(1-\beta_k)\eta_i\eta_k v_i v_k \mathbb{E}\left[\delta^2 g'(\lambda_i)g'(\lambda_k)\right], \end{aligned} \tag{A8}$$

$$\begin{aligned} P_{ik}^{\mu+1} &= m_i^{\mu+1} \cdot a_k \\ &= \left(N\beta_i \Gamma_i^\mu - N(1-\beta_i)\eta_i \frac{\partial L}{\partial w_i}\right) \cdot a_k \\ &= \beta_i P_{ik}^\mu - N(1-\beta_i)\eta_i v_i \mathbb{E}[\delta^\mu g'(\lambda_i)\rho_k]. \end{aligned} \tag{A9}$$

In the above expressions we implicitly simplify all expectations involving $\mathbb{E}[\zeta] = 0$ and $\mathbb{E}[\zeta^2] = 1$ by definition.

Appendix B: Closing Mean-Field Dynamics

Closing the mean-field dynamics requires careful consideration of scaling behavior to ensure that each order parameter stays $O(1)$ in the appropriate mean-field limit. First, we introduce the normalized time-scale $\Delta\alpha = 1/N$. We also assume that our chosen local fields are $O(1)$ (true under self-consistency).

$$\lambda_i = w_i \cdot x \sim O(1) \quad \rho_i = a_i \cdot x \sim O(1) \quad \theta_i = m_i \cdot x \sim O(1) \tag{B1}$$

Recall that $x_i \sim N(0, \Delta^2)$ and from initialization, we set $w_i \sim N(0, 1/N)$ and analogously for $a_i \sim N(0, 1/N)$. We begin by supposing that $v_i \sim O(1)$ where we know $v_i^{(0)} \sim O(1)$ from initialization. This implies that $\delta^0 \sim O(1)$. If we can validate that $v \sim O(1)$ under the equations of motion, this carries forward to δ by construction. Under the rescaled time, we find that

$$\frac{dv_i}{d\alpha} = \lim_{N\to\infty} \frac{v_i^{\mu+1} - v_i^{\mu}}{\Delta\alpha} = \ell^{\mu+1} \tag{B2}$$

$$\frac{d\ell_i}{d\alpha} = \lim_{N\to\infty} \frac{\ell_i^{\mu+1} - \ell_i^{\mu}}{\Delta\alpha} = N(\beta_i - 1)\ell_i - (1-\beta_i)N^2\eta_i\mathbb{E}[\delta g(\lambda_i)]) \tag{B3}$$

The expectation $\mathbb{E}[\delta g(\lambda_i)] \sim O(1)$ under the initialization assumptions. We can introduce a change/rescaling of variables $\beta_i = 1 - \gamma_i/N$ and $\eta \to \bar{\eta}/N$. Under this conditions,

$$\frac{d\ell}{d\alpha} = -\gamma_i\ell_i - \gamma_i\bar{\eta}_i\mathbb{E}[\delta g(\lambda_i)] \sim O(1) \tag{B4}$$

$$\frac{dv_i}{d\alpha} = \ell \sim O(1) \tag{B5}$$

Self-consistency is satisfied for $\ell_i \sim O(1)$ and $v_i \sim O(1)$. We thus proceed with the following scaling relations:

$$\beta_i = 1 - \frac{\gamma_i}{N} \qquad \bar{\eta}_i = \eta/N \tag{B6}$$

We can now derive the equations of motion for the remaining order parameters:

$$\begin{aligned}
\frac{dQ_{ik}}{d\alpha} = \lim_{N\to\infty} \frac{Q_{ik}^{\mu+1} - Q_{ik}^{\mu}}{\Delta\alpha} &= (1 - \frac{\gamma_i}{N})U_{ik} + (1 - \frac{\gamma_k}{N})U_{ki} + (\frac{1}{N} - \frac{\gamma_i}{N^2} - \frac{\gamma_k}{N^2} + \frac{\gamma_i\gamma_k}{N^3})S_{ik} && \text{(B7)}\\
&\quad - \frac{\gamma_i\bar{\eta}_i}{N}v_i\mathbb{E}[\delta^\mu g'(\lambda_i)\lambda_k] - \frac{\gamma_k\bar{\eta}_k}{N}v_k\mathbb{E}[\delta^\mu g'(\lambda_k)\lambda_i] && \text{(B8)}\\
&\quad - (1 - \frac{\gamma_k}{N})\frac{\gamma_i\bar{\eta}_i}{N}v_i\mathbb{E}[\delta^\mu g'(\lambda_i)\theta_k] - (1 - \frac{\gamma_i}{N})\frac{\gamma_k\bar{\eta}_k}{N}v_k\mathbb{E}[\delta^\mu g'(\lambda_k)\theta_i] && \text{(B9)}\\
&\quad + \Delta^2\frac{\gamma_i\gamma_k\bar{\eta}_i\bar{\eta}_k}{N^2}v_iv_k\left(\mathbb{E}[\delta^2 g'(\lambda_i)g'(\lambda_k)] + \sigma^2\mathbb{E}[g'(\lambda_i)g'(\lambda_k)]\right) && \text{(B10)}\\
&= U_{ki} + U_{ik} && \text{(B11)}\\
&\sim O(1) && \text{(B12)}
\end{aligned}$$

$$\frac{dS_{ik}}{d\alpha} = \lim_{N\to\infty} \frac{S_{ik}^{\mu+1} - S_{ik}^{\mu}}{\Delta\alpha} = N(-\frac{\gamma_i}{N} - \frac{\gamma_k}{N} + \frac{\gamma_i\gamma_k}{N^2})S_{ik} - N^2(1-\frac{\gamma_k}{N})\frac{\gamma_i\bar{\eta}_i}{N^2}v_i\mathbb{E}[\delta g'(\lambda_i)\theta_k] \tag{B13}$$

$$- N^2(1-\frac{\gamma_i}{N})\frac{\gamma_k\bar{\eta}_k}{N^2}v_k\mathbb{E}[\delta g'(\lambda_k)\theta_i] \tag{B14}$$

$$+ \Delta^2 N^4 \frac{\gamma_i\gamma_k\bar{\eta}_i\bar{\eta}_k}{N^4}v_iv_k\left(\mathbb{E}[\delta^2 g'(\lambda_i)g'(\lambda_k)] + \sigma^2\mathbb{E}[g'(\lambda_i)g'(\lambda_k)]\right) \tag{B15}$$

$$= (-\gamma_i - \gamma_k)S_{ik} - \gamma_i\bar{\eta}_i v_i\mathbb{E}[\delta g'(\lambda_i)\theta_k] - \gamma_k\bar{\eta}_k v_k\mathbb{E}[\delta g'(\lambda_k)\theta_i] \tag{B16}$$

$$+ \Delta^2\gamma_i\bar{\eta}_i\gamma_k\bar{\eta}_k v_iv_k\left(\mathbb{E}[\delta^2 g'(\lambda_i)g'(\lambda_k)] + \sigma^2\mathbb{E}[g'(\lambda_i)g'(\lambda_k)]\right) \tag{B17}$$

$$\sim O(1) \tag{B18}$$

$$\frac{dR}{d\alpha} = \lim_{N\to\infty} \frac{R_{ik}^{\mu+1} - R_{ik}^{\mu}}{\Delta\alpha} = (1-\gamma_i/N)P_{ik} - \frac{\gamma_i\bar{\eta}_i}{N}v_i\mathbb{E}[\delta g'(\lambda_i)\rho_k] = P_{ik} \sim O(1) \tag{B19}$$

$$\frac{dP_{ik}}{d\alpha} = \lim_{N\to\infty} \frac{P_{ik}^{\mu+1} - P_{ik}^{\mu}}{\Delta\alpha} = -\gamma_i P_{ik} - \gamma_i\bar{\eta}_i v_i\mathbb{E}[\delta g'(\lambda_i)\rho_k] \sim O(1) \tag{B20}$$

$$\frac{dU_{ik}}{d\alpha} = \lim_{N\to\infty} \frac{U_{ik}^{\mu+1} - U_{ik}^{\mu}}{\Delta\alpha} = -\gamma_i U_{ik} + (1-\frac{\gamma_i}{N}-\frac{\gamma_k}{N}+\frac{\gamma_i\gamma_k}{N^2})S_{ik} - \frac{\gamma_k\bar{\eta}_k}{N}(1-\frac{\gamma_i}{N})v_k\mathbb{E}[\delta g'(\lambda_k)\theta_i] \tag{B21}$$

$$- \bar{\eta}_i v_i\mathbb{E}[\delta g'(\lambda_i)\lambda_k] - \frac{\gamma_i\bar{\eta}_i}{N}(1-\frac{\gamma_k}{N})v_i\mathbb{E}[\delta g'(\lambda_i)\theta_k] \tag{B22}$$

$$+ \Delta^2\frac{\gamma_i\gamma_k\bar{\eta}_i\bar{\eta}_k}{N}v_iv_k\mathbb{E}[\delta^2 g'(\lambda_i)g'(\lambda_k)] \tag{B23}$$

$$= -\gamma_i U_{ik} + S_{ik} - \gamma_i\bar{\eta}_i v_i\mathbb{E}[\delta g'(\lambda_i)\lambda_k] \tag{B24}$$

$$\sim O(1) \tag{B25}$$

Appendix C: Gaussian Integrals

We can write closed form expressions for the above dynamical equations by solving the gaussian integrals for each expectation. At this point it is helpful to define primitives

$$F1_{ik} = \mathbb{E}[\delta g'(\lambda_i)\theta_k] \tag{C1}$$
$$F2_{ik} = \mathbb{E}[\delta g'(\lambda_i)\lambda_k] \tag{C2}$$
$$F3_{ik} = \mathbb{E}[\delta g'(\lambda_i)\rho_k] \tag{C3}$$
$$G_{ik} = \mathbb{E}[\delta^2 g'(\lambda_i)g'(\lambda_k)] \tag{C4}$$
$$H_i = \mathbb{E}[\delta g(\lambda_i)] \tag{C5}$$
$$J2_{ik} = \mathbb{E}[g'(\lambda_i)g'(\lambda_k)] \tag{C6}$$
$$\tag{C7}$$

Recall that our local fields are jointly distributed via a multivariate gaussian. With the erf activation function, we can compute closed forms of each of the primitives as originally done in [1]. The following definitions come from [2]. For a generic random variable drawn from a jointly gaussian distribution with covariance $\Sigma \in \mathbb{R}^{N\times N}$

$$x \sim N(0, \Sigma) \tag{C8}$$

We can write the following closed form expressions where $g = erf(x/\sqrt{2})$

$$I_2(r,s) = \mathbb{E}_x[g(x_r)g(x_s))] = \frac{2}{\pi}\arcsin\frac{\Sigma_{r,s}}{\sqrt{1+\Sigma_{s,s}}\sqrt{1+\Sigma_{r,r}}} \tag{C9}$$

$$I_3(r,s,t) = \mathbb{E}_x[g'(x_r)x_s g(x_t)] = \frac{2\Sigma_{st}(1+\Sigma_{rr}) - 2\Sigma_{rs}\Sigma_{rt}}{\pi\sqrt{\Lambda_3}(1+\Sigma_{rr})} \tag{C10}$$

$$I_4(r,s,t,u) = \mathbb{E}_x[g(x_r)g(x_s)g'(x_t)g'(x_u)] = \frac{4}{\pi^2\sqrt{\Lambda_4}}\arcsin\frac{\Lambda_0}{\sqrt{\Lambda_1\Lambda_2}} \tag{C11}$$

$$I_5(r,s) = \mathbb{E}_x[g'(x_r)g'(x_s)] = \frac{2}{\pi}v_i v_k \frac{1}{\sqrt{1+\Sigma_{r,r}+\Sigma_{s,s}+\Sigma_{r,r}\Sigma_{s,s}-\Sigma_{r,s}^2}} \tag{C12}$$

With the following definitions:

$$\Lambda_0 = \Lambda_4\Sigma_{r,s} - \Sigma_{r,u}\Sigma_{s,u}(1+\Sigma_{tt}) - \Sigma_{rt}\Sigma_{st}(1+\Sigma_{uu}) + \Sigma_{tu}\Sigma_{rt}\Sigma_{su} + \Sigma_{tu}\Sigma_{st}\Sigma_{ru} \tag{C13}$$

$$\Lambda_1 = \Lambda_4(1+\Sigma_{r,r}) - \Sigma_{r,u}^2(1+\Sigma_{tt}) - \Sigma_{rt}^2(1+\Sigma_{uu}) + 2\Sigma_{tu}\Sigma_{rt}\Sigma_{ru} \tag{C14}$$

$$\Lambda_2 = \Lambda_4(1+\Sigma_{s,s}) - \Sigma_{s,u}^2(1+\Sigma_{tt}) - \Sigma_{st}^2(1+\Sigma_{uu}) + 2\Sigma_{tu}\Sigma_{st}\Sigma_{su} \tag{C15}$$

$$\Lambda_3 = (1+\Sigma_{r,r})(1+\Sigma_{ss}) - \Sigma_{rs}^2 \tag{C16}$$

$$\Lambda_4 = (1+\Sigma_{t,t})(1+\Sigma_{uu}) - \Sigma_{tu}^2 \tag{C17}$$

Now, we can expand each primitive. First notice, that the structure of each F is identical. For a generic index in x,

$$F_{ik} = \mathbb{E}[\delta g'(\lambda_i)x_k] = \sum_{j=1}^{K} v_j \mathbb{E}[x_k g'(\lambda_i)g(\lambda_j)] - \sum_{n=1}^{M} b_n \mathbb{E}[x_k g'(\lambda_i)g(\rho_n)] \tag{C18}$$

$$= \sum_{j=1}^{K} v_j I_3(i,k,j) - \sum_{n=1}^{M} b_n I_3(i,k,K+n) \tag{C19}$$

$$G_{ik} = \mathbb{E}[\delta^2 g'(\lambda_i)g'(\lambda_k)] = \sum_{j,l=1}^{K} v_j v_l \,\mathbb{E}\left[g(\lambda_j)g(\lambda_l)g'(\lambda_i)g'(\lambda_k)\right] + \sum_{n,m=1}^{M} b_n b_m \,\mathbb{E}\left[g(\rho_n)g(\rho_m)g'(\lambda_i)g'(\lambda_k)\right] \tag{C20}$$

$$- 2\sum_{j=1}^{K}\sum_{n=1}^{M} v_j b_n \,\mathbb{E}\left[g(\lambda_j)g(\rho_n)g'(\lambda_i)g'(\lambda_k)\right] \tag{C21}$$

$$= \sum_{j,l=1}^{K} v_j v_l \, I_4(K+n, K+m, i, k) + \sum_{n,m=1}^{M} b_n b_m I_4(j,l,i,k) \tag{C22}$$

$$- 2\sum_{j=1}^{K}\sum_{n=1}^{M} v_j b_n I_4(j, K+m, i, k) \tag{C23}$$

$$H_i = \mathbb{E}[\delta g(\lambda_i)] = \sum_{j=1}^{K} v_j \mathbb{E}[g(\lambda_i)g(\lambda_j)] - \sum_{n=1}^{M} b_n \mathbb{E}[g(\lambda_i)g(\rho_n)] \tag{C24}$$

$$= \sum_{j=1}^{K} v_j I_2(i,j) - \sum_{n=1}^{M} b_n I_2(i, K+n) \tag{C25}$$

$$J2_{ik} = \mathbb{E}[g'(\lambda_i)g'(\lambda_k)] = I_5(i,k) \tag{C26}$$

Generalization Error

We can compute the generalization error as the expectation of the squared loss over the data distribution as follows:

$$\varepsilon_g = \mathbb{E}[\delta^2] = \sum_{j,l=1}^{K} v_j v_l \, E\left[g(\lambda_j) g(\lambda_l)\right] + \sum_{n,m=1}^{M} b_n b_m \, E\left[g(\rho_n) g(\rho_m) g\right] - \tag{C27}$$

$$2 \sum_{j=1}^{K} \sum_{n=1}^{M} v_j b_n \, E\left[g(\lambda_j) g(\rho_n)\right] \tag{C28}$$

$$= \sum_{j,l=1}^{K} v_j v_l \, I_2(j,l) + \sum_{n,m=1}^{M} b_n b_m \, I_2(K+n, K+m) - \tag{C29}$$

$$2 \sum_{j=1}^{K} \sum_{n=1}^{M} v_j b_n, I_2(j, K+n) \tag{C30}$$

Where I_2 is a guassian integral with the form:

$$I_2(r,s) = \frac{2}{\pi} \arcsin \frac{\Sigma_{r,s}}{\sqrt{1+\Sigma_{s,s}}\sqrt{1+\Sigma_{r,r}}} \tag{C31}$$

Consequently, we can write generalization error as a function of our order parameters.

$$\varepsilon_g = \frac{2}{\pi} \Bigg(\sum_{j,l=1}^{K} v_j v_l \arcsin \frac{\Delta^2 Q_{jl}}{\sqrt{1+\Delta^2 Q_{ll}}\sqrt{1+\Delta^2 Q_{j,j}}} + \sum_{n,m=1}^{M} b_n b_m \arcsin \frac{\Delta^2 T_{nm}}{\sqrt{1+\Delta^2 T_{nn}}\sqrt{1+\Delta^2 T_{m,m}}} \tag{C32}$$

$$- 2 \sum_{j=1}^{K} \sum_{n=1}^{M} v_j b_n \arcsin \frac{\Delta^2 R_{jn}}{\sqrt{1+\Delta^2 R_{jj}}\sqrt{1+\Delta^2 R_{nn}}} \Bigg) \tag{C33}$$

Gaussian Primitive Expansion

I acknowledge ChatGPT for the assistance with the expansions.
We order the local fields as

$$z = (\lambda_1, \ldots, \lambda_K, \rho_1, \ldots, \rho_M, \theta_1, \ldots, \theta_K),$$

with covariance

$$\Sigma = \Delta^2 \begin{pmatrix} Q & R & U^\top \\ R^\top & T & P^\top \\ U & P & S \end{pmatrix}.$$

Equivalently,

$$\Sigma_{\lambda_i \lambda_j} = \Delta^2 Q_{ij}, \qquad \Sigma_{\lambda_i \rho_n} = \Delta^2 R_{in}, \qquad \Sigma_{\rho_n \rho_m} = \Delta^2 T_{nm}, \tag{C34}$$

$$\Sigma_{\theta_k \lambda_i} = \Delta^2 U_{ki}, \qquad \Sigma_{\theta_k \rho_n} = \Delta^2 P_{kn}, \qquad \Sigma_{\theta_k \theta_\ell} = \Delta^2 S_{k\ell}. \tag{C35}$$

Recall that

$$\delta = \sum_{j=1}^{K} v_j g(\lambda_j) - \sum_{n=1}^{M} b_n g(\rho_n).$$

Define the following denominators:

$$D^{\theta}_{ik} = \sqrt{(1+\Delta^2 Q_{ii})(1+\Delta^2 S_{kk}) - \Delta^4 U_{ki}^2}, \tag{C36}$$

$$D^{\lambda}_{ik} = \sqrt{(1+\Delta^2 Q_{ii})(1+\Delta^2 Q_{kk}) - \Delta^4 Q_{ik}^2}, \tag{C37}$$

$$D^{\rho}_{ik} = \sqrt{(1+\Delta^2 Q_{ii})(1+\Delta^2 T_{kk}) - \Delta^4 R_{ik}^2}. \tag{C38}$$

Using the closed form for I_3, we obtain

$$F1_{ik} = \mathbb{E}[\delta g'(\lambda_i)\theta_k] \tag{C39}$$

$$= \sum_{j=1}^{K} v_j \frac{2\Delta^2 U_{kj}(1+\Delta^2 Q_{ii}) - 2\Delta^4 U_{ki} Q_{ij}}{\pi(1+\Delta^2 Q_{ii}) D^{\theta}_{ik}} - \sum_{n=1}^{M} b_n \frac{2\Delta^2 P_{kn}(1+\Delta^2 Q_{ii}) - 2\Delta^4 U_{ki} R_{in}}{\pi(1+\Delta^2 Q_{ii}) D^{\theta}_{ik}}, \tag{C40}$$

$$F2_{ik} = \mathbb{E}[\delta g'(\lambda_i)\lambda_k] \tag{C41}$$

$$= \sum_{j=1}^{K} v_j \frac{2\Delta^2 Q_{kj}(1+\Delta^2 Q_{ii}) - 2\Delta^4 Q_{ik} Q_{ij}}{\pi(1+\Delta^2 Q_{ii}) D^{\lambda}_{ik}} - \sum_{n=1}^{M} b_n \frac{2\Delta^2 R_{kn}(1+\Delta^2 Q_{ii}) - 2\Delta^4 Q_{ik} R_{in}}{\pi(1+\Delta^2 Q_{ii}) D^{\lambda}_{ik}}, \tag{C42}$$

$$F3_{ik} = \mathbb{E}[\delta g'(\lambda_i)\rho_k] \tag{C43}$$

$$= \sum_{j=1}^{K} v_j \frac{2\Delta^2 R_{jk}(1+\Delta^2 Q_{ii}) - 2\Delta^4 R_{ik} Q_{ij}}{\pi(1+\Delta^2 Q_{ii}) D^{\rho}_{ik}} - \sum_{n=1}^{M} b_n \frac{2\Delta^2 T_{kn}(1+\Delta^2 Q_{ii}) - 2\Delta^4 R_{ik} R_{in}}{\pi(1+\Delta^2 Q_{ii}) D^{\rho}_{ik}}. \tag{C44}$$

For the readout primitive,

$$H_i = \mathbb{E}[\delta g(\lambda_i)] \tag{C45}$$

$$= \sum_{j=1}^{K} v_j \frac{2}{\pi} \arcsin\left(\frac{\Delta^2 Q_{ij}}{\sqrt{(1+\Delta^2 Q_{ii})(1+\Delta^2 Q_{jj})}}\right) - \sum_{n=1}^{M} b_n \frac{2}{\pi} \arcsin\left(\frac{\Delta^2 R_{in}}{\sqrt{(1+\Delta^2 Q_{ii})(1+\Delta^2 T_{nn})}}\right). \tag{C46}$$

For the derivative–derivative primitive,

$$J2_{ik} = \mathbb{E}[g'(\lambda_i) g'(\lambda_k)] \tag{C47}$$

$$= \frac{2}{\pi} \frac{1}{\sqrt{1+\Delta^2 Q_{ii} + \Delta^2 Q_{kk} + \Delta^4 Q_{ii} Q_{kk} - \Delta^4 Q_{ik}^2}}. \tag{C48}$$

Now we expand

$$G_{ik} = \mathbb{E}[\delta^2 g'(\lambda_i) g'(\lambda_k)].$$

First define

$$\Lambda_4^{ik} = (1+\Delta^2 Q_{ii})(1+\Delta^2 Q_{kk}) - \Delta^4 Q_{ik}^2.$$

The student–student contribution is

$$G^{SS}_{ik} = \sum_{j=1}^{K} \sum_{\ell=1}^{K} v_j v_\ell \frac{4}{\pi^2 \sqrt{\Lambda_4^{ik}}} \arcsin\left(\frac{\Lambda^{SS}_{0;j\ell ik}}{\sqrt{\Lambda^{SS}_{1;j\ell ik} \Lambda^{SS}_{2;j\ell ik}}}\right), \tag{C49}$$

where

$$\Lambda^{SS}_{0;j\ell ik} = \Lambda_4^{ik} \Delta^2 Q_{j\ell} - \Delta^4 Q_{jk} Q_{\ell k}(1+\Delta^2 Q_{ii}) - \Delta^4 Q_{ji} Q_{\ell i}(1+\Delta^2 Q_{kk}) \tag{C50}$$

$$+ \Delta^6 Q_{ik} Q_{ji} Q_{\ell k} + \Delta^6 Q_{ik} Q_{\ell i} Q_{jk}, \tag{C51}$$

$$\Lambda^{SS}_{1;j\ell ik} = \Lambda_4^{ik}(1+\Delta^2 Q_{jj}) - \Delta^4 Q_{jk}^2(1+\Delta^2 Q_{ii}) - \Delta^4 Q_{ji}^2(1+\Delta^2 Q_{kk}) \tag{C52}$$

$$+ 2\Delta^6 Q_{ik} Q_{ji} Q_{jk}, \tag{C53}$$

$$\Lambda^{SS}_{2;j\ell ik} = \Lambda_4^{ik}(1+\Delta^2 Q_{\ell\ell}) - \Delta^4 Q_{\ell k}^2(1+\Delta^2 Q_{ii}) - \Delta^4 Q_{\ell i}^2(1+\Delta^2 Q_{kk}) \tag{C54}$$

$$+ 2\Delta^6 Q_{ik} Q_{\ell i} Q_{\ell k}. \tag{C55}$$

The teacher–teacher contribution is

$$G_{ik}^{TT} = \sum_{n=1}^{M}\sum_{m=1}^{M} b_n b_m \frac{4}{\pi^2\sqrt{\Lambda_4^{ik}}} \arcsin\left(\frac{\Lambda_{0;nmik}^{TT}}{\sqrt{\Lambda_{1;nmik}^{TT}\Lambda_{2;nmik}^{TT}}}\right), \tag{C56}$$

where

$$\Lambda_{0;nmik}^{TT} = \Lambda_4^{ik}\Delta^2 T_{nm} - \Delta^4 R_{kn}R_{km}(1+\Delta^2 Q_{ii}) - \Delta^4 R_{in}R_{im}(1+\Delta^2 Q_{kk}) \tag{C57}$$
$$+ \Delta^6 Q_{ik}R_{in}R_{km} + \Delta^6 Q_{ik}R_{im}R_{kn}, \tag{C58}$$
$$\Lambda_{1;nmik}^{TT} = \Lambda_4^{ik}(1+\Delta^2 T_{nn}) - \Delta^4 R_{kn}^2(1+\Delta^2 Q_{ii}) - \Delta^4 R_{in}^2(1+\Delta^2 Q_{kk}) \tag{C59}$$
$$+ 2\Delta^6 Q_{ik}R_{in}R_{kn}, \tag{C60}$$
$$\Lambda_{2;nmik}^{TT} = \Lambda_4^{ik}(1+\Delta^2 T_{mm}) - \Delta^4 R_{km}^2(1+\Delta^2 Q_{ii}) - \Delta^4 R_{im}^2(1+\Delta^2 Q_{kk}) \tag{C61}$$
$$+ 2\Delta^6 Q_{ik}R_{im}R_{km}. \tag{C62}$$

The student–teacher cross contribution is

$$G_{ik}^{ST} = -2\sum_{j=1}^{K}\sum_{n=1}^{M} v_j b_n \frac{4}{\pi^2\sqrt{\Lambda_4^{ik}}} \arcsin\left(\frac{\Lambda_{0;jnik}^{ST}}{\sqrt{\Lambda_{1;jnik}^{ST}\Lambda_{2;jnik}^{ST}}}\right), \tag{C63}$$

where

$$\Lambda_{0;jnik}^{ST} = \Lambda_4^{ik}\Delta^2 R_{jn} - \Delta^4 Q_{jk}R_{kn}(1+\Delta^2 Q_{ii}) - \Delta^4 Q_{ji}R_{in}(1+\Delta^2 Q_{kk}) \tag{C64}$$
$$+ \Delta^6 Q_{ik}Q_{ji}R_{kn} + \Delta^6 Q_{ik}R_{in}Q_{jk}, \tag{C65}$$
$$\Lambda_{1;jnik}^{ST} = \Lambda_4^{ik}(1+\Delta^2 Q_{jj}) - \Delta^4 Q_{jk}^2(1+\Delta^2 Q_{ii}) - \Delta^4 Q_{ji}^2(1+\Delta^2 Q_{kk}) \tag{C66}$$
$$+ 2\Delta^6 Q_{ik}Q_{ji}Q_{jk}, \tag{C67}$$
$$\Lambda_{2;jnik}^{ST} = \Lambda_4^{ik}(1+\Delta^2 T_{nn}) - \Delta^4 R_{kn}^2(1+\Delta^2 Q_{ii}) - \Delta^4 R_{in}^2(1+\Delta^2 Q_{kk}) \tag{C68}$$
$$+ 2\Delta^6 Q_{ik}R_{in}R_{kn}. \tag{C69}$$

Therefore the fully expanded primitive is

$$G_{ik} = G_{ik}^{SS} + G_{ik}^{TT} + G_{ik}^{ST}. \tag{C70}$$

Appendix D: Strong Symmetry Ansatz Fixed Point Calculation

Recall the plateau fixed point conditions:

$$\frac{dQ_{ik}}{d\alpha} = 0 \tag{D1}$$
$$\frac{dR_{ik}}{d\alpha} = 0 \tag{D2}$$
$$\frac{dS_{ik}}{d\alpha} = -2\gamma S_{ik} - \gamma\eta F1_{ik} - \gamma\eta F1_{ki} + \Delta^2\gamma^2\eta^2\left(G_{ik} + \sigma^2 J2_{ik}\right) \tag{D3}$$
$$U = \begin{cases} S_{ii} = \gamma\bar{\eta}_i F2_{ii} \\ 0 = -\gamma(U_{ik}+U_{ki}) + S_{ik} + S_{ki} - 2\gamma\eta F2_{ki} \qquad i \neq k \end{cases} \tag{D4}$$
$$0 = \frac{dP_{ik}}{d\alpha} = F3_{ik} \tag{D5}$$

We can apply the strong symmetry ansatz to produce fixed point relations between Q, C, and R.

Stationarity of P: $F3_{ik} = 0$

We can expand the definition of $F3_{ik}$ for arbitrary i and k.

$$0 = F3_{ik} = \mathbb{E}[\delta g'(\lambda_i)\rho_k] \tag{D6}$$

$$= \sum_{j=1}^{K} \frac{2\Delta^2 R_{jk}(1+\Delta^2 Q_{ii}) - 2\Delta^4 R_{ik}Q_{ij}}{\pi(1+\Delta^2 Q_{ii})D^{\rho}_{ik}} - \sum_{n=1}^{M} \frac{2\Delta^2 T_{kn}(1+\Delta^2 Q_{ii}) - 2\Delta^4 R_{ik}R_{in}}{\pi(1+\Delta^2 Q_{ii})D^{\rho}_{ik}} \tag{D7}$$

where

$$D^{\rho}_{ik} = \sqrt{(1+\Delta^2 Q_{ii})(1+\Delta^2 T_{kk}) - \Delta^4 R^2_{ik}}. \tag{D8}$$

This implies that

$$0 = K\Delta^2 R(1+\Delta^2 C) - (K-1)\Delta^4 Rq - \Delta^4 RC - \Delta^2(1+\Delta^2 C) + M\Delta^4 R^2 \tag{D9}$$

We can rearrange this expression to relate our symmetry parameters:

$$q = \frac{KR + (K-1)\Delta^2 RC - 1 - \Delta^2 C + M\Delta^2 R^2}{(K-1)\Delta^2 R} \tag{D10}$$

or

$$C = \frac{KR - (K-1)\Delta^2 Rq - 1 + M\Delta^2 R^2}{\Delta^2(1 - RK + R)} \tag{D11}$$

Equation 2: Anti-Symmetry of U

Recall that S is always symmetric by construction (defined as an overlap). Under the strong symmetry ansatz, we assume that Q and R are symmetric so we it follows that $F2$ must also be symmetric. Any asymmetric evolution of U must arise from asymmetry in U itself. We've previously shown that at the plateau point $U_{ii} = 0$. Consider

$$U_{ik} = \frac{U_{ik} + U_{ki}}{2} + \frac{U_{ik} - U_{ki}}{2} \tag{D12}$$

At the fixed point, we our anti-symmetry condition simplifies to

$$\frac{dU_{ik}}{d\alpha} = \frac{1}{2}\left(\frac{dU_{ik}}{d\alpha} - \frac{dU_{ki}}{d\alpha}\right) \tag{D13}$$

$$\frac{dU_{ik}}{d\alpha} = \frac{1}{2}\left(\frac{dU_{ik}}{d\alpha} - \frac{dU_{ki}}{d\alpha}\right) = \frac{1}{2}\left(-\gamma U_{ik} + S_{ik} - \gamma\eta F2_{ik} + \gamma U_{ki} - S_{ki} + \gamma\eta F2_{ki}\right) \tag{D14}$$

Using the symmetry of S and $F2$, this equation simplifies to

$$\frac{dU_{ik}}{d\alpha} = \frac{1}{2}\left(\frac{dU_{ik}}{d\alpha} - \frac{dU_{ki}}{d\alpha}\right) = -\frac{\gamma}{2}\left(U_{ik} - U_{ki}\right) \tag{D15}$$

This implies that an off-diagonal element of U should decay corresponding to the difference between the symmetric pairs. For a non-transient fixed point, this collapses to $U_{ik} = U_{ki} = 0$. Therefore, we argue that $U = 0$ at the

symmetric fixed point under strong symmetry assumptions. Note that if the learning rate and memory parameters were not identical across sites this would introduce a driving parameter that would allow for an anti-symmetric U.

With the constraint that $U = 0$, we can drastically simplifies $F1$.

$$F1_{ik} = \mathbb{E}[\delta g'(\lambda_i)\theta_k] \tag{D16}$$

$$= \sum_{j=1}^{K} \frac{2\Delta^2 U_{kj}(1+\Delta^2 C) - 2\Delta^4 U_{ki} Q_{ij}}{\pi(1+\Delta^2 C) D^{\theta}_{ik}} + \sum_{n=1}^{M} \frac{2\Delta^4 U_{ki} R}{\pi(1+\Delta^2 C) D^{\theta}_{ik}} \tag{D17}$$

$$= 0 \tag{D18}$$

where we use the definition

$$D^{\theta}_{ik} = \sqrt{(1+\Delta^2 C)(1+\Delta^2 \gamma_k \bar{\eta}_k F2_{kk}) - \Delta^4 U^2_{ki}} \tag{D19}$$

Generically, we now have that for any indices i and k

$$S_{ik} = \gamma \eta F2_{ik} \tag{D20}$$

$$\frac{dS_{ik}}{d\alpha} = -2\gamma S_{ik} + \Delta^2 \gamma^2 \eta^2 \left(G_{ik} + \sigma^2 J2_{ik}\right) \tag{D21}$$

Notice that $F2$ is a function of C, R, q which are all static at the plateau points. Therefore, $F2$ and consequently S_{ik} are also static. We conclude that from the plateau point conditions, under strong symmetry we derive true fixed points. Now, we can derive the form of the final equations:

$$2F2_{ik} = \Delta^2 \eta (G_{ik} + \sigma^2 J2_{ik}) \tag{D22}$$

Interestingly, we preserve η dependence!

First, consider the diagonal entries:

$$\frac{4}{\pi} \frac{\Delta^2}{(1+\Delta^2 C)\sqrt{1+2\Delta^2 C}} \left(q(K-1) + C - MR \right) = \Delta^2 \eta G_{ii} + \frac{2\Delta^2 \eta \sigma^2}{\pi} \frac{1}{\sqrt{1+2\Delta^2 C}} \tag{D23}$$

Off-Diagonal:

$$\frac{4}{\pi} \frac{1}{(1+\Delta^2 C)\sqrt{1+2\Delta^2 C + \Delta^4 C^2 - \Delta^4 q^2}} \Big((K-1) q(1+\Delta C^2) + C(1+\Delta^2 C) \tag{D24}$$

$$- (K-1)\Delta^4 q^2 - \Delta^4 qC - MR(1+\Delta^2 C) + M\Delta^4 qR \Big) \tag{D25}$$

$$= \Delta^2 \eta G_{ik} + \frac{2\Delta^2 \eta \sigma^2}{\pi} \frac{1}{\sqrt{1+2\Delta^2 C + \Delta^4 C^2 - \Delta^4 q^2}} \tag{D26}$$

G has a non trivial expansion so it might be easiest to copmute this numerically.

Fixed Point Self-Consistency Equations

In conclusion, we get the following fixed point self-consistency equations:

$$q = \frac{KR + (K-1)\Delta^2 RC - 1 - \Delta^2 C + M\Delta^2 R^2}{(K-1)\Delta^2 R} \tag{D27}$$

$$\frac{4}{\pi}\frac{\Delta^2}{(1+\Delta^2 C)\sqrt{1+2\Delta^2 C}}\Big(q(K-1)+C-MR\Big)=\Delta^2\eta G_{ii}+\frac{2\Delta^2\eta\sigma^2}{\pi}\frac{1}{\sqrt{1+2\Delta^2 C}} \tag{D28}$$

$$\frac{4}{\pi}\frac{1}{(1+\Delta^2 C)\sqrt{1+2\Delta^2 C+\Delta^4 C^2-\Delta^4 q^2}}\Big((K-1)q(1+\Delta C^2)+C(1+\Delta^2 C) \tag{D29}$$

$$-(K-1)\Delta^4 q^2-\Delta^4 qC-MR(1+\Delta^2 C)+M\Delta^4 qR\Big) \tag{D30}$$

$$=\Delta^2\eta G_{ik}+\frac{2\Delta^2\eta\sigma^2}{\pi}\frac{1}{\sqrt{1+2\Delta^2 C+\Delta^4 C^2-\Delta^4 q^2}} \tag{D31}$$