Asymptotic Dynamics for Delayed Feature Learning in a Toy Model

Author Names Withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2024

Abstract

We consider a toy model that exhibits grokking, recently advanced by [9], and take advantage of the simple setting to derive the *dynamics* of the train and test loss using Dynamical Mean Field Theory (DMFT). This gives a closed-form expression for the gap between train and test loss that characterizes grokking in this toy model, illustrating how two parameters of interest – NTK alignment and network laziness – control the size of this gap and how grokking emerges as a uniquely *offline* property during repeated training over the same dataset. This is the first quantitative characterization of grokking dynamics in a general setting that makes no assumptions about weight decay, weight norm, etc.

1. Introduction

Grokking is an empirical phenomenon discovered by [15] where there is a sharp decrease in test error long after train loss reaches a low value. It is an example of abrupt and unexpected change in a neural network's performance, and has been under close empirical study since its discovery [6, 7, 10, 11, 14, 17, 18]. A number of theories have been proposed to explain the causes of grokking; initially, most attributed grokking to weight decay and weight norm decrease at late time, [11, 18]. Here, we consider a toy model from [9] that groks without weight decay or adaptive optimizers, and derive the test and train loss dynamics to identify the source of the separation between test and train loss.

Our toy model is a two-layer MLP trained on a polynomial regression task using vanilla gradient descent. Here, we use a technique from statistical physics called Dynamical Mean Field Theory to derive the full train and test loss *dynamics* of this toy model in terms of two parameters: alignment of the Neural Tangent Kernel [8] and network laziness [5]. Solving the resulting DMFT equations illustrates the origins of the test-train loss gap due to these two parameters.

2. Related work

Grokking. Since discovery by [15], there has been much work on understanding why it happens. [14] mechanistically interpret the algorithm learned during grokking for modular addition tasks, [17] attribute grokking to an optimization anomaly of adaptive optimizers. [6] observe that grokking and double descent share similarities in dynamics, [11] find empirically that parameter weight norm at initialization can control grokking, and [18] propose a heuristic for cross-entropy training with weight decay. [9] suggest grokking is when a network transitions from a lazy, kernel regime to a rich, feature learning regime.

Dynamical Mean Field Theory. DMFT is a mathematical technique that has its origin in condensed matter physics, but it has since then shown to be a useful tool to describe the weight and loss dynamics of neural networks [3, 4, 13] in the "thermodynamic limit" where dataset size and data dimension jointly go to infinity. [12] study the transition from a lazy to rich regime in a simple setting with weight decay, whereas our dynamical theory make no assumptions about weight decay and instead takes a proportional asymptotic in data dimension and number of data points.

3. The Setting: Polynomial regression in a two layer perceptron

We consider a one-hidden layer MLP with polynomial activations on a high-dimensional polynomial regression task. Because grokking persists when we fix the readout weighs, we do so in pursuit of the simplest setting in which to study grokking; this is the model studied by [9], though they do not provide any theory for the dynamics at play.

The model $f(\boldsymbol{w}, \boldsymbol{x})$ and target function $y(\boldsymbol{x})$ are defined in terms of input $\boldsymbol{x} \in \mathbb{R}^D$ as

$$f(\boldsymbol{w}, \boldsymbol{x}) = \frac{\alpha}{N} \sum_{i=1}^{N} \phi(\boldsymbol{w}_i \cdot \boldsymbol{x}) \,, \quad \phi(h) = h + \frac{\epsilon}{2} h^2 \,, \quad y(\boldsymbol{x}) = \frac{1}{2} (\boldsymbol{\beta}_\star \cdot \boldsymbol{x})^2 \tag{1}$$

The value of α controls the scale of the output, and consequently the speed of feature learning [5]. The value of ϵ alters how difficult the task is for the initial NTK since it controls the power the network puts in quadratic (vs linear) functions, which is the form the target takes. We train on a fixed dataset $\{(\boldsymbol{x}_\mu, y_\mu)\}_{\mu=1}^P$ of P samples. The inputs $\boldsymbol{x}$ are drawn from an isotropic Gaussian distribution $\boldsymbol{x} \sim \mathcal{N}(0, \frac{1}{D}\boldsymbol{I})$. We introduce the following two summary statistics (known in statistical physics as "order parameters") $\bar{\boldsymbol{w}} = \frac{1}{N}\sum_{i=1}^N \boldsymbol{w}_i \in \mathbb{R}^D \,, \quad \boldsymbol{M} = \frac{1}{N}\sum_{i=1}^N \boldsymbol{w}_i \boldsymbol{w}_i^\top \in \mathbb{R}^{D\times D}$ Using these two moments of the weights, we can write the neural network function as $f(\boldsymbol{x}) = \alpha \bar{\boldsymbol{w}} \boldsymbol{x} + \frac{\alpha\epsilon}{2} \boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{x}$. Then, we have that the NTK can be written as $K(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x} \cdot \boldsymbol{x}' + \epsilon (\boldsymbol{x} \cdot \boldsymbol{x}')\, \bar{\boldsymbol{w}} \cdot (\boldsymbol{x} + \boldsymbol{x}') + \epsilon^2 (\boldsymbol{x} \cdot \boldsymbol{x}')\, \boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{x}'$. At initialization in a wide network, $\bar{\boldsymbol{w}} = 0$ and $\boldsymbol{M} = \boldsymbol{I}$. With this setup, we proceed to derive the dynamics of this network to see what is the regime in α, ϵ, and the dataset size in which it exhibits grokking. 1 illustrates the learning curves of this toy model during training, showcasing a gap of several thousand epochs between the fall in train and test loss.

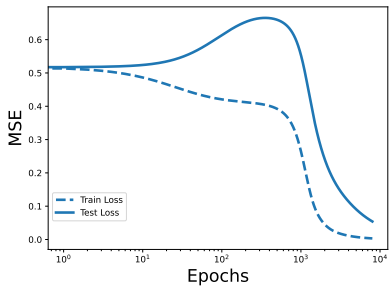


Figure 1: Grokking in our toy model. Fall in test loss happens thousands of epochs after fall in train loss (note the logarithmic scale in epochs). This used $\alpha = 0.5, \epsilon = 0.05$, and this test-train gap (i.e., time to grok) is increasing in α, for theoretical reasons we explain in the following section.

4. Train and Test Error for Grokking from DMFT

To gain insight into the typical train and test loss dynamics in this setting, we investigate the high dimensional limit $D, P \to \infty$ with $P = \gamma D$ with hidden neurons N held fixed, see Appendix A for details of the following derivation.

Intuitively, grokking happens in a data regime where we don't have too much data (otherwise the network learns features immediately), so the equations that result predict a gap between test and train can emerge as γ becomes smaller. We will see that α, ϵ show up in the resulting Equation (5) in a way that explains why they have the effects they do on grokking dynamics. The derivation of the DMFT equations is deferred to Appendix A, and we state and interpret the resulting equations here instead. We also present numerical solutions to these equations and show they illustrate where grokking comes from in this toy model. The DMFT is summarized by self-averaging (concentrating) correlation and response functions, which are defined as

$$C^w_{ij}(t,s) \equiv \frac{1}{D}\boldsymbol{w}_i(t)\cdot\boldsymbol{w}_j(t)\ ,\ A_i(t) \equiv \frac{1}{D}\boldsymbol{w}_i(t)\cdot\boldsymbol{\beta}_\star\ ,\ R^w_{ij}(t,s) \equiv \frac{1}{D}\sum_{k=1}^{D}\frac{\delta w_{ik}(t)}{\delta v^w_{jk}(s)}$$
$$C^g_{ij}(t,s) \equiv \frac{1}{P}\sum_{\mu=1}^{P} g^\mu_i(t)g^\mu_j(s)\ ,\ R^h_{ij}(t,s) \equiv \frac{1}{P}\sum_{\mu=1}^{P}\frac{\delta g^\mu_i(t)}{\delta h^\mu_j(s)}\ ,\ R^{h_\star}_i(t) \equiv \frac{1}{P}\sum_{\mu=1}^{P}\frac{\delta g^\mu_i(t)}{\delta h^\mu_\star}, \tag{2}$$

where $g^\mu_i(t) = \dot\phi(h^\mu_i(t))(y_\mu - f_\mu)$ are the gradients and the $v_{ik}(t)$ are fictitous source terms added to the weight dynamics (see Appendix A.5 for more details). In the high dimensional limit, the weights for each input dimension become statistically independent and identically distributed, allowing for a reduced (lower) dimensional description of the network dynamics. We provide some theoretical solutions in Figure 2 and Appendix Figure 3 obtained by integrating the DMFT equations.

Using solved dynamics to identify the source of grokking on our model. While the resulting equations coupling the correlation and response functions are complicated, we provide them below to see how grokking can emerge as these three introduced quantities of interest $(\alpha, \gamma, \epsilon)$ vary.

$$\frac{d}{dt}w_i(t) = u^w_i(t) + \int_0^t ds\sum_{j=1}^{N} R^h_{ij}(t,s)w_j(s) + R^{h_\star}_i(t)\beta\ ,\ u^w \sim \mathcal{GP}(0,\gamma^{-1}C_g) \tag{3}$$

$$h_i(t) = u^h_i(t) + \frac{1}{\gamma}\int_0^t ds\sum_{j=1}^{N} R^w_{ij}(t,s)g_j(s)\ ,\ [u^h_i, u^{h_\star}] \sim \mathcal{GP}\left(0, Q\right)\ ,\ Q = \begin{bmatrix} C_w & A \\ A^\top & 1\end{bmatrix} \tag{4}$$

where C^w, A, C^g are correlations computed over the statistics of the above stochastic processes.

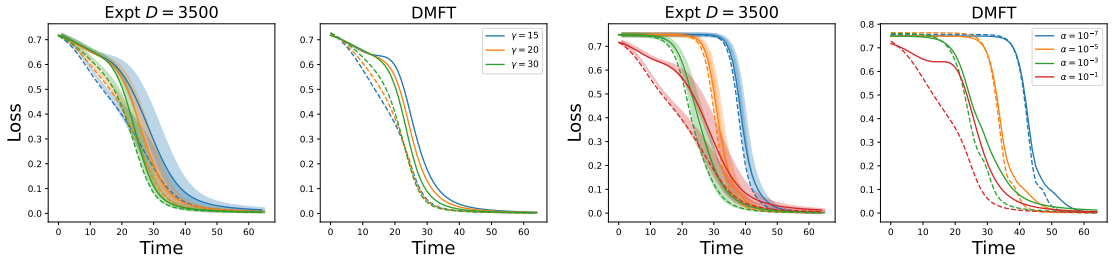
Overfitting effects at finite γ Now we analyze what (3) and (4) tells us about the effect of γ on the initial memorization (overfitting) phase of grokking. While the $\gamma \to \infty$ limit recovers the dynamics of online learning from [16], at finite γ we have the following complications

- Noise $u^w(t)$ in the weight dynamics which is correlated across time and neurons
- Non-Gaussian correction to $h_i(t)$ (the integral in h equation) which accelerates overfitting.
- The response function $R^h_{ij}(t,s)$ becomes non-local in time (not proportional to $\delta(t-s)$), generating non-Markovian weight updates.

Overfitting is caused by the finite γ correction to $h_i(t)$, generating a gap in train and test loss

$$\mathcal{L}_{\text{test}}(t) - \mathcal{L}_{\text{train}}(t) \sim \frac{2\alpha}{\gamma N} \sum_{ij=1}^{N} \int_0^t ds R_{ij}^w(t,s) C_{ij}^g(t,s) + \mathcal{O}(\gamma^{-2}) \tag{5}$$

This quantitative characterization of the test-train gap is a key result of the DMFT because it allows us to see exactly where grokking comes from as a function of γ, α, ϵ, since grokking takes place exactly when $\mathcal{L}_{\text{test}}(t) - \mathcal{L}_{\text{train}}(t)$ is large. The key idea is that grokking cannot happen in the online limit $\gamma \to \infty$ because train loss tracks test loss, and grokking is precisely when train loss is egregiously unrepresentative of progress on improving test risk. Equation 5 illustrates how this gap can become large (grokking) in the middle of training as we move from the online to the offline training regime. But γ is not the only parameter in this model that quantitatively controls grokking, as we will see now.



(*a*) $\alpha = 0.1$, Vary γ (*b*) $\gamma = 15$, Vary α

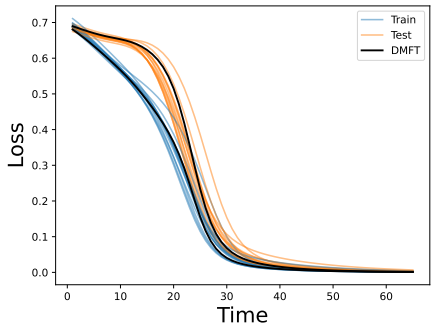
Figure 2: We compare the train (dashed) and test (solid) losses of different networks with $N = 2$ hidden neurons and data of dimension $D = 3500$ in a randomly initialized network (with constant initialization) and random training sets of size $P = \gamma D$ to the DMFT train and test loss predictions. Error bars represent standard deviation over 10 datasets. See Appendix Figure 3 for additional plots of DMFT correlation and response functions.

Effect of ϵ and α The scale parameter α and quadratic component ϵ enter into the dynamics through the gradient signals $g_i(t) = [1 + \epsilon h_i(t)] \left(\frac{1}{2} h_\star^2 - \frac{\alpha}{N} \sum_k \left[h_k + \frac{\epsilon}{2} h_k^2 \right] \right)$ which are averaged to compute C^g. This indicates that ϵ and α both have interesting effects on response functions $R^h, R^{h_\star}$.

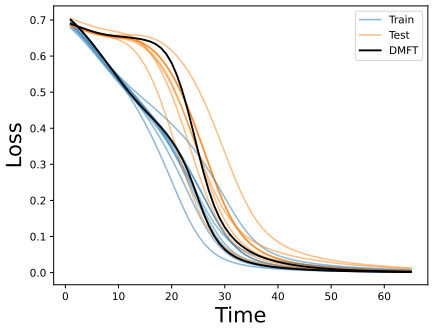
- First, we see that if α is small, the train and test losses will be closer to one another (no grokking) based on equation (5). Early in training the weight dynamics are dominated by updates in the signal β direction, representing useful feature learning. The losses will both begin to drop at time $t \sim \frac{2}{3\epsilon} \ln \left(\frac{1}{A_0 \alpha} \right)$, where A_0 is the scale of initial alignment (see App. A.3 and Figure 2 b). For large α, R_h will have a large effect on the dynamics, delaying alignment and causing grokking.

- If ϵ is small, then we expect the alignment to initially decrease followed by a subsequent increase at later time (more grokking). If ϵ is large, then we expect it is possible for h_k to evolve close to an interpolation condition before w aligns to β. All these predictions are tested in Figure 2.

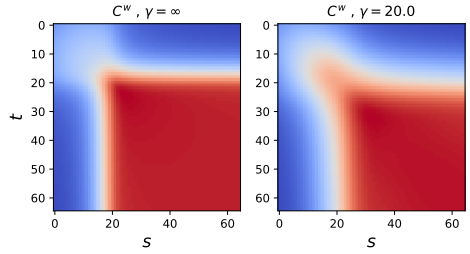
Above in 2(a), we can see first the offline effects described earlier of $\gamma \ll \infty$. In 2(b) we can see how making α bigger (and thus the network lazier, so it tracks its linearized dynamics more closely for longer during training [5]) causes a gap to emerge between the train and test loss. We include more detailed derivations in the Appendix that discuss more of the interpretation of the DMFT equations and their origins. This is the first quantitative characterization of grokking *dynamics*, to the best of our knowledge, that makes no assumptions about weight norm, weight decay, etc.



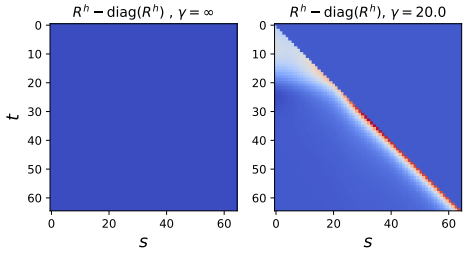
(*a*) $\gamma = 20$



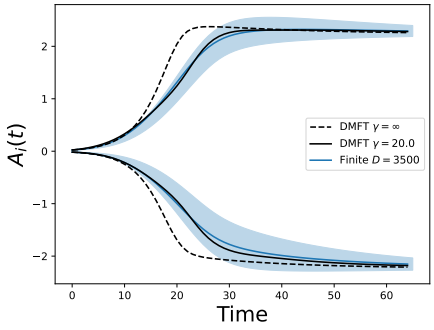
(*b*) $\gamma = 15$



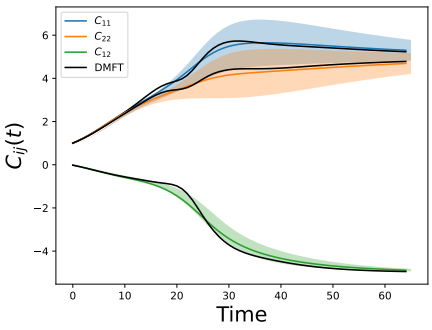
(*c*) Weight Correlation Online vs Offline



(*d*) Preactivation Response Online vs Offline



(*e*) Alignment Dynamics



(*f*) Weight Correlations

Figure 3: DMFT captures the effect of data repetition in the Gaussian data model. We see that data repetition (finite γ) causes (a) a change in the weight correlations, and the (b) preactivation response functions become non-diagonal reflecting a transition to non-Markovian dynamics. (c) The alignment of each of the two neurons to $\beta_\star$ is well predicted by DMFT but not the online limit. (d) The time-time diagonal entries in C^w are well predicted by DMFT.

References

[1] Luca Arnaboldi, Ludovic Stephan, Florent Krzakala, and Bruno Loureiro. From high-dimensional & mean-field dynamics to dimensionless odes: A unifying approach to sgd in two-layers networks. *arXiv preprint arXiv:2302.05882*, 2023.

[2] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *The Journal of Machine Learning Research*, 22(1):4788–4838, 2021.

[3] Blake Bordelon and Cengiz Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. *Advances in Neural Information Processing Systems*, 35: 32240–32256, 2022.

[4] Michael Celentano, Chen Cheng, and Andrea Montanari. The high-dimensional asymptotics of first order methods with random data. *arXiv preprint arXiv:2112.07572*, 2021.

[5] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

[6] Xander Davies, Lauro Langosco, and David Krueger. Unifying grokking and double descent. *arXiv preprint arXiv:2303.06173*, 2023.

[7] Andrey Gromov. Grokking modular arithmetic. *arXiv preprint arXiv:2301.02679*, 2023.

[8] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[9] Tanishq Kumar, Blake Bordelon, Samuel J Gershman, and Cengiz Pehlevan. Grokking as the transition from lazy to rich training dynamics. *arXiv preprint arXiv:2310.06110*, 2023.

[10] Ziming Liu, Ouail Kitouni, Niklas S Nolte, Eric Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. *Advances in Neural Information Processing Systems*, 35:34651–34663, 2022.

[11] Ziming Liu, Eric J Michaud, and Max Tegmark. Omnigrok: Grokking beyond algorithmic data. *arXiv preprint arXiv:2210.01117*, 2022.

[12] Kaifeng Lyu, Jikai Jin, Zhiyuan Li, Simon Shaolei Du, Jason D Lee, and Wei Hu. Dichotomy of early and late phase implicit biases can provably induce grokking. In *The Twelfth International Conference on Learning Representations*, 2023.

[13] Francesca Mignacco, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Dynamical mean-field theory for stochastic gradient descent in gaussian mixture classification. *Advances in Neural Information Processing Systems*, 33:9540–9550, 2020.

[14] Neel Nanda, Lawrence Chan, Tom Liberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

[15] Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

[16] David Saad and Sara A Solla. On-line learning in soft committee machines. *Physical Review E*, 52(4):4225, 1995.

[17] Vimal Thilak, Etai Littwin, Shuangfei Zhai, Omid Saremi, Roni Paiss, and Joshua Susskind. The slingshot mechanism: An empirical study of adaptive optimizers and the grokking phenomenon. *arXiv preprint arXiv:2206.04817*, 2022.

[18] Vikrant Varma, Rohin Shah, Zachary Kenton, János Kramár, and Ramana Kumar. Explaining grokking through circuit efficiency. *arXiv preprint arXiv:2309.02390*, 2023.

Appendix A. Dynamical Mean Field Theory for Asymptotic Risk Evolution

In this section we derive a set of DMFT equations which describe the dynamics of a two-layer network with $N = \mathcal{O}(1)$ hidden neurons in a proportional limit $P, D \to \infty$ with $P = \gamma D$. The general framework for DMFT for models trained on random data can be found in prior works [4, 13]. Our goal is to analyze the following gradient flow dynamics

$$\frac{d}{dt}\boldsymbol{w}_i(t) = \frac{1}{\gamma}\sum_{\mu=1}^{P} g_i^\mu(t)\boldsymbol{x}^\mu \ , \ g_i^\mu(t) = (y_\mu - f_\mu)\dot{\phi}(h_i^\mu(t)) \ , \ h_i^\mu(t) = \boldsymbol{w}_i(t)\cdot\boldsymbol{x}^\mu \tag{6}$$

where the labels $y^\mu = \sigma(h_\star^\mu) = \sigma(\boldsymbol{\beta}\cdot\boldsymbol{x})$ are determined by the target function and $f_\mu(t) = \frac{\alpha}{N}\sum_{i=1}^{N}\phi(h_i^\mu(t))$ is the model's output on data point $\boldsymbol{x}^\mu$. The vector $\boldsymbol{\beta}$ can be regarded as a random vector independent entries with $\langle\beta^2\rangle = 1$. The key idea of DMFT is that the above dynamics on ND weights $\{\boldsymbol{w}_i\}_{i=1}^N$ can be reduced to $2N+1$ stochastic processes (One weight $w_i(t)$ for each hidden neuron and one gradient for each hidden neuron $g_i(t)$ and one additional for the target field $h_\star$) which are summarized by correlation and response functions. Each entry of the weight vector $\boldsymbol{w}_i$ becomes an independent stochastic process with the following dynamics

$$\begin{aligned}
\frac{d}{dt}w_i(t) &= u_i^w(t) + \int_0^t ds\ \sum_{j=1}^{N} R_{ij}^h(t,s)w_j(s) + R_i^{h_\star}(t)\beta \ , \ \boldsymbol{u}_w \sim \mathcal{GP}\left(0, \gamma^{-1}\boldsymbol{C}^g\right) \\
h_i(t) &= u_i^h(t) + \frac{1}{\gamma}\sum_{j=1}^{N}\int_0^t ds R_{ij}^w(t,s)g_j(s) \ , \ h_\star = u^{h^\star} \\
C_{ij}^g(t,s) &= \langle g_i(t)g_j(s)\rangle \ , \ C_{ij}^w(t,s) = \langle w_i(t)w_j(s)\rangle
\end{aligned} \tag{7}$$

where $u_i^h(t), u_{h^\star}$ are jointly Gaussian with covariance

$$\left\langle u_i^h(t)u_j^h(s)\right\rangle = C^w(t,s) \ , \ \left\langle u_i^h(t)u^{h^\star}\right\rangle = A_i(t) \equiv \langle w_i(t)\beta\rangle \ , \ \left\langle\left[u^{h^\star}\right]^2\right\rangle = \langle\beta^2\rangle = 1 \tag{8}$$

Lastly, we can compute the response functions R as averages over the above stochastic processes

$$R^w_{ij}(t,s) = \left\langle \frac{\delta w_i(t)}{\delta u^w_j(s)} \right\rangle \ , \ R^h_{ij}(t,s) = \left\langle \frac{\delta g_i(t)}{\delta u^h_j(s)} \right\rangle \ , \ R^{h_\star}_i(t) = \left\langle \frac{\delta g_i(t)}{\delta u^{h_\star}} \right\rangle \tag{9}$$

These equations close self-consistently. We derive these equations in Appendix A.5 using a dynamical cavity method. In the next section, we discuss some special limits of this theory where we can gain some insight. We note that while the stochastic process for $w_i(t)$ is Gaussian for any choice of nonlinearity, the preactivations $h_i(t)$ are non-Gaussian if we use nonlinear link functions ϕ.

A.1. The Large γ Limit Coincides With Online Learning Dynamics

To gain insight into the equations, we first investigate the large data limit $\gamma \to \infty$. Unfortunately, this limit cannot exhibit grokking as train and test losses coincide, however it will prove a useful starting point for our analysis. In this case, we have the following simplifications

$$\begin{aligned} &R^h_{ij}(t,s) \to R^h_{ij}(t)\delta(t-s) \ , \ h_i(t) \to u^h_i(t) \\ &\frac{d}{dt} w_i(t) \to \sum_{j=1}^N R^h_{ij}(t) w_j(t) + R^{h_\star}_i(t)\beta \end{aligned} \tag{10}$$

Using these simplifications, we can close the equations for C_w and A_i as

$$\frac{d}{dt} A_i(t) = \sum_j R^h_{ij}(t) A_j(t) + R^{h_\star}_i(t) \tag{11}$$

$$\frac{d}{dt} C^w_{ij}(t,s) = \sum_k R^h_{ik}(t) C^h_{kj}(t,s) + R^{h_\star}_i(t) A_j(s) \tag{12}$$

where we have the equal-time response functions

$$R^h_{ij}(t,s) = \left\langle \frac{\delta g_i(t)}{\delta u^h_j(t)} \right\rangle \ , \ R^{h_\star}_i(t) = \left\langle \frac{\delta g_i(t)}{\delta u^{h_\star}} \right\rangle \tag{13}$$

This exactly coincides with the online learning equations of [16]. For our choice of noninearities $\phi(h) = h + \frac{\epsilon}{2}h^2$ and $\sigma(h) = \frac{1}{2}h^2$, we have the following response functions

$$\begin{aligned} R^h_{ij}(t) &= \delta_{ij}\epsilon \left\langle \left(\frac{1}{2}h^2_\star - \frac{1}{N}\sum_k \left[h_k(t) + \frac{\epsilon}{2} h_k(t)^2 \right] \right) \right\rangle \\ &\quad - \frac{1}{N} \left\langle \left(1 + \epsilon h_i(t)\right)\left(1 + \epsilon h_j(t)\right) \right\rangle \\ &= \frac{\epsilon}{2}\delta_{ij} \left[1 - \frac{\epsilon}{N}\sum_k C^w_{kk}(t) \right] - \frac{1}{N}\left[1 + \epsilon^2 C_{ij}(t) \right] \\ R^{h_\star}(t) &= \langle (1 + \epsilon h_i(t)) h_\star \rangle = \epsilon A_i(t) \end{aligned} \tag{14}$$

Thus, we obtain the following closed ordinary differential equations for $\{A_i(t), C^w_{ij}(t,s)\}$

$$\frac{d}{dt}A_i(t) = \frac{\epsilon}{2}\left[1 - \frac{1}{N}\sum_k C^w_{kk}(t)\right]A_i(t) - \frac{1}{N}\sum_j [1+\epsilon^2 C^w_{ij}(t)]A_j(t) + \epsilon A_i(t)$$
$$\frac{d}{dt}C^w_{ij}(t) = \epsilon\left[1 - \frac{1}{N}\sum_k C^w_{kk}(t)\right]C^w_{ij}(t) - \frac{1}{N}\sum_k C^w_{ik}(t)C^w_{kj}(t)[2+\epsilon^2 C_{ik}(t) + \epsilon^2 C_{kj}(t)]$$
$$+ 2\epsilon A_i(t)A_j(t) \tag{15}$$

We see that this dynamical system has fixed points for A_i at $\boldsymbol{A} = 0$. This is the cause for potential slow timescales for alignment from generic initial conditions in high dimensions even in online learning since the initial value for A_i is generally $\mathcal{O}(D^{-1/2})$ [1, 2]. We note however, that this is not a statistical *sample complexity* issue as we are still in the $P = \gamma D$ scaling but with $\gamma \gg 1$.

A.2. Leading Expression for Test/Train Loss Gap

We can use the following expressions for the train and test losses to derive a formula for their gap valid at large but finite γ

$$\mathcal{L}_{\text{test}} = \left\langle \left(\frac{\alpha}{N}\sum_{i=1}^N \phi(u^h_i(t)) - \sigma(h_\star) \right)^2 \right\rangle \ , \ \mathcal{L}_{\text{train}} = \left\langle \left(\frac{\alpha}{N}\sum_{i=1}^N \phi(h_i(t)) - \sigma(h_\star) \right)^2 \right\rangle \tag{16}$$

Using the fact that $u^h_i(t) = h_i(t) - \frac{1}{\gamma}\int ds \sum_j R^w_{ij}(t,s)g_j(s)$ we can expand the test loss around the training loss at large γ this gives

$$\mathcal{L}_{\text{test}}(t) \sim \left\langle \left(\frac{\alpha}{N}\sum_{i=1}^N \phi(h_i(t)) - \frac{\alpha}{\gamma N}\sum_{i=1}^N \dot\phi(h_i(t)) \int ds \sum_j R^w_{ij}(t,s)g_j(s) - \sigma(h_\star) \right)^2 \right\rangle$$
$$\sim \mathcal{L}_{\text{train}} + \frac{2\alpha}{\gamma N}\sum_{i=1}^N \int ds \sum_j R^w_{ij}(t,s)\left\langle \dot\phi(h_i(t))\Delta(t)g_j(s) \right\rangle + \mathcal{O}(\gamma^{-2}) \tag{17}$$

Recognizing that $g_i(t) = \dot\phi(h_i(t))\Delta(t)$, we can express the leading order correction as

$$\mathcal{L}_{\text{test}} - \mathcal{L}_{\text{train}} \sim \frac{2\alpha}{\gamma N}\int_0^t ds \sum_{ij} R^w_{ij}(t,s)C^g_{ij}(t,s) + \mathcal{O}(\gamma^{-2}) \tag{18}$$

We could now evaluate this expression at large γ so that R^w, C^g can be replaced with its value at $\gamma \to \infty$ (the solution in the Saad-Solla limit).

A.3. Early dynamics at Small α

We can also make some progress analytically in the case where $\alpha \ll 1$ and $\gamma \gg 1$. The early dynamics are

$$\frac{d}{dt}w_i(t) \sim u^w_i(t) + \frac{\epsilon}{2}w_i(t) + \epsilon A_i(t)\beta$$
$$\implies \frac{d}{dt}A_i(t) = \frac{3\epsilon}{2}A_i(t) \implies A_i(t) = \exp\left(\frac{3\epsilon}{2}t\right)A_i(0) \tag{19}$$

In order to reduce the loss appreciably, we need $A(t)$ to reach a scale of α^{-1} since $f = \alpha\phi(h)$. To reach an alignment of scale $\frac{1}{\alpha}$, we need to train for time

$$t \sim \frac{2}{3\epsilon} \ln\left(\frac{1}{\alpha\, A(0)}\right) \tag{20}$$

This is why the small α experiments train more slowly in Figure 2. We note that in high dimension under random initial conditions $A(0) \sim \frac{1}{\sqrt{D}}$ so the time to escape the saddle point at initialization is $\mathcal{O}(\ln D)$, consistent with prior works on one-pass SGD [1, 2].

A.4. General Response Function Recursions

We need a way to compute $R^h_{ij}(t,s)$. This requires knowledge of the linear responses in the non-Markovian $h_i(t)$ system. To obtain these we note the following implicit relations.

$$\begin{aligned}
\frac{\delta h_i(t)}{\delta u_j(s)} &= \delta(t-s)\delta_{ij} + \frac{1}{\gamma}\int_0^t dt' \sum_k R^w_{ik}(t,t')\frac{\delta g_k(t')}{\delta u_j(s)} \\
\frac{\delta g_i(t)}{\delta u_j(s)} &= \ddot{\phi}(h_i(t))\frac{\delta h_i(t)}{\delta u_j(s)}\Delta(t) \\
&\quad - \dot{\phi}(h_i(t))\left[\frac{1}{N}\sum_k \dot{\phi}(h_k(t))\frac{\delta h_k(t)}{\delta u_j(s)}\right] \\
\frac{\delta h_i(t)}{\delta u^{h_\star}} &= \frac{1}{\gamma}\int_0^t dt' \sum_j R^w_{ij}(t,t')\frac{\delta g_j(t)}{\delta u^{h_\star}} \\
\frac{\delta g_i(t)}{\delta u^{h_\star}} &= \dot{\phi}(h_i(t))\left[\dot{\sigma}(h_\star) - \frac{1}{N}\sum_k \frac{\delta h_k(t)}{\delta u^{h_\star}}\dot{\phi}(h_k(t))\right] + \ddot{\phi}(h_i(t))\frac{\delta h_i(t)}{\delta u^{h_\star}}\Delta
\end{aligned} \tag{21}$$

where $\Delta(t) = \sigma(h_\star) - \frac{1}{N}\sum_k \phi(h_k(t))$ are the instantaneous errors.

A.5. Cavity Derivation of DMFT Equations

In this section we quickly derive the limiting DMFT equations for the infinite data and dimension limit with $P = \gamma D$ with $\gamma = \mathcal{O}(1)$. This computation consists of two parts. First, we consider the effect of adding a single data point to the training set and computing the marginal statistics on this new training point. Then we proceed to add a single dimension to the data and then calculate the dynamics for the weight associated with this dimension. These lead to scalar stochastic processes which can be sampled and solved for in terms of correlation and response functions. To make the calculation simpler, we consider adding fictitious source terms $\boldsymbol{v}^w(t)$ and $v^h_\mu(t)$ which perturb the dynamics

$$\frac{d}{dt}\boldsymbol{w}_i(t) = \frac{1}{\gamma}\sum_{\mu=1}^P g^\mu_i(t)\boldsymbol{x}^\mu + \boldsymbol{v}^w_i(t) \tag{22}$$

$$h^\mu_i(t) = \boldsymbol{w}_i(t)\cdot\boldsymbol{x}^\mu + v^h_{\mu,i}(t) \tag{23}$$

We will define the following two response functions

$$\boldsymbol{R}^w_{ij}(t,s) = \lim_{\boldsymbol{v}^w \to 0} \frac{\delta \boldsymbol{w}_i(t)}{\delta \boldsymbol{v}^w_j(s)} \ , \ R^h_{ij,\mu\nu}(t,s) = \lim_{\{v^h_{\mu i}\}\to 0} \frac{\delta g^\mu_i(t)}{\delta v^h_{\nu j}(s)} \ , \ R^{h_\star}_{i,\mu\nu}(t) = \frac{\delta g^\mu_i(t)}{\delta h^\nu_\star} \tag{24}$$

Adding One Data Point We consider the effect of adding a data point $\boldsymbol{x}^0 \sim \mathcal{N}(0, \frac{1}{D}\boldsymbol{I})$ to the dataset. This addition leads to a small perturbation in the dynamics of the weights. Let $\tilde{\boldsymbol{w}}(t)$ represent the weights in the $P+1$ data point system (including $\boldsymbol{x}^0$) while $\boldsymbol{w}(t)$ is the weights in the P data system. By linear response theory, we have the following perturbed dynamics for $\tilde{\boldsymbol{w}}$

$$\tilde{\boldsymbol{w}}_i(t) \sim \boldsymbol{w}_i(t) + \frac{1}{\gamma}\int_0^t ds \sum_{j=1}^N g^0_j(s) \boldsymbol{R}^w_{ij}(t,s)\boldsymbol{x}^0. \tag{25}$$

Now, we can compute the perturbed preactivation statistics $h^0_i(t)$ on the new data point

$$h^0_i(t) \sim \boldsymbol{w}_i(t)\cdot \boldsymbol{x}^0 + \frac{1}{\gamma}\int_0^t ds \sum_{j=1}^N g^0_j(s) \boldsymbol{x}^{0\top}\boldsymbol{R}^w_{ij}(t,s)\boldsymbol{x}^0. \tag{26}$$

In the first, term, the random variables $\boldsymbol{w}_i(t)$ are statistically independent from $\boldsymbol{x}^0$ so this first term is just a Gaussian with mean zero and variance

$$\boldsymbol{w}_i(t)\left\langle \boldsymbol{x}^0\boldsymbol{x}^{0\top}\right\rangle \boldsymbol{w}_j(s) = \frac{1}{D}\boldsymbol{w}_i(t)\cdot\boldsymbol{w}_j(s) \equiv C^w_{ij}(t,s) \tag{27}$$

This covariance matrix will concentrate as $D\to\infty$. Next, we consider the second term, which also concentrates around its average

$$\boldsymbol{x}^{0\top}\boldsymbol{R}^w_{ij}(t,s)\boldsymbol{x}^0 \sim \frac{1}{D}\text{Tr }\boldsymbol{R}^w_{ij}(t,s) \equiv R^w_{ij}(t,s) \tag{28}$$

Thus we have the following marginal stochastic process for $h^0_i(t)$

$$h^0_i(t) = u^h_i(t) + \frac{1}{\gamma}\int ds \sum_j R^w_{ij}(t,s) g^0_j(s) \tag{29}$$

where $u^h \sim \mathcal{GP}(0, C^w)$.

Adding a feature dimension We next consider adding a dimension so that we have $D+1$ instead of a D dimensional weight vector and data vectors. Upon the addition of a single dimension, the gradient signals are perturbed as

$$\tilde{g}^\mu_i(t) = g^\mu_i(t) + \int_0^t ds \sum_{\nu=1}^P\sum_{j=1}^N R^h_{ij,\mu\nu}(t,s)x^\nu_0 w_{j0}(s) + \sum_{\nu=1}^P R^{h_\star}_{i,\mu\nu}(t)x^\nu_0\beta_0. \tag{30}$$

As a consequence the dynamics of the new trainable weight $w_{i0}(t)$ have the form

$$\begin{aligned}
\frac{d}{dt}w_{i0}(t) &= \frac{1}{\gamma}\sum_{\mu=1}^P g^\mu_i(t)x^0_\mu + \frac{1}{\gamma}\int_0^t ds \sum_{j\mu\nu} R^h_{ij,\mu\nu}(t,s)x^\mu_0 x^\nu_0 w_{j0}(s) + \frac{1}{\gamma}\sum_{\mu\nu} R^{h_\star}_{i,\mu\nu}(t)x^\mu_0 x^\nu_0\beta_0 \\
&\sim u^w_i(t) + \int_0^t ds \sum_j R^h_{ij}(t,s)w_{j0}(s) + R^{h\star}_i(t)\beta_0 \\
u^w &\sim \mathcal{GP}\left(0, \gamma^{-1}C^g\right) \ , \ C^g_{ij}(t,s) \equiv \langle g_i(t)g_j(s)\rangle
\end{aligned} \tag{31}$$

where we invoked the same central limit theorem and concentration ideas as the previous section. These equations and corresponding averages give us the DMFT presented in the previous sections.

A.6. How to Solve DMFT Equations

We are currently using a fixed point iteration scheme to solve the DMFT equations numerically.

1. Solve the online limit $\gamma \to \infty$ of the DMFT equations for $\{C^w, A, C^g, R^w, R^h, R^{h_\star}\}$ directly, leveraging the Gaussianity of $w, h, h_\star$.

2. Using these correlations and response functions, draw many Monte Carlo samples for u^h. Integrate the equations for the preactivations, resulting in many non-Gaussian samples for h. Use these samples to construct a Monte Carlo estimate of $C^g, R^h, R^{h_\star}$.

3. Solve for the weight covariance C^w, alignment A and response R^w directly utilizing the Gaussianity of w.

4. Update all of the order parameters $\{C^w, A, C^g, R^w, R^h, R^{h_\star}\}$ in the direction of the new estimate.

When this iteration procedure converges, it usually takes around $20 - 50$ iterations, though it takes longer for smaller γ (since the initial guess for order parameters is farther away from the fixed point). In the provided figures, we used 8000 Monte Carlo samples at each step and updated each order parameter C_n at step n of the above procedure as

$$C_{n+1} = \frac{1}{2} C_n + \frac{1}{2} C_n^{MC}. \tag{32}$$

where C^{MC} is the Monte-Carlo estimate of C at step n.