Structured Identity Mapping: A Model for Out-of-Distribution Generalization Dynamics

Anonymous Author(s)
Affiliation
Address
email

Abstract

Modern multi-modal generative models exhibit remarkable out-of-distribution generalization capabilities, combining concepts in ways not observed in the training data. While a rich body of literature theoretically studies learning dynamics in in-distribution generalization settings, the dynamics of out-of-distribution generation remain underexplored. In this work, we introduce and analyze the **Structured Identity Mapping** task, demonstrating how this simple model yields rich learning dynamics. Specifically, we analyze a one-hidden-layer network learning the identity map, using a training set composed of Gaussian point clouds structurally positioned at nodes of concept graphs. Our analysis of this model yields solutions that explain various empirical observations previously reported in text-conditioned diffusion models, including: (i) wave-like progression of compositional generalization dynamics, respecting hierarchical compositional structures; (ii) the impact of concept-centric data structures on concept learning speed; and (iii) non-monotonic progress of out-of-distribution generalization. In conclusion, our analytical model of concept learning establishes a theoretical foundation for investigating the dynamics of concept acquisition and combination in generative models.

1 Introduction

Concept learning and compositional generalization are essential features of modern generative models [1, 2, 3, 4, 5, 6]. These models can learn abstract concepts like shape, size and color from a limited set of training data and use them to generate images with novel combinations of concepts. In this paper, we focus on prompt conditioned image generative models, where the model is trained to generate images given a conditioning, e.g. a text prompt. Many papers have shown that this kind of model, especially ones based on diffusion models, often show some extent of concept learning [7, 8, 9].

One natural question which arises is: How do these capabilities emerges from training? It is generally hard to track the model behavior in terms of concept learning throughout training. Park et al. [6] proposed a principled way of studying it: instead of considering learning dynamics in parameter space as most previous work do, the authors study the learning dynamics in "concept space". Briefly speaking, concept space is vector space that serves as an abstraction of real concepts. For each concept (e.g. color), a binary number can be used to represent its value (e.g. 0 for red and 1 for blue). In this way, a binary string can be mapped to a text condition (e.g. $(1, 0, 1)$ might represent "big blue triangle") and then be fed into the generative model as a conditioning vector. After that, a pre-trained classifier is used to check whether the model has indeed generated the specified combination of concepts. The idea of concept space is illustrated in Figure 1 as well as the middle figure in Figure 2.

Park et al. [6] found some interesting phenomena from the training dynamics in the concept space, such as control variables for concept learning and non monotonic trajectories. However, they only gave a description of these phenomena and thus the underlying causal mechanisms are still unclear. In this paper, we make a first step of establishing a theoretical framework that explains those phenomena.

Submitted to Workshop on Scientific Methods for Understanding Deep Learning, NeurIPS 2024.

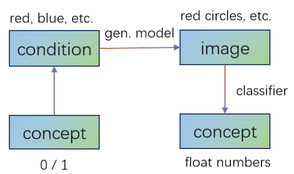


Figure 1: An schematic description of Concept Space.

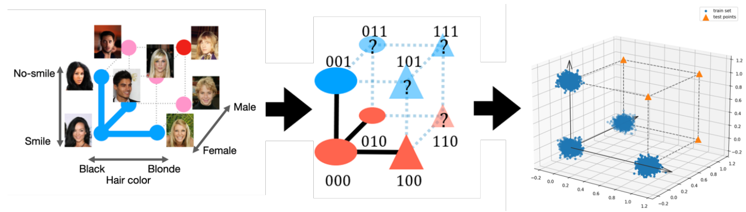


Figure 2: The abstraction process. Left: Actual Tasks (images from [10]) Middle: Concept Space considered in [6]; Right: SIM task considered in this work.

We would like to argue that the concept space framework established a very important idea: **In concept space, generation is essentially identity mapping**, as the final classifier output should match the input concept space specification. To this end, we make a further abstraction of the concept learning process: learning identity mapping in Euclidean space. Each concept is represented as a Gaussian cluster centered on an axes and the test point is a combination of several clusters means (see the right figure in Figure 2). We call this task the Structured Identity Mapping learning (SIM) task. As an analogue of concept learning, we claim that the seemingly simple SIM task has those key features:

1. It is an out-of-distribution task, and theoretically non-generalizing solutions exists, so it is non-trivial that a model solves this task;
2. As we will show later, this task captures many phenomena observed on real datasets;
3. This task masks out distracting terms and is simple enough that we can elaborate the behaviour of the model on it.

In this work, we give a detailed description of how the model behaves throughout training on the SIM task and how those behaviors corresponds to the behavior of actual diffusion models. After that, we study the learning dynamics of a specific model on this task, and give a rigorous proof of the existence of the phenomena. More specifically, we make two major contributions:

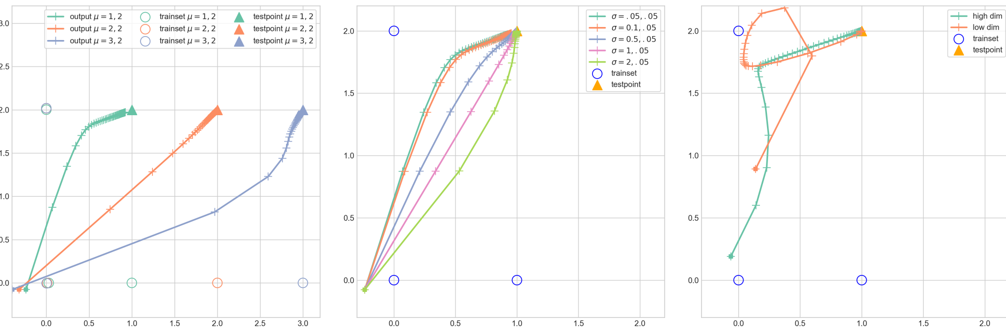
1. In Section 3, we train MLP regression models on the SIM task and empirically reveal some key features of the training dynamics, including: 1) how the order of concepts learning is controlled by the signal strength and diversity of the dataset; 2) the deceleration of concept learning with training progress and 3) a "Transient Memorization" phenomena where the model shows a trend of generalization at the beginning of the training soon followed by memorization before heading back to generalization again. The last phenomena leads to a double descent-like loss curve with respect to optimization steps even in an under-parameterized setting.
2. In Appendix A, we explain the empirical phenomena by a detailed analysis of two simplified models. Interestingly, we show that *Transient Memorization* is a phenomenon only observed in multi-layer models, which reveals a key difference between one-layer model and deep models. Our novel analysis of dynamics reveals a multi-stage evolution process of the Jacobian of a two layer symmetric linear model ($f(\boldsymbol{x};\boldsymbol{U}) = \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{x}$), and we show that each stage of the Jacobian evolution precisely corresponds to the stages of the *Transient Memorization*.

2 Preliminaries and Problem Setting

Throughout the paper, we use bold lowercase letters (e.g. $\boldsymbol{x}$) to represent vectors, and bold uppercase letters (e.g. $\boldsymbol{A}$) for matrices. Non-bold versions with subscripts represent corresponding entries of the vectors or matrices, e.g. x_i represent the i-th entry of $\boldsymbol{x}$ and $a_{i,j}$ represent the (i,j)-th entry of $\boldsymbol{A}$.

$[a]$ represents the set of all natural numbers that is smaller or equal to a, i.e. $[a] = \{1, 2, \cdots, a\}$.

$\mathbf{1}_k$ represents a one-hot vector which is 0 at every entry except the k-th entry being 1. The dimensionality of the vector is determined by the context if not specified. $\boldsymbol{I}_{:k}$ represents a diagonal matrix whose first k diagonal entries are 1 and others are 0.



(a) (b) (c)

Figure 3: The two dimensional ($s = 2$) output dynamics under different settings, evaluated for the test point, $(1, 1)$. We only show the center of the training classes as a circle, but the actual training set can have varied shapes based on the configuration of $\boldsymbol{\sigma}$. (a) one layer linear model with $\boldsymbol{\sigma}_{:2} = (.05, .05)$ and varying $\boldsymbol{\mu}$; (b) one layer linear model with $\boldsymbol{\mu}_{:2} = (1, 2)$ and varying $\boldsymbol{\sigma}$; (c) 4 layer linear models for different model dimensions. high dim: $d = 64$, low dim: $d = 2$. Note that (a) and (b) are both in high dimensional model setting.

2.1 Problem Setting

Data. The SIM dataset is composed of several Gaussian clusters, each occupying a coordinate direction. Figure 2 (right) illustrates the SIM dataset.

Let $d \in \mathbb{N}$ be the dimensionality of the input space, $s \in [d]$ be the number of clusters, and $n \in \mathbb{N}$ be the number of samples from each cluster. The training set $\mathcal{D} = \bigcup_{p \in [s]} \left\{ \boldsymbol{x}_k^{(p)} \right\}_{k=1}^{n}$ is generated by the following process: for each $p \in [s]$, the data sample is sampled i.i.d. from a Gaussian distribution $x_k^{(p)} \sim \mathcal{N}\left[\mu_p \mathbf{1}_p, \operatorname{diag}(\boldsymbol{\sigma})^2\right]$, where $\mu_p \geq 0$ is the distance of the p-th cluster center from the origin, and $\boldsymbol{\sigma}$ is a vector with only the first s entries being non-zero, where σ_i represents the variance on the i-th direction. Notice that we allow $\mu_p = 0$ for a specific p to create a cluster that centers at $\mathbf{0}$ while keeping the terminology as simple as possible.

Loss function. The training problem is to learn identity mapping on $\mathbb{R}^d$. For a model $f : \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}^d$ and a parameter vector $\boldsymbol{\theta} \in \mathbb{R}^m$, we train the model parameter $\boldsymbol{\theta}$ with the mean square error loss.

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2sn} \sum_{p=1}^{s} \sum_{k=1}^{n} \left\| f\left(\boldsymbol{\theta}; \boldsymbol{x}_k^{(s)}\right) - \boldsymbol{x}_k^{(s)} \right\|^2 . \tag{2.1}$$

Evaluation. In the main paper, we focus on a single test point $\widehat{\boldsymbol{x}} = \sum_{p=1}^{s} \mu_p \mathbf{1}_p$. Notice that this point is outside of the training distribution. In Appendix B, we report additional results for the case of multiple test points.

3 Observations on the SIM Task

In this section, we summarize some of the interesting empirical findings of the model behavior when we conduct experiments on the SIM. We note that these findings reported in this section are in one-to-one correspondence with results with diffusion models.

In all experiments, we use MLP models. We perform experiments with both linear activations and non-linear ReLU activations, as well as model with different number of layers. Due to the limited space, we only show the results of some settings here but the phenomena reported are consistent with different hyperparameters.

3.1 Generalization Order is Controlled by Signal Strength and Diversity

One interesting findings from previous work is that if we alter the strength of one signal from small to large, the concur of the learning dynamics would dramatically change [6]. Moreover, it is also commonly hypothesised that with more diverse data, the model should also generalize better [11, 12].

In the SIM task the distance μ_k of each cluster can be viewed as the corresponding signal strength, and the covariance σ_k can be viewed as the data diversity. In Figure 3, we train models with $s = 2$ allowing us to directly plot the trajectory of the model output at each timestep. In this case, there are two components, x and y, to be learned and the order of learning can be seen from the trajectory.

In Figure 3 (a)[1], $\boldsymbol{\sigma}$ is fixed, and we can see when $\mu_1 < \mu_2$, the dynamics shows an upward bulging, showing a preference for the direction of stronger signal. As we gradually increase μ_1, this trajectory gradually transitions from an upward bulge to downward one, consistent with the stronger signal strength.

In Figure 3 (b), the $\boldsymbol{\mu}$ is fixed to have one signal stronger than the other, and the model, as expected, prefers the direction with stronger signal when the data diversity is balanced. However, if we tune the data diversity of one side from weak to strong while keeping the other side unchanged, we can override the preference coming from the mean signal.

The results in Figure 3 (a) and (b) gives us a very concrete conclusion: The learning direction is a competition driven by signal strength and diversity, and the model prefers direction that has stronger signal and more diversity.

3.2 Transient Memorization

The results in Figure 3 (a) and (b) are both performed with one layer models and under a high dimensional setting ($d = 64$). Despite the overall trend is similar in other settings, it is worth exploring the change of trajectory as we increase the number of layers, and / or reduce the dimension.

In Figure 3 (c), we perform experiments with deeper models, and optionally with a lower dimension. Under these changes, we find that the model shows an interesting irregular behavior, where it initially heads towards the right direction, but soon turns toward the training set cluster with the strongest signal, exhibiting distributional memorization the training set. However, with enough training, the model correctly moves towards the intended target and thus generalizes. We call this behavior **Transient Memorization**.

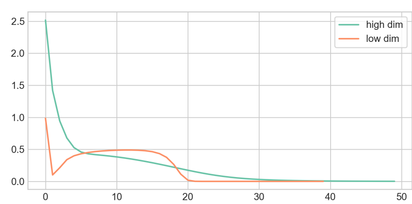


Figure 4: The loss function of multi layer models.

This trajectory could be suggestive of a non-monotonic generalization. We track the value of the loss function during training in Figure 4, demonstrating a double descent-like curve. We note that the *Transient Memorization* phenomena seems to be strongest when the dimensionality is low, and is rather modest with high dimensional settings. In the high dimensional setting the loss descent slows down at some point but doesn't actually exhibit non-monotonic behavior.. This low dimensional preference can also be explained perfectly by our theory, further described in Appendix A.

3.3 Convergence Rate Slow Down In Terminal Phase

In Figure 3, the markers on the curve corresponds to equal time intervals. One can observe that the model's generation's evolution in concept space slows down as training progresses. This phenomena is observed in diffusion models as well [6].

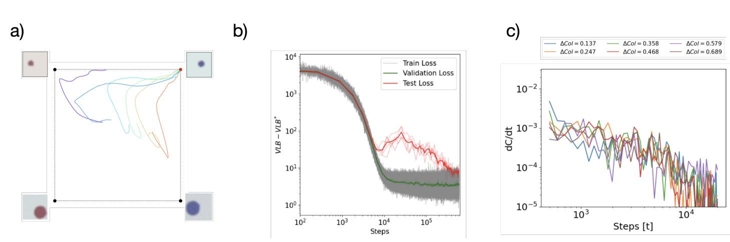


Figure 5: **Diffusion Model Results** a) Concept signal controls learning order and speed b) Transient Memorization observed in vector space diffusion models c) Deceleration of concept learning with time

[1] In Figure 3 (a) some training set centroids are overlapping and we perturb them a little for visibility.

References

[1] Maya Okawa, Ekdeep S Lubana, Robert Dick, and Hidenori Tanaka. Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task. *Advances in Neural Information Processing Systems*, 36, 2024.

[2] Parikshit Ram, Tim Klinger, and Alexander G Gray. What makes models compositional? a theoretical view: With supplement. *arXiv preprint arXiv:2405.02350*, 2024.

[3] Thaddäus Wiedemer, Prasanna Mayilvahanan, Matthias Bethge, and Wieland Brendel. Compositional generalization from first principles. *Advances in Neural Information Processing Systems*, 36, 2024.

[4] Xindi Wu, Dingli Yu, Yangsibo Huang, Olga Russakovsky, and Sanjeev Arora. Conceptmix: A compositional image generation benchmark with controllable difficulty, 2024. URL `https://arxiv.org/abs/2408.14339`.

[5] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models, 2023. URL `https://arxiv.org/abs/2206.01714`.

[6] Core Francisco Park, Maya Okawa, Andrew Lee, Ekdeep Singh Lubana, and Hidenori Tanaka. Emergence of hidden capabilities: Exploring learning dynamics in concept space. *arXiv preprint arXiv:2406.19370*, 2024.

[7] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021.

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[9] Core Francisco Park, Maya Okawa, Andrew Lee, Ekdeep Singh Lubana, and Hidenori Tanaka. Emergence of hidden capabilities: Exploring learning dynamics in concept space, 2024. URL `https://arxiv.org/abs/2406.19370`.

[10] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018.

[11] Zhiqiang Gong, Ping Zhong, and Weidong Hu. Diversity in machine learning. *Ieee Access*, 7: 64323–64350, 2019.

[12] José F Díez-Pastor, Juan J Rodríguez, César I García-Osorio, and Ludmila I Kuncheva. Diversity techniques improve the performance of the best imbalance learning ensembles. *Information Sciences*, 325:98–117, 2015.

[13] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

[14] Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. *arXiv preprint arXiv:1810.02281*, 2018.

[15] Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032*, 2018.

[16] Dominik Stöger and Mahdi Soltanolkotabi. Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. *Advances in Neural Information Processing Systems*, 34:23831–23843, 2021.

[17] Jikai Jin, Zhiyuan Li, Kaifeng Lyu, Simon Shaolei Du, and Jason D Lee. Understanding incremental learning of gradient descent: A fine-grained analysis of matrix sensing. In *International Conference on Machine Learning*, pages 15200–15238. PMLR, 2023.

A Theoretical Explanation

In this section, we study the training dynamics of a specific type of linear models which is tractable on the SIM task, and we provide explanations of the behaviors on the SIM task. We first consider a linear model and show that despite it can explain some of the phenomena, the linear model can not actually reproduce every phenomenon, which suggests that some phenomena are intrinsic for deep models, which highlights the difference of shallow and deep models.

Throughout this section, we assume $\boldsymbol{f}(\boldsymbol{\theta};\boldsymbol{x})$ is a linear operator of $\boldsymbol{x}$. In this case the Jacobian of $\boldsymbol{f}$ w.r.t. $\boldsymbol{x}$ is a matrix that is completely determined by $\boldsymbol{\theta}$, which we denote by $\boldsymbol{W_\theta} = \frac{\partial f(\boldsymbol{\theta};\boldsymbol{x})}{\partial \boldsymbol{x}}$. It's easy to see that we have $\boldsymbol{f}(\boldsymbol{\theta};\boldsymbol{x}) = \boldsymbol{W_\theta x}$. Through the trace trick, it's easy to show that the overall loss function is equal to

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\left\|\left(\boldsymbol{W_\theta} - \boldsymbol{I}\right)\boldsymbol{A}^{1/2}\right\|_{\mathcal{F}}^2, \tag{A.1}$$

where $\boldsymbol{A} = \frac{1}{sn}\sum_{p=1}^{s}\sum_{k=1}^{n}\boldsymbol{x}_k^{(p)}\boldsymbol{x}_k^{(p)\top}$ is the empirical covariance. When n is large, $\boldsymbol{A}$ converges to the true covariance of the dataset $\boldsymbol{A} \to \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\boldsymbol{x}\boldsymbol{x}^\top$. To avoid extra distractions, through out this section we assume $\boldsymbol{A}$ equals the the true covariance, which is a diagonal matrix $\boldsymbol{A} = \operatorname{diag}(\boldsymbol{a})$ defined by $a_p = \begin{cases}\sigma_p^2 + \frac{\mu_p^2}{s} & p \le s \\ 0 & p > s\end{cases}$, for any $p \in [s]$. For completeness, we write the full derivation of eq. (A.1) and $\boldsymbol{A}$ in Appendix C.1.

Remark. Notice that in the linear setting we might not directly train $\boldsymbol{W_\theta}$, instead we train its components. For example we might have $\boldsymbol{\theta} = (\boldsymbol{W}_1, \boldsymbol{W}_2)$ and have $\boldsymbol{W_\theta} = \boldsymbol{W}_1\boldsymbol{W}_2$. What we actually train is $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ instead of $\boldsymbol{W_\theta}$. As many previous work have emphasized [13, 14, 15], although the deep linear model has the same capacity as a one-layer linear model, their dynamics can be vastly different and the loss landscape of deep linear models can be non-convex.

A.1 The Failure of One Layer Model Theory

As a warm-up, we first study the dynamics of one layer linear models, i.e. $f(\boldsymbol{W};\boldsymbol{x}) = \boldsymbol{W}\boldsymbol{x}$, in which case the Jacobian $\boldsymbol{W_\theta}$ is simply $\boldsymbol{W}$. We will show that it can explain some of the phenomenons but fail to capture other interesting behaviours.

Theorem A.1. *Let $\boldsymbol{W}(t) \in \mathbb{R}^{d\times d}$ be initialized as $\boldsymbol{W}(0) = \boldsymbol{W}^{(0)}$, and updated by*

$$\frac{\mathrm{d}\boldsymbol{W}(t)}{\mathrm{d}t} = -\nabla\mathcal{L}(W), \tag{A.2}$$

with $\mathcal{L}$ be defined by eq. (A.1) *with $f(\boldsymbol{W},\boldsymbol{z}) = \boldsymbol{W}\boldsymbol{z}$, then we have for any $\boldsymbol{z} \in \mathbb{R}^d$,*

$$f(\boldsymbol{W}(t),\boldsymbol{z})_k = \underbrace{\mathbb{1}_{\{k\le s\}}\left[1-\exp\left(-a_k t\right)\right] z_k}_{\widetilde{G}_k(t)} + \underbrace{\sum_{i=1}^{s}\exp\left(-a_i t\right) w_{k,i}(0) z_i}_{\widetilde{N}_k(t)}. \tag{A.3}$$

See Appendix C.2 for a proof of Theorem A.1.

The output of the one-layer model $f(\boldsymbol{\theta}(t);\boldsymbol{z})$ can be decomposed into two terms: the *growth term* $\widetilde{G}_k(t;\boldsymbol{z}) = \mathbb{1}_{\{k\le s\}}\left[1-\exp\left(-a_k t\right)\right] z_k$ and the *noise term* $\widetilde{N}_k(t;\boldsymbol{z}) = \sum_{i=1}^{s}\exp\left(-a_i t\right) w_{k,i}(0) z_i$. By observing these two terms we can find the following properties: 1) the growth term converges to x_k when $k \le s$ and 0 when $k > s$, and the noise term always converges to 0; 2) both terms converges in an exponential rate; 3) the noise term is upper bounded by $\sum_{i=1}^{s} w_{k,i}(0) x_i$.

If the model is initialized small, specifically $w_{k,i}(0) \ll \frac{1}{s\max_{i\in[s]}\{x_i\}}$, then the $\widetilde{N}_k(t)$ will always be small, and thus can be omitted. With this assumption in effect, the model output is dominated by the growth term. A closer look at the growth term reveals the origin of some of the phenomena observed before.

Generalization Order. In eq. (A.3) we can see that the growth term $\widetilde{G}_k(t;\boldsymbol{z})$ converges in an exponential speed, with the exponential term controlled by a_k, which is $a_k = \frac{s\sigma_k^2+\mu_k^2}{s}$. Therefore, the direction with larger a_k, i.e. larger μ_k and / or σ_k, converges faster. The equation also reveals the proportional relationship of μ_k and σ_k.

Terminal Phase Slowing Down. By taking derivative of the growth term $\widetilde{G}_k(t;\boldsymbol{z})$ with $k \le s$, we have

$$\dot{\widetilde{G}}_k(t;\boldsymbol{z}) = a_k z_k \exp\left(-a_k t\right), \tag{A.4}$$

which monotonically decays w.r.t. t, and reveals an (exponentially) slowing down of the convergence rate when t becomes large.

The Failure of One Layer Model Theory. We have shown that Theorem A.1 can explain many observations. However, in eq. (A.3) the growth term $\widetilde{G}_k(t;\boldsymbol{z})$ is independent and monotonic for each layer, which only produces monotonic and rather regular traces (this is verified by the experiments in Section 3.1. However, as the experiments in Section 3.2 show, when the number of layer becomes larger, the model actually shows a non-monotonic trace that can have detours. The theory based on one layer model fails in capturing that phenomenon. In the subsequent section, we introduce a more complex theory by considering a deeper model, and we will show that this theory explains all the phenomena observed in Section 3.

A.2 A Symmetric Two Layer Model Theory

In this subsection, we introduce a two layer model with symmetric weights. Despite its simplicity, we show that it perfectly captures every observations presented in Section 3. More importantly, the theory derived from this model draws a clear picture of the evolution of the evolution of the model Jacobian and provides us with a clear and understandable explanation of the origin of the seemingly irregular behaviours of the model.

Due to the space limitation, in this subsection we focus on providing an intuitive explanation of the model behaviour and delay the formal proof to later sections.

Formally, in this subsection, we consider a linear model that has two layers with shared weights, namely

$$f(\boldsymbol{U};\boldsymbol{x}) = \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{x}, \tag{A.5}$$

where $\boldsymbol{U} \in \mathbb{R}^{d\times d'}$ and $d' > d$. Notice that this is a model commonly studied in theory [16, 17], and our analysis goes beyond the existing ones by providing an analysis for the early stage phenomenon, Transient Memorization. For simplicity, in this subsection we denote the Jacobian of f at time point t by $\boldsymbol{W}(t) = \boldsymbol{W}_{\boldsymbol{U}(t)}$, then the update of the i,j-th entry of $\boldsymbol{W}$ is given by

$$\dot{w}_{i,j} = \underbrace{w_{i,j}(a_i+a_j)}_{G_{i,j}(t)} - \underbrace{\frac{1}{2}w_{i,j}\left[w_{i,i}(3a_i+a_j) + \mathbb{1}_{\{i\neq j\}}w_{j,j}(3a_j+a_i)\right]}_{S_{i,j}(t)} - \underbrace{\frac{1}{2}\sum_{\substack{k\neq i\\ k\neq j}} w_{k,i}w_{k,j}(a_i+a_j+2a_k)}_{N_{i,j}(t)}. \tag{A.6}$$

As noted in eq. (A.6), we decompose the update of $w_{i,j}$ into three terms. We call $G_{i,j}(t) = w_{i,j}(t)(a_i+a_j)$ the *growth term*, $S_{i,j}(t) = \frac{1}{2}w_{i,j}\left[w_{i,i}(3a_i+a_j) + \mathbb{1}_{\{i\neq j\}}w_{j,j}(3a_j+a_i)\right]$ the *suppression term*, and $N_{i,j}(t) = \frac{1}{2}\sum_{\substack{k\neq i\\ k\neq j}} w_{k,i}(t)w_{k,j}(t)(a_i+a_j+2a_k)$ the *noise term*. The name of the three terms suggests their role in the evolution of the Jacobian: the growth term $G_{i,j}$ always has the same sign as $w_{i,j}$, and has a positive contribution to the update, so it always leads to the direction that **enlarges the absolute value** of $w_{i,j}$; the suppression term $S_{i,j}$ also has the same sign[2] as $w_{i,j}$,

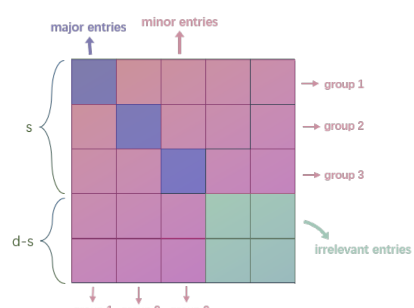


Figure 6: An illustration of the entries of the Jacobian.

[2]Notice that since $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{U}^\top$ is a PSD matrix, the diagonal entries are always non-negative.

but has a negative contribution in the update function of $w_{i,j}$, so it always leads to the direction that **shrinks the absolute value** of $w_{i,j}$; the effect of noise term is random since it depends on the sign of $w_{i,j}$ and other terms. It is proved in Lemma D.8 that under specific assumptions, the noise term will never be too large so for the sake of brevity, we ignore it in the following discussion and delay the treatment of it to the rigorous proofs in later sections.

A.2.1 The Evolution of Entries of Jacobian

In order to better present the evolution of the Jacobian, we divide the entries of the Jacobian into three types: the **major entries** are the first s diagonal entries, and the **minor entries** are the off-diagonal entries who are in the first s rows or first s columns, and other entries are **irrelevant entries**. Notice that the irrelevant entries doesn't contributes to the output of the test points so we ignore them. Moreover, we also divide minor entries to groups. The minor entry in the p-th row or column belongs to the p-th group (each entry can belong to at most two groups). See Figure 6 for an illustration of the division of the entries.

Initial Growth. In this section, we assume $w_{i,i}$-s are initialized around a very small value ω such that $\omega \ll \frac{1}{d \max_{i\in[s]} a_i}$ (See Appendix D.1 for specific assumptions). One can easily notice that when all $w_{i,j}$-s are around ω, the growth term is $o(\omega)$ and the suppression term and the noise term are both $o(\omega^2)$. This indicates when all entries are closed to initialization, the suppression term is negligible and the evolution of $w_{i,j}$ is dominated by the growth term. Therefore, in this stage, every value in the Jacobian grows towards the direction of enlarging its absolute value, with the speed determined by $a_i + a_j$. Since we assumed that $\boldsymbol{a}$ is ordered in a descending order, it's not hard to see that each entry grows faster than all entries below or on the right of it. The initial growth stage is characterized by Lemmas D.1 to D.3.

First Suppression. We say an entry that is close to its initialization is in the "initial phase". In the Initial Growth stage, the first major entry will be the one that grows exponentially faster than all other entries, thus it will be the first one that leaves the initial phase. Once the first major entry becomes significant and non-negligible, it will effect on the suppression term of all minor entries in the first group. When the difference between a_1 and a_2 is large enough, the first major entry is able to change the growth direction of the first group of minor entries and push their value to 0. The suppression stages are characterized by Lemma D.7

Second Growth and Cycle. Once the suppression of the first group of minor entries takes into effect, the second major entry becomes the one that grows fastest. Thus, the second major entry will be the second one that leaves the initial stage. When the second major entry becomes large enough, again it will suppress the second group of minor entries and push their value to 0. The process continues like this: A major entry growth is followed by the suppression of the corresponding group of minor entries, and the suppression leaves space for the growth of the next major entry. The general growth stages are characterized by Lemma D.4 and the fate of off-diagonal entries are characterized by Lemma D.8.

Growth Slow Down and Stop. Notice that the suppression term of a major entry is also determined by itself. Thus when a major entries becomes significantly large, it also suppresses itself, and causes the slow down of growth. Note that this won't reverse its growth direction since the suppression term is always smaller than the growth term, until $w_{i,i}$ becomes one where the growth and suppression equal and the evolution stops. The terminal stage of the growth of major entries are characterized by Lemma D.5.

A.2.2 Explaining Model Behavior

Recall that we have $f\left(\boldsymbol{U}(t); \hat{\boldsymbol{x}}\right)_k = \sum_{p=1}^{s} w_{k,p}(t) v_p \mu_p$. In this subsection we explain how the Jacobian evolution predicted in Appendix A.2.1 are reflected in the model output evolution.

Learning Order and Terminal Slowing Down. From the discussions in Appendix A.2.1, we see that at the end of the learning, all the major entries converges to 1 and all minor entries converges to

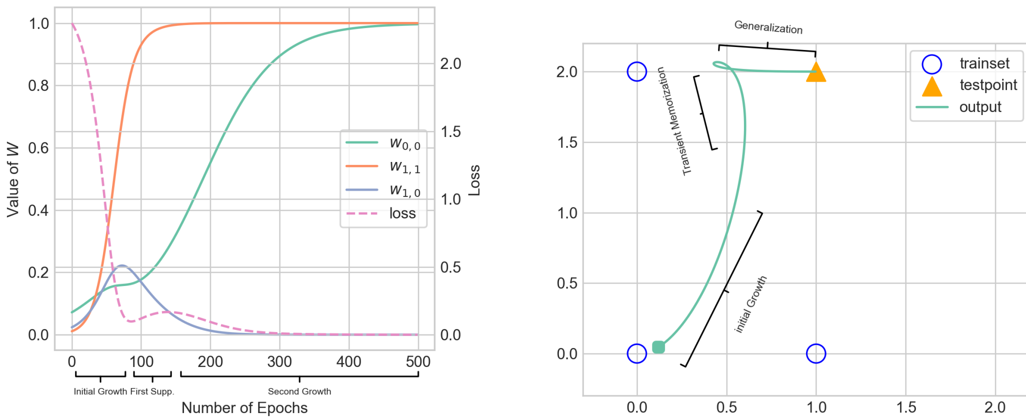


Figure 7: The learning dynamics of a two layer symmetric linear mode. Left: The change of the loss and the Jacobian entries with time predicted by the theory; Right: the corresponding model output curve. The figures are plotted under $s = 2$ and all entries of $\boldsymbol{W}$ are initialized positive.

0, and the major entries grows in the order of corresponding a_p, which depends on the μ_p and σ_p, and slows down when approaching the terminal. This explains our observation that directions with larger μ_p and / or σ_p is learned first, as well as the terminal phase slowing down of the learning.

Transient Memorization. We argue that, the Transient Memorization is caused by the initial growth. Notice that in the initial stage $f\left(\boldsymbol{U}(t); \hat{\boldsymbol{x}}\right)_k$ is dominated by $w_{k,1}(t)v_1\mu_1$, since $w_{k,1}$ grows fastest among all the entries. If $w_{k,1}$ happens to be initialized positive, it growth towards the positive direction. When s is small, this is actually easy to satisfy[3]. This causes an illusion that the model is going towards the right direction, because the target point has all positive coordinates.

Figure 7 shows the loss curve and Jacobian entry evolution predicted by the theory with all entries of $\boldsymbol{W}$ initialized positive. Notice that how the first and second descending of loss accurately corresponds to the initial and second growth of the major entries, and the ascending of the loss corresponds to the suppression of the minor entries.

Remark. We note that, for a major entry $w_{i,i}$ in the Jacobian $\boldsymbol{W}$, Lemma D.4 proves that when $w_{i,i} = \lambda$, the growth rate of $w_{i,i}$ is $\lambda a_i \exp(-2\lambda a_k t)$, which although not exactly the same, also suggests an exponential growth and slow down rate, and thus coincides with the theory prediction in the one-layer model discussed in Appendix A.1. Thus, we prefer to view the theory in this subsection as a refinement of the theory derived by one layer model, instead of a refutation. If we take into account more layer and non-linearities, we might be able to find more refinements but the predictions should follow the same trend as described in this subsection, since all the predicted behaviours of the theory presented in this subsection are experimentally verified with more complex models.

B Model Compositionally Generalize in Topologically Constrained Order

In this section, we introduce another phenomenon observed on SIM task learning that we don't put in the main paper: the order of compositional generalization happens in a topologically constrained order.

In this section, instead of the single test point $\widehat{\boldsymbol{x}}$, we introduce a hierarchy of test points. Specifically, let $\mathcal{I} = \{0,1\}^s$ be the index set of test points. For each $\boldsymbol{v} \in \mathcal{I}$, we define a test point

$$\widehat{\boldsymbol{x}}^{(\boldsymbol{v})} = \sum_{p=1}^{s} v_p \mu_p \mathbf{1}_p, \tag{B.1}$$

[3]And in opposite, when s is large it's unlikely that all entries are initialized positive, thus the Transient Memorization happens rarer.

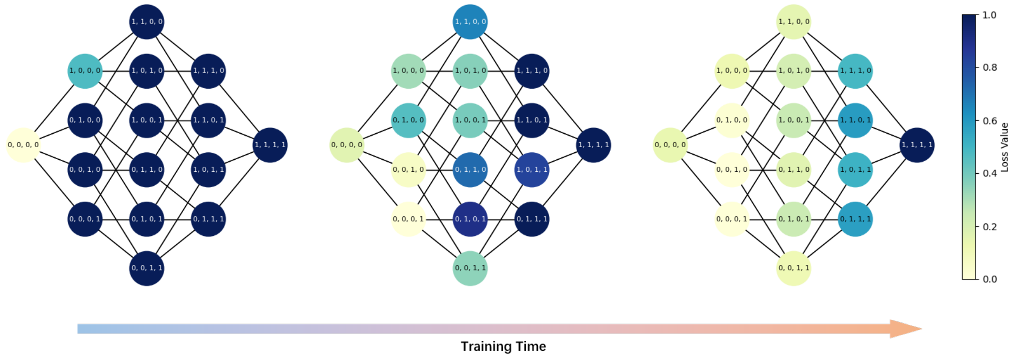


Figure 8: The loss at each test point in different timepoints during training for a 2-layer MLP with ReLU activation. Each graph represents a timepoint. Each node in the graph represents a test point, with index printed on it, and edges connecting nodes with Hamming distance 1. The color of the graph represents the loss of corresponding test point. Notice that we truncate the loss at 1 in order to unify the scale. From lest to right: epoch = $1, 3, 5$.

and call $\widehat{\boldsymbol{x}}^{(\boldsymbol{v})}$ the test point with the index $\boldsymbol{v}$. Intuitively, the index $\boldsymbol{v}$ describes which training sets are combined into the current test point. If $\|\boldsymbol{v}\| = 1$ then $\widehat{\boldsymbol{x}}^{(\boldsymbol{v})}$ is the center of one of the training clusters.

We assign the component-wise ordering $\preceq$ to the index set $\mathcal{I}$, i.e. for $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{I}$, we say $\boldsymbol{u} \preceq \boldsymbol{v}$ if and only if $\forall i \in [n], u_i \leq v_i$. It's easy to see that $\preceq$ is a partial-ordering.

Interestingly, in the SIM experiment, the order of the generalization in different test points strictly follow the component-wise order. This finding can be described formally in the following way: the loss function is an order homomorphism between $\preceq$ on the index set, and $\leq$ on the real number. Let $\ell(\boldsymbol{z})$ be the loss function of the test point $\boldsymbol{z}$, then we have the following empirical observation:

$$\forall \boldsymbol{u}, \boldsymbol{v} \in \mathcal{I}, \boldsymbol{u} \preceq \boldsymbol{v} \implies \ell\left(\tilde{\boldsymbol{x}}^{(u)}\right) \leq \ell\left(\tilde{\boldsymbol{x}}^{(v)}\right). \tag{B.2}$$

In Figure 8 we show the loss of each test point in several timepoints, with $\boldsymbol{\mu} = (1, 2, 3, 4)$, $\boldsymbol{\sigma} = \left\{\frac{1}{2}\right\}^4$. There is a clear trend that the test points that are on the right of the graph (larger in the component-wise order) will only be learned after all of its predecessors are all learned. We call this phenomenon the *topological constraint* since the constraint is based on the topology of the graph in Figure 8.

C Proofs and Calculations

In the main text we have omitted some critical proofs and calculations due to space limitation. In this section we provide the complete derivations. Notice that we delay the calculations of Appendix A.2 to Appendix D due to its length.

C.1 The Loss Function with Linear Model and Infinite Data Limit

In this subsection we derive the transformed loss function eq. (A.1), as well as the expression of the data matrix $\boldsymbol{A}$. For convenience we denote $\boldsymbol{W}_{\boldsymbol{\theta}}$ by $\boldsymbol{W}$. We have

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \frac{1}{2ns}\sum_{p=1}^{s}\sum_{k=1}^{n}\left\|(\boldsymbol{W}-\boldsymbol{I})\boldsymbol{x}_k^{(p)}\right\|^2 && \text{(C.1)}\\
&= \frac{1}{2ns}\operatorname{Tr}\left[\boldsymbol{x}_k^{(p)\top}(\boldsymbol{W}-\boldsymbol{I})^\top(\boldsymbol{W}-\boldsymbol{I})\boldsymbol{x}_k^{(p)}\right] && \text{(C.2)}\\
&= \frac{1}{2ns}\operatorname{Tr}\left[(\boldsymbol{W}-\boldsymbol{I})^\top(\boldsymbol{W}-\boldsymbol{I})\boldsymbol{x}_k^{(p)}\boldsymbol{x}_k^{(p)\top}\right] && \text{(C.3)}\\
&= \frac{1}{2}\operatorname{Tr}\left[(\boldsymbol{W}-\boldsymbol{I})^\top(\boldsymbol{W}-\boldsymbol{I})\frac{1}{ns}\boldsymbol{x}_k^{(p)}\boldsymbol{x}_k^{(p)\top}\right] && \text{(C.4)}\\
&= \frac{1}{2}\operatorname{Tr}\left[\boldsymbol{A}^{1/2}(\boldsymbol{W}-\boldsymbol{I})^\top(\boldsymbol{W}-\boldsymbol{I})\boldsymbol{A}^{1/2}\right] && \text{(C.5)}\\
&= \frac{1}{2}\left\|(\boldsymbol{W}-\boldsymbol{I})\,\boldsymbol{A}^{1/2}\right\|_{\mathcal{F}}^2. && \text{(C.6)}
\end{aligned}$$

Let $\mathcal{G}$ be the data generating process. It can be viewed as two components: first assign one of the s clusters, and then draw a Gaussian vector from a Gaussian distribution in that cluster. Specifically, let $\boldsymbol{x}$ be an arbitrary sample from the traning set, then the distribuition of $\boldsymbol{x}$ is equal to

$$\boldsymbol{x} \simeq \boldsymbol{\mu}^{(\eta)} + \operatorname{diag}(\boldsymbol{\sigma})\boldsymbol{\xi}, \tag{C.7}$$

where η is a uniform random variable taking values in $[s]$ and $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ is a random Gaussian vector that is independent from η. Here $\simeq$ represents having the same distribution.

When $n \to \infty$, the data matrix $\boldsymbol{A}$ converges to the true covariance, which is is

$$\begin{aligned}
\boldsymbol{A} &\to \mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}^\top\right) && \text{(C.8)}\\
&= \mathbb{E}\left[\left(\boldsymbol{\mu}^{(\eta)} + \operatorname{diag}(\boldsymbol{\sigma})\boldsymbol{\xi}\right)\left(\boldsymbol{\mu}^{(\eta)} + \operatorname{diag}(\boldsymbol{\sigma})\boldsymbol{\xi}\right)^\top\right] && \text{(C.9)}\\
&= \mathbb{E}\left(\boldsymbol{\mu}^{(\eta)}\boldsymbol{\mu}^{(\eta)\top}\right) + \mathbb{E}\operatorname{diag}(\boldsymbol{\sigma})\boldsymbol{\xi}\boldsymbol{\xi}^\top\operatorname{diag}(\boldsymbol{\sigma}) && \text{(C.10)}\\
&= \frac{1}{s}\sum_{p=1}^{s}\boldsymbol{\mu}^{(p)}\boldsymbol{\mu}^{(p)\top} + \operatorname{diag}(\boldsymbol{\sigma})^2 && \text{(C.11)}\\
&= \frac{1}{s}\sum_{p=1}^{s}\mu_p^2\boldsymbol{1}_p\boldsymbol{1}_p^\top + \operatorname{diag}(\boldsymbol{\sigma})^2 && \text{(C.12)}\\
&= \frac{1}{s}\operatorname{diag}(\boldsymbol{\mu})^2 + \operatorname{diag}(\boldsymbol{\sigma})^2. && \text{(C.13)}
\end{aligned}$$

C.2 Proof of Theorem A.1

In this subsection for the notation-wise convenience we denote $\boldsymbol{W} = \boldsymbol{\theta}$. Since the model is one layer, the loss function eq. (A.1) becomes

$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{2}\left\|(\boldsymbol{W}-\boldsymbol{I})\boldsymbol{A}^{1/2}\right\|_{\mathcal{F}}^2, \tag{C.14}$$

and the gradient is

$$\nabla\mathcal{L}(\boldsymbol{W}) = (\boldsymbol{W}-\boldsymbol{I})\boldsymbol{A} = \boldsymbol{W}\boldsymbol{A} - \boldsymbol{A}. \tag{C.15}$$

We denote the k-th row of $\boldsymbol{W}$ and $\boldsymbol{A}$ by $\boldsymbol{w}_k$ and $\boldsymbol{A}_k$ respectively. Then we have

$$\dot{\boldsymbol{w}}_k = -\boldsymbol{A}\boldsymbol{w}_k + \boldsymbol{a}_k. \tag{C.16}$$

The solution of this differential equation is

$$\boldsymbol{w}_k(t) = \exp\left(-\boldsymbol{A}t\right)\left[w_k(0) - \boldsymbol{A}^{-1}\boldsymbol{a}_k\right] + \boldsymbol{A}^{-1}\boldsymbol{a}_k, \tag{C.17}$$

where we use the convention $0 \times \left(0^{-1}\right) = 0$ to avoid the non-invertible case of $\boldsymbol{A}$.

Thus for any $\boldsymbol{z} \in \mathbb{R}^d$ we have

$$
\begin{aligned}
f(\boldsymbol{W}(t);\boldsymbol{z})_k &= \langle \boldsymbol{w}_k(t), \boldsymbol{z} \rangle && \text{(C.18)}\\
&= \left\langle \left(\boldsymbol{I} - e^{-\boldsymbol{A}t}\right) \boldsymbol{A}^{-1}\boldsymbol{a}_k, \boldsymbol{z} \right\rangle + \left\langle e^{-\boldsymbol{A}t} w_k(0), \boldsymbol{z} \right\rangle && \text{(C.19)}\\
&= \sum_{p=1}^{n} \frac{1 - e^{-a_p t}}{a_p} \mathbb{1}_{\{k=p\}} a_p z_p + \sum_{i=1}^{n} e^{-a_i t} w_{k,i}(0) z_i && \text{(C.20)}\\
&= \mathbb{1}_{\{k \le s\}} \left(1 - e^{-a_k t}\right) z_k + \sum_{i=1}^{n} e^{-a_i t} w_{k,i}(0) z_i, && \text{(C.21)}
\end{aligned}
$$

and this proves the claim.

D Theoretical Analysis of the Two Layer Model

In this section we provide a detailed analysis of the symmetric two layer model described in Appendix A.2.

In the theory part we frequently consider functions of a variable t which is explained as time. If a function $g(t)$ is a function of time t, we denote the derivative of g w.r.t. t by $\dot{g}(t) = \left.\frac{\mathrm{d}g}{\mathrm{d}t}\right|_{t=t}$. Moreover, we sometimes omit the argument t, i.e. g means $g(t)$ for any time t.

For a statement ϕ, we define $\mathbb{1}_{\{\phi\}} = \begin{cases} 1 & \phi \text{ is true} \\ 0 & \phi \text{ is false} \end{cases}$ that indicates the Boolean value of ϕ.

In this section we assume a finite step size, i.e. $\boldsymbol{W} : \mathbb{N} \to \mathbb{R}^{d \times d}$ is initialized by $\boldsymbol{W}(0)$ and updated by

$$
\begin{aligned}
\frac{\boldsymbol{W}(t+1) - \boldsymbol{W}(t)}{\eta} &= -\boldsymbol{U}(t)\nabla\mathcal{L}(\boldsymbol{U}(t))^\top - \nabla\mathcal{L}(\boldsymbol{U}(t))\boldsymbol{U}(t)^\top && \text{(D.1)}\\
&= \boldsymbol{W}(t)\boldsymbol{A} + \boldsymbol{A}\boldsymbol{W}(t) - \frac{1}{2}\left[\boldsymbol{A}\boldsymbol{W}(t)^2 + \boldsymbol{W}(t)^2\boldsymbol{A} + 2\boldsymbol{W}(t)\boldsymbol{A}\boldsymbol{W}(t)\right]. && \text{(D.2)}
\end{aligned}
$$

The update of each entry $w_{i,j}(t)$ can be decomposed into three terms, as we described in the main text:

$$
\begin{aligned}
\frac{w_{i,j}(t+1) - w_{i,j}(t)}{\eta} =& w_{i,j}(t)(a_i + a_j) - \frac{1}{2}\sum_{k=1}^{d} w_{k,i} w_{k,j} (a_i + a_j + 2a_k) && \text{(D.3)}\\
=& \underbrace{w_{i,j}(t)(a_i + a_j)}_{G_{i,j}(t)} && \text{(D.4)}\\
& \underbrace{- \frac{1}{2} w_{i,j}\left[w_{i,i}(3a_i + a_j) + \mathbb{1}_{\{i \neq j\}} w_{j,j}(3a_j + a_i)\right]}_{S_{i,j}(t)} && \text{(D.5)}\\
& \underbrace{- \frac{1}{2} \sum_{\substack{k \neq i \\ k \neq j}} w_{k,i}(t) w_{k,j}(t)(a_i + a_j + 2a_k)}_{N_{i,j}(t)}. && \text{(D.6)}
\end{aligned}
$$

D.1 Assumptions

We need make several assumptions to prove the results. Below we make several assumptions that all commonly hold in the practice. The first assumption to make is that both the value of a_k and the initialization of $\boldsymbol{W}$ is bounded.

Assumption D.1 (Bounded Initialization and Signal Strength)**.** *There exists* $\alpha > 0, \gamma > 1, \beta > 1$ *such that*

$$\forall k, \alpha \leq a_k \leq \gamma\alpha, \tag{D.7}$$

$$\forall i, j, \omega \leq |w_{i,j}(0)| \leq \beta\omega. \tag{D.8}$$

The second assumption is that the step size is small enough.

Assumption D.2 (Small Step Size)**.** *There exists a constant* $K \geq 20$*, such that* $\eta \leq \frac{1}{9K\gamma\alpha}$.

Next, we define a concept called initial phase. The definition of initial phase is related to a constant $P > 0$.

Definition D.1. *Assume there is a constant* $P > 0$*. For an entry* (i, j) *and time* t*, if* $|w_{i,j}(t)| \leq P\beta\omega$*, we say this entry is in* ***initial phase****.*

The next assumption to make the that the boundary of the initial phase should not be too large.

Assumption D.3 (Small Initial Phase)**.** $P\omega\beta \leq 0.4$.

The next assumption to make are that the intialization value (ω) should not be too large.

Assumption D.4 (Small Initialization)**.**

$$\omega \leq \min\left\{\frac{\min\{\kappa - 1, 1 - \kappa^{-1/2}\}}{PK\gamma d\beta^2}, \frac{1}{\sqrt{2\beta}}\right\} \tag{D.9}$$

and $\kappa > 1.1$*, and* $\kappa \leq 1 + \frac{1}{2}KC^{-1}$*,* $P \geq 2$*.*

Finally, we also assume that the signal strength difference is significant enough.

Assumption D.5 (Significant Signal Strength Difference)**.** *For any* $i > j$*, we have*

$$\frac{a_i + a_i}{2a_i} \leq \frac{\log P}{10\kappa^2 \log \frac{1}{P\beta\omega} + \log P\beta}. \tag{D.10}$$

and there exists a constant $C > 1$ *such that* $a_i - 3a_j \geq C^{-1}\alpha$*.*

D.2 The Characterization of the Evolution of the Jacobian

In this subsection, we provide a series of lemmas that characterize each stage the evolution of the Jacobian matrix $\boldsymbol{W}$.

The whole proof is based on induction, and in order to avoid a too complicated induction, we make the following assertion, which obviously holds at initialization.

Assertion D.1. *For all* $t \in \mathbb{N}$*, if* $i \neq j$*, then the entry* (i, j) *stays in the initial phase for all time.*

We will use Assertion D.1 as an assumption throughout the proves and prove it at the end. This is essentially another way of writing inductions.

We have the following corollary that directly followed by Assertion D.1.

Corollary D.1. *For all* $t \in \mathbb{N}$ *and all* i, j*,* $|N_{i,j}(t)| \leq 2P\gamma\alpha d\beta^2\omega^2$*.*

Now, we are ready to present and prove the major lemmas. The first lemma is to post a (rather loose) upper bound of the value of the entries.

Lemma D.1 (Upper Bounded Growth)**.** *Consider entry* (i, j)*. We have for all* $t \in \mathbb{N}$*, at timepoint* t *the absolute value of the* (i, j)*-th entry satisfies*

$$|w_{i,j}(t)| \leq |w_{i,j}(0)| \exp\left[\eta t(a_i + a_j)\kappa\right]. \tag{D.11}$$

Proof. Since of the $N_{i,j}$ term we only use its absolute value, the positive case and negative case are symmetric. WLOG we only consider the case where $w_{i,j}(0) > 0$ here.

The claim is obviously satisfied at initialization. We use it as the inductive hypothesis. Suppose at timepoint $t \leq T - 1$ the claim is satisfied, we consider the time step $t + 1$.

Since Assertion D.1 guaranteed that every non-diagonal entry is in the initial phase, and the $S_{i,j}$ term has different symbol with $w_{i,j}(0)$, we have

$$S_{i,j}(t) + N_{i,j}(t) \leq 2P\gamma\alpha d\beta^2\omega^2. \tag{D.12}$$

We have

$$\begin{aligned} w_{i,j}(t+1) - w_{i,j}(t) &\leq \eta w_{i,j}(t)(a_i + a_j) + 4\eta\gamma\alpha d\beta_0\omega^2 && \text{(D.13)}\\ &\leq \eta(a_i + a_j) w_{i,j}(0) \exp\left[\eta t(a_i + a_j)\kappa\right] + 2P\eta\gamma\alpha d\beta^2\omega^2 && \text{(D.14)}\\ &= w_{i,j}(0) \exp\left[\eta t(a_i + a_j)\kappa\right] \left[\eta(a_i + a_j) + \frac{2P\gamma\alpha d\beta^2\omega^2}{w_{i,j}(0)\exp\left[\eta t(a_i + a_j)\kappa\right]}\right] && \text{(D.15)} \end{aligned}$$

From Assumption D.4, we have

$$\begin{aligned} \eta(a_i + a_j) + \frac{2P\gamma\alpha d\beta^2\omega^2}{w_{i,j}(0)\exp\left[\eta t(a_i + a_j)\kappa\right]} &\leq \eta(a_i + a_j) + 2P\gamma\alpha d\beta^2\omega && \text{(D.16)}\\ &\leq \eta(a_i + a_j) + 2(\kappa - 1)\eta\alpha && \text{(D.17)}\\ &\leq \kappa\eta(a_i + a_j) && \text{(D.18)}\\ &\leq \exp(\kappa\eta[a_i + a_j]) - 1, && \text{(D.19)} \end{aligned}$$

thus we have

$$\begin{aligned} w_{i,j}(t+1) \leq & w_{i,j}(t) + \left[\exp(\kappa\eta[a_i + a_j]) - 1\right] w_{i,j}(t) && \text{(D.20)}\\ & \leq w_{i,j}(0) \exp\left[\eta(t+1)(a_i + a_j)\kappa\right]. && \text{(D.21)} \end{aligned}$$

Finally, notice that since $T_1 = \frac{\kappa \log P}{2\eta\gamma\alpha} \leq \frac{\kappa \log 2}{\eta(a_i + a_j)}$, we have

$$\exp\left[\eta T(a_i + a_j)\kappa^{-1}\right] \leq P. \tag{D.22}$$

□

Next, we prove that Lemma D.1 is tight in the initial stage of the training, up to a constant κ in the exponential term.

Lemma D.2 (Lower Bounded Initial Growth). *Let $T_1 = \frac{\log P}{2\eta\gamma\alpha\kappa}$. We have for all $t \in [T_1]$, at timepoint t every entry (i, j) is in the initial phase, and the absolute value of the (i, j)-th entry satisfies*

$$|w_{i,j}(t)| \geq |w_{i,j}(0)| \exp\left[\eta t(a_i + a_j)\kappa^{-1}\right] \tag{D.23}$$

and $w_{i,j}(t)w_{i,j}(0) > 0$.

Proof. Similar to the proof of Lemma D.1, we may just assume $w_{i,j}(0) > 0$.

Moreover, we also use the claim as an inductive hypothesis and prove it by induction. Since here the inductive hypothesis states that every entry is in the initial phase, we have

$$|S_{i,j}(t) + N_{i,j}(t)| \leq 4\gamma\alpha d\beta^2\omega^2. \tag{D.24}$$

We have

$$\begin{aligned} w_{i,j}(t+1) - w_{i,j}(t) &\geq \eta(a_i + a_j) w_{i,j}(0) \exp\left[\eta t(a_i + a_j)\kappa^{-1}\right] - 2P\eta\gamma\alpha d\beta^2\omega^2 && \text{(D.25)}\\ &= w_{i,j}(0) \exp\left[\eta t(a_i + a_j)\kappa^{-1}\right] \left[\eta(a_i + a_j) - \frac{2P\eta\gamma\alpha d\beta^2\omega^2}{w_{i,j}(0)\exp\left[\eta t(a_i + a_j)\kappa^{-1}\right]}\right] && \text{(D.26)} \end{aligned}$$

From Assumption D.4, we have

$$\begin{aligned} \frac{2P\eta\gamma\alpha d\beta^2\omega^2}{w_{i,j}(0)\exp\left[\eta t(a_i + a_j)\kappa^{-1}\right]} &\leq 2P\eta\gamma\alpha d\beta^2\omega && \text{(D.27)}\\ &\leq \left(1 - \kappa^{-1/2}\right)\eta(a_i + a_j). && \text{(D.28)} \end{aligned}$$

Moreover, notice that when $\kappa > 1.1$, for any $x < 0.1$, we have $\kappa^{-1/2}x + 1 \geq e^{\kappa^{-1}x}$. Since Assumption D.2 ensured that $\eta \leq \frac{1}{10(a_i+a_j)}$, we have

$$w_{i,j}(t+1) \geq w_{i,j}(t) + w_{i,j}(t)\left[\kappa^{-1/2}\eta(a_i+a_j)\right] \tag{D.29}$$
$$\geq w_{i,j}(t)\exp\left(\eta(a_i+a_j)\kappa^{-1}\right) \tag{D.30}$$
$$\geq w_{i,j}(0)\exp\left[\eta(t+1)(a_i+a_j)\kappa^{-1}\right]. \tag{D.31}$$

Finally, from lemma D.1, we have when

$$w_{i,j}(t) \leq |w_{i,j}(0)|\exp\left(\eta t(a_i+a_j)\kappa\right) \tag{D.32}$$
$$\leq \beta\omega\exp\left(2\eta T_1\gamma\alpha\kappa\right) \tag{D.33}$$
$$\leq P\beta\omega, \tag{D.34}$$

which confirms that every entry (i, j) stays in the initial phase before time T_1.

□

Notice that the time bound in Lemma D.2 is a uniform one which applies to all entries. For the major entries, we might want to consider a finer bound of the time that it leaves the initial phase. This can be proved by essentially repeating the same proof idea of Lemma D.2.

Lemma D.3 (Lower Bounded Initial Growth for Diagonal Entries)**.** *Consider an diagonal entry (i, i). Let $T_1^{(i)} = \frac{\log\frac{P\beta\omega}{w_{i,i}(0)}}{2\eta a_i\kappa}$. We have for all $t \in \left[T_1^{(i)}\right]$, at timepoint t the entry (i, i) is in the initial phase, and the absolute value of the (i, i)-th entry satisfies*

$$w_{i,i}(t) \geq w_{i,i}(0)\exp\left[2\eta t a_i\kappa^{-1}\right]. \tag{D.35}$$

We omit the proof of Lemma D.3 since it is almost identical to the proof of Lemma D.2, only with replacing $\gamma\alpha$ by a_i and $\beta\omega$ by $w_{i,i}(0)$.

Next, we characterize the behavior of one diagonal entry after it leaves the initial phase.

Lemma D.4 (Lower Bounded After-Initial Growth for Diagonal Entries)**.** *Consider a diagonal entry (i, i). If at time t_0 we have $|w_{i,i}(t_0)| \geq P\beta\omega$, and for a $\lambda \in (P\beta\omega, 1-K^{-1})$, before time $T^{(\lambda)}$ we have $w_{i,i}(t+t_0) < \lambda$ for all $t \in [T^{(\lambda)}]$, then we have*

$$w_{i,i}(t+t_0) \geq w_{i,i}(t_0)\exp\left[2\eta t a_i(1-\lambda)\kappa^{-1}\right]. \tag{D.36}$$

Moreover, $w_{i,i}(0), w_{i,i}(t_0), w_{i,i}(t_0+t) \geq 0$.

Proof. Notice that since $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{U}^\top$ is a PSD matrix, its diagonal entries are always non-negative, this ensures that $w_{i,i}(0), w_{i,i}(t_0), w_{i,i}(t_0+t) \geq 0$.

For the time after t_0 and before $t_0 + T^{(\lambda)}$, we use an induction to prove the claim, with the claim itself as the inductive hypothesis. It clearly holds when $t = 1$.

Notice that when $w_{i,j}(t') < \lambda$, we have

$$G_{i,j}(t') - S_{i,j}(t') = 2a_i w_{i,i}(t')\left[1 - w_{i,i}(t')\right] \geq 2a_i w_{i,i}(t')(1-\lambda). \tag{D.37}$$

Thus we have

$$w_{i,i}(t_0+t+1) - w_{i,i}(t_0+t) \tag{D.38}$$
$$\geq\ 2\eta a_i(1-\lambda)w_i(t_0)\exp\left[\eta t(a_i+a_j)(1-\lambda)\kappa^{-1}\right] - 2P\eta\gamma\alpha d\beta^2\omega^2 \tag{D.39}$$
$$=\ w_{i,i}(t_0)\exp\left[2\eta t a_i(1-\lambda)\kappa^{-1}\right]\left[2\eta a_i(1-\lambda) - \frac{2P\eta\gamma\alpha d\beta^2\omega^2}{w_{i,i}(t_0)\exp\left[2\eta t a_i(1-\lambda)\kappa^{-1}\right]}\right] \tag{D.40}$$

Since $\lambda < 1 - K^{-1}$, and $w_{i,i}(t_0) \geq 2\beta\omega \geq \omega$, from Assumption D.4, we have

$$\frac{2P\eta\gamma\alpha d\beta^2\omega^2}{w_{i,i}(t_0)\exp\left[2\eta t a_i(1-\lambda)\kappa^{-1}\right]} \leq 2P\eta\gamma\alpha d\beta^2\omega \tag{D.41}$$
$$\leq 2K^{-1}\left(1-\kappa^{-1/2}\right)\eta\alpha \tag{D.42}$$
$$\leq 2\left(1-\kappa^{-1/2}\right)\eta a_i(1-\lambda). \tag{D.43}$$

Moreover, since Assumption D.2 ensured that $\eta \leq \frac{1}{2Ka_i(1-\lambda)} \leq \frac{1}{20a_i(1-\lambda)}$, using the fact that if $\kappa > 1.1$ then $\kappa^{-1/2}x + 1 \geq e^{\kappa^{-1}x}$ for any $x < 0.1$, we can get

$$w_{i,i}(t+1) \geq w_{i,i}(t) + w_{i,i}(t)\left[\kappa^{-1/2}2\eta a_i(1-\lambda)\right] \tag{D.44}$$

$$\geq w_{i,i}(t)\exp\left(2\eta a_i\kappa^{-1}(1-\lambda)\right) \tag{D.45}$$

$$\geq w_{i,i}(t_0)\exp\left[2\eta(t+1)\kappa^{-1}(1-\lambda)\right]. \tag{D.46}$$

□

Next, we provide an uniform upper bound (over time) of the diagonal entries. Remember that we mentioned in the gradient flow case, the diagonal term stops evolving when it reaches 1. In the discrete case, since the step size is not infinitesimal, Lemma D.5 shows that it can actually exceed 1 a little bit but not too much since the step size is small.

Lemma D.5 (Upper Bounded Diagonal Entry)**.** *For any diagonal entry* (i,i) *and any time* t*,* $0 \leq w_{i,i}(t) \leq 1 + 2K^{-1}$.

Proof. First notice that since $\boldsymbol{W}(t)$ is PSD, its diagonal entry $w_{i,i}(t)$ should always be non-negative, thus $w_{i,i}(t) \geq 0$ is always satisfied. In the following we prove $w_{i,i}(t) \leq 1 + 2K^{-1}$.

We use induction to prove this claim. The inductive hypothesis is the claim it self. It is obviously satisfied at initialization. In the following we assume the claim is satisfied at timepoint t and prove it for timepoint $t+1$. Notice that since $K \leq 10$, we have $1 + K^{-1} \leq 2$.

Notice that by Assertion D.1 and Assumption D.4,

$$|N_{i,i}(t)| \leq 2P\gamma\alpha d\beta^2\omega^2 \leq \frac{(\kappa-1)^2}{K^2\gamma d\beta^2}\alpha \leq K^{-1}a_i. \tag{D.47}$$

If $w_{i,i}(t) \geq 1 + K^{-1}$, we have

$$G_{i,i}(t) - S_{i,i}(t) = 2a_i w_{i,i}(1 - w_{i,i}) \leq -4a_iK^{-1}. \tag{D.48}$$

Therefore,

$$w_{i,i}(t+1) = w_{i,i}(t) + \eta\left[G_{i,i}(t) - S_{i,i}(t) - N_{i,i}(t)\right] \tag{D.49}$$

$$\leq w_{i,i}(t) - 3a_iK^{-1}\eta \tag{D.50}$$

$$\leq w_{i,i}(t) \tag{D.51}$$

$$\leq 1 + 2K^{-1}. \tag{D.52}$$

Moreover, since $w_{i,i}(t) \leq 1 + 2K^{-1} \leq 2$, we have

$$|G_{i,i}(t)| + |S_{i,i}(t)| + |N_{i,i}(t)| \leq 4a_i + 4a_i + K^{-1}a_i \leq 9\gamma\alpha \leq \frac{1}{K\eta}. \tag{D.53}$$

When $w_{i,i}(t) \leq 1 + K^{-1}$, we have

$$w_{i,i}(t+1) \leq w_{i,i}(t) + \eta\left(|G_{i,i}(t)| + |S_{i,i}(t)| + |N_{i,i}(t)|\right) \leq 1 + 2K^{-1}. \tag{D.54}$$

The above results together shows that $w_{i,i}(t+1) \leq 1 + 2K^{-1}$.

□

Corollary D.2 (Upper Bounded Diagonal Update)**.** *For any diagonal entry* (i,i) *and any time* t*,* $|w_{i,i}(t+1) - w_{i,i}(t)| \leq K^{-1}$.

Corollary D.2 is a direct consequence of Lemma D.5 (and we actually proved Corollary D.2 in the proof of Lemma D.5).

The next lemma lower bounds the final value of diagonal entries. Together with Lemma D.5 we show that in the terminal stage of training the diagonal entries oscillate around 1 by the amplitude not exceeding $2K^{-1}$.

Lemma D.6. *Consider a diagonal entry (i,i). If at time t_0 we have $w_{i,i}(t_0) \geq 1 - 2K^{-1}$, then for all $t' \geq t_0$ we have $w_{i,i}(t') \geq 1 - 2K^{-1}$.*

Proof. we use an induction. The inductive hypothesis the claim itself. This obviously holds when $t' = t_0$. We assume $w_{i,i}(t') \geq 1 - 2K^{-1}$ at timepoint t' and prove the claim for $t' + 1$.

If $w_{i,i}(t') < 1 - K^{-1}$, then from Lemma D.4 we know

$$w_{i,i}(t'+1) \geq w_{i,i}(t') \geq 1 - 2K^{-1}. \tag{D.55}$$

If $w_{i,i}(t') > 1 - K^{-1}$, then from Corollary D.2 we have

$$w_{i,i}(t'+1) \geq w_{i,i}(t') - K^{-1} \geq 1 - 2K^{-1}. \tag{D.56}$$

□

Now, we are ready to prove Assertion D.1 by considering the suppression. We first prove a lemma that upper bounds the absolute value of the minor entries after its corresponding major entry becomes significant.

Lemma D.7 (Suppression)**.** *Consider an off-diagonal entry (i,j) where $i > j$. If there exists a time t_0 such that $w_{i,i}(t_0) > 0.8$, then for any $t' \geq t_0$ we have*

$$|w_{i,j}(t')| \leq \max\left\{|w_{i,j}(t_0)|, \omega\right\}. \tag{D.57}$$

Proof. Since $K > 10$, from Lemma D.6 and Lemma D.4 we know $w_{i,i}(t') > 0.8$ for all $t' \geq t_0$.

In this proof, we use an induction with the inductive hypothesis being the claim itself, i.e. we assume the claim is true at timepoint t' and prove it for $t' + 1$. The claim obviously holds for $t' = t_0$.

Since in this proof we only use the absolute value of $N_{i,j}$, WLOG we may assume that $w_{i,j}(t') > 0$.

If $w_{i,j}(t') < \omega$ then we have proved the claim. In the following we may assume $w_{i,j}(t') \geq \omega$.

We have

$$\begin{aligned}
G_{i,j}(t') - S_{i,j}(t') &\leq w_{i,j}(t')(a_i + a_j) - \frac{1}{2} w_{i,j}(t') w_{i,i}(3a_i + a_j) && \text{(D.58)}\\
&\leq w_{i,j}(t')(a_i + a_j) - w_{i,j}(t')\left[0.4(3a_i + a_j)\right] && \text{(D.59)}\\
&= -\frac{1}{5} w_{i,j}(t') a_i + \frac{3}{5} w_{i,j}(t') a_j && \text{(D.60)}\\
&\overset{\text{(i)}}{\leq} -C^{-1}\omega\alpha, && \text{(D.61)}
\end{aligned}$$

where in (i) we use Assumption D.5.

Thus we have

$$\begin{aligned}
G_{i,j}(t') - S_{i,j}(t') - N_{i,j}(t') &\leq G_{i,j}(t') - S_{i,j}(t') + |N_{i,j}(t')| && \text{(D.62)}\\
&\leq -C^{-1}\omega\alpha + 2P\gamma\alpha d\beta^2\omega^2 && \text{(D.63)}\\
&\overset{(i)}{<} 0, && \text{(D.64)}
\end{aligned}$$

where (i) is from Assumption D.4 and Assumption D.5. This confirms that $w_{i,j}(t'+1) < w_{i,j}(t') \leq \max\{|w_{i,j}(t_0), \omega\}$.

Next, we prove $w_{i,j}(t'+1) \geq -\max\{|w_{i,j}(t_0)|, \omega\}$. Notice that Lemma D.5 stated that $|w_{i,i}| \leq 2$. Notice that we also have $w_{i,j}(t') \leq K^{-1}$, thus

$$\begin{aligned}
|G_{i,j}(t')| + |S_{i,j}(t')| + |N_{i,j}(t')| &\leq 10\gamma\alpha|w_{i,j}(t')| + 2P\gamma\alpha d\beta^2\omega^2 && \text{(D.65)}\\
&\leq \frac{10|w_{i,j}(t')| + 2Pd\beta^2\omega^2}{9K\eta} && \text{(D.66)}\\
&\leq \frac{10|w_{i,j}(t')| + 2\omega}{9K\eta} && \text{(D.67)}\\
&\leq \frac{|w_{i,j}(t')| + \omega}{2\eta}. && \text{(D.68)}
\end{aligned}$$

We have

$$
\begin{aligned}
w_{i,j}(t'+1) &\geq w_{i,j}(t') - \eta(|G_{i,j}(t')| + |S_{i,j}(t')| + |N_{i,j}(t)'|) && \text{(D.69)}\\
&\geq -\eta(|G_{i,j}(t')| + |S_{i,j}(t')| + |N_{i,j}(t')|) && \text{(D.70)}\\
&\geq -\frac{1}{2}(|w_{i,j}(t')| + \omega) && \text{(D.71)}\\
&\geq -\max\{|w_{i,j}(t')|, \omega\}. && \text{(D.72)}
\end{aligned}
$$

□

With all the lemmas proved above, we are now ready to prove Assertion D.1.

Lemma D.8 (Assertion D.1)**.** *For all $t \in \mathbb{N}$, if $i \neq j$, then the entry (i, j) stays in the initial phase for all time.*

Proof. Notice that since $\boldsymbol{W}$ is symmetric, we only need to prove the claim for $i > j$. Moreover, From Lemma D.7, we only need to prove that there exists a timepoint t^*, such that $w_{i,i}(t^*) \geq 0.8$, and $|w_{i,j}(t^*)| \leq P\beta\omega$.

Let $t_0 = \frac{\log \frac{P\beta\omega}{w_{i,i}(0)}}{2\eta a_i \kappa}$, by Lemma D.3, we have $w_{i,i}(t_0) \geq P\beta\omega$. By Lemma D.3 and Lemma D.4, we have for any $t \geq t_0$ such that $w_{i,i}(t) \leq \lambda$, where $\lambda = 0.85$,

$$
\begin{aligned}
w_{i,i}(t) &\geq w_{i,i}(t_0) \exp\left[0.3\eta(t - t_0)a_i\kappa^{-1}\right] && \text{(D.73)}\\
&\geq P\beta\omega \exp\left[0.3\eta(t - t_0)a_i\kappa^{-1}\right] && \text{(D.74)}
\end{aligned}
$$

Let t' be the first time that $w_{i,i}(t')$ arrives above 0.8. Let $t^* = \min\left\{\frac{\kappa \log \frac{0.8}{P\beta\omega}}{0.3\eta a_i} + t_0, t'\right\} \geq t_0$. If $t^* = t'$, we have $w_{i,i}(t^*) \geq 0.8$. If $t^* = \frac{\log \frac{P\beta\omega}{w_{i,i}(0)}}{2\eta a_i \kappa} + t_0$, we have

$$
\begin{aligned}
w_{i,i}(t^*) &\geq w_{i,i}(0) \exp\left(0.3\eta t^* a_i \kappa^{-1}\right) && \text{(D.75)}\\
&\geq P\beta\omega \exp\left(\log \frac{0.8}{P\beta\omega}\right) && \text{(D.76)}\\
&\geq 0.8. && \text{(D.77)}
\end{aligned}
$$

Moreover, from Lemma D.1 and Assumption D.5, we have

$$
\begin{aligned}
|w_{i,j}|(t^*) &\leq |w_{i,j}(0)| \exp\left[\eta t^* \kappa(a_i + a_j)\right] && \text{(D.78)}\\
&\leq \beta\omega \exp\left[\left(\frac{\kappa^2 \log \frac{0.8}{P\beta\omega}}{0.15} + \log \frac{P\beta\omega}{w_{i,i}(0)}\right) \times \frac{a_i + a_j}{2a_i}\right] && \text{(D.79)}\\
&\leq \beta\omega \exp\left[\left(10\kappa^2 \log \frac{1}{P\beta\omega} + \log P\beta\right) \times \frac{a_i + a_j}{2a_i}\right] && \text{(D.80)}\\
&\leq \beta\omega \exp\left[\log(P)\right] && \text{(D.81)}\\
&\leq P\omega\beta. && \text{(D.82)}
\end{aligned}
$$

The claim is thus proved by combining the above bounds on $|w_{i,j}(t^*)|$ and $w_{i,i}(t^*)$ with Lemma D.7.

□

E Debunking Challenge Submission

E.1 What commonly-held position or belief are you challenging?

Provide a short summary of the body of work challenged by your results. Good summaries should outline the state of the literature and be reasonable, e.g. the people working in this area will agree with your overview. You can cite sources beside published work (e.g., blogs, talks, etc).

People generally believe that double descent with respect to training time (or "epochwise double descent") only happens either 1) with large step size or noisy training process; or 2) under over-parameterized settings.

E.2 How are your results in tension with this commonly-held position?

Detail how your submission challenges the belief described in (1). You may cite or synthesize results (e.g. figures, derivations, etc) from the main body of your submission and/or the literature.

In Section 3.2, we show a epochwise double descent phenomenon that happens when using large training set and small step size. Our theoretical analysis in Appendix A.2 shows that even in gradient flow (infinitesimal stepsize) and infinite data limit, this epochwise double descent still exists (see Figure 7). This suggests that the epochwise double descent found in our setting is not introduced by noise in the training or overfitting the data but an inherent property of the task itself.

We note that this contradictory comes from the out-of-distribution nature of the SIM task we consider: when training is noiseless and model is under-parameterized, generally there will not be in-distribution epochwise double descent, but once we consider an out-of-distribution task, even if it is as simple as learning identity mapping, there can still be epochwise double descent.

E.3 How do you expect your submission to affect future work?

Perhaps the new understanding you are proposing calls for new experiments or theory in the area, or maybe it casts doubt on a line of research.

We would like to call for the awareness of this kind of epochwise double descent that is intrinsic to certain OOD tasks. Since SIM is a very simple toy task but still show epochwise double descent, we hypothesize that this kind of epochwise double descent might presence across many tasks. It is also an important direction for future work that to (either empirically or theoretically) fully characterize the scenarios that has this kind of epochwise double descent. Specifically, since it is not caused by training or (in-distribution) generalization issues, it is likely caused by the structure of the input data. Therefore, to determine what kind of input data structure will / will not cause this kind of epochwise double descent is a very interesting direction to explore.