Provable Effects of Data Replay in Continual Learning: A Feature Learning Perspective

Meng Ding[2] **Jinhui Xu**[1] **Kaiyi Ji**[2]
[1] School of Information Science and Technology, USTC and Institute of Artificial Intelligence, HCNSC
[2] Department of Computer Science and Engineering, SUNY at Buffalo

Abstract

Continual learning (CL) aims to train models on a sequence of tasks while retaining performance on previously learned ones. A core challenge in this setting is *catastrophic forgetting*, where new learning interferes with past knowledge. Among various mitigation strategies, data-replay methods—where past samples are periodically revisited—are considered simple yet effective, especially when memory constraints are relaxed. However, the theoretical effectiveness of full data replay, where all past data is accessible during training, remains largely unexplored. In this paper, we present a comprehensive theoretical framework for analyzing full data-replay training in continual learning from a feature learning perspective. Adopting a multi-view data model, we identify the signal-to-noise ratio (SNR) as a critical factor affecting forgetting. Focusing on task-incremental binary classification across M tasks, our analysis verifies two key conclusions: (1) forgetting can still occur under full replay when the cumulative noise from later tasks dominates the signal from earlier ones; and (2) with sufficient signal accumulation, data replay can recover earlier tasks-even if their initial learning was poor. Notably, we uncover a novel insight into task ordering: prioritizing higher-signal tasks not only facilitates learning of lower-signal tasks but also helps prevent catastrophic forgetting. We validate our theoretical findings through synthetic and real-world experiments that visualize the interplay between signal learning and noise memorization across varying SNRs and task correlation regimes.

1 INTRODUCTION

Continual learning (CL) is a paradigm in machine learning where models learn sequentially from a stream of tasks or datasets, continually adapting to new information while preserving performance on previously learned tasks (Parisi et al., 2019; Wang et al., 2024). The key challenge in continual learning is *catastrophic forgetting*, a phenomenon where modern models drastically lose previously acquired knowledge when learning new tasks (McCloskey and Cohen, 1989; Kirkpatrick et al., 2017; Korbak et al., 2022).

Previous empirical research alleviating catastrophic forgetting in continual learning can be broadly classified into five categories (Wang et al., 2024): regularization-, replay-, optimization-, representation-, and architecture-based approaches. Regularization-based methods (Ritter et al., 2018; Aljundi et al., 2018; Titsias et al., 2019; Pan et al., 2020; Benzing, 2022; Lin et al., 2022) introduce explicit regularizers to balance learning across tasks, often relying on a frozen copy of the old model for reference. Replay-based methods (Lopez-Paz and Ranzato, 2017; Riemer et al., 2018; Chaudhry et al., 2019; Yoon et al., 2021; Shim et al., 2021; Tiwari et al., 2022; Van de Ven et al., 2020; Liu et al., 2020; Zheng et al., 2024) approximate and recover past data distributions to reinforce old knowledge. Optimization-based methods (Lopez-Paz and Ranzato, 2017; Chaudhry et al., 2018; Tang et al., 2021; Liu et al., 2020; Wang et al., 2022a) focus on modifying the learning dynamics, such as through gradient projection, to avoid interference. Representation-based methods (Wu et al., 2022; Shi et al., 2022; Wang et al., 2022b; McDonnell et al., 2023; Le et al., 2024) aim to develop and leverage task-robust representations via the advantages of pre-training, while architecture-based methods (Gurbuz and Dovrolis, 2022; Douillard et al., 2022; Miao et al., 2021; Ostapenko et al., 2021) design adaptable model structures that share parameters across tasks to retain knowledge.

Among these approaches, data-replay methods are often regarded as the most straightforward to implement—particularly when buffer constraints are ignored—since they rely on storing and periodically retraining on past task samples to preserve prior knowledge. However, their empirical success typically hinges on careful sample selection (Chaudhry et al., 2019; Riemer et al., 2018). When full data replay is employed, exposing the model to all historical data, the effectiveness of this strategy remains an open question: does it still reliably counteract forgetting under such conditions?

To address this, we present a comprehensive theoretical analysis showing that full data-replay training does not always effectively mitigate forgetting.

Our contribution can be summarized as follows:

- We develop a thorough theoretical framework that rigorously analyzes full data-replay training within the theoretical continual learning community. Prior studies have primarily focused on simplified linear regression models, two-task setups, or naive sequential training, leaving fundamental gaps in understanding the behavior of replay-based methods in general multi-task settings (see section 2 for details). More specifically: (1) we adopt a multi-view data model (following Allen-Zhu and Li (2020)), where each data point consists of both feature signals and noise, allowing us to introduce the signal-to-noise ratio as a key factor governing whether forgetting occurs; and (2) we focus on task-incremental binary classification in a general M-task setting, where each task is associated with a distinct feature signal vector. This formulation enables us to characterize how task ordering and inter-task correlation influence forgetting.

- Based on the above data model, our results formally show two interesting findings: (1) Even with full data replay, forgetting of task k after replaying up to task m ($m > k$) can still occur under certain SNR regimes, particularly when the cumulative noise from later tasks outweighs the signal intensity of task k. (2) Even if the performance on task k is initially unsatisfactory, data replay can help amplify the signal intensity, enabling the model to recover task k 's information in later stages-provided the accumulated signal outweighs the noise. Furthermore, by incorporating task correlation, we uncover a key insight into task ordering: prioritizing higher-signal tasks not only facilitates learning for lower-signal tasks but can also help prevent catastrophic forgetting. This observation suggests a promising direction for designing order-aware replay strategies in future continual learning frameworks.

- We complement our theory with synthetic experiments that examine the dynamics of signal learning and noise memorization during continual training under full data replay, comparing different task orderings across varying levels of task correlation and SNR conditions.

2 RELATED WORK

Replay-based Continual Learning. Replay-based approaches mitigate catastrophic forgetting by approximating the original data distribution during continual training. Specifically, they can be categorized based on how they reconstruct previous data: (1) Experience replay. A small subset of historical samples is stored in a memory buffer and replayed alongside new data. Early work stored a fixed or class-balanced share of examples from each batch to enforce simple selection rules (Lopez-Paz and Ranzato, 2017; Riemer et al., 2018; Chaudhry et al., 2019). Later studies introduced gradient-aware or optimizable selection schemes to maximize sample diversity (Yoon et al., 2021; Shim et al., 2021; Tiwari et al., 2022), and used data-augmentation techniques to improve storage efficiency (Ebrahimi et al., 2021; Kumari et al., 2022). (2) Generative replay (pseudo-rehearsal). Instead of storing raw inputs, an auxiliary generative model is trained to synthesise data from previous tasks, and these pseudo-examples are replayed alongside new data during subsequent training. To mitigate forgetting in the generative model itself, additional strategies are often employed, such as weight regularization to preserve past knowledge (Nguyen et al., 2017; Wang et al., 2021), task-specific parameter allocation (e.g., binary masks) (Ostapenko et al., 2019; Cong et al., 2020) to reduce inter-task interference, and feature-level replay to simplify conditional generation by replaying intermediate features instead of raw data (Van de Ven et al., 2020; Liu et al., 2020). In practice, replay methods must work with a limited memory buffer. For analytical clarity, however, we assume an unlimited buffer that stores all past data; extending the theory to constrained-memory settings will be left for future work.

Theoretical Continual Learning. Recent theoretical work on catastrophic forgetting has focused mainly on linear regression models, leaving more complex settings largely unexplored. Evron et al. (2022) analyzed catastrophic forgetting under two task-ordering schemes—cyclic and random—using alternating projections and the Kaczmarz method to

pinpoint both the worst-case and the no-forgetting scenarios. Building on this, Swartworth et al. (2023) tightened nearly optimal forgetting bounds for cyclic orderings, and Evron et al. (2025) further improved the rates for random orderings with replacement. Additionally, Goldfarb and Hand (2023) provided analysis that overparameterization accounts for most of the performance loss caused by catastrophic forgetting. Lin et al. (2023) examined how overparameterization, task similarity, and task ordering jointly influence both forgetting and generalization error in continual learning, and Li et al. (2024b) extended this analysis by characterizing the role of Mixture-of-Experts (MoE) architectures. Ding et al. (2024) developed a general theoretical framework for catastrophic forgetting under Stochastic Gradient Descent, revealing that the task order shapes the extent of forgetting in continual learning. Zhao et al. (2024) offered a statistical perspective on regularization-based continual learning, showing how various regularizers affect model performance.

Beyond linear-regression settings, several studies have investigated catastrophic forgetting in neural networks settings. Doan et al. (2021) investigated catastrophic forgetting in the Neural Tangent Kernel (NTK) regime and showed that projected-gradient algorithms can mitigate forgetting by introducing a task-similarity measure called the NTK overlap matrix. Cao et al. (2022a) demonstrated that, for any target accuracy, one can keep the learned representation's dimension nearly as small as the true underlying representation with the proposed CL algorithm. The most relevant works to ours with data-replay strategies are (Banayeeanzade et al., 2024; Zheng et al.), where Banayeeanzade et al. (2024) primarily focuses on the comparison between multi-task learning and continual learning, while Zheng et al. extends previous continual learning theory to memory-based methods. Both works are limited to the linear regression setting and leave the behavior of more complex models unexplored. We will provide more discussion in section 4.

3 PRELIMINARIES

Problem Setup. In our setup, we consider a sequence of tasks denoted by $\mathbb{M} = \{1, 2, \ldots, M\}$. For each task m in this sequence, let $\{\mathbf{v}_m^*\}_{m\in[M]} \subseteq \mathbb{R}^d$ represent the feature vectors, where $\|\mathbf{v}_m^*\| = 1$ for all $m \in [M]$, and $\langle \mathbf{v}_m^*, \mathbf{v}_{m'}^* \rangle = A_{(m,m')} \geq 0$ whenever $m \neq m'$. Then, we define the data distributions for each task as follows.

Definition 1 (Data Distribution for Task m)**.** *For the task m, let $\mathbf{v}_m^* \in \mathbb{R}^d$ be a fixed vector representing the feature signal contained in each data point. Each data point $(\mathbf{x}_m, y_m)$ with input $\mathbf{x}_m = [\mathbf{x}_m^1, \mathbf{x}_m^2] \in (\mathbb{R}^d)^2$ and label $y \in \{+1, -1\}$ is generated from a data distribution $\mathcal{D}_m$ as follows:*

(1) The label $y \in \{-1, 1\}$ is sampled uniformly;

(2) The input $\mathbf{x}_m$ is generated as a vector of 2 patches, i.e., $\mathbf{x}_m = [\mathbf{x}_m^1, \mathbf{x}_m^2] \in (\mathbb{R}^d)^2$, where

- *Feature patch. The first patch is given by $\mathbf{x}_m^1 = \alpha_m y_m \cdot \mathbf{v}_m^*$, where $\alpha_m > 0$ indicates the signal intensity.*
- *Noise patch. The second patch is given by $\mathbf{x}_m^2 = \boldsymbol{\xi}_m$, where $\boldsymbol{\xi}_m \sim \mathcal{N}(0, \sigma_\xi^2 \cdot \mathbf{H})$ and is independent of the label y_m, where $\mathbf{H} = \mathbf{I}_d - \sum_{m=1}^M \mathbf{v}_m^* (\mathbf{v}_m^*)^\top$.*

Our data generation model is inspired by the structure of image data, which has been widely utilized in the feature learning theory area (Allen-Zhu and Li, 2020; Cao et al., 2022b; Jelassi and Li, 2022; Kou et al., 2023; Zou et al., 2023; Ding et al., 2025; Han et al., 2024; Li et al., 2024a; Bu et al., 2024, 2025; Han et al., 2025). Specifically, the input data comprises two patches, among which only a subset is relevant to the class label of the image. We denote this relevant part as $y_m \alpha_m \mathbf{v}_m^*$, where y_m represents the label, $\mathbf{v}_m^*$ is the corresponding feature signal vector, and $\alpha_m > 0$ indicates the intensity of the feature signal. As described in Definition 1, we assume that each task m has its own unique feature signal vector and that the feature vectors across tasks are correlated with the correlation strength $A_{(m,m')} > 0$. For instance, in a continual learning setting where the model first classifies cars and later bicycles, the initial task may use the car's wheel as a key feature and the subsequent task may use the bicycle's wheel. Because both wheels share similar shapes, this overlap promotes feature reuse and helps the model recognize both objects as forms of transportation. In contrast, the irrelevant patches, referred to as noise, are independent of the data label and do not contribute to prediction. We denote such noise as $\boldsymbol{\xi}$, which is assumed to follow a Gaussian distribution $\mathcal{N}(0, \sigma_\xi^2 \cdot \mathbf{H})$. For simplicity, the noise follows the same independent distribution for each task, and the noise vector is orthogonal to any feature signal vector $\mathbf{v}_m^*$.

Learner Model. Following existing work (Jelassi and Li, 2022; Bao et al.), we consider a one-hidden-layer convolutional neural network architecture equipped with the cubic activation function $\sigma(z) = z^3$:

$$F(\mathbf{W}, \mathbf{x}_m) = \sum_{r\in[R]} \sigma(\langle \mathbf{w}_r, \mathbf{x}_m^1 \rangle) + \sigma(\langle \mathbf{w}_r, \mathbf{x}_m^2 \rangle) = \sum_{r\in[R]} \sigma(\langle \mathbf{w}_r, \alpha_m y_m \mathbf{v}_m^* \rangle) + \sigma(\langle \mathbf{w}_r, \boldsymbol{\xi}_m \rangle), \tag{1}$$

Figure 1: Illustration of feature signals across multiple tasks using images from Salient ImageNet.

where R is the number of hidden neurons and $\mathbf{W} = \{\mathbf{w}_1, \ldots, \mathbf{w}_R\}$ represents the model weights. We denote the logistic loss function evaluated for the m-th task as

$$L(\mathbf{W}; D_m) = \frac{1}{n_m} \sum_{j \in [n_m]} \log\{1 + e^{-y_{mj} F(\mathbf{W}, \mathbf{x}_{mj})}\}. \quad (2)$$

Here, D_m is the training data set for task m with sample size n_m. To keep the analysis clean, we assume all tasks share the same sample size, i.e., $n_m = n$ for every m. We train the model from a Gaussian initialization, drawing each hidden weight $\mathbf{w}_r^{(0)}$ independently from $\mathcal{N}\left(0, \sigma_0^2 \mathbf{I}_d\right)$.

Data Replay Training. Starting with the randomly initialized point $\mathbf{W}_0$ and employing a constant step size η, the model is updated by data-replay training for task m over T iterations, with $t = 1, .., T$:

$$\mathbf{W}_m^{(t+1)} = \mathbf{W}_m^{(t)} - \frac{\eta}{m n_m} \sum_{j \in [n_m]} \nabla L(\mathbf{W}_m^{(t)}; D_1, D_2, ..., D_m). \quad (3)$$

Here, $\mathbf{W}_m^{(T)}$ denotes the parameter state after the completion of training on task m, which subsequently serves as the starting point for training on task $m+1$. In contrast to classical sequential training, the fully data-replay training incorporates all previous task datasets, $D_1, D_2, ..., D_m$, into the training of the current task model.

Catastrophic Forgetting. Catastrophic forgetting refers to the phenomenon where modern models substantially lose previously acquired knowledge when learning new tasks (McCloskey and Cohen, 1989). In the following, we provide a formal definition of this behavior in the context of continual learning over M tasks.

Definition 2 (Catastrophic Forgetting)**.** *Given a test data $(\mathbf{x}_k, y_k)$ drawn from the data distribution $\mathcal{D}_k$ of the k-th task, we claim Catastrophic Forgetting occurs if the following conditions hold:*

1. *After training on the k-th task (i.e., at iteration T_k), with high probability, the model correctly classifies the sample:*

$$\mathbb{P}\left\{y_k F\left(\mathbf{W}^{(T_k)}, \mathbf{x}_k\right) < 0\right\} \leq \frac{1}{\text{poly}(d)}.$$

2. *After training on the m-th task ($m > k$, at iteration T_m), with high probability, the model's performance on task k deteriorates:*

$$\mathbb{P}\left\{y_k F\left(\mathbf{W}^{(T_m)}, \mathbf{x}_k\right) < 0\right\} \geq \frac{1}{2} - \frac{1}{\text{polylog(d)}}.$$

4 MAIN RESULTS

In this section, we present our main results on the generalization performance for task k, evaluated after training on the k-th task and again after training on the m-th task ($m > k$) based on SNR $= \alpha_p / \sigma_\xi \sqrt{d}$, respectively. Before stating the theorems, we first introduce the conditions that underlie our analysis.

Condition 1. *For the data model described in Definition 1, we assume that the noise standard deviation scales as $\sigma_\xi = \Theta(d^{-0.51})$. For the random initialization of the model weights, we assume $\sigma_0 = \Theta\left((n/R)^{1/3} d^{-0.52}\right)$. Furthermore, we assume the model is overparameterized, with both the hidden dimension R and the sample size n are bounded by* $\text{polylog}(d)$.

Our conditions follow those in existing work (Jelassi and Li, 2022; Bao et al.), but without imposing assumptions on the signal intensity. This relaxation allows us to explicitly investigate how the signal-to-noise ratio (SNR) influences the behavior of data replay training in continual learning.

Theorem 1. *Suppose the setting in Condition 1 holds, and the SNR satisfies* $\frac{k^2}{R^{2/3}\sigma_0^2\sigma_\xi^2 d^{13/6}} \lesssim \frac{\sum_{p=1}^{k}(1-\frac{p-1}{k})\alpha_p^3 A_{(p,k)}}{(\sigma_\xi\sqrt{d})^3} \lesssim \frac{1}{n}$. *Consider full data-replay training with learning rate $\eta \in (0, \widetilde{O}(1)]$, and let $(\mathbf{x}_k, y_k) \sim \mathcal{D}_k$ be a test sample from the task k. Then, with high probability, there exist training times T_k and T_m $(m > k)$ such that*

- *The model fails to correctly classify task k immediately after learning it:*

$$\mathbb{P}\left\{y_k F\left(\mathbf{W}^{(T_k)}, \mathbf{x}_k\right) < 0\right\} \geq \frac{1}{2} - \frac{1}{\text{polylog}(d)}. \tag{4}$$

- *(Persistent Learning Failure on Task k) If the additional SNR condition holds $\frac{m^2-k^2}{R^{1/3}\sigma_0\sigma_\xi d} \lesssim \frac{\sum_{p=1}^{m}(1-\frac{p-1}{m})\alpha_p^3 A_{(p,k)}}{(\sigma_\xi\sqrt{d})^3} \lesssim \frac{1}{n}$, then the model still fails to correctly classify task k after subsequent training to task m:*

$$\mathbb{P}\left\{y_k F\left(\mathbf{W}^{(T_m)}, \mathbf{x}_k\right) < 0\right\} \geq \frac{1}{2} - \frac{1}{\text{polylog}(d)}. \tag{5}$$

- *(Enhanced Signal Learning on Task k) If the additional SNR conditions holds $\frac{\sum_{p=1}^{m}\alpha_p^3 A_{(p,k)}}{(\sigma_\xi\sqrt{d})^3} \gtrsim \frac{1}{nR^{1/3}\sigma_0\sigma_\xi\sqrt{d}}$, then the model can correctly classify task k after subsequent training to task m:*

$$\mathbb{P}\left\{y_k F\left(\mathbf{W}^{(T_m)}, \mathbf{x}_k\right) < 0\right\} \leq \frac{1}{\text{poly}(d)}. \tag{6}$$

Theorem 1 shows that if the cumulative signal from the first k tasks related to task k is not sufficiently strong, the model fails to correctly classify task k even immediately after learning it, as shown in eq. (4). This reflects poor generalization under low-SNR conditions and aligns with observations in standard (non-continual) learning settings (Cao et al., 2022b). Moreover, if the cumulative signal from the first m tasks remains weak with respect to task k, the model continues to misclassify task k, indicating a persistent failure to learn its features. However, if the cumulative signal from the first m tasks becomes sufficiently strong, the model can eventually classify task k correctly–potentially even better than immediately after learning it-highlighting that learning subsequent tasks can help transfer useful features and improve generalization on earlier tasks. In addition, noticed that when analyzing learning failure, the SNR condition involves not only an upper bound but also a lower bound. This lower bound arises from the need to control the magnitude of noise memorization—even if effective signal learning does not occur. The model must still control the magnitude of noise memorization to ensure stable training, a principle that also holds in standard (non-continual) training settings Cao et al. (2022b).

Prioritizing Higher-Signal Tasks Facilitates Learning of Task k. When evaluating the generalization performance for task k under the SNR conditions, it can be observed that the cumulative signal depends on three key components: the coefficient $(1 - \frac{p-1}{k})$, the signal intensity α_p^3, and the correlation strength $A_{(p,k)}$. The coefficient reflects that tasks appearing earlier (i.e., smaller p) contribute more heavily to the accumulation of signal relevant to task k. The term $\alpha_p^3 A_{(p,k)}$ quantifies how much task p contributes to the effective signal aligned with task k. Therefore, placing tasks with stronger signal intensity and higher alignment to task k earlier in the sequence may help prevent persistent learning failure on task k, by boosting the overall cumulative signal in its favor.

Theorem 2. *Suppose the setting in Condition 1 holds, and the SNR satisfies $\frac{\sum_{p=1}^{k}\alpha_p^3 A_{(p,k)}}{(\sigma_\xi\sqrt{d})^3} \gtrsim \frac{1+nk^2/\sqrt{d}}{kn}$. Consider full data-replay training with learning rate $\eta \in (0, \widetilde{O}(1)]$, and let $(\mathbf{x}_k, y_k) \sim \mathcal{D}_k$ be a test sample from the task k. Then, with high probability, there exist training times T_k and T_m $(m > k)$ such that*

- *The model can correctly classify task k immediately after learning it:*

$$\mathbb{P}\left\{y_k F\left(\mathbf{W}^{(T_k)}, \mathbf{x}_k\right) < 0\right\} \leq \frac{1}{\text{poly}(d)}. \tag{7}$$

- *(Catastrophic Forgetting on Task k) If the additional SNR conditions holds $\frac{m^2}{R^{2/3}\sigma_0^2\sigma_\xi^2 d^{13/6}} \lesssim \frac{\sum_{p=1}^{m}\alpha_p^3 A_{(p,k)}}{(\sigma_\xi\sqrt{d})^3} \lesssim \frac{\alpha_k R^{1/3}}{n}$, then it occurs Catastrophic Forgetting on task k after subsequent training to task m:*

$$\mathbb{P}\left\{y_k F\left(\mathbf{W}^{(T_m)}, \mathbf{x}_k\right) < 0\right\} \geq \frac{1}{2} - \frac{1}{\text{polylog}(d)}. \tag{8}$$

- *(Continual Learning on Task k) If the additional SNR conditions holds $\frac{\sum_{p=1}^{m}\alpha_p^3 A_{(p,k)}}{(\sigma_\xi\sqrt{d})^3} \gtrsim \frac{\alpha_k R^{1/3}\sigma_0\left((1-\frac{k-1}{m})+nm/\sqrt{d}\right)}{n}$, then the model can still correctly classify task k after subsequent training to task m:*

$$\mathbb{P}\left\{y_k F\left(\mathbf{W}^{(T_m)}, \mathbf{x}_k\right) < 0\right\} \leq \frac{1}{\text{poly}(d)}. \tag{9}$$

In contrast to Theorem 1, Theorem 2 considers the case where the model successfully learns task k after training on it, due to a sufficiently strong cumulative signal from the first k tasks, as shown in eq. (7). This success may be maintained throughout continual learning if subsequent tasks continue to contribute meaningful signal toward task k (see eq. (9)). However, if the cumulative signal from later tasks is insufficient or misaligned, the model may still experience forgetting of task k despite its initial success—resulting in *catastrophic forgetting* (refer to eq. (8)).

Prioritizing Higher-Signal Tasks Mitigates Forgetting of Task k. Similar to Theorem 1, task ordering and signal intensity also play crucial roles in

the subsequent learning and retention of task k. For instance, when evaluation occurs shortly after training task k (i.e., when $m > k$ is close to k), a smaller amount of cumulative signal is required to satisfy the relaxed SNR condition in eq. (6). Furthermore, placing tasks with stronger signal intensity and higher alignment to task k between tasks k and m increases the cumulative signal, making it more likely to meet the continual learning condition and prevent *catastrophic forgetting*.

Comparison with Existing Work. Existing work shows that task ordering affects forgetting behavior from both empirical (Lesort et al., 2022; Hemati et al., 2025; Li and Hiratani, 2025) and analytical perspectives (Evron et al., 2022; Swartworth et al., 2023; Lin et al., 2023; Ding et al., 2024; Evron et al., 2025; Li and Hiratani, 2025). Specifically, Evron et al. (2022) demonstrate that forgetting diminishes over time when task ordering is cyclic or random. Swartworth et al. (2023) and Evron et al. (2025) provide tighter forgetting bounds for cyclic and random orderings, respectively. Lin et al. (2023), Ding et al. (2024), and Li and Hiratani (2025) show that forgetting can be influenced by the arrangement of task orderings based on task similarity. Our work shares similar insights but offers a novel feature-signal perspective: prioritizing higher-signal tasks not only facilitates the learning of lower-signal tasks but also mitigates forgetting. Moreover, prior analyses mainly focus on linear regression models, two-task settings, and naive sequential training, whereas our approach is developed under a more general two-layer neural network model and a more challenging data replay training setup, making it more relevant to realistic continual learning scenarios.

5 DATA REPLAY WITH M TASKS

In this section, we provide a proof sketch of the theoretical results introduced earlier. Our analysis focuses on understanding when and how a model trained via full data replay can either memorize noise or successfully learn meaningful features across multiple tasks. Before diving into the technical lemmas, we first establish the following notation:

- The signal learning of task k's feature at time t under task m: $\Gamma_{(m,r)}^{(t,k)} := \langle \mathbf{w}_{(m,r)}^{(t)}, \mathbf{v}_k^* \rangle$.
- The noise memorization of sample j from task k at time t under task m: $\Phi_{(m,r)}^{(t,k,j)} := \langle \mathbf{w}_{(m,r)}^{(t)}, \boldsymbol{\xi}_k^j \rangle$.

In section 6, we will illustrate the dynamics of signal learning and noise memorization during the continual training process under full data-replay.

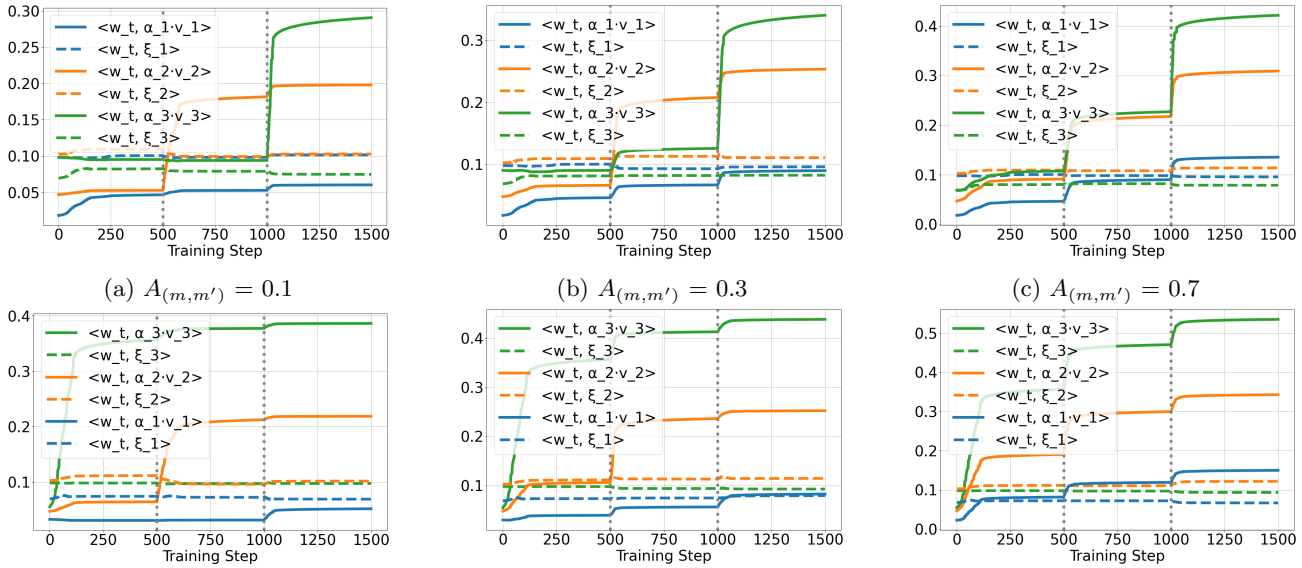
Lemma 1 (Continual Noise Memorization)**.** *Suppose the SNR condition satisfying* $\frac{k^2}{R^{2/3}\sigma_0^2\sigma_\xi^2 d^{13/6}} \lesssim \frac{\sum_{p=1}^k (1-\frac{p-1}{k})\alpha_p^3 A_{(p,k)}}{(\sigma_\xi\sqrt{d})^3} \lesssim \frac{1}{n}$*, and there exists an iteration* $\tau_{kj}^k \le \mathrm{T}_\xi^k = \mathrm{T}_\xi^- + O(\log(d))$ *such that* τ_{kj}^k *is the first iteration for which* $\max_{r\in[R]}(y_{kj}\Phi_{(k,r)}^{(t,k,j)}) \ge \Theta(R^{-\frac{1}{3}})$*, and for any* $t \le \mathrm{T}_\xi^k$ *it holds that* $\max_{r\in[R]} |\Gamma_{(k,r)}^{(t,k)}| \le \widetilde{O}(\sigma_0)$*. Then, if the additional SNR condition* $\frac{m^2-k^2}{R^{1/3}\sigma_0\sigma_\xi d} \lesssim \frac{\sum_{p=1}^m (1-\frac{p-1}{m})\alpha_p^3 A_{(p,k)}}{(\sigma_\xi\sqrt{d})^3} \lesssim \frac{1}{n}$ *also holds, there exists an iteration* τ_{kj}^m *such that* τ_{kj}^m *is the first iteration satisfying* $\max_{r\in[R]}(y_{kj}\Phi_{(m,r)}^{(t,k,j)}) \ge \Theta(R^{-\frac{1}{4}})$*. In this case, we can also guarantee that* $\max_{r\in[R]} |\Gamma_{(m,r)}^{(t,k)}| \le \widetilde{O}(\sigma_0)$ *for any* $t \le \mathrm{T}_\xi^m$*.*

Lemma 1 shows that the signal alignment for task k remains bounded by $\widetilde{O}(\sigma_0)$, indicating that the model fails to learn sufficient features of task k even by the end of its training. Instead, noise memorization dominates the learning process with a lower bound by $\Theta(R^{-\frac{1}{3}})$. This issue persists through subsequent training up to task m, suggesting that when the cumulative signal contribution from the first m tasks is insufficient, the model consistently fails to learn task k. As a result, task k suffers from continual learning failure and poor performance.

Lemma 2 (Enhanced Signal Learning)**.** *Suppose the SNR satisfying* $\frac{k^2}{R^{2/3}\sigma_0^2\sigma_\xi^2 d^{13/6}} \lesssim \frac{\sum_{p=1}^k (1-\frac{p-1}{k})\alpha_p^3 A_{(p,k)}}{(\sigma_\xi\sqrt{d})^3} \lesssim \frac{1}{n}$*, and there exists an iteration* $\tau_{kj}^k \le \mathrm{T}_\xi^k = \mathrm{T}_\xi^- + O(\log(d))$ *such that* τ_{kj}^k *is the first iteration where* $\max_{r\in[R]}(y_{kj}\Phi_{(k,r)}^{(t,k,j)}) \ge \Theta(R^{-\frac{1}{3}})$*, and for any* $t \le \mathrm{T}_\xi^k$ *it holds that* $\max_{r\in[R]} |\Gamma_{(k,r)}^{(t,k)}| \le \widetilde{O}(\sigma_0)$*. Then, if the additional SNR condition* $\frac{\sum_{p=1}^m \alpha_p^3 A_{(p,k)}}{(\sigma_\xi\sqrt{d})^3} \gtrsim \frac{1}{nR^{1/3}\sigma_0\sigma_\xi\sqrt{d}}$ *also holds, there exists* $\tau_{kv}^m \le \mathrm{T}_v^m = \mathrm{T}_v^k + O(\log(d))$ *such that* τ_{kv}^m *be the first iteration satisfying* $\max_{r\in[R]} |\Gamma_{(m,r)}^{(t,k)}| \ge \Theta(\frac{1}{\alpha_k R^{1/5}})$*.*

Similar to Lemma 1, Lemma 2 shows that the model fails to learn task k 's feature signal during its own training phase. However, in this case, tasks in later stages $p \in (k, m]$ possess strong alignment with task k, contributing sufficient signal to compensate for the earlier deficiency. This cumulative reinforcement enables the model to gradually build up the correct representation of task k, and by time T_v^m, it can successfully classify samples from task k 's distribution.

Lemma 3 (Amplified Noise Memorization)**.** *Suppose the SNR satisfying* $\frac{\sum_{p=1}^k \alpha_p^3 A_{(p,k)}}{(\sigma_\xi\sqrt{d})^3} \gtrsim \frac{1+nk^2/\sqrt{d}}{kn}$*, and there exists an iteration* $\tau_{kv}^k \le \mathrm{T}_v^k = \mathrm{T}_v^- + O(\log(d))$ *such that* τ_{kv}^k *is the first iteration where* $\max_{r\in[R]} |\Gamma_{(k,r)}^{(t,k)}| \ge \Theta(\frac{1}{\alpha_k R^{1/3}})$*, and for any* $t \le$



(a) $A_{(m,m')} = 0.1$ (b) $A_{(m,m')} = 0.3$ (c) $A_{(m,m')} = 0.7$

(d) $A_{(m,m')} = 0.1$ (e) $A_{(m,m')} = 0.3$ (f) $A_{(m,m')} = 0.7$

Figure 2: Dynamics of signal learning and noise memorization during full data-replay continual training across different task orderings and correlation strengths.

T_v^k *it holds that* $\max_{r\in[R]} |\Phi_{(k,r)}^{(t,k,j)}| \leq \widetilde{O}(\sigma_0\sigma_\xi\sqrt{d})$. *Then, if the additional SNR condition* $\frac{m^2}{R^{2/3}\sigma_0^2\sigma_\xi^2 d^{13/6}} \lesssim \frac{\sum_{p=1}^m \alpha_p^3 A_{(p,k)}}{(\sigma_\xi\sqrt{d})^3} \lesssim \frac{\alpha_k R^{1/3}}{n}$ *also holds, there exists* $\tau_{kj}^m \leq \mathrm{T}_\xi^m = \mathrm{T}_\xi^k + O(\log(d))$ *such that* τ_{kj}^m *be the first iteration satisfying* $\max_{r\in[R]}(y_{kj}\Phi_{(m,r)}^{(t,k,j)}) \geq \Theta(R^{-\frac{1}{5}})$.

In contrast to Lemmas 1 and 2, Lemma 3 presents a case where the model initially succeeds in learning the feature of task k. However, this learned signal is not preserved-subsequent training phases are dominated by noise memorization, and the cumulative signal contribution from tasks k to m is insufficient to maintain the representation. As a result, the model gradually forgets task k, leading to catastrophic forgetting as characterized in Theorem 2.

Lemma 4 (Continual Signal Learning)**.** *Suppose the SNR satisfying* $\frac{\sum_{p=1}^k \alpha_p^3 A_{(p,k)}}{(\sigma_\xi\sqrt{d})^3} \gtrsim \frac{1+nk^2/\sqrt{d}}{kn}$, *and there exists an iteration* $\tau_{kv}^k \leq \mathrm{T}_v^k = \mathrm{T}_v^- + O(\log(d))$ *such that* τ_{kv}^k *is the first iteration where* $\max_{r\in[R]} |\Gamma_{(k,r)}^{(t,k)}| \geq \Theta(\frac{1}{\alpha_k R^{1/3}})$, *and for any* $t \leq \mathrm{T}_v^k$ *it holds that* $\max_{r\in[R]} |\Phi_{(k,r)}^{(t,k,j)}| \leq \widetilde{O}(\sigma_0\sigma_\xi\sqrt{d})$. *Then, if the additional SNR condition* $\frac{\sum_{p=1}^m \alpha_p^3 A_{(p,k)}}{(\sigma_\xi\sqrt{d})^3} \gtrsim \frac{\alpha_k R^{1/3}\sigma_0\left((1-\frac{k-1}{m})+nm/\sqrt{d}\right)}{n}$ *also holds, there exists* $\tau_{kv}^m \leq \mathrm{T}_v^m = \mathrm{T}_v^k + O(\log(d))$ *such that* τ_{kv}^m *be the first iteration satisfying* $\max_{r\in[R]} |\Gamma_{(m,r)}^{(t,k)}| \geq \Theta(\frac{1}{\alpha_k R^{1/5}})$.

To achieve successful continual learning of task k, the model must consistently prioritize signal learning over noise memorization-not only during the training of task k but also throughout subsequent tasks up to task m. Lemma 4 formalizes this by showing that the signal intensity aligned with task k must remain above a certain threshold, while noise memorization must be kept under control. This balance ensures that the feature of task k is both learned and retained over time.

6 EXPERIMENT

In this section, we present synthetic experimental results to support our theoretical findings. Additional results are provided in the Appendix due to space limitations.

Experimental Setup. We design a synthetic continual learning experiment using a two-layer neural network with cubic activation. The model takes an input of dimension $2d$ (with $d = 1000$) and projects it to a hidden layer of size $R = 10$. The network is trained to solve three binary classification tasks sequentially, each associated with a distinct signal sampled from a multivariate Gaussian with varying correlation levels (off-diagonal entries set to 0.1, 0.3, and 0.7 to represent low, medium, and high correlation). For each task k, the input is generated from definition 1, comprising signal and noise components. The signal strength α_k is scaled based on a task-specific SNR (set to $[0.1, 0.2, 0.3]$), and the noise is drawn from a distribution orthogonal to all signal directions, with fixed deviation $\sigma_\xi = 0.1$. Training is performed using SGD with a fixed learning rate $\eta = 0.1$ and Gaussian initialization ($\sigma_0 = 0.1$). Each task is trained for 50 epochs

with 10 samples. To assess learning dynamics, we track the alignment between hidden weights and both signal and noise across tasks. Notably, the dynamics of signal learning and noise memorization are closely consistent with accuracy performance—stronger signal learning generally corresponds to higher accuracy. Due to space limitations, we present the detailed accuracy figures in the Appendix.

Prioritizing Higher-Signal Tasks May Enhance Lower-Signal Tasks Learning. Figures 2 shows the dynamics of signal learning and noise memorization during continual training under full data replay, comparing different task orderings across varying levels of task correlation. In Figures 2a-2c, Task 3—which has the highest signal intensity (corresponding to the highest SNR = 0.3 under fixed noise scale)—is placed earlier in the task sequence. In contrast, Figures 2d-2f reverse the task order, placing lower-SNR tasks earlier. When the correlation strength is low ($A_{(m,m')} = 0.1$, implying near-orthogonality between task vectors and low task similarity), prioritizing the high-signal Task 3 has limited effect: the cumulative signal for the lower-signal Task 1 remains insufficient in both orderings (see Figures 2a and 2d). However, as correlation strength increases, the effect of task ordering becomes more pronounced. For instance, in the moderate correlation setting (Figures 2b and 2e), prioritizing Task 3 improves signal acquisition for the other tasks—Task 2 achieves higher signal learning in the ordered setting. Furthermore, in Figure 2e, the signal learning of Task 1 eventually exceeds its noise memorization, while in the non-prioritized setting (Figure 2b), Task 1 continues to struggle. This effect becomes even more evident under high correlation ($A_{(m,m')} = 0.7$), where prioritizing high-signal tasks yields better signal learning for lower-SNR tasks, as shown in Figure 2f. These empirical observations also validate our theoretical conclusions in Theorem 1 and 2.

Higher Correlation Enhances Signal Learning. Figures 2a, 2b, and 2c (and their reordered counterparts) illustrate that increasing the correlation between tasks significantly improves signal learning across the board. When the correlation strength is low ($A_{(m,m')} = 0.1$), tasks contribute little to one another, resulting in limited signal accumulation for earlier, lower-SNR tasks—regardless of ordering. However, as the correlation increases to 0.3 and 0.7, tasks—especially those with stronger signals—can contribute more effectively to the overall feature representation, improving the learning of other tasks in the sequence. For example, under the high-correlation setting ($A_{(m,m')} = 0.7$), even lower-signal tasks (e.g., Task 1) can accumulate sufficient signal to surpass noise memorization, demonstrating that strong task correlation amplifies the benefits of both task ordering and feature sharing in continual learning.

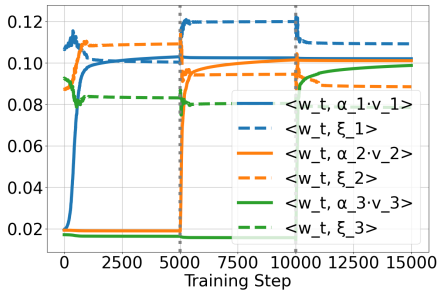


Figure 3: Dynamics of signal learning and noise memorization under lower SNR.

Competition between Noise Memorization and Signal Learning. In Figure 2, it is clear that noise memorization remains relatively stable, which may be attributed to the model focusing more on signal learning during training. To further investigate the behavior of noise memorization, we increase the sample size to 100, reduce the signal intensity to 0.06 for all tasks, and set the correlation strength to 0.01 to simulate a low-correlation regime. As shown in Figure 3, Task 1 performs well during its initial training phase, as the signal learning surpasses noise memorization. However, as new tasks are introduced—each weakly correlated with Task 1—the model fails to reinforce Task 1's features, ultimately leading to catastrophic forgetting of Task 1. We further explore the impact of correlation by increasing the correlation strength to 0.3 and 0.7. As expected, higher correlation allows the model to benefit from the features learned in Tasks 2 and 3, effectively contributing to Task 1's signal and mitigating forgetting. These results demonstrate that *catastrophic forgetting tends to occur when tasks are orthogonal*, consistent with Theorem 2, where the SNR conditions fail to hold due to near-zero correlation $A_{(m,m')} = 0$. Due to space limitations, the corresponding figures are deferred to the Appendix.

7 CONCLUSION

In this work, we provide a comprehensive theoretical framework for understanding full data-replay training in continual learning through the lens of feature learning. By adopting a multi-view data model, task-specific signal structures and inter-task correlations, we identify the SNR as a fundamental factor driving forgetting. A particularly novel insight from our study is the impact of task ordering—prioritizing higher-signal tasks not only improves learning for subsequent tasks but also mitigates forgetting of earlier ones. This highlights the need for order-aware replay strategies in the design of continual learning systems.

ACKNOWLEDGMENT

We thank the AISTATS reviewers and community for their valuable suggestions, which motivated us to conduct and include additional empirical verification on real-world CIFAR-100 data in Appendix. The research of Jinhui Xu was partially supported by startup funds from USTC and a grant from IAI.

References

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pages 139–154, 2018.

Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020.

Zeyuan Allen-Zhu and Yuanzhi Li. Feature purification: How adversarial training performs robust deep learning. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 977–988. IEEE, 2022.

Amin Banayeeanzade, Mahdi Soltanolkotabi, and Mohammad Rostami. Theoretical insights into overparameterized models in multi-task and replay-based continual learning. *arXiv preprint arXiv:2408.16939*, 2024.

Yajie Bao, Michael Crawshaw, and Mingrui Liu. Provable benefits of local steps in heterogeneous federated learning for neural networks: A feature learning perspective. In *Forty-first International Conference on Machine Learning*.

Frederik Benzing. Unifying importance based regularisation methods for continual learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2372–2396. PMLR, 2022.

Dake Bu, Wei Huang, Andi Han, Atsushi Nitanda, Taiji Suzuki, Qingfu Zhang, and Hau-San Wong. Provably transformers harness multi-concept word semantics for efficient in-context learning. *Advances in Neural Information Processing Systems*, 37:63342–63405, 2024.

Dake Bu, Wei Huang, Andi Han, Atsushi Nitanda, Qingfu Zhang, Hau-San Wong, and Taiji Suzuki. Provable in-context vector arithmetic via retrieving task concepts. In *Forty-second International Conference on Machine Learning*, 2025.

Xinyuan Cao, Weiyang Liu, and Santosh Vempala. Provable lifelong learning of representations. In *International Conference on Artificial Intelligence and Statistics*, pages 6334–6356. PMLR, 2022a.

Yuan Cao, Zixiang Chen, Misha Belkin, and Quanquan Gu. Benign overfitting in two-layer convolutional neural networks. *Advances in neural information processing systems*, 35:25237–25250, 2022b.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*, 2018.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.

Yulai Cong, Miaoyun Zhao, Jianqiao Li, Sijia Wang, and Lawrence Carin. Gan memory with no forgetting. *Advances in neural information processing systems*, 33:16481–16494, 2020.

Meng Ding, Kaiyi Ji, Di Wang, and Jinhui Xu. Understanding forgetting in continual learning with linear regression. In *Forty-first International Conference on Machine Learning*, 2024.

Meng Ding, Mingxi Lei, Shaopeng Fu, Shaowei Wang, Di Wang, and Jinhui Xu. Understanding private learning from feature perspective. *arXiv preprint arXiv:2511.18006*, 2025.

Thang Doan, Mehdi Abbana Bennani, Bogdan Mazoure, Guillaume Rabusseau, and Pierre Alquier. A theoretical analysis of catastrophic forgetting through the ntk overlap matrix. In *International Conference on Artificial Intelligence and Statistics*, pages 1072–1080. PMLR, 2021.

Arthur Douillard, Alexandre Ramé, Guillaume Couairon, and Matthieu Cord. Dytox: Transformers for continual learning with dynamic token expansion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9285–9295, 2022.

Sayna Ebrahimi, Suzanne Petryk, Akash Gokul, William Gan, Joseph E Gonzalez, Marcus Rohrbach, and Trevor Darrell. Remembering for the right reasons: Explanations reduce catastrophic forgetting. *Applied AI letters*, 2(4):e44, 2021.

Itay Evron, Edward Moroshko, Rachel Ward, Nathan Srebro, and Daniel Soudry. How catastrophic can catastrophic forgetting be in linear regression? In *Conference on Learning Theory*, pages 4028–4079. PMLR, 2022.

Itay Evron, Ran Levinstein, Matan Schliserman, Uri Sherman, Tomer Koren, Daniel Soudry, and Nathan Srebro. Better rates for random task orderings in continual linear models. *arXiv preprint arXiv:2504.04579*, 2025.

Daniel Goldfarb and Paul Hand. Analysis of catastrophic forgetting for random orthogonal transformation tasks in the overparameterized regime. In *International Conference on Artificial Intelligence and Statistics*, pages 2975–2993. PMLR, 2023.

Mustafa Burak Gurbuz and Constantine Dovrolis. Nispa: Neuro-inspired stability-plasticity adaptation for continual learning in sparse networks. *arXiv preprint arXiv:2206.09117*, 2022.

Andi Han, Wei Huang, Yuan Cao, and Difan Zou. On the feature learning in diffusion models. *arXiv preprint arXiv:2412.01021*, 2024.

Andi Han, Wei Huang, Zhanpeng Zhou, Gang Niu, Wuyang Chen, Junchi Yan, Akiko Takeda, and Taiji Suzuki. On the role of label noise in the feature learning process. *arXiv preprint arXiv:2505.18909*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Hamed Hemati, Lorenzo Pellegrini, Xiaotian Duan, Zixuan Zhao, Fangfang Xia, Marc Masana, Benedikt Tscheschner, Eduardo Veas, Yuxiang Zheng, Shiji Zhao, et al. Continual learning in the presence of repetition. *Neural Networks*, 183:106920, 2025.

Wei Huang, Yuan Cao, Haonan Wang, Xin Cao, and Taiji Suzuki. Graph neural networks provably benefit from structural information: A feature learning perspective. *arXiv preprint arXiv:2306.13926*, 2023a.

Wei Huang, Ye Shi, Zhongyi Cai, and Taiji Suzuki. Understanding convergence and generalization in federated learning through feature learning theory. In *The Twelfth International Conference on Learning Representations*, 2023b.

Samy Jelassi and Yuanzhi Li. Towards understanding how momentum improves generalization in deep learning. In *International Conference on Machine Learning*, pages 9965–10040. PMLR, 2022.

Samy Jelassi, Michael Sander, and Yuanzhi Li. Vision transformers provably learn spatial structure. *Advances in Neural Information Processing Systems*, 35:37822–37836, 2022.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526, 2017.

Tomasz Korbak, Hady Elsahar, German Kruszewski, and Marc Dymetman. Controlling conditional language models without catastrophic forgetting. In *International Conference on Machine Learning*, pages 11499–11528. PMLR, 2022.

Yiwen Kou, Zixiang Chen, Yuanzhou Chen, and Quanquan Gu. Benign overfitting in two-layer relu convolutional neural networks. In *International Conference on Machine Learning*, pages 17615–17659. PMLR, 2023.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Lilly Kumari, Shengjie Wang, Tianyi Zhou, and Jeff A Bilmes. Retrospective adversarial replay for continual learning. *Advances in neural information processing systems*, 35:28530–28544, 2022.

Minh Le, Huy Nguyen, Trang Nguyen, Trang Pham, Linh Ngo, Nhat Ho, et al. Mixture of experts meets prompt-based continual learning. *Advances in Neural Information Processing Systems*, 37:119025–119062, 2024.

Timothée Lesort, Oleksiy Ostapenko, Diganta Misra, Md Rifat Arefin, Pau Rodríguez, Laurent Charlin, and Irina Rish. Challenging common assumptions about catastrophic forgetting. *arXiv preprint arXiv:2207.04543*, 2022.

Bingrui Li, Wei Huang, Andi Han, Zhanpeng Zhou, Taiji Suzuki, Jun Zhu, and Jianfei Chen. On the optimization and generalization of two-layer transformers with sign gradient descent. *arXiv preprint arXiv:2410.04870*, 2024a.

Hongbo Li, Sen Lin, Lingjie Duan, Yingbin Liang, and Ness B Shroff. Theory on mixture-of-experts in continual learning. *arXiv preprint arXiv:2406.16437*, 2024b.

Hongkang Li, Meng Wang, Sijia Liu, and Pin-Yu Chen. A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity. *arXiv preprint arXiv:2302.06015*, 2023.

Ziyan Li and Naoki Hiratani. Optimal task order for continual learning of multiple tasks. *arXiv preprint arXiv:2502.03350*, 2025.

Guoliang Lin, Hanlu Chu, and Hanjiang Lai. Towards better plasticity-stability trade-off in incremental learning: A simple linear connector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 89–98, 2022.

Sen Lin, Peizhong Ju, Yingbin Liang, and Ness Shroff. Theory on forgetting and generalization of continual learning. In *International Conference on Machine Learning*, pages 21078–21100. PMLR, 2023.

Xialei Liu, Chenshen Wu, Mikel Menta, Luis Herranz, Bogdan Raducanu, Andrew D Bagdanov, Shangling Jui, and Joost van de Weijer. Generative feature replay for class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 226–227, 2020.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

Mark D McDonnell, Dong Gong, Amin Parvaneh, Ehsan Abbasnejad, and Anton Van den Hengel. Ranpac: Random projections and pre-trained models for continual learning. *Advances in Neural Information Processing Systems*, 36:12022–12053, 2023.

Zichen Miao, Ze Wang, Wei Chen, and Qiang Qiu. Continual learning with filter atom swapping. In *International Conference on Learning Representations*, 2021.

Cuong V Nguyen, Yingzhen Li, Thang D Bui, and Richard E Turner. Variational continual learning. *arXiv preprint arXiv:1710.10628*, 2017.

Oleksiy Ostapenko, Mihai Puscas, Tassilo Klein, Patrick Jahnichen, and Moin Nabi. Learning to remember: A synaptic plasticity driven framework for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11321–11329, 2019.

Oleksiy Ostapenko, Pau Rodriguez, Massimo Caccia, and Laurent Charlin. Continual learning via local module composition. *Advances in Neural Information Processing Systems*, 34:30298–30312, 2021.

Pingbo Pan, Siddharth Swaroop, Alexander Immer, Runa Eschenhagen, Richard Turner, and Mohammad Emtiyaz E Khan. Continual deep learning by functional regularisation of memorable past. *Advances in neural information processing systems*, 33: 4453–4464, 2020.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural networks*, 113:54–71, 2019.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. *arXiv preprint arXiv:1810.11910*, 2018.

Hippolyt Ritter, Aleksandar Botev, and David Barber. Online structured laplace approximations for overcoming catastrophic forgetting. *Advances in Neural Information Processing Systems*, 31, 2018.

Yujun Shi, Kuangqi Zhou, Jian Liang, Zihang Jiang, Jiashi Feng, Philip HS Torr, Song Bai, and Vincent YF Tan. Mimicking the oracle: An initial phase decorrelation approach for class incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16722–16731, 2022.

Dongsub Shim, Zheda Mai, Jihwan Jeong, Scott Sanner, Hyunwoo Kim, and Jongseong Jang. Online class-incremental continual learning with adversarial shapley value. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9630–9638, 2021.

William Swartworth, Deanna Needell, Rachel Ward, Mark Kong, and Halyun Jeong. Nearly optimal bounds for cyclic forgetting. *Advances in Neural Information Processing Systems*, 36:68197–68206, 2023.

Shixiang Tang, Dapeng Chen, Jinguo Zhu, Shijie Yu, and Wanli Ouyang. Layerwise optimization by gradient decomposition for continual learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 9634–9643, 2021.

Michalis K Titsias, Jonathan Schwarz, Alexander G de G Matthews, Razvan Pascanu, and Yee Whye Teh. Functional regularisation for continual learning with gaussian processes. *arXiv preprint arXiv:1901.11356*, 2019.

Rishabh Tiwari, Krishnateja Killamsetty, Rishabh Iyer, and Pradeep Shenoy. Gcr: Gradient coreset based replay buffer selection for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 99–108, 2022.

Gido M Van de Ven, Hava T Siegelmann, and Andreas S Tolias. Brain-inspired replay for continual learning with artificial neural networks. *Nature communications*, 11(1):4069, 2020.

Liyuan Wang, Bo Lei, Qian Li, Hang Su, Jun Zhu, and Yi Zhong. Triple-memory networks: A brain-inspired method for continual learning. *IEEE Transactions on Neural Networks and Learning Systems*, 33(5):1925–1934, 2021.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Runqi Wang, Yuxiang Bao, Baochang Zhang, Jianzhuang Liu, Wentao Zhu, and Guodong Guo. Anti-retroactive interference for lifelong learning. In *European Conference on Computer Vision*, pages 163–178. Springer, 2022a.

Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning. *Advances in Neural Information Processing Systems*, 35:5682–5695, 2022b.

Zixin Wen and Yuanzhi Li. Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference on Machine Learning*, pages 11112–11122. PMLR, 2021.

Tz-Ying Wu, Gurumurthy Swaminathan, Zhizhong Li, Avinash Ravichandran, Nuno Vasconcelos, Rahul Bhotika, and Stefano Soatto. Class-incremental learning with strong pre-trained models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9601–9610, 2022.

Jaehong Yoon, Divyam Madaan, Eunho Yang, and Sung Ju Hwang. Online coreset selection for rehearsal-based continual learning. *arXiv preprint arXiv:2106.01085*, 2021.

Xuyang Zhao, Huiyuan Wang, Weiran Huang, and Wei Lin. A statistical theory of regularization-based continual learning. *arXiv preprint arXiv:2406.06213*, 2024.

Bowen Zheng, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Multi-layer rehearsal feature augmentation for class-incremental learning. In *Forty-first International Conference on Machine Learning*, 2024.

Guodong Zheng, Peng Wang, and Li Shen. Towards understanding memory buffer based continual learning.

Difan Zou, Yuan Cao, Yuanzhi Li, and Quanquan Gu. The benefits of mixup for feature learning. In *International Conference on Machine Learning*, pages 43423–43479. PMLR, 2023.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board

(IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A Additional Related Work

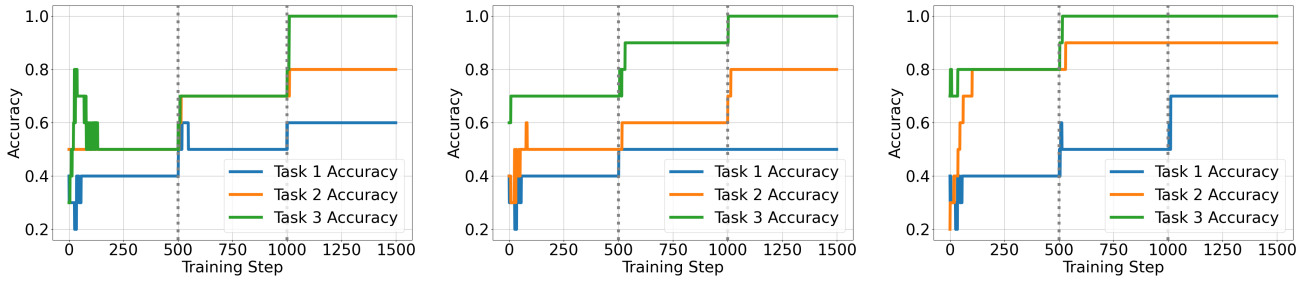
Feature Learning Theory Allen-Zhu and Li (2022) first introduced the feature learning framework to explain the benefits of adversarial training in robust learning. This was further extended by Allen-Zhu and Li (2020), who incorporated a multi-view data structure to show how ensemble methods can enhance generalization. Since then, feature learning has been studied across a range of model architectures, including graph neural networks Huang et al. (2023a), convolutional neural networks Cao et al. (2022b); Kou et al. (2023), vision transformers Jelassi et al. (2022); Li et al. (2023), and diffusion models Han et al. (2024). Beyond model architectures, the framework has also been used to analyze the behavior of optimization algorithms and training techniques—such as Adam Zou et al. (2023), momentum Jelassi and Li (2022), and Mixup Zou et al. (2023). Furthermore, feature learning provides new insights into broader learning paradigms, including federated learning Huang et al. (2023b); Bao et al., contrastive learning Wen and Li (2021), and in-context learning Bu et al. (2024). To the best of our knowledge, this work is the first to investigate the effects of data replay in continual learning from the perspective of feature learning. Compared to standard learning settings, continual learning introduces additional challenges—such as task-specific feature vectors, and complex interactions between signal and noise across sequential tasks—which make theoretical analysis significantly more intricate.

B Additional Experimental

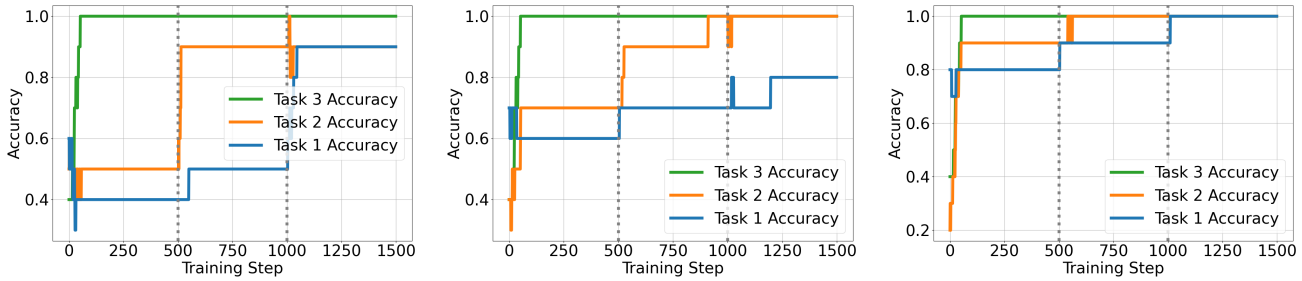
B.1 Synthetic Data

Accuracy Reflects Learning Dynamics. Figure 4 highlights how both task ordering and inter-task similarity influence model accuracy during continual learning, with trends that align closely with the signal and noise dynamics presented in Figure 2. When the task with the strongest signal (i.e., highest α_k) is placed earlier in the sequence—such as Task 3 in subplots (4d–4f)—the model is better able to acquire meaningful representations, resulting in higher accuracy even for subsequent lower-signal tasks. In contrast, when lower-signal tasks are prioritized (subplots 4a–4c), signal learning for those tasks becomes less effective, and overall accuracy suffers. Specifically, when the alignment with task-specific signal directions dominates over noise components , task accuracy exceeds 50%. Conversely, when noise memorization exceeds signal learning, accuracy deteriorates to near-random levels. For instance, under low task correlation ($A_{(m,m')} = 0.1$), Task 1 performs poorly when it appears last in the training sequence (Figure 4a), but its performance significantly improves when prioritized earlier (Figure 4d), confirming that task ordering matters. Additionally, across all orderings, stronger inter-task correlations (e.g., $A_{(m,m')} = 0.7$) facilitate signal transfer across tasks, allowing lower-signal tasks to benefit from earlier learned features. These patterns underscore the consistency between accuracy outcomes and the learning dynamics: accuracy increases when signal learning outweighs noise memorization, and fails when the noise dominates the representation.

Catastrophic Forgetting Occurs with Lower Task Similarity. Figure 5 investigates catastrophic forgetting under full data-replay continual learning by varying the inter-task correlation $A_{(m,m')}$. When the correlation is extremely low or near zero (e.g., $A_{(m,m')} = 0.01$), the tasks are nearly orthogonal—meaning their signal directions share no meaningful relationship. In this regime, newly introduced tasks overwrite earlier ones, and previously learned signal components decay, resulting in forgetting. As the correlation increases to 0.1, tasks begin to share overlapping features, which helps stabilize the representations and retain earlier task knowledge over time. These results highlight that task similarity, measured through correlation, is critical for mitigating forgetting: when tasks are orthogonal (i.e., $A \approx 0$), they compete destructively during training, whereas higher similarity allows for constructive feature reuse and knowledge retention.

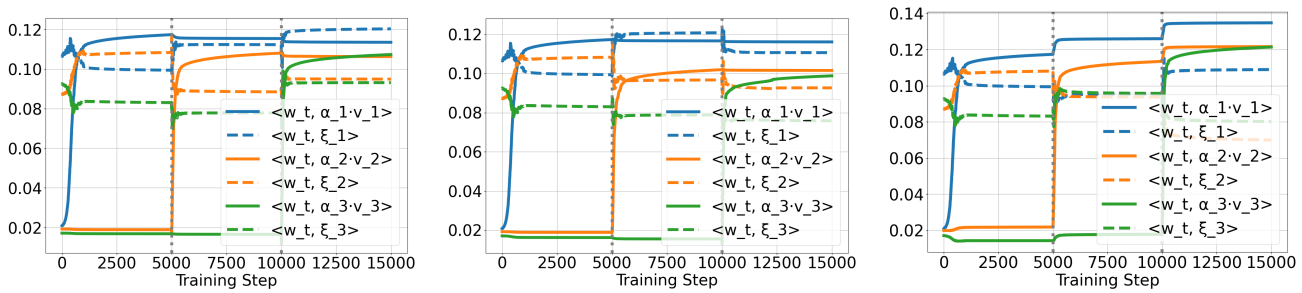


(a) $A_{(m,m')} = 0.1$ (b) $A_{(m,m')} = 0.3$ (c) $A_{(m,m')} = 0.7$

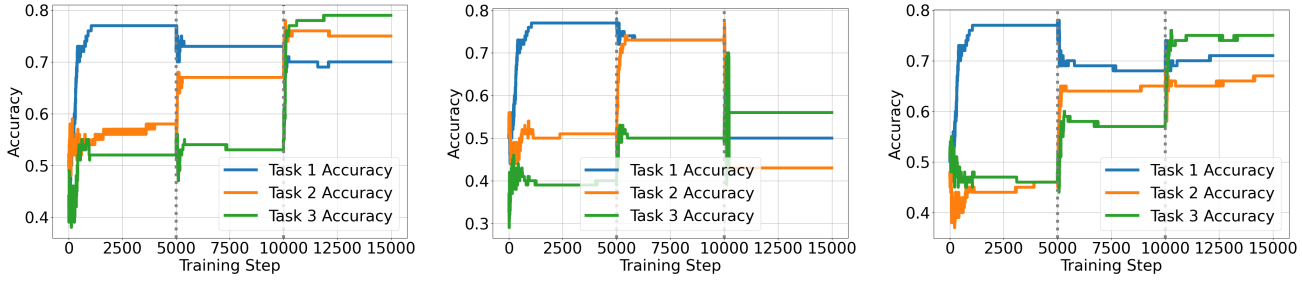


(d) $A_{(m,m')} = 0.1$ (e) $A_{(m,m')} = 0.3$ (f) $A_{(m,m')} = 0.7$

Figure 4: Accuracy under full data-replay continual training across different task orderings and correlation strengths.



(a) $A_{(m,m')} = 0$ (b) $A_{(m,m')} = 0.01$ (c) $A_{(m,m')} = 0.1$



(d) $A_{(m,m')} = 0$ (e) $A_{(m,m')} = 0.01$ (f) $A_{(m,m')} = 0.1$

Figure 5: Catastrophic forgetting under full data-replay continual training across various correlation strengths.

B.2 Empirical Verification on Real-World Data

To address the limitations of synthetic data and shallow networks, and to further validate our theoretical findings in a realistic deep learning scenario, we conduct experiments using the CIFAR-100 benchmark Krizhevsky et al. (2009) with a ResNet-18 architecture He et al. (2016).

Crucially, to ensure a rigorous alignment with our theoretical framework—which analyzes task-incremental binary classification (see Definition 1 and Section 4), we adapt the CIFAR-100 tasks into binary classification problems

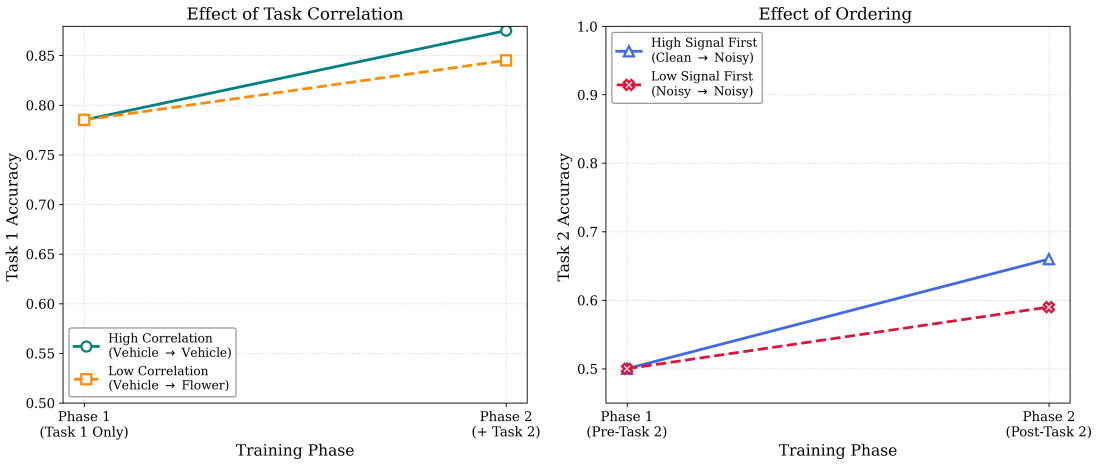


Figure 6: Empirical Verification of Correlation and Ordering Effects on CIFAR-100 (ResNet-18)

(e.g., "Class A vs. Rest"). This setup allows us to strictly verify the impact of signal-to-noise ratio (SNR) and task correlation ($A_{(m,m')}$) on feature learning and forgetting.

Experimental Setup. We construct binary tasks from CIFAR-100 superclasses. For a target class C (e.g., *Bicycle*), positive samples are drawn from C, and negative samples are randomly sampled from disjoint classes to create a balanced binary dataset.

- **Model:** We employ a ResNet-18 backbone. To isolate feature transfer from classifier interference, we utilize a multi-head architecture where the backbone is shared across tasks, but each task possesses an independent binary linear classifier.
- **Training:** Consistent with our theoretical premise, we employ *Full Data Replay*. When training on Task m, the model is optimized on the union of all datasets $\mathcal{D}_1 \cup \cdots \cup \mathcal{D}_m$.

Impact of Task Correlation. Theorem 1 suggests that high inter-task correlation ($A_{(m,m')} > 0$) facilitates signal accumulation. When tasks share feature subspaces, training on a subsequent Task m should reinforce the features relevant to Task 1. We design two sequences:

1. **High Correlation:** Task 1 (*Bicycle*) → Task 2 (*Motorcycle*). Both belong to the *Vehicles 1* superclass and share semantic features (e.g., wheels).
2. **Low Correlation:** Task 1 (*Bicycle*) → Task 2 (*Orchid*). The classes belong to disjoint superclasses and represent orthogonal tasks.

Results: As illustrated in Figure 6 (Left), while full replay allows both models to maintain performance, the High Correlation sequence (Teal line) exhibits superior retention and positive backward transfer compared to the Low Correlation sequence (Orange dashed line). The introduction of the semantically related *Motorcycle* task reinforces the feature subspace used by *Bicycle*, validating our theoretical insight that feature sharing is critical for robust signal accumulation.

Impact of Task Ordering and SNR. Theorem 2 uncovers that prioritizing higher-signal tasks facilitates the learning of subsequent tasks. To simulate varying SNR in real-world images, we inject strong Gaussian noise (σ_{noise}) into the inputs. We focus on two aligned tasks from the *Fruit* superclass: *Apple* (Task 1) and *Pear* (Task 2). We investigate whether a high-signal Task 1 facilitates the learning of a low-signal Task 2:

1. **High-Signal First (Setup A):** Task 1 is *Clean Apple* ($\sigma = 0$) → Task 2 is *Noisy Pear* ($\sigma = 5.0$).
2. **Low-Signal First (Setup B):** Task 1 is *Noisy Apple* ($\sigma = 5.0$) → Task 2 is *Noisy Pear* ($\sigma = 5.0$).

Results: Figure 6 (Right) demonstrates the critical role of ordering. In Setup A (Blue line), the model learns

robust "fruit" features from the Clean Apple task in Phase 1. When the Noisy Pear task arrives in Phase 2, the model leverages these pre-learned features to achieve significantly higher accuracy ($\sim 66\%$). In contrast, in Setup B (Red dashed line), the model struggles to learn meaningful features from the initial Noisy Apple task; consequently, its ability to learn the subsequent Noisy Pear task is impaired ($\sim 59\%$). This empirically confirms prioritizing high-signal tasks is essential for effective feature transfer to downstream low-signal tasks.

C Proof of Main Results

C.1 Notations.

Given the iterate $\mathbf{W}^{(t)}$ in sequential training, we define the following notations during the training process:

- The learning dynamics of task k's feature at time t under current task m: $\Gamma_{(m,r)}^{(t,k)} := \langle \mathbf{w}_{(m,r)}^{(t)}, \mathbf{v}_k^* \rangle$.
- The learning dynamics of task k's noise at time t under current task m: $\Phi_{(m,r)}^{(t,k,j)} := \langle \mathbf{w}_{(m,r)}^{(t)}, \boldsymbol{\xi}_k^j \rangle$.
- Derivative: $\ell_{mj}^{(t)} = \ell_{mj}(\mathbf{W}^{(t)}) = 1/(1 + e^{y_{mj} F(\mathbf{W}^{(t)}, \mathbf{x}_{mj})})$ for $j \in [n]$.
- Maximum signal intensity: $\Gamma_{(m,r^*)}^{(t,k)} \equiv \Gamma_{(m,r_k^*)}^{(t,k)}$, where $r_k^* = \arg\max_{r \in [R]} \Gamma_{(0,r)}^{(0,k)}$.
- Maximum noise memorization: $\Phi_{(m,r^*)}^{(t,k,j)} \equiv \Phi_{(m,r_{kj}^*)}^{(t,k,j)}$, where $r_{kj}^* = \arg\max_{r \in [R]} y_{kj} \Phi_{(0,r)}^{(0,k,j)}$.

C.2 Learning dynamics of task k's feature and noise at time t under current task m.

According to Definition 1, we assume that tasks share common features, i.e.,$A_{(m,m')} > 0$. As a result, even without direct training on the target task, the model can still accumulate relevant features through similar tasks. Furthermore, based on the gradient computation, the learned signal can be characterized as follows:

$$
\begin{aligned}
\Gamma_{(m,r)}^{(t,k)} &= \langle \mathbf{w}_{(m,r)}^{(t)}, \mathbf{v}_k^* \rangle \\
&= \langle \mathbf{w}_{(m,r)}^{(t-1)} - \eta \nabla_{\mathbf{w}_r} L(\mathbf{W}_m^{(t-1)}, D_1, ..., D_m), \mathbf{v}_k^* \rangle \\
&= \langle \mathbf{w}_{(m,r)}^{(t-1)} + \frac{\eta}{nm} \sum_{p \in [m]} \sum_{j \in [n]} y_{kj} \ell_{pj}(\mathbf{W}_m^{(t-1)}) [3 \langle \mathbf{w}_{(m,r)}^{(t-1)}, \alpha_p y_{pj} \mathbf{v}_p^* \rangle^2 \cdot \alpha_p y_{pj} \mathbf{v}_p^*], \mathbf{v}_k^* \rangle \\
&= \Gamma_{(m,r)}^{(t-1,k)} + \frac{\eta}{nm} \sum_{j \in [n]} \sum_{p \in [m]} 3 \alpha_p^3 A_{(p,k)} \ell_{pj}(\mathbf{W}_m^{(t-1)}) (\Gamma_{(m,r)}^{(t-1,p)})^2 \\
&= \Gamma_{(0,r)}^{(0,k)} + \frac{\eta}{nm} \sum_{j \in [n]} \sum_{p \in [m]} \sum_{s \in [T_m]} 3 \alpha_p^3 A_{(p,k)} \ell_{pj}(\mathbf{W}_m^{(s-1)}) (\Gamma_{(m,r)}^{(s-1,p)})^2 \\
&= \Gamma_{(0,r)}^{(0,k)} + \frac{\eta}{nm} \sum_{j \in [n]} \sum_{q \in [m]} \sum_{p \in [q]} \sum_{s \in [T_q]} 3 \alpha_p^3 A_{(p,k)} \ell_{pj}(\mathbf{W}_q^{(s-1)}) (\Gamma_{(q,r)}^{(s-1,p)})^2.
\end{aligned} \tag{10}
$$

When considering noise memorization, it can be observed that the noise also continues to accumulate regardless of the relationship between task m and k.

$$
\begin{aligned}
&\Phi_{(m,r)}^{(t,k,j)} = \langle \mathbf{w}_{(m,r)}^{(t)}, \boldsymbol{\xi}_{kj} \rangle = \langle \mathbf{w}_{(m,r)}^{(t)}, \boldsymbol{\xi}_{kj} \rangle \\
&= \langle \mathbf{w}_{(m,r)}^{(t-1)} - \eta \nabla_{\mathbf{w}_r} L(\mathbf{W}_m^{(t-1)}, D_1, ..., D_m), \boldsymbol{\xi}_{kj} \rangle \\
&= \langle \mathbf{w}_{(m,r)}^{(t-1)} + \frac{\eta}{nm} \sum_{j' \in [n]} y_{mj'} \ell_{mj'}(\mathbf{W}_m^{(t-1)}) [3 \langle \mathbf{w}_{(m,r)}^{(t-1)}, \alpha y_{mj'} \mathbf{v}_m^* \rangle^2 \cdot \alpha y_{mj'} \mathbf{v}_m^* + 3 \langle \mathbf{w}_{(m,r)}^{(t-1)}, \boldsymbol{\xi}_{mj'} \rangle^2 \cdot \boldsymbol{\xi}_{mj'}], \boldsymbol{\xi}_{kj} \rangle \\
&+ \langle \sum_{p=1}^{m} \frac{\eta}{nm} \sum_{j' \in [n]} y_{pj'} \ell_{pj'}(\mathbf{W}_m^{(t-1)}) [3 \langle \mathbf{w}_{(m,r)}^{(t-1)}, \alpha y_{pj'} \mathbf{v}_p^* \rangle^2 \cdot \alpha y_{pj'} \mathbf{v}_p^* + 3 \langle \mathbf{w}_{(m,r)}^{(t-1)}, \boldsymbol{\xi}_{pj'} \rangle^2 \cdot \boldsymbol{\xi}_{pj'}], \boldsymbol{\xi}_{kj} \rangle \\
&= \Phi_{(m,r)}^{(t-1,k,j)} + \sum_{p=1}^{m} \frac{\eta}{nm} \sum_{j' \in [n]} y_{pj'} \ell_{pj'}(\mathbf{W}_m^{(t-1)}) (\Phi_{(m,r)}^{(t-1,p,j)})^2 \langle \boldsymbol{\xi}_{pj'}, \boldsymbol{\xi}_{kj} \rangle.
\end{aligned} \tag{11}
$$

C.3 Proof of Theorem 1.

In this section, we present the proof of Theorem 1 in two parts. The first part analyzes the failure of signal learning after training on k tasks (i.e., before task $k+1$). The second part focuses on noise memorization after training on $m > k$ tasks (i.e., before task $m+1$) and further considers two scenarios in the later phase: one where learning continues to fail, and another where signal learning is enhanced.

In the following, we show that the signal learning is always under control before training task $m+1$.

Lemma 5. *In the data replay training process on task m, with probability at least $1 - 1/\operatorname{poly}(d)$, it holds that $\max_{r\in[R],p\in[k]} |\Gamma_{(k,r)}^{(t,p)}| \leq \widetilde{O}(\sigma_0)$ for any $t \in [T_k], p \in [k]$.*

Proof of lemma 5. We consider the induction process to prove the statement. We assume that, for any $s \leq t$, it holds that $\max_{r\in[R],p\in[k]} |\Gamma_{(k,r)}^{(t,p)}| \leq \widetilde{O}(\sigma_0)$. Then, we proceed to analyze the case for $s = t+1$. According to eq. (10), we have:

$$
\begin{aligned}
|\Gamma_{(k,r)}^{(t,k)}| &= |\langle \mathbf{w}_{(m,r)}^{(t)}, \mathbf{v}_k^* \rangle| \\
&\leq |\Gamma_{(k,r)}^{(0,k)}| + |\frac{\eta}{nm} \sum_{j\in[n]} \sum_{p\in[k]} \sum_{s\in[T_k]} 3\alpha_p^3 A_{(p,k)} \ell_{pj}(\mathbf{W}_m^{(s-1)})(\Gamma_{(k,r)}^{(s-1,p)})^2| \\
&= |\Gamma_{(0,r)}^{(0,k)}| + |\frac{\eta}{nm} \sum_{j\in[n]} \sum_{q\in[k]} \sum_{p\in[q]} \sum_{s\in[T_q]} 3\alpha_p^3 A_{(p,k)} \ell_{pj}(\mathbf{W}_q^{(s-1)})(\Gamma_{(q,r)}^{(s-1,p)})^2| \\
&\leq \widetilde{O}(\sigma_0) + |\frac{\eta}{m} \sum_{q\in[k]} \sum_{p\in[q]} 3\alpha_p^3 A_{(p,k)} T_q \widetilde{O}(\sigma_0^2)| \\
&\overset{(i)}{\leq} \widetilde{O}(\sigma_0) + |\frac{\eta}{m} T_v \widetilde{O}(\sigma_0^2) \cdot \sum_{p=1}^{k} (k-p+1)\alpha_p^3 A_{(p,k)}| \\
&\overset{(ii)}{\leq} \widetilde{O}(\sigma_0).
\end{aligned}
\tag{12}
$$

Here, (i) follows from the assumption that every task before k is trained for the same number of iterations T_v; (ii) drives from the choice of $T_v \leq \widetilde{O}\left(\frac{m}{\eta\sigma_0 \sum_{p=1}^{k}(k-p+1)\alpha_p^3 A_{(p,k)}}\right)$. □

Lemma 6. *Let $T_\xi^- = \frac{nm}{\eta\sigma_0(\sigma_\xi\sqrt{d})^3}$. In the data replay training process on task k, with probability at least $1 - 1/\operatorname{poly}(d)$, it holds that:*

$$
\max_{r\in[R],p\in[k],j\in[n]} y_{mj}\Phi_{(k,r)}^{(t,p,j)} \leq (Rd)^{-1/3} \quad \text{for any} \quad t \leq \mathrm{T}_\xi^-, p \in [k].
$$

Proof of lemma 6. We first assume lemma 6 holds for any $t \leq T_\xi^- - 1$, then the following can be obtained:

$$
\begin{aligned}
\ell_{pj}^{(t)} &= \frac{1}{1 + \exp\{\sum_{r=1}^{R}[\alpha_p^3(\Gamma_{(k,r)}^{(t,p)})^3 + (y_{pj}\Phi_{(k,r)}^{(t,p,j)})^3]\}} \\
&\overset{(i)}{\geq} \frac{1}{1 + \exp\{\widetilde{O}(d^{-1}) + \widetilde{O}(\alpha_p^3 R\widetilde{O}(\sigma_0^3)\}} \\
&\overset{(ii)}{\geq} \frac{1}{1 + \exp\{\widetilde{O}(d^{-1}) + \widetilde{O}(d^{-3/2})\}} \\
&\geq \frac{1}{2} - \frac{e^{2d^{-1}} - 1}{2(1 + e^{2d^{-1}})} \\
&= \frac{1}{2} - \widetilde{O}(d^{-1}),
\end{aligned}
$$

where the inequality (i) derives from the induction hypothesis and lemma 5 and (ii) holds due to the Condition 1 and SNR choices.

Therefore, using recursion eq. (11), with high probability $1-1/\operatorname{poly}(d)$, we have

$$
\begin{aligned}
y_{kj}\Phi_{(k,r^*)}^{(t+1,k,j)} &= y_{kj}\Phi_{(k,r^*)}^{(t,k,j)} + \sum_{p=1}^{k}\frac{3\eta}{nm}\sum_{j'\in[n]} y_{kj}y_{pj'}\ell_{pj'}(\mathbf{W}_k^{(t)})(\Phi_{(k,r)}^{(t,p,j)})^2\langle\boldsymbol{\xi}_{pj'},\boldsymbol{\xi}_{kj}\rangle \\
y_{kj}\Phi_{(k,r^*)}^{(t+1,k,j)} &= y_{kj}\Phi_{(k,r^*)}^{(0,k,j)} + \sum_{p=1}^{k}\sum_{s=1}^{t-1}\frac{3\eta}{nm}\sum_{j'\in[n]} y_{kj}y_{pj'}\ell_{pj'}(\mathbf{W}_k^{(s)})(\Phi_{(k,r)}^{(s,p,j)})^2\langle\boldsymbol{\xi}_{pj'},\boldsymbol{\xi}_{kj}\rangle \\
y_{kj}\Phi_{(k,r^*)}^{(t+1,k,j)} &= y_{kj}\Phi_{(0,r^*)}^{(0,k,j)} + \sum_{q=1}^{k}\sum_{p=1}^{q}\sum_{s=1}^{t-1}\frac{3\eta}{nm}\sum_{j'\in[n]} y_{kj}y_{pj'}\ell_{pj'}(\mathbf{W}_k^{(s)})(\Phi_{(k,r)}^{(s,p,j)})^2\langle\boldsymbol{\xi}_{pj'},\boldsymbol{\xi}_{kj}\rangle \\
y_{kj}\Phi_{(k,r^*)}^{(t+1,k,j)} &\overset{(i)}{=} y_{kj}\Phi_{(0,r^*)}^{(0,k,j)} + \Theta\left(\frac{3\eta d\sigma_\xi^2}{nm}\right)\sum_{s=1}^{t-1}\left(y_{kj}\Phi_{(k,r^*)}^{(s,k,j)}\right)^2 \\
&\quad \pm \Theta\left(\frac{3\eta\sqrt{d}\sigma_\xi^2}{nm}\right)\sum_{q=1}^{k}\sum_{\substack{p\in[q],\, j'\in[n]\\(p,j')\neq(k,j)}}\sum_{s=1}^{t-1}\ell_{pj'}(\mathbf{W}_k^{(s)})\left(y_{kj}\Phi_{(k,r^*)}^{(s,p,j)}\right)^2 \\
&\overset{(ii)}{=} y_{kj}\Phi_{(0,r^*)}^{(0,k,j)} + \Theta\left(\frac{3\eta d\sigma_\xi^2}{nm}\right)\sum_{s=1}^{t-1}\left(y_{kj}\Phi_{(k,r^*)}^{(s,p,j)}\right)^2 \pm \widetilde{O}\left(\frac{3\sqrt{d}\sigma_\xi^2 k^2}{\sigma_0\sum_{p\in[k]}3\alpha_p^3 A_{(p,k)}}\cdot\frac{1}{(Rd)^{2/3}}\right)
\end{aligned}
\tag{13}
$$

where equality (i) holds due to lemma 14 and (ii) comes from lemma 7. Let $T_{\xi_{pj}}^- = \Theta(\frac{nm}{\eta M_{pj}})$ with $M_{pj} = (y_{kj}\Phi_{(k,r^*)}^{(T_1,p,j)}\sqrt{d}\sigma_\xi^2)$. Then, according to lemma 20, it holds $y_{kj}\Phi_{(k,r^*)}^{(T_1,p,j)} \le (Rd)^{-1/3}$ for any $t \le T_{\xi_{pj}}^-$ since $(Rd)^{-1/3} \ge \sigma_0\sigma_\xi\sqrt{d} \ge 2y_{kj}\Phi_{(k,r)}^{(T_1,p,j)}$. Moreover, we also know $\max_{ij}\ T_{\xi_{pj}}^- + 1 \le T_\xi^-$ by concentration, which indicates that $(y_{kj}\Phi_{(k,r^*)}^{(t,p,j)} \le (Rd)^{-1/3}$. □

Lemma 7. *Given any $p\in[k]$ and $k\in[M]$, with high probability $1-1/\operatorname{poly}(d)$, it holds that:*

$$
\sum_{s=1}^{t}\frac{1}{n}\sum_{j'}\ell_{pj'}(\mathbf{W}_k^{(s)}) \le \widetilde{O}\left(\frac{m}{\eta\sigma_0\sum_{p\in[k]}\alpha_p^3 A_{(p,k)}}\right).
$$

Proof of lemma 7. Applying lemma 21 with $z^{(0)} = \Gamma_{k,r^*}^{(\tau_{kv}^k,k)}, h = H = \frac{3\eta}{nm}\sum_{j'\in[n]}\sum_{p\in[k]}\alpha_p^3 A_{(p,k)}$ and $\frac{1}{nm}\sum_{j'}\ell_{pj'} \le \frac{1}{m}$, then it holds that

$$
\sum_{s=1}^{t}\frac{1}{n}\sum_{j'}\ell_{pj'}(\mathbf{W}_k^{(t)}) \le \log d + \widetilde{O}\left(\frac{m}{\eta\sigma_0\sum_{p\in[k]}\alpha_p^3 A_{(p,k)}}\right). \tag{14}
$$

□

Lemma 8 (Restatement of Lemma 1)**.** *Suppose the SNR condition satisfying $\frac{k^2}{R^{2/3}\sigma_0^2\sigma_\xi^2 d^{13/6}} \lesssim \frac{\sum_{p=1}^{k}(1-\frac{p-1}{k})\alpha_p^3 A_{(p,k)}}{(\sigma_\xi\sqrt{d})^3} \lesssim \frac{1}{n}$, and there exists an iteration $\tau_{kj}^k \le \mathrm{T}_\xi^k = \mathrm{T}_\xi^- + O(\log(d))$ such that τ_{kj}^k is the first iteration for which $\max_{r\in[R]}(y_{kj}\Phi_{(k,r)}^{(t,k,j)}) \ge \Theta(R^{-\frac{1}{3}})$, and for any $t \le \mathrm{T}_\xi^k$ it holds that $\max_{r\in[R]}|\Gamma_{(k,r)}^{(t,k)}| \le \widetilde{O}(\sigma_0)$. Then, if the additional SNR condition $\frac{m^2-k^2}{R^{1/3}\sigma_0\sigma_\xi d} \lesssim \frac{\sum_{p=1}^{m}(1-\frac{p-1}{m})\alpha_p^3 A_{(p,k)}}{(\sigma_\xi\sqrt{d})^3} \lesssim \frac{1}{n}$ also holds, there exists an iteration τ_{kj}^m such that τ_{kj}^m is the first iteration satisfying $\max_{r\in[R]}(y_{kj}\Phi_{(m,r)}^{(t,k,j)}) \ge \Theta(R^{-\frac{1}{4}})$. In this case, we can also guarantee that $\max_{r\in[R]}|\Gamma_{(m,r)}^{(t,k)}| \le \widetilde{O}(\sigma_0)$ for any $t \le \mathrm{T}_\xi^m$.*

Proof of lemma 8. According to the definition of τ_{kj}^k, it is clear that $\max_{r\in[R]} y_{kj}\Phi_{(k,r)}^{(s,k,j)} \le \Theta(R^{-\frac{1}{3}})$ for $s \le \tau_{kj}^k$.

Furthermore, it holds that $\tau_{kj}^k \geq T_\xi^-$ due to lemma 6. For any $s \leq \tau_{kj}^k$, we also have

$$\begin{aligned}
\ell_{kj}^{(s)} &= \frac{1}{1+\exp\{\sum_{r=1}^R[\alpha_k^3(\Gamma_{(k,r)}^{(s,k)})^3 + (y_{kj}\Phi_{(k,r)}^{(s,k,j)})^3]\}} \\
&\geq \frac{1}{1+\exp\{R\Theta(1/R) + \widetilde{O}(\alpha_k^3 R\sigma_0^3)\}} \\
&\geq \frac{1}{1+\exp\{\Theta(1)\}} \\
&= \Theta(1).
\end{aligned} \tag{15}$$

Let $\tau_{r^*,kj}^-$ be the first iteration such that $y_{kj}\Phi_{(k,r^*)}^{(t,k,j)} \geq \Theta((Rd)^{-\frac{1}{3}})$, then it follows $\tau_{r^*,kj}^- > \mathrm{T}_\xi^-$. After enrolling update rule with $r = r_{kj}^*$, for any $\tau_{r^*,kj}^- \leq t \leq \min\{\tau_{kj}^k, T_\xi^k\}$, it holds that

$$\begin{aligned}
y_{kj}\Phi_{(k,r^*)}^{(t+1,k,j)} &= y_{kj}\Phi_{(k,r^*)}^{(\tau_{r^*,kj}^-,k,j)} + \frac{3\eta}{nm}\sum_{p=1}^k\sum_{s=\tau_{r^*,kj}^-}^{t-1}\frac{3\eta}{nm}\sum_{j'\in[n]} y_{kj}y_{pj'}\ell_{pj'}(\mathbf{W}_k^{(s)})(\Phi_{(k,r)}^{(s,p,j)})^2\langle\boldsymbol{\xi}_{pj'},\boldsymbol{\xi}_{kj}\rangle \\
&\overset{(i)}{=} y_{kj}\Phi_{(k,r^*)}^{((\tau_{r^*,kj}^-,k,j),k,j)} + \Theta\left(\frac{3\eta d\sigma_\xi^2}{nm}\right)\sum_{s=\tau_{r^*,kj}^-}^{t-1}\left(y_{kj}\Phi_{(k,r^*)}^{(s,k,j)}\right)^2 \\
&\pm \Theta\left(\frac{3\sqrt{d}\sigma_\xi^2 k^2}{\sigma_0\sum_{p\in[k]}3\alpha_p^3 A_{(p,k)}}\cdot\frac{1}{(Rd)^{2/3}}\right) \\
&\overset{(ii)}{=} y_{kj}\Phi_{(k,r^*)}^{((\tau_{r^*,1j}^-,k,j),k,j)} + \Theta\left(\frac{3\eta d\sigma_\xi^2}{nm}\right)\sum_{s=\tau_{r^*,kj}^-}^{t-1}\left(y_{kj}\Phi_{(k,r^*)}^{(s,k,j)}\right)^2 \pm o\left(R^{-\frac{1}{3}}d^{-\frac{1}{3}}\right).
\end{aligned}$$

The inequality (i) holds due to Lemma 14 and Lemma 7 and (ii) comes from SNR choices.

Let $A = \Theta(\frac{\eta d\sigma_\xi^2}{nm}), C = o(R^{-\frac{1}{3}}d^{-\frac{1}{3}}), v = \Theta(R^{-\frac{1}{3}})$. By applying the tensor power method via Lemma 19, we have:

$$\begin{aligned}
\tau_{r^*,kj}^k &\leq \tau_{r^*,kj}^- + \frac{21}{Ay_{kj}\Phi_{r^*,kj}^{\left(\tau_{r^*,kj}^-\right)}} + 8\left\lceil\frac{\log\left(v/\left\lfloor y_{ij}\Phi_{r^*,kj}^{\left(\tau_{r^*,kj}^-\right)}\right\rfloor\right)}{\log(2)}\right\rceil \\
&\leq \Theta\left(\frac{1}{\eta}\frac{nm}{\left(\sqrt{d}\sigma_\xi\right)^3\sigma_0}\right) + \Theta\left(\frac{1}{\eta}\frac{nmR^{1/3}}{d^{2/3}\sigma_\xi^2}\right) + O(\log d) \\
&\leq O\left(\frac{1}{\eta}\frac{mn}{\left(\sqrt{d}\sigma_\xi\right)^3\sigma_0} + \frac{1}{\eta}\frac{mnR^{1/3}}{d^{2/3}\sigma_\xi^2} + \log d\right) = T_\xi^k.
\end{aligned}$$

Next, we will show that the above also holds when training task m for the first scenario. First, we have:

$$\begin{aligned}
\ell_{mj}^{(t)} &= \frac{1}{1+\exp\{\sum_{r=1}^R[\alpha_m^3(\Gamma_{(m,r)}^{(t,m)})^3 + (y_{mj}\Phi_{(m,r)}^{(t,m,j)})^3]\}} \\
&\overset{(i)}{\geq} \frac{1}{1+\exp\{\Theta(1) + \widetilde{O}(\alpha_m^3 R\widetilde{O}(\sigma_0^3)\}} \\
&\geq \Theta(1).
\end{aligned}$$

Then, when training task $m \geq k$, noise memorization satisfies:

$$\begin{aligned} y_{kj}\Phi_{(m,r^*)}^{(t+1,k,j)} &= y_{kj}\Phi_{(m,r^*)}^{(t,k,j)} + \sum_{p=1}^{k}\frac{3\eta}{nm}\sum_{j'\in[n]} y_{kj}y_{pj'}\ell_{pj'}(\mathbf{W}_m^{(t)})(\Phi_{(m,r)}^{(t,p,j)})^2\langle\boldsymbol{\xi}_{pj'},\boldsymbol{\xi}_{kj}\rangle \\ &= y_{kj}\Phi_{(m,r^*)}^{(0,k,j)} + \sum_{p=1}^{m}\sum_{s=1}^{t-1}\frac{3\eta}{nm}\sum_{j'\in[n]} y_{kj}y_{pj'}\ell_{pj'}(\mathbf{W}_m^{(s)})(\Phi_{(m,r)}^{(s,p,j)})^2\langle\boldsymbol{\xi}_{pj'},\boldsymbol{\xi}_{kj}\rangle \end{aligned} \tag{16}$$

Then, according to Lemma 14, it also holds that:

$$\begin{aligned} y_{kj}\Phi_{(m,r^*)}^{(t+1,k,j)} =& y_{kj}\Phi_{(k+1,r^*)}^{(0,k,j)} + \Theta\left(\frac{3\eta d\sigma_\xi^2}{nm}\right)\sum_{s=1}^{t-1}\left(y_{kj}\Phi_{(m,r^*)}^{(s,m,j)}\right)^2 \\ &\pm\Theta\left(\frac{3\eta\sqrt{d}\sigma_\xi^2}{nm}\right)\sum_{q=k}^{m}\sum_{\substack{p\in[q],\,j'\in[n]\\(p,j')\neq(k,j)}}\sum_{s=1}^{t-1}\ell_{pj'}(\mathbf{W}_m^{(s)})\left(y_{kj}\Phi_{(m,r^*)}^{(s,p,j)}\right)^2 \\ \overset{(i)}{=}& y_{kj}\Phi_{(k+1,r^*)}^{(0,k,j)} + \Theta\left(\frac{3\eta d\sigma_\xi^2}{nm}\right)\sum_{s=1}^{t-1}\left(y_{kj}\Phi_{(m,r^*)}^{(s,p,j)}\right)^2 \pm\widetilde{O}\left(\frac{3\sqrt{d}\sigma_\xi^2(m^2-k^2)}{\sigma_0\sum_{p\in[m]}3\alpha_p^3A_{(p,k)}}\cdot\frac{1}{(R)^{2/3}}\right) \\ \overset{(ii)}{=}& y_{kj}\Phi_{(k+1,r^*)}^{(0,k,j)} + \Theta\left(\frac{3\eta d\sigma_\xi^2}{nm}\right)\sum_{s=T_1}^{t}\left(y_{1j}\Phi^{(s,1,j)}\right)^2 \pm o\left(R^{-1/3}\right). \end{aligned} \tag{17}$$

Here, (i) follows from Lemma 7 with the range of $p\in[m]$ adjusted accordingly; (ii) is derived from the robustness of the SNR choices.

Let $A=\Theta(\frac{\eta d\sigma_\xi^2}{mn}), C=o(R^{-\frac{1}{3}}), v=\Theta(R^{-\frac{1}{4}})$. By applying the tensor power method via Lemma 19, we also have:

$$\begin{aligned} \tau_{r^*,kj}^m &\leq T_k + \frac{21}{Ay_{kj}\Phi_{r^*,kj}^{(T_k)}} + 8\left[\frac{\log\left(v/\left[y_{kj}\Phi_{r^*,kj}^{(T_k)}\right]\right)}{\log(2)}\right] \\ &\leq \Theta\left(\frac{1}{\eta}\frac{mn}{\left(\sqrt{d}\sigma_\xi\right)^3\sigma_0}\right) + \Theta\left(\frac{1}{\eta}\frac{nmR^{1/3}}{d^{2/3}\sigma_\xi^2}\right) + \Theta\left(\frac{1}{\eta}\frac{nmR^{1/3}}{d\sigma_\xi^2}\right) + \log d \\ &\leq \Theta\left(\frac{1}{\eta}\frac{mn}{\left(\sqrt{d}\sigma_\xi\right)^3\sigma_0} + \frac{1}{\eta}\frac{mnR^{1/3}}{d^{2/3}\sigma_\xi^2} + \frac{1}{\eta}\frac{nmR^{1/3}}{d\sigma_\xi^2}\right) + \log d = T_\xi^m. \end{aligned}$$

□

Lemma 9. *For any $0\leq t\leq T_\xi^-$, with probability at least $1-1/\operatorname{poly}(d)$, it holds that:*

$$\min_{r\in[R],m\in[M],p,k\in[m],j\in[n]} y_{kj}\Phi_{(m,r)}^{(t,p,j)} \geq -(d)^{-1/2}.$$

Proof of Lemma 9. According to Lemma 15, we know $\min_{r\in[m]} y_{kj}\Phi_{(r,0)}^{(0,k,j)} \geq -\widetilde{O}(\sqrt{d}\sigma_\xi\sigma_0)$ holds for any $j\in[n]$. By Lemma 6, we know $\ell_{kj}^{(s)}\geq\frac{1}{2}-O\left(d^{-1}\right)$ for any $s\leq \mathrm{T}_\xi^-$. Similar to Lemma 13, we can obtain that for any $t\leq \mathrm{T}_\xi^-$,

$$\begin{aligned} y_{kj}\Phi_{(k,r)}^{(t+1,k,j)} &= y_{kj}\Phi_{(0,r^*)}^{(0,k,j)} + \Theta\left(\frac{3\eta d\sigma_\xi^2}{nm}\right)\sum_{s=1}^{t-1}\left(y_{kj}\Phi_{(k,r^*)}^{(s,p,j)}\right)^2 \pm o\left(\sqrt{d}\sigma_\xi\sigma_0\right) \\ &\overset{(i)}{\geq} -\widetilde{O}\left(\sqrt{d}\sigma_\xi\sigma_0\right) - o\left(\sqrt{d}\sigma_\xi\sigma_0\right) \\ &= -\widetilde{O}\left(\sqrt{d}\sigma_\xi\sigma_0\right), \end{aligned}$$

where (i) holds due to the second term being always positive. □

Lemma 10 (Restatement of Lemma 2)**.** *Suppose the SNR satisfying* $\frac{k^2}{R^{2/3}\sigma_0^2\sigma_\xi^2 d^{13/6}} \lesssim \frac{\sum_{p=1}^k (1-\frac{p-1}{k})\alpha_p^3 A_{(p,k)}}{(\sigma_\xi\sqrt{d})^3} \lesssim \frac{1}{n}$*, and there exists an iteration* $\tau_{kj}^k \le \mathrm{T}_\xi^k = \mathrm{T}_\xi^- + O(\log(d))$ *such that* τ_{kj}^k *is the first iteration where* $\max_{r\in[R]}(y_{kj}\Phi_{(k,r)}^{(t,k,j)}) \ge \Theta(R^{-\frac{1}{3}})$*, and for any* $t \le \mathrm{T}_\xi^k$ *it holds that* $\max_{r\in[R]}|\Gamma_{(k,r)}^{(t,k)}| \le \widetilde{O}(\sigma_0)$*. Then, if the additional SNR condition* $\frac{\sum_{p=1}^m \alpha_p^3 A_{(p,k)}}{(\sigma_\xi\sqrt{d})^3} \gtrsim \frac{1}{nR^{1/3}\sigma_0\sigma_\xi\sqrt{d}}$ *also holds, there exists* $\tau_{kv}^m \le \mathrm{T}_v^m = \mathrm{T}_v^k + O(\log(d))$ *such that* τ_{kv}^m *be the first iteration satisfying* $\max_{r\in[R]}|\Gamma_{(m,r)}^{(t,k)}| \ge \Theta(\frac{1}{\alpha_k R^{1/5}})$.

Proof of Lemma 10. The proof of the first part of Lemma 10 follows directly from Lemma 8 for the initial training phase. Therefore, we focus on the second training phase, beginning with the analysis of enhanced signal learning, followed by a demonstration that noise memorization remains controlled under certain SNR conditions.

It is clear that before $T_k = kT_v$ in Lemma 5, we have $\max_{r\in[R],p\in[k]}|\Gamma_{(k,r)}^{(t,p)}| \le \widetilde{O}(\sigma_0)$ for any $t \in [T_k], p \in [k]$. Then, according to eq. (10), we have the following since T_k:

$$\begin{aligned}\Gamma_{(m,r^*)}^{(t,k)} &= \Gamma_{(m,r^*)}^{(t-1,k)} + \frac{\eta}{nm}\sum_{j\in[n]}\sum_{p\in[m]} 3\alpha_p^3 A_{(p,k)}\ell_{pj}(\mathbf{W}_m^{(t-1)})(\Gamma_{(m,r^*)}^{(t-1,p)})^2 \\ &= \Gamma_{(m,r^*)}^{(t-1,k)} + \Theta\left(\frac{\eta}{m}\sum_{p\in[m]} 3\alpha_p^3 A_{(p,k)}\right)(\Gamma_{(m,r^*)}^{(t-1,p)})^2.\end{aligned} \tag{18}$$

Then, by applying the tensor power method from Lemma 18 to the sequence $\{\Gamma_{(m,r^*)}^{(s,k)}\}_{s\ge T_k}$, let $h = H = 3\frac{\eta}{m}(\sum_{p\in[m]} 3\alpha_p^3 A_{(p,k)})$, $z^{(0)} = \Gamma_{(k+1,r^*)}^{(0,k)} \ge O(\sigma_0)$, $v = \Theta(\frac{1}{\alpha_k R^{1/5}})$, then we obtain:

$$\begin{aligned}\tau_{kv}^m &\le T_k + \frac{m}{3\eta\sigma_0\sum_{p\in[m]} 3\alpha_p^3 A_{(p,k)}} + 8\left[\frac{\log\left(v/z^{(0)}\right)}{\log(2)}\right] \\ &\le O\left(\frac{1}{\eta}\frac{mn}{\left(\sqrt{d}\sigma_\xi\right)^3\sigma_0} + \frac{1}{\eta}\frac{mnR^{1/3}}{d^{2/3}\sigma_\xi^2}\right) + \frac{m}{3\eta\sigma_0\sum_{p\in[m]} 3\alpha_p^3 A_{(p,k)}} + \log d \\ &\le O\left(\frac{1}{\eta}\frac{mn}{\left(\sqrt{d}\sigma_\xi\right)^3\sigma_0} + \frac{1}{\eta}\frac{mnR^{1/3}}{d^{2/3}\sigma_\xi^2} + \frac{m}{3\eta\sigma_0\sum_{p\in[m]} 3\alpha_p^3 A_{(p,k)}}\right) + \log d = T_v^m.\end{aligned}$$

Then, it is noticed that if the additional SNR condition $\frac{\sum_{p=1}^m \alpha_p^3 A_{(p,k)}}{(\sigma_\xi\sqrt{d})^3} \gtrsim \frac{1}{nR^{1/3}\sigma_0\sigma_\xi\sqrt{d}}$ also holds, we will have $T_v^m \le T_\xi^m$ according to Lemma 8, which indicates that noise memorization remain controlled within $\Theta(R^{-1/4})$ and is slower than the signal learning in the second training phase. □

Theorem 3 (Restatement of Theorem 1)**.** *Suppose the setting in Condition 1 holds, and the SNR satisfies* $\frac{k^2}{R^{2/3}\sigma_0^2\sigma_\xi^2 d^{13/6}} \lesssim \frac{\sum_{p=1}^k (1-\frac{p-1}{k})\alpha_p^3 A_{(p,k)}}{(\sigma_\xi\sqrt{d})^3} \lesssim \frac{1}{n}$*. Consider full data-replay training with learning rate* $\eta \in (0, \widetilde{O}(1)]$*, and let* $(\mathbf{x}_k, y_k) \sim \mathcal{D}_k$ *be a test sample from the task* k*. Then, with high probability, there exist training times* T_k *and* T_m *(*$m > k$*) such that*

- *The model fails to correctly classify task* k *immediately after learning it:*

$$\mathbb{P}\left\{y_k F\left(\mathbf{W}^{(T_k)}, \mathbf{x}_k\right) < 0\right\} \ge \frac{1}{2} - \frac{1}{\mathrm{polylog}(d)}. \tag{19}$$

- *(Persistent Learning Failure on Task* k*) If the additional SNR condition holds* $\frac{m^2-k^2}{R^{1/3}\sigma_0\sigma_\xi d} \lesssim \frac{\sum_{p=1}^m (1-\frac{p-1}{m})\alpha_p^3 A_{(p,k)}}{(\sigma_\xi\sqrt{d})^3} \lesssim \frac{1}{n}$*, then the model still fails to correctly classify task* k *after subsequent training to task* m*:*

$$\mathbb{P}\left\{y_k F\left(\mathbf{W}^{(T_m)}, \mathbf{x}_k\right) < 0\right\} \ge \frac{1}{2} - \frac{1}{\mathrm{polylog}(d)}. \tag{20}$$

- *(Enhanced Signal Learning on Task k) If the additional SNR conditions holds* $\frac{\sum_{p=1}^{m}\alpha_p^3 A_{(p,k)}}{(\sigma_\xi\sqrt{d})^3} \gtrsim \frac{1}{nR^{1/3}\sigma_0\sigma_\xi\sqrt{d}}$, *then the model can correctly classify task k after subsequent training to task m:*

$$\mathbb{P}\left\{y_k F\left(\mathbf{W}^{(T_m)}, \mathbf{x}_k\right) < 0\right\} \leq \frac{1}{\operatorname{poly}(d)}. \tag{21}$$

Proof of Theorem 3. We first prove the training phase one (before training task $k+1$). For the new test data $(\mathbf{x}_k, y_k) \sim \mathcal{D}_k$, with probability at least $1 - 1/\operatorname{poly}(d)$, we have

$$\begin{aligned} y_k F(\mathbf{W}^{(T_k)}, \mathbf{x}_k) &= y_k \sum_{r\in[R]} (\langle \mathbf{w}_r^{(T_k)}, \mathbf{x}_k^1\rangle^3 + \langle \mathbf{w}_r^{(T_k)}, \mathbf{x}_k^2\rangle^3) \\ &= y_k \sum_{r\in[R]} (\langle \mathbf{w}_r^{(T_k)}, \alpha_k y_k \mathbf{v}_k^*\rangle^3 + \langle \mathbf{w}_r^{(T_k)}, \boldsymbol{\xi}_k\rangle^3) \\ &= \sum_{r\in[R]} [\alpha_k^3 (\Gamma_{(k,r)}^{(T_k)})^3 + y\langle \mathbf{w}_r^{(T_k)}, \boldsymbol{\xi}_k\rangle^3] \\ &\leq \sum_{r\in[R]} y_k \langle \mathbf{w}_r^{(T_k)}, \boldsymbol{\xi}_k\rangle^3 + \widetilde{O}(R\alpha_k^3\sigma_0^3) \end{aligned} \tag{22}$$

Let $\mathbf{P}_v = \sum_{m\in[M]} \mathbf{v}_m^* \left(\mathbf{v}_m^*\right)^\top$ and $\mathbf{P}_v^\perp = \mathbf{I}_d - \mathbf{P}_v$. Since $\boldsymbol{\xi} \sim \mathcal{N}\left(0, \sigma_\xi^2 \mathbf{P}_v^\perp\right)$, there exists a vector $\boldsymbol{\xi}_d \sim \mathcal{N}\left(0, \sigma_\xi^2 \mathbf{I}_d\right)$ such that $\boldsymbol{\xi} = \mathbf{P}_v^\perp \boldsymbol{\xi}_d$. Now, decompose $\mathbf{w}_r^{(T_k)}$ as: $\mathbf{w}_r^{(T_k)} = \mathbf{P}_v \mathbf{w}_r^{(T_k)} + \mathbf{P}_v^\perp \mathbf{w}_r^{(T_k)}$. According to the definition of $\Phi_{(k,r)}^{(T_k,k,j)} = \left\langle \mathbf{w}_r^{(T_k)}, \boldsymbol{\xi}_{kj}\right\rangle = \left\langle \mathbf{P}_v^\perp \mathbf{w}_r^{(T_k)}, \boldsymbol{\xi}_{kj}\right\rangle$ and Lemma 8, for Task 1's data $(\mathbf{x}_k, y_k)$, we have

$$\Theta\left(R^{-\frac{1}{3}}\right) \leq \max_{r\in[R]} y_{kj} \Phi_{(k,r)}^{(T_k,k,j)} = \max_{r\in[m]} \left\langle \mathbf{P}_v^\perp \mathbf{w}_r^{(T_k)}, y_{kj}\boldsymbol{\xi}_{kj}\right\rangle.$$

Denote $r^* = \arg\max y_{kj} \Phi_{(k,r)}^{(T_k,k,j)}$, then it holds that

$$\begin{aligned} \sum_{r\in[R]} \left\langle \mathbf{P}_v^\perp \mathbf{w}_r^{(T_k)}, \frac{y_{kj}\boldsymbol{\xi}_{kj}}{\|\boldsymbol{\xi}_{kj}\|}\right\rangle^3 &\geq \frac{1}{\|\boldsymbol{\xi}_{kj}\|^3} \left[\left(y_{kj}\Phi_{k,r^*}^{(T_k,k,j)}\right)^3 - \sum_{r\neq r^*} \left(y_{kj}\Phi_{(k,r)}^{(T_k,k,j)}\right)^3\right] \\ &\overset{(i)}{\geq} \tilde{\Omega}\left(\frac{1}{d^{3/2}\sigma_\xi^3}\right)\left[\Theta\left(R^{-1}\right) - \widetilde{O}\left(R\left(d^{-1/2}\right)^3\right)\right] \\ &= \widetilde{\Omega}\left(\frac{1}{d^{3/2}\sigma_\xi^3}\right) \overset{(ii)}{\geq} 1. \end{aligned} \tag{23}$$

Here, (i) comes from Lemma 9 and Lemma 8 and (ii) holds due to the assumption on σ_ξ. Given that the model $\mathbf{W}^{(T_k)}$ and the test label y_k are independent of the noise $\boldsymbol{\xi}_d$, it follows that the distribution of $\sum_{r\in[R]} y_k \langle \mathbf{P}_v^\perp \mathbf{w}_r^{(T_k)}, \boldsymbol{\xi}_d\rangle^3$ is symmetric. This holds under the condition that $\mathbf{W}^{(T_k)}$ and $y\boldsymbol{\xi}_d$ are distributed as $\mathcal{N}(0, \sigma_\xi^2 \mathbf{I}_d)$, where $y \in \{-1, +1\}$. According to Lemma 16, let $\mathbf{w}_r = \mathbf{P}_v^\perp \mathbf{w}_r^{(T_k)}$ and $\boldsymbol{u} = y_{kj}\boldsymbol{\xi}_{kj}/\|\boldsymbol{\xi}_{kj}\|$, then we derive:

$$\begin{aligned} &\mathbb{P}_{\boldsymbol{\xi}_d}\left(\sum_{r\in[R]} \left\langle \mathbf{P}_v^\perp \mathbf{w}_r^{(T_k)}, y_k\boldsymbol{\xi}_d\right\rangle^3 < -\epsilon\sigma_\xi^3\right) \\ &\geq \frac{1}{2} - \mathbb{P}_{\boldsymbol{\xi}_d}\left(\left|\sum_{r\in[R]} \left\langle \mathbf{P}_v^\perp \mathbf{w}_r^{(T_k)}, y_k\boldsymbol{\xi}_d\right\rangle^3\right| \leq \epsilon\sigma_\xi^3 \left|\sum_{r\in[R]} \left\langle \mathbf{P}_v^\perp \mathbf{w}_r^{(T_k)}, \boldsymbol{u}\right\rangle^3\right|\right) \\ &\geq \frac{1}{2} - O\left(\epsilon^{1/3}\right). \end{aligned} \tag{24}$$

Taking $\epsilon = 1/\operatorname{polylog}(d)$, it holds that

$$\mathbb{P}\left(\sum_{r\in[R]} \left\langle \mathbf{P}_v^\perp \mathbf{w}_r^{(T_k)}, y_k\boldsymbol{\xi}_d\right\rangle^3 < -\widetilde{O}\left(\sigma_\xi^3\right)\right) \geq \frac{1}{2} - \frac{1}{\operatorname{polylog}(d)}.$$

Moreover, along with eq. (22), we can further obtain the following:

$$\begin{aligned} y_k F\left(\boldsymbol{W}^{(T_k)}, \mathbf{x}_k\right) &\leq \sum_{r\in[R]} y_k \left\langle \mathbf{w}_r^{(T_k)}, \boldsymbol{\xi} \right\rangle^3 + \widetilde{O}(R\alpha_k^3\sigma_0^3) \\ &\leq -\widetilde{O}\left(\sigma_\xi^3\right) + \widetilde{O}(R\alpha_k^3\sigma_0^3) \\ &\overset{(i)}{\leq} 0, \end{aligned}$$

where (i) comes from the Condition 1 and the SNR choices. The proofs for T_k and T_m are identical, where $t = T_m$, it still holds that $\Gamma \leq \widetilde{O}(\sigma_0)$ and $\max_{r\in[R]} y_{1j}\Phi_{(m,r)}^{(T_m,k,j)} \geq \Theta(R^{-\frac{1}{4}})$. Hence, the remainder of the proof proceeds exactly as in the case $t = T_m$.

For the scenario of enhanced signal learning, the noise memorization of training phase 2 will be under control and the signal learning will increase as stated in Lemmar 10. Thus, given the new test data $(\mathbf{x}_k, y_k) \sim \mathcal{D}_k$ for task k, with probability at least $1 - 1/\text{ poly }(d)$, we have

$$\begin{aligned} y_k F\left(\boldsymbol{W}^{(T_m)}, \mathbf{x}\right) &= y_k \sum_{r\in[R]} (\langle \mathbf{w}_r^{(T_m)}, \mathbf{x}_k^1\rangle^3 + \langle \mathbf{w}_r^{(T_m)}, \mathbf{x}_k^2\rangle^3) \\ &= y_k \sum_{r\in[R]} \left\langle \boldsymbol{W}^{(T_m)}, y_k\alpha_k\mathbf{v}_k^* \right\rangle^3 + \left\langle \boldsymbol{W}^{(T_m)}, \boldsymbol{\xi}_k \right\rangle^3 \\ &\overset{(i)}{\geq} \Theta(\alpha_k^3 \cdot R \cdot \alpha_k^{-3} R^{-3/5}) \pm \Theta(R \cdot R^{-3/4}) \\ &\geq \widetilde{\Omega}(1). \end{aligned}$$

Here, (i) follows from Lemma 10, which shows that, under the SNR condition stated in Theorem 3, noise memorization is slower than signal learning. □

C.4 Proof of Theorem 2

In this section, we present the proof of Theorem 2 in two parts. The first part analyzes the success of signal learning after training on k tasks (i.e., before task $k+1$). The second part focuses on noise memorization after training on $m > k$ tasks (i.e., before task $m+1$) and further considers two scenarios in the later phase: one where learning fails to retain previously acquired features, and another where signal learning continues to improve.

Lemma 11. *During the data replay training process, with probability at least* $1 - 1/\operatorname{poly}(d)$*, it holds that:*

$$\max_{r\in[R],k\in[m],j\in[n]} |\Phi_{(k,r)}^{(t,k,j)}| \leq \widetilde{O}(\sigma_0\sigma_\xi\sqrt{d}) \quad \textit{for any} \quad t \leq \mathrm{T}_k.$$

Proof of Lemma 11. According to the initialization and the concentration by Lemma 15, with probability at least $1 - 1/\operatorname{poly}(d)$, it holds that

$$\bar{\Phi}_{(0,r)}^{(0)} := \max_{k\in[M],j\in[n]} |\Phi_{(0,r)}^{(0,k,j)}| = \max_{j\in[n]} |\langle \mathbf{w}_{(0,r)}^{(0)}, \boldsymbol{\xi}_{kj}\rangle| \leq \widetilde{O}(\sigma_0\sigma_\xi\sqrt{d}).$$

Next, we consider the induction process to prove the statement. First, we assume that $\Phi_{(k,r)}^{(s)} \leq \widetilde{O}(\sigma_0\sigma_\xi\sqrt{d})$ holds for any $s \leq t$. Then, we proceed to analyze the case for $s = t+1$. Denote $\bar{\Phi}_{(k,r)}^{(s,k,j)} = \max_{k\in[M],j\in[n]} |\Phi_{(k,r)}^{(s,k,j)}|$,

according to the update rule (11), we have

$$\begin{aligned}
\bar{\Phi}_{(k,r)}^{(s+1)} &\leq \max_{k\in[M],j\in[n]} \Phi_{(k,r)}^{(s,k,j)} + \frac{3\eta}{nm}\sum_{p=1}^{k}\sum_{j'\in[n]} y_{pj'}\ell_{pj'}(\mathbf{W}_k^{(t-1)})(\Phi_{(k,r)}^{(t-1,p,j)})^2\langle \boldsymbol{\xi}_{pj'}, \boldsymbol{\xi}_{kj}\rangle \\
&\overset{(i)}{\leq} \bar{\Phi}_{(k,r)}^{(s)} + \frac{3\eta\sqrt{d}\sigma_\xi^2(n-1)k}{nm}(\bar{\Phi}_{(k,r)}^{(s)})^2 + \frac{3\eta d\sigma_\xi^2}{nm}(\Phi_{(k,r)}^{(s,k,j)})^2 \\
&\overset{(ii)}{\leq} \bar{\Phi}_{(k,r)}^{(s)} + \frac{3k(n-1)\eta d\sqrt{d}\sigma_\xi^4\sigma_0^2}{nm} + \frac{3\eta d^2\sigma_\xi^4\sigma_0^2}{nm} \\
&\leq \Phi_{(0,r)}^{0} + \frac{3k(s-1)(n-1)\eta d\sqrt{d}\sigma_\xi^4\sigma_0^2}{nm} + \frac{3(s-1)\eta d^2\sigma_\xi^4\sigma_0^2}{nm} \\
&\overset{(iii)}{\leq} \widetilde{O}(\sigma_0\sigma_\xi\sqrt{d}) + O(\frac{3T_k k(n-1)\eta d\sqrt{d}\sigma_\xi^4\sigma_0^2}{nm} + \frac{3T_k\eta d^2\sigma_\xi^4\sigma_0^2}{nm}) \\
&\overset{(iv)}{\leq} \widetilde{O}(\sigma_0\sigma_\xi\sqrt{d}),
\end{aligned}$$

where (i) holds due to the concentration in Lemma 14; (ii) derives from the induction hypothesis; (iii) comes from $s+1 \leq T_k$; (iv) holds due to $T_k \leq \frac{m}{\eta\sigma_0\sigma_\xi}$. □

Lemma 12 (Restatement of Lemma 3)**.** *Suppose the SNR satisfying* $\frac{\sum_{p=1}^k \alpha_p^3 A_{(p,k)}}{(\sigma_\xi\sqrt{d})^3} \gtrsim \frac{1+nk^2/\sqrt{d}}{kn}$*, and there exists an iteration* $\tau_{kv}^k \leq \mathrm{T}_v^k = \mathrm{T}_v^- + O(\log(d))$ *such that* τ_{kv}^k *is the first iteration where* $\max_{r\in[R]}|\Gamma_{(k,r)}^{(t,k)}| \geq \Theta(\frac{1}{\alpha_k R^{1/3}})$*, and for any* $t \leq \mathrm{T}_v^k$ *it holds that* $\max_{r\in[R]}|\Phi_{(k,r)}^{(t,k,j)})| \leq \widetilde{O}(\sigma_0\sigma_\xi\sqrt{d})$*. Then, if the additional SNR condition* $\frac{m^2}{R^{2/3}\sigma_0^2\sigma_\xi^2 d^{13/6}} \lesssim \frac{\sum_{p=1}^m \alpha_p^3 A_{(p,k)}}{(\sigma_\xi\sqrt{d})^3} \lesssim \frac{\alpha_k R^{1/3}}{n}$ *also holds, there exists* $\tau_{kj}^m \leq \mathrm{T}_\xi^m = \mathrm{T}_\xi^k + O(\log(d))$ *such that* τ_{kj}^m *be the first iteration satisfying* $\max_{r\in[R]}(y_{kj}\Phi_{(m,r)}^{(t,k,j)}) \geq \Theta(R^{-\frac{1}{5}})$.

Proof of Lemma 12. In the first training phase, the noise memorization can be controlled by Lemma 11. Thus, we only need to consider the signal learning process here. By learning dynamic of signal in eq. (10), we have:

$$\begin{aligned}
\Gamma_{(k,r^*)}^{(t,k)} &= \Gamma_{(k,r^*)}^{(t-1,k)} + \frac{\eta}{n}\sum_{j\in[n]}\sum_{p\in[k]} 3\alpha_p^3 A_{(p,k)}\ell_{pj}(\mathbf{W}_k^{(t-1)})(\Gamma_{(k,r^*)}^{(t-1,p)})^2 \\
&= \Gamma_{(k,r^*)}^{(t-1,k)} + \Theta\left(\eta\sum_{p\in[k]} 3\alpha_p^3 A_{(p,k)}\right)(\Gamma_{(k,r^*)}^{(t-1,p)})^2.
\end{aligned} \tag{25}$$

Then, by applying the tensor power method from Lemma 18 to the sequence $\{\Gamma_{(k,r^*)}^{(s,k)}\}_{s\geq T_k}$, let $h = H = 3\eta(\sum_{p\in[k]} 3\alpha_p^3 A_{(p,k)})$, $z^{(0)} = \Gamma_{(k+1,r^*)}^{(0,k)} \geq O(\sigma_0)$, $v = \Theta(\frac{1}{\alpha_k R^{1/3}})$, then we obtain:

$$\begin{aligned}
\tau_{kv}^k &\leq \frac{m}{3\eta\sigma_0\sum_{p\in[k]} 3\alpha_p^3 A_{(p,k)}} + 8\left[\frac{\log\left(v/z^{(0)}\right)}{\log(2)}\right] \\
&\leq \frac{m}{3\eta\sigma_0\sum_{p\in[k]} 3\alpha_p^3 A_{(p,k)}} + \log d = T_v^k.
\end{aligned}$$

When considering the second phase training, we first consider the signal learning and denote τ_{kv}^m as the first time that $\Gamma_{(m,r^*)}^{(t,k)}$ exceeds $(\alpha_k R)^{-\frac{1}{4}}$. Then, the signal learning dynamic will be:

$$\begin{aligned}
\Gamma_{(m,r^*)}^{(t,k)} &= \Gamma_{(k,r^*)}^{(t-1,k)} + \frac{\eta}{mn}\sum_{j\in[n]}\sum_{p\in[m]} 3\alpha_p^3 A_{(p,k)}\ell_{pj}(\mathbf{W}_k^{(t-1)})(\Gamma_{(m,r^*)}^{(t-1,p)})^2 \\
&= \Gamma_{(m,r^*)}^{(t-1,k)} + \Theta\left(\frac{\eta}{m}\sum_{p\in[m]} 3\alpha_p^3 A_{(p,k)}\right)(\Gamma_{(m,r^*)}^{(t-1,p)})^2.
\end{aligned} \tag{26}$$

Then, we still apply the tensor power method from Lemma 18 to the sequence $\{\Gamma_{(m,r^*)}^{(s,k)}\}_{s\geq T_m}$, but with modified parameters, such that: $h = H = 3\frac{\eta}{m}(\sum_{p\in[m]} 3\alpha_p^3 A_{(p,k)})$, $z^{(0)} = \Gamma_{(k+1,r^*)}^{(0,k)} \geq \Theta(\frac{1}{\alpha_k R^{1/3}})$, $v = \Theta(\frac{1}{\alpha_k R^{1/4}})$, then we obtain:

$$\begin{aligned}\tau_{kv}^k &\leq T_v^k + \frac{m}{3\eta\sigma_0\sum_{p\in[m]} 3\alpha_p^3 A_{(p,k)}} + 8\left[\frac{\log\left(v/z^{(0)}\right)}{\log(2)}\right] \\ &\leq \frac{m}{3\eta\sigma_0\sum_{p\in[k]} 3\alpha_p^3 A_{(p,k)}} + \frac{m}{3\eta\sigma_0\sum_{p\in[m]} 3\alpha_p^3 A_{(p,k)}} + \log d = T_v^m.\end{aligned}$$

Therefore, as long as $t \leq T_v^m$, signal learning remains bounded by $\Theta\left(\frac{1}{\alpha_k R^{1/4}}\right)$. In the sequel, we show that noise memorization can accumulate to $R^{-1/5}$, making the noise term larger than the signal.

Similar to Lemma 2, it can be derived that $y_{kj}\Phi_{(m,r^*)}^{(t+1,k,j)} \leq (Rd)^{-1/3}$ for any $t \leq \tau_{r^*,kj}^k := T_k + \Theta\left(\frac{1}{\eta}\frac{mn}{\left(\sqrt{d}\sigma_\xi\right)^3\sigma_0}\right)$. Then, it can be shown that the following holds:

$$\begin{aligned}y_{kj}\Phi_{(m,r^*)}^{(t+1,k,j)} =& y_{kj}\Phi_{(k+1,r^*)}^{(0,k,j)} + \Theta\left(\frac{3\eta d\sigma_\xi^2}{nm}\right)\sum_{s=1}^{t-1}\left(y_{kj}\Phi_{(m,r^*)}^{(s,m,j)}\right)^2 \\ &\pm \Theta\left(\frac{3\eta\sqrt{d}\sigma_\xi^2}{nm}\right)\sum_{q=k}^{m}\sum_{\substack{p\in[q],\, j'\in[n]\\(p,j')\neq(k,j)}}\sum_{s=1}^{t-1}\ell_{pj'}(\mathbf{W}_m^{(s)})\left(y_{kj}\Phi_{(m,r^*)}^{(s,p,j)}\right)^2 \\ &\overset{(i)}{=} y_{kj}\Phi_{(k+1,r^*)}^{(0,k,j)} + \Theta\left(\frac{3\eta d\sigma_\xi^2}{nm}\right)\sum_{s=1}^{t-1}\left(y_{kj}\Phi_{(m,r^*)}^{(s,p,j)}\right)^2 \\ &\pm \widetilde{O}\left(\frac{3(\alpha_k R)^{1/3}\sqrt{d}\sigma_\xi^2(m^2-k^2)}{\sum_{p\in[m]} 3\alpha_p^3 A_{(p,k)}}\cdot\frac{1}{(Rd)^{2/3}}\right) \\ &\overset{(ii)}{=} y_{kj}\Phi_{(k+1,r^*)}^{(0,k,j)} + \Theta\left(\frac{3\eta d\sigma_\xi^2}{nm}\right)\sum_{s=T_1}^{t}\left(y_{1j}\Phi^{(s,1,j)}\right)^2 \pm o\left((Rd)^{-1/3}\right).\end{aligned} \tag{27}$$

Here, (i) follows from Lemma 7 with the range of $p \in [m]$ and $z^{(0)} = \Gamma_{(k,r^*)}^{\tau_{kv}^k,k}$ adjusted accordingly; (ii) is derived from the robustness of the SNR choices.

Let $A = \Theta(\frac{\eta d\sigma_\xi^2}{mn}), C = o((Rd)^{-\frac{1}{3}}), v = \Theta(R^{-\frac{1}{5}})$. By applying the tensor power method via Lemma 19, we also have:

$$\begin{aligned}\tau_{r^*,kj}^m &\leq \tau_{r^*,kj}^k + \frac{21}{Ay_{kj}\Phi_{r^*,kj}^{\left(\tau_{r^*,kj}^k\right)}} + 8\left[\frac{\log\left(v/\left[y_{kj}\Phi_{r^*,kj}^{\left(\tau_{r^*,kj}^k\right)}\right]\right)}{\log(2)}\right] \\ &\leq T_k + \Theta\left(\frac{1}{\eta}\frac{mn}{\left(\sqrt{d}\sigma_\xi\right)^3\sigma_0}\right) + \Theta\left(\frac{1}{\eta}\frac{nmR^{1/3}}{d^{2/3}\sigma_\xi^2}\right) + \log d \\ &\leq \Theta\left(\frac{1}{\eta}\frac{mn}{\left(\sqrt{d}\sigma_\xi\right)^3\sigma_0} + \frac{1}{\eta}\frac{mnR^{1/3}}{d^{2/3}\sigma_\xi^2}\right) + \log d = T_\xi^m.\end{aligned}$$

Based on the condition of SNR, we have $T_\xi^m \leq T_v^m$, which indicates that noise memorization exceeds signal learning during the second phase. □

Lemma 13 (Restatement of Lemma 4)**.** *Suppose the SNR satisfying* $\frac{\sum_{p=1}^k \alpha_p^3 A_{(p,k)}}{(\sigma_\xi\sqrt{d})^3} \gtrsim \frac{1+nk^2/\sqrt{d}}{kn}$*, and there exists an iteration* $\tau_{kv}^k \leq \mathrm{T}_v^k = \mathrm{T}_v^- + O(\log(d))$ *such that* τ_{kv}^k *is the first iteration where* $\max_{r\in[R]}|\Gamma_{(k,r)}^{(t,k)}| \geq \Theta(\frac{1}{\alpha_k R^{1/3}})$*, and for any* $t \leq \mathrm{T}_v^k$ *it holds that* $\max_{r\in[R]}|\Phi_{(k,r)}^{(t,k,j)})| \leq \widetilde{O}(\sigma_0\sigma_\xi\sqrt{d})$*. Then, if the additional SNR condition*

$\frac{\sum_{p=1}^{m}\alpha_p^3 A_{(p,k)}}{(\sigma_\xi\sqrt{d})^3} \gtrsim \frac{\alpha_k R^{1/3}\sigma_0\left((1-\frac{k-1}{m})+nm/\sqrt{d}\right)}{n}$ *also holds, there exists* $\tau_{kv}^m \le \mathrm{T}_v^m = \mathrm{T}_v^k + O(\log(d))$ *such that* τ_{kv}^m *be the first iteration satisfying* $\max_{r\in[R]}|\Gamma_{(m,r)}^{(t,k)}| \ge \Theta(\frac{1}{\alpha_k R^{1/5}})$.

Proof of Lemma 13. The proof of the first training phase is identical to Lemma 12. Thus, we only focus on the second training phase. Similarly, we have the update for signal learning as follows:

$$\begin{aligned}\Gamma_{(m,r^*)}^{(t,k)} &= \Gamma_{(k,r^*)}^{(t-1,k)} + \frac{\eta}{mn}\sum_{j\in[n]}\sum_{p\in[m]} 3\alpha_p^3 A_{(p,k)}\ell_{pj}(\mathbf{W}_k^{(t-1)})(\Gamma_{(m,r^*)}^{(t-1,p)})^2 \\ &= \Gamma_{(m,r^*)}^{(t-1,k)} + \Theta\left(\frac{\eta}{m}\sum_{p\in[m]} 3\alpha_p^3 A_{(p,k)}\right)(\Gamma_{(m,r^*)}^{(t-1,p)})^2.\end{aligned} \tag{28}$$

Then, we still apply the tensor power method from Lemma 18 to the sequence $\{\Gamma_{(m,r^*)}^{(s,k)}\}_{s\ge T_m}$, but with modified parameters, such that: $h = H = 3\frac{\eta}{m}(\sum_{p\in[m]} 3\alpha_p^3 A_{(p,k)})$, $z^{(0)} = \Gamma_{(k+1,r^*)}^{(0,k)} \ge \Theta(\frac{1}{\alpha_k R^{1/3}})$, $v = \Theta(\frac{1}{\alpha_k R^{1/5}})$, then we obtain:

$$\begin{aligned}\tau_{kv}^k &\le T_v^k + \frac{m}{3\eta\sigma_0\sum_{p\in[m]} 3\alpha_p^3 A_{(p,k)}} + 8\left[\frac{\log\left(v/z^{(0)}\right)}{\log(2)}\right] \\ &\le \frac{m}{3\eta\sigma_0\sum_{p\in[k]} 3\alpha_p^3 A_{(p,k)}} + \frac{m}{3\eta\sigma_0\sum_{p\in[m]} 3\alpha_p^3 A_{(p,k)}} + \log d = T_v^m.\end{aligned}$$

Therefore, as long as $t \le T_v^m$, signal learning remains bounded by $\Theta\left(\frac{1}{\alpha_k R^{1/5}}\right)$. Moreover, according to the SNR condition, we have $T_v^m \le T_\xi^m$, which indicates that during the training phase the noise memorization will not exceed $\Theta(\frac{1}{R^{1/3}})$. □

Theorem 4 (Restatement of Theorem 4)**.** *Suppose the setting in Condition 1 holds, and the SNR satisfies* $\frac{\sum_{p=1}^{k}\alpha_p^3 A_{(p,k)}}{(\sigma_\xi\sqrt{d})^3} \gtrsim \frac{1+nk^2/\sqrt{d}}{kn}$. *Consider full data-replay training with learning rate* $\eta \in (0, \widetilde{O}(1)]$, *and let* $(\mathbf{x}_k, y_k) \sim \mathcal{D}_k$ *be a test sample from the task* k. *Then, with high probability, there exist training times* T_k *and* T_m $(m > k)$ *such that*

- *The model can correctly classify task* k *immediately after learning it:*

$$\mathbb{P}\left\{y_k F\left(\mathbf{W}^{(T_k)}, \mathbf{x}_k\right) < 0\right\} \le \frac{1}{\mathrm{poly}(d)}. \tag{29}$$

- *(Catastrophic Forgetting on Task* k*) If the additional SNR conditions holds* $\frac{m^2}{R^{2/3}\sigma_0^2\sigma_\xi^2 d^{13/6}} \lesssim \frac{\sum_{p=1}^{m}\alpha_p^3 A_{(p,k)}}{(\sigma_\xi\sqrt{d})^3} \lesssim \frac{\alpha_k R^{1/3}}{n}$, *then it occurs Catastrophic Forgetting on task* k *after subsequent training to task* m*:*

$$\mathbb{P}\left\{y_k F\left(\mathbf{W}^{(T_m)}, \mathbf{x}_k\right) < 0\right\} \ge \frac{1}{2} - \frac{1}{\mathrm{polylog}(d)}. \tag{30}$$

- *(Continual Learning on Task* k*) If the additional SNR conditions holds* $\frac{\sum_{p=1}^{m}\alpha_p^3 A_{(p,k)}}{(\sigma_\xi\sqrt{d})^3} \gtrsim \frac{\alpha_k R^{1/3}\sigma_0\left((1-\frac{k-1}{m})+nm/\sqrt{d}\right)}{n}$, *then the model can still correctly classify task* k *after subsequent training to task* m*:*

$$\mathbb{P}\left\{y_k F\left(\mathbf{W}^{(T_m)}, \mathbf{x}_k\right) < 0\right\} \le \frac{1}{\mathrm{poly}(d)}. \tag{31}$$

Proof of Theorem 4. We first present the analysis for the initial training phase; the results for the second phase in the continual learning scenario follow analogously, with the primary difference lying in the bound on noise memorization. Given the new test data $(\mathbf{x}_k, y_k) \sim \mathcal{D}_k$ for task k, with probability at least $1 - 1/\mathrm{poly}(d)$,

we have

$$\begin{aligned} y_k F\left(\boldsymbol{W}^{(T_k)}, \mathbf{x}\right) &= y_k \sum_{r\in[R]} (\langle \mathbf{w}_r^{(T_k)}, \mathbf{x}_k^1\rangle^3 + \langle \mathbf{w}_r^{(T_k)}, \mathbf{x}_k^2\rangle^3) \\ &= y_k \sum_{r\in[R]} \left\langle \boldsymbol{W}^{(T_k)}, y_k\alpha_k \mathbf{v}_k^*\right\rangle^3 + \left\langle \boldsymbol{W}^{(T_k)}, \boldsymbol{\xi}_k\right\rangle^3 \\ &\overset{(i)}{\geq} \Theta(\alpha_k^3 \cdot R \cdot \alpha_k^{-3} R^{-1}) \pm \Theta(R \cdot \sigma_0^3 \sigma_\xi^3 d^{3/2}) \\ &\geq \widetilde{\Omega}(1). \end{aligned}$$

Here, (i) follows from Lemma 13 and the SNR condition stated in Theorem 2. The second phase differs from $\Gamma \geq \Theta(\frac{1}{\alpha_k R^{1/5}})$ and $\Phi \leq \Theta(R^{-1/3})$.

Next, we present the proof of Catastrophic Forgetting during the second phase. Given a new test sample $(\mathbf{x}_k, y_k) \sim \mathcal{D}_k$ from task k, and noting that we consider binary classification with labels $y = \pm 1$, it follows that, with probability at least $1/2 - 1/\mathrm{poly}(d)$, the label y_k will interact oppositely with the Φ, which implies that:

$$\begin{aligned} y_k F\left(\boldsymbol{W}^{(T_k)}, \mathbf{x}\right) &= y_k \sum_{r\in[R]} (\langle \mathbf{w}_r^{(T_k)}, \mathbf{x}_k^1\rangle^3 + \langle \mathbf{w}_r^{(T_k)}, \mathbf{x}_k^2\rangle^3) \\ &= y_k \sum_{r\in[R]} \left\langle \boldsymbol{W}^{(T_k)}, y_k\alpha_k \mathbf{v}_k^*\right\rangle^3 + \left\langle \boldsymbol{W}^{(T_k)}, \boldsymbol{\xi}_k\right\rangle^3 \\ &\overset{(i)}{\leq} \Theta(\alpha_k^3 \cdot R \cdot \alpha_k^{-3} R^{-1}) - \Theta(R \cdot R^{-3/4}) \\ &\leq 0. \end{aligned}$$

Here, (i) holds due to Lemma 12 and the SNR condition stated in Theorem 2. □

D Supplementary Lemmas

Lemma 14. *Suppose that $\delta_\xi > 0$ and $d = \Omega(\log(4n/\delta_\xi))$. Then, for all $i, i' \in [n]$, with probability at least $1 - \delta_\xi$,*

$$\begin{aligned} &\sigma_\xi^2 d/2 \leq \|\boldsymbol{\xi}_i\|_2^2 \leq 3\sigma_\xi^2 d/2 \\ &|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'}\rangle| \leq 2\sigma_\xi^2 \cdot \sqrt{d\log(4n^2/\delta_\xi)}. \end{aligned}$$

Lemma 15. *Under the Gaussian initialization, with probability $1 - 1/\ \mathrm{poly}\ (d)$, we have*

- *Given any $m \in [M]$, $\max_{r\in[R]} \Gamma_{(m,r)}^{(0)} > \Omega\left(\sigma_0\right)$. In addition, $\max_{r\in[R], m\in[M]} \left|\Gamma_{(m,r)}^{(0)}\right| \leq O\left(\sigma_0\sqrt{\log d}\right)$.*
- *Given any $k \in [M]$ and $j \in [n], \max_{r\in[R]} y_{kj}\Phi_{(0,r)}^{(0,k,j)} > \Omega\left(\sqrt{d}\sigma_\xi\sigma_0\right)$. In addition, $\max_{r\in[R],k\in[M],j\in[n]} \left|\Phi_{(0,r)}^{(0,k,j)}\right| \leq O\left(\sigma_\xi\sigma_0\sqrt{d\log d}\right)$. for all $r \in [R]$ and $m \in [M]$.*

The proof of Lemma 14 and Lemma 15 can be derived directly from the properties of the Gaussian distribution. In the following, we will provide some tensor power lemmas that can be extended to m cases.

Lemma 16 (Lemma K. 12 in Jelassi and Li (2022))**.** *Let $\{\mathbf{w}_r\}_{r=1}^R$ be vectors in $\mathbb{R}^d$ and $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma_\xi^2 \mathbf{I}_d)$. If there exists a unit norm vector $\boldsymbol{u}$ such that $|\sum_{r=1}^R \langle \mathbf{w}_r, \boldsymbol{u}\rangle^3| \geq 1$, then for any $\epsilon \in (0,1)$, we have*

$$\mathbb{P}\left(\left|\sum_{r=1}^R \langle \mathbf{w}_r, \boldsymbol{\xi}\rangle^3\right| \leq \epsilon\sigma_\xi^3\right) \leq O\left(\epsilon^{1/3}\right).$$

Lemma 17 (Lemma K. 15 in Jelassi and Li (2022))**.** *Let $\left\{z^{(t)}\right\}_{t=0}^T$ be a positive sequence defined by the following recursions:*

$$\begin{aligned} z^{(t+1)} &\geq z^{(t)} + h\left[z^{(t)}\right]^2, \\ z^{(t+1)} &\leq z^{(t)} + H\left[z^{(t)}\right]^2, \end{aligned}$$

where $z^{(0)} > 0$ is the initialization and $h, H > 0$. Let $v > 0$ such that $z^{(0)} \le v$ and t_0 be the first iteration $z^{(t)} \ge v$. Then, we have

$$t_0 \le \frac{3}{hz^{(0)}} + \frac{8H}{h}\left\lceil \frac{\log\left(v/z^{(0)}\right)}{\log(2)} \right\rceil.$$

Lemma 18. *Let $\left\{z_i^{(t)}\right\}_{t=0}^{T}$ be a positive sequence defined by the following recursions:*

$$z_i^{(t+1)} \ge z_i^{(t)} + h\sum_{j=1}^{m}\left[z_j^{(t)}\right]^2,$$
$$z_i^{(t+1)} \le z_i^{(t)} + H\sum_{j=1}^{m}\left[z_j^{(t)}\right]^2,$$

where $z_j^{(0)} > 0 (j \in [m])$ is the initialization and $h, H > 0$. Let $v > 0$ such that $\max_j z_j^{(0)} \le v$ and t_0 be the first iteration $z_j^{(t)} \ge v$. Then, we have

$$t_0 \le \frac{3}{h\max_j z_j^{(0)}} + \frac{8Hm}{h}\left\lceil \frac{\log\left(v/z^{(0)}\right)}{\log(2)} \right\rceil.$$

Proof of Lemma 18. Let$M^{(t)} = \max\left(z_1^{(t)}, \ldots, z_m^{(t)}\right)$. Due to symmetry, it suffices to analyze $M^{(t)}$. Fix any time step t. Suppose $M^{(t)} = z_k^{(t)}$, then the lower bound is:

$$z_k^{(t+1)} \ge z_k^{(t)} + h\left(\left[z_k^{(t)}\right]^2 + \sum_{j\neq k}\left[z_j^{(t)}\right]^2\right) \ge M^{(t)} + h\left[M^{(t)}\right]^2$$

Therefore, we have $M^{(t+1)} \ge M^{(t)} + h\left[M^{(t)}\right]^2$. The sum of squares of all variables satisfies:

$$\sum_{j=1}^{m}\left[z_j^{(t)}\right]^2 \le m\left[M^{(t)}\right]^2.$$

Therefore, for any $z_i^{(t+1)}$, we have:

$$z_i^{(t+1)} \le z_i^{(t)} + H\cdot m\left[M^{(t)}\right]^2.$$

Hence,

$$M^{(t+1)} \le M^{(t)} + Hm\left[M^{(t)}\right]^2.$$

Replace H in Lemma 17 with Hm, and let the initial value be $M^{(0)}$. Applying the result directly yields:

$$t_0 \le \frac{3}{hM^{(0)}} + \frac{8Hm}{h}\left\lceil \frac{\log\left(v/M^{(0)}\right)}{\log 2} \right\rceil.$$

□

Lemma 19 (Lemma K. 16 in Jelassi and Li (2022))**.** *Let $\left\{z^{(t)}\right\}_{t=0}^{T}$ be a positive sequence defined by the following recursions*

$$z^{(t)} \ge z^{(0)} + A\sum_{s=0}^{t-1}\left[z^{(s)}\right]^2 - C,$$
$$z^{(t)} \le z^{(0)} + A\sum_{s=0}^{t-1}\left[z^{(s)}\right]^2 + C,$$

where $A, C > 0$ and $z^{(0)} > 0$ is the initialization. Assume that $C \leq z^{(0)}/8$. Let t_0 be the first iteration $z^{(t)} \geq v$. If $v > z^{(0)}$, we have the following upper bound

$$t_0 \leq \frac{21}{Az^{(0)}} + 8\left\lceil \frac{\log\left(v/z^{(0)}\right)}{\log(2)} \right\rceil.$$

Lemma 20 (Lemma E. 7 in Bao et al.)**.** *For the same sequence $\left\{z^{(t)}\right\}_{t\geq 0}$ be a positive sequence satisfying the recursive upper bound in lemma 19 Let $v > 0$ such that $z^{(0)} \leq v$ and t_0 be the first iteration $z^{(t)} \geq v$. For any $v \geq 2z^{(0)}$, we have the following lower bound*

$$t_0 \geq \frac{1}{8Az^{(0)}}.$$

Lemma 21 (Lemma E. 8 in Bao et al.)**.** *Let $\left\{z^{(t)}\right\}_{t=0}^{T}$ and $\left\{a^{(t)}\right\}_{t=0}^{T}$ be two positive sequences admitting the following recursions*

$$z^{(t+1)} \geq z^{(t)} + ha^{(t)}\left[z^{(t)}\right]^2,$$
$$z^{(t+1)} \leq z^{(t)} + Ha^{(t)}\left[z^{(t)}\right]^2,$$

where $0 < h < H$ and $z^{(0)} > 0$. If $\max_{t\leq T} a^{(t)} \leq A$, we have

$$\sum_{s=0}^{T} a^{(s)} \leq \frac{4}{hz^{(0)}} + \frac{8HA}{h}\left\lceil \frac{\log\left(z^{(T)}/z^{(0)}\right)}{\log(2)} \right\rceil,$$

and

$$\sum_{s=0}^{T} a^{(s)} \geq \frac{z^{(T)} - z^{(0)}}{H\left[z^{(T)}\right]^2}.$$