Understanding Forgetting in Continual Learning with Linear Regression

Meng Ding[1] **Kaiyi Ji**[1] **Di Wang**[2] **Jinhui Xu**[1]

Abstract

Continual learning, focused on sequentially learning multiple tasks, has gained significant attention recently. Despite the tremendous progress made in the past, the theoretical understanding, especially factors contributing to *catastrophic forgetting*, remains relatively unexplored. In this paper, we provide a general theoretical analysis of forgetting in the linear regression model via Stochastic Gradient Descent (SGD) applicable to both under-parameterized and overparameterized regimes. Our theoretical framework reveals some interesting insights into the intricate relationship between task sequence and algorithmic parameters, an aspect not fully captured in previous studies due to their restrictive assumptions. Specifically, we demonstrate that, given a sufficiently large data size, the arrangement of tasks in a sequence—where tasks with larger eigenvalues in their population data covariance matrices are trained later—tends to result in increased forgetting. Additionally, our findings highlight that an appropriate choice of step size will help mitigate forgetting in both under-parameterized and over-parameterized settings. To validate our theoretical analysis, we conducted simulation experiments on both linear regression models and Deep Neural Networks (DNNs). Results from these simulations substantiate our theoretical findings.

1. Introduction

Continual learning, also known as lifelong learning, is a subfield of machine learning that focuses on developing a model capable of learning continuously from a stream of data, which are i.i.d sampled from different tasks and presented sequentially to the model. A primary challenge in continual learning is the *catastrophic forgetting* phenomenon (McCloskey & Cohen, 1989), wherein the model forgets previously acquired knowledge when exposed to new data.

Previous research addressing catastrophic forgetting in continuous learning primarily focuses on empirical studies, which can be broadly classified into three categories: expansion-based methods, regularization-based methods, and memory-based methods. Expansion-based methods (Yoon et al., 2017; 2019; Yang et al., 2021) mitigate catastrophic forgetting by allocating distinct subsets of network parameters to individual tasks. Regularization-based methods (Kirkpatrick et al., 2017; Aljundi et al., 2018; Serra et al., 2018; Liu & Liu, 2022) employee structural regularization in fixed capacity models to counteract forgetting, which penalize significant changes in parameters that are crucial for previous tasks. Memory-based methods (Shin et al., 2017; Chaudhry et al., 2018; Riemer et al., 2018; Saha et al., 2021; Lin et al., 2022; Hao et al., 2023) alleviate forgetting by storing subsets of previous task data or synthesizing pseudo-data without data-replay.

Recently, there has been a growing body of work focused on understanding the behavior of catastrophic forgetting from a theoretical standpoint. For example, Bennani et al. 2020; Doan et al. 2021 analyze the generalization of continual learning for Orthogonal Gradient Descent (OGD) (Farajtabar et al., 2020) in the Neural Tangent Kernel (NTK) (Jacot et al., 2018) regime. Lee et al. 2021; Asanuma et al. 2021 explore the impact of task similarity in a teacher-student setting. Evron et al. 2022; Lin et al. 2023 provide a detailed forgetting analysis of the *minimum-norm interpolator* for the overparameterized linear regression model. However, the existing analyses of forgetting often rely on relatively stringent assumptions that may not be applicable in many scenarios. For example, Bennani et al. 2020; Doan et al. 2021; Evron et al. 2022; Lin et al. 2023 necessitate an over-parameterized regime for their analysis, which may be invalid when involving large datasets. Moreover, Lee et al. 2021; Asanuma et al. 2021; Lin et al. 2023; Swartworth et al. 2023 assume that data follows a Gaussian distribution that may not hold in real-world datasets exhibiting more complex distributions. Evron et al. 2022; Lin et al. 2023 focus on the minimum-norm interpolator, where each task requires achieving zero loss on its training samples and

[1]Department of Computer Science and Engineering, State University of New York at Buffalo, Buffalo, USA [2]Division of CEMSE, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. Correspondence to: Kaiyi Ji <kaiyiji@buffalo.edu>, Jinhui Xu <jinhui@buffalo.edu>.

hence can find a closed-form solution.

In this paper, we investigate the behavior of forgetting under the linear regression model via the more practical Stochastic Gradient Descent (SGD) method and provide a general theoretical analysis that is applicable to both over-parameterized and under-parameterized regimes. Our main contributions can be summarized as follows:

Firstly, our work provides a theoretical analysis for multi-step SGD algorithms in both underparameterized and over-parameterized regimes, with the population data covariance matrix satisfying the general fourth moment instead of Gaussian distribution as in existing studies. In specific, we provide a novel upper bound on the model forgetting, as well as a matching lower bound that shows the tightness of our characterization. Our bounds derive the forgetting bound that is stated as a function of **1**) the spectrum of the population data covariance matrices for each task, **2**) the step size, **3**) the number of training samples and **4**) the effective dimensions on the forgetting.

Second, our study provides some interesting insights into the impact of task sequence and algorithmic parameters on the degree of forgetting. Specifically, we show that when the data size is sufficiently large, forgetting tends to escalate when we postpone the training of tasks, whose population data covariance matrices possess larger eigenvalues. It is intuitive that when tasks with larger eigenvalues are trained later, the model might overfit these tasks due to their high variance. In addition, our findings reveal that an appropriate choice of step size can help mitigate forgetting in both underparameterized and overparameterized settings. Note that these results cannot be derived from existing works due to their restrictive data distribution assumptions or closed-form updating rules. More detailed discussions can be found in Section 4.

Finally, we conducted simulation experiments on both linear regression models and Deep Neural Networks (DNNs) to validate our theoretical analysis. Our simulation results indicate that both linear regression models and DNNs exhibit increased forgetting when tasks with larger eigenvalues are encountered later. Additionally, we demonstrate that smaller step sizes in training can also mitigate forgetting across task sequences, especially in under-parameterized settings. Interestingly, we observe that in over-parameterized DNNs, higher dimensionality does not necessarily equate to more forgetting if the dataset size is fixed, as opposite to the linear regression case.

1.1. Related Work

In this section, we discuss related work on Covariate Shift, SGD analysis in linear regression, and theoretical studies for catastrophic forgetting.

Covariate Shift Covariate shift is a specific set-up in machine learning (Pan & Yang, 2009; Sugiyama & Kawanabe, 2012), referring to a distribution mismatch between the training and test data. The concept is typically applied in transfer learning, which can be seen as a particular instance of continual learning, generally involving two tasks. For example, Mohri & Medina 2012; Cortes & Mohri 2014; Kpotufe & Martinet 2018; Cortes et al. 2019; Hanneke & Kpotufe 2020; Ma et al. 2023; Wu et al. 2022b examine the (regularized) empirical risk minimizer, which focuses on minimizing the empirical and generalization error across accessible datasets. Nevertheless, the standard covariate shift is defined over two distinct data distributions, which can not be directly applied to our case. Consequently, we propose an extended version in Definition 2.2 to better suit our context.

SGD Analysis Recently, several studies have investigated the behavior of Stochastic Gradient Descent (SGD) in linear regression models through the lens of bias-variance decomposition (Défossez & Bach, 2015; Dieuleveut et al., 2017; Jain et al., 2017; 2018) and the eigen-decomposition of the covariance matrix (Chen et al., 2020; Zou et al., 2021; Wu et al., 2022a;b). Our work closely relates to the studies in Zou et al. 2021; Wu et al. 2022b that also characterized the SGD dynamic in linear regression with respect to the full eigenspectrum of the data covariance matrix. However, they focused on either the single-task setting or the pretraining-finetuning setting, while we studied the more challenging continual learning problem that involves a sequence of tasks with different data distributions. More discussion in Section 4.

Theoretical Studies in Continual Learning Although significant progress has been made in empirical studies addressing the issue of forgetting in continual learning, theoretical insights into this area are still largely unexplored. In this context, Bennani et al. 2020 established a theoretical framework to study continual learning algorithms in the NTK regime, and provided the first generalization bound dependent on task similarity for SGD and OGD. Doan et al. 2021 introduced the NTK overlap matrix as a task similarity metric and proposed a data-structure-informed variant of OGD that utilizes Principal Component Analysis (PCA). Asanuma et al. 2021 utilized the teacher-student framework on a single neural network and demonstrated that catastrophic forgetting can be circumvented when the similarity among input distributions is small and the similarity among teacher networks is large. Lee et al. 2021 expanded an earlier analysis of two-layer networks within the teacher-student setup to the setting with multiple teachers and revealed that the highest level of forgetting occurs when tasks have intermediate similarity with each other. Evron et al. 2022; Swartworth et al. 2023 explained the behavior of forgetting in the linear regression model from the perspectives of alternating projections and

the Kaczmarz method (Karczmarz, 1937). Lin et al. 2023 investigated the impact of overparameterization, task similarity, and task ordering on forgetting and generalization in the overparameterized linear regression model.

The works most relevant to our study include (Evron et al., 2022; Lin et al., 2023), both of which also studied the behavior of forgetting in the linear regression model. However, our work differs from their studies in several aspects.

Firstly, with regard to assumptions, Evron et al. 2022 assumed all data are bounded with 1 and the model is noiseless, and Lin et al. 2023 assumed all data are sampled from a Gaussian distribution. In contrast, our assumptions cover more data distributions and are much milder than theirs (see Remark 2 and Section 4 for more details). Secondly, in terms of methods, both Evron et al. 2022 and Lin et al. 2023 analyze the problem of forgetting using the minimum norm solution, which presupposes zero training error—a requirement not necessary in our approach with SGD (see Section 2 for further discussions). Third, Evron et al. 2022; Lin et al. 2023 considered only the overparameterized case where the data dimension is larger than the data size, while our analysis holds for both the underparameterized and overparameterized settings.

Notations: In this paper, we adhere to a consistent notation style for clarity. We use boldface lower letters such as $\mathbf{x}, \mathbf{w}$ for vectors, and boldface capital letters (e.g. $\mathbf{A}, \mathbf{H}$) for matrices. Let $\|\mathbf{A}\|_2$ denote the spectral norm of $\mathbf{A}$ and $\|\mathbf{v}\|_2$ denote the Euclidean norm of $\mathbf{v}$. For two vectors $\mathbf{u}$ and $\mathbf{v}$, their inner product is denoted by $\langle \mathbf{u}, \mathbf{v} \rangle$ or $\mathbf{u}^\top \mathbf{v}$. For two matrices $\mathbf{A}$ and $\mathbf{B}$ of appropriate dimension, their inner product is defined as $\langle \mathbf{A}, \mathbf{B} \rangle := \mathrm{tr}(\mathbf{A}^\top \mathbf{B})$. For a positive semi-definite (PSD) matrix $\mathbf{A}$ and a vector $\mathbf{v}$ of appropriate dimension, we write $\|\mathbf{v}\|_{\mathbf{A}}^2 := \mathbf{v}^\top \mathbf{A} \mathbf{v}$. The outer product is denoted by $\otimes$.

2. Preliminaries

In our setup, we consider a sequence of tasks, denoted as $\mathbb{M} = \{1, 2, \ldots, M\}$. For each task m in this sequence, we have a corresponding dataset D_m, which consists of N data points. Each of these data points, denoted as $(\mathbf{x}_{m,i}, y_{m,i})$, is drawn independently and identically distributed (i.i.d.) from a specific distribution $\mathcal{D}_m = \mathcal{X}_m \times \mathcal{Y}_m \subset \mathbb{R}^d \times \mathbb{R}$. Here, $\mathbf{x}_{m,i}$ represents the feature vector, and $y_{m,i}$ is the response variable for each data point in the dataset D_m. Assume that $\{(\mathbf{x}_{m,i}, y_{m,i})\}_{i=1}^N$ are i.i.d. sampled from a linear regression model, i.e., each pair $(\mathbf{x}_{m,i}, y_{m,i})$ is a realization of the linear regression model $y_m = (\mathbf{x}_m^\top \mathbf{w}_*) + z_m$, where z_m is some randomized noise and $\mathbf{w}_* \in \mathbb{R}^d$ is the optimal model parameter.

Our goal is to output a model $\mathbf{w}_{MN}$ minimizing the degree

of *forgetting* (Evron et al., 2022) for M tasks, i.e.

$$G(M) = \frac{1}{M} \sum_{m=1}^M \mathcal{L}_m(\mathbf{w}_{MN}), \quad \text{where} \qquad (1)$$

$$\mathcal{L}_m(\mathbf{w}) = \frac{1}{2} \mathbb{E}_{(\mathbf{x}_m, y_m) \sim \mathcal{D}_m} \|\mathbf{x}_m^\top \mathbf{w} - y_m\|^2, \quad m \in \mathbb{M}$$

$\mathbf{w}_{MN}$ represents the final output after sequentially training on M tasks, each updated via SGD over N iterations for each task. Equation (1) quantifies an average excess population risk on the final output $\mathbf{w}_{MN}$ across all tasks. For each task m, the loss $\mathcal{L}_m$ evaluate how well $\mathbf{w}_{MN}$ performs on it, thus assessing the degree of the model's forgetting on previous tasks in continual learning scenarios.

Definition 2.1 (Data Covariance). Assume that each entry and the trace of the $\mathbb{E}[\mathbf{x}_m \mathbf{x}_m^\top]$ are finite. Define $\mathbf{H}_m := \mathbb{E}[\mathbf{x}_m \mathbf{x}_m^\top]$ as data covariance matrix.

Let $\mathbf{H}_m$ denote the eigen decomposition of the data covariance for task m, given by $\mathbf{H}_m = \sum_i \lambda_m^i \mathbf{v}_m^i \mathbf{v}_m^{i \top}$, where $(\lambda_m^i)_{i \geq 1}$ are eigenvalues in a nonincreasing order and $(\mathbf{v}_m^i)_{i \geq 1}$ are the corresponding eigenvectors. Define $\mathbf{H}_{m,k_1:k_2}$ as $\mathbf{H}_{m,k_1:k_2} := \sum_{k_1 < i \leq k_2} \lambda_m^i \mathbf{v}_m^i \mathbf{v}_m^{i \top}$, and allow $k_2 = \infty$ to imply that $\mathbf{H}_{m,k:\infty} = \sum_{i > k} \lambda_m^i \mathbf{v}_m^i \mathbf{v}_m^{i \top}$.

Definition 2.2 (Covariate Shift). For each task m, the covariates $\mathbf{x}_{m,1}, \ldots, \mathbf{x}_{m,N}$ are i.i.d. drawn from $\mathcal{D}_m$.

Compared to the concept of covariate shift in transfer learning (Pathak et al., 2022), Definition 2.2 provides a more general scenario applicable to a series of tasks $M \geq 2$. For simplicity, in our analysis, we assume that each task m in our model consists of N data points, differentiating it from transfer learning approaches that typically consider the total dataset size as N.

Assumption 2.3 (Fourth moment conditions). Assume that for each task m, the expected fourth moment of covariates, denoted as $\mathcal{M} := \mathbb{E}[\mathbf{x}_m \otimes \mathbf{x}_m \otimes \mathbf{x}_m \otimes \mathbf{x}_m]$, and the expected covariance matrix $\mathbf{H}_m$ are finite. Moreover:

(A) There exists a constant $\alpha_m > 0$ such that for any Positive Semi-Definite (PSD) matrix $\mathbf{A}$, the following holds:

$$\mathbb{E}[\mathbf{x}_m \mathbf{x}_m^\top \mathbf{A} \mathbf{x}_m \mathbf{x}_m^\top] \preceq \alpha_m \cdot \mathrm{tr}(\mathbf{H}_m \mathbf{A}) \mathbf{H}_m.$$

(B) There exists a constant $\beta_m > 0$, such that for every PSD matrix $\mathbf{A}$, the following holds:

$$\mathbb{E}[\mathbf{x}_m \mathbf{x}_m^\top \mathbf{A} \mathbf{x}_m \mathbf{x}_m^\top] - \mathbf{H}_m \mathbf{A} \mathbf{H}_m \succeq \beta_m \cdot \mathrm{tr}(\mathbf{H}_m \mathbf{A}) \mathbf{H}_m.$$

Remark 1. Assumption 2.3 is a commonly employed assumption in the linear regression analysis utilizing SGD methods (Zou et al., 2021; Wu et al., 2022a;b), which is

much weaker than the assumptions on the aforementioned related work. Specifically, it can be verified that Assumption 2.3 holds with $\alpha_m = 3$ and $\beta_m = 1$ for Gaussian distribution discussed in (Asanuma et al., 2021; Lee et al., 2021; Lin et al., 2023). Additionally, Assumption 2.3(A) can be relaxed to $\mathbb{E}\|\mathbf{x}_m\|_2^2 \leq \alpha_m \operatorname{tr}(\mathbf{H}_m)$ with $\mathbf{A} = \mathbf{I}$, where $\alpha_m \operatorname{tr}(\mathbf{H}_m) = 1$ is assumed in Evron et al. 2022.

Assumption 2.4 (Well-specified noise). Assume that for each distribution of task m, the response (conditional on input covariates) is given by $y_m = \mathbf{x}_m^\top \mathbf{w}^* + z_m$, where $z_m \sim \mathcal{N}(0, \sigma^2)$ and z_m is independent with $\mathbf{x}_m$.

Similar to previous works, we assume that z_m is some randomized noise that satisfies $\mathbb{E}[z_m|\mathbf{x}] = 0$ and $\mathbb{E}[z_m^2] = \sigma^2$ for each task m.

Continual Learning via SGD Suppose we train the model parameter $\mathbf{w}$ sequentially. Let $\mathbf{w}_{(m-1)N+N}$ represent the parameter state after the completion of training on task m, which also serves as the initial condition for the training of task $m+1$. Starting with $\mathbf{w}_0$ and employing a constant step size η, the model is updated by SGD for each task $m \in \mathbb{M}$ over N iterations, with $t = 1, \ldots, N$:

$$\mathbf{w}_{(m-1)N+t} = \mathbf{w}_{(m-1)N+t-1} - \eta \cdot \mathbf{g}_{m,t}, \quad \text{and}$$
$$\mathbf{g}_{m,t} := (\mathbf{x}_{m,t}^\top \mathbf{w}_{(m-1)N+t-1} - y_{m,t})\mathbf{x}_{m,t}, \tag{2}$$

where $\mathbf{g}_{m,t}$ represents the gradient of the loss function at task m and iteration t for a given data point $(\mathbf{x}_{m,t}, y_{m,t})$.

Contrastingly, the minimum norm solution in linear regression, particularly relevant in overparameterized settings, aims to find a weight vector $\mathbf{w}$ that not only achieves zero training error but also possesses the minimal possible norm. Here, $\mathbf{w}_m$ represents the outcome post-training for task m, and it also serves as the starting point for training task $m+1$. The objective, beginning from an initial condition $\mathbf{w}_0 = \mathbf{0}$, is defined by the following optimization problem:

$$\min_{\mathbf{w}} \|\mathbf{w} - \mathbf{w}_{m-1}\|_2, \quad \text{s.t. } (\mathbf{X}_m)^\top \mathbf{w} = \boldsymbol{y}_m,$$

where $\mathbf{X}_m := [\mathbf{x}_{m,1}, \ldots, \mathbf{x}_{m,N}] \in \mathbb{R}^{d \times N}$ and $\boldsymbol{y}_m = [y_{m,1}, \ldots, y_{m,N}] \in \mathbb{R}^{1 \times N}$. The update rules for each iteration follow as:

$$\mathbf{w}_m = \mathbf{w}_{m-1} + \mathbf{X}_m(\mathbf{X}_m^\top \mathbf{X}_m)^{-1}(\boldsymbol{y}_m - \mathbf{X}_m^\top \mathbf{w}_{m-1}), \tag{3}$$

where highlights the computational intensity of inverting the matrix $(\mathbf{X}_m^\top \mathbf{X}_m)^{-1}$. This is particularly challenging for large datasets or overparameterized feature spaces. Unlike the minimum norm solution, SGD does not assume the existence of a unique, exact solution and is more adaptable to a variety of problems, including those with non-linear dynamics.

3. Main Results

Before presenting our upper bound, we shall establish the following notations to facilitate comprehension of the results.

$$\begin{cases} \Gamma^i_{(p,q)} := \prod_{j=p}^q (1 - \eta \lambda_j^i)^{2N}, \quad \boldsymbol{\Gamma}^q_p := \prod_{j=p}^q (\mathbf{I} - \eta \mathbf{H}_j)^{2N}, \\ \\ \mathbf{U}_{k_m^*} := \mathbf{I}_{m,0:k_m^*} + N\eta \mathbf{H}_{m,k_m^*:\infty}, \quad \Lambda^i := \sum_{m=1}^M \lambda_m^i, \end{cases} \tag{4}$$

where $(\lambda_m^i)_{i \geq 1}$ are eigenvalues of $\mathbf{H}_m$ in a nonincreasing order and $k_m^* = \max\{i : \lambda_m^i \geq \frac{1}{N\eta}\}$ represents the cut-off index for $\mathbf{H}_m$. Here, $\Gamma^i_{(p,q)}$ and $\boldsymbol{\Gamma}^q_p$ can be regarded as a projection accumulation from task p to task q, and basically capture the impact of the learning dynamic of previous tasks on the subsequent task. $\mathbf{U}_{k_m^*}$ is defined with respect to the cut-off index k_m^* for each task's data covariance matrix $\mathbf{H}_m$ that captures both the dominant eigenvalues and the tail of the spectrum, and Λ^i denotes the sum of the i-th eigenvalue across all tasks.

In the following, we first provide our upper bound for the behavior of forgetting via SGD in the linear regression model.

Theorem 3.1 (Upper Bound). *Consider a scenario where the model $\mathbf{w}$ undergoes training via SGD for M distinct tasks, following a sequence $1, \ldots, M$. With a constant step size of $\eta \leq 1/R^2$ given that $R^2 = \max\{\alpha_m \operatorname{tr}(\mathbf{H}_m)\}_{m=1}^M$, each task m is executed for N iterations. Given that Assumptions (A) and 2.4 are satisfied, the following will hold:*

$$G(M) \leq err_{var} + err_{bias},$$

where the variance and bias errors are upper-bounded by

$$err_{var} \leq \frac{\sum_{m=1}^M}{M} \cdot \frac{\eta \sigma^2}{(1 - \eta R^2)} \cdot D_1^{eff},$$

$$err_{bias} \leq \frac{\sum_{k=1}^M}{M} \|\mathbf{w}_0 - \mathbf{w}^*\|_{\boldsymbol{\Gamma}_1^M \mathbf{H}_k}^2$$
$$+ \frac{\sum_{m=1}^M \frac{2\alpha_m \eta^2 \cdot (D_2^{eff} + \Phi_1^{m-1} D_3^{eff})}{1 - \eta \alpha_m \operatorname{tr}(\mathbf{H}_m)} \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{U}_{k_m^*}}^2}{M}$$
$$+ \frac{\sum_{m=1}^M}{M} \alpha_m \eta \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|_{\boldsymbol{\Gamma}_1^M \mathbf{H}_k(\mathbf{H}_m + \Phi_1^{m-1}\mathbf{I}) \cdot \mathbf{U}_{k_m^*}}^2,$$

where the effective dimensions are given by

$$D_1^{eff} := \sum_{i < k_m^*} \Gamma^i_{(m+1,M)} \Lambda^i + N\eta \sum_{i > k_m^*} \Gamma^i_{(m+1,M)} \lambda_m^i \Lambda^i$$
$$D_2^{eff} := \sum_{i < k_m^*} \Gamma^i_{(1,M)} (\lambda_m^i)^2 \Lambda^i + N\eta \sum_{i > k_m^*} \Gamma^i_{(1,M)} (\lambda_m^i)^3 \Lambda^i$$
$$D_3^{eff} := \sum_{i < k_m^*} \Gamma^i_{(m,M)} (\lambda_m^i) \Lambda^i + \eta N \sum_{i > k_m^*} \Gamma^i_{(m,M)} (\lambda_m^i)^2 \Lambda^i,$$

$$\tag{5}$$

with $k_m^*, \Gamma^i_{(p,q)}$ and $\boldsymbol{\Gamma}^q_p$ defined as in Equation (4) and denoting $\Phi_1^{m-1} := \sum_{j=1}^{m-1} \prod_{k=1}^j \alpha_k \eta^j \cdot \langle \mathbf{H}_{k-1}, \mathbf{I} - (\mathbf{I} - \eta \mathbf{H}_{m-1})^N \rangle \cdot \langle \mathbf{H}_j, \mathbf{H}_m \rangle.$

In Theorem 3.1, we establish an upper bound on the forgetting behavior of a model trained using SGD in the continual learning with various data distribution settings. It highlights that the model's performance is influenced by both err_{var} and err_{bias}, where err_{var} stems from the inherent noise intrinsic to the model itself and err_{bias} represents the bias associated with the initial value during the learning process. Notice that both of them are determined jointly by the spectrum of the covariance matrices as well as the stepsizes for continual learning.

To provide a more intuitive explanation, we explore a simplified scenario by setting $\eta = 0$. Specifically, this setting simplifies our analysis by reducing the error terms to only the first term in bias error, which appears to depend solely on the initial weight $\mathbf{w}_0$ and the data. However, this simplification might misleadingly imply that a minimal η would result in optimal learning outcomes. A crucial aspect overlooked in this interpretation is the role of the projection term $\Gamma_1^M = \prod_{j=1}^M (\mathbf{I} - \eta \mathbf{H}_j)^{2N}$, which becomes an identity matrix $\mathbf{I}$ when $\eta = 0$. Thus, while setting $\eta = 0$ eliminates other error terms, it also exacerbates the first term of bias error, potentially making it the most significant error contributor. Consequently, there exists a trade-off in choosing the step size.

The subsequent theorem presents a nearly matching lower bound.

Theorem 3.2 (Lower Bound). *Consider a scenario where the model $\mathbf{w}$ undergoes training via SGD for M distinct tasks, following a sequence $1, \ldots, M$. With a constant step size of $\eta \leq 1/R^2$ given that $R^2 = \max\{\alpha_m \operatorname{tr}(\mathbf{H}_m)\}_{m=1}^M$, each task m is executed for N iterations. Given that Assumptions (B) and 2.4 are satisfied, the following will hold:*

$$G(M) \geq \text{err}_{var} + \text{err}_{bia},$$

where the variance and bias errors are lower bounded by

$$\text{err}_{var} \geq \frac{\sum_{m=1}^M}{M} \cdot \frac{9\eta^2\sigma^2}{20} \cdot D_1^{eff},$$

$$\text{err}_{bias} \geq \frac{\sum_{k=1}^M}{M} \|\mathbf{w}_0 - \mathbf{w}^*\|_{\Gamma_1^M \mathbf{H}_k}^2$$
$$+ \frac{\sum_{m=1}^M}{M} \cdot \frac{\beta_m^2 \eta^2}{25} \cdot (D_2^{eff} + \hat{\Phi}_1^{m-1} D_3^{eff}) \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{U}_{k_m^*}}^2$$
$$+ \frac{\sum_{m=1}^M}{M} \frac{\beta_m \eta^2}{5} \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|_{(\mathbf{I}-\eta\mathbf{H}_m)^{2N}\Gamma_1^M \mathbf{H}_k(\mathbf{H}_m + \hat{\Phi}_1^{m-1}\mathbf{I}) \cdot \mathbf{U}_{k_m^*}}^2$$

where the effective dimensions k_m^*, $\Gamma_{(p,q)}^i$ and Γ_p^q are the same as in Theorem 3.1, and $\hat{\Phi}_1^{m-1} := \sum_{j=1}^{m-1} \prod_{k=1}^j \beta_k (\frac{\eta}{2})^j \cdot \langle \mathbf{H}_{k-1}, (\mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_{m-1})^{2N}) \rangle \cdot \langle \mathbf{H}_j, \mathbf{H}_m \rangle$.

Analogous to the Theorem 3.1, our lower bound also consists of the bias term and the variance term. It is noteworthy that our lower bound is tight with the upper bound in terms of variance term, differing only by absolute constants. Additionally, our lower bound closely matches the upper bound

in terms of the bias term, with some differences arising from the following quantities

$$\hat{\Phi}_1^{m-1} \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{U}_{k_m^*}}^2, \quad \|\mathbf{w}_0 - \mathbf{w}^*\|_{(\mathbf{I}-\eta\mathbf{H}_m)^{2N}}.$$

Specifically, $\hat{\Phi}_1^{m-1}$ here differs from Φ_1^{m-1} in Theorem 3.1 only by a factor of constants (i.e. α_k and β_k defined in Assumption 2.3). The term $\|\mathbf{w}_0 - \mathbf{w}^*\|_{(\mathbf{I}-\eta\mathbf{H}_m)^{2N}}$ has a different subscript of $(\mathbf{I} - \eta\mathbf{H}_m)^{2N}$ compared to that of the upper bound. Nevertheless, it can be regarded as a part of the projection accumulation Γ_1^M that exists in the subscript of both results simultaneously.

More importantly, we show that the upper and lower bounds converge, ignoring constant factors, under the conditions

$$\|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{U}_{k_m^*}}^2 \lesssim \sigma^2, \quad \hat{\Phi}_1^{m-1} \lesssim O(1),$$

which can be satisfied that the signal-to-noise ratios $\|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{U}_{k_m^*}}^2 / \sigma^2$ is bounded and the step size is appropriate small.

4. Discussion

Building on Theorem 3.1 and Theorem 3.2, we aim to offer a more comprehensive understanding of our findings from three key perspectives: 1) Technical Understanding Under Simplified Cases; 2) Comparison with Existing Work; 3) The Impact of Task Ordering and Parameters on Forgetting.

4.1. Technical Understanding Under Simplified Cases

In this section, we demonstrate how to achieve a vanishing bound in the overparameterized regime.

Based on Theorem 3.1, we consider a scenario where $\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2, \sigma^2 \lesssim 1$ and $\operatorname{tr}(\mathbf{H}_m) \simeq 1$ for each task m, implying a rapid decay in the spectrum of $\mathbf{H}_m$. To obtain a vanishing bound in the overparameterized regime, the effective dimension should hold that

$$D_1^{\text{eff}} \simeq D_3^{\text{eff}} = o(\frac{MN}{e^{(M-m)}}),$$
$$D_2^{\text{eff}} = o(\frac{MN}{e^M}). \tag{6}$$

To meet the condition in Equation (6), for each task $\tilde{m}$, let $k^\dagger = \min\{k_m^*, k_{\tilde{m}}^*\}$ and $k^\star = \max\{k_m^*, k_{\tilde{m}}^*\}$. It necessarily holds that

$$\sum_{i<k^\star} \lambda_{\tilde{m}}^i \simeq \sum_{i<k^\star} \lambda_m^i \lambda_{\tilde{m}}^i \simeq \sum_{i<k^\star} (\lambda_m^i)^2 \lambda_{\tilde{m}}^i = o(N),$$
$$\sum_{i>k^\dagger} \lambda_{\tilde{m}}^i \lambda_m^i \simeq \sum_{i>k^\dagger} (\lambda_m^i)^2 \lambda_{\tilde{m}}^i \simeq \sum_{i>k^\dagger} (\lambda_m^i)^3 \lambda_{\tilde{m}}^i = o(\frac{1}{N}). \tag{7}$$

To clarify Equation (7), let notice the crucial cut-off index $k^\star$ and $k^\dagger$, which divide the entire feature space into two

$k^\star$-dimensional and $k^\dagger$-dimensional subspaces. For achieving a diminishing bound in overparameterized setting, it is necessary that the sum of eigenvalues for indices less than $k^\star$, denoted as $\sum_{i<k^\star}$, should be $o(N)$, and the sum of the tail eigenvalues for indices greater than $k^\dagger$, $\sum_{i>k^\dagger}$, should be $o(\frac{1}{N})$. These conditions are typically met when the dataset size N is sufficiently large, or when a smaller step size η is chosen dependent on N. Additionally, We note that the condition in Equation (7) can be relaxed. In light of the definition of k_m^*, the eigenvalues for task $\widetilde{m}$ are truncated based on the following two scenarios: **1**) $k_m^* < k_{\widetilde{m}}^*$: Here, the cut-off for task $\widetilde{m}$ occurs earlier, resulting in an additional $(k_{\widetilde{m}}^* - k_m^*)$ dimensions of eigenvalues such that $\lambda_{\widetilde{m}}^i \geq 1/(N\eta)$. To achieve a diminishing bound under this condition, it is necessary that $\sum_{k_m^* \leq i \leq k_{\widetilde{m}}^*} \lambda_{\widetilde{m}}^i = o(N)$. **2**) $k_m^* \geq k_{\widetilde{m}}^*$: In this case, the cut-off for task $\widetilde{m}$ occurs later, involving an additional $(k_m^* - k_{\widetilde{m}}^*)$ dimensions of eigenvalues where $\lambda_{\widetilde{m}}^i \leq 1/(N\eta)$, achieving the same results.

In the under-parameterized regime, we even account for the worst-case scenario where $\lambda_m^i \geq \frac{1}{N\eta}$ for all index i and task m, leading to a bound of $D_1^{\text{eff}} \simeq D_3^{\text{eff}} = o(\frac{Md\lambda_m^1}{e^{(M-m)}})$, $D_2^{\text{eff}} = o(\frac{Md\lambda_m^1}{e^M})$.

4.2. Comparison with Existing work

In this section, we will first explore the challenges and parallels between traditional/transfer learning and continual learning. Secondly, we examine how restrictive assumptions in previous studies might overshadow the impact of key factors, thereby affecting the overall understanding of forgetting in continual learning.

Our results reveal that compared to traditional learning (Zou et al., 2021), which typically involves a single task, and transfer learning (Wu et al., 2022b), which usually incorporates two data distributions, the effective dimension in continual learning scenarios is more complex. Specifically, in our analysis, the term Λ^i arises from a distinct measurement perspective (i.e. *forgetting*), which requires us to consider how the final output aligns with all previously encountered tasks in the continual learning (i.e. $\mathbf{H}_m$ for all m). This is in contrast to both traditional training and transfer learning, where the evaluation metric is uniformly focused on performance against a single dataset (i.e. $\mathbf{H}_M$). Moreover, the multi-task nature of continual learning introduces unique challenges considering the bias iterates and variance iterates, where we refer to the proof in Appendix for more details.

Given that our analysis, similar to theirs, characterizes bounds with the full eigenspectrum of the data covariance matrix, it follows that our derived results match their findings in several aspects: **1**) The cutoff index k_m^* is uniquely determined for each task m in continual learning, akin to

the one in Zou et al. 2021; Wu et al. 2022b, where they identify corresponding indices k_{training}^* and k_{test}^* . **2**) The projection terms $\Gamma_{(p,q)}^i$ and $\mathbf{\Gamma}_p^q$ also occur in transfer learning (Wu et al., 2022b), showing how previous iterations/past learning is projected onto the future updates.

Previous work (Evron et al., 2022) also explored the dynamics of forgetting through the perspective of projection. We first revisit the findings presented by Evron et al. 2022. Considering a scenario where the number of iterations $N = 1$, the update rule in their analysis can be reformulated as follows:

$$\mathbf{w}_m - \mathbf{w}^* = (\mathbf{I} - \eta_m \mathbf{x}_m \mathbf{x}_m^\top)(\mathbf{w}_{m-1} - \mathbf{w}^*), \quad (8)$$

where they incorporate the noiseless model assumption that $y_m = \mathbf{x}_m^\top \mathbf{w}^*$. As a result, the forgetting in Evron et al. 2022 holds that

$$G(M) = \frac{1}{M} \sum_{m=1}^{M} \|\mathbf{x}_m(\mathbf{w}_M - \mathbf{w}^*)\|_2^2, \quad \text{given} \quad \|\mathbf{x}_m\|_2 \leq 1$$

$$\leq \frac{1}{M} \sum_{m=1}^{M} \|(\mathbf{I} - \eta_m \mathbf{x}_M \mathbf{x}_M^\top) \ldots (\mathbf{I} - \eta_m \mathbf{x}_1 \mathbf{x}_1^\top)(\mathbf{w}_0 - \mathbf{w}^*)\|_2^2,$$

indicating that the forgetting dynamic can be determined by the projection of $(\mathbf{I} - \eta_m \mathbf{x}_m \mathbf{x}_m^\top)$, where $\eta_m = \|\mathbf{x}_m\|^{-2}$. However, compared to our analysis, their study exhibits several key differences in comparison to ours. **1**) The inherent model noise: Evron et al. 2022 considers a noiseless model, where results in the absence of an additional iterative term $\mathbf{x}_m \cdot z_m$ related to noise in Equation (8). This omission leads to a lack of accumulative variance error in the evaluation of forgetting performance (i.e. err_{var} in our analysis). It is noteworthy to mention that in numerous learning problems, the variance error often plays a dominant role in the total error (Jain et al., 2018; Zou et al., 2021; Wu et al., 2022b). **2**) The bounded norm $\|\mathbf{x}\|_2$: the assumption of the bounded norm, which omits the interaction with projection effects, is crucial in our analysis as the factor Λ^i in Theorem 3.1 and Theorem 3.2. **3**) Last iterate SGD results: Evron et al. 2022 shows that, with a $\|\mathbf{x}_m\|_2^{-2}$ step size, their worst-case expected forgetting will become a dimension-dependent bound of $O(d/M)$. This analysis, conducted under the overparameterized regime, suggests the occurrence of *catastrophic forgetting*. In contrast, our results, as discussed earlier, offer a different perspective, suggesting the possibility of achieving a vanishing forgetting bound in overparameterized settings with certain conditions met.

It is noticed that Lin et al. 2023 also investigates the relationship between catastrophic forgetting and factors such as task sequence (order) and dimensionality. However, their results will tend to be vacuous in the under-parameterized setting since $(\mathbf{X}_m \mathbf{X}_m^\top)^{-1}$, data matrix for task m, is non-invertible when employing minimum norm solution, as we discussed earlier in Section 2. Due to space constraints, a more extensive discussion will be provided in Appendix D.
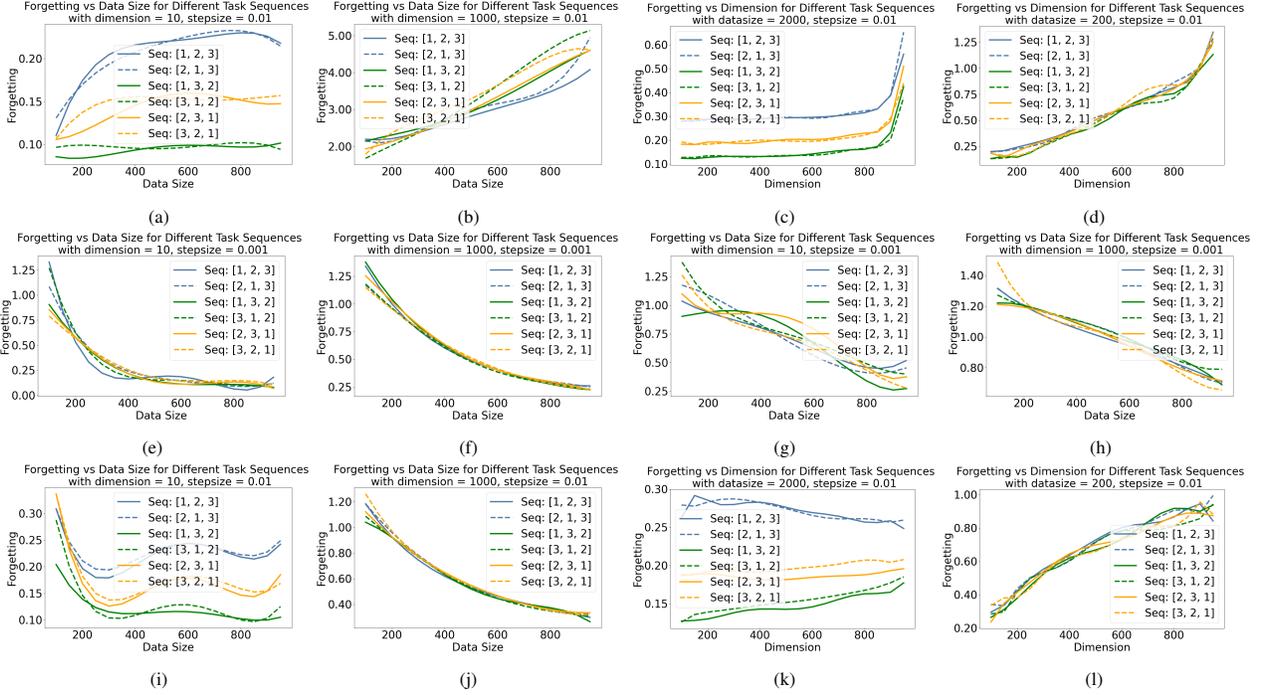
Figure 1: Impact of Task Sequence and Algorithmic Parameters on Forgetting Behavior with Linear Regression Model and Deep Neural Networks. This figure presents the relationship between task sequence order and algorithmic parameters (data size, dimensionality, and step size) on the forgetting behavior observed in linear regression models (figures (a)-(f)) and deep neural networks (figures (g)-(l)). Figures (a), (b), (i), and (j) illustrate how varying data sizes impact forgetting behavior for different task sequences, while figures (c), (d), (k), and (l) demonstrate the effect of changing dimensionality on forgetting. Lastly, Figures (e)-(h) demonstrate the influence of stepsize on the rate of forgetting across different model configurations.

4.3. The Impact of Task Ordering and Parameters on Forgetting

In the upcoming discussion, we will present theoretical insights derived from our results.

Notice that the bounds in Theorem 3.1 and Theorem 3.2 contain two crucial factors: the effective dimension D^{eff} and the covariance accumulation $\hat{\Phi}_1^{m-1}/\Phi_1^{m-1}$. We first discuss the effective dimension. Each D^{eff} is consist of a projection term $\Gamma^i_{(m,M)}$ and the eigenvalues λ^i_m, with Λ^i serving as the constant. It can be observed that when data size N approaches infinity, the projection term converges to $\frac{1}{e^{M-m}}$, implying that the eigenvalue will predominantly dictate the larger effective dimension with respect to $\frac{\lambda^i_m}{e^{M-m}}$. This observation highlights the substantial influence of eigenvalues on task sequence in continual learning. Specifically, it shows that *when data size is sufficiently large, task sequences organized in a way, where tasks associated with larger eigenvalues in their population data covariance matrix are trained later, exhibit more forgetting*. Additionally, if the step size is appropriately small, the projection term stabilizes to a constant of less than 1, leading to similar outcomes as in the first scenario. It is noteworthy that these insights can not be derived from the existing work analysis

due to their restrictive assumptions, such as Gaussian data distribution in Lee et al. 2021; Asanuma et al. 2021; Lin et al. 2023 and minimum norm solution in Evron et al. 2022; Lin et al. 2023; Swartworth et al. 2023.

The covariance accumulation term, $\hat{\Phi}_1^{m-1}/\Phi_1^{m-1}$, which includes the covariance matrices $\mathbf{H}_{j \leq m}$ and the step size η, plays a crucial role in demonstrating how previously acquired information is retained and influences the model's adaptability to new tasks. Notably, there is an interesting contradiction in the optimal accumulation order within $\hat{\Phi}_1^{m-1}/\Phi_1^{m-1}$ compared to the projection term in $\Gamma^i_{(m,M)}$. Specifically, earlier occurrence of $\mathbf{H}_j$ with larger expected eigenvalues tends to increase the degree of forgetting. Meanwhile, an important observation is that if the step size is sufficiently small, the impact of the covariance accumulation term becomes less significant. This interplay between the effective dimension and covariance accumulation elucidates the complexities inherent in continual learning scenarios.

5. Empirical Stimulation

In this section, we conduct experiments using synthetic data to validate our theoretical results and shed light on the intricate interplay between *eigenvalues*, *step size*, and

dimensionality.

Experimental Setup In our study, we designed three distinct tasks, denoted as Tasks 1,2, and 3, each with a different feature space. During the initial simulations, the eigenvalues for the feature values of Tasks 1, 2, and 3 were set according to $\lambda_i = i^{-3}, \lambda_i = i^{-2}$, and $\lambda_i = i^{-1}$ respectively. To mimic real-world data imperfections, Gaussian noise with a standard deviation of 0.1 was added to the labels. We assessed the impact of task sequence on the model's tendency to forget by evaluating six different task orders: [1,2,3], [2,1,3], [1, 3, 2], [3, 1, 2], [2, 3, 1], and [3, 2, 1].

5.1. Linear Regression

Training and Evaluation For this experiment, a linear regression model was trained using Stochastic Gradient Descent (SGD) with a learning rate of 0.01 or 0.001. The model was tested in both low-dimensional (10 input features) and high-dimensional (1000 input features) settings. Each task sequence underwent training with various data sizes, ranging from 100 to 950 in increments of 50, and each task was trained for five epochs. The performance of the model was evaluated on each task to calculate the average excess risk (Equation (1)), quantifying the degree of forgetting the model experienced.

Impact of Eigenvalue Sequencing The observations from Figure 1a and Figure 1c reveal the significant impact of eigenvalue sequencing on forgetting behavior in the under-parameterzied regime. Notably, task sequences that are arranged such that tasks with larger eigenvalues (i.e. Task 3 in our case, characterized by $\lambda_i = i^{-1}$) are trained later in the learning process tend to result in increased forgetting. This empirical finding aligns well with our theoretical analysis (the term $\frac{\lambda_m^i}{e^{M-m}}$ discussed in Section 4.3). In an under-parameterized setting, or when the eigenvalues decay rapidly, the effective dimension — crucial in determining the model's forgetting performance - is largely influenced by the eigenvalues. Such a pattern is intuitive as when tasks with larger eigenvalues are trained later, the model might overfit these tasks due to their high variance.

Impact of Dimensionality Our results, depicted in Figure 1c and Figure 1d, show that in under-parameterized scenarios, performance remains relatively unaffected by an increase in dimensionality. However, in over-parameterized settings, the model tends to exhibit increased forgetting as dimensionality rises, particularly when the data size is kept constant. This highlights the varying impact of dimensionality on model performance in different parameterization contexts. In higher-dimensional settings, the influence of the projection term $\Gamma_{(p,q)}^i$, as shown in Theorem 3.1, diminishes in comparison to the impact of Λ^i and λ_i. Consequently, as the number of features in the model increases, the sequence

in which tasks are presented becomes less significant in determining the model's forgetting behavior. This shift implies that, in high-dimensional scenarios, the inherent complexity and the distribution of eigenvalues of the feature space play a more critical role than the sequence of tasks, influencing the model's learning and retention capabilities.

Impact of Step-size Our results, depicted in Figure 1e and Figure 1f, reveal that a smaller step size effectively reduces forgetting in various task sequences and across different dimensionalities. This trend is especially noticeable in high-dimensional feature spaces, where a reduced step size markedly lowers the rate of forgetting. This observation is in line with the theoretical insights provided in Theorem 3.1 and Theorem 3.2, as smaller step sizes may lead to more refined updates during training, allowing the model to incrementally adjust to new tasks while preserving knowledge from previous ones.

5.2. Implication on DNNs

Intriguingly, our next discussion will adopt the same data generation and task setup as outlined in Section 5.1, but shift our focus to a different Neural Network model. This model comprises an input layer, a hidden layer with ten neurons, and an output layer, and it undergoes a training process akin to that of linear regression.

Impact of Eigenvalue Sequencing In our studies with Deep Neural Networks (DNNs), we still find that task sequences, ending with tasks having larger eigenvalues, tend to exhibit increased forgetting, especially in under-parameterized settings, similar to linear regression models. This indicates that the tendency of overfitting observed in linear models, particularly when tasks with larger eigenvalues are trained later in the sequence, may occur in DNNs as well.

Impact of Dimensionality Our results also reveal the consistent behaviors between DNNs and linear regression concerning dimensionality. In under-parameterized scenarios (Figure 1k), forgetting remains stable despite increased dimensionality, while in over-parameterized settings (Figure 1l), higher dimensionality leads to more forgetting when data size is fixed. However, the adverse effects of higher dimensions can be alleviated by expanding the dataset size, as demonstrated in Figure 1j. It is a notable contrast to linear regression, which suggests that the complex structures of DNNs are better suited to manage and learn from high-dimensional data in continual learning scenarios. The different behaviors observed between DNNs and linear regression models will be a potentially interesting direction for future work.

Impact of Step Size Our results, depicted in Figure 1e and Figure 1f, indicate that in under-parameterized settings, a smaller step size significantly lessens the influence of

task sequences on forgetting, while in models with high-dimensional features, forgetting can be mitigated even without adjusting the step size.

6. Conclusion

In this work, we contribute to the understanding of catastrophic forgetting in continual learning via a multi-step SGD algorithm. Our theoretical analysis establishes bounds that illustrate the impact of various factors on forgetting such as data covariance matrix spectrum, step size, data size, and dimensionality, which can not be fully captured in previous studies due to their restrictive assumptions. This theoretical understanding is further substantiated through simulations conducted in linear regression models and Deep Neural Networks, which corroborate our theoretical insights.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Acknowledgments

References

Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., and Tuytelaars, T. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 139–154, 2018.

Asanuma, H., Takagi, S., Nagano, Y., Yoshida, Y., Igarashi, Y., and Okada, M. Statistical mechanical analysis of catastrophic forgetting in continual learning with teacher and student networks. *Journal of the Physical Society of Japan*, 90(10):104001, 2021.

Bennani, M. A., Doan, T., and Sugiyama, M. Generalisation guarantees for continual learning with orthogonal gradient descent. *arXiv preprint arXiv:2006.11942*, 2020.

Chaudhry, A., Ranzato, M., Rohrbach, M., and Elhoseiny, M. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*, 2018.

Chen, X., Liu, Q., and Tong, X. T. Dimension independent generalization error by stochastic gradient descent. *arXiv preprint arXiv:2003.11196*, 2020.

Cortes, C. and Mohri, M. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519:103–126, 2014.

Cortes, C., Mohri, M., and Medina, A. M. Adaptation based on generalized discrepancy. *The Journal of Machine Learning Research*, 20(1):1–30, 2019.

Défossez, A. and Bach, F. Averaged least-mean-squares: Bias-variance trade-offs and optimal sampling distributions. In *Artificial Intelligence and Statistics*, pp. 205–213. PMLR, 2015.

Dieuleveut, A., Flammarion, N., and Bach, F. Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research*, 18(1): 3520–3570, 2017.

Doan, T., Bennani, M. A., Mazoure, B., Rabusseau, G., and Alquier, P. A theoretical analysis of catastrophic forgetting through the ntk overlap matrix. In *International Conference on Artificial Intelligence and Statistics*, pp. 1072–1080. PMLR, 2021.

Evron, I., Moroshko, E., Ward, R., Srebro, N., and Soudry, D. How catastrophic can catastrophic forgetting be in linear regression? In *Conference on Learning Theory*, pp. 4028–4079. PMLR, 2022.

Farajtabar, M., Azizan, N., Mott, A., and Li, A. Orthogonal gradient descent for continual learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3762–3773. PMLR, 2020.

Hanneke, S. and Kpotufe, S. On the value of target data in transfer learning, 2020.

Hao, J., Ji, K., and Liu, M. Bilevel coreset selection in continual learning: A new formulation and algorithm. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Jain, P., Kakade, S. M., Kidambi, R., Netrapalli, P., Pillutla, V. K., and Sidford, A. A markov chain theory approach to characterizing the minimax optimality of stochastic gradient descent (for least squares). *arXiv preprint arXiv:1710.09430*, 2017.

Jain, P., Kakade, S., Kidambi, R., Netrapalli, P., and Sidford, A. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. *Journal of machine learning research*, 18, 2018.

Karczmarz, S. Angenaherte auflosung von systemen linearer glei-chungen. *Bull. Int. Acad. Pol. Sic. Let., Cl. Sci. Math. Nat.*, pp. 355–357, 1937.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Kpotufe, S. and Martinet, G. Marginal singularity, and the benefits of labels in covariate-shift. In *Conference On Learning Theory*, pp. 1882–1886. PMLR, 2018.

Lee, S., Goldt, S., and Saxe, A. Continual learning in the teacher-student setup: Impact of task similarity. In *International Conference on Machine Learning*, pp. 6109–6119. PMLR, 2021.

Lin, S., Yang, L., Fan, D., and Zhang, J. Trgp: Trust region gradient projection for continual learning. *arXiv preprint arXiv:2202.02931*, 2022.

Lin, S., Ju, P., Liang, Y., and Shroff, N. Theory on forgetting and generalization of continual learning. *arXiv preprint arXiv:2302.05836*, 2023.

Liu, H. and Liu, H. Continual learning with recursive gradient optimization. *arXiv preprint arXiv:2201.12522*, 2022.

Ma, C., Pathak, R., and Wainwright, M. J. Optimally tackling covariate shift in rkhs-based nonparametric regression, 2023.

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

Mohri, M. and Medina, A. M. New analysis and algorithm for learning with drifting distributions, 2012.

Pan, S. J. and Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10): 1345–1359, 2009.

Pathak, R., Ma, C., and Wainwright, M. A new similarity measure for covariate shift with applications to nonparametric regression. In *International Conference on Machine Learning*, pp. 17517–17530. PMLR, 2022.

Riemer, M., Cases, I., Ajemian, R., Liu, M., Rish, I., Tu, Y., and Tesauro, G. Learning to learn without forgetting by maximizing transfer and minimizing interference. *arXiv preprint arXiv:1810.11910*, 2018.

Saha, G., Garg, I., and Roy, K. Gradient projection memory for continual learning. *arXiv preprint arXiv:2103.09762*, 2021.

Serra, J., Suris, D., Miron, M., and Karatzoglou, A. Overcoming catastrophic forgetting with hard attention to the task. In *International conference on machine learning*, pp. 4548–4557. PMLR, 2018.

Shin, H., Lee, J. K., Kim, J., and Kim, J. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017.

Sugiyama, M. and Kawanabe, M. *Machine learning in nonstationary environments: Introduction to covariate shift adaptation*. MIT press, 2012.

Swartworth, W. J., Needell, D., Ward, R., Kong, M., and Jeong, H. Nearly optimal bounds for cyclic forgetting. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=X25L5AjHig.

Wu, J., Zou, D., Braverman, V., Gu, Q., and Kakade, S. Last iterate risk bounds of sgd with decaying stepsize for over-parameterized linear regression. In *International Conference on Machine Learning*, pp. 24280–24314. PMLR, 2022a.

Wu, J., Zou, D., Braverman, V., Gu, Q., and Kakade, S. The power and limitation of pretraining-finetuning for linear regression under covariate shift. *Advances in Neural Information Processing Systems*, 35:33041–33053, 2022b.

Yang, L., Lin, S., Zhang, J., and Fan, D. Grown: Grow only when necessary for continual learning. *arXiv preprint arXiv:2110.00908*, 2021.

Yoon, J., Yang, E., Lee, J., and Hwang, S. J. Lifelong learning with dynamically expandable networks. *arXiv preprint arXiv:1708.01547*, 2017.

Yoon, J., Kim, S., Yang, E., and Hwang, S. J. Scalable and order-robust continual learning with additive parameter decomposition. *arXiv preprint arXiv:1902.09432*, 2019.

Zou, D., Wu, J., Braverman, V., Gu, Q., and Kakade, S. Benign overfitting of constant-stepsize sgd for linear regression. In *Conference on Learning Theory*, pp. 4633–4635. PMLR, 2021.

A. Support Lemmas

Notations

For two matrices $\mathbf{A}$ and $\mathbf{B}$, their inner product is defined as $\langle \mathbf{A}, \mathbf{B} \rangle := \operatorname{tr}(\mathbf{A}^\top \mathbf{B})$. For each task $m \in [M]$, we define the following linear operators:

$$\mathcal{I} = \mathbf{I} \otimes \mathbf{I}, \quad \mathcal{M}_m = \mathbb{E}[\mathbf{x}_m \otimes \mathbf{x}_m \otimes \mathbf{x}_m \otimes \mathbf{x}_m], \quad \widetilde{\mathcal{M}}_m = \mathbf{H}_m \otimes \mathbf{H}_m,$$

$$\mathcal{T} = \mathbf{H} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{H}_m - \eta \mathcal{M}_m, \quad \widetilde{\mathcal{T}} = \mathbf{H}_m \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{H}_m - \eta \mathbf{H}_m \otimes \mathbf{H}_m.$$

We use the notation $\mathcal{O} \circ \mathbf{A}$ to denote the operator $\mathcal{O}$ acting on a symmetric matrix $\mathbf{A}$. For example, with these definitions, we have that for a symmetric matrix $\mathbf{A}$,

$$\mathcal{I} \circ \mathbf{A} = \mathbf{A}, \quad \mathcal{M}_m \circ \mathbf{A} = \mathbb{E}[(\mathbf{x}_m{}^\top \mathbf{A} \mathbf{x}_m) \mathbf{x}_m \mathbf{x}_m{}^\top], \quad \widetilde{\mathcal{M}}_m \circ \mathbf{A} = \mathbf{H}_m \mathbf{A} \mathbf{H}_m$$

$$(\mathcal{I} - \eta \widetilde{\mathcal{T}}_m) \circ \mathbf{A} = (\mathbf{I} - \eta \mathbf{H}_m) \mathbf{A} (\mathbf{I} - \eta \mathbf{H}_m)$$

$$(\mathcal{I} - \eta \mathcal{T}_m) \circ \mathbf{A} = \mathbb{E}[(\mathbf{I} - \eta \mathbf{x}_m \mathbf{x}_m{}^\top) \mathbf{A} (\mathbf{I} - \eta \mathbf{x}_m \mathbf{x}_m{}^\top)]$$

It can be readily understood that the following properties are satisfied:

Lemma A.1 ((Zou et al., 2021)). *An operator $\mathcal{O}$, when defined on symmetric matrices, is termed a Positive Semi-Definite (PSD) mapping if $\mathbf{A} \succeq 0$ implies $\mathcal{O} \circ \mathbf{A} \succeq 0$. Consequently, for each task $m \in [M]$ we have:*

1. *$\mathcal{M}_m$ and $\widetilde{\mathcal{M}}_m$ are both PSD mappings.*

2. *$\mathcal{M}_m - \widetilde{\mathcal{M}}_m$ and $\widetilde{\mathcal{T}}_m - \mathcal{T}_m$ are both PSD mappings.*

3. *$\mathcal{I} - \eta \mathcal{T}_m$ and $\mathcal{I} - \eta \widetilde{\mathcal{T}}_m$ are both PSD mappings.*

4. *If $0 < \eta < 1/\lambda_m^1$, then $\widetilde{\mathcal{T}}^{-1}$ exists, and is a PSD mapping.*

5. *If $0 < \eta < 1/(\alpha_m \operatorname{tr}(\mathbf{H}_m))$, then $\mathcal{T}_m^{-1} \circ \mathbf{A}$ exists for PSD matrix $\mathbf{A}$, and $\mathcal{T}_m^{-1}$ is a PSD mapping.*

Then for the SGD iterates, we can consider their associated bias iterates and variance iterates:

$$\begin{cases} \mathbf{B}_0 = (\mathbf{w}_0 - \mathbf{w}^*)(\mathbf{w}_0 - \mathbf{w}^*)^\top, \\ \mathbf{B}_{(m-1)N+t+1} = (\mathcal{I} - \eta \mathcal{T}_m(\eta)) \circ \mathbf{B}_{(m-1)N+t}; \end{cases} \tag{9}$$

$$\begin{cases} \mathbf{C}_0 = \mathbf{0}, \\ \mathbf{C}_{(m-1)N+t+1} = (\mathcal{I} - \eta \mathcal{T}_{\mathbf{H}_m}(\eta)) \circ \mathbf{C}_{(m-1)N+t} + \eta^2 \mathbf{\Sigma}_{\mathbf{H}_m}; \end{cases} \tag{10}$$

where $t = 0, \ldots, N-1$ and $m = 1, \ldots, M$.

Lemma A.2 (Bias-variance decomposition). *Suppose that Assumption 2.4 holds. Then we have:*

$$\mathbb{E}[\operatorname{ExcessRisk}(\mathbf{w}_{MN})] = \frac{1}{2} \langle \mathbf{H}, \mathbf{B}_{MN} \rangle + \frac{1}{2} \langle \mathbf{H}, \mathbf{C}_{MN} \rangle.$$

B. Variance Error

B.1. Upper Bound

The assumption presented below can be inferred from 2.3 by setting $\mathbf{A} = \mathbf{I}$, given that $R^2 = \max\{\alpha_m \operatorname{tr}(\mathbf{H}_m)\}_{m=1}^M$.

Assumption B.1 (Relaxed version). *For each task m, there exists a constant $R \geq 0$ such that:*

$$\mathbb{E}_{\mathbf{x} \sim \mathcal{D}_m}[\mathbf{x}\mathbf{x}^\top \mathbf{x}\mathbf{x}^\top] \preceq R^2 \mathbf{H}_m.$$

Lemma B.2. *Suppose Assumptions 2.3 and 2.4 hold with step size $\eta \leq 1/R^2$, then it holds that:*

$$\mathbf{C}_t \leq \frac{\eta \sigma^2}{1 - \eta R^2} \mathbf{I}, \quad \text{for every } t = 0, 1, \ldots, MN$$

Proof. This lemma is derived directly from the Lemmas in (Jain et al., 2018; Zou et al., 2021). To ensure completeness, we include a proof as follows.

We prove the lemma via induction. Initially, for $t = 0$, it is evident that $\mathbf{C}_0 = \mathbf{0} \preceq \frac{\eta\sigma^2}{1-\eta R^2}\mathbf{I}$. Now, assuming that $\mathbf{C}_t \preceq \frac{\eta\sigma^2}{1-\eta R^2}\mathbf{I}$, let us examine $\mathbf{C}_{t+1}$ in light of Equation (9). When $0 \le t \le N-1$, for each task m, it implies:

$$
\begin{aligned}
\mathbf{C}_{(m-1)N+t+1} &= (\mathcal{I} - \eta\mathcal{T}_{\mathbf{H}_m}(\eta)) \circ \mathbf{C}_{(m-1)N+t} + \eta^2\mathbf{\Sigma}_{\mathbf{H}_m} \\
&\preceq \frac{\eta\sigma^2}{1-\eta R^2}\mathbf{I} \cdot (\mathcal{I} - \eta\mathcal{T}_{\mathbf{H}_m}(\eta)) \circ \mathbf{I} + \eta^2\sigma^2\mathbf{H}_m \\
&\preceq \frac{\eta\sigma^2}{1-\eta R^2} \cdot (\mathbf{I} - 2\eta\mathbf{H}_m + \eta^2 R^2\mathbf{H}_m) + \eta^2\sigma^2\mathbf{H}_m \\
&\preceq \frac{\eta\sigma^2}{1-\eta R^2} \cdot \mathbf{I}.
\end{aligned}
\tag{11}
$$

$\square$

Lemma B.3. *Suppose Assumptions 2.3 and 2.4 hold with step size $\eta \le 1/R^2$, then it holds that:*

$$
\mathbf{C}_{MN} \preceq \prod_{m=2}^{M}(\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta))^N \circ \mathbf{C}_N + \sum_{m=1}^{M-1}\prod_{j=m+1}^{M}(\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_j}(\eta))^N\mathbf{P}_m + \mathbf{P}_M,
$$

where $\mathbf{P}_m = \frac{\eta^2\sigma^2}{1-\eta R^2}\sum_{t=0}^{N-1}(\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta))^t \circ \mathbf{H}_m$ and $\mathbf{C}_N \preceq \frac{\eta\sigma^2}{1-\eta R^2} \cdot (\mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_1)^N)$.

Proof. We first examine the recursion from $t = 0$ to $t = N-1$ for each task m:

$$
\begin{aligned}
\mathbf{C}_{(m-1)N+t+1} &= (\mathcal{I} - \eta\mathcal{T}_{\mathbf{H}_m}(\eta)) \circ \mathbf{C}_{(m-1)N+t} + \eta^2\mathbf{\Sigma}_{\mathbf{H}_m} \\
&\preceq (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta)) \circ \mathbf{C}_{(m-1)N+t} + \eta^2\mathcal{M}_m \circ \mathbf{C}_{(m-1)N+t} + \eta^2\sigma^2\mathbf{H}_m \\
&\preceq (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta)) \circ \mathbf{C}_{(m-1)N+t} + \eta^2 R^2 \cdot \frac{\eta\sigma^2}{1-\eta R^2} \cdot \mathbf{H}_m + \eta^2\sigma^2\mathbf{H}_m \\
&\preceq (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta)) \circ \mathbf{C}_{(m-1)N+t} + \frac{\eta^2\sigma^2}{1-\eta R^2}\mathbf{H}_m,
\end{aligned}
$$

where the penultimate inequality is derived from the Lemma B.2.

Hence, after N iterations, we could have the following results for task m:

$$
\mathbf{C}_{(m-1)N+N} \preceq (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta))^N \circ \mathbf{C}_{(m-1)N} + \frac{\eta^2\sigma^2}{1-\eta R^2}\sum_{t=0}^{N-1}(\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta))^t \circ \mathbf{H}_m.
$$

Now, we consider the first task incorporating with the Lemma B.5 in (Zou et al., 2021), which implies:

$$
\mathbf{C}_N \preceq \frac{\eta\sigma^2}{1-\eta R^2} \cdot (\mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_1)^N).
$$

By combining the aforementioned results and denoting $\mathbf{P}_m = \frac{\eta^2\sigma^2}{1-\eta R^2}\sum_{t=0}^{N-1}(\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta))^t \circ \mathbf{H}_m$, we obtain:

$$
\mathbf{C}_{MN} \preceq \prod_{m=2}^{M}(\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta))^N \circ \mathbf{C}_N + \sum_{m=1}^{M-1}\prod_{j=m+1}^{M}(\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_j}(\eta))^N\mathbf{P}_m + \mathbf{P}_M.
$$

$\square$

Based on Lemma A.2, the upper bound of the variance error can be expressed as follows:

$$\sum_{k=1}^{M} \langle \mathbf{H}_k, \mathbf{C}_{MN} \rangle \leq \underbrace{\sum_{k=1}^{M} \frac{\eta\sigma^2}{1-\eta R^2} \langle \mathbf{H}_k, \prod_{m=2}^{M} (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta))^N (\mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_1)^N) \rangle}_{\text{variance term 1}}$$

$$+ \underbrace{\sum_{k=1}^{M} \langle \mathbf{H}_k, \sum_{m=1}^{M-1} \prod_{j=m+1}^{M} (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_j}(\eta))^N \mathbf{P}_m \rangle}_{\text{variance term 2}} + \underbrace{\sum_{k=1}^{M} \langle \mathbf{H}_k, \mathbf{P}_M \rangle}_{\text{variance term 3}}. \qquad (12)$$

Let us consider the variance terms separately.

$$\begin{aligned}
\text{variance term 1} &= \sum_{k=1}^{M} \frac{\eta\sigma^2}{1-\eta R^2} \langle \mathbf{H}_k, \prod_{m=2}^{M} (\mathbf{I} - \eta\mathbf{H}_m)^N (\mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_1)^N)(\mathbf{I} - \eta\mathbf{H}_m)^N \rangle \\
&\leq \sum_{k=1}^{M} \frac{\eta\sigma^2}{1-\eta R^2} \langle \prod_{m=2}^{M} (\mathbf{I} - \eta\mathbf{H}_m)^N \mathbf{H}_k, (\mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_1)^N) \rangle \\
&= \frac{\eta\sigma^2}{1-\eta R^2} \sum_{k=1}^{M} \sum_{i} [\prod_{m=2}^{M} (1 - \eta\lambda_m^i)^N \lambda_k^i (1 - (1 - \eta\lambda_1^i)^N)] \\
&\leq \frac{\eta\sigma^2}{1-\eta R^2} (\sum_{i<k_1^*} \Gamma_{(2,M)}^i \Lambda^i + N\eta \sum_{i>k_1^*} \Gamma_{(2,M)}^i \lambda_1^i \Lambda^i),
\end{aligned} \qquad (13)$$

where we use the facts that $1 - (1 - \eta\lambda_m^i)^N \leq \min\{1, \eta N \lambda_m^i\}$ hold for all $i \geq 1$ in the last inequality.

Before we turn our attention to the second term, we first consider the $\mathbf{P}_m$:

$$\begin{aligned}
\mathbf{P}_m &= \frac{\eta^2\sigma^2}{1-\eta R^2} \sum_{t=0}^{N-1} (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta))^t \circ \mathbf{H}_m \\
&= \frac{\eta^2\sigma^2}{1-\eta R^2} \sum_{t=0}^{N-1} (\mathbf{I} - \eta\mathbf{H}_m)^t \mathbf{H}_m (\mathbf{I} - \eta\mathbf{H}_m)^t \\
&\preceq \frac{\eta\sigma^2}{1-\eta R^2} (\mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_m)^N).
\end{aligned}$$

Substituting the above to the variance term 2, we have:

$$\begin{aligned}
\text{variance term 2} &= \sum_{k=1}^{M} \langle \mathbf{H}_k, \sum_{m=1}^{M-1} \prod_{j=m+1}^{M} (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_j}(\eta))^N \mathbf{P}_m \rangle \\
&\leq \frac{\eta\sigma^2}{1-\eta R^2} \sum_{k=1}^{M} \langle \mathbf{H}_k, \sum_{m=1}^{M-1} \prod_{j=m+1}^{M} (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_j}(\eta))^N (\mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_m)^N) \rangle \\
&\leq \frac{\eta\sigma^2}{1-\eta R^2} \sum_{k=1}^{M} \langle \mathbf{H}_k, \sum_{m=1}^{M-1} \prod_{j=m+1}^{M} (\mathbf{I} - \eta\mathbf{H}_j)^N (\mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_m)^N) \rangle \\
&\leq \sum_{m=1}^{M-1} \frac{\eta\sigma^2}{1-\eta R^2} (\sum_{i<k_m^*} \Gamma_{(m+1,M)}^i \Lambda^i + N\eta \sum_{i>k_m^*} \Gamma_{(m+1,M)}^i \lambda_m^i \Lambda^i)
\end{aligned} \qquad (14)$$

Similarly, for the last term, we have:

$$\text{variance term 3} = \sum_{k=1}^{M} \langle \mathbf{H}_k, \mathbf{P}_M \rangle \leq \frac{\eta\sigma^2}{1-\eta R^2} (\sum_{i<k_M^*} \Gamma_{(M+1,M)}^i \Lambda^i + N\eta \sum_{i>k_M^*} \Gamma_{(M+1,M)}^i \lambda_m^i \Lambda^i) \qquad (15)$$

13

B.2. Lower Bound

Now, we shift our focus to the lower bound of variance. Similarly, we have the following lemma hold:

Lemma B.4. *Suppose Assumptions 2.3 and 2.4 hold with step size $\eta \leq 1/R^2$, then it holds that:*

$$\mathbf{C}_{MN} \succeq \prod_{m=2}^{M}(\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta))^N \circ \mathbf{C}_N + \sum_{m=1}^{M-1}\prod_{j=m+1}^{M}(\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_j}(\eta))^N\mathbf{P}'_m + \mathbf{P}'_M,$$

where $\mathbf{P}'_m = \eta^2\sigma^2\sum_{t=0}^{N-1}(\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta))^t \circ \mathbf{H}_m$ and $\mathbf{C}_N \succeq \frac{\eta\sigma^2}{2}\cdot(\mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_1)^{2N})$.

Proof. In a similar fashion, let's first examine the recursion of $\mathbf{C}$ from $t = 0$ to $t = N - 1$ for each task m.

$$
\begin{aligned}
\mathbf{C}_{(m-1)N+t+1} &= (\mathcal{I} - \eta\mathcal{T}_{\mathbf{H}_m}(\eta)) \circ \mathbf{C}_{(m-1)N+t} + \eta^2\mathbf{\Sigma}_{\mathbf{H}_m} \\
&= (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta)) \circ \mathbf{C}_{(m-1)N+t} + \eta^2(\mathcal{M}_m - \widetilde{\mathcal{M}}_m) \circ \mathbf{C}_{(m-1)N+t} + \eta^2\sigma^2\mathbf{H}_m \\
&\succeq (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta)) \circ \mathbf{C}_{(m-1)N+t} + \eta^2\sigma^2\mathbf{H}_m,
\end{aligned}
$$

where we utilize the fact that $\mathcal{M}_m - \widetilde{\mathcal{M}}_m$ is a PSD mapping, as established by A.1.

Consequently, after N iterations, the following results can be deduced for task m:

$$\mathbf{C}_{(m-1)N+N} \succeq (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta))^N \circ \mathbf{C}_{(m-1)N} + \eta^2\sigma^2\sum_{t=0}^{N-1}(\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta))^t \circ \mathbf{H}_m.$$

Now, we consider the first task incorporating the Lemma C.2 in (Zou et al., 2021), which implies:

$$\mathbf{C}_N \succeq \frac{\eta\sigma^2}{2}\cdot(\mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_1)^{2N}).$$

By combining the aforementioned results and denoting $\mathbf{P}'_m = \eta^2\sigma^2\sum_{t=0}^{N-1}(\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta))^t \circ \mathbf{H}_m$, we obtain:

$$\mathbf{C}_{MN} \succeq \prod_{m=2}^{M}(\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta))^N \circ \mathbf{C}_N + \sum_{m=1}^{M-1}\prod_{j=m+1}^{M}(\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_j}(\eta))^N\mathbf{P}'_m + \mathbf{P}'_M,$$

which completes the proof. $\square$

Drawing from Lemma A.2, the lower bound of the variance error is expressed as follows:

$$
\begin{aligned}
\sum_{k=1}^{M}\langle\mathbf{H}_k, \mathbf{C}_{MN}\rangle \geq{}& \underbrace{\sum_{k=1}^{M}\frac{\eta\sigma^2}{2}\langle\mathbf{H}_k, \prod_{m=2}^{M}(\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta))^N(\mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_1)^{2N})\rangle}_{\text{variance term } 1'} \\
&+ \underbrace{\sum_{k=1}^{M}\langle\mathbf{H}_k, \sum_{m=1}^{M-1}\prod_{j=m+1}^{M}(\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_j}(\eta))^N\mathbf{P}'_m\rangle}_{\text{variance term } 2'} + \underbrace{\sum_{k=1}^{M}\langle\mathbf{H}_k, \mathbf{P}'_M\rangle}_{\text{variance term } 3'}. \quad (16)
\end{aligned}
$$

Analogous to the approach for the upper bound, we will examine the terms one by one.

$$\text{variance term } 1' = \sum_{k=1}^{M} \frac{\eta\sigma^2}{2} \langle \mathbf{H}_k, \prod_{m=2}^{M} (\mathbf{I} - \eta\mathbf{H}_m)^N (\mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_1)^{2N})(\mathbf{I} - \eta\mathbf{H}_m)^N \rangle$$

$$= \sum_{k=1}^{M} \frac{\eta\sigma^2}{2} \langle \prod_{m=2}^{M} (\mathbf{I} - \eta\mathbf{H}_m)^{2N} \mathbf{H}_k, (\mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_1)^{2N}) \rangle$$

$$= \frac{\eta\sigma^2}{2} \sum_{k=1}^{M} \sum_i [\prod_{m=2}^{M} (1 - \eta\lambda_m^i)^{2N} \lambda_k^i (1 - (1 - \eta\lambda_1^i)^{2N})]$$

$$\geq \frac{\eta\sigma^2}{2} \sum_i [\prod_{m=2}^{M} (1 - \eta\lambda_m^i)^{2N} (\sum_{k=1}^{M} \lambda_k^i)(1 - (1 - \eta\lambda_1^i)^{2N})] \tag{17}$$

To further lower bound the two terms, noticing the following inequality:

$$1 - (1 - \eta\lambda_1^i)^{2N} \geq \begin{cases} 1 - (1 - \frac{1}{N})^{2N} \geq 1 - e^{-2} \geq \frac{9}{10}, & \lambda_1^i \geq \frac{1}{\eta N}, \\ 2N \cdot \eta\lambda_1^i - \frac{2N(N-1)}{2} \cdot \eta^2 \lambda_1^{i\,2} \geq \frac{9N}{10} \cdot \eta\lambda_1^i, & \lambda_1^i < \frac{1}{\eta N}. \end{cases}$$

Hence, the first term, we have:

$$\text{variance term } 1' \geq \frac{9\eta^2\sigma^2}{20} (\sum_{i<k_1^*} \Gamma_{(2,M)}^i \Lambda^i + N\eta \sum_{i>k_1^*} \Gamma_{(2,M)}^i \lambda_1^i \Lambda^i).$$

For the variance term $2'$, we notice that:

$$\mathbf{P}'_m = \eta^2\sigma^2 \sum_{t=0}^{N-1} (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta))^t \circ \mathbf{H}_m = \eta^2\sigma^2 \sum_{t=0}^{N-1} (\mathbf{I} - \eta\mathbf{H}_m)^{2t} \mathbf{H}_m$$

$$\geq \frac{\eta^2\sigma^2}{2} (\mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_m)^{2N})$$

Substituting the above to the variance term 2', we have:

$$\text{variance term } 2' = \sum_{k=1}^{M} \langle \mathbf{H}_k, \sum_{m=1}^{M-1} \prod_{j=m+1}^{M} (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_j}(\eta))^N \mathbf{P}'_m \rangle$$

$$\geq \frac{\eta^2\sigma^2}{2} \sum_{k=1}^{M} \langle \mathbf{H}_k, \sum_{m=1}^{M-1} \prod_{j=m+1}^{M} (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_j}(\eta))^N (\mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_m)^{2N}) \rangle$$

$$= \frac{\eta^2\sigma^2}{2} \sum_{k=1}^{M} \langle \mathbf{H}_k, \sum_{m=1}^{M-1} \prod_{j=m+1}^{M} (\mathbf{I} - \eta\mathbf{H}_j)^{2N} (\mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_m)^{2N}) \rangle$$

$$\geq \frac{9\eta^2\sigma^2}{20} \sum_{m=1}^{M-1} (\sum_{i<k_m^*} \Gamma_{(m+1,M)}^i \Lambda^i + N\eta \sum_{i>k_m^*} \Gamma_{(m+1,M)}^i \lambda_m^i \Lambda^i).$$

Also, similar to the variance term 3', it holds that:

$$\text{variance term 3'} \geq \frac{9\eta^2\sigma^2}{20} (\sum_{i<k_M^*} \Gamma_{(M+1,M)}^i \Lambda^i + N\eta \sum_{i>k_M^*} \Gamma_{(M+1,M)}^i \lambda_M^i \Lambda^i).$$

C. Bias Error

Before providing the proof of bias bound, we first introduce the following lemmas for tradition SGD training in Zou et al. 2021.

Lemma C.1 (Summation of bias iterates (Zou et al., 2021)). *Suppose that Assumption 2.3 holds. Suppose that $\eta < 1/(\alpha \operatorname{tr}(\mathbf{H}_m))$. Then for every $N \geq 1$ and each task m, it holds that:*

$$\frac{1}{2\eta} \cdot (\mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_m)^{2N}) \preceq \sum_{t=0}^{N-1} (\mathcal{I} - \eta \cdot \mathcal{T}_{\mathbf{H}_m}(\eta))^t \circ \mathbf{H}_m \preceq \frac{1}{\eta} \cdot (\mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_m)^{2N})$$

Lemma C.2. *Under Assumptions 2.3, let $\mathbf{B}_{a,b} = \mathbf{B}_a - (\mathbf{I} - \eta\mathbf{H}_m)^{b-a}\mathbf{B}_a(\mathbf{I} - \eta\mathbf{H}_m)^{b-a}$, if the stepsize satisfies $\eta < 1/(\alpha_m \operatorname{tr}(\mathbf{H}_m))$, then for any $t \leq N$, it holds that for each task m:*

$$\mathbf{S}_t \preceq \sum_{k=0}^{t-1} (\mathbf{I} - \eta\mathbf{H}_m)^k \big(\frac{\eta\alpha_m \operatorname{tr}(\mathbf{B}_{0,N})}{1 - \eta\alpha_m \operatorname{tr}(\mathbf{H}_m)} \cdot \mathbf{H}_m + \mathbf{B}_0 \big)(\mathbf{I} - \eta\mathbf{H}_m)^k,$$

where denoting $\mathbf{S}_t = \sum_{k=0}^{t-1}(\mathcal{I} - \mathcal{T}(\eta) \circ \mathbf{B}_0$.

Lemma C.3. *Suppose Assumptions 2.3 and 2.4 hold with step size $\eta \leq 1/R^2$, then it holds that:*

$$\mathbf{S}_t \succeq \frac{\beta_m}{4} \operatorname{tr} \left(\left(\mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_m)^{t/2} \right) \mathbf{B}_0 \right) \cdot \left(\mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_m)^{t/2} \right) + \sum^{t-1}(\mathbf{I} - \eta\mathbf{H}_m)^t \cdot \mathbf{B}_0 \cdot (\mathbf{I} - \eta\mathbf{H}_m)^t,$$

where denoting $\mathbf{S}_t = \sum_{k=0}^{t-1}(\mathcal{I} - \mathcal{T}(\eta) \circ \mathbf{B}_0$.

C.1. Upper Bound

Lemma C.4. *Suppose Assumptions 2.3 and 2.4 hold with step size $\eta \leq 1/R^2$, then it holds that:*

$$\mathbf{B}_{MN} \preceq \prod_{m=1}^{M} (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta))^N \circ \mathbf{B}_0 + \sum_{m=1}^{M} \prod_{j=m}^{M} (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_j}(\eta))^N \mathbf{P}_m,$$

where $\mathbf{P}_m = \alpha_m \eta^2 \sum_{t=0}^{N-1} (\mathbf{I} - \eta\mathbf{H}_m)^{2t} \mathbf{H}_m \langle \mathbf{H}_m, \mathbf{B}_{(m-1)N+t} \rangle$ and $\prod_{k_1}^{k_2} = 1$ if $k_1 > k_2$.

We first examine the recursion from $t = 0$ to $t = N - 1$ for each task m:

$$\begin{aligned}
\mathbf{B}_{(m-1)N+t+1} &= (\mathcal{I} - \eta\mathcal{T}_{\mathbf{H}_m}(\eta)) \circ \mathbf{B}_{(m-1)N+t} \\
&= (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta)) \circ \mathbf{B}_{(m-1)N+t} + \eta^2 (\mathcal{M}_m - \widetilde{\mathcal{M}}_m) \circ \mathbf{B}_{(m-1)N+t} \\
&\preceq (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta)) \circ \mathbf{B}_{(m-1)N+t} + \alpha_m \eta^2 \cdot \mathbf{H}_m \cdot \langle \mathbf{H}_m, \mathbf{B}_{(m-1)N+t} \rangle.
\end{aligned} \tag{18}$$

where the penultimate inequality is derived from the assumption 2.3.

Hence, after N iterations, we could have the following results for task m:

$$\begin{aligned}
\mathbf{B}_{(m-1)N+N} &\preceq (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta))^N \circ \mathbf{B}_{(m-1)N} + \alpha_m \eta^2 \sum_{t=0}^{N-1} (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta))^t \mathbf{H}_m \langle \mathbf{H}_m, \mathbf{B}_{(m-1)N+t} \rangle \\
&= (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta))^N \circ \mathbf{B}_{(m-1)N} + \alpha_m \eta^2 \sum_{t=0}^{N-1} (\mathbf{I} - \eta\mathbf{H}_m)^{2t} \mathbf{H}_m \langle \mathbf{H}_m, \mathbf{B}_{(m-1)N+t} \rangle \\
&= (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta))^N \circ \mathbf{B}_{(m-1)N} + \alpha_m \eta^2 \sum_{t=0}^{N-1} (\mathbf{I} - \eta\mathbf{H}_m)^{2t} \mathbf{H}_m \langle \mathbf{H}_m, (\mathcal{I} - \eta\mathcal{T}_{\mathbf{H}_m}(\eta))^t \mathbf{B}_{(m-1)N} \rangle \\
&\preceq (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta))^N \circ \mathbf{B}_{(m-1)N} + \alpha_m \eta^2 \sum_{t=0}^{N-1} \mathbf{H}_m \langle \mathbf{H}_m, (\mathcal{I} - \eta\mathcal{T}_{\mathbf{H}_m}(\eta))^t \mathbf{B}_{(m-1)N} \rangle
\end{aligned}$$

We now examine the second term for each m:

$$\sum_{t=0}^{N-1} \langle \mathbf{H}_m, (\mathcal{I} - \eta\mathcal{T}_{\mathbf{H}_m}(\eta))^t \mathbf{B}_{(m-1)N} \rangle$$

$$= \sum_{t=0}^{N-1} \langle \mathbf{H}_m, (\mathcal{I} - \eta\mathcal{T}_{\mathbf{H}_m}(\eta))^t (\mathcal{I} - \eta\mathcal{T}_{\mathbf{H}_{m-1}}(\eta))^N \dots (\mathcal{I} - \eta\mathcal{T}_{\mathbf{H}_1}(\eta))^N \mathbf{B}_0 \rangle$$

$$= \sum_{t=0}^{N-1} \langle (\mathcal{I} - \eta\mathcal{T}_{\mathbf{H}_{m-1}}(\eta))^N \dots (\mathcal{I} - \eta\mathcal{T}_{\mathbf{H}_1}(\eta))^N \mathbf{H}_m, (\mathcal{I} - \eta\mathcal{T}_{\mathbf{H}_m}(\eta))^t \mathbf{B}_0 \rangle,$$

where we know the following holds:

$$(\mathcal{I} - \eta\mathcal{T}_{\mathbf{H}_{m-1}}(\eta)) \circ \mathbf{H}_m = (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_{m-1}}(\eta)) \circ \mathbf{H}_m + (\mathcal{M} - \widetilde{\mathcal{M}}) \circ \mathbf{H}_m$$

$$\preceq (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_{m-1}}(\eta)) \circ \mathbf{H}_m + \alpha_{m-1}\eta^2 \cdot \mathbf{H}_{m-1} \cdot \langle \mathbf{H}_{m-1}, \mathbf{H}_m \rangle.$$

Moreover, we have $\sum_{t=0}^{N-1} \eta \cdot \mathbf{H}_{m-1} \cdot (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_{m-1}}(\eta))^t \preceq \mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_{m-1})^N \preceq \mathbf{I}$. Therefore, it holds that:

$$(\mathcal{I} - \eta\mathcal{T}_{\mathbf{H}_{m-1}}(\eta))^N \circ \mathbf{H}_m \preceq (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_{m-1}}(\eta))^N \circ \mathbf{H}_m + \alpha_{m-1}\eta \cdot \mathbf{I} \cdot \langle \mathbf{H}_{m-1}, \mathbf{H}_m \rangle.$$

It implies:

$$(\mathcal{I} - \eta\mathcal{T}_{\mathbf{H}_1}(\eta))^N \dots (\mathcal{I} - \eta\mathcal{T}_{\mathbf{H}_{m-1}}(\eta))^N \circ \mathbf{H}_m \preceq (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_1}(\eta))^N \dots (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_{m-1}}(\eta))^N \circ \mathbf{H}_m$$

$$+ \sum_{j=1}^{m-1} \prod_{k=1}^{j} \alpha_k \eta^j \cdot \langle \mathbf{H}_{k-1}, \mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_{m-1})^N \rangle \cdot \langle \mathbf{H}_j, \mathbf{H}_m \rangle \cdot \mathbf{I},$$

where we denote $\mathbf{H}_0 = \mathbf{I}$ and define $\Phi_1^{m-1} := \sum_{j=1}^{m-1} \prod_{k=1}^{j} \alpha_k \eta^j \cdot \langle \mathbf{H}_{k-1}, \mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_{m-1})^N \rangle \cdot \langle \mathbf{H}_j, \mathbf{H}_m \rangle$. Therefore, Equation (19) can be represented as follows:

$$\sum_{t=0}^{N-1} \langle (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_1}(\eta))^N \dots (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_{m-1}}(\eta))^N \circ \mathbf{H}_m + \Phi_1^{m-1} \cdot \mathbf{I}, (\mathcal{I} - \eta\mathcal{T}_{\mathbf{H}_m}(\eta))^t \mathbf{B}_0 \rangle$$

$$\leq \underbrace{\sum_{t=0}^{N-1} \langle (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_1}(\eta))^N \dots (\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_{m-1}}(\eta))^N \circ \mathbf{H}_m, (\mathbf{I} - \eta\mathbf{H}_m)^t (\frac{\eta\alpha_m \operatorname{tr}(\mathbf{B}_{0,N})}{1 - \eta\alpha_m \operatorname{tr}(\mathbf{H}_m)} \cdot \mathbf{H}_m + \mathbf{B}_0)(\mathbf{I} - \eta\mathbf{H}_m)^t \rangle}_{\text{term 1}}$$

$$+ \underbrace{\sum_{t=0}^{N-1} \langle \Phi_1^{m-1} \cdot \mathbf{I}, (\mathbf{I} - \eta\mathbf{H}_m)^t (\frac{\eta\alpha_m \operatorname{tr}(\mathbf{B}_{0,N})}{1 - \eta\alpha_m \operatorname{tr}(\mathbf{H}_m)} \cdot \mathbf{H}_m + \mathbf{B}_0)(\mathbf{I} - \eta\mathbf{H}_m)^t \rangle}_{\text{term 2}}.$$

We first consider the term 1 with Lemma C.2.

term 1

$$= \sum_{t=0}^{N-1} \langle \prod_{j=1}^{m-1} (\mathbf{I} - \eta\mathbf{H}_j)^{2N} (\mathbf{I} - \eta\mathbf{H}_m)^{2t} \mathbf{H}_m, (\frac{\eta\alpha_m \operatorname{tr}(\mathbf{B}_{0,N})}{1 - \eta\alpha_m \operatorname{tr}(\mathbf{H}_m)} \cdot \mathbf{H}_m + \mathbf{B}_0) \rangle$$

$$= \sum_{t=0}^{N-1} \frac{\eta\alpha_m \operatorname{tr}(\mathbf{B}_{0,N})}{1 - \eta\alpha_m \operatorname{tr}(\mathbf{H}_m)} \langle \prod_{j=1}^{m-1} (\mathbf{I} - \eta\mathbf{H}_j)^{2N} (\mathbf{I} - \eta\mathbf{H}_m)^{2t} \mathbf{H}_m, \mathbf{H}_m \rangle + \sum_{t=0}^{N-1} \langle (\mathbf{I} - \eta\mathbf{H}_m)^{2t} \mathbf{H}_m, \mathbf{B}_0 \rangle$$

$$\leq \frac{\alpha_m \operatorname{tr}(\mathbf{B}_{0,N})}{1 - \eta\alpha_m \operatorname{tr}(\mathbf{H}_m)} \langle \prod_{j=1}^{m-1} (\mathbf{I} - \eta\mathbf{H}_j)^{2N} (\mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_m)^N), \mathbf{H}_m \rangle + \frac{1}{\eta} \langle \prod_{j=1}^{m-1} (\mathbf{I} - \eta\mathbf{H}_j)^{2N} (\mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_m)^N), \mathbf{B}_0 \rangle$$

$$= \frac{\alpha_m \operatorname{tr}(\mathbf{B}_{0,N})}{1 - \eta\alpha_m \operatorname{tr}(\mathbf{H}_m)} \sum_{i} \Gamma_{(1,m-1)}^i [1 - (1 - \eta\lambda_m^i)^N] \lambda_m^i + \frac{1}{\eta} \sum_{i} \Gamma_{(1,m-1)}^i \omega_i^2 [1 - (1 - \eta\lambda_m^i)^N]$$

$$\leq \frac{\alpha_m \operatorname{tr}(\mathbf{B}_{0,N})}{1 - \eta\alpha_m \operatorname{tr}(\mathbf{H}_m)} \sum_{i} \Gamma_{(1,m-1)}^i \min\{1, \eta N\lambda_m^i\} + \frac{1}{\eta} \sum_{i} \Gamma_{(1,m-1)}^i \omega_i^2 \min\{1, \eta N\lambda_m^i\}$$

$$\leq \frac{\alpha_m \operatorname{tr}(\mathbf{B}_{0,N})}{1 - \eta\alpha_m \operatorname{tr}(\mathbf{H}_m)} (\sum_{i \leq k_m^*} \frac{\Gamma_{(1,m-1)}^i \lambda_m^i}{N\eta} + N\eta \sum_{i > k_m^*} \Gamma_{(1,m-1)}^i (\lambda_m^i)^2) + \frac{1}{\eta} \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{\Gamma}_1^{m-1} \mathbf{I}_{m,0:k_m^*}}^2 + N\|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{\Gamma}_1^{m-1} \mathbf{H}_{m,k_m^*:\infty}}^2$$

where k_m^* is the index of the smallest eigenvalue of $\mathbf{H}_m$ satisfying $\lambda_{k_m^*}^i \geq 1/(\eta N)$, and denotes $U_m = \frac{\alpha_m \operatorname{tr}(\mathbf{B}_{0,N})}{1-\eta\alpha_m \operatorname{tr}(\mathbf{H}_m)}(\sum_{i \leq k_m^*}\frac{\Gamma_{(1,m-1)}^i}{N\eta} + N\eta\sum_{i>k_m^*}(\lambda_m^i)^2) + \frac{1}{\eta}\|\mathbf{w}_0 - \mathbf{w}^*\|_{\Gamma_1^{m-1}\mathbf{I}_{m,0:k_m^*}}^2 + N\|\mathbf{w}_0 - \mathbf{w}^*\|_{\Gamma_1^{m-1}\mathbf{H}_{m,k_m^*:\infty}}^2$.

Moreover, $. \operatorname{tr}(\mathbf{B}_{0,N}) = \operatorname{tr}(\mathbf{B}_0 - (\mathbf{I} - \eta\mathbf{H}_m)^N \mathbf{B}_0(\mathbf{I} - \eta\mathbf{H}_m)^N)) = \sum(1 - (1-\eta\Lambda^i)^{2N}) \cdot (\langle\mathbf{w}_0 - \mathbf{w}^*, \mathbf{v}_i\rangle)^2$, Hence:

$$\operatorname{tr}(\mathbf{B}_{0,N}) \leq 2\sum_i \min\{1, N\eta\Lambda^i\}(\langle\mathbf{w}_0 - \mathbf{w}^*, \mathbf{v}_i\rangle)^2 \leq 2(\|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{I}_{m,0:k_m^*}}^2 + N\eta\|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}_{m,k_m^*:\infty}}^2).$$

Now we are ready to examine the term 2.

$$
\begin{aligned}
\text{term 2} &= \sum_{t=0}^{N-1}\langle\Phi_1^{m-1}\cdot\mathbf{I}\cdot(\mathbf{I}-\eta\mathbf{H}_m)^{2t}, (\frac{\eta\alpha_m\operatorname{tr}(\mathbf{B}_{0,N})}{1-\eta\alpha_m\operatorname{tr}(\mathbf{H}_m)}\cdot\mathbf{H}_m + \mathbf{B}_0)\rangle \\
&\leq \sum_{t=0}^{N-1}\frac{\eta\alpha_m\operatorname{tr}(\mathbf{B}_{0,N})}{1-\eta\alpha_m\operatorname{tr}(\mathbf{H}_m)}\langle\Phi_1^{m-1}\cdot(\mathbf{I}-\eta\mathbf{H}_m)^{2t},\mathbf{H}_m\rangle + \sum_{t=0}^{N-1}\langle\Phi_1^{m-1}\cdot(\mathbf{I}-\eta\mathbf{H}_m)^{2t},\mathbf{B}_0\rangle \\
&\leq \frac{\alpha_m\operatorname{tr}(\mathbf{B}_{0,N})}{1-\eta\alpha_m\operatorname{tr}(\mathbf{H}_m)}\langle\Phi_1^{m-1}\mathbf{I}, (\mathbf{I}-(\mathbf{I}-\eta\mathbf{H}_m)^N))\rangle + \frac{1}{\eta}\langle\sum_{j=1}^{m-1}\langle\mathbf{H}_j, \alpha^j\Gamma_1^{j-1}\cdot\mathbf{H}_m\rangle\mathbf{H}_m^{-1}(\mathbf{I}-(\mathbf{I}-\eta\mathbf{H}_m)^N), \mathbf{B}_0\rangle \\
&= \frac{\alpha_m\operatorname{tr}(\mathbf{B}_{0,N})}{1-\eta\alpha_m\operatorname{tr}(\mathbf{H}_m)}\sum_i\Phi_1^{m-1}[1-(1-\eta\lambda_m^i)^N] + \frac{1}{\eta}\sum_i\Phi_1^{m-1^i}\omega_i^2(\lambda_m^i)^{-1}[1-(1-\eta\lambda_m^i)^N] \\
&\leq \frac{\alpha_m\operatorname{tr}(\mathbf{B}_{0,N})}{1-\eta\alpha_m\operatorname{tr}(\mathbf{H}_m)}\sum_i\Phi_1^{m-1}\min\{1, \eta N\lambda_m^i\} + \frac{1}{\eta}\sum_i\Phi_1^{m-1}\omega_i^2(\lambda_m^i)^{-1}\min\{1, \eta N\lambda_m^i\} \\
&\leq \frac{\alpha_m\operatorname{tr}(\mathbf{B}_{0,N})\Phi_1^{m-1}}{1-\eta\alpha_m\operatorname{tr}(\mathbf{H}_m)}(k_m^* + N\eta\sum_{i>k_m^*}(\lambda_m^i)) + \frac{\Phi_1^{m-1}}{\eta}\|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}_{m,0:k_m^*}^{-1}}^2 + N\Phi_1^{m-1}\|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{I}_{m,k_m^*:\infty}}^2.
\end{aligned}
$$

Let us denote $V_m = \frac{\alpha_m\operatorname{tr}(\mathbf{B}_{0,N})\Phi_1^{m-1}}{1-\eta\alpha_m\operatorname{tr}(\mathbf{H}_m)}(k_m^* + N\eta\sum_{i>k_m^*}(\lambda_m^i)) + \frac{\Phi_1^{m-1}}{\eta}\|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}_{m,0:k_m^*}^{-1}}^2 + N\Phi_1^{m-1}\|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{I}_{m,k_m^*:\infty}}^2$.

By combining the aforementioned results, we obtain:

$$\mathbf{B}_{MN} \preceq \prod_{m=1}^{M}(\mathcal{I}-\eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta))^N \circ \mathbf{B}_0 + \sum_{m=1}^{M}\prod_{j=m}^{M}(\mathcal{I}-\eta\widetilde{\mathcal{T}}_{\mathbf{H}_j}(\eta))^N\mathbf{P}_m,$$

where denoting $\mathbf{P}_m = \alpha_m\eta^2(U_m + V_m)\cdot\mathbf{H}_m$.

Based on Lemma A.2, the upper bound of the bias error can be expressed as follows:

$$\sum_{k=1}^{M}\langle\mathbf{H}_k, \mathbf{B}_{MN}\rangle \leq \underbrace{\sum_{k=1}^{M}\langle\mathbf{H}_k, \prod_{m=1}^{M}(\mathcal{I}-\eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta))^N\circ\mathbf{B}_0\rangle}_{\text{bias term 1}} + \underbrace{\sum_{k=1}^{M}\langle\mathbf{H}_k, \sum_{m=1}^{M}\prod_{j=m}^{M}(\mathcal{I}-\eta\widetilde{\mathcal{T}}_{\mathbf{H}_j}(\eta))^N\mathbf{P}_m\rangle}_{\text{bias term 2}}.$$

For each k:

$$\langle \mathbf{H}_k, \mathbf{B}_{MN} \rangle = \langle \prod_{m=1}^{M} (\mathbf{I} - \eta \mathbf{H}_m)^{2N} \mathbf{H}_k, \mathbf{B}_0 \rangle + \langle \mathbf{H}_k, \sum_{m=1}^{M} \prod_{j=m}^{M} (\mathbf{I} - \eta \mathbf{H}_j)^{2N} \alpha_m \eta^2 (U_m + V_m) \cdot \mathbf{H}_m \rangle$$

$$\leq \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\prod_{m=1}^{M}(\mathbf{I}-\eta\mathbf{H}_m)^{2N}\mathbf{H}_k}$$

$$+ \sum_{m=1}^{M} \alpha_m \eta^2 \left(\frac{\alpha_m \operatorname{tr}(\mathbf{B}_{0,N})}{1 - \eta\alpha_m \operatorname{tr}(\mathbf{H}_m)} \left(\sum_{i<k_m^*} \frac{\Gamma^i_{(1,M)}(\lambda_m^i)^2 \lambda_k^i}{N\eta} + N\eta \sum_{i>k_m^*} \Gamma^i_{(1,M)}(\lambda_m^i)^3 \lambda_k^i \right) \right)$$

$$+ \sum_{m=1}^{M} \alpha_m \eta^2 \left(\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{(\boldsymbol{\Gamma}_1^M \mathbf{H}_m \mathbf{H}_k)_{0:k_m^*}} + N\eta \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{(\boldsymbol{\Gamma}_1^M \mathbf{H}_m^2 \mathbf{H}_k)_{k_m^*:\infty}} \right) \tag{19}$$

$$+ \sum_{m=1}^{M} \alpha_m \eta^2 \Phi_1^{m-1} \left(\frac{\alpha_m \operatorname{tr}(\mathbf{B}_{0,N})}{1 - \eta\alpha_m \operatorname{tr}(\mathbf{H}_m)} \left(\sum_{i<k_m^*} \Gamma^i_{(m,M)}(\lambda_m^i) + N\eta \sum_{i>k_m^*} \Gamma^i_{(m,M)}(\lambda_m^i)^2 \right) \right)$$

$$+ \sum_{m=1}^{M} \alpha_m \eta^2 \Phi_1^{m-1} \left(\frac{1}{\eta} \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{(\boldsymbol{\Gamma}_m^M \mathbf{H}_m^{-1})0:k_m^*} + N \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{(\boldsymbol{\Gamma}_m^M)_{k_m^*:\infty}} \right)$$

Hence,

$$\sum_{k=1}^{M} \langle \mathbf{H}_k, \mathbf{B}_{MN} \rangle \leq \sum_{k=1}^{M} \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\prod_{m=1}^{M}(\mathbf{I}-\eta\mathbf{H}_m)^{2N}\mathbf{H}_k}$$

$$+ \sum_{k=1}^{M} \sum_{m=1}^{M} \alpha_m \eta^2 \frac{\alpha_m \operatorname{tr}(\mathbf{B}_{0,N})}{1 - \eta\alpha_m \operatorname{tr}(\mathbf{H}_m)} \left(\sum_{i<k_m^*} \frac{\Gamma^i_{(1,M)}(\lambda_m^i)^2 \lambda_k^i}{N\eta} + N\eta \sum_{i>k_m^*} \Gamma^i_{(1,M)}(\lambda_m^i)^3 \lambda_k^i \right)$$

$$+ \sum_{k=1}^{M} \sum_{m=1}^{M} \alpha_m \eta^2 \left(\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{(\boldsymbol{\Gamma}_1^M \mathbf{H}_m \mathbf{H}_k)_{0:k_m^*}} + N\eta \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{(\boldsymbol{\Gamma}_1^M \mathbf{H}_k \mathbf{H}_m^2)_{k_m^*:\infty}} \right)$$

$$+ \sum_{k=1}^{M} \sum_{m=1}^{M} \alpha_m \eta^2 \Phi_1^{m-1} \left(\frac{\alpha_m \operatorname{tr}(\mathbf{B}_{0,N})}{1 - \eta\alpha_m \operatorname{tr}(\mathbf{H}_m)} \left(\sum_{i<k_m^*} \Gamma^i_{(m,M)} \lambda_k^i (\lambda_m^i) + N\eta \sum_{i>k_m^*} \Gamma^i_{(m,M)}(\lambda_m^i)^2 \lambda_k^i \right) \right)$$

$$+ \sum_{k=1}^{M} \sum_{m=1}^{M} \alpha_m \eta \Phi_1^{m-1} \left(\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{(\boldsymbol{\Gamma}_m^M \mathbf{H}_k)0:k_m^*} + N \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{(\boldsymbol{\Gamma}_m^M \mathbf{H}_k \mathbf{H}_m)_{k_m^*:\infty}} \right)$$

C.2. Lower Bound

We first examine the recursion from $t = 0$ to $t = N - 1$ for each task m:

$$\mathbf{B}_{(m-1)N+t+1} = (\mathcal{I} - \eta \mathcal{T}_{\mathbf{H}_m}(\eta)) \circ \mathbf{B}_{(m-1)N+t} \tag{20}$$

$$= (\mathcal{I} - \eta \widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta)) \circ \mathbf{B}_{(m-1)N+t} + \eta^2 (\mathcal{M}_m - \widetilde{\mathcal{M}}_m) \circ \mathbf{B}_{(m-1)N+t}$$

$$\succeq (\mathcal{I} - \eta \widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta)) \circ \mathbf{B}_{(m-1)N+t} + \beta_m \eta^2 \cdot \mathbf{H}_m \cdot \langle \mathbf{H}_m, \mathbf{B}_{(m-1)N+t} \rangle.$$

Hence, after N iterations, we could have the following results for task m:

$$\mathbf{B}_{(m-1)N+N} \succeq (\mathcal{I} - \eta \widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta))^N \circ \mathbf{B}_{(m-1)N} + \beta_m \eta^2 \sum_{t=0}^{N-1} (\mathcal{I} - \eta \widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta))^t \mathbf{H}_m \langle \mathbf{H}_m, \mathbf{B}_{(m-1)N+t} \rangle$$

$$= (\mathcal{I} - \eta \widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta))^N \circ \mathbf{B}_{(m-1)N} + \beta_m \eta^2 \sum_{t=0}^{N-1} (\mathbf{I} - \eta \mathbf{H}_m)^{2t} \mathbf{H}_m \langle \mathbf{H}_m, \mathbf{B}_{(m-1)N+t} \rangle$$

$$= (\mathcal{I} - \eta \widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta))^N \circ \mathbf{B}_{(m-1)N} + \beta_m \eta^2 \sum_{t=0}^{N-1} (\mathbf{I} - \eta \mathbf{H}_m)^{2t} \mathbf{H}_m \langle \mathbf{H}_m, (\mathcal{I} - \eta \mathcal{T}_{\mathbf{H}_m}(\eta))^t \mathbf{B}_{(m-1)N} \rangle$$

$$\tag{21}$$

19

We now examine the second term for each m:

$$\beta_m \eta^2 \sum_{t=0}^{N-1} (\mathbf{I} - \eta \mathbf{H}_m)^{2t} \mathbf{H}_m \langle \mathbf{H}_m, (\mathcal{I} - \eta \mathcal{T}_{\mathbf{H}_m}(\eta))^t \mathbf{B}_{(m-1)N} \rangle$$

$$= \beta_m \eta^2 \sum_{t=0}^{N-1} (\mathbf{I} - \eta \mathbf{H}_m)^{2t} \mathbf{H}_m \langle \mathbf{H}_m, (\mathcal{I} - \eta \mathcal{T}_{\mathbf{H}_m}(\eta))^t (\mathcal{I} - \eta \mathcal{T}_{\mathbf{H}_{m-1}}(\eta))^N \dots (\mathcal{I} - \eta \mathcal{T}_{\mathbf{H}_1}(\eta))^N \mathbf{B}_0 \rangle \qquad (22)$$

Acccording to Assumption 2.3 (B), we have:

$$(\mathcal{I} - \eta \mathcal{T}_{\mathbf{H}_{m-1}}(\eta)) \circ \mathbf{H}_m = (\mathcal{I} - \eta \widetilde{\mathcal{T}}_{\mathbf{H}_{m-1}}(\eta)) \circ \mathbf{H}_m + (\mathcal{M} - \widetilde{\mathcal{M}}) \circ \mathbf{H}_m$$

$$\succeq (\mathcal{I} - \eta \widetilde{\mathcal{T}}_{\mathbf{H}_{m-1}}(\eta)) \circ \mathbf{H}_m + \beta_{m-1} \eta^2 \cdot \mathbf{H}_{m-1} \cdot \langle \mathbf{H}_{m-1}, \mathbf{H}_m \rangle$$

$$\to (\mathcal{I} - \eta \mathcal{T}_{\mathbf{H}_{m-1}}(\eta))^N \circ \mathbf{H}_m \succeq (\mathcal{I} - \eta \widetilde{\mathcal{T}}_{\mathbf{H}_{m-1}}(\eta))^N \circ \mathbf{H}_m + \beta_{m-1} \eta^2 \cdot \sum_{t=0}^{N-1} (\mathcal{I} - \eta \widetilde{\mathcal{T}}_{\mathbf{H}_{m-1}}(\eta))^t \circ \mathbf{H}_{m-1} \cdot \langle \mathbf{H}_{m-1}, \mathbf{H}_m \rangle$$

$$\succeq (\mathcal{I} - \eta \widetilde{\mathcal{T}}_{\mathbf{H}_{m-1}}(\eta))^N \circ \mathbf{H}_m + \frac{\beta_{m-1} \eta}{2} \cdot (\mathbf{I} - (\mathbf{I} - \eta \mathbf{H}_{m-1})^{2N}) \cdot \langle \mathbf{H}_{m-1}, \mathbf{H}_m \rangle.$$

Therefore, we have iterations that:

$$(\mathcal{I} - \eta \mathcal{T}_{\mathbf{H}_1}(\eta))^N \dots (\mathcal{I} - \eta \mathcal{T}_{\mathbf{H}_{m-1}}(\eta))^N \circ \mathbf{H}_m \succeq (\mathcal{I} - \eta \widetilde{\mathcal{T}}_{\mathbf{H}_1}(\eta))^N \dots (\mathcal{I} - \eta \widetilde{\mathcal{T}}_{\mathbf{H}_{m-1}}(\eta))^N \circ \mathbf{H}_m$$

$$+ \sum_{j=1}^{m-1} \prod_{k=1}^{j} \beta_k (\frac{\eta}{2})^j \cdot \langle \mathbf{H}_{k-1}, (\mathbf{I} - (\mathbf{I} - \eta \mathbf{H}_{m-1})^{2N}) \rangle \cdot \langle \mathbf{H}_j, \mathbf{H}_m \rangle \cdot \mathbf{I}.$$

Subsituting the above to Equation (22) and denoting $\hat{\Phi}_1^{m-1} := \sum_{j=1}^{m-1} \prod_{k=1}^{j} \beta_k (\frac{\eta}{2})^j \cdot \langle \mathbf{H}_{k-1}, (\mathbf{I} - (\mathbf{I} - \eta \mathbf{H}_{m-1})^{2N}) \rangle \cdot \langle \mathbf{H}_j, \mathbf{H}_m \rangle$, we have:

$$\sum_{t=0}^{N-1} \langle \mathbf{H}_m, (\mathcal{I} - \eta \mathcal{T}_{\mathbf{H}_m}(\eta))^t \mathbf{B}_{(m-1)N} \rangle \succeq \sum_{t=0}^{N-1} \langle (\mathcal{I} - \eta \widetilde{\mathcal{T}}_{\mathbf{H}_{m-1}}(\eta))^N \dots (\mathcal{I} - \eta \widetilde{\mathcal{T}}_{\mathbf{H}_1}(\eta))^N \mathbf{H}_m, (\mathcal{I} - \eta \mathcal{T}_{\mathbf{H}_m}(\eta))^t \mathbf{B}_0 \rangle$$

$$+ \sum_{t=0}^{N-1} \langle \hat{\Phi}_1^{m-1} \mathbf{I}, (\mathcal{I} - \eta \mathcal{T}_{\mathbf{H}_m}(\eta))^t \mathbf{B}_0 \rangle$$

$$= \underbrace{\langle \prod_{p=1}^{m-1} (\mathbf{I} - \eta \mathbf{H}_p)^{2N} \mathbf{H}_m, \sum_{t=0}^{N-1} (\mathcal{I} - \eta \mathcal{T}_{\mathbf{H}_m}(\eta))^t \mathbf{B}_0 \rangle}_{\text{term 1}}$$

$$+ \underbrace{\langle \hat{\Phi}_1^{m-1} \mathbf{I}, \sum_{t=0}^{N-1} (\mathcal{I} - \eta \mathcal{T}_{\mathbf{H}_m}(\eta))^t \mathbf{B}_0 \rangle}_{\text{term 2}}$$

From the Lemma, we have:

$$\sum_{t=0}^{N-1} (\mathcal{I} - \eta \mathcal{T}_{\mathbf{H}_m}(\eta))^t \mathbf{B}_0 \succeq \frac{\beta_m}{4} \text{tr}((\mathbf{I} - (\mathbf{I} - \eta \mathbf{H}_m)^{N/2}) \mathbf{B}_0) \cdot (\mathbf{I} - (\mathbf{I} - \eta \mathbf{H}_m)^{N/2})$$

$$+ \sum_{t=0}^{N-1} (\mathbf{I} - \eta \mathbf{H}_m)^t \cdot \mathbf{B}_0 \cdot (\mathbf{I} - \eta \mathbf{H}_m)^t.$$

20

Then, for each task m, we examine the term 1:

$$\text{term } 1 = \langle \prod_{p=1}^{m-1} (\mathbf{I} - \eta \mathbf{H}_p)^{2N} \mathbf{H}_m, \sum_{t=0}^{N-1} (\mathcal{I} - \eta \mathcal{T}_{\mathbf{H}_m}(\eta))^t \mathbf{B}_0 \rangle$$

$$\geq \langle \prod_{p=1}^{m-1} (\mathbf{I} - \eta \mathbf{H}_p)^{2N} \mathbf{H}_m, \frac{\beta_m}{4} \text{tr}((\mathbf{I} - (\mathbf{I} - \eta \mathbf{H}_m)^{N/2}) \mathbf{B}_0) \cdot (\mathbf{I} - (\mathbf{I} - \eta \mathbf{H}_m)^{N/2}) \rangle$$

$$+ \langle \prod_{p=1}^{m-1} (\mathbf{I} - \eta \mathbf{H}_p)^{2N} \mathbf{H}_m, \sum_{t=0}^{N-1} (\mathbf{I} - \eta \mathbf{H}_m)^t \cdot \mathbf{B}_0 \cdot (\mathbf{I} - \eta \mathbf{H}_m)^t \rangle$$

$$= \underbrace{\frac{\beta_m}{4} \text{tr}((\mathbf{I} - (\mathbf{I} - \eta \mathbf{H}_m)^{N/2}) \mathbf{B}_0) \cdot \langle \prod_{p=1}^{m-1} (\mathbf{I} - \eta \mathbf{H}_p)^{2N} \mathbf{H}_m, (\mathbf{I} - (\mathbf{I} - \eta \mathbf{H}_m)^{N/2}) \rangle}_{\text{bias term } b_1^m}$$

$$+ \underbrace{\frac{1}{2\eta} \langle \prod_{p=1}^{m-1} (\mathbf{I} - \eta \mathbf{H}_p)^{2N} \cdot (\mathbf{I} - (\mathbf{I} - \eta \mathbf{H}_m)^{2N}), \mathbf{B}_0 \rangle}_{\text{bias term } b_2^m}$$

The first bias item is lower bounded by:

$$\text{bias term } b_1^m = \frac{\beta_m}{4} (\sum_i (1 - (1 - \eta \lambda_m^i)^{N/2}) \omega_i^2) \cdot (\sum_i \prod_{p=1}^{m-1} (1 - \eta \lambda_p^i)^{2N} \lambda_m^i (1 - (1 - \eta \lambda_m^i)^{N/2})),$$

The second bias item is lower bounded by:

$$\text{bias term } b_2^m \geq (\sum_i \prod_{p=1}^{m-1} (1 - \eta \lambda_p^i)^{2N} (1 - (1 - \eta \lambda_m^i)^{2N}) \omega_i^2)$$

To further lower bound the two terms, we notice that:

$$1 - (1 - \eta \lambda_m^i)^{\frac{N}{2}} \geq \begin{cases} 1 - (1 - \frac{1}{N})^{\frac{N}{2}} \geq 1 - e^{-\frac{1}{2}} \geq \frac{1}{5}, & \lambda_m^i \geq \frac{1}{\eta N} \\ \frac{N}{2} \cdot \eta \lambda_m^i - \frac{N(N-2)}{8} \cdot \eta^2 \lambda_m^i{}^2 \geq \frac{N}{5} \cdot \eta \lambda_m^i, & \lambda_m^i < \frac{1}{\eta N} \end{cases}$$

Substituting to the previous results, we have:

$$\text{bias term } b_1^m \geq \frac{\beta_m}{4} (\frac{1}{5} \cdot \sum_{i \leq k_m^*} \omega_i^2 + \frac{\eta N}{5} \sum_{i > k_m^*} (\lambda_m^i) \omega_i^2) \cdot (\frac{1}{5} \cdot \sum_{i \leq k_m^*} \Gamma_{(1,m-1)}^i \lambda_m^i + \frac{\eta N}{5} \sum_{i > k_m^*} \Gamma_{(1,m-1)}^i (\lambda_m^i)^2)$$

$$= \frac{\beta_m}{25} \cdot (\|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{I}_{m,0:k_m^*}}^2 + N\eta \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}_{m,k_m^*:\infty}}^2) \cdot (\sum_{i \leq k_m^*} \Gamma_{(1,m-1)}^i (\lambda_m^i) + \eta N \sum_{i > k_m^*} \Gamma_{(1,m-1)}^i (\lambda_m^i)^2)$$

and

$$\text{bias term } b_2^m \geq (\frac{1}{5} \cdot \sum_{i \leq k_m^*} \Gamma_{(1,m-1)}^i \omega_i^2 + \frac{\eta N}{5} \sum_{i > k_m^*} \Gamma_{(1,m-1)}^i \lambda_m^i \omega_i^2)$$

$$= \frac{1}{5} \cdot (\|\mathbf{w}_0 - \mathbf{w}^*\|_{(\prod_{p=1}^{m-1} (\mathbf{I} - \eta \mathbf{H}_p)^{2N})_{0:k_m^*}}^2 + N\eta \|\mathbf{w}_0 - \mathbf{w}^*\|_{(\prod_{p=1}^{m-1} (\mathbf{I} - \eta \mathbf{H}_p)^{2N} \mathbf{H}_m)_{k_m^*:\infty}}^2)$$

Now we are ready to examine term 2.

$$
\begin{aligned}
\text{term 2} =& \langle \hat{\Phi}_1^{m-1}\mathbf{I}, \sum_{t=0}^{N-1}(\mathcal{I} - \eta\mathcal{T}_{\mathbf{H}_m}(\eta))^t \mathbf{B}_0 \rangle \\
\geq& \langle \hat{\Phi}_1^{m-1}\mathbf{I}, \frac{\beta_m}{4}\operatorname{tr}((\mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_m)^{N/2})\mathbf{B}_0) \cdot (\mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_m)^{N/2}) \rangle \\
&+ \langle \hat{\Phi}_1^{m-1}\mathbf{I}, \sum_{t=0}^{N-1}(\mathbf{I} - \eta\mathbf{H}_m)^t \cdot \mathbf{B}_0 \cdot (\mathbf{I} - \eta\mathbf{H}_m)^t \rangle \\
=& \underbrace{\frac{\beta_m}{4}\operatorname{tr}((\mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_m)^{N/2})\mathbf{B}_0) \cdot \langle \hat{\Phi}_1^{m-1}\mathbf{I}, (\mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_m)^{N/2}) \rangle}_{\text{bias term } d_1^m} \\
&+ \underbrace{\frac{1}{2\eta}\langle \hat{\Phi}_1^{m-1}\mathbf{H}_m^{-1} \cdot (\mathbf{I} - (\mathbf{I} - \eta\mathbf{H}_m)^{2N}), \mathbf{B}_0 \rangle}_{\text{bias term } d_2^m}
\end{aligned}
$$

Analogous to term 1, we have:

$$
\begin{aligned}
\text{bias term } d_1^m =& \frac{\beta_m}{4}(\sum_i(1 - (1 - \eta\lambda_m^i)^{N/2})\omega_i^2) \cdot (\sum_i \hat{\Phi}_1^{m-1}\lambda_m^i(1 - (1 - \eta\lambda_m^i)^{N/2})) \\
\geq& \frac{\beta_m}{25} \cdot (\|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{I}_{m,0:k_m^*}}^2 + N\eta\|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}_{m,k_m^*:\infty}}^2) \cdot \Phi_1^{m-1} \cdot (\sum_{i<k_m^*}(\lambda_m^i) + \eta N\sum_{i>k_m^*}(\lambda_m^i)^2)
\end{aligned}
$$

and

$$
\begin{aligned}
\text{bias term } d_2^m \geq& (\sum_i \hat{\Phi}_1^{m-1}(\lambda_m^i)^{-1}(1 - (1 - \eta\lambda_m^i)^{2N})\omega_i^2) \\
\geq& \hat{\Phi}_1^{m-1}(\frac{1}{5} \cdot \sum_{i\leq k_m^*}(\lambda_m^i)^{-1}\omega_i^2 + \frac{\eta N}{5}\sum_{i>k_m^*}\omega_i^2) \\
=& \frac{\hat{\Phi}_1^{m-1}}{5} \cdot (\|\mathbf{w}_0 - \mathbf{w}^*\|_{(\mathbf{H}_m^{-1})_{0:k_m^*}}^2 + N\eta \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{I}_{k_m^*:\infty}}^2)
\end{aligned}
$$

After MN iterations, it holds that:

$$
\mathbf{B}_{MN} \succeq \prod_{m=1}^{M}(\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta))^N \circ \mathbf{B}_0 + \sum_{m=1}^{M}\prod_{j=m}^{M}(\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_j}(\eta))^N \mathbf{P}_m,
$$

where denoting $\mathbf{P}_m = \beta_m\eta^2(b_1^m + b_2^m + d_1^m + d_2^m) \cdot (\mathbf{I} - \eta\mathbf{H}_m)^{2N}\mathbf{H}_m$.

Then, the bias error can be represented as follows:

$$
\begin{aligned}
\sum_{k=1}^{M}\langle \mathbf{H}_k, \mathbf{B}_{MN} \rangle \geq& \underbrace{\sum_{k=1}^{M}\langle \mathbf{H}_k, \prod_{m=1}^{M}(\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_m}(\eta))^N \circ \mathbf{B}_0 \rangle}_{\text{bias term 1'}} + \underbrace{\sum_{k=1}^{M}\langle \mathbf{H}_k, \sum_{m=1}^{M}\prod_{j=m}^{M}(\mathcal{I} - \eta\widetilde{\mathcal{T}}_{\mathbf{H}_j}(\eta))^N \mathbf{P}_m \rangle}_{\text{bias term 2'}} \cdot \\
\geq& \sum_{k=1}^{M}\|\mathbf{w}_0 - \mathbf{w}^*\|_{\prod_{m=1}^{M}(\mathbf{I}-\eta\mathbf{H}_m)^{2N}\mathbf{H}_k}^2 \\
&+ \sum_{k=1}^{M}\langle \mathbf{H}_k, \sum_{m=1}^{M}\prod_{j=m}^{M}(\mathbf{I} - \eta\mathbf{H}_j)^{2N}\beta_m\eta^2(b_1^m + b_2^m + d_1^m + d_2^m) \cdot (\mathbf{I} - \eta\mathbf{H}_m)^{2N}\mathbf{H}_m \rangle.
\end{aligned}
$$

It follows that:

$$\sum_{k=1}^{M} \langle \mathbf{H}_k, \mathbf{B}_{MN} \rangle \geq \sum_{k=1}^{M} \| \mathbf{w}_0 - \mathbf{w}^* \|^2_{\prod_{m=1}^{M} (\mathbf{I} - \eta \mathbf{H}_m)^{2N} \mathbf{H}_k}$$

$$+ \sum_{k=1}^{M} \sum_{i} \lambda_k^i \cdot \sum_{m=1}^{M} \prod_{j=m}^{M} (1 - \eta \lambda_j^i)^{2N} \beta_m \eta^2 (b_1^m + b_2^m + d_1^m + d_2^m) \cdot (1 - \eta \lambda_m^i)^{2N} \lambda_m^i$$

$$\geq \sum_{k=1}^{M} \| \mathbf{w}_0 - \mathbf{w}^* \|^2_{\prod_{m=1}^{M} (\mathbf{I} - \eta \mathbf{H}_m)^{2N} \mathbf{H}_k}$$

$$+ \sum_{k=1}^{M} \sum_{m=1}^{M} (b_1^{m\prime} + b_2^{m\prime} + d_1^{m\prime} + d_2^{m\prime})$$

where

$$b_1^{m\prime} = \frac{\beta_m^2 \eta^2}{25} \cdot (\| \mathbf{w}_0 - \mathbf{w}^* \|^2_{\mathbf{I}_{m,0:k_m^*}} + N\eta \| \mathbf{w}_0 - \mathbf{w}^* \|^2_{\mathbf{H}_{m,k_m^*:\infty}})$$

$$\cdot (\sum_{i<k_m^*} \Gamma_{(1,M)}^i \lambda_k^i (\lambda_m^i)^2 + \eta N \sum_{i>k_m^*} \Gamma_{(1,M)}^i (\lambda_m^i)^3 \lambda_k^i)$$

$$b_2^{m\prime} = \frac{\beta_m \eta^2}{5} \cdot (\| \mathbf{w}_0 - \mathbf{w}^* \|^2_{(\mathbf{\Gamma}_{(1,M)} \mathbf{H}_m \mathbf{H}_k)_{0:k_m^*}} + N\eta \| \mathbf{w}_0 - \mathbf{w}^* \|^2_{(\mathbf{\Gamma}_{(1,M)} (\mathbf{I} - \eta \mathbf{H}_m)^{2N} \mathbf{H}_m^2 \mathbf{H}_k)_{k_m^*:\infty}}),$$

and

$$d_1^{m\prime} = \frac{\beta_m}{25} \cdot (\| \mathbf{w}_0 - \mathbf{w}^* \|^2_{\mathbf{I}_{m,0:k_m^*}} + N\eta \| \mathbf{w}_0 - \mathbf{w}^* \|^2_{\mathbf{H}_{m,k_m^*:\infty}})$$

$$\cdot \hat{\Phi}_1^{m-1} \cdot (\sum_{i<k_m^*} \Gamma_{(m,M)}^i \lambda_k^i (\lambda_m^i) + \eta N \sum_{i>k_m^*} \Gamma_{(m,M)}^i (\lambda_m^i)^2 \lambda_k^i)$$

$$d_2^{m\prime} = \frac{\beta_m \eta^2 \hat{\Phi}_1^{m-1}}{5} \cdot (\| \mathbf{w}_0 - \mathbf{w}^* \|^2_{(\mathbf{\Gamma}_{(m,M)} \mathbf{H}_k)_{0:k_m^*}} + N\eta \| \mathbf{w}_0 - \mathbf{w}^* \|^2_{(\mathbf{\Gamma}_{(m,M)} (\mathbf{I} - \eta \mathbf{H}_m)^{2N} \mathbf{H}_m \mathbf{H}_k)_{k_m^*:\infty}}).$$

D. Extension work

It is noticed that when the step size is set to $\| \mathbf{x}_m \|^{-2}$, the update rule for the minimum norm solution can be considered equivalent to that of the last iterate SGD. Consequently, in this subsection, we will focus on a particular case (akin to the setting in Lin et al. 2023) that involves this specific step size, allowing us to draw direct comparisons and insights under a defined set of conditions.

Consider a series of tasks $\mathbb{M} = \{1, 2, \ldots, M\}$. Given M datasets, for each dataset $m \in \mathbb{M}$, $D_m = \{(\mathbf{x}_{m,i}, y_{m,i})\}_{i=1}^{N}$ drawn i.i.d from some fixed distribution $\mathcal{D}_m = \mathcal{X}_m \times \mathcal{Y}_m \subset \mathbb{R}^d \times \mathbb{R}$. Assume that $\{(\mathbf{x}_{m,i}, y_{m,i})\}_{i=1}^{N}$ are i.i.d. sampled from a linear regression model, i.e., each $(\mathbf{x}_{m,i}, y_{m,i})$ is a realization of the linear regression model $y_m = (\mathbf{x}_m^\top \mathbf{w}_m^*) + z_m$, where z_m is some randomized noise satisfing well-specified condition and $\mathbf{w}_m^* \in \mathbb{R}^d$ is the optimal model parameter for task m.

We adopt the same learning procedure with specific step size, aiming to output a model $\mathbf{w}_M^N$ minimizing the *performance* (Lin et al., 2023), i.e.

$$G(\mathbf{w}_M^N) = \frac{1}{M} \sum_{i=1}^{M} \| \mathbf{w}_M^N - \mathbf{w}_i^* \|^2. \tag{23}$$

Therefore, our results can be restated as follows

Theorem D.1. *Consider a scenario where the model $\mathbf{w}$ undergoes training via SGD for M distinct tasks, following a sequence $1, \ldots, M$. With a specific step size of $\eta_{m,t} = \| \mathbf{x}_{m,t} \|^{-2}$, each task is executed for N iterations. Given that*

Assumption 2.4 are satisfied, the following will hold:

$$\mathbb{E}[G(\mathbf{w}_M^N)] = \frac{1}{M} \sum_{i=1}^{M} \|\mathbf{w}_0^0 - \mathbf{w}_i^*\|_{\Pi_{m=1}^{M} \Pi_{t=1}^{N}(\mathbf{I}-\mathbf{H}_m\eta_{m,t})}^2$$

$$+ \frac{1}{M} \sum_{i=1}^{M} \sum_{m=1}^{M} \sum_{t=0}^{N-1} \|\mathbf{w}_m^* - \mathbf{w}_i^*\|_{\Pi_{p=1}^{M-m} \Pi_{j=1}^{N}(\mathbf{I}-\mathbf{H}_p\eta_{p,j}) \Pi_{j=q}^{N-t}(\mathbf{I}-\mathbf{H}_m\eta_{m,q})\mathbf{H}_m\eta_{m,q}}^2$$

$$+ \frac{1}{M} \sum_{i=1}^{M} \sum_{m=1}^{M} \sum_{t=0}^{N-1} \|\mathbf{z}_{m,t}\|_{\Pi_{p=1}^{M-m} \Pi_{j=1}^{N}(\mathbf{I}-\mathbf{H}_p\eta_{p,j}) \Pi_{q=1}^{N-t}(\mathbf{I}-\mathbf{H}_m\eta_{m,q})\eta_{m,q}}^2.$$

Remark 2. In contrast to the approach in Theorem 3.1 and Theorem 3.2, here we do not rely on the decomposition of bias and variance error while considering that the projection $(\mathbf{I} - \eta_{m,t}\mathbf{x}_{m,t}\mathbf{x}_{m,t}^\top)$ is orthogonal to $\eta_{m,t}\mathbf{x}_{m,t}\mathbf{x}_{m,t}^\top$ with a specific stepsize $\eta_{m,t} = \|\mathbf{x}_{m,t}\|_2^{-2}$. This perspective allows us to derive a closed-form expression for the expected performance, which integrates the impact of initial parameter deviations, task-specific parameter variations, and random noise. Furthermore, Theorem D.1 in our study explores the performance behavior on general data distributions, expanding beyond the Gaussian distribution context discussed in Lin et al. 2023. In scenarios where there is only a single sample per training iteration, our results could cover their findings.

Proof. For each iteration, according to the update rule of SGD, it holds that

$$\mathbf{w}_m^N = \mathbf{w}_m^{N-1} - \eta(\boldsymbol{x}_{m,N}((\boldsymbol{x}_{m,N})^\top \mathbf{w}_m^{N-1} - y_{m,N})).$$

which can be rewritten as:

$$\mathbf{w}_m^N - \mathbf{w}_i^* = (\mathbf{I} - \eta\boldsymbol{x}_{m,N}(\boldsymbol{x}_{m,N})^\top)(\mathbf{w}_m^{N-1} - \mathbf{w}_i^*) + \eta z_{m,N}\boldsymbol{x}_{m,N}.$$

We consider the expectation norm for both sides:

$$\mathbb{E}[\|\mathbf{w}_m^N - \mathbf{w}_i^*\|^2]$$
$$= \mathbb{E}[(\mathbf{w}_m^N - \mathbf{w}_i^*)^\top(\mathbf{w}_m^N - \mathbf{w}_i^*)]$$
$$= \mathbb{E}[(\mathbf{w}_m^{N-1} - \mathbf{w}_i^*)^\top(\mathbf{I} - \eta\boldsymbol{x}_{m,N}(\boldsymbol{x}_{m,N})^\top)^\top(\mathbf{I} - \eta\boldsymbol{x}_{m,N}(\boldsymbol{x}_{m,N})^\top)(\mathbf{w}_m^{N-1} - \mathbf{w}_i^*) + \eta^2(z_{m,N}\boldsymbol{x}_{m,N})^\top(z_{m,N}\boldsymbol{x}_{m,N})]$$
$$(*) = \mathbb{E}[\|(\mathbf{I} - \eta_{m,N}\boldsymbol{x}_{m,N}(\boldsymbol{x}_{m,N})^\top)(\mathbf{w}_m^{N-1} - \mathbf{w}_i^*)\|^2 + \|\eta_{m,N}\boldsymbol{x}_{m,N}(\boldsymbol{x}_{m,N})^\top(\mathbf{w}_m^* - \mathbf{w}_i^*)\|^2 + \|\eta_{m,N}\boldsymbol{x}_{m,N}z_{m,N}\|^2]$$
$$= \mathbb{E}[\|\mathbf{w}_m^{N-1} - \mathbf{w}_i^*\|_{(\mathbf{I}-\mathbf{H}_m\eta_{m,N})}^2] + \eta_{m,N}\sigma^2 + \|\mathbf{w}_m^* - \mathbf{w}_i^*\|_{\mathbf{H}_m\eta_{m,N}}^2$$
$$= \|\mathbf{w}_m^0 - \mathbf{w}_i^*\|_{\Pi_{t=1}^{N}(\mathbf{I}-\mathbf{H}_m\eta_{m,t})}^2 + \sum_{t=1}^{N-1} \|\mathbf{w}_m^* - \mathbf{w}_i^*\|_{\Pi_{j=1}^{N-t}(\mathbf{I}-\mathbf{H}_m\eta_{m,j})^j\mathbf{H}_m\eta_{m,N}}^2 + \sum_{t=1}^{N-1} \|\mathbf{z}_{m,t}\|_{\Pi_{j=1}^{N-t}(\mathbf{I}-\mathbf{H}_m\eta_{m,j})^j\eta_{m,N}}^2,$$

where the (*) equation comes from the choice of step size such that $(\mathbf{I} - \eta_{m,t}\boldsymbol{x}_{m,t}(\boldsymbol{x}_{m,t})^\top)$ and $\eta_{m,t}\boldsymbol{x}_{m,t}(\boldsymbol{x}_{m,t})^\top$ are orthogonal projection, which equals the minimum norm solution with one sample.

Considering M tasks, it holds that

$$\mathbb{E}\|\mathbf{w}_{m,} - \mathbf{w}_i^*\|^2 = \|\mathbf{w}_0^0 - \mathbf{w}_i^*\|_{\Pi_{m=1}^{M} \Pi_{t=1}^{N}(\mathbf{I}-\mathbf{H}_m\eta_{m,t})}^2$$

$$+ \sum_{m=1}^{M-1} \sum_{t=0}^{N-1} \|\mathbf{w}_m^* - \mathbf{w}_i^*\|_{\Pi_{p=1}^{M-m} \Pi_{j=1}^{N}(\mathbf{I}-\mathbf{H}_p\eta_{p,j}) \Pi_{j=q}^{N-t}(\mathbf{I}-\mathbf{H}_m\eta_{m,q})\mathbf{H}_m\eta_{m,q}}^2$$

$$+ \sum_{m=1}^{M-1} \sum_{t=0}^{N-1} \|\mathbf{z}_m\|_{\Pi_{p=1}^{M-m} \Pi_{j=1}^{N}(\mathbf{I}-\mathbf{H}_p\eta_{p,j}) \Pi_{q=1}^{N-t}(\mathbf{I}-\mathbf{H}_m\eta_{m,q})\eta_{m,q}}^2.$$

In conclusion, we aggregate the performance metrics across tasks, ranging from $i = 1$ to $i = M$, to derive the final result. $\qquad\square$