The Impact of Overparameterization on Knowledge Transfer and Generalization in Multi-Task and Replay-Based Continual Learning

Chris Junchi Li[◇]

Department of Electrical Engineering and Computer Sciences[◇]
University of California, Berkeley

September 17, 2024

Abstract

This paper provides theoretical insights into the behavior of overparameterized models in the context of multi-task learning (MTL) and replay-based continual learning (CL). We derive explicit expressions for the generalization error and knowledge transfer of linear models with i.i.d. Gaussian features and additive noise. Our analysis reveals the effects of task similarity, model complexity, and sample size on the generalization error, highlighting the existence of error peaks caused by both label noise and task dissimilarity. We extend these results to replay-based CL methods, illustrating the role of memory buffer size in mitigating catastrophic forgetting. Our findings are validated through experiments using deep neural networks, demonstrating that the observed behaviors in linear models can generalize to deep learning settings. These results enhance the understanding of overparameterized models in MTL and CL and provide practical insights into optimizing model and memory buffer sizes for real-world applications.

Keywords: Multi-Task Learning (MTL); Continual Learning (CL); Overparameterization; Catastrophic Forgetting; Knowledge Transfer

1 Introduction

In recent years, the field of machine learning has seen a paradigm shift towards models that are overparameterized, where the number of parameters in a model exceeds the number of training samples. This trend is especially prevalent in deep neural networks (DNNs), which often possess more parameters than data points, yet exhibit strong generalization performance. Despite achieving near-zero training error, overparameterized models have been shown to generalize well to unseen data, a phenomenon that defies classical learning theory, where overfitting is typically expected in such cases. This intriguing behavior, termed *benign overfitting*, is key to the success of modern neural networks but lacks comprehensive theoretical understanding in the context of multi-task learning (MTL) and continual learning (CL).

Multi-task learning aims to train a single model on multiple related tasks simultaneously, with the potential to improve the overall efficiency and generalization by leveraging shared information between tasks. However, in overparameterized regimes, the interplay between task similarity, model capacity, and knowledge transfer remains poorly understood. Additionally, continual learning, where tasks arrive sequentially, faces the challenge of *catastrophic forgetting*, where learning new tasks can lead to the loss of knowledge from previously learned tasks. Replay-based methods, which store and reuse a subset of past data, are among the most effective techniques to counteract forgetting, but their theoretical underpinnings in overparameterized models have yet to be fully explored.

In this paper, we focus on the theoretical aspects of overparameterized models in MTL and CL. We aim to characterize the generalization performance, task interference, and knowledge transfer mechanisms in these settings. Specifically, we provide closed-form expressions for the generalization error of linear models under Gaussian features and additive noise, highlighting the role of model complexity, task similarity, and sample size. Additionally, we extend our analysis to replay-based continual learning, studying the effects of memory buffer size and task dissimilarity on catastrophic forgetting and generalization error.

Backgrounds. To overcome the issue of the lack of generalization ability across related tasks, Multi-Task Learning (MTL) methods train a single model on multiple tasks simultaneously with the goal of benefiting from the similarities between a collection of related tasks to improve the efficiency of learning [Caruana(1997)]. MTL not only enhances the performance of individual tasks but also facilitates the development of models that are more adaptable and capable of transferring knowledge from one task to another. This transferability is particularly valuable in real-world applications of deep learning where data can be scarce or imbalanced [Zhang & Yang(2022)].

In many real-world applications, the tasks are not available for offline training, and instead, the agent needs to learn new tasks sequentially as they are encountered. Learning in these settings is particularly challenging due to the phenomenon known as catastrophic forgetting, where learning new tasks can lead to a significant loss of previously acquired knowledge [French(1999)]. Continual Learning (CL) is a solution to this challenge that builds on the foundation of MTL but with a focus on the model's ability to learn continuously over time without forgetting [Parisi et al.(2019)].

Despite the practical success of integrating MTL and CL with deep learning, there still remains a critical need for a theoretical understanding of learning mechanisms in these methods [Crawshaw(2020)]. Several efforts have been previously made to theoretically understand MTL [Baxter(2000), Ben-David & Borbely(20 However, these endeavors are inapplicable to contemporary deep neural networks (DNN), where the models are heavily parameterized. In the overparameterized regime, DNNs exhibit peculiar generalization behaviors, where despite having more parameters than training samples, they can still generalize well to unseen test data [Zhang et al.(2017)]. An important unexplored area, however, is the effect of overparametrization on the generalizability of MTL models. More importantly, it is crucial to understand the impact of overparametrization on the possibility of effective cross-task knowledge transfer.

DNN models optimized with SGD inherit some specific behaviors of linear models, specifically in overparameterized regimes [Chizat et al.(2020)]. As a result, linear models have been widely studied as the first step toward understanding the deep double descent and benign overfitting [Hastie et al.(2022)]. However, previous investigations are limited to single-task learning paradigms with additive label noise. In this work, we provide theoretical characterizations of the linear overparameterized models in an MTL configuration for the first time. More specifically:

- We provide explicit expressions describing the expected generalization error and knowledge transfer of multi-task learners and compare them to single-task learners in a linear regression setup with i.i.d. Gaussian features and additive noise. Our results help to understand the system's behavior by highlighting the role and the interplay of various system parameters, including task similarity, model size, and sample size. We demonstrate that a peak exists in the generalization error curve of the multi-task learner when increasing model complexity. The error peak stems not only from label noise but also from the dissimilarity across the tasks. We also highlight the effect of each task's sample size on the strength and location of

the test error peak. Additionally, we measure knowledge transfer and indicate the conditions under which the tasks can be effectively learned together or interfere with each other.

- We provide similar results for continual learners and use them to explain the characteristics of replay-based CL methods. These results demonstrate the impact of replay buffer size on forgetting and the generalization error. As a practical fruit, our results shed light on state-of-the-art replay-based CL methods by illustrating the effect of model size on the effectiveness of memory buffer, especially in the presence of practical limitations where the replay buffer and model size can not be arbitrarily large. Furthermore, we complete the theoretical results by analyzing the CL methods that use a combined strategy of regularization and rehearsal techniques.

- By conducting experiments on practical datasets using DNNs, we empirically show that our findings for the linear models are generalizable to DNNs and can enlighten some of their characteristics. Specifically, we perform experiments using DNNs with varying model architectures on different practical datasets such as CIFAR-100, Imagenet-R, and CUB-200, to emphasize that the test error follows a similar trend observed in linear models, demonstrating that our results can help understand MTL in DNNs. We also perform experiments with replay-based deep continual learning methods and study the effect of memory buffer size in more depth.

1.1 Related Work

Multi-Task Learning Several theoretical efforts have been made to understand the benefits of multi-task learning. In one branch, the statistical learning theory (SLT) toolkit was used to derive generalization bounds for MTL models. These efforts include using VC dimension [Baxter(2000)], covering number [Pentina & Ben-David(2015)], Rademacher complexity [Yousefi et al.(2018)], and algorithmic stability [Zhang(2015)]. This line of work mostly focuses on defining notions of task similarity in the SLT sense [Ben-David & Borbely(2008)] and deriving generalization bounds as a function of the number of tasks, training set size, and task similarity [Zhang & Yang(2022)]. However, these methods are inapplicable to the contemporary paradigm of overparametrized DNNs [Allen-Zhu et al.(2019)], mainly because these models achieve zero training error and perfectly memorize the training data, yet they still generalize well to the unseen test set.

Additionally, some previous work studied multi-variate regression models, but their analysis or setup fundamentally differs from ours. For instance, they study sparse systems of equations [Lounici et al.(2009)] with the goal of union support recovery [Kolar et al.(2011), Obozinski et al.(2011)]. As another example, [Wu et al.(2020)] establishes an upper bound on the error of regularized underparameterized models. As opposed to our work, none of these works are able to explicitly characterize the exact generalization performance of the system based on its fundamental components in the overparameterized regime.

Continual Learning Most CL methods primarily use either model regularization [Schaul et al.(2015)] or experience replay techniques [Kirkpatrick et al.(2017)] to tackle catastrophic forgetting. Model regularization is based on consolidating the model parameters by penalizing drifts in the parameter space during model updates with the goal of preserving the older knowledge [Kirkpatrick et al.(2017), Zenke et al.(2017), Aljundi et al.(2018)]. The core idea for experience replay is to keep a portion of the training set in a separate memory buffer [Schaul et al.(2015)]. These samples are representative of previously encountered distributions and are replayed back during learning subsequent tasks

along with the current task data. The stored samples enable the model to revisit and learn from the distributions it has encountered before, facilitating the retention of acquired knowledge from past tasks [Robins(1995), Shin et al.(2017), Goodfellow et al.(2013)]. In practice, a combination of both techniques often leads to an optimal performance.

Several theoretical papers exist on understanding continual learning models [Doan et al.(2021), Yin et al.(2021), Chen et al.(2022)]. The most relevant study to our work is the investigation of regularization-based linear CL models. [Goldfarb & Hand(2023)] derive an upper bound for the generalization error of such models, and [Lin et al.(2023)] enhances this result by proposing a closed-form characterization of the error. However, these studies focus only on regularization-based techniques and are not helpful when studying replay-based CL methods. Besides, these results can be seen as a specific case of our studies in Section 3 by setting the memory size to zero.

Overparameterization and Double Descent In classical learning theory, it was widely accepted that increasing a model's complexity would initially reduce test error, but beyond a certain point, the error would rise again due to overfitting. However, the empirically observed double descent curve reveals that after the model complexity surpasses a certain threshold, making the model overparameterized, the test error surprisingly starts to decrease again [Nakkiran et al.(2021)]. In fact, state-of-the-art DNNs operate in the overparameterized regime, meaning that the model can perfectly fit the training data and achieve near-zero training errors while paradoxically still being able to generalize well to new unseen data, a phenomenon known as benign overfitting [Zhang et al.(2017), Cao et al.(2022)].

Double descent is not specific to DNNs [Belkin et al.(2019)]. For instance, linear regression exhibits similar behaviors under certain assumptions [Belkin et al.(2020)]. On the other hand, recent literature has pointed out a direct connection between linear models and more complex models such as neural networks optimized with SGD [Jacot et al.(2018), Gunasekar et al.(2018), Oymak & Soltanolkotabi(2019)]. Consequently, several studies investigated linear models as a proxy for more complex models such as DNNs [Hastie et al.(2022)]. However, these works are primarily focused on single-task setups and are not capable of capturing task notions in MTL setups. In contrast, our work analyzes the MTL case where multiple tasks are learned together. Considering that even a simple multi-class classification problem is a special case of MTL, our work is essential in understanding commonly used overparameterized DNN models.

2 Theoretical Results on MTL in Overparameterized Regimes

We first define the problem setting and then offer our theoretical results.

2.1 Problem Formulation

Consider a set of T learning tasks with a linear ground truth model for each task. Specifically, for task t, assume an input feature vector $\bar{x}_t \in \mathbb{R}^{s_t}$ and an output $y \in \mathbb{R}$ given by

$$y = \bar{x}_t^\top \bar{w}_t^* + z \tag{1}$$

where $\bar{w}_t^* \in \mathbb{R}^{s_t}$ denotes the optimal parameters, and $z \in \mathbb{R}$ is the random noise [Lin et al.(2023), Belkin et al.(2018), Evron et al.(2022)]. The specific true feature space for each task is unknown, and the features for a specific task may not be useful to other tasks. Since we seek to use a single

MTL model to learn all tasks, we consider a larger parameter space with size p that encompasses all true features.

More formally, consider a global set of features indexed by $1, 2, ...$ We assume that the true set of features for task t is denoted by S_t such that $|S_t| = s_t$. Among all features, we chose a set of p features denoted by $\mathcal{W}$ to train our model, assuming that $\cup_{t=1}^{\top} S_t \subseteq \mathcal{W}$. With this in mind, we define the expanded ground truth for task t by introducing $w_t^* \in \mathbb{R}^p$, where w_t^* is the same as $\bar{w}_t^*$ in the S_t indices and filled by zero in the remaining $p - s_t$ places.

Data model. We consider a training dataset of size n_t for task t represented as $\mathcal{D}_t = \{(x_{t,i}, y_{t,i})\}_{i=1}^{n_t}$, where $x_{t,i} \in \mathbb{R}^p$ and $y_{t,i} = x_{t,i}^{\top} w_t^* + z_{t,i}$. We also assume the features and noise are i.i.d. according to the following distributions:

$$x_{t,i} \sim \mathcal{N}(0, \mathbf{I}_p) \quad \text{and} \quad z_{t,i} \sim \mathcal{N}(0, \sigma^2) \tag{2}$$

Here, σ^2 denotes the noise strength.

Data model in matrix form. Consider a matrix representation by stacking the training data as $\mathbf{X}_t := [x_{t,1}, x_{t,2}, ..., x_{t,n_t}] \in \mathbb{R}^{p \times n_t}$, $\mathbf{y}_t := [y_{t,1}, y_{t,2}, ..., y_{t,n_t}]^{\top}$, and the noise vector as $\mathbf{z}_t = [z_{t,1}, z_{t,2}, ..., z_{t,n_t}]^{\top}$, to summarize the data generation process as

$$\mathbf{y}_t = \mathbf{X}_t^{\top} w_t^* + \mathbf{z}_t \tag{3}$$

Similarly, we might stack the training data of all tasks to build $\mathbf{X}_{1T} \in \mathbb{R}^{p \times \bar{n}}$, $\mathbf{y}_{1T} \in \mathbb{R}^{\bar{n}}$ and $\mathbf{z}_{1T} \in \mathbb{R}^{\bar{n}}$ where $\bar{n} = \sum_{t=1}^{\top} n_t$ is the total number of training samples.

Single-task learning (STL) To train a single-task learner for task t, we consider the standard setting in which the mean-squared-error loss function on $\mathcal{D}_t$ is optimized:

$$w_t = \underset{w \in \mathbb{R}^p}{\arg\min} \frac{1}{n_t} \|\mathbf{X}_t^{\top} w - \mathbf{y}_t\|_2^2 \tag{4}$$

In the underparameterized regime where $p < n_t$, there is a unique solution to minimizing the loss, given by $w_t = (\mathbf{X}_t \mathbf{X}_t^{\top})^{-1} \mathbf{X}_t \mathbf{y}_t$. In contrast, in the overparameterized regime where $p > n_t$, there are infinite solutions with zero training error. Among all solutions, we are particularly interested in the solution with minimum ℓ_2-norm, since it is the corresponding convergence point of applying stochastic gradient descent (SGD) on Equation 4 [Gunasekar et al.(2018)]. In fact, w_t in this case is obtained by equivalently solving the following constrained optimization:

$$w_t = \underset{w \in \mathbb{R}^p}{\arg\min} \|w\|_2^2 \quad \text{s.t.} \quad \mathbf{X}_t^{\top} w = \mathbf{y}_t \tag{5}$$

Since we are interested in the overparameterized regime, our focus is on solving Equation 5.

Single-task generalization error. To evaluate the generalization performance of w on task t, we use the following test loss:

$$\mathcal{L}_t(w) = \mathbb{E}_{x,z}[\|x^{\top} w - y\|_2^2] = \sigma^2 + \|w - w_t^*\|_2^2. \tag{6}$$

In what follows, we drop σ^2 and only use $\mathcal{L}_t(w) = \|w - w_t^*\|_2^2$ as a comparison criterion. Prior studies have characterized the behavior of the generalization error w.r.t. different model sizes in

STL [Hastie et al.(2022)]. In fact, the error increases by increasing p in the underparameterized regime until the training interpolation threshold $p \approx n_t$ is reached. After this point, the training error becomes zero and the model perfectly overfits the training noise. However, the model is benignly overfitted and increasing p further will reduce the noise effect. This surprising behavior is associated with the implicit regularization of the SGD, as discussed in Section 1.1. A more detailed discussion of prior results is presented in Section 4.1.

2.2 Main Results on Multi-Task Learning (MTL)

In this section, we present our main results for MTL and the deductions our results imply. Consider a multi-task learner that simultaneously learns all tasks by solving the optimization given in Equation 5 using the training data of all tasks:

$$w_{1T} = \underset{w \in \mathbb{R}^p}{\arg\min} \|w\|_2^2 \quad \text{s.t.} \quad \mathbf{X}_{1T}^\top w = \mathbf{y}_{1T} \tag{7}$$

To evaluate model performance, we utilize two metrics:

• *Average generalization error* reflects the overall generalization of the model, averaged across all tasks, i.e.,

$$G(w_{1T}) := \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}_t(w_{1T}) \tag{8}$$

• *Average knowledge transfer* is a metric to measure the gain of mutual cross-task knowledge transfer when comparing MTL to STL, i.e.,

$$K(w_{1T}) := \frac{1}{T} \sum_{t=1}^{T} [\mathcal{L}_t(w_t) - \mathcal{L}_t(w_{1T})] \tag{9}$$

Notably, a lower generalization and a higher knowledge transfer are more desirable. In what comes next, we offer the main results of the paper on the generalization error and knowledge transfer of an overparameterized multi-task learner.

Theorem 1. *The multi-task learner described in Equation 7 in the overparameterized regime, where $p \geq \bar{n} + 2$, has the following exact generalization error and knowledge transfer:*

$$\mathbb{E}[G(w_{1T})] = \underbrace{\sum_{t=1}^{\top} \sum_{t'=1}^{\top} \frac{n_{t'}}{Tp} \left(1 + \frac{Tn_t}{2(p - \bar{n} - 1)}\right) \|w_t^* - w_{t'}^*\|_2^2}_{term\ G_1} + \underbrace{\frac{1}{T}\left(1 - \frac{\bar{n}}{p}\right) \sum_{t=1}^{\top} \|w_t^*\|_2^2}_{term\ G_2} + \underbrace{\frac{\bar{n}\sigma^2}{p - \bar{n} - 1}}_{term\ G_3} \tag{10}$$

$$\begin{aligned} \mathbb{E}[K(w_{1T})] = & \underbrace{2 \sum_{t=1}^{\top} \sum_{t'=1}^{\top} \frac{n_{t'}}{Tp} \left(1 + \frac{Tn_t}{2(p - \bar{n} - 1)}\right) \langle w_t^*, w_{t'}^* \rangle}_{term\ K_1} \\ & \underbrace{- \left(1 + \frac{1}{T} + \frac{\bar{n}}{p - \bar{n} - 1}\right) \sum_{t=1}^{\top} \frac{n_t}{p} \|w_t^*\|_2^2}_{term\ K_2} \underbrace{- \frac{\bar{n}\sigma^2}{p - \bar{n} - 1} + \frac{1}{T} \sum_{t=1}^{\top} \frac{n_t \sigma^2}{p - n_t - 1}}_{term\ K_3} \end{aligned} \tag{11}$$

Proofs for all theorems are provided in Section 4.3. To the best of our knowledge, this is the first theoretical result on multi-task learning in overparameterized linear models that establishes a closed-form exact expression under non-asymptotic conditions, capturing the dependency of various system parameters. Setting $T = 1$ recovers the prior STL results [Hastie et al.(2022)], which means that our theorem is a more generalized version.

Theorem 1 precisely characterizes various phenomena that occur in the multi-task setting. Terms G_2 and G_3, which also appear in the STL setup, correspond to the task norm and noise strength. However, term G_1, which is specific to the MTL configuration, appears due to the distance between the optimal task vectors and is directly affected by task similarities. Concretely, we have the the following observations:

(i) **Interpolation threshold shifts**. Similar to the single-task learner, an interpolation threshold exists for the multi-task learner. For both models, the average error increases at the threshold and decreases afterward. However, the interpolation threshold for the single-task learner occurs at $p = n_t$, while for the multi-task learner, it happens at $p = \bar{n}$.

(ii) **Not all tasks can be learned together**. The presence of the term $\|w_t^* - w_{t'}^*\|_2^2$ in G_1 indicates the impact of task similarity on the overall generalization performance. More specifically, if the tasks are highly similar, meaning that their optimal parameters are close, utilizing the training dataset of one task can improve the test generalization of the other task. However, when using a multi-task learner to simultaneously learn tasks with considerable gaps, interference happens in the parameters space and leads to poor performance.

The knowledge transfer also tightly depends on task similarity. In fact, the sign of the K_1 term is closely correlated to the pairwise cosine similarity of the tasks. With conflicting tasks, negative knowledge transfer [Wang et al.(2019)] happens and the multi-task learner underperforms the single-task learner. In addition to that, the K_3 term in Equation 11 reveals that with the same p, the multi-task learner performs worse against the noise. However, to fairly compare the models, notice that the multi-task learner, which has only p parameters, is being compared against an average of T single-task learners with $T \times p$ overall parameters.

Finally, collaboration or conflict across the tasks is observable in models with limited capacity. In other words, when $p \to \infty$, both G_1 and K_1 terms vanish, meaning that neither collaboration nor interference happens anymore in infinite wide models.

(iii) **Descent floor exists when tasks are collaborative**. When tasks are collaborative and the noise is not too strong, increasing the p in the overparameterized regime can be a double-edged sword. On the one hand, increasing the number of parameters can help by reducing the effect of noise through the term G_3. On the other hand, increasing p to ∞ can kill the positive knowledge transfer across the tasks. There exists a point in the middle where the effect of both mechanisms matches and a descent floor appears in the generalization error

(iv) **Task dissimilarity and noise intensify the error peak**. As previously mentioned, an error peak exists at the interpolation threshold. This behavior corresponds to the existence of the label noise and intensifies as σ grows larger. However, in the single-task model, when $\sigma = 0$, meaning that there exists no label noise in the training set, no error peak happens at the interpolation threshold. Meanwhile, there are two sources for the error peak in MTL. Not only can the noise cause the peak through the term G_3, but task dissimilarity can also contribute through the term G_1, meaning that highly dissimilar tasks can produce larger error peaks. However, the benign overfitting is still observable since both terms converge to zero

as $p \to \infty$. This observation is consistent with earlier empirical findings where the double-descent phenomenon was observed even in the absence of label noise [Nakkiran et al.(2021)].

3 Theoretical Results on CL in Overparameterized Regimes

Continual learning can be considered a special type of MTL, with an extra constraint that the tasks arrive sequentially during a continual episode. At each timestep, t, the model is only exposed to $\mathcal{D}_t$, and the goal is to continually train a model that performs well across all tasks at the end of the episode. Note that the data for past learned tasks is not accessible when the current task is learned.

The most naive continual learner can be implemented using Equation 5, meaning that the learner only adapts to the last task at hand without any constraints. Such a learner performs well on the most recent task, but its performance on previous tasks degrades, a phenomenon called catastrophic forgetting in the CL literature. The forgetting effects emerge from the fact that the model is adapted to fit the current task data, which would make the learnable parameters sub-optimal for the past learned tasks. More formally, let $w_{\boldsymbol{v}t}$ denote the weights of the continual learner at timestep t after sequentially observing tasks from 1 to t. *Average forgetting* (lower is better) reflects the amount of negative backward knowledge transfer in a continual learner, defined as $F(w_{\boldsymbol{v}T}) := \frac{1}{T-1}\sum_{t=1}^{T-1}\left[\mathcal{L}_t(w_{\boldsymbol{v}T}) - \mathcal{L}_t(w_{\boldsymbol{v}t})\right]$. For simplicity, in the rest of this section, we assume that $n_t = n$ for all tasks.

Replay-based continual learning. Most successful methods in deep continual learning benefit from memory replay buffer [van de Ven et al.(2022)]. These methods store a small portion of the training data from each task. When learning new tasks, the stored samples are used to either regularize or retrain the model. In this section, we seek to theoretically study replay-based continual learning in the context of overparameterized linear models. Specifically, assume we have access to m_i samples for task i in the memory buffer when solving task $t+1$. Let the memory size be $\bar{m}_t = \sum_{i=1}^{t} m_i$. We use the notation $\widehat{\mathbf{X}}_{1t} \in \mathbb{R}^{p\times\bar{m}_t}$ and $\widehat{\mathbf{y}}_{1t} \in \mathbb{R}^{\bar{m}_t}$ to denote the training data and their labels in the memory. Then, the rehearsal-based continual learning model can be formulated as:

$$w_{\boldsymbol{v}t+1} = \underset{w\in\mathbb{R}^p}{\arg\min}\, \|w\|_2^2 \quad \text{s.t.} \quad \mathbf{X}_{t+1}^\top w = \mathbf{y}_{t+1}, \quad \widehat{\mathbf{X}}_{1t}^\top w = \widehat{\mathbf{y}}_{1t} \tag{12}$$

Replay+Regularization methods. Another family of CL methods are regularization techniques that aim to mitigate catastrophic forgetting by imposing constraints during the learning process. These constraints typically involve modifying the learning objective by penalizing the drift in the parameters space to preserve previously acquired knowledge. Previously, regularization-based continual learning has been studied in linear models [Lin et al.(2023)]. However, pure regularization is often not sufficient on its own for optimal continual learning. In practice, it is combined with memory replay methods, which can be formulated as:

$$w_{\boldsymbol{v}t+1} = \underset{w\in\mathbb{R}^p}{\arg\min}\, \|w - w_{\boldsymbol{v}t}\|_2^2 \quad \text{s.t.} \quad \mathbf{X}_{t+1}^\top w = \mathbf{y}_{t+1}, \quad \widehat{\mathbf{X}}_{1t}^\top w = \widehat{\mathbf{y}}_{1t} \tag{13}$$

where $w_{\boldsymbol{v}0} = \mathbf{0} \in \mathbb{R}^p$.

3.1 Theoretical Results on CL

In this section, we theoretically study the continual learners described in Equations 12 and 13. With a closer look, solving Equation 12 at $t = T$ is a specific case of multi-task learning described in Equation 7, achieved by setting $n_1 = m_1, ..., n_{T-1} = m_{T-1}$ and $n_T = n$. Therefore, we avoid repeating the theoretical results, and instead, we provide a corollary for the two-task case.

Theorem 2. *Assume $T = 2$, $\sigma = 0$, $m_1 = m$, $n_1 = n_2 = n$ and $\|w_1^*\| = \|w_2^*\|$. When $p \geq n+m+2$, for the continual learner described in Equation 12, it holds that*

$$\mathbb{E}[G(w_{\boldsymbol{v}T})] = \underbrace{\frac{n}{2p}(1 + \frac{m}{n} + \frac{2m}{p-(n+m)-1})\|w_1^* - w_2^*\|_2^2}_{term\ G_1} + \underbrace{(1 - \frac{n+m}{p})\|w_1^*\|_2^2}_{term\ G_2} \tag{14}$$

$$\mathbb{E}[F(w_{\boldsymbol{v}T})] = \underbrace{\frac{n}{p}(1 + \frac{m}{p-(n+m)-1})\|w_1^* - w_2^*\|_2^2}_{term\ F_1} - \underbrace{\frac{m}{p}\|w_1^*\|_2^2}_{term\ F_2} \tag{15}$$

The generalization error of the regularized learner is more complicated compared to the pure replay-based model and requires more delicate investigation. In fact, the dependence of $w_{\boldsymbol{v}t}$ to the samples in the memory buffer is the source of such complication and produces many cross-terms in the final expression. To keep the results understandable and intuitive, we present the results for the two tasks case here and leave the full form for Appendix 4.3.5. Appendix 4.4 also offers an in-depth discussion dedicated to the effect of regularization.

Theorem 3. *Assume $T = 2$, $\sigma = 0$, $m_1 = m$, $n_1 = n_2 = n$ and $\|w_1^*\| = \|w_2^*\|$. When $p \geq n+m+2$, for the continual learner described in Equation 13, it holds that*

$$\mathbb{E}[G(w_{\boldsymbol{v}T})] = \underbrace{\frac{n}{2p}(2 - \frac{n-m}{p-m} + \frac{2m}{p-(n+m)-1})\|w_1^* - w_2^*\|_2^2}_{term\ G_1} + \underbrace{(1 - \frac{n}{p-m})(1 - \frac{n}{p})\|w_1^*\|_2^2}_{term\ G_2} \tag{16}$$

$$\mathbb{E}[F(w_{\boldsymbol{v}T})] = \underbrace{\frac{n}{p}(1 + \frac{m}{p-(n+m)-1})\|w_1^* - w_2^*\|_2^2}_{term\ F_1} - \underbrace{\frac{n}{p-m}(1 - \frac{n}{p})\|w_1^*\|_2^2}_{term\ F_2} \tag{17}$$

Equations 14 to 17 reveal the effect of memory capacity. Setting $m = 0$ yields the naive (regularized) sequential learner. Increasing the memory size reduces the forgetting and the generalization error through the terms G_2 and F_2, respectively. However, a larger memory size can be harmful through the terms G_1 and F_1 if the tasks are highly dissimilar. Additionally, the full MTL model is recovered when $m = n$, meaning that all samples from the previous tasks are stored in the buffer.

An interesting observation is the effect of model size on the forgetting and the generalization error. With a larger p, both the positive and negative terms in the forgetting vanish. This suggests that bigger models with more capacity are less vulnerable to forgetting [Goldfarb & Hand(2023)]. The next observation is that a large p reduces the effect of memory size. This means a natural trade-off exists in practical applications with limited physical memory. In other words, one may consider investing the hardware in deploying larger models or consider a larger memory buffer for training samples. In fact, this is an essential result that sheds light on several state-of-the-art continual learning models where they reported that sample memory is not enough and the best solution is

to also keep some parts of the architecture in the memory [Zhou et al.(2023), Wang et al.(2022), Douillard et al.(2022)].

Although we presented the $T = 2$ versions of the theorems for better comprehension, the mentioned phenomena are also observable in the $T > 2$ case as presented in Section 4.3.4 and 4.3.5.

4 Appendix

4.1 Prelude: Single-Task Learning

In this section, we review the theoretical results on the single-task learners as a prelude to our work. The following theorem analyses the error of a single-task learner:

Theorem 4. *([Hastie et al.(2022)]) When $n_t \geq p+2$, the single-task learner described in Equation 4 achieves*

$$\mathbb{E}[\mathcal{L}_t(w_t)] = \frac{p\sigma^2}{n_t - p - 1} \tag{18}$$

and when $p \geq n_t + 2$, the single-task learner described in Equation 5 obtains

$$\mathbb{E}[\mathcal{L}_t(w_t)] = (1 - \frac{n_t}{p})\|w_t^*\|_2^2 + \frac{n_t\sigma^2}{p - n_t - 1} \tag{19}$$

where the expectation is due to randomness of X_t and z_t.

This theorem resembles the double descent phenomena in neural networks where the error increases by increasing p in the underparameterized regime until the training interpolation threshold $p \approx n_t$ is hit. After this point, the training error becomes zero and the model perfectly overfits the training noise. However, the model is benignly overfitted, meaning that increasing p will reduce the noise effect ($\frac{n_t\sigma^2}{p-n_t-1}$ converges to 0 as $p \to \infty$).

4.2 Naive-Sequential Learner

In this section, we present our theoretical results on the naive-sequential learner of sequentially optimizing Equation 5 on the last train set and ignoring the earlier tasks. We will use this simple theorem as a baseline for comparison with more sophisticated continual learning methods discussed in the coming sections. For simplicity, we assume that $n_t = n$ for all tasks.

Theorem 5. *For a naive sequential learner described in Equation 5, when $p \geq n + 2$, it holds that:*

$$\mathbb{E}[G(w_T)] = \frac{n}{Tp}\sum_{i=1}^{T}\|w_T^* - w_i^*\|^2 + \frac{1}{T}(1 - \frac{n}{p})\sum_{i=1}^{\top}\|w_i^*\|^2 + \frac{n\sigma^2}{p - n - 1} \tag{20}$$

$$\mathbb{E}[F(w_T)] = \frac{n}{p(T-1)}\sum_{i=1}^{T-1}\|w_T^* - w_i^*\|^2. \tag{21}$$

As this theorem suggests, even in the most naive sequential learner, increasing the number of parameters p can reduce the forgetting, as also reported by [Goldfarb & Hand(2023)]. However, the generalization error does not behave monotonically with the number of parameters and highly depends on the similarity of the tasks.

4.3 Proof of Theorems

4.3.1 Useful Lemmas

In this section, we start by providing some useful lemmas and then provide proofs for the theorems in the main text.

Lemma 1. *Consider a square invertible matrix $S \in \mathbb{R}^{n\times n}$. Assume S can be partitioned into four smaller blocks as*

$$S = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

where A and D are invertible square matrices with arbitrary relative sizes. Denote $H = D - CA^{-1}B$ and assume that H is invertible. Then, the inverse of S can be written as

$$S^{-1} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}BH^{-1}CA^{-1} & -A^{-1}BH^{-1} \\ -H^{-1}CA^{-1} & H^{-1} \end{bmatrix}$$

Proof. The lemma can be proved by simply examining that $SS^{-1} = I$. □

Lemma 2. *Consider a square invertible matrix $S \in \mathbb{R}^{n\times n}$. Assume S can be partitioned into 3×3 blocks:*

$$S = \begin{bmatrix} E & F & G \\ H & J & K \\ L & M & N \end{bmatrix}$$

Then the inverse of S is

$$S^{-1} = \begin{bmatrix} E^{-1} + E^{-1}(FA^{-1}H + UR^{-1}V)E^{-1} & -E^{-1}(F - UR^{-1}C)A^{-1} & -E^{-1}UR^{-1} \\ -A^{-1}(H - BR^{-1}V)E^{-1} & A^{-1} + A^{-1}BR^{-1}CA^{-1} & -A^{-1}BR^{-1} \\ -R^{-1}VE^{-1} & -R^{-1}CA^{-1} & R^{-1} \end{bmatrix}$$

where we define

$$\begin{aligned} A &= J - HE^{-1}F \\ B &= K - HE^{-1}G \\ C &= M - LE^{-1}F \\ D &= N - LE^{-1}G \\ U &= G - FA^{-1}B \\ V &= L - CA^{-1}H \\ R &= D - CA^{-1}B \end{aligned}$$

and we assume that the matrix inverses are defined wherever necessary.

Proof. Similarly this lemma is proved by examining $SS^{-1} = I$. □

Notations. We have already introduced some matrix notations in Section 2.1. In this section, we continue to introduce new helpful notations. Remember that for task $t \in [T]$ with n_t samples, we had $X_t \in \mathbb{R}^{p\times n_t}$, $w_t^* \in \mathbb{R}^p$, $z_t \in \mathbb{R}^{n_t}$ and $y_t = X_t^\top w_t^* + z_t$.

Now consider indices $i, j \in [T]$ such that $i < j$ and denote $n_{ij} = \sum_{t=i}^{j} n_t$. We introduce the concatenated data matrix as $X_{ij} \in \mathbb{R}^{p\times n_{ij}}$ which is constructed by concatenating data matrices from X_i to X_j in the second dimension, i.e. $X_{ij} = \begin{bmatrix} X_i & X_{i+1} & \dots & X_j \end{bmatrix}$. With this notation, the matrix representation of all training samples is $X_{1T} \in \mathbb{R}^{p\times \bar{n}}$ where $\bar{n} = \sum_{t=1}^{T} n_t$. Additionally, we introduce the matrix $X_{0t} \in \mathbb{R}^{p\times \bar{n}}$ which is built from X_{1T} by replacing all entries with zeros except at the columns corresponding to the task t. In other words, $X_{0t} = \begin{bmatrix} 0_{p\times r} & X_t & 0_{p\times q} \end{bmatrix}$, where $0_{p\times r}$ is an all zero matrix with size $p \times r$, $r = \sum_{i=1}^{t-1} n_i$ and $q = \sum_{i=t+1}^{T} n_i$.

Since our study mainly focuses on overparameterized regimes, the data points $X_t \in \mathbb{R}^{p\times n_t}$ are random matrices with more rows than columns. Therefore, $(X_t^\top X_t)^{-1}$ exists almost surely and we define $X_t^\dagger = X_t(X_t^\top X_t)^{-1}$ and the projection matrix $P_t = X_t(X_t^\top X_t)^{-1}X_t^\top$. Additionally, let $P_{ij} = X_{ij}(X_{ij}^\top X_{ij})^{-1}X_{ij}^\top$, $X_{ij}^\dagger = X_{ij}(X_{ij}^\top X_{ij})^{-1}$, and $P_{0t} = X_{1T}(X_{1T}^\top X_{1T})^{-1}X_{0t}^\top$. With this in mind, we provide some useful lemmas.

Lemma 3. *Let $X_{12} \in \mathbb{R}^{p\times(n_1+n_2)}$ be the result of concatenating $X_1 \in \mathbb{R}^{p\times n_1}$ and $X_2 \in \mathbb{R}^{p\times n_2}$. Also let $X_{01} \in \mathbb{R}^{p\times(n_1+n_2)}$ be result of concatenating X_1 with a zero matrix. Similarly, build $X_{02} \in \mathbb{R}^{p\times(n_1+n_2)}$ by concatenating a zero matrix with X_2. Assuming $p > n_1 + n_2$ and that all inverses exist, it holds that*

(i) $P_{12} = X_{12}(X_{12}^\top X_{12})^{-1}X_{12}^\top = P_1 + (I - P_1)X_2(X_2^\top(I - P_1)X_2)^{-1}X_2^\top(I - P_1)$

(ii) $P_{01} = X_{12}(X_{12}^\top X_{12})^{-1}X_{01}^\top = P_1 - (I - P_1)X_2(X_2^\top(I - P_1)X_2)^{-1}X_2^\top P_1$

(iii) $P_{02} = X_{12}(X_{12}^\top X_{12})^{-1}X_{02}^\top = (I - P_1)X_2(X_2^\top(I - P_1)X_2)^{-1}X_2^\top$

(iv) $P_{01}^\top P_{01} = X_{01}(X_{12}^\top X_{12})^{-1}X_{01}^\top = P_1 + P_1X_2(X_2^\top(I - P_1)X_2)^{-1}X_2^\top P_1$

(v)
$$\begin{aligned} X_{12}^\dagger &= X_{12}(X_{12}^\top X_{12})^{-1} \\ &= \begin{bmatrix} X_1^\dagger - (I - P_1)X_2(X_2^\top(I - P_1)X_2)^{-1}X_2^\top X_1^\dagger & (I - P_1)X_2(X_2^\top(I - P_1)X_2)^{-1} \end{bmatrix} \end{aligned}$$

where $P_1 = X_1(X_1^\top X_1)^{-1}X_1^\top$, $X_1^\dagger = X_1(X_1^\top X_1)^{-1}$ and I is the identity matrix with size $p \times p$.

Proof. We start by rewriting $X_{12}^\top X_{12}$ as a block matrix and then finding its inverse:

$$(X_{12}^\top X_{12})^{-1} = \left(\begin{bmatrix} X_1^\top X_1 & X_1^\top X_2 \\ X_2^\top X_1 & X_2^\top X_2 \end{bmatrix}\right)^{-1}$$

Now using Lemma 1

$$\begin{aligned} &(X_{12}^\top X_{12})^{-1} = \\ &\begin{bmatrix} (X_1^\top X_1)^{-1} + (X_1^\top X_1)^{-1}X_1^\top X_2 H^{-1} X_2^\top X_1 (X_1^\top X_1)^{-1} & -(X_1^\top X_1)^{-1}X_1^\top X_2 H^{-1} \\ -H^{-1}X_2^\top X_1(X_1^\top X_1)^{-1} & H^{-1} \end{bmatrix}, \end{aligned}$$

where $H = X_2^\top X_2 - X_2^\top X_1(X_1^\top X_1)^{-1}X_1^\top X_2 = X_2^\top(I - P_1)X_2$. Now, by noticing that $X_{12} = \begin{bmatrix} X_1 & X_2 \end{bmatrix}$, we can prove what was desired by doing multiplications:

(i)

$$\begin{aligned}
P_{12} &= \begin{bmatrix} X_1 & X_2 \end{bmatrix} \left(\begin{bmatrix} X_1^\top X_1 & X_1^\top X_2 \\ X_2^\top X_1 & X_2^\top X_2 \end{bmatrix} \right)^{-1} \begin{bmatrix} X_1^\top \\ X_2^\top \end{bmatrix} \\
&= P_1 + P_1 X_2 H^{-1} X_2^\top P_1 - X_2 H^{-1} X_2^\top P_1 - P_1 X_2 H^{-1} X_2^\top + X_2 H^{-1} X_2^\top \\
&= P_1 + (I - P_1) X_2 (X_2^\top (I - P_1) X_2)^{-1} X_2^\top (I - P_1)
\end{aligned}$$

(ii)

$$\begin{aligned}
P_{01} &= \begin{bmatrix} X_1 & X_2 \end{bmatrix} \left(\begin{bmatrix} X_1^\top X_1 & X_1^\top X_2 \\ X_2^\top X_1 & X_2^\top X_2 \end{bmatrix} \right)^{-1} \begin{bmatrix} X_1^\top \\ 0 \end{bmatrix} \\
&= P_1 + P_1 X_2 H^{-1} X_2^\top P_1 - X_2 H^{-1} X_2^\top P_1 \\
&= P_1 - (I - P_1) X_2 (X_2^\top (I - P_1) X_2)^{-1} X_2^\top P_1
\end{aligned}$$

(iii)

$$\begin{aligned}
P_{02} &= \begin{bmatrix} X_1 & X_2 \end{bmatrix} \left(\begin{bmatrix} X_1^\top X_1 & X_1^\top X_2 \\ X_2^\top X_1 & X_2^\top X_2 \end{bmatrix} \right)^{-1} \begin{bmatrix} 0 \\ X_2^\top \end{bmatrix} \\
&= -P_1 X_2 H^{-1} X_2^\top + X_2 H^{-1} X_2^\top \\
&= (I - P_1) X_2 (X_2^\top (I - P_1) X_2)^{-1} X_2^\top
\end{aligned}$$

(iv)

$$\begin{aligned}
P_{01}^\top P_{01} &= \begin{bmatrix} X_1 & 0 \end{bmatrix} \left(\begin{bmatrix} X_1^\top X_1 & X_1^\top X_2 \\ X_2^\top X_1 & X_2^\top X_2 \end{bmatrix} \right)^{-1} \begin{bmatrix} X_1^\top \\ 0 \end{bmatrix} \\
&= P_1 + P_1 X_2 H^{-1} X_2^\top P_1 = P_1 + P_1 X_2 (X_2^\top (I - P_1) X_2)^{-1} X_2^\top P_1
\end{aligned}$$

(v)

$$\begin{aligned}
& X_{12}^\dagger \\
&= \begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix} (X_1^\top X_1)^{-1} + (X_1^\top X_1)^{-1} X_1^\top X_2 H^{-1} X_2^\top X_1^\dagger & -(X_1^\top X_1)^{-1} X_1^\top X_2 H^{-1} \\ -H^{-1} X_2^\top X_1 (X_1^\top X_1)^{-1} & H^{-1} \end{bmatrix} \\
&= \begin{bmatrix} X_1^\dagger - (I - P_1) X_2 (X_2^\top (I - P_1) X_2)^{-1} X_2^\top X_1^\dagger & (I - P_1) X_2 (X_2^\top (I - P_1) X_2)^{-1} \end{bmatrix}
\end{aligned}$$

□

Lemma 4. *Let $X_{13} \in \mathbb{R}^{p \times (n_1+n_2+n_3)}$ be the result of concatenating $X_1 \in \mathbb{R}^{p \times n_1}$, $X_2 \in \mathbb{R}^{p \times n_2}$ and $X_3 \in \mathbb{R}^{p \times n_3}$. Also let $X_{01}, X_{02}, X_{03} \in \mathbb{R}^{p \times (n_1+n_2+n_3)}$ be the result of concatenating X_1, X_2 and X_3 with zero matrices such that $X_{01} = \begin{bmatrix} X_1 & 0 & 0 \end{bmatrix}$, $X_{02} = \begin{bmatrix} 0 & X_2 & 0 \end{bmatrix}$ and $X_{03} = \begin{bmatrix} 0 & 0 & X_3 \end{bmatrix}$. Define $P_{01} = X_{13}(X_{13}^\top X_{13})^{-1} X_{01}^\top$ and $P_{03} = X_{13}(X_{13}^\top X_{13})^{-1} X_{03}^\top$. Assuming $p > n_1 + n_2 + n_3$ and that all inverses exist, it holds that*

$$P_{01}^\top P_{03} = -P_1 (I - \widehat{P}_{12}) X_3 (X_3^\top (I - p_1)(I - \widehat{P}_{12}) X_3)^{-1} X_3^\top$$

where $P_1 = X_1 (X_1^\top X_1)^{-1} X_1^\top$ and we denote $\widehat{P}_{12} = X_2 (X_2^\top (I - P_1) X_2)^{-1} X_2^\top (I - P_1)$.

Proof.

$$\begin{aligned} P_{01}^\top P_{03} &= X_{01}(X_{13}^\top X_{13})^{-1}X_{13}^\top X_{13}(X_{13}^\top X_{13})^{-1}X_{03}^\top \\ &= X_{01}(X_{13}^\top X_{13})^{-1}X_{03}^\top \\ &= X_{01}\left(\begin{bmatrix} X_1^\top X_1 & X_1^\top X_2 & X_1^\top X_3 \\ X_2^\top X_1 & X_2^\top X_2 & X_2^\top X_3 \\ X_3^\top X_1 & X_3^\top X_2 & X_3^\top X_3 \end{bmatrix}\right)^{-1} X_{03}^\top \end{aligned}$$

Now we use Lemma 2 to find the matrix inverse. With the notation introduced in that lemma, we are looking for $-X_1E^{-1}UR^{-1}X_3^\top$. Thus we have,

$$\begin{aligned} A &= J - HE^{-1}F = X_2^\top(I-P_1)X_2 \\ B &= K - HE^{-1}G = X_2^\top(I-P_1)X_3 \\ C &= M - LE^{-1}F = X_3^\top(I-P_1)X_2 \\ D &= N - LE^{-1}G = X_3^\top(I-P_1)X_3 \\ U &= G - FA^{-1}B = X_1^\top(I-\widehat{P}_{12})X_3 \\ R &= D - CA^{-1}B = X_3^\top(I-P_1)(I-\widehat{P}_{12})X_3 \end{aligned}$$

$$\Rightarrow P_{01}^\top P_{03} = -X_1E^{-1}UR^{-1}X_3^\top = -P_1(I-\widehat{P}_{12})X_3(X_3^\top(I-P_1)(I-\widehat{P}_{12})X_3)^{-1}X_3^\top$$

□

Lemma 5. *Assume $X \in \mathbb{R}^{p\times n}$ is the training data matrix and $y \in \mathbb{R}^n$ are the corresponding labels. In the underparameterized regime where $p < n$, we seek to solve an optimization of the form*

$$w^* = \underset{w\in\mathbb{R}^p}{\arg\min} \|X^\top w - y\|_2^2$$

The optimal solution to this optimization is given by

$$w^* = (XX^\top)^{-1}Xy$$

Proof. By setting the derivative of the objective to zero, we obtain:

$$X(X^\top w^* - y) = 0 \Rightarrow w^* = (XX^\top)^{-1}Xy$$

noticing that $(XX^\top)^{-1}$ exist almost surely if X is a Gaussian random matrix with $p < n$. □

Lemma 6. *Let $X \in \mathbb{R}^{p\times n}$ be the training data matrix and $y \in \mathbb{R}^n$ be the corresponding labels. Assume $w_0 \in \mathbb{R}^p$ is an arbitrary fixed vector. In the overparameterized regime where $p > n$, we seek to solve optimizations of the form*

$$w^* = \underset{w\in\mathbb{R}^p}{\arg\min} \|w - w_0\|_2^2 \quad s.t. \quad X^\top w = y$$

The optimal solution to this optimization is given by

$$w^* = (I-P)w_0 + X^\dagger y$$

where $P = X(X^\top X)^{-1}X$ and $X^\dagger = X(X^\top X)^{-1}$.

Proof. Using the Lagrange multipliers and by setting the derivatives to 0, we can get:

$$w^*, \lambda^* = \arg\min_{w,\lambda} \frac{1}{2}\|w - w_0\|_2^2 + \lambda^\top (X^\top w - y)$$

$$\Rightarrow w^* - w_0 + X\lambda^* = 0 \Rightarrow w^* = -X\lambda^* + w_0$$

$$X^\top w^* = y \Rightarrow -X^\top X\lambda^* + X^\top w_0 = y \Rightarrow \lambda^* = (X^\top X)^{-1}X^\top w_0 - (X^\top X)^{-1}y$$

$$\Rightarrow w^* = -X(X^\top X)^{-1}X^\top w_0 + X(X^\top X)^{-1}y + w_0 = (I - P)w_0 + X^\dagger y$$

□

Lemma 7. *Assume matrices $X_1 \in \mathbb{R}^{p\times n_1}$ and $X_2 \in \mathbb{R}^{p\times n_2}$ to be random matrices with entries being independently sampled from the standard Gaussian distribution $\mathcal{N}(0,1)$. Assume $p > n_1 + n_2 + 1$ and let $P_1 = X_1(X_1^\top X_1)^{-1}X_1^\top$ be the orthogonal projection matrix that projects onto the column space of X_1. Also assume $X_{12} \in \mathbb{R}^{p\times(n_1+n_2)}$ be the result of concatenation of X_1 and X_2, and $X_{01} \in \mathbb{R}^{p\times(n_1+n_2)}$ be the result of concatenation of X_1 with a zero matrix. Additionally, let $P_{12} = X_{12}(X_{12}^\top X_{12})^{-1}X_{12}^\top$ and $P_{01} = X_{12}(X_{12}^\top X_{12})^{-1}X_{01}^\top$. Assuming $w \in \mathbb{R}^p$ is a fixed given vector, the following equalities hold:*

(i) $\mathbb{E}[\|P_1 w\|^2] = \frac{n_1}{p}\|w\|^2$

(ii) $\mathbb{E}[\|(I - P_1)w\|^2] = (1 - \frac{n_1}{p})\|w\|^2$

(iii) $\mathbb{E}[\|P_{01}w\|^2] = \frac{n_1}{p}(1 + \frac{n_2}{p-(n_1+n_2)-1})\|w\|^2$

Proof.

(i) Without loss of generality, we can focus on finding $\mathbb{E}[\|P_1 u\|^2]$ where $u \in \mathbb{R}^p$ and $\|u\|^2 = 1$, since X_1 is a Gaussian matrix with rank n_1, and $P_1 = X_1(X_1^\top X_1)^{-1}X_1^\top$ is an orthogonal projection matrix with a similar rank. Due to the rotational invariance of the standard normal distribution, we can assume that P_1 is a fixed matrix and instead, u is a random vector uniformly sampled from the unit sphere in $\mathbb{R}^p$. Using the rotational invariance again, we may assume without the loss of generality that P_1 is the coordinate projection onto the first n_1 coordinates in $\mathbb{R}^p$. Then it holds that:

$$\mathbb{E}_{X_1}[\|P_1 u\|^2] = \mathbb{E}_u[\sum_{i=1}^{n_1} u_i^2] = \frac{n_1}{p}$$

(ii) $I - P_1$ is a projection orthogonal to P_1. Therefore, by Pythagorean theorem,

$$\mathbb{E}[\|(I - P_1)w\|^2] = \|w\|^2 - \mathbb{E}[\|P_1 w\|^2] = (1 - \frac{n_1}{p})\|w\|^2$$

(iii)

$$\begin{aligned}
\mathbb{E}[\|P_{01}w\|^2] &= \mathbb{E}[w^\top P_{01}^\top P_{01} w] \\
&= \mathbb{E}[w^\top P_1 w] + \mathbb{E}[w^\top P_1 X_2 (X_2^\top (I - P_1) X_2)^{-1} X_2^\top P_1 w] && \text{(Lemma 3 part (iv))} \\
&= \frac{n_1}{p}\|w\|^2 + \mathbb{E}[\mathrm{tr}(w^\top P_1 X_2 (X_2^\top (I - P_1) X_2)^{-1} X_2^\top P_1 w)]
\end{aligned}$$

$$= \frac{n_1}{p}\|w\|^2 + \mathbb{E}_{X_1}[\mathrm{tr}(\mathbb{E}_{X_2}[X_2^\top P_1 w w^\top P_1 X_2 (X_2^\top (I-P_1)X_2)^{-1}])]$$

To find the above expectation over X_2, notice that P_1 is an orthogonal projection matrix and, therefore, for a fixed P_1, $P_1 X_2$ is independent of $(I-P_1)X_2$.

$$\begin{aligned}
\Rightarrow \mathbb{E}[\|P_{01}w\|^2] &= \frac{n_1}{p}\|w\|^2 + \mathbb{E}_{X_1}[\mathrm{tr}(\mathbb{E}_{X_2}[X_2^\top P_1 w w^\top P_1 X_2]\mathbb{E}_{X_2}[(X_2^\top (I-P_1)X_2)^{-1}])] \\
&= \frac{n_1}{p}\|w\|^2 + \mathbb{E}_{X_1}[\mathrm{tr}(\mathrm{tr}(P_1 w w^\top P_1)\mathbb{E}_{X_2}[(X_2^\top (I-P_1)X_2)^{-1}])] \\
&= \frac{n_1}{p}\|w\|^2 + \mathbb{E}_{X_1}[\|P_1 w\|^2 \,\mathrm{tr}(\mathbb{E}_{X_2}[(X_2^\top (I-P_1)X_2)^{-1}])]
\end{aligned}$$

We focus on finding the inner expectation first. Notice that for a fixed X_1, $I-P_1$ is an orthogonal projection matrix with rank $p-n_1$. Due to the rotational invariance of the standard normal distribution, we may assume without loss of generality that $I-P_1$ is the projection matrix that projects onto the first $p-n_1$ coordinates. With this in mind, $(X_2^\top (I-P_1)X_2)^{-1}$ follows an inverse-Wishart distribution with an identity scale matrix $I_{n_2\times n_2} \in \mathbb{R}^{n_2}$ and $p-n_1$ degrees of freedom.

$$\begin{aligned}
\Rightarrow \mathbb{E}[\|P_{01}w\|^2] &= \frac{n_1}{p}\|w\|^2 + \mathbb{E}_{X_1}[\|P_1 w\|^2 \,\mathrm{tr}(\frac{I_{n_2\times n_2}}{p-n_1-n_2-1})] \\
&= \frac{n_1}{p}\|w\|^2 + \frac{n_2}{p-(n_1+n_2)-1}\mathbb{E}_{X_1}[\|P_1 w\|^2] \\
&= \frac{n_1}{p}(1 + \frac{n_2}{p-(n_1+n_2)-1})\|w\|^2
\end{aligned}$$

□

Lemma 8. *Assume matrices $X_1 \in \mathbb{R}^{p\times n_1}$, $X_2 \in \mathbb{R}^{p\times n_2}$ and $X_3 \in \mathbb{R}^{p\times n_3}$ to be random matrices with entries being independently sampled from the standard Gaussian distribution $\mathcal{N}(0,1)$. Assume $p > n_1+n_2+n_3+1$ and let $X_{13} \in \mathbb{R}^{p\times(n_1+n_2+n_3)}$ be the result of concatenation of X_1, X_2, X_3. Also let $X_{01}, X_{02}, X_{03} \in \mathbb{R}^{p\times(n_1+n_2+n_3)}$ be the result of concatenating X_1, X_2 and X_3 with zero matrices such that $X_{01} = \begin{bmatrix} X_1 & 0 & 0 \end{bmatrix}$, $X_{02} = \begin{bmatrix} 0 & X_2 & 0 \end{bmatrix}$ and $X_{03} = \begin{bmatrix} 0 & 0 & X_3 \end{bmatrix}$. Define $P_{01} = X_{13}(X_{13}^\top X_{13})^{-1}X_{01}^\top$ and $P_{03} = X_{13}(X_{13}^\top X_{13})^{-1}X_{03}^\top$. Assuming $w, w' \in \mathbb{R}^p$ are fixed given vectors, it holds that*

$$\mathbb{E}[w^\top P_{01}^\top P_{03} w'] = -\frac{n_1 n_3}{p(p-(n_1+n_2+n_3)-1)}\langle w, w'\rangle$$

Proof. Based on Lemma 4, we have

$$\begin{aligned}
\mathbb{E}[w^\top P_{01}^\top P_{03} w'] &= -\mathbb{E}[w^\top P_1(I-\widehat{P}_{12})X_3(X_3^\top (I-P_1)(I-\widehat{P}_{12})X_3)^{-1}X_3^\top w'] \\
&= -\mathbb{E}_{X_1,X_2}[\mathbb{E}_{X_3}[w^\top P_1(I-\widehat{P}_{12})X_3(X_3^\top (I-P_1)(I-\widehat{P}_{12})X_3)^{-1}X_3^\top w']]
\end{aligned}$$

where $\widehat{P}_{12} = X_2(X_2^\top (I-P_1)X_2)^{-1}X_2^\top (I-P_1)$. Now denote $P = (I-P_1)(I-\widehat{P}_{12})$. With a fixed X_1 and X_2, P is an orthogonal projection matrix, meaning that $P^\top = P^2 = P$. Therefore,

$$\begin{aligned}
\mathbb{E}[w^\top P_{01}^\top P_{03} w'] &= -\mathbb{E}_{X_1,X_2}[\mathbb{E}_{X_3}[w^\top P_1(I-\widehat{P}_{12})X_3(X_3^\top P X_3)^{-1}X_3^\top (P+I-P)w']] \\
&= -\mathbb{E}_{X_1,X_2}[\mathbb{E}_{X_3}[w^\top P_1(I-\widehat{P}_{12})X_3(X_3^\top P X_3)^{-1}X_3^\top P w']]
\end{aligned}$$

$$- \mathbb{E}_{X_1,X_2}[\mathbb{E}_{X_3}[w^\top P_1(I-\widehat{P}_{12})X_3(X_3^\top PX_3)^{-1}X_3^\top(I-P)w']]$$

Notice that for a fixed X_1 and X_2, P_1 and P are two orthogonal projection matrices. Therefore, the following independence relations hold:

$$(I-P)X_3 \perp\!\!\!\perp PX_3$$
$$P_1(I-\widehat{P}_{12})X_3 \perp\!\!\!\perp (I-P_1)(I-\widehat{P}_{12})X_3$$

With this in mind, we can write

$$\begin{aligned}
\mathbb{E}[w^\top P_{01}^\top P_{03}w'] &= -\mathbb{E}_{X_1,X_2}[w^\top P_1 \mathbb{E}_{X_3}[(I-\widehat{P}_{12})X_3]\mathbb{E}_{X_3}[(X_3^\top PX_3)^{-1}X_3^\top Pw']] \\
&\quad - \mathbb{E}_{X_1,X_2}[\mathbb{E}_{X_3}[\mathrm{tr}(X_3^\top(I-P)w'w^\top P_1(I-\widehat{P}_{12})X_3(X_3^\top PX_3)^{-1})]] \\
&= 0 - \mathbb{E}_{X_1,X_2}[\mathrm{tr}(\mathbb{E}_{X_3}[X_3^\top(I-P)w'w^\top P_1(I-\widehat{P}_{12})X_3]\mathbb{E}_{X_3}[(X_3^\top PX_3)^{-1}])] \\
&= -\mathbb{E}_{X_1,X_2}[\mathrm{tr}((I-P)w'w^\top P_1(I-\widehat{P}_{12}))\,\mathrm{tr}(\mathbb{E}_{X_3}[(X_3^\top PX_3)^{-1}])]
\end{aligned}$$

Here we again use a technique similar to the one we used in Lemma 7 part (iii) by first focusing on the inner expectation. Assume a fixed X_1 and X_2. Then $I-P_1$ and P_1 are orthogonal projections with ranks $p-n_1$ and n_1 respectively. Additionally, $(I-P_1)\widehat{P}_{12}$ is an orthogonal projection matrix with rank n_2 and $p-n_1 > n_2$. On the other hand, P_1 and $(I-P_1)\widehat{P}_{12}$ are orthogonal projections. Therefore, $P_1+(I-P_1)\widehat{P}_{12}$ is a projection matrix with rank n_1+n_2 and $P = I-(P_1+(I-P_1)\widehat{P}_{12})$ is a projection matrix with rank $p-(n_1+n_2)$. When taking the inner expectation only w.r.t. X_3, due to the rotational invariance property, we may assume that $(I-P_1)(I-\widehat{P}_{12})$ is the coordinate projection onto the first $p-(n_1+n_2)$ coordinates. Therefore, $(X_3^\top(I-P_1)(I-\widehat{P}_{12})X_3)^{-1} \sim \mathcal{IW}(I_{n_3\times n_3}, p-(n_1+n_2))$ and we have:

$$\begin{aligned}
\Rightarrow \mathbb{E}[w^\top P_{01}^\top P_{03}w'] &= -\mathbb{E}_{X_1,X_2}[\mathrm{tr}((I-P)w'w^\top P_1(I-\widehat{P}_{12}))\,\mathrm{tr}(\frac{I_{n_3\times n_3}}{p-(n_1+n_2)-n_3-1})] \\
&= -\frac{n_3}{p-(n_1+n_2+n_3)-1}\mathbb{E}_{X_1,X_2}[\mathrm{tr}(w^\top P_1(I-\widehat{P}_{12})(I-P)w')]
\end{aligned}$$

Now, using the definition of P and $\widehat{P}_{12}$, we have

$$\begin{aligned}
&\mathbb{E}[w^\top P_{01}^\top P_{03}w'] = -\frac{n_3}{p-(n_1+n_2+n_3)-1}\mathbb{E}[\mathrm{tr}(w^\top P_1(I-\widehat{P}_{12})w')] \\
&= -\frac{n_3}{p-(n_1+n_2+n_3)-1}\mathbb{E}[\mathrm{tr}(w^\top P_1 w')] \\
&+ \frac{n_3}{p-(n_1+n_2+n_3)-1}\mathbb{E}[\mathrm{tr}(w^\top P_1X_2(X_2^\top(I-P_1)X_2)^{-1}X_2^\top(I-P_1)w')]
\end{aligned}$$

Remember that for a given X_1, P_1X_2 and $(I-P_1)X_2$ are independent. Therefore,

$$\begin{aligned}
&\mathbb{E}[w^\top P_{01}^\top P_{03}w'] = -\frac{n_3}{p-(n_1+n_2+n_3)-1}\mathbb{E}[\mathrm{tr}(w^\top P_1 w')] \\
&= -\frac{n_3}{p-(n_1+n_2+n_3)-1}\times\frac{1}{4}\mathbb{E}[(w+w')^\top P_1(w+w') - (w-w')^\top P_1(w-w')] \\
&= -\frac{n_3}{p-(n_1+n_2+n_3)-1}\times\frac{1}{4}\mathbb{E}[\|P_1(w+w')\|^2 - \|P_1(w-w')\|^2]
\end{aligned}$$

$$= -\frac{n_1 n_3}{p-(n_1+n_2+n_3)-1} \times \frac{1}{4}(\|w+w'\|^2 - \|w-w'\|^2) \qquad \text{(Lemma 7 part (i))}$$
$$= -\frac{n_3 n_1}{p(p-(n_1+n_2+n_3)-1)}\langle w, w'\rangle$$

□

4.3.2 Proof of Theorem 4

Now that we have all the required building blocks, we are ready to provide the proofs. We start by proving the theorem related to the single-task learner and continue by proving the theorems on multi-task learning and continual learning in this and next subsubsections.

Proof. **underparameterized regime.** In the underparameterized regime where $n_t \geq p+2$, the signle-task learner is the unique solution to the Equation 4 and is obtained by $w_t = (X_t X_t^\top)^{-1} X_t y_t$ according to Lemma 5. Therefore, we must have

$$w_t = (X_t X_t^\top)^{-1} X_t y_t = (X_t X_t^\top)^{-1} X_t (X_t^\top w_t^* + z_t) = w_t^* + (X_t X_t^\top)^{-1} X_t z_t$$
$$\Rightarrow \mathbb{E}[\mathcal{L}_t(w)] = \mathbb{E}[\|w_t - w_t^*\|^2] = \mathbb{E}[\|(X_t X_t^\top)^{-1} X_t z_t\|^2]$$

where the expectation is due to randomness of both X_t and z_t. First, we take the expectation w.r.t. $z_t \sim \mathcal{N}_{n_t}(0, \sigma^2 I_{n_t \times n_t})$:

$$\begin{aligned}
\mathbb{E}[\mathcal{L}_t(w)] &= \mathbb{E}[z_t^\top X_t^\top (X_t X_t^\top)^{-1}(X_t X_t^\top)^{-1} X_t z_t] \\
&= \sigma^2 \mathbb{E}[\mathrm{tr}(X_t^\top (X_t X_t^\top)^{-2} X_t)] \\
&= \sigma^2 \mathbb{E}[\mathrm{tr}((X_t X_t^\top)^{-2} X_t X_t^\top)] \\
&= \sigma^2 \mathbb{E}[\mathrm{tr}((X_t X_t^\top)^{-1})] \\
&= \sigma^2 \,\mathrm{tr}(\mathbb{E}[(X_t X_t^\top)^{-1}])) \\
&= \sigma^2 \,\mathrm{tr}(\frac{I_{p\times p}}{n_t - p - 1}) \\
&= \frac{p\sigma^2}{n_t - p - 1}
\end{aligned}$$

where the two last line comes from the fact that $(X_t X_t^\top)^{-1}$ follows the inverse-Wishart distribution as $\mathcal{IW}(I_{p\times p}, n)$.

overparameterized regime. In the overparameterized regime where $p \geq n_t + 2$, we look for the solution of optimization 5. Therefore, based on Lemma 6, we can have

$$\begin{aligned}
\mathbb{E}[\mathcal{L}_t(w)] &= \mathbb{E}[\|w_t - w_t^*\|^2] \\
&= \mathbb{E}[\|X_t^\dagger y_t - w_t^*\|^2] \\
&= \mathbb{E}[\|X_t^\dagger (X_t^\top w_t^* + z_t) - w_t^*\|^2] \\
&= \mathbb{E}[\|(I - P_t) w_t^* - X_t^\dagger z_t\|^2] \\
&= \mathbb{E}[\|(I - P_t) w_t^*\|^2] + \mathbb{E}[\|X_t^\dagger z_t\|^2] - 2 w_t^{*T} \mathbb{E}[(I - P_t) X_t^\dagger] \mathbb{E}[z_t] \\
&= \mathbb{E}[\|(I - P_t) w_t^*\|^2] + \mathbb{E}[z_t^\top (X_t^\top X_t)^{-1} z_t] \qquad (\mathbb{E}[z_t] = 0)
\end{aligned}$$

$$= \mathbb{E}[\|(I - P_t)w_t^*\|^2] + \frac{n_t\sigma^2}{p - n_t - 1} \qquad ((X_t^\top X_t)^{-1} \sim \mathcal{IW}(I_{n_t \times n_t}, p))$$
$$= (1 - \frac{n_t}{p})\|w_t^*\|^2 + \frac{n_t\sigma^2}{p - n_t - 1} \qquad \text{(Lemma 7 part (ii))}$$

□

4.3.3 Proof of Theorem 1

Proof. Based on Lemma 6 we can write

$$w_{1T} = X_{1T}(X_{1T}^\top X_{1T})^{-1} y_{1T}$$

Using the notations introduced in the previous section, we have

$$w_{1T} = X_{1T}(X_{1T}^\top X_{1T})^{-1}(\sum_{s=1}^{T} X_{0s}^\top w_s^* + z_{1T}) = \sum_{s=1}^{T} P_{0s} w_s^* + X_{1T}^\dagger z_{1T}$$

Now we start by calculating the multi-task learner's loss for the task i:

$$\begin{aligned}
&\mathbb{E}[\mathcal{L}_i(w_{1T})] = \mathbb{E}[\|w_{1T} - w_i^*\|^2] \\
&= \mathbb{E}[\|\sum_{s=1}^{T} P_{0s} w_s^* + X_{1T}^\dagger z_{1T} - w_i^*\|^2] \\
&= \mathbb{E}[\|\sum_{s=1}^{T} P_{0s}(w_s^* - w_i^*) + (I - P_{1T})w_i^* + X_{1T}^\dagger z_{1T}\|^2] \\
&= \mathbb{E}[\|\sum_{s=1}^{T} P_{0s}(w_s^* - w_i^*) + (I - P_{1T})w_i^*\|^2] + \mathbb{E}[\|X_{1T}^\dagger z_{1T}\|^2] \qquad (z_{1T} \perp\!\!\!\perp X_{1T}, \mathbb{E}[z_{1T}] = 0) \\
&= \mathbb{E}[\|\sum_{s=1}^{T} P_{0s}(w_s^* - w_i^*) + (I - P_{1T})w_i^*\|^2] + \sigma^2 \operatorname{tr}(\mathbb{E}[(X_{1T}X_{1T})^{-1}]) \qquad (z_{1T} \sim \mathcal{N}(0, \sigma^2 I)) \\
&= \mathbb{E}[\|\sum_{s=1}^{T} P_{0s}(w_s^* - w_i^*) + (I - P_{1T})w_i^*\|^2] + \frac{\bar{n}\sigma^2}{p - \bar{n} - 1} \qquad ((X_{1T}X_{1T})^{-1} \sim \mathcal{IW}(I_{\bar{n} \times \bar{n}}, p))
\end{aligned}$$

Therefore, we focus on finding the first term by expanding it. First notice that

$$\begin{aligned}
(I - P_{1T})P_{0s} &= (I - X_{1T}(X_{1T}^\top X_{1T})^{-1} X_{1T}^\top) X_{1T}(X_{1T}^\top X_{1T})^{-1} X_{0s}^\top \\
&= X_{1T}(X_{1T}^\top X_{1T})^{-1} X_{0s}^\top - X_{1T}(X_{1T}^\top X_{1T})^{-1} X_{0s}^\top \\
&= 0
\end{aligned}$$

Thus,

$$\mathbb{E}[\|\sum_{s=1}^{T} P_{0s}(w_s^* - w_i^*) + (I - P_{1T})w_i^*\|^2]$$

$$= \mathbb{E}[\| \sum_{s=1}^{T} P_{0s}(w_s^* - w_i^*)\|^2] + \mathbb{E}[\|(I - P_{1T})w_i^*\|^2]$$
$$= \mathbb{E}[\| \sum_{s=1}^{T} P_{0s}(w_s^* - w_i^*)\|^2] + (1 - \frac{\bar{n}}{p})\|w_i^*\|^2 \qquad \text{(Lemma 7 part (ii))}$$
$$= \sum_{s=1}^{T} \mathbb{E}[\|P_{0s}(w_s^* - w_i^*)\|^2] + \sum_{s=1}^{T} \sum_{\substack{s'=1 \\ s' \neq s}}^{T} \mathbb{E}[(w_s^* - w_i^*)^\top P_{0s}^\top P_{0s'}(w_{s'}^* - w_i^*)] + (1 - \frac{\bar{n}}{p})\|w_i^*\|^2$$

Since the distribution of P_{0s} is invariant to the permutation of columns of X_{1T}, we can focus on finding $\mathbb{E}[\|P_{01}(w_1^* - w_i^*)\|^2]$ and $\mathbb{E}[(w_1^* - w_i^*)^\top P_{01}^\top P_{0T}(w_T^* - w_i^*)]$ without loss of generality. In Lemmas 7 and 8, we have already calculated similar quantities. Therefore,

$$\mathbb{E}[\|P_{01}(w_1^* - w_i^*)\|^2] = \frac{n_1}{p}(1 + \frac{\bar{n} - n_1}{p - \bar{n} - 1})\|w_1^* - w_i^*\|^2$$

and

$$\mathbb{E}[(w_1^* - w_i^*)^\top P_{01}^\top P_{0T}(w_T^* - w_i^*)] = -\frac{n_T n_1}{p(p - \bar{n} - 1)}\langle w_1^* - w_i^*, w_T^* - w_i^* \rangle$$

Overall, we can write :

$$\mathbb{E}[\mathcal{L}_i(w_{1T})]$$
$$= -\frac{1}{p}\sum_{s=1}^{T}\sum_{\substack{s'=1 \\ s' \neq s}}^{T} \frac{n_s n_{s'}}{p - \bar{n} - 1}\langle w_s^* - w_i^*, w_{s'}^* - w_i^* \rangle + \sum_{s=1}^{\top} \frac{n_s}{p}(1 + \frac{\bar{n} - n_s}{p - \bar{n} - 1})\|w_s^* - w_i^*\|^2$$
$$+ (1 - \frac{\bar{n}}{p})\|w_i^*\|^2 + \frac{\bar{n}\sigma^2}{p - \bar{n} - 1}$$
$$= -\frac{1}{p}\sum_{s=1}^{T}\sum_{s'=1}^{T} \frac{n_s n_{s'}}{p - \bar{n} - 1}\langle w_s^* - w_i^*, w_{s'}^* - w_i^* \rangle + \sum_{s=1}^{\top} \frac{n_s}{p}(1 + \frac{\bar{n}}{p - \bar{n} - 1})\|w_s^* - w_i^*\|^2$$
$$+ (1 - \frac{\bar{n}}{p})\|w_i^*\|^2 + \frac{\bar{n}\sigma^2}{p - \bar{n} - 1}$$
$$= \frac{1}{2p}\sum_{s=1}^{T}\sum_{s'=1}^{T} \frac{n_s n_{s'}}{p - \bar{n} - 1}(\|w_s^* - w_{s'}^*\|^2 - \|w_s^* - w_i^*\|^2 - \|w_{s'}^* - w_i^*\|^2)$$
$$+ \sum_{s=1}^{\top} \frac{n_s}{p}(1 + \frac{\bar{n}}{p - \bar{n} - 1})\|w_s^* - w_i^*\|^2 + (1 - \frac{\bar{n}}{p})\|w_i^*\|^2 + \frac{\bar{n}\sigma^2}{p - \bar{n} - 1}$$
$$= \frac{1}{2p}\sum_{s=1}^{T}\sum_{s'=1}^{T} \frac{n_s n_{s'}}{p - \bar{n} - 1}\|w_s^* - w_{s'}^*\|^2 + \sum_{s=1}^{\top} \frac{n_s}{p}\|w_s^* - w_i^*\|^2 + (1 - \frac{\bar{n}}{p})\|w_i^*\|^2 + \frac{\bar{n}\sigma^2}{p - \bar{n} - 1} \qquad (22)$$

Now we calculate the desired metrics as

$$\mathbb{E}[G(w_{1T})] = \frac{1}{T}\sum_{i=1}^{T} \mathbb{E}[\mathcal{L}_i(w_{1T})]$$

$$
\begin{aligned}
&= \frac{1}{2p}\sum_{s=1}^{T}\sum_{s'=1}^{T}\frac{n_s n_{s'}}{p-\bar{n}-1}\|w_s^*-w_{s'}^*\|^2 + \frac{1}{T}\sum_{i=1}^{T}\sum_{s=1}^{\top}\frac{n_s}{p}\|w_s^*-w_i^*\|^2 \\
&+ \frac{1}{T}\sum_{i=1}^{T}(1-\frac{\bar{n}}{p})\|w_i^*\|^2 + \frac{\bar{n}\sigma^2}{p-\bar{n}-1} \\
&= \frac{1}{T}\sum_{i=1}^{T}\sum_{s=1}^{\top}\frac{n_s}{p}(1+\frac{\frac{T}{2}n_i}{p-\bar{n}-1})\|w_s^*-w_i^*\|^2 + \frac{1}{T}(1-\frac{\bar{n}}{p})\sum_{i=1}^{T}\|w_i^*\|^2 + \frac{\bar{n}\sigma^2}{p-\bar{n}-1}
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}[K(w_{1T})] &= \frac{1}{T}\sum_{i=1}^{T}\mathbb{E}[\mathcal{L}_i(w_i)] - \mathbb{E}[\mathcal{L}_i(w_{1T})] \\
&= \frac{1}{T}\sum_{i=1}^{T}(1-\frac{n_i}{p})\|w_i^*\|^2 + \frac{1}{T}\sum_{i=1}^{T}\frac{n_i\sigma^2}{p-n_i-1} \\
&- \frac{1}{T}\sum_{i=1}^{T}\sum_{s=1}^{\top}\frac{n_s}{p}(1+\frac{\frac{T}{2}n_i}{p-\bar{n}-1})\|w_s^*-w_i^*\|^2 - \frac{1}{T}(1-\frac{\bar{n}}{p})\sum_{i=1}^{T}\|w_i^*\|^2 - \frac{\bar{n}\sigma^2}{p-\bar{n}-1} \\
&= \frac{1}{T}\sum_{i=1}^{T}(1-\frac{n_i}{p})\|w_i^*\|^2 + \frac{1}{T}\sum_{i=1}^{T}\frac{n_i\sigma^2}{p-n_i-1} \\
&- \frac{1}{T}\sum_{i=1}^{T}\sum_{s=1}^{\top}\frac{n_s}{p}(1+\frac{\frac{T}{2}n_i}{p-\bar{n}-1})(\|w_s^*\|^2+\|w_i^*\|^2-2\langle w_s^*,w_i^*\rangle) \\
&- \frac{1}{T}(1-\frac{\bar{n}}{p})\sum_{i=1}^{T}\|w_i^*\|^2 - \frac{\bar{n}\sigma^2}{p-\bar{n}-1} \\
&= \frac{2}{T}\sum_{i=1}^{T}\sum_{s=1}^{\top}\frac{n_s}{p}(1+\frac{\frac{T}{2}n_i}{p-\bar{n}-1})\langle w_s^*,w_i^*\rangle - \frac{1}{T}\sum_{s=1}^{\top}\frac{n_s}{p}(T+\frac{\frac{T}{2}\bar{n}}{p-\bar{n}-1})\|w_s^*\|^2 \\
&- \frac{1}{T}\sum_{i=1}^{T}\frac{\bar{n}}{p}(1+\frac{\frac{T}{2}n_i}{p-\bar{n}-1})\|w_i^*\|^2 - \frac{1}{T}(1-\frac{\bar{n}}{p})\sum_{i=1}^{T}\|w_i^*\|^2 - \frac{\bar{n}\sigma^2}{p-\bar{n}-1} \\
&+ \frac{1}{T}\sum_{i=1}^{T}(1-\frac{n_i}{p})\|w_i^*\|^2 + \frac{1}{T}\sum_{i=1}^{T}\frac{n_i\sigma^2}{p-n_i-1} \\
&= \frac{2}{T}\sum_{i=1}^{T}\sum_{s=1}^{\top}\frac{n_s}{p}(1+\frac{\frac{T}{2}n_i}{p-\bar{n}-1})\langle w_s^*,w_i^*\rangle \\
&- \frac{1}{T}\sum_{s=1}^{\top}(\frac{Tn_s}{p}+\frac{n_s}{p}\frac{\frac{T}{2}\bar{n}}{p-\bar{n}-1}+\frac{\bar{n}}{p}\frac{\frac{T}{2}n_s}{p-\bar{n}-1}+\frac{n_s}{p})\|w_s^*\|^2 \\
&+ \frac{1}{T}\sum_{i=1}^{T}\frac{n_i\sigma^2}{p-n_i-1} - \frac{\bar{n}\sigma^2}{p-\bar{n}-1}
\end{aligned}
$$

$$= \frac{2}{T}\sum_{i=1}^{T}\sum_{s=1}^{\top}\frac{n_s}{p}(1+\frac{\frac{T}{2}n_i}{p-\bar{n}-1})\langle w_s^*, w_i^*\rangle - \sum_{s=1}^{\top}\frac{n_s}{p}(1+\frac{1}{T}+\frac{\bar{n}}{p-\bar{n}-1})\|w_s^*\|^2$$
$$+\frac{1}{T}\sum_{i=1}^{T}\frac{n_i\sigma^2}{p-n_i-1} - \frac{\bar{n}\sigma^2}{p-\bar{n}-1}$$

□

Next, let $\sigma = 0$. Assume the equal number of samples $n_1 = n_2 = ... = n$ and the same task norm for all tasks $\|w_1^*\|^2 = ... = \|w_T^*\|^2$. Also assume $\langle w_s^*, w_i^*\rangle = \|w_1^*\|^2\cos\theta$ for all pair of tasks. The knowledge transfer then simplifies to:

$$\mathbb{E}[K(w_{1T})] = \frac{2}{T}\frac{n}{p}(1+\frac{\frac{T}{2}n}{p-Tn-1})\sum_{i=1}^{T}\sum_{s=1}^{\top}\langle w_s^*, w_i^*\rangle - \frac{n}{p}(1+\frac{1}{T}+\frac{Tn}{p-Tn-1})\sum_{s=1}^{\top}\|w_s^*\|^2$$
$$= \frac{2}{T}\frac{n}{p}(1+\frac{\frac{T}{2}n}{p-Tn-1})((T^2-T)\cos\theta+T)\|w_1^*\|^2 - \frac{Tn}{p}(1+\frac{1}{T}+\frac{Tn}{p-Tn-1})\|w_1^*\|^2$$

Now, solving $\mathbb{E}[K(w_{1T})] = 0$ yields

$$\frac{2}{T}(1+\frac{\frac{T}{2}n}{p-Tn-1})((T-1)\cos\theta+1) = (1+\frac{1}{T}+\frac{Tn}{p-Tn-1})$$
$$\Rightarrow (2+\frac{Tn}{p-Tn-1})((T-1)\cos\theta+1) = (T+1+\frac{T^2n}{p-Tn-1})$$
$$\Rightarrow (2+\frac{Tn}{p-Tn-1})(T-1)\cos\theta = (T-1+\frac{T(T-1)n}{p-Tn-1})$$
$$\Rightarrow (2p-Tn-2)\cos\theta = p-1$$
$$\Rightarrow \cos\theta = \frac{p-1}{2p-Tn-2}$$

4.3.4 Proof of Theorem 5 and 2

To provide proof for both theorems, we start by presenting a more general expression and derive both theorems as special cases. We start by finding the loss of task i at timestep t, i.e. $\mathbb{E}[\mathcal{L}_i(w_{\boldsymbol{vt}})]$, using an intermediate result in the proof provided in section 4.3.3. In fact, we refer to Equation 22 and assume $n_1 = n_2 = ... = n_{t-1} = m$, $n_t = n$, $n_{t+1} = ... = n_T = 0$ and denote $\bar{n}_t = (t-1)m + n$ to achieve:

$$\mathbb{E}[\mathcal{L}_i(w_{\boldsymbol{vt}})] = \frac{1}{2p}\sum_{s=1}^{t-1}\sum_{s'=1}^{t-1}\frac{m^2}{p-\bar{n}_t-1}\|w_s^*-w_{s'}^*\|^2 + \frac{1}{p}\sum_{s=1}^{t-1}\frac{mn}{p-\bar{n}_t-1}\|w_t^*-w_s^*\|^2$$
$$+\sum_{s=1}^{t-1}\frac{m}{p}\|w_s^*-w_i^*\|^2 + \frac{n}{p}\|w_t^*-w_i^*\|^2 + (1-\frac{\bar{n}_t}{p})\|w_i^*\|^2 + \frac{\bar{n}_t\sigma^2}{p-\bar{n}_t-1} \tag{23}$$

Proof of Theorem 5

Proof. Set $m = 0$ in Equation 23, to retrieve the loss for the naive sequential learner:

$$\mathbb{E}[\mathcal{L}_i(w_t)] = \frac{n}{p}\|w_t^* - w_i^*\|^2 + (1 - \frac{n}{p})\|w_i^*\|^2 + \frac{n\sigma^2}{p-n-1}$$

Therefore,

$$\begin{aligned}\mathbb{E}[G(w_T)] &= \frac{1}{T}\sum_{i=1}^{T}\mathbb{E}[\mathcal{L}_i(w_T)] \\ &= \frac{n}{Tp}\sum_{i=1}^{T-1}\|w_T^* - w_i^*\|^2 + \frac{1}{T}(1-\frac{n}{p})\sum_{i=1}^{T}\|w_i^*\|^2 + \frac{n\sigma^2}{p-n-1}\end{aligned}$$

and for the forgetting, we have:

$$\begin{aligned}\mathbb{E}[F(w_T)] &= \frac{1}{T-1}\sum_{i=1}^{T}\mathbb{E}[\mathcal{L}_i(w_T)] - \mathbb{E}[\mathcal{L}_i(w_i)] \\ &= \frac{n}{(T-1)p}\sum_{i=1}^{T-1}\|w_T^* - w_i^*\|^2 + \frac{1}{T-1}(1-\frac{n}{p})\sum_{i=1}^{T}\|w_i^*\|^2 + \frac{n\sigma^2}{p-n-1} \\ &- \frac{1}{T-1}(1-\frac{n}{p})\sum_{i=1}^{T}\|w_i^*\|^2 - \frac{n\sigma^2}{p-n-1} \\ &= \frac{n}{(T-1)p}\sum_{i=1}^{T-1}\|w_T^* - w_i^*\|^2\end{aligned}$$

□

Proof of Theorem 2

Proof. We set $T = 2$ in Equation 23 to get:

$$\mathbb{E}[\mathcal{L}_1(w_{\boldsymbol{v}2})] = \frac{n}{p}(1 + \frac{m}{(p-(m+n)-1)})\|w_1^* - w_2^*\|^2 + (1 - \frac{m+n}{p})\|w_1^*\|^2 + \frac{(m+n)\sigma^2}{p-(m+n)-1}$$

and

$$\mathbb{E}[\mathcal{L}_2(w_{\boldsymbol{v}2})] = \frac{n}{p}(\frac{m}{n} + \frac{m}{(p-(m+n)-1)})\|w_2^* - w_1^*\|^2 + (1 - \frac{m+n}{p})\|w_2^*\|^2 + \frac{(m+n)\sigma^2}{p-(m+n)-1}$$

Therefore, the generalization error is

$$\begin{aligned}\mathbb{E}[G(w_{\boldsymbol{v}2})] &= \frac{1}{2}(\mathbb{E}[\mathcal{L}_1(w_{\boldsymbol{v}2})] + \mathbb{E}[\mathcal{L}_2(w_{\boldsymbol{v}2})]) \\ &= \frac{n}{2p}(1 + \frac{m}{n} + \frac{2m}{(p-(m+n)-1)})\|w_2^* - w_1^*\|^2 + (1 - \frac{m+n}{p})\|w_1^*\|^2 + \frac{(m+n)\sigma^2}{p-(m+n)-1}\end{aligned}$$

Similarly,

$$\mathbb{E}[\mathcal{L}_1(w_{\boldsymbol{v}1})] = (1 - \frac{n}{p})\|w_1^*\|^2 + \frac{n\sigma^2}{p-n-1}$$

Therefore,

$$\begin{aligned}\mathbb{E}[F(w_{\boldsymbol{v}2})] &= \mathbb{E}[\mathcal{L}_1(w_{\boldsymbol{v}2})] - \mathbb{E}[\mathcal{L}_1(w_{\boldsymbol{v}1})] \\ &= \frac{n}{p}(1 + \frac{m}{(p-(m+n)-1)})\|w_1^* - w_2^*\|^2 - \frac{m}{p}\|w_1^*\|^2 + \frac{(m+n)\sigma^2}{p-(m+n)-1} - \frac{n\sigma^2}{p-n-1}\end{aligned}$$

□

4.3.5 Proof of Theorem 3

We start by proving a theorem for the general case:

Theorem 6. *Assume $m_1 = m_2 = ... = m_t = m$ and $n_{t+1} = n$. Also denote $\bar{n}_{t+1} = tm + n$ and let $\alpha_t = 1 - \frac{n}{p-tm}$. Considering the continual learner described in Equation 13 when $p \geq \bar{n}_t + 2$, the loss of task i at timestep t, can be recursively calculated as:*

$$\begin{aligned}\mathbb{E}[\mathcal{L}_i(w_{\boldsymbol{v}t+1})] &= \mathbb{E}[\|w_{\boldsymbol{v}t+1} - w_i^*\|^2] \\ &= \frac{m}{p}(1-\alpha_t)\sum_{s=1}^{t}\|w_s^* - w_i^*\|^2 + \frac{n}{p}\|w_{t+1}^* - w_i^*\|^2 + \frac{mn}{p(p-\bar{n}_t-1)}\sum_{s=1}^{t}\|w_{t+1}^* - w_s^*\|^2 \\ &+ \frac{m^2}{2p}(\frac{1}{p-\bar{n}_t-1} - \frac{\alpha_t}{p-(\bar{n}_t-n)-1})\sum_{s=1}^{t}\sum_{s'=1}^{t}\|w_s^* - w_{s'}^*\|^2 \\ &+ \frac{\bar{n}_t\sigma^2}{p-\bar{n}_t-1} - \frac{\alpha_t(\bar{n}_t-n)\sigma^2}{p-(\bar{n}_t-n)-1} \\ &+ \alpha_t\mathbb{E}[\|w_{\boldsymbol{v}t} - w_i^*\|^2] \end{aligned} \tag{24}$$

Proof. For notation simplicity, denote $r := t + 1$. Using Lemma 6 and the notation introduced in Section 4.3.1, we know that:

$$w_{\boldsymbol{v}t+1} = (I - P_{1r})w_{\boldsymbol{v}t} + X_{1r}^{\dagger}y_{1r}$$

where $P_{1r} = X_{1r}(X_{1r}^{\top}X_{1r})^{-1}X_{1r}^{\top}$ and $X_{1r}^{\dagger} = X_{1r}(X_{1r}^{\top}X_{1r})^{-1}$. Notice that $X_{1r} \in \mathbb{R}^{p\times(tm+n)}$ is the result of the concatenation of all data points in the memory to the new data for task $t+1$. Importantly, X_{1r} contains exactly m columns from task t. Next we can write:

$$w_{\boldsymbol{v}t+1} = (I - P_{1r})w_{\boldsymbol{v}t} + \sum_{s=1}^{r}P_{0s}w_s^* + X_{1r}^{\dagger}z_{1r}$$

where $P_{0s} = X_{1r}(X_{1r}^{\top}X_{1r})^{-1}X_{0s}$.Therefore,

$$\begin{aligned}&\mathbb{E}[\mathcal{L}_i(w_{\boldsymbol{v}t+1})] = \mathbb{E}[\|w_{\boldsymbol{v}t+1} - w_i^*\|^2] \\ &= \mathbb{E}[\|\sum_{s=1}^{r}P_{0s}(w_s^* - w_i^*) + (I - P_{1r})(w_{\boldsymbol{v}t} - w_i^*) + X_{1r}^{\dagger}z_{1r}\|^2] && (\sum_{s=1}^{r}P_{0s} = P_{1r}) \\ &= \mathbb{E}[\|\sum_{s=1}^{r}P_{0s}(w_s^* - w_i^*) + (I - P_{1r})(w_{\boldsymbol{v}t} - w_i^*)\|^2] + \mathbb{E}[\|X_{1r}^{\dagger}z_{1r}\|^2] && (z_{1r} \perp\!\!\!\perp X_{1r}, \mathbb{E}[z_{1r}] = 0)\end{aligned}$$

$$= \mathbb{E}[\| \sum_{s=1}^{r} P_{0s}(w_s^* - w_i^*)\|^2] + \mathbb{E}[\|(I - P_{1r})(w_{\boldsymbol{vt}} - w_i^*)\|^2] + \mathbb{E}[\|X_{1r}^\dagger z_{1r}\|^2] \quad ((I - P_{1r})P_{0s} = 0)$$

The form of this expression is very similar to what we used in the proof of Theorem 1 in Section 4.3.3. The only considerable difference that causes complications is the term $\mathbb{E}[\|(I - P_{1r})(w_{\boldsymbol{vt}} - w_i^*)\|^2]$. The reason is that the term $w_{\boldsymbol{vt}}$ is no longer independent of P_{1r}, and it is not straightforward to calculate this expectation. Therefore, we avoid repeating the other steps and only focus on finding this term:

$$\mathbb{E}[\|(I - P_{1r})(w_{\boldsymbol{vt}} - w_i^*)\|^2] = \mathbb{E}[\|w_{\boldsymbol{vt}} - w_i^*\|^2] - \mathbb{E}[\|P_{1r}(w_{\boldsymbol{vt}} - w_i^*)\|^2]$$

Notice that $w_{\boldsymbol{vt}}$ is independent of X_r. Therefore, we decompose the second term using Lemma 3 part (i):

$$\begin{aligned}\mathbb{E}[\|P_{1r}(w_{\boldsymbol{vt}} - w_i^*)\|^2] &= \mathbb{E}[\|P_{1t}(w_{\boldsymbol{vt}} - w_i^*) + (I - P_{1t})X_r(X_r^\top(I - P_{1t})X_r)^{-1}X_r^\top(I - P_{1t})(w_{\boldsymbol{vt}} - w_i^*)\|^2] \\ &= \mathbb{E}[\|P_{1t}(w_{\boldsymbol{vt}} - w_i^*)\|^2] + \mathbb{E}[\|(I - P_{1t})X_r(X_r^\top(I - P_{1t})X_r)^{-1}X_r^\top(I - P_{1t})(w_{\boldsymbol{vt}} - w_i^*)\|^2]\end{aligned}$$

Recall that $X_r \in \mathbb{R}^{p\times n}$ follows the standard normal distribution and is independent of $(w_{\boldsymbol{vt}} - w_i^*)$ and P_{1t}. Due to the rotational invariance property, we can write:

$$\begin{aligned}\mathbb{E}[\|P_{1r}(w_{\boldsymbol{vt}} - w_i^*)\|^2] &= \mathbb{E}[\|P_{1t}(w_{\boldsymbol{vt}} - w_i^*)\|^2] + (\frac{n}{p - tm})\mathbb{E}[\|(I - P_{1t})(w_{\boldsymbol{vt}} - w_i^*)\|^2] \\ &= (1 - \frac{n}{p - tm})\mathbb{E}[\|P_{1t}(w_{\boldsymbol{vt}} - w_i^*)\|^2] + (\frac{n}{p - tm})\mathbb{E}[\|w_{\boldsymbol{vt}} - w_i^*\|^2]\end{aligned}$$

Therefore, it suffices to focus on $\mathbb{E}[\|P_{1t}(w_{\boldsymbol{vt}} - w_i^*)\|^2]$ as the next step. We go back to the definition of $w_{\boldsymbol{vt}}$ and use Lemma 6 again. Notice that $w_{\boldsymbol{vt}}$ was trained on the samples in the memory from tasks 1 to $t-1$ (each task with size m) in addition to the data from task t with size n. Let's build a new matrix $\widehat{X}_{1t} \in \mathbb{R}^{p\times((t-1)m+n)}$ to represent all of the training data used in that step. Similarly, build $\widehat{y}_{1t}$ and $\widehat{z}_{1t}$ by concatenating all the label and noise values. Also denote $\widehat{X}_{0s} \in \mathbb{R}^{p\times((t-1)m+n)}$ with zeros at all columns except the columns corresponding to task s. Then we can write:

$$w_{\boldsymbol{vt}} = (I - \widehat{P}_{1t})w_{\boldsymbol{vt}-1} + \sum_{s=1}^{t} \widehat{P}_{0s}w_s^* + \widehat{X}_{1t}^\dagger \widehat{z}_{1t}$$

where $\widehat{X}_{1t}^\dagger = \widehat{X}_{1t}(\widehat{X}_{1t}^\top \widehat{X}_{1t})^{-1}$, $\widehat{P}_{1t} = \widehat{X}_{1t}(\widehat{X}_{1t}^\top \widehat{X}_{1t})^{-1}\widehat{X}_{1t}^\top$ and $\widehat{P}_{0s} = \widehat{X}_{1t}(\widehat{X}_{1t}^\top \widehat{X}_{1t})^{-1}\widehat{X}_{0s}^\top$. Notice that the first tm columns of $\widehat{X}_{1t}$ is exactly X_{1t} and there are exactly $n - m$ extra columns in $\widehat{X}_{1t}$ that we threw away before proceeding to the task $t+1$. Let $\widehat{X}_t \in \mathbb{R}^{p\times(n-m)}$ represent this portion of the training set. In other words, $\widehat{X}_{1t} = \begin{bmatrix} X_{1t} & \widehat{X}_t \end{bmatrix}$. With this in mind, based on Lemma 3 part (v), it holds that

$$\begin{aligned}&\widehat{X}_{1t}^\dagger = \begin{bmatrix} X_{1t}^\dagger - (I - P_{1t})\widehat{X}_t(\widehat{X}_t^\top(I - P_{1t})\widehat{X}_t)^{-1}\widehat{X}_t^\top X_{1t}^\dagger & (I - P_{1t})\widehat{X}_t(\widehat{X}_t^\top(I - P_{1t})\widehat{X}_t)^{-1} \end{bmatrix} \\ &\Rightarrow P_{1t}\widehat{P}_{1t} = P_{1t}\widehat{X}_{1t}^\dagger \widehat{X}_{1t}^\top = \begin{bmatrix} X_{1t}^\dagger & 0 \end{bmatrix} \widehat{X}_{1t}^\top = P_{1t}\end{aligned}$$

and also

$$P_{1t}\widehat{P}_{0s} = P_{1t}\widehat{X}_{1t}^\dagger \widehat{X}_{0s}^\top = P_{0s}$$

Therefore,

$$\mathbb{E}[\|P_{1t}(w_{\boldsymbol{v}t} - w_i^*)\|^2] = \mathbb{E}[\|P_{1t}(I - \widehat{P}_{1t})(w_{\boldsymbol{v}t-1} - w_i^*) + \sum_{s=1}^{t} P_{1t}\widehat{P}_{0s}(w_s^* - w_i^*) + P_{1t}\widehat{X}_{1t}^{\dagger}\widehat{z}_{1t}\|^2]$$

$$= \mathbb{E}[\|\sum_{s=1}^{t} P_{0s}(w_s^* - w_i^*) + X_{1t}^{\dagger}z_{1t}\|^2]$$

We have previously calculated terms like this in Section 4.3.3. Therefore, we can write overall:

$$\mathbb{E}[\|(I - P_{1r})(w_{\boldsymbol{v}t} - w_i^*)\|^2] = \mathbb{E}[\|w_{\boldsymbol{v}t} - w_i^*\|^2] - (1 - \frac{n}{p - tm})\mathbb{E}[\|P_{1t}(w_{\boldsymbol{v}t} - w_i^*)\|^2] - (\frac{n}{p - tm})\mathbb{E}[\|w_{\boldsymbol{v}t} - w_i^*\|^2]$$

$$= \alpha_t\mathbb{E}[\|w_{\boldsymbol{v}t} - w_i^*\|^2] - \alpha_t\mathbb{E}[\|P_{1t}(w_{\boldsymbol{v}t} - w_i^*)\|^2]$$

$$= \alpha_t\mathbb{E}[\|w_{\boldsymbol{v}t} - w_i^*\|^2] - \alpha_t[\frac{m^2}{2p(p - tm - 1)}\sum_{s=1}^{t}\sum_{s'=1}^{t}\|w_s^* - w_{s'}^*\|^2$$

$$+ \frac{m}{p}\sum_{s=1}^{t}\|w_s^* - w_i^*\|^2 + \frac{tm\sigma^2}{p - tm - 1}]$$

Finally, we have

$$\mathbb{E}[\mathcal{L}_i(w_{\boldsymbol{v}t+1})] = \mathbb{E}[\|w_{\boldsymbol{v}t+1} - w_i^*\|^2]$$

$$= \frac{m^2}{2p(p - \bar{n}_t - 1)}\sum_{s=1}^{t}\sum_{s'=1}^{t}\|w_s^* - w_{s'}^*\|^2 + \frac{mn}{p(p - \bar{n}_t - 1)}\sum_{s=1}^{t}\|w_{t+1}^* - w_s^*\|^2$$

$$+ \frac{m}{p}\sum_{s=1}^{t}\|w_s^* - w_i^*\|^2 + \frac{n}{p}\|w_{t+1}^* - w_i^*\|^2 + \mathbb{E}[\|(I - P_{1r})(w_{\boldsymbol{v}t} - w_i^*)\|^2] + \frac{\bar{n}_t\sigma^2}{p - \bar{n}_t - 1}$$

$$= \frac{m}{p}(1 - \alpha_t)\sum_{s=1}^{t}\|w_s^* - w_i^*\|^2 + \frac{n}{p}\|w_{t+1}^* - w_i^*\|^2 + \frac{mn}{p(p - \bar{n}_t - 1)}\sum_{s=1}^{t}\|w_{t+1}^* - w_s^*\|^2$$

$$+ \frac{m^2}{2p}(\frac{1}{p - \bar{n}_t - 1} - \frac{\alpha_t}{p - (\bar{n}_t - n) - 1})\sum_{s=1}^{t}\sum_{s'=1}^{t}\|w_s^* - w_{s'}^*\|^2$$

$$+ \frac{\bar{n}_t\sigma^2}{p - \bar{n}_t - 1} - \frac{\alpha_t(\bar{n}_t - n)\sigma^2}{p - (\bar{n}_t - n) - 1}$$

$$+ \alpha_t\mathbb{E}[\|w_{\boldsymbol{v}t} - w_i^*\|^2]$$

□

Using the recursive form in Equation 24, one can exactly find the loss of a replay+regularization-based continual learner. However, the full form is not intuitive. Therefore, we focus on the two-task case instead.

Proof of Theorem 3

Proof. We use Equation 24 and substitute $\sigma = 0$ and $T = 2$. Therefore, $\alpha_0 = 1 - \frac{n}{p}$ and $\alpha_1 = 1 - \frac{n}{p-m}$. For the loss of the first task, we have

$$\mathbb{E}[\mathcal{L}_1(w_{\boldsymbol{v}1})] = \mathbb{E}[\|w_{\boldsymbol{v}1} - w_1^*\|^2] = (1 - \frac{n}{p})\|w_1^*\|^2$$

$$
\begin{aligned}
\Rightarrow \mathbb{E}[\mathcal{L}_1(w_{\boldsymbol{v}2})] &= \mathbb{E}[\|w_{\boldsymbol{v}2} - w_1^*\|^2] \\
&= \frac{n}{p}(1 + \frac{m}{p-(n+m)-1})\|w_2^* - w_1^*\|^2 + (1 - \frac{n}{p-m})\mathbb{E}[\|w_{\boldsymbol{v}1} - w_1^*\|^2] \\
&= \frac{n}{p}(1 + \frac{m}{p-(n+m)-1})\|w_2^* - w_1^*\|^2 + (1 - \frac{n}{p-m})(1 - \frac{n}{p})\|w_1^*\|^2
\end{aligned}
$$

And for the second task:

$$
\begin{aligned}
\mathbb{E}[\mathcal{L}_2(w_{\boldsymbol{v}1})] &= \mathbb{E}[\|w_{\boldsymbol{v}1} - w_2^*\|^2] \\
&= \frac{n}{p}\|w_1^* - w_2^*\|^2 + (1 - \frac{n}{p})\|w_2^*\|^2
\end{aligned}
$$

$$
\begin{aligned}
\Rightarrow \mathbb{E}[\mathcal{L}_2(w_{\boldsymbol{v}2})] &= \mathbb{E}[\|w_{\boldsymbol{v}t+1} - w_i^*\|^2] \\
&= \frac{m}{p}(\frac{n}{p-m} + \frac{n}{p-(m+n)-1})\|w_1^* - w_2^*\|^2 + (1 - \frac{n}{p-m})\mathbb{E}[\|w_{\boldsymbol{v}1} - w_2^*\|^2] \\
&= \frac{n}{p}(1 - \frac{n-m}{p-m} + \frac{m}{p-(m+n)-1})\|w_1^* - w_2^*\|^2 + (1 - \frac{n}{p-m})(1 - \frac{n}{p})\|w_2^*\|^2
\end{aligned}
$$

Finally, we can find the desired metrics:

$$
\begin{aligned}
\mathbb{E}[G(w_{\boldsymbol{v}2})] &= \frac{1}{2}(\mathbb{E}[\mathcal{L}_1(w_{\boldsymbol{v}2})] + \mathbb{E}[\mathcal{L}_2(w_{\boldsymbol{v}2})]) \\
&= \frac{n}{2p}(2 - \frac{n-m}{p-m} + \frac{2m}{(p-\bar{n}_t-1)})\|w_1^* - w_2^*\|^2 + (1 - \frac{n}{p-m})(1 - \frac{n}{p})\|w_1^*\|^2
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}[F(w_{\boldsymbol{v}2})] &= \mathbb{E}[\mathcal{L}_1(w_{\boldsymbol{v}2})] - \mathbb{E}[\mathcal{L}_1(w_{\boldsymbol{v}1})] \\
&= \frac{n}{p}(1 + \frac{m}{p-(n+m)-1})\|w_2^* - w_1^*\|^2 - \frac{n}{p-m}(1 - \frac{n}{p})\|w_1^*\|^2
\end{aligned}
$$

□

4.4 Effect of Regularization in CL Models

In this section, we study the effect of regularization on the performance of replay-based continual models by using the results of Theorems 2 and 3. To that end, let's reconsider the Equations 14 to 17 and denote the G terms in these equations as G^{rep} and G^{reg} to respectively represent the generalization error of the pure replay-based and regularization+replay CL models. We apply similar notations to the F terms to have F^{rep} and F^{reg} which lets us write

$$
G^{\text{rep}} = \underbrace{\frac{n}{2p}(1 + \frac{m}{n} + \frac{2m}{p-(n+m)-1})\|w_1^* - w_2^*\|_2^2}_{\text{term } G_1^{\text{rep}}} + \underbrace{(1 - \frac{n+m}{p})\|w_1^*\|_2^2}_{\text{term } G_2^{\text{rep}}}
$$

$$
F^{\text{rep}} = \underbrace{\frac{n}{p}(1 + \frac{m}{p-(n+m)-1})\|w_1^* - w_2^*\|_2^2}_{\text{term } F_1^{\text{rep}}} - \underbrace{\frac{m}{p}\|w_1^*\|_2^2}_{\text{term } F_2^{\text{rep}}}
$$

$$G^{\text{reg}} = \underbrace{\frac{n}{2p}(2 - \frac{n-m}{p-m} + \frac{2m}{p-(n+m)-1})\|w_1^* - w_2^*\|_2^2}_{\text{term } G_1^{\text{reg}}} + \underbrace{(1 - \frac{n}{p-m})(1 - \frac{n}{p})\|w_1^*\|_2^2}_{\text{term } G_2^{\text{reg}}}$$

$$F^{\text{reg}} = \underbrace{\frac{n}{p}(1 + \frac{m}{p-(n+m)-1})\|w_1^* - w_2^*\|_2^2}_{\text{term } F_1^{\text{reg}}} - \underbrace{\frac{n}{p-m}(1 - \frac{n}{p})\|w_1^*\|_2^2}_{\text{term } F_2^{\text{reg}}}$$

We start by comparing the forgetting terms:

$$\begin{aligned}
F^{\text{rep}} - F^{\text{reg}} &= F_1^{\text{rep}} + F_2^{\text{rep}} - F_1^{\text{reg}} - F_2^{\text{reg}} \\
&= -\frac{m}{p}\|w_1^*\|_2^2 + \frac{n}{p-m}(1 - \frac{n}{p})\|w_1^*\|_2^2 \\
&= \frac{1}{p(p-m)}(-m(p-m) + n(p-n))\|w_1^*\|_2^2 \\
&= \frac{(n-m)}{p(p-m)}(p-(n+m))\|w_1^*\|_2^2 > 0
\end{aligned}$$

Noticing that $p > n + m$ and $m \leq n$, reveals that always $F^{\text{rep}} > F^{\text{reg}}$ which means that the regularized continual learner always has lower forgetting. Similarly, we can compare the G terms:

$$\begin{aligned}
G_1^{\text{rep}} - G_1^{\text{reg}} &= \frac{n}{2p}(-1 + \frac{m}{n} + \frac{n-m}{p-m})\|w_1^* - w_2^*\|_2^2 \\
&= \frac{-(n-m)}{2p(p-m)}(p-m-n)\|w_1^* - w_2^*\|_2^2 < 0
\end{aligned}$$

This shows that the first generalization error is better for the unregularized continual learner. For the second term,

$$\begin{aligned}
G_2^{\text{rep}} - G_2^{\text{reg}} &= (1 - \frac{n+m}{p} - (1 - \frac{n}{p-m})(1 - \frac{n}{p}))\|w_1^*\|_2^2 \\
&= (-\frac{m}{p} + \frac{n}{p-m} - \frac{n}{p-m}\frac{n}{p})\|w_1^*\|_2^2 \\
&= \frac{1}{p(p-m)}(-m(p-m) + np - n^2)\|w_1^*\|_2^2 \\
&= \frac{n-m}{p(p-m)}(p-(m+n))\|w_1^*\|_2^2 > 0
\end{aligned}$$

This difference is a positive value since $p > m + n$ and $n \geq m$. Therefore, for the second term, the regularization does not benefit. With this in mind, we can not clearly express whether the regularization is helpful in the overall generalization error of the continual learner and it depends on the task similarity value $\|w_1^* - w_2^*\|_2^2$ to decide on the effectiveness of the regularization.

4.5 Bridging Linear Models and Overparameterized DNNs

Recent literature has pointed out several connections between linear models and overparameterized DNNs. In this section, we briefly review this connection to highlight the importance of studying

linear models and refer the reader to Section 1.2 of [Hastie et al.(2022)] for a broader discussion. We establish the connection using the concept of lazy training regime.

Assume an i.i.d. data as $\mathcal{D} = \{(z_i, y_i)\}_{i=1}^n$ with inputs $z_i \in \mathbb{R}^d$ and labels $y_i \in \mathbb{R}$. Consider a possibly non-linear neural network $f(\cdot;\theta) : \mathbb{R}^d \to \mathbb{R}$ parameterized by $\theta \in \mathbb{R}^p$. Under certain conditions [Jacot et al.(2018), Allen-Zhu et al.(2019), Chizat et al.(2020)], including overparameterization, the neural network $f(\cdot;\theta)$ can be approximated by its first-order Taylor expansion around the initial parameters θ_0. Furthermore, by supposing that the initialization is such that $f(z;\theta_0) \approx 0$, we can write the following linear form:

$$f(z;\theta) \approx f(z;\theta_0) + \nabla_\theta f(z;\theta_0)^\top(\theta - \theta_0) \approx \nabla_\theta f(z;\theta_0)^\top(\theta - \theta_0)$$

This approximation is still non-linear in z but linear in the parameters θ, and it implies that in the lazy training regime, the neural network behaves similarly to a linear model where the features are given by the gradients of the network with respect to its parameters. Specifically, the features are the Jacobian matrix $x_i := \nabla_\theta f(z_i;\theta_0)$.

Additionally, this approximation allows us to define a Neural Tangent Kernel (NTK) [Jacot et al.(2018)] as $\Theta(z_i, z_j) = \nabla_\theta f(z_i;\theta_0)^\top \nabla_\theta f(z_j;\theta_0)$, which captures the inner product of the gradients of the neural network with respect to the parameters at different data points. As training progresses, if the parameters remain close to their initialization (lazy training), the predictions of the neural network can be effectively described by a linear model in this high-dimensional feature space defined by the NTK. This connection elucidates how overparameterized deep neural networks can exhibit behavior akin to kernel methods, where the NTK serves as the kernel function that determines the similarity between data points.

5 Conclusion

In this paper, we investigated overparameterized models in both multi-task learning (MTL) and continual learning (CL) settings. We derived exact theoretical results to characterize the impact of key system parameters, such as model size, dataset size, and task similarity, on generalization error and knowledge transfer. Additionally, we provided insights into the performance of replay-based continual learning models, analyzing how buffer size and model capacity influence forgetting rates. Our theoretical framework uses overparameterized linear models as proxies for more complex architectures, offering precise, closed-form results that enhance our understanding of MTL and CL with overparameterized systems.

We validated our theoretical findings through empirical experiments with deep neural networks (DNNs), demonstrating that similar behaviors hold in practice, thus bridging theoretical insights with real-world applications. These results offer valuable guidance for the design and optimization of overparameterized models in both MTL and CL setups.

Our analysis, while offering novel closed-form results, is based on several simplifying assumptions, such as i.i.d. Gaussian features and linear models, which may limit the generalizability to more complex or non-Gaussian data. Moreover, while we provided connections between linear models and DNNs, these links hold under certain assumptions that may not always apply. Future work could focus on extending these theoretical tools to deep models, particularly in understanding task similarity and its role in the double descent phenomenon observed in DNNs.

References

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

[Allen-Zhu et al.(2019)] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[Baxter(2000)] Jonathan Baxter. A model of inductive bias learning. *J. Artif. Int. Res.*, 12(1): 149–198, March 2000.

[Belkin et al.(2018)] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 541–549. PMLR, 10–15 Jul 2018.

[Belkin et al.(2019)] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[Belkin et al.(2020)] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

[Ben-David & Borbely(2008)] Shai Ben-David and Reba Schuller Borbely. A notion of task relatedness yielding provable multiple-task learning guarantees. *Machine Learning*, 73(3):273–287, Dec 2008.

[Cao et al.(2022)] Yuan Cao, Zixiang Chen, Misha Belkin, and Quanquan Gu. Benign overfitting in two-layer convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 25237–25250. Curran Associates, Inc., 2022.

[Caruana(1997)] Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, Jul 1997.

[Chen et al.(2022)] Xi Chen, Christos Papadimitriou, and Binghui Peng. Memory bounds for continual learning. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 519–530, 2022.

[Chizat et al.(2020)] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming, 2020.

[Crawshaw(2020)] Michael Crawshaw. Multi-task learning with deep neural networks: A survey. *ArXiv*, abs/2009.09796, 2020.

[Deng(2012)] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[Doan et al.(2021)] Thang Doan, Mehdi Abbana Bennani, Bogdan Mazoure, Guillaume Rabusseau, and Pierre Alquier. A theoretical analysis of catastrophic forgetting through the ntk overlap matrix. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1072–1080. PMLR, 13–15 Apr 2021.

[Douillard et al.(2022)] Arthur Douillard, Alexandre Ramé, Guillaume Couairon, and Matthieu Cord. Dytox: Transformers for continual learning with dynamic token expansion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[Evron et al.(2022)] Itay Evron, Edward Moroshko, Rachel Ward, Nathan Srebro, and Daniel Soudry. How catastrophic can catastrophic forgetting be in linear regression? In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 4028–4079. PMLR, 02–05 Jul 2022.

[French(1999)] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999.

[Goldfarb & Hand(2023)] Daniel Goldfarb and Paul Hand. Analysis of catastrophic forgetting for random orthogonal transformation tasks in the overparameterized regime. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 2975–2993. PMLR, 25–27 Apr 2023.

[Goodfellow et al.(2013)] Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013.

[Gunasekar et al.(2018)] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1832–1841. PMLR, 10–15 Jul 2018.

[Hastie et al.(2022)] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2): 949, 2022.

[Hendrycks et al.(2021)] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021.

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

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

[Kolar et al.(2011)] Mladen Kolar, John Lafferty, and Larry Wasserman. Union support recovery in multi-task learning. *Journal of Machine Learning Research*, 12(72):2415–2435, 2011.

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

[Lin et al.(2023)] Sen Lin, Peizhong Ju, Yingbin Liang, and Ness Shroff. Theory on forgetting and generalization of continual learning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[Lounici et al.(2009)] Karim Lounici, Massimiliano Pontil, Alexandre B. Tsybakov, and Sara van de Geer. Taking advantage of sparsity in multi-task learning, 2009.

[Nakkiran et al.(2021)] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: where bigger models and more data hurt*. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, dec 2021.

[Obozinski et al.(2011)] Guillaume Obozinski, Martin J. Wainwright, and Michael I. Jordan. Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, 39 (1):1–47, 2011.

[Oymak & Soltanolkotabi(2019)] Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4951–4960. PMLR, 09–15 Jun 2019.

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

[Paszke et al.(2019)] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[Pentina & Ben-David(2015)] Anastasia Pentina and Shai Ben-David. Multi-task and lifelong learning of kernels. In *Algorithmic Learning Theory*, pp. 194–208, Cham, 2015. Springer International Publishing.

[Robins(1995)] Anthony Robins. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146, 1995.

[Schaul et al.(2015)] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.

[Shin et al.(2017)] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems*, pp. 2990–2999, 2017.

[van de Ven et al.(2022)] Gido M. van de Ven, Tinne Tuytelaars, and Andreas S. Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197, Dec 2022.

[Wah et al.(2011)] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset, 2011. California Institute of Technology.

[Wang et al.(2022)] Fu-Yun Wang, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Foster: Feature boosting and compression for class-incremental learning. *arXiv preprint arXiv:2204.04662*, 2022.

[Wang et al.(2019)] Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11293–11302, 2019.

[Wu et al.(2020)] Sen Wu, Hongyang R. Zhang, and Christopher Ré. Understanding and improving information transfer in multi-task learning. In *International Conference on Learning Representations*, 2020.

[Yin et al.(2021)] Dong Yin, Mehrdad Farajtabar, Ang Li, Nir Levine, and Alex Mott. Optimization and generalization of regularization-based continual learning: a loss approximation viewpoint, 2021.

[Yousefi et al.(2018)] Niloofar Yousefi, Yunwen Lei, Marius Kloft, Mansooreh Mollaghasemi, and Georgios C. Anagnostopoulos. Local rademacher complexity-based learning guarantees for multi-task learning. *Journal of Machine Learning Research*, 19(38):1–47, 2018.

[Zenke et al.(2017)] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pp. 3987–3995. PMLR, 2017.

[Zhang et al.(2017)] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

[Zhang(2015)] Yu Zhang. Multi-task learning and algorithmic stability. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), Feb. 2015.

[Zhang & Yang(2022)] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2022.

[Zhou et al.(2023)] Da-Wei Zhou, Qi-Wei Wang, Han-Jia Ye, and De-Chuan Zhan. A model or 603 exemplars: Towards memory-efficient class-incremental learning. In *International Conference on Learning Representations*, 2023.