A Statistical Theory of Regularization-Based Continual Learning

Xuyang Zhao [1] **Huiyuan Wang** [2] **Weiran Huang** [† 3 4] **Wei Lin** [† 1]

Abstract

We provide a statistical analysis of regularization-based continual learning on a sequence of linear regression tasks, with emphasis on how different regularization terms affect the model performance. We first derive the convergence rate for the oracle estimator obtained as if all data were available simultaneously. Next, we consider a family of generalized ℓ_2-regularization algorithms indexed by matrix-valued hyperparameters, which includes the minimum norm estimator and continual ridge regression as special cases. As more tasks are introduced, we derive an iterative update formula for the estimation error of generalized ℓ_2-regularized estimators, from which we determine the hyperparameters resulting in the optimal algorithm. Interestingly, the choice of hyperparameters can effectively balance the trade-off between forward and backward knowledge transfer and adjust for data heterogeneity. Moreover, the estimation error of the optimal algorithm is derived explicitly, which is of the same order as that of the oracle estimator. In contrast, our lower bounds for the minimum norm estimator and continual ridge regression show their suboptimality. A byproduct of our theoretical analysis is the equivalence between early stopping and generalized ℓ_2-regularization in continual learning, which may be of independent interest. Finally, we conduct experiments to complement our theory.

[1]School of Mathematical Sciences and Center for Statistical Science, Peking University, Beijing, China [2]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA [3]MIFA Lab, Qing Yuan Research Institute, SEIEE, Shanghai Jiao Tong University, Shanghai, China [4]Shanghai AI Laboratory, Shanghai, China. [†]Correspondence to: Weiran Huang <weiran.huang@outlook.com>, Wei Lin <weilin@math.pku.edu.cn>. This work was done while the first author was visiting MIFA Lab.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

Continual learning (CL) in machine learning involves training a model continuously across multiple tasks, constrained by limited memory. As more tasks are introduced and additional data samples are collected, it is expected that the model will exhibit enhanced performance on both old and new tasks. However, due to memory limits, not all past data can be retained; typically, only a subset of the data or summary statistics are stored. This makes continual learning more challenging than single-task learning, as it prohibits the simple pooling of all samples (Parisi et al., 2019). Alternatively, without using exceedingly large long-term memory, we can view continual learning as an online multi-task problem where a model is sequentially fitted to data provided for each task. However, such an approach may result in poor performance of the current model on previous tasks, a phenomenon known as *catastrophic forgetting* (McCloskey & Cohen, 1989; Goodfellow et al., 2014). Clearly, forgetting information from earlier tasks undermines the overall effectiveness of the model.

There are two goals of continual learning algorithms. One is the *forward knowledge transfer*, which focuses on transferring knowledge from previous tasks to make learning on new tasks simpler. The other is the *backward knowledge transfer* (Lin et al., 2023), which aims to address the issue of catastrophic forgetting when learning new tasks and keep the overall performance improving over time. From a statistical perspective, the main difficulty in these two goals is *heterogeneity* among tasks, i.e., the data distribution can vary across different tasks. In the presence of heterogeneity, the forward and backward knowledge transfer can contradict each other, between which a trade-off will arise (Lin et al., 2023; Wang et al., 2024). An ideal CL algorithm should properly balance the knowledge extracted from old tasks and the information contained in new samples to achieve both forward and backward knowledge transfer.

To resolve the conflict, many algorithms have been proposed recently. Roughly speaking, these algorithms fall into three categories: regularization-based methods (Kirkpatrick et al., 2017; Aljundi et al., 2018; Liu & Liu, 2022), replay-based methods (Chaudhry et al., 2019; Riemer et al., 2019; Jin et al., 2021), and expansion-based

methods (Serra et al., 2018; Yoon et al., 2020; Yang et al., 2021). The common intuition underlying these algorithms is applying different techniques that can use old information to constrain the model's change on new tasks, thereby achieving forward and backward knowledge transfer simultaneously. However, the theoretical understanding of CL algorithms is still underdeveloped. In particular, none of the existing work shows an explicit trade-off between forward and backward knowledge transfer, let alone offering a guidance on how to balance them properly. Also, the roles of heterogeneity and noise are not fully discussed, which are crucial aspects of practical continual learning.

In this paper, we enrich the existing literature by establishing theoretical properties of regularization-based continual learning algorithms within the linear regression framework. Our analysis includes considerations for heterogeneity, noise, and overparametrization, and offers an in-depth investigation of the trade-off between forward and backward knowledge transfer. Specifically, our contributions are summarized as follows.

- We provide lower bounds for two continual learning algorithms, i.e., the minimum norm estimator (Lin et al., 2023) and continual ridge regression (Li et al., 2023). These bounds reveal their suboptimality compared to the oracle estimator, which motivates us to study some new algorithms.
- We point out two main reasons for the failure of the above two methods: forward–backward trade-off and information heterogeneity. The former is essentially the trade-off between the information carried in old tasks and that in the new task, and the latter means that the knowledge carried in different tasks varies.
- Inspired by our findings, we propose a generalized ℓ_2-regularized estimator. By choosing its hyperparameters properly to deal with the forward–backward trade-off and information heterogeneity, we show that our estimator attains the error rate of the oracle estimator and hence avoids catastrophic forgetting.
- We establish the relationship between early stopping between ℓ_2-regularization in continual linear regression. We show that, if the learning rate of gradient descent takes a more general form as in our generalized ℓ_2-regularization, then these two methods are actually equivalent. This can be viewed as an extension of similar results shown for learning a single task.
- We conduct simulation experiments to complement our theory. We obtain a practical algorithm based on the above theoretical results, which has a close connection with elastic weighted consolidation (EWC). We illustrate its performance through simulations.

1.1. Related Work

Continual learning algorithms. Over the past several years, continual learning has attracted considerable attention, leading to the proposal of numerous empirical algorithms aimed at mitigating catastrophic forgetting. Broadly speaking, these methods can be categorized into three groups: (1) regularization-based methods (Kirkpatrick et al., 2017; Aljundi et al., 2018; Liu & Liu, 2022), which regularize modifications to the importance weights for old tasks when learning the new task; (2) expansion-based methods (Serra et al., 2018; Yoon et al., 2020; Yang et al., 2021), which learn a mask to fix the importance weights for old tasks during the new task learning and further expand the neural network when needed; (3) memory-based methods, which either store and replay the data from old tasks when learning the new task, i.e., experience-replay based methods (Chaudhry et al., 2019; Riemer et al., 2019; Jin et al., 2021), or store the gradient information from old tasks and learn the new task in the direction orthogonal to old tasks, i.e., orthogonal-projection based methods (Farajtabar et al., 2020; Saha et al., 2021; Lin et al., 2022).

Theoretical studies in CL. McCloskey & Cohen (1989) proposed a unified framework for the performance analysis of regularization-based CL methods, by formulating them as a second-order Taylor approximation of the loss function for each task. Bennani et al. (2020) and Doan et al. (2021) analyzed generalization error and forgetting for the orthogonal gradient descent (OGD) approach (Yin et al., 2020) based on NTK models, and further proposed variants of OGD to address forgetting. Lee et al. (2021) and Asanuma et al. (2021) studied CL in the teacher–student setup to characterize the impact of task similarity on forgetting performance. Cao et al. (2022) and Li et al. (2022) investigated continual representation learning with dynamically expanding feature spaces, and developed provably efficient CL methods with a characterization of the sample complexity.

Besides, there are some theoretical works on regularization-based methods. Evron et al. (2022) studied the minimum norm estimator in CL under an overparameterized and noise-free setup. Li et al. (2023) gave a fixed design analysis of continual ridge regression for two-task linear regression.

Chen et al. (2022) characterized the lower memory bound in CL using the PAC framework. Andle & Yasaei Sekeh (2022) analyzed the selection of frozen filters based on layer sensitivity to maximize the performance of CL. Wen et al. (2024) studied the contrastive CL methods and provided upper and lower performance bounds. Yang et al. (2022) presented a CL algorithm based on supervised PCA

and gave a theoretical analysis. Denevi et al. (2019) proposed to add a bias term to SGD and showed improved performance theoretically.

2. Continual Linear Regression

Data. We consider a standard continual learning problem where a sequence of tasks indexed by $t = 1, \ldots, T$ arrives sequentially. Suppose that each task t holds a dataset $\mathcal{D}_t = \{(\boldsymbol{x}_i^{(t)}, y_i^{(t)}) \in \mathbb{R}^p \times \mathbb{R}\}_{i=1}^{n_t}$, where n_t denotes its sample size. We assume that all of the T tasks are generated by a linear model with the same regression coefficient, i.e., for all $t \in [T]$ and $i \in [n_t]$,

$$y_i^{(t)} = (\boldsymbol{x}_i^{(t)})^\top \boldsymbol{w}_* + \varepsilon_i^{(t)}, \tag{1}$$

where $\boldsymbol{w}_* \in \mathbb{R}^p$ is the true parameter and $\varepsilon_i^{(t)}$ are independent random noises with variance σ^2. By stacking the data as $\boldsymbol{X}_t := (\boldsymbol{x}_1^{(t)}, \ldots, \boldsymbol{x}_{n_t}^{(t)})^\top \in \mathbb{R}^{n_t \times p}$, $\boldsymbol{y}_t := (y_1^{(t)}, \ldots, y_{n_t}^{(t)}) \in \mathbb{R}^{n_t}$, and $\boldsymbol{\varepsilon}_t := (\varepsilon_1^{(t)}, \ldots, \varepsilon_{n_t}^{(t)}) \in \mathbb{R}^{n_t}$, we can rewrite (1) as

$$\boldsymbol{y}_t = \boldsymbol{X}_t \boldsymbol{w}_* + \boldsymbol{\varepsilon}_t.$$

We define $\boldsymbol{\Sigma}_t := \boldsymbol{X}_t^\top \boldsymbol{X}/n_t \in \mathbb{R}^{p \times p}$ as the covariance matrix for task t. Note that we do not require $n_t > p$, i.e., we allow for overparametrization in any single task.

Evaluation metric. Our goal is to estimate $\boldsymbol{w}_*$ in the continual learning setting. For any estimator $\widehat{\boldsymbol{w}}$, we use $\mathcal{L}(\widehat{\boldsymbol{w}}) := E\|\widehat{\boldsymbol{w}} - \boldsymbol{w}_*\|^2$ to denote its estimation error. Note that the definition of $\mathcal{L}$ applies to each task, since they share a common true parameter $\boldsymbol{w}_*$. Based on $\mathcal{L}$, two key metrics, forgetting and generalization error, can be defined respectively as

$$\mathcal{F}_t := \frac{1}{t-1} \sum_{\tau=1}^{t-1} (\mathcal{L}(\widehat{\boldsymbol{w}}_t) - \mathcal{L}(\widehat{\boldsymbol{w}}_\tau)),$$

$$\mathcal{G}_t := \frac{1}{t} \sum_{\tau=1}^{t} \mathcal{L}(\widehat{\boldsymbol{w}}_t) = \mathcal{L}(\widehat{\boldsymbol{w}}_t),$$

for each $t \in [T]$, where $\widehat{\boldsymbol{w}}_\tau$ denotes the output of a continual learning algorithm after the arrival of task τ. Small $\mathcal{F}_t$ means that the estimator learned after task t still has good performance on previous tasks. If $\mathcal{F}_t < 0$ for every $t \in [T]$, the continual learning algorithm achieves consistently increasing performance and avoids catastrophic forgetting.

Oracle estimator. Without the constraint of continual learning, i.e., data of all tasks are available simultaneously, we can estimate $\boldsymbol{w}_*$ by simply pooling all samples together and solving the offline optimization problem

$$\min_{\boldsymbol{w}} \left\{ \sum_{t=1}^{T} \|\boldsymbol{X}_t \boldsymbol{w} - \boldsymbol{y}_t\|^2 \right\}.$$

We call its solution the oracle estimator (ORA) and denote it by

$$\widehat{\boldsymbol{w}}_T^{(\mathrm{ORA})} := \arg\min_{\boldsymbol{w}} \left\{ \sum_{t=1}^{T} \|\boldsymbol{X}_t \boldsymbol{w} - \boldsymbol{y}_t\|^2 \right\}. \tag{2}$$

The oracle estimator cannot be used in continual learning practice since it requires simultaneous availability of all data. Nevertheless, it serves as an ideal baseline to gauge the accuracy of estimating $\boldsymbol{w}_*$ without continual learning constraint. If a continual learning algorithm exhibits comparable performance to the oracle estimator, then we can assert the superiority of that algorithm.

3. Learning Algorithms

In this paper, our primary objective is to investigate the *generalized ℓ_2-regularization* algorithm (GR), which is a family of regularization-based continual learning algorithms. Specifically, it sequentially produces an estimate of w_* as depicted in Algorithm 1, where $\{\boldsymbol{H}_t\}_{t=1}^T$ are user-specified regularization weight matrices and $\|\boldsymbol{w} - \widehat{\boldsymbol{w}}_{t-1}^{(\mathrm{GR})}\|_{\boldsymbol{H}_t}^2 := (\boldsymbol{w} - \widehat{\boldsymbol{w}}_{t-1}^{(\mathrm{GR})})^\top \boldsymbol{H}_t (\boldsymbol{w} - \widehat{\boldsymbol{w}}_{t-1}^{(\mathrm{GR})})$.

Algorithm 1 Generalized ℓ_2-regularization method

Initialization: $\widehat{\boldsymbol{w}}_0^{(\mathrm{GR})} = 0$

Iterative update for each task $t \in [T]$**:**

$$\widehat{\boldsymbol{w}}_t^{(\mathrm{GR})} := \arg\min_{\boldsymbol{w}} \left\{ \frac{1}{n} \|\boldsymbol{X}_t \boldsymbol{w} - \boldsymbol{y}_t\|^2 + \|\boldsymbol{w} - \widehat{\boldsymbol{w}}_{t-1}^{(\mathrm{GR})}\|_{\boldsymbol{H}_t}^2 \right\} \tag{3}$$

The choice of $\{\boldsymbol{H}_t\}_{t=1}^T$ determines how we navigate the balance between forward and backward knowledge transfer. With different choices of $\{\boldsymbol{H}_t\}_{t=1}^T$, the GR algorithm encompasses several commonly studied algorithms as special cases. For example, when $\boldsymbol{H}_t = \lambda_t \boldsymbol{I}_p$ for some $\lambda_t > 0$, GR becomes the conventional continual ridge regression algorithm (Li et al., 2023). In the overparameterized scenario where $p > n_t$, if $\boldsymbol{H}_t \to 0$, then GR is equivalent to the minimum norm estimator (Evron et al., 2022; Lin et al., 2023).

In the rest of this section, we give an in-depth discussion of these two algorithms.

3.1. Minimum Norm Estimator

Let $\gamma_j^{(t)}$ be the jth eigenvalue of $\boldsymbol{\Sigma}_t$. If $|\{j : \gamma_j^{(t)} > 0\}| = n_t < p$, then there always exists some $\boldsymbol{w}$ that interpolates the training data of task t, i.e., $\boldsymbol{X}_t \boldsymbol{w} = \boldsymbol{y}_t$. In this overparameterized regime, some recent works (Evron et al., 2022; Lin et al., 2023) studied the the minimum norm estimator (MN), which is defined in Algorithm 2.

Algorithm 2 Minimum norm estimator

Initialization: $\widehat{\boldsymbol{w}}_0^{(\mathrm{MN})} = 0$
Iterative update for each task $t \in [T]$**:**

$$\widehat{\boldsymbol{w}}_t^{(\mathrm{MN})} = \arg\min_{\boldsymbol{w}} \Big\{ \|\boldsymbol{w} - \widehat{\boldsymbol{w}}_{t-1}^{(\mathrm{MN})}\|^2 \quad \text{s.t. } \boldsymbol{X}_t \boldsymbol{w} = \boldsymbol{y}_t, \Big\}$$

Algorithm 3 Continual ridge regression

Initialization: $\widehat{\boldsymbol{w}}_0^{(\mathrm{CRR})} = 0$
Iterative update for each task $t \in [T]$**:**

$$\widehat{\boldsymbol{w}}_t^{(\mathrm{CRR})} = \arg\min_{\boldsymbol{w}} \Big\{ \frac{1}{n}\|\boldsymbol{X}_t \boldsymbol{w} - \boldsymbol{y}_t\|^2 + \lambda_t \|\boldsymbol{w} - \widehat{\boldsymbol{w}}_{t-1}\|^2 \Big\}$$

Compared to ℓ_2-regularization methods, MN can be regarded as the limit of the ℓ_2-regularized estimator when the penalty strength tends to 0. From this perspective, it might overly prioritize the data from the new task and underestimate the knowledge embedded in old tasks. Given that $\boldsymbol{y}_t = \boldsymbol{X}_t \boldsymbol{w}_* + \boldsymbol{\varepsilon}_t \neq \boldsymbol{X}_t \boldsymbol{w}_*$, imposing the condition $\boldsymbol{X}_t \boldsymbol{w} = \boldsymbol{y}_t$ on the estimators inevitably introduces the noise term, which in reality dominates the information when $p > n$.

Specifically, the following theorem provides a lower bound showing that the estimation error of the MN estimator cannot converge to 0.

Theorem 3.1 (Lower bound for the minimum norm estimator)**.** *Suppose that* $\boldsymbol{\Sigma}_t$ *satisfies* $|\{j : \gamma_j^{(t)} > 0\}| = n_t < p$*. Then we have*

$$\mathcal{L}(\widehat{\boldsymbol{w}}_t^{(\mathrm{MN})}) \geq \frac{\sigma^2}{\max_{j \in [p]} \gamma_j^{(t)}}.$$

From Theorem 3.1 we see that the estimation error of the minimum norm estimator is lower bounded by a term independent of the old tasks.

Consequently, even if the old tasks provide sufficient samples for an accurate estimate of $\boldsymbol{w}_*$ or the number of tasks increases infinitely, the estimation error of the MN estimator is always lower bounded by a constant that is not approaching 0. Indeed, irrespective of the accuracy of $\widehat{\boldsymbol{w}}_{t-1}$, even if it precisely matches $\boldsymbol{w}_*$, the MN estimator cannot leverage it to obtain a better estimate. This is because the estimator attempts to interpolate the newly encountered deficient data and hence does not balance the trade-off between old and new tasks, which we refer to as the *forward–backward trade-off*. As a result, the MN estimator is highly susceptible to catastrophic forgetting.

3.2. Continual Ridge Regression

Continual ridge regression (CRR) (Li et al., 2023) uses ridge regularization to constrain the parameter's change when fitting new tasks. Specifically, it updates the estimate using the iterations defined in Algorithm 3.

Clearly, CRR is a special case of our generalized ℓ_2-regularized estimator, which uses the conventional ridge penalty by setting $\lambda_t^{(1)} = \cdots = \lambda_t^{(p)} = \lambda_t$. CRR treats each coordinate of $\boldsymbol{w}_*$ equally, i.e., it potentially assumes that $|(\widehat{\boldsymbol{w}}_t^{(\mathrm{CRR})})_j - (\boldsymbol{w}_*)_j|^2$ are the same for different j.

However, such homogeneity does not always exist in continual learning setting since the information introduced by different tasks can vary across various directions of $\boldsymbol{w}_*$, especially in the scenario where the data distributions differ across tasks. For example, there may exist some $i \neq j$ such that $|(\widehat{\boldsymbol{w}}_t^{(\mathrm{CRR})})_j - (\boldsymbol{w}_*)_j|^2 = o(1)$ while $|(\widehat{\boldsymbol{w}}_t^{(\mathrm{CRR})})_i - (\boldsymbol{w}_*)_i|^2 = O(1)$ if previous tasks contain very little information about $(\boldsymbol{w}_*)_i$. In this case, the suitable values for λ_i and λ_j might differ. Consequently, the CRR estimator, which cannot address such *information heterogeneity*, may be suboptimal.

More specifically, we have the following lower bound for the CRR estimator, which shows that its worst-case performance is much worse than that of GR.

Theorem 3.2 (Lower bound for continual ridge regression)**.** *Consider a two-task and two-dimensional continual learning problem with covariance matrices* $\boldsymbol{\Sigma}_1 = \mathrm{diag}(1, \epsilon)$ *and* $\boldsymbol{\Sigma}_2 = \mathrm{diag}(\epsilon, 1)$ *and sample sizes* n_1 *and* n_2*. Then we have*

$$\sup_{n_1, n_2, \epsilon} \inf_{\lambda} \frac{\mathcal{L}\left(\widehat{\boldsymbol{w}}_2^{(\mathrm{CRR})}\right)}{\mathcal{L}\left(\widehat{\boldsymbol{w}}_2^{(\mathrm{GR})}\right)} = +\infty,$$

where λ *is the regularization hyperparameter of CRR.*

4. Generalized ℓ_2-Regularization Attains Oracle Rate

In this section, we provide a theoretical analysis of the generalized ℓ_2-regularized estimator (GR) defined in (3). Our theory shows that, through the proper selection of the regularization weight matrix $\boldsymbol{H}_t$, it is possible to avoid catastrophic forgetting, and the resulting estimation error can even be comparable with that of the oracle estimator.

Before establishing our main results, we first present some assumptions for our analysis.

4.1. Assumptions

Assumption 4.1 (Fixed design)**.** *The features $\{\boldsymbol{X}_t\}_{t=1}^T$ are fixed while the noises $\boldsymbol{\varepsilon}_t$ are random with mean 0 and variance $\sigma^2 > 0$.*

Assumption 4.2 (Commutable covariance matrices)**.** *The set of covariance matrices $\{\boldsymbol{\Sigma}_t\}_{t=1}^T$ are commutable.*

These two assumptions ensure that the GR estimator have explicit solutions, which helps to deliver our messages concisely. Similar assumptions are commonly made in related literature (Lei et al., 2021; Wu et al., 2022; Li et al., 2023). In Section 6, we will show that without these assumptions, similar results still hold.

By simple linear algebra, Assumption 4.2 is equivalent to the fact that $\{\boldsymbol{\Sigma}_t\}_{t=1}^T$ are simultaneously diagonalizable. Therefore, there exists a single orthogonal matrix $\boldsymbol{U} \in \mathbb{R}^{p\times p}$ such that $\boldsymbol{\Sigma}_t = \boldsymbol{U}\boldsymbol{\Gamma}_t\boldsymbol{U}^\top$, where $\boldsymbol{\Gamma}_t = \mathrm{diag}\{\gamma_1^{(t)}, \ldots, \gamma_p^{(t)}\}$ denotes the diagonal matrix consisting of the eigenvalues of $\boldsymbol{\Sigma}_t$. In this case, the heterogeneity among different tasks is solely encoded by the different eigenvalues in $\boldsymbol{\Gamma}_t$.

Assumption 4.3 (Sufficient sample size)**.** *For each $j \in [p]$, $\sum_{t=1}^T \gamma_j^{(t)} > 0$.*

This assumption is imposed to simplify the analysis of $\widehat{\boldsymbol{w}}^{(\mathrm{ORA})}$. Under this assumption, when the data of all T tasks are pooled together, there is no overparameterization, i.e., $\sum_{t=1}^T \boldsymbol{\Sigma}_t$ has full rank. Therefore, the oracle estimator $\widehat{\boldsymbol{w}}^{(\mathrm{ORA})}$ defined by (2) has a unique solution, whose estimation error can be calculated directly.

Indeed, the following lemma gives an explicit expression for the estimation error of the oracle estimator.

Lemma 4.1 (Estimation error of the oracle estimator)**.** *Suppose that Assumptions 4.1–4.3 hold. Then the estimator error of the oracle estimator is*

$$\mathcal{L}(\widehat{\boldsymbol{w}}_T^{(\mathrm{ORA})}) = \sum_{j=1}^p \frac{\sigma^2}{\gamma_j^{(1)} n_1 + \cdots + \gamma_j^{(t)} n_T}.$$

As the task number T increases, the estimation error of ORA is monotonically decreasing. Therefore, it does not suffer from the issue of catastrophic forgetting.

We remark that Assumption 4.3 still allows a single task to be overparameterized.

4.2. Main Results

In this section, we consider a set of specific choices of $\boldsymbol{H}_t = \boldsymbol{U}\boldsymbol{\Lambda}_t\boldsymbol{U}^\top$, where $\boldsymbol{\Lambda}_t = \mathrm{diag}\{\lambda_1^{(t)}, \ldots, \lambda_p^{(t)}\}$ is some diagonal matrix. We show that if $\boldsymbol{\Lambda}_t$ is selected properly, the estimation error of GR is compatible with that of the oracle estimator; as a result, catastrophic forgetting is avoided.

We first decompose the estimation error into components along different directions. Let $\boldsymbol{u}_j \in \mathbb{R}^p$ be the jth column of $\boldsymbol{U}$. Define $e_j^{(t)} := (\boldsymbol{u}_j^\top(\widehat{\boldsymbol{w}}_t^{(\mathrm{GR})} - \boldsymbol{w}_*))^2$ to be the projected estimation error of $\widehat{\boldsymbol{w}}_t^{(\mathrm{GR})}$ onto $\boldsymbol{u}_j$ for $j \geq 1$ and $e_j^{(0)} := \left(\boldsymbol{u}_j^\top \boldsymbol{w}_*\right)^2$. Since $\boldsymbol{U}$ is orthogonal, we have $\mathcal{L}(\widehat{\boldsymbol{w}}_t^{(\mathrm{GR})}) = \sum_{j=1}^p e_j^{(t)}$.

We are ready to present our main result regarding the estimation error of the GR estimator.

Theorem 4.2. *Suppose that Assumptions 4.1–4.3 hold. Consider $\boldsymbol{H}_t = \boldsymbol{U}\boldsymbol{\Lambda}_t\boldsymbol{U}^\top$, where $\boldsymbol{\Lambda}_t = \mathrm{diag}\{\lambda_1^{(t)}, \ldots, \lambda_p^{(t)}\}$ is some diagonal matrix. Then the projected estimation error satisfies*

$$\begin{aligned}\mathbb{E}\left[e_j^{(t)}\right] = \mathbb{E}\left[e_j^{(t-1)}\right] - 2\frac{\gamma_j^{(t)}\mathbb{E}\left[e_j^{(t-1)}\right]}{\lambda_j^{(t)} + \gamma_j^{(t)}} \\ + \frac{(\gamma_j^{(t)})^2\mathbb{E}\left[e_j^{(t-1)}\right] + \gamma_j^{(t)}\sigma^2/n}{(\lambda_j^{(t)} + \gamma_j^{(t)})^2}.\end{aligned} \tag{4}$$

If we set $\boldsymbol{\Lambda}_t$ by

$$\lambda_j^{(t)} = \frac{\sigma^2/e_j^{(0)} + \gamma_j^{(1)} n_1 + \cdots + \gamma_j^{(t-1)} n_{t-1}}{n_t} \tag{5}$$

for each $j \in [p]$ and $t \in [T]$, then (4) is minimized and we have

$$\mathbb{E}\left[e_j^{(t)}\right] = \frac{\sigma^2}{\sigma^2/e_j^{(0)} + \gamma_j^{(1)} n_1 + \cdots + \gamma_j^{(t)} n_t},$$

which further implies

$$\mathcal{L}(\widehat{\boldsymbol{w}}_t^{(\mathrm{GR})}) = \sum_{j=1}^p \frac{\sigma^2}{\sigma^2/e_j^{(0)} + \gamma_j^{(1)} n_1 + \cdots + \gamma_j^{(t)} n_t}. \tag{6}$$

Under the choices of regularization weight matrices given in Theorem 4.2, we see that the estimation error of the GR estimator is monotonically nonincreasing with task index t. Indeed, as long as the covariance matrices $\boldsymbol{\Sigma}_t$ are positive definite, the estimation error is strictly decreasing. Therefore, the forgetting error $\mathcal{F}_t \leq 0$ for every $t \in [T]$ and hence catastrophic forgetting is eliminated, even though we allow a single task to be overparameterized and the covariance matrices to be different across tasks.

Compared with Lemma 4.1, the estimation errors of the GR and oracle estimators are asymptotically equivalent as T increases, even though the latter can only be calculated when pooling data of all tasks together. Indeed, the only difference between them is the additional term $\sigma^2/e_j^{(0)}$ in the

denominator of the estimation error of GR. Therefore, the estimation error of GR is even slightly smaller than that of the oracle estimator. This is because GR has an extra ridge term when learning the first task, whereas the oracle estimator has no regularization term. We also remark that given a fixed set of tasks, the final estimation error $\mathcal{L}(\widehat{\boldsymbol{w}}_T^{(\mathrm{GR})})$ is independent of the task ordering, although the choice of $\boldsymbol{H}_t$ is dependent on it.

The key to achieving these desirable properties lies in the specific form of $\{\boldsymbol{H}_t\}_{t=1}^T$. From the proof of Theorem 4.2, we identify two crucial considerations in choosing $\{\boldsymbol{H}_t\}_{t=1}^T$.

(1) The first consideration concerns balancing the trade-off between the information carried in $\widehat{\boldsymbol{w}}_{t-1}$ and that in $\mathcal{D}_t$, i.e., the forward–backward trade-off. For example, if the estimation error of $\widehat{\boldsymbol{w}}_{t-1}$ is relatively small (larger n_τ for $\tau \le t-1$) compared with the error of the new task, σ^2/n_t, we should increase the regularization strength $\lambda_j^{(t)}$.

(2) The second one involves addressing the information heterogeneity among different tasks. As the covariance matrices vary, the amount of information pertaining to different directions of $\boldsymbol{w}_*$ within different tasks may differ. Therefore, $\boldsymbol{\Lambda}_t$ should adapt to this information heterogeneity, allowing $\lambda_i^{(t)}$ and $\lambda_j^{(t)}$ to be different for $i \neq j$.

The choice of hyperparameters specified in Theorem 4.2 effectively addresses the forward–backward trade-off and information heterogeneity, thereby avoiding catastrophic forgetting and achieving an estimation error comparable with that of the oracle estimator.

We remark that Theorem 4.2 does not necessitate $p < n_t$; it allows any individual task to be overparameterized. As long as aggregating all the data leads to an underparameterized linear regression problem, we can progressively improve the estimation of $\boldsymbol{w}_*$ as new tasks are continuously introduced using generalized ℓ_2-regularization. Ultimately, we achieve the error rate of the oracle estimator after completing the final task.

4.3. A Practical Algorithm

Now we take a closer look at the optimal choice of $\{\boldsymbol{H}_t\}_{t=1}^T$ developed in Theorem 4.2. Substituting (5) into the definition of $\boldsymbol{H}_t$ gives

$$\boldsymbol{H}_t = \frac{1}{n_t}(n_1\boldsymbol{\Sigma}_1 + \cdots + n_{t-1}\boldsymbol{\Sigma}_{t-1} + \sigma^2\boldsymbol{U}\boldsymbol{E_0}\boldsymbol{U}^\top),$$

which is the summation of the covariance matrices of old tasks weighted by sample sizes plus an additional error term. Tasks with larger sample size will be allocated with larger weights in the optimal regularization matrix, which is reasonable since they contain more information about $\boldsymbol{w}_*$.

If n_t is sufficiently large, the term $\sigma^2\boldsymbol{U}\boldsymbol{E_0}\boldsymbol{U}^\top/n_t$ in $\boldsymbol{H}_t$ is negligible and we can approximate $\boldsymbol{H}_t$ by

$$\tilde{\boldsymbol{H}}_t := \frac{1}{n_t}(n_1\boldsymbol{\Sigma}_1 + \cdots + n_{t-1}\boldsymbol{\Sigma}_{t-1}) \approx \boldsymbol{H}_t, \qquad (7)$$

which can be easily computed in practice. This approximation makes the generalized ℓ_2-regularized estimator a practical algorithm, which can be implemented without any underlying knowledge about the true parameter.

Connection with other regularization methods. Note that in linear regression, the covariance matrix is just the Hessian matrix (or Fisher information matrix) of the loss function. This links our GR estimator to some other popular regularization-based algorithms such as EWC and its variants (Kirkpatrick et al., 2017; Huszár, 2018; Schwarz et al., 2018). Specifically, if all tasks have the same sample size, our method recovers the online EWC proposed by Schwarz et al. (2018) with the hyperparameter $\gamma = 1$. Our theory gives a precise characterization of how to combine the Fisher information of old tasks properly in continual linear regression.

Approximate weight matrices. We now present a result demonstrating that using the approximate optimal weight matrices has minimal impact on the estimation error when certain conditions are met. To this end, we define $\rho_j^{(t)} := \gamma_j^{(t)}/(e_0^{(j)} + \gamma_j^{(1)}n_1 + \ldots \gamma_j^{(t-1)}n_{t-1})$, which can be viewed as the information ratio between the new task t and the old tasks. A larger ρ indicates that the new task contains more information.

Theorem 4.3. *Suppose that Assumptions 4.1–4.3 hold. Assume that we use $\tilde{\boldsymbol{H}}_t := \boldsymbol{U}\tilde{\boldsymbol{\Lambda}}_t\boldsymbol{U}^\top$ instead of $\boldsymbol{H}_t$ defined in Theorem 4.2 as the regularization weight matrices. Let $\Delta_j^{(t)} = 1/(\widetilde{\lambda}_j + \gamma_j) - 1/(\lambda_j + \gamma_j)$. Suppose that there exists some constant $C > 0$ such that*

$$(\gamma_j^{(t)}\Delta_j^{(t)})^2 \le \frac{C(C-1)(\rho_j^{(t)})^2}{(1+\rho_j^{(t)})(1+C\rho_j^{(t)})^2} \qquad (8)$$

for each t. Then for every $t \in [T]$, we have

$$\mathcal{L}(\widehat{\boldsymbol{w}}_t^{(\mathrm{GR})}) \le \frac{C}{e_j^{(0)} + \gamma_j^{(1)}n_1 + \cdots + \gamma_j^{(t)}n_t}.$$

Since $\rho_j^{(t)}$ is of order $o(1)$, the right-hand side of (8) is roughly $O((\rho_j^{(t)})^2)$. Therefore, as $\rho_j^{(t)}$ becomes larger, which means that there is relatively more information of

$\boldsymbol{u}_j^\top \boldsymbol{w}_*$ contained in the new task t, the requirement on the approximation accuracy of $\widetilde{\lambda}_j^{(t)}$ becomes looser. In this case, we can still attain the oracle rate without calculating the optimal regularization matrix very accurately. In Section 7, we will conduct experiments to illustrate the performance of the generalized ℓ_2-regularized estimator using $\tilde{\boldsymbol{H}}_t$ defined in (7) instead of $\boldsymbol{H}_t$.

5. Connection Between Early Stopping and ℓ_2-Regularization

Besides adding a penalty term to the loss function, another commonly used regularization method is early stopping. When training a single task, several works (Raskutti et al., 2014; Ali et al., 2019) have shown that applying gradient descent with early stopping is equivalent to ridge regression, in both classification and regression tasks. However, in continual learning where there is a sequence of tasks to be learned, similar results are still limited. In this section, we show that such equivalence also exists in continual linear regression.

Specifically, we formulate the early stopping estimator (ES) for continual linear regression in the following algorithm. Specifically, let $\widehat{\boldsymbol{w}}_0^{(\mathrm{ES})} = 0$ be the initial value. At each task t, we set $\widehat{\boldsymbol{w}}_{t-1}^{(\mathrm{ES})}$ as the initial point and apply m_t-step gradient descent to the loss function of this new task, where $\boldsymbol{A}_t$ is a positive definite matrix used to control the learning rate and m_t is the number of gradient descent iterations.

Algorithm 4 Early stopping estimator

Initialization: $\widehat{\boldsymbol{w}}_0^{(\mathrm{ES})} = 0$
for each task $t = 1$ **to** T **do**
  $\boldsymbol{w}_t^{(0)} = \widehat{\boldsymbol{w}}_{t-1}^{(\mathrm{ES})}$;
  for $\tau = 1$ **to** m_t **do**
    $\boldsymbol{w}_t^{(\tau)} = \boldsymbol{w}_t^{(\tau-1)} - (\boldsymbol{A}_t/n)\boldsymbol{X}_t^\top(\boldsymbol{X}_t\boldsymbol{w}_t^{(\tau-1)} - \boldsymbol{y}_t)$;
  end for
  $\widehat{\boldsymbol{w}}_t^{(\mathrm{ES})} = \boldsymbol{w}_t^{(m_t)}$;
end for

Note that in ordinary gradient descent, $\boldsymbol{A}_t$ is simply $s_t\boldsymbol{I}_p$ for some $s_t > 0$, which we refer to as *vanilla early stopping* (vanilla ES). In contrast, here we take a more general form of the learning rate matrix in order to capture the information heterogeneity and align with the generalized ℓ_2-regularization studied above.

The following theorem establishes the equivalence between the ES and GR estimators.

Theorem 5.1. *Assume that* $\boldsymbol{\Sigma}_t = \boldsymbol{U}_t\boldsymbol{\Gamma}_t\boldsymbol{U}_t^\top$, $\boldsymbol{A}_t = \boldsymbol{U}_t\boldsymbol{S}_t\boldsymbol{U}_t^\top$, *and* $\boldsymbol{H}_t = \boldsymbol{U}_t\boldsymbol{\Lambda}_t\boldsymbol{U}_t^\top$ *for some positive definite diagonal matrices* $\boldsymbol{\Gamma}_t = \mathrm{diag}\{\gamma_1^{(t)}, \ldots, \gamma_p^{(t)}\}$, $\boldsymbol{S}_t = \mathrm{diag}\{s_1^{(t)}, \ldots, s_p^{(t)}\}$ *and* $\boldsymbol{\Lambda}_t = \mathrm{diag}\{\lambda_1^{(t)}, \ldots, \lambda_p^{(t)}\}$ *satisfying*

$$\lambda_j^{(t)} = \frac{\gamma_j^{(t)}(1 - s_j^{(t)}\gamma_j^{(t)})^{m_t}}{1 - (1 - s_j^{(t)}\gamma_j^{(t)})^{m_t}} \tag{9}$$

for each $j \in [p]$ *and* $t \in [T]$. *Then for each* $t \in [T]$, *we have*

$$\widehat{\boldsymbol{w}}_t^{(\mathrm{ES})} = \widehat{\boldsymbol{w}}_t^{(\mathrm{GR})},$$

where $\widehat{\boldsymbol{w}}_t^{(\mathrm{ES})}$ *is the ES estimator using the learning rate matrix* $\boldsymbol{A}_t$, *and* $\widehat{\boldsymbol{w}}_t^{(\mathrm{GR})}$ *is the GR estimator using the regularization weight matrix* $\boldsymbol{H}_t$.

Note that this result does not require commutable covaraince matrices in Assumption 4.2. From Theorem 5.1 we conclude that with some proper choices of the learning rate matrix $\boldsymbol{A}_t$ and regularization weight matrix $\boldsymbol{H}_t$, the ES estimator $\widehat{\boldsymbol{w}}_t^{(\mathrm{ES})}$ and the GR estimator $\widehat{\boldsymbol{w}}_t^{(\mathrm{GR})}$ output exactly the same estimates for each t. Indeed, the errors $\widehat{\boldsymbol{w}}_t - \boldsymbol{w}_*$ of these two estimators are both the weighted average of the error of the $(t-1)$th task $\widehat{\boldsymbol{w}}_{t-1} - \boldsymbol{w}_*$ and the variance term for the new task, $\boldsymbol{X}_t^\top\boldsymbol{\varepsilon}_t/n$, where the weights are determined by the learning rate matrix $\boldsymbol{A}_t$, iteration number m_t, and regularization weight matrix $\boldsymbol{H}_t$.

We remark that (9) is required to hold for each $j \in [p]$. Therefore, vanilla ES with $\boldsymbol{A}_t = s_t\boldsymbol{I}_p$ and vanilla ℓ_2-regularization with $\boldsymbol{H}_t = \lambda_t\boldsymbol{I}_p$ may not be equivalent since $\gamma_j^{(t)}$ could be different for different j and a single λ_t and s_t could not make (9) hold for every $j \in [p]$. It could happen that vanilla ES is equivalent to some generalized ℓ_2-regularized estimator or vice verse.

Similar to ℓ_2-regularization, early stopping with proper learning rate matrix $\boldsymbol{A}_t$ can also avoid catastrophic forgetting and attain the oracle rate.

Corollary 5.2. *Suppose that Assumption 4.1–4.3 hold. Assume that* $\boldsymbol{A}_t = \boldsymbol{U}\boldsymbol{S}_t\boldsymbol{U}^\top$ *for some diagonal matrix* $\boldsymbol{S}_t = \mathrm{diag}\{s_1^{(t)}, \ldots, s_p^{(t)}\}$ *satisfying*

$$\left(1 - s_j^{(t)}\gamma_j^{(t)}\right)^{m_t} = 1 - \frac{\gamma_j^{(t)} n_t}{\sigma^2/e_j^{(0)} + \gamma_j^{(1)} n_1 + \cdots + \gamma_j^{(t)} n_t} \tag{10}$$

for each $j \in [p]$. *Then the estimation error of* $\widehat{\boldsymbol{w}}_t^{(\mathrm{ES})}$ *is*

$$\mathcal{L}(\widehat{\boldsymbol{w}}_t^{(\mathrm{ES})}) = \sum_{j=1}^{p} \frac{\sigma^2}{\sigma^2/e_j^{(0)} + \gamma_j^{(1)} n_1 + \cdots + \gamma_j^{(t)} n_t}.$$

If the new task t has a larger sample size n_t, the term $(1 - s_j\gamma_j)^{m_t}$ should decrease by (10), implying that both the learning rate $s_j^{(t)}$ and the iteration number m_t should be increased. This means that when task t provides more information, we should traverse a more extensive path in the gradient descent process, allowing for a deeper utilization of the new data.

6. Extensions

In this section, we discuss some possible extensions to relax our model assumptions.

Commutable covariance matrices. The main purpose of Assumption 4.2 is to obtain explicit forms for some crucial quantities of ℓ_2-regularized estimators, such as the optimal regularization matrix $\boldsymbol{H}_t$ and the corresponding optimal estimation error.

Without this assumption, even though the optimal estimation error does not have an explicit form, we can still show that there exist some regularization weight matrices such that the estimation error is monotonically nonincreasing with t. Therefore, catastrophic forgetting can still be avoided.

Specifically, we have the following result without Assumption 4.2.

Theorem 6.1. *There exist* $\{\lambda_j^{(t)}, j = 1, \ldots, p, t = 1 \ldots, T\}$ *such that for each* $t \in [T]$,

$$\mathcal{L}(\widehat{\boldsymbol{w}}_t^{(\mathrm{GR})}) \leq \mathcal{L}(\widehat{\boldsymbol{w}}_{t-1}^{(\mathrm{GR})}),$$

where the strict inequality holds for t *satisfying* $\sum_{j=1}^p \gamma_j^{(t)} > 0$.

Intuitively, the condition $\sum_{j=1}^p \gamma_j^{(t)}$ means that task t has nonzero information about the true parameter $\boldsymbol{w}_*$. Therefore, there always exists some choice of the regularization weight matrix under which we can leverage the new information and improve on the existing estimator.

Moreover, in Section 7 we will empirically show that violating this assumption will not cause a significant performance degradation for our method.

Other loss functions. Our theory can be extended to general convex loss functions. In this scenario, the Hessian matrix of the loss function at the true parameter plays the role of the data covariance matrix in linear regression. The heterogeneity among different tasks is encoded by the differences in the Hessian matrices. Our analysis can then proceed with some modifications.

Common true parameters. Our model (1) assumes that all tasks share the same true parameter $\boldsymbol{w}_*$. In real-world continual learning, new challenges may arise when the true parameters are different across tasks. Analyzing the setting with distinct true parameters may need to introduce more trade-offs and insights. For example, if the parameters are not too far apart, our results may still hold with an additional error term. On the other hand, if the parameters differ significantly, negative transfer may dominate and continual learning might not work at all. We leave a comprehensive analysis of this problem to future work.

7. Experiments

We conduct simulation experiments to illustrate the performance of continual ridge regression (CRR), the minimum norm estimator (MN), and the generalized ℓ_2-regularized estimator (GR).

Data generation. We consider two data generating settings, namely with and without covariate shift. The difference between them is whether the covariance matrices are the same for different tasks.

(1) *Without covariate shift.* The true parameter $\boldsymbol{w}_*$ is sampled from $\mathcal{N}(0, \boldsymbol{I}_p)$ and is fixed for each task. The features $\boldsymbol{x}_i^{(t)}$ are independently sampled from $\mathcal{N}(0, \boldsymbol{I}_p)$ and the noises $\varepsilon_i^{(t)}$ are independently sampled from $\mathcal{N}(0, \sigma^2)$. Then the labels are generated by $y_i^{(t)} = \boldsymbol{w}_*^\top \boldsymbol{x}_i^{(t)} + \varepsilon_i^{(t)}$.

(2) *With covariate shift.* The true parameter $\boldsymbol{w}_*$ is sampled from $\mathcal{N}(0, \boldsymbol{I}_p)$ and is fixed for each task. The covariance matrices of the features are generated as follows. We first randomly sample the eigenvalues $\gamma_t^{(j)}$ by $P(\gamma_t^{(j)} = 1) = 0.99$ and $P(\gamma_t^{(j)} = 100) = 0.01$. Then the covariance matrices are set by $\boldsymbol{\Sigma}_t := \mathrm{diag}\{\gamma_t^{(1)}, \ldots, \gamma_t^{(p)}\}$. After the covariance matrices are generated, the features $\boldsymbol{x}_i^{(t)}$ are independently sampled from $\mathcal{N}(0, \boldsymbol{\Sigma}_t)$ and the noises $\varepsilon_i^{(t)}$ are independently sampled from $\mathcal{N}(0, \sigma^2)$. Finally, the labels are generated by $y_i^{(t)} = \boldsymbol{w}_*^\top \boldsymbol{x}_i^{(t)} + \varepsilon_i^{(t)}$.

Experimental configuration. We compare the performance of the CRR, MN, and GR algorithms with that of the oracle estimator. The regularization weight matrices of GR are set to $\tilde{\boldsymbol{H}}_t$ as discussed in Section 4.3.

We set the task number $T = 20$ and sample size $n_1 = \cdots = n_t = 150$. The parameter dimension $p = 200$, and hence each single task is overparameterized. We consider two noise levels: $\sigma^2 = 1$ or 5. We repeated our experiments 100 times and present the average results.

Simulation results. The simulation results for different noise levels are depicted in Figure 1. We observe that the estimation error of the MN estimator remains nearly constant as the task number t increases. Furthermore, a higher noise level makes the MN estimator perform worse than the other methods. This highlights the sensitivity of MN to noise.

The simulation results with and without covariate shift are contrasted in Figure 2, from which we find that covariate shift makes the CRR estimator worse. In the absence of covariate shift, CRR exhibits a decreasing loss, even though

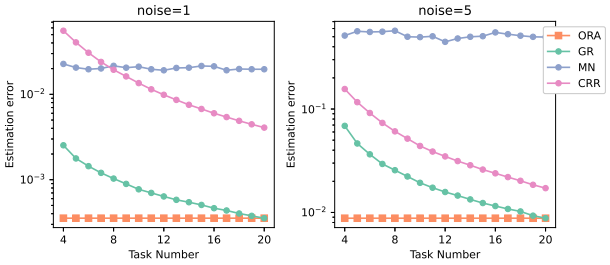

Figure 1. Simulation results for different noise levels: $T = 20$, $n_t = 150$, $p = 200$, $\sigma^2 = 1$ or 5, and no covariate shift.

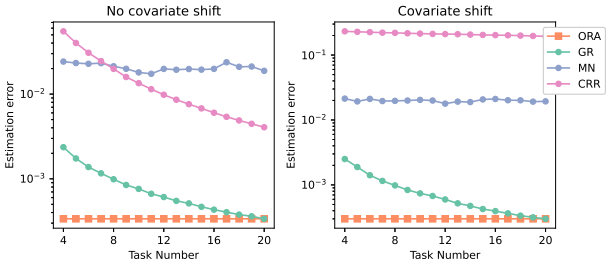

Figure 2. Simulation results with and without covariate shift: $T = 20$, $n_t = 150$, $p = 200$, and $\sigma^2 = 1$.

it is inferior to the GR estimator. In the presence of covariate shift, the performance of CRR deteriorates significantly, and its estimation error remains approximately constant.

In each case, the GR estimator consistently demonstrates a decreasing estimation error, which eventually converges to the oracle estimator. It is noteworthy that, due to the random generating process for sampling the features, Assumption 4.2 does not hold for the empirical covariance matrices. Nevertheless, this departure does not adversely impact the performance of our method.

8. Conclusion

Our analysis focuses on regularization-based continual learning across a series of linear regression tasks. We establish the estimation error of the oracle estimator with access to all data concurrently. We then explore a set of generalized ℓ_2-regularization algorithms characterized by matrix-valued hyperparameters. We develop an iterative formula to update the estimation error for these generalized ℓ_2-regularized estimators when new tasks are introduced. This allows us to identify the hyperparameters that optimize the performance of the algorithm. Remarkably, selecting the optimal hyperparameters achieves a balanced trade-off between forward and backward knowledge transfer and accommodates the variability in data distribution. Furthermore, we explicitly derive the estimation error of the optimal algorithm, which is found to match the order for the oracle estimator. Finally, we show that early stopping and generalized ℓ_2-regularization, rather than the conventional ridge regression, are equivalent in the context of continual learning, thereby addressing a question raised by Evron et al. (2023) on the connection between early stopping and explicit regularization in continual learning.

Acknowledgments

Xuyang Zhao and Wei Lin are supported by National Natural Science Foundation of China grants 12171012, 12292980, and 12292981. Weiran Huang is supported by 2023 CCF-Baidu Open Fund and Microsoft Research Asia. We thank three reviewers for valuable comments that have helped improve the paper.

Impact Statement

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

References

Ali, A., Kolter, J. Z., and Tibshirani, R. J. A continuous-time view of early stopping for least squares regression. In *International Conference on Artificial Intelligence and Statistics*, pp. 1370–1378, 2019.

Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., and Tuytelaars, T. Memory aware synapses: Learning what (not) to forget. In *European Conference on Computer Vision*, pp. 139–154, 2018.

Andle, J. and Yasaei Sekeh, S. Theoretical understanding of the information flow on continual learning performance. In *European Conference on Computer Vision*, pp. 86–101, 2022.

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

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

Cao, X., Liu, W., and Vempala, S. Provable lifelong learning of representations. In *International Conference on Artificial Intelligence and Statistics*, pp. 6334–6356, 2022.

Chaudhry, A., Ranzato, M., Rohrbach, M., and Elhoseiny, M. Efficient lifelong learning with A-GEM. In *International Conference on Learning Representations*, 2019.

Chen, X., Papadimitriou, C., and Peng, B. Memory bounds for continual learning. In *IEEE Symposium on Foundations of Computer Science*, pp. 519–530, 2022.

Denevi, G., Ciliberto, C., Grazzi, R., and Pontil, M. Learning-to-learn stochastic gradient descent with biased regularization. In *International Conference on Machine Learning*, pp. 1566–1575, 2019.

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

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

Evron, I., Moroshko, E., Buzaglo, G., Khriesh, M., Marjieh, B., Srebro, N., and Soudry, D. Continual learning in linear classification on separable data. In *International Conference on Machine Learning*, pp. 9440–9484, 2023.

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

Goodfellow, I. J., Mirza, M., Xiao, D., Courville, A., and Bengio, Y. An empirical investigation of catastrophic forgetting in gradient-based neural networks. In *International Conference on Learning Representations*, 2014.

Huszár, F. Note on the quadratic penalties in elastic weight consolidation. *Proceedings of the National Academy of Sciences*, 115(11):E2496–E2497, 2018.

Jin, X., Sadhu, A., Du, J., and Ren, X. Gradient-based editing of memory examples for online task-free continual learning. *Advances in Neural Information Processing Systems*, 34:29193–29205, 2021.

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

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

Lei, Q., Hu, W., and Lee, J. Near-optimal linear regression under distribution shift. In *International Conference on Machine Learning*, pp. 6164–6174, 2021.

Li, H., Wu, J., and Braverman, V. Fixed design analysis of regularization-based continual learning. In *Conference on Lifelong Learning Agents*, 2023.

Li, Y., Li, M., Asif, M. S., and Oymak, S. Provable and efficient continual representation learning. *arXiv preprint arXiv:2203.02026*, 2022.

Lin, S., Yang, L., Fan, D., and Zhang, J. Trgp: Trust region gradient projection for continual learning. In *International Conference on Learning Representations*, 2022.

Lin, S., Ju, P., Liang, Y., and Shroff, N. Theory on forgetting and generalization of continual learning. In *International Conference on Machine Learning*, pp. 21078–21100, 2023.

Liu, H. and Liu, H. Continual learning with recursive gradient optimization. In *International Conference on Learning Representations*, 2022.

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

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., and Wermter, S. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

Raskutti, G., Wainwright, M. J., and Yu, B. Early stopping and non-parametric regression: An optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15(11):335–366, 2014.

Riemer, M., Cases, I., Ajemian, R., Liu, M., Rish, I., Tu, Y., and Tesauro, G. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *International Conference on Learning Representations*, 2019.

Saha, G., Garg, I., and Roy, K. Gradient projection memory for continual learning. In *International Conference on Learning Representations*, 2021.

Schwarz, J., Czarnecki, W., Luketina, J., Grabska-Barwinska, A., Teh, Y. W., Pascanu, R., and Hadsell, R. Progress & compress: A scalable framework for continual learning. In *International Conference on Machine Learning*, pp. 4528–4537, 2018.

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

Wang, L., Zhang, X., Su, H., and Zhu, J. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Wen, Y. W., Tan, Z., Zheng, K., Xie, C., and Huang, W. Provable contrastive continual learning. In *International Conference on Machine Learning*, 2024.

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

Yang, C., Tiomoko, M., and Wang, Z. Optimizing spca-based continual learning: A theoretical approach. In *International Conference on Learning Representations*, 2022.

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

Yin, D., Farajtabar, M., Li, A., Levine, N., and Mott, A. Optimization and generalization of regularization-based continual learning: A loss approximation viewpoint. *arXiv preprint arXiv:2006.10974*, 2020.

Yoon, J., Kim, S., Yang, E., and Hwang, S. J. Scalable and order-robust continual learning with additive parameter decomposition. In *International Conference on Learning Representations*, 2020.

Appendix

A. Proofs for Section 3

Proof of Theorem 3.1. By the definition of minimum norm estimator, we have

$$\boldsymbol{X}_t\widehat{\boldsymbol{w}}_t^{(\mathrm{MN})} = \boldsymbol{y}_t$$

for each $t \in [T]$. Therefore,

$$\boldsymbol{X}_t(\widehat{\boldsymbol{w}}_t^{\mathrm{MN}} - \boldsymbol{w}_*) = \boldsymbol{y}_t - \boldsymbol{X}_t\boldsymbol{w}_* = \boldsymbol{\varepsilon}_t,$$

which implies

$$(\widehat{\boldsymbol{w}}_t^{(\mathrm{MN})} - \boldsymbol{w}_*)^\top \boldsymbol{\Sigma}_t(\widehat{\boldsymbol{w}}_t^{\mathrm{MN}} - \boldsymbol{w}_*) = \frac{1}{n_t}\|\boldsymbol{X}_t(\widehat{\boldsymbol{w}}_t - \boldsymbol{w}_*)\|^2 = \frac{1}{n_t}\|\boldsymbol{\varepsilon}_t\|^2.$$

Taking expectation with respect to $\boldsymbol{\varepsilon}_t$ on both sides gives

$$E(\widehat{\boldsymbol{w}}_t^{(\mathrm{MN})} - \boldsymbol{w}_*)^\top \boldsymbol{\Sigma}_t(\widehat{\boldsymbol{w}}_t^{(\mathrm{MN})} - \boldsymbol{w}_*) = \boldsymbol{\sigma}^2.$$

By the property of eigenvalues, we have

$$E(\widehat{\boldsymbol{w}}_t^{(\mathrm{MN})} - \boldsymbol{w}_*)^\top \boldsymbol{\Sigma}_t(\widehat{\boldsymbol{w}}_t^{(\mathrm{MN})} - \boldsymbol{w}_*) \le \max_{j\in[p]} \gamma_j^{(t)} E\|\widehat{\boldsymbol{w}}_t - \boldsymbol{w}_*\|^2.$$

Therefore, we finally conclude that

$$E\|\widehat{\boldsymbol{w}}_t - \boldsymbol{w}_*\|^2 \ge \frac{\sigma^2}{\max_{j\in[p]} \gamma_j^{(t)}}.$$

□

Proof of Theorem 3.2. Without loss of generality, we assume $\boldsymbol{w}_{*,1}^2 = \boldsymbol{w}_{*,1}^2 = 1$. In this two-task problem, the definition of CRR estimator is

$$\widehat{\boldsymbol{w}}_1 = \arg\min_{\boldsymbol{w}} \left\{ \frac{1}{n_1}\|\boldsymbol{X}_1\boldsymbol{w} - \boldsymbol{y}_1\|^2 + \lambda_1\|\boldsymbol{w}\|^2 \right\},$$

$$\widehat{\boldsymbol{w}}_2 = \arg\min_{\boldsymbol{w}} \left\{ \frac{1}{n_2}\|\boldsymbol{X}_2\boldsymbol{w} - \boldsymbol{y}_2\|^2 + \lambda_2\|\boldsymbol{w} - \widehat{\boldsymbol{w}}_1\|^2 \right\},$$

where λ_1 and λ_2 are the hyperparameters.

Task 1 By taking derivatives, we can explicitly obtain the solution of $\widehat{\boldsymbol{w}}_1$:

$$\widehat{\boldsymbol{w}}_1 = (\boldsymbol{\Sigma}_1 + \lambda_1\mathbf{I}_2)^{-1}(\boldsymbol{X}_1\boldsymbol{y}_1/n_1).$$

By the definition of $\boldsymbol{\Sigma}_1$, we further have

$$\widehat{\boldsymbol{w}}_{1,1} = \frac{1}{1+\lambda_1}\frac{\boldsymbol{X}_{1,1}^\top\boldsymbol{y}_1}{n_1}$$

and

$$\widehat{\boldsymbol{w}}_{1,2} = \frac{1}{\epsilon+\lambda_1}\frac{\boldsymbol{X}_{1,2}^\top\boldsymbol{y}_1}{n_1},$$

where $\widehat{\boldsymbol{w}}_{1,j}$ is the jth coordinate of $\widehat{\boldsymbol{w}}$ and $\boldsymbol{X}_{1,j}$ is the jth column of $\boldsymbol{X}$. Therefore, using the definition of $\boldsymbol{\Sigma}_1$ again we obtain

$$\begin{aligned}
\mathbb{E}\,(\widehat{\boldsymbol{w}}_{1,1} - \boldsymbol{w}_{*,1})^2 &= \mathbb{E}\left(\frac{1}{1+\lambda_1}\frac{\boldsymbol{X}_{1,1}^\top(\boldsymbol{X}_1\boldsymbol{w}_* + \boldsymbol{\varepsilon}_1)}{n_1} - \boldsymbol{w}_{*,1}\right)^2 \\
&= \mathbb{E}\left(\frac{\boldsymbol{w}_{*,1}}{1+\lambda_1} - \boldsymbol{w}_{*,1} + \frac{1}{1+\lambda_1}\frac{\boldsymbol{X}_{1,1}^\top\boldsymbol{\varepsilon}_1}{n_1}\right)^2 \\
&= \left(\frac{\lambda_1}{1+\lambda_1}\right)^2 + \left(\frac{1}{1+\lambda_1}\right)^2\frac{\sigma^2}{n_1} \\
&\ge \frac{\sigma^2}{n_1+\sigma^2},
\end{aligned}$$

where the last equation holds if and only if $\lambda_1 = \sigma^2/n_1$. Similarly, for $\widehat{\boldsymbol{w}}_{1,2}$ we have

$$\begin{aligned}\mathbb{E}\left(\widehat{\boldsymbol{w}}_{1,2} - \boldsymbol{w}_{*,2}\right)^2 &= \mathbb{E}\left(\frac{1}{\epsilon+\lambda_1}\frac{\boldsymbol{X}_{1,2}^\top(\boldsymbol{X}_1\boldsymbol{w}_* + \boldsymbol{\varepsilon}_1)}{n_1} - \boldsymbol{w}_{*,2}\right)^2 \\ &= \mathbb{E}\left(\frac{\epsilon}{\epsilon+\lambda_1}\boldsymbol{w}_{*,2} - \boldsymbol{w}_{*,2} + \frac{1}{\epsilon+\lambda_1}\frac{\boldsymbol{X}_{1,2}^\top\boldsymbol{\varepsilon}_1}{n_1}\right)^2 \\ &= \left(\frac{\lambda_1}{\epsilon+\lambda_1}\right)^2 + \frac{\epsilon}{(\epsilon+\lambda_1)^2}\frac{\sigma^2}{n_1} \\ &\geq \frac{\sigma^2}{\epsilon n_1 + \sigma^2},\end{aligned}$$

where the last equation holds if and only if $\lambda_1 = \sigma^2/\epsilon n_1$.

Task 2 Through almost the same analysis, for $\widehat{\boldsymbol{w}}_2$ we have

$$\begin{aligned}\mathbb{E}\left[(\widehat{\boldsymbol{w}}_{2,1} - \boldsymbol{w}_{*,1})^2 | \widehat{\boldsymbol{w}}_1\right] &= \mathbb{E}\left[\frac{1}{\epsilon+\lambda_2}\left(\frac{\boldsymbol{X}_{2,1}^\top(\boldsymbol{X}_2\boldsymbol{w}_* + \boldsymbol{\varepsilon}_2)}{n_2} + \lambda_2\widehat{\boldsymbol{w}}_{1,1}\right) - \boldsymbol{w}_{*,1}\right]^2 \\ &= \left(\frac{\lambda_2}{\epsilon+\lambda_2}\right)^2 (\widehat{\boldsymbol{w}}_{1,1} - \boldsymbol{w}_{*,1})^2 + \frac{\epsilon}{(\epsilon+\lambda_2)^2}\frac{\sigma^2}{n_2}\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\left[(\widehat{\boldsymbol{w}}_{2,2} - \boldsymbol{w}_{*,2})^2 | \widehat{\boldsymbol{w}}_1\right] &= \mathbb{E}\left[\frac{1}{1+\lambda_2}\left(\frac{\boldsymbol{X}_{2,2}^\top(\boldsymbol{X}_2\boldsymbol{w}_* + \boldsymbol{\varepsilon}_2)}{n_2} + \lambda_2\widehat{\boldsymbol{w}}_{1,2}\right) - \boldsymbol{w}_{*,2}\right]^2 \\ &= \left(\frac{\lambda_2}{1+\lambda_2}\right)^2 (\widehat{\boldsymbol{w}}_{1,2} - \boldsymbol{w}_{*,2})^2 + \frac{1}{(1+\lambda_2)^2}\frac{\sigma^2}{n_2}.\end{aligned}$$

From Theorem 4.2, we know that

$$\mathbb{E}(\widehat{\boldsymbol{w}}_{2,1}^{(\mathrm{GR})} - \boldsymbol{w}_{*,1})^2 = O\left(\frac{\sigma^2}{n_1 + \epsilon n_2}\right)$$

and

$$\mathbb{E}(\widehat{\boldsymbol{w}}_{2,2}^{(\mathrm{GR})} - \boldsymbol{w}_{*,2})^2 = O\left(\frac{\sigma^2}{\epsilon n_1 + n_2}\right).$$

By some calculations, if

$$\frac{\mathbb{E}(\widehat{\boldsymbol{w}}_{2,1}^{(\mathrm{CRR})} - \boldsymbol{w}_{*,1})^2}{\mathbb{E}(\widehat{\boldsymbol{w}}_{2,1}^{(\mathrm{GR})} - \boldsymbol{w}_{*,1})^2} < \infty$$

when $n_1, n_2 \to \infty$ and $\epsilon \to 0$, we need

$$\epsilon\frac{1 - \sqrt{\frac{\epsilon n_2}{n_1+\epsilon n_2}}}{\sqrt{\frac{\epsilon n_2}{n_1+\epsilon n_2}}} \lesssim \lambda_2 \lesssim \epsilon\frac{\sqrt{\frac{n_1}{n_1+\epsilon n_2}}}{1 - \sqrt{\frac{n_1}{n_1+\epsilon n_2}}}.$$

If

$$\frac{\mathbb{E}(\widehat{\boldsymbol{w}}_{2,2}^{(\mathrm{CRR})} - \boldsymbol{w}_{*,2})^2}{\mathbb{E}(\widehat{\boldsymbol{w}}_{2,2}^{(\mathrm{GR})} - \boldsymbol{w}_{*,2})^2} < \infty$$

when $n_1, n_2 \to \infty$ and $\epsilon \to 0$, we need

$$\sqrt{\frac{\epsilon n_1 + n_2}{n_2}} - 1 \lesssim \lambda_2 \lesssim \frac{\sqrt{\frac{1}{\epsilon n_1+n_2}}}{1 - \sqrt{\frac{1}{\epsilon n_1+n_2}}}.$$

We consider a special case, where $\epsilon n_2/n_1 \to \infty$. Then the above requirements become

$$\epsilon \frac{n_1}{n_1 + \epsilon n_2} \lesssim \lambda_2 \lesssim \epsilon \sqrt{\frac{n_1}{n_1 + \epsilon n_2}}$$

and

$$\frac{\epsilon n_1}{n_2} \lesssim \lambda_2 \lesssim \sqrt{\frac{1}{\epsilon n_1 + n_2}}.$$

However, if $n_1 = O(n^2)$, $n_2 = O(n^3)$ and $\epsilon = O(n^{-0.5})$ for some $n \to \infty$, the lower bound of the first inequality if greater than the upper bound of the second inequality:

$$\epsilon \frac{n_1}{n_1 + \epsilon n_2} = O(n^{-1})$$

while

$$\sqrt{\frac{1}{\epsilon n_1 + n_2}} = O(n^{-1.5}).$$

Therefore, by contradiction we have

$$\sup_{n_1, n_2, \epsilon} \inf_{\lambda_2} \frac{\mathcal{L}(\widehat{\boldsymbol{w}}_2^{(\mathrm{CRR})})}{\mathcal{L}(\widehat{\boldsymbol{w}}_2^{(\mathrm{GR})})} = \infty.$$

□

B. Proofs for Section 4

Proof of Lemma 4.1. The oracle estimator $\widehat{\boldsymbol{w}}_T^{(\mathrm{ORA})}$ satisfies

$$\sum_{t=1}^{T} X_t^\top (\boldsymbol{X}_t \widehat{\boldsymbol{w}}_T^{(\mathrm{ORA})} - \boldsymbol{y}_t) = 0,$$

which implies

$$\left(\sum_{t=1}^{T} \boldsymbol{X}_t^\top \boldsymbol{X}_t \right) \widehat{\boldsymbol{w}}_T^{(\mathrm{ORA})} = \sum_{t=1}^{T} \boldsymbol{X}_t^\top \boldsymbol{y}_t.$$

By Assumption 4.2 and 4.3, we have that $\sum_{t=1}^{T} \boldsymbol{X}_t^\top \boldsymbol{X}_t$ is invertible. Therefore, the ORA has the following explicit form solution

$$\begin{aligned}
\widehat{\boldsymbol{w}}_T^{(\mathrm{ORA})} &= \left(\sum_{t=1}^{T} \boldsymbol{X}_t^\top \boldsymbol{X}_t \right)^{-1} \left(\sum_{t=1}^{T} \boldsymbol{X}_t^\top \boldsymbol{y}_t \right) \\
&= \left(\sum_{t=1}^{T} \boldsymbol{X}_t^\top \boldsymbol{X}_t \right)^{-1} \left(\sum_{t=1}^{T} \boldsymbol{X}_t^\top (\boldsymbol{X}_t \boldsymbol{w}_* + \boldsymbol{\varepsilon}_t) \right) \\
&= \boldsymbol{w}_* + \left(\sum_{t=1}^{T} \boldsymbol{X}_t^\top \boldsymbol{X}_t \right)^{-1} \left(\sum_{t=1}^{T} \boldsymbol{X}_t^\top \boldsymbol{\varepsilon}_t \right).
\end{aligned}$$

Taking expectation with respect to $\{\boldsymbol{\varepsilon}_t\}_{t=1}^T$, we obtain

$$\begin{aligned}
\mathbb{E} \left\| \widehat{\boldsymbol{w}}_T^{(\mathrm{ora})} - \boldsymbol{w}_* \right\|^2 &= E \left\| \left(\sum_{t=1}^{T} \boldsymbol{X}_t^\top \boldsymbol{X}_t \right)^{-1} \left(\sum_{t=1}^{T} \boldsymbol{X}_t^\top \boldsymbol{\varepsilon}_t \right) \right\|^2 \\
&= tr \left\{ \left(\sum_{t=1}^{T} \boldsymbol{X}_t^\top \boldsymbol{X}_t \right)^{-1} \right\}.
\end{aligned}$$

By Assumption 4.2, we can further have

$$\operatorname{tr}\left\{\left(\sum_{t=1}^{T}\boldsymbol{X}_t^\top\boldsymbol{X}_t\right)^{-1}\right\}=\operatorname{tr}\left\{\left(\sum_{t=1}^{T}n_t\boldsymbol{\Gamma}_t\right)^{-1}\right\}$$
$$=\sum_{j=1}^{p}\frac{1}{n_1\gamma_j^{(1)}+\ldots n_T\gamma_j^{(T)}},$$

which completes the proof.

□

Proof of Theorem 4.2. For each $t=1,\ldots,T$, the solution of GR estimator $\widehat{\boldsymbol{w}}_t^{(\mathrm{GR})}$ satisfies

$$\frac{1}{n_t}\boldsymbol{X}_t^\top(\boldsymbol{X}_t\widehat{\boldsymbol{w}}_t^{(\mathrm{GR})}-\boldsymbol{y}_t)+\boldsymbol{H}_t(\widehat{\boldsymbol{w}}_t^{(\mathrm{GR})}-\widehat{\boldsymbol{w}}_{t-1}^{(\mathrm{GR})})=0$$

Since $\boldsymbol{y}_t=\boldsymbol{X}_t\boldsymbol{w}_*+\boldsymbol{\varepsilon}_t$, it can be written explicitly as

$$\widehat{\boldsymbol{w}}_t^{(\mathrm{GR})}=\boldsymbol{w}_*+(\boldsymbol{X}_t^\top\boldsymbol{X}_t+n\boldsymbol{H}_t)^{-1}\boldsymbol{X}_t^\top\boldsymbol{\varepsilon}_t+(\boldsymbol{X}_t^\top\boldsymbol{X}_t+n_t\boldsymbol{H}_t)^{-1}n_t\boldsymbol{H}_t(\widehat{\boldsymbol{w}}_{t-1}^{(\mathrm{GR})}-\boldsymbol{w}_*).$$

Therefore, for each $j=1,\ldots,p$,

$$\boldsymbol{u}_j^\top(\widehat{\boldsymbol{w}}_t^{(\mathrm{GR})}-\boldsymbol{w}_*)=\boldsymbol{u}_j^\top(n_t\boldsymbol{\Sigma}_t+n_t\boldsymbol{H}_t)^{-1}\boldsymbol{X}_t^\top\boldsymbol{\varepsilon}_t+(n_t\boldsymbol{\Sigma}_t+n_t\boldsymbol{H}_t)^{-1}n_t\boldsymbol{\Lambda}_t\boldsymbol{U}^\top(\widehat{\boldsymbol{w}}_{t-1}^{(\mathrm{GR})}-\boldsymbol{w}_*),$$

which implies that

$$\begin{aligned}
\mathbb{E}\left[e_j^{(t)}\right]&=\mathbb{E}\left[\left\|\boldsymbol{u}_j^\top(\widehat{\boldsymbol{w}}_t^{(\mathrm{GR})}-\boldsymbol{w}_*)\right\|^2\right]\\
&\overset{\text{(i)}}{=}\mathbb{E}\left[(\boldsymbol{u}_j^\top(\boldsymbol{X}_t^\top\boldsymbol{X}_t+n_t\boldsymbol{H}_t)^{-1}\boldsymbol{X}_t\boldsymbol{\varepsilon}_t)^2\right]+\mathbb{E}\left[(\boldsymbol{u}_j^\top(\boldsymbol{X}_t^\top\boldsymbol{X}_t+n_t\boldsymbol{H}_t)^{-1}n_t\boldsymbol{H}_t(\widehat{\boldsymbol{w}}_{t-1}-\boldsymbol{w}_*))^2\right]\\
&\overset{\text{(ii)}}{=}\mathbb{E}\left[(\boldsymbol{u}_j^\top\boldsymbol{U}(n_t\boldsymbol{\Gamma}_t+n_t\boldsymbol{\Lambda}_t)^{-1}\boldsymbol{U}^\top\boldsymbol{X}_t\boldsymbol{\varepsilon}_t)^2\right]+n_t^2\mathbb{E}\left[(\boldsymbol{u}_j^\top\boldsymbol{U}(n_t\boldsymbol{\Gamma}_t+n_t\boldsymbol{\Lambda}_t)^{-1}\boldsymbol{\Lambda}\boldsymbol{U}^\top(\widehat{\boldsymbol{w}}_{t-1}-\boldsymbol{w}_*))^2\right]\\
&\overset{\text{(iii)}}{=}\frac{\gamma_j^{(t)}\sigma^2/n+(\lambda_j^{(t)})^2\mathbb{E}\left[e_j^{(t-1)}\right]}{(\lambda_j^{(t)}+\gamma_j^{(t)})^2}\\
&=\frac{\gamma_j^{(t)}\sigma^2/n+(\lambda_j^{(t)}+\gamma_j^{(t)}-\gamma_j^{(t)})^2\mathbb{E}\left[e_j^{(t-1)}\right]}{(\lambda_j^{(t)}+\gamma_j^{(t)})^2}\\
&=\mathbb{E}\left[e_j^{(t-1)}\right]-2\frac{\gamma_j^{(t)}\mathbb{E}\left[e_j^{(t-1)}\right]}{\lambda_j^{(t)}+\gamma_j^{(t)}}+\frac{(\gamma_j^{(t)})^2\mathbb{E}\left[e_j^{(t-1)}\right]+\gamma_j^{(t)}\sigma^2/n}{(\lambda_j^{(t)}+\gamma_j^{(t)})^2},
\end{aligned}\tag{11}$$

where (i) comes from the independence between $\boldsymbol{\varepsilon}_t$ and $\widehat{\boldsymbol{w}}_{t-1}^{(\mathrm{GR})}$, (ii) comes from Assumption 4.2 and (iii) is obtained by the property of eigenvalues and eigenvectors. To derive the optimal value of $\lambda_j^{(t)}$ Now we consider two different cases:

1. Consider the case $\gamma_j=0$. Then as long as $\lambda_j>0$, $\boldsymbol{\Sigma}_t+\boldsymbol{H}_t$ is invertible and we have

$$\mathbb{E}\left[e_j^{(t)}\right]=\mathbb{E}\left[e_j^{(t-1)}\right].$$

This means that data of task t do not bring new information about the direction j of $\boldsymbol{w}_*$, which makes the jth projected error $e_j^{(t)}$ unchanged.

2. Consider the case $\gamma_j > 0$. In this case, the last formula of Equation 11 can be regarded as a quadratic function of $1/(\lambda_j^{(t)} + \gamma_j^{(t)})$ as $\lambda_j^{(t)}$ changes. Therefore, the optimal λ_j is obtained by

$$\frac{1}{\lambda_j^{(t)} + \gamma_j^{(t)}} = \frac{\mathbb{E}\left[e_j^{(t-1)}\right]}{\gamma_j^{(t)}\mathbb{E}\left[e_j^{(t-1)}\right] + \sigma^2/n_t},$$

which is the minimum of the quadratic function. This further implies that $\lambda_j^{(t)} = \frac{\sigma^2/n_t}{\mathbb{E}(e_j^{(t-1)})^2}$, where we have

$$\begin{aligned}
\mathbb{E}\left[e_j^{(t)}\right] &= \frac{\left(E\left[e_j^{(t-1)}\right]\right)^2 \gamma_j^2 + \mathbb{E}\left[e_j^{(t-1)}\right]\gamma_j\sigma^2/n_t - \gamma_j^2\left(\mathbb{E}\left[e_j^{(t-1)}\right]\right)^2}{\gamma_j^2\left(\mathbb{E}\left[e_j^{(t-1)}\right]\right)^2 + \gamma_j\sigma^2/n_t} \\
&= \frac{\mathbb{E}\left[e_j^{(t-1)}\right]\cdot\sigma^2/(\gamma_j n_t)}{\mathbb{E}\left[e_j^{(t-1)}\right] + \sigma^2/(\gamma_j n_t)} \\
&= \frac{1}{\left(\mathbb{E}\left[e_j^{(t-1)}\right]\right)^{-1} + (\sigma^2/(\gamma_j n_t))^{-1}}.
\end{aligned} \tag{12}$$

To prove the final results, we consider mathematical induction. By Assumption 4.3, for each $j \in [p]$, there exists $\tau_j \in [T]$ such that $\tau_j > 0$. Therefore, by the above derivation, $e_j^{(\tau_j)}$ satisfies

$$\begin{aligned}
\mathbb{E}\left[e_j^{(\tau_j)}\right] &= \frac{1}{(e_j^{(0)})^{-1} + (\sigma^2/(\gamma_j^{(\tau_j)} n_{\tau_j}))^{-1}} \\
&= \frac{\sigma^2}{\sigma^2/e_j^{(0)} + \gamma_j^{(\tau_j)} n_{\tau_j}} \\
&= \frac{\sigma^2}{\sigma^2/e_j^{(0)} + \gamma_j^{(1)} n_1 + \cdots + \gamma_j^{(\tau_j)} n_{\tau_j}}.
\end{aligned}$$

For $t < \tau_j$, by Case (1) discussed above we have

$$\mathbb{E}\left[e_j^{(t)}\right] = \frac{\sigma^2}{\sigma^2/e_j^{(0)} + \gamma_j^{(1)} n_1 + \cdots + \gamma_j^{(t-1)} n_t}$$

since $\gamma_j^{(t)} = 0$ for every $t < \tau_j$.

Now suppose

$$\mathbb{E}\left[e_j^{t}\right] = \frac{\sigma^2}{\sigma^2/e_j^{(0)} + \gamma_j^{(1)} n_1 + \cdots + \gamma_j^{(t)} n_t}$$

holds for some $t \geq \tau_j$. If $\gamma_j^{(t+1)} = 0$, by Case (1) we have

$$\mathbb{E}\left[e_j^{(t+1)}\right] = \mathbb{E}e_j^{(t)} = \frac{\sigma^2}{\sigma^2/e_j^{(0)} + \gamma_j^{(1)} n_1 + \cdots + \gamma_j^{(t+1)} n_{t+1}}$$

since $\gamma_j^{(t+1)} = 0$. If $\gamma_j^{(t+1)} > 0$, by Case (2) we have

$$\begin{aligned}
\mathbb{E}\left[e_j^{(t+1)}\right] &= \frac{1}{\left(Ee_j^{(t)}\right)^{-1} + \left(\sigma^2/\gamma_j^{(t+1)} n_{t+1}\right)^{-1}} \\
&= \frac{\sigma^2}{\sigma^2/e_j^{(0)} + \gamma_j^{(1)} n_1 + \cdots + \gamma_j^{(t+1)} n_{t+1}}.
\end{aligned}$$

Therefore, we conclude that for each $t \in [T]$,

$$\mathbb{E}[e_j^{(t)}] = \frac{\sigma^2}{\sigma^2/e_j^{(0)} + \gamma_j^{(1)} n_1 + \cdots + \gamma_j^{(t)} n_t}$$

holds if Λ_t is chosen by

$$\begin{aligned}\lambda_j^{(t)} &= \frac{\sigma^2/n_t}{Ee_j^{(t-1)}} \\ &= \frac{\sigma^2/e_j^{(0)} + \gamma_j^{(1)} n_1 + \cdots + \gamma_j^{(t-1)} n_{t-1}}{n_t}.\end{aligned}$$

Finally, since $\|\widehat{\boldsymbol{w}}_t^{\mathrm{GR}}) - \boldsymbol{w}_*\|^2 = \sum_{j=1}^p e_j^{(t)}$, taking summation of all $e_j^{(t)}$ gives

$$\mathcal{L}(\widehat{\boldsymbol{w}}_t^{(\mathrm{GR})}) = \sum_{j=1}^p \frac{\sigma^2}{\sigma^2/e_j^{(0)} + \gamma_j^{(1)} n_1 + \cdots + \gamma_j^{(t)} n_t}.$$

□

Proof of Theorem 4.3. Recall from the proof of Theorem 4.2 that

$$\begin{aligned}\mathbb{E}\left[e_j^{(t)}\right] &= \mathbb{E}(u_j^\top (X_t^\top X_t + nH_t)^{-1} X_t \varepsilon_t)^2 + \mathbb{E}(u_j^\top (X_t^\top X_t + nH_t)^{-1} nH_t(\widehat{w}_{t-1} - w_*))^2 \\ &= \mathbb{E}(u_j^\top U(n\Gamma_t + n\Lambda_t)^{-1} U^\top X_t \varepsilon_t)^2 + n^2 \mathbb{E}(u_j^\top U(n\Gamma_t + n\Lambda_t)^{-1} \Lambda U^\top (\widehat{w}_{t-1} - w_*))^2 \\ &= \frac{\gamma_j \sigma^2/n + \lambda_j^2 \mathbb{E} e_j^{(t-1)}}{(\lambda_j + \gamma_j)^2} \\ &= \frac{\gamma_j \sigma^2/n + (\lambda_j + \gamma_j - \gamma_j)^2 \mathbb{E} e_j^{(t-1)}}{(\lambda_j + \gamma_j)^2} \\ &= \mathbb{E} e_j^{(t-1)} - 2\frac{\gamma_j \mathbb{E} e_j^{(t-1)}}{\lambda_j + \gamma_j} + \frac{\gamma_j^2 \mathbb{E} e_j^{(t-1)} + \gamma_j \sigma^2/n}{(\lambda_j + \gamma_j)^2}.\end{aligned}$$

The optimal λ_j that minimize the above equation satisfies

$$\frac{1}{\lambda_j + \gamma_j} = \frac{\mathbb{E} e_j^{(t-1)}}{\gamma_j \mathbb{E} e_j^{(t-1)} + \sigma^2/n_t},$$

namely

$$\lambda_j = \frac{\sigma^2/n_t}{\mathbb{E} e_j^{(t-1)}}.$$

Now suppose we use its approximated version instead:

$$\frac{1}{\widetilde{\lambda}_j + \gamma_j} = \frac{1}{\lambda_j + \gamma_j} + \Delta$$

for some $\widetilde{\lambda}_j$. Then we have

$$\mathbb{E} e_j^{(t)} \le \frac{\mathbb{E} e_j^{(t-1)} \cdot \sigma^2/(\gamma_j n_t)}{\mathbb{E} e_j^{(t-1)} + \sigma^2/(\gamma_j n_t)} + (\gamma_j^2 \mathbb{E} e_j^{(t-1)} + \gamma_j \sigma^2/n)\Delta^2.$$

Suppose that

$$\mathbb{E} e_j^{(t-1)} \le \frac{C\sigma^2}{e_j^{(0)} + \gamma_j^{(1)} n_1 + \ldots \gamma_j^{(t-1)} n_{t-1}}.$$

If we want

$$\mathbb{E}e_j^{(t)} \le \frac{C\sigma^2}{e_j^{(0)} + \gamma_j^{(1)} n_1 + \dots \gamma_j^{(t)} n_t},$$

holds true, we only need to make sure

$$\begin{aligned}
&\frac{\mathbb{E}e_j^{t-1} \cdot \sigma^2/(\gamma_j n_t)}{\mathbb{E}e_j^{t-1} + \sigma^2/(\gamma_j n_t)} + (\gamma_j^2 \mathbb{E}e_j^{(t-1)} + \gamma_j \sigma^2/n)\Delta^2 \\
\le &\frac{C\sigma^2}{e_j^{(0)} + \gamma_j^{(1)} n_1 + \dots + \gamma_j^{(t-1)} n_{t-1} + C\gamma_j^{(t)} n_t} + \gamma_j^2 \Delta^2 \left(\frac{C\sigma^2}{e_j^0 + \gamma_j^1 n_1 + \dots \gamma_j^{t-1} n_{t-1}} + \frac{\sigma^2}{\gamma_j^t n_t} \right) \\
\le &\frac{C\sigma^2}{e_j^{(0)} + \gamma_j^{(1)} n_1 + \dots + \gamma_j^{(t)} n_t}.
\end{aligned}$$

Define

$$\rho_j^{(t)} := \frac{\gamma_j^{(t)} n_t}{e_j^{(0)} + \gamma_j^{(1)} n_1 + \dots + \gamma_j^{(t)} n_t},$$

then the above inequality becomes

$$\frac{C}{1 + C\rho_j^{(t)}} + (\gamma_j^{(t)})^2 \Delta^2 \left(C + \frac{1}{\rho_j^{(t)}} \right) \le \frac{C}{1 + \rho_j^{(t)}},$$

which is indeed

$$(\gamma_j^{(t)})^2 \Delta^2 \le \frac{C(C-1)(\rho_j^{(t)})^2}{(1 + \rho_j^{(t)})(1 + C\rho_j^{(t)})^2}.$$

□

C. Proofs for Section 5

Proof of Theorem 5.1. For each $t \in [T]$ and $\tau \in [m_t]$, using the update iteration of MN estimator we have

$$\begin{aligned}
\boldsymbol{w}_t^{(\tau)} - \boldsymbol{w}_* &= (\boldsymbol{I}_p - \boldsymbol{A}_t \boldsymbol{X}_t^\top \boldsymbol{X}_t / n_t) \boldsymbol{w}_t^{(\tau-1)} + (\boldsymbol{A}_t / n_t) \boldsymbol{X}_t^\top (\boldsymbol{X}_t \boldsymbol{w}_* + \boldsymbol{\varepsilon}_t) - \boldsymbol{w}_* \\
&= (\boldsymbol{I}_p - \boldsymbol{A}_t \boldsymbol{X}_t^\top \boldsymbol{X}_t / n_t)(\boldsymbol{w}_t^{(\tau-1)} - \boldsymbol{w}_*) + \boldsymbol{A}_t \boldsymbol{X}_t^\top \boldsymbol{\varepsilon}_t / n_t \\
&= (\boldsymbol{I}_p - \boldsymbol{A}_t \boldsymbol{X}_t^\top \boldsymbol{X}_t / n_t)^\tau (\boldsymbol{w}_{t-1}^{(\mathrm{ES})} - \boldsymbol{w}_*) + (\boldsymbol{I} - (\boldsymbol{I} - \boldsymbol{A}_t \boldsymbol{X}_t^\top \boldsymbol{X}_t / n_t)^\tau)(\boldsymbol{A}_t \boldsymbol{X}_t^\top \boldsymbol{X}_t / n_t)^{-1} \frac{\boldsymbol{A}_t}{n_t} \boldsymbol{X}_t^\top \boldsymbol{\varepsilon}_t \\
&= \boldsymbol{U}(\boldsymbol{I}_p - \boldsymbol{S}_t \boldsymbol{\Gamma}_t)^\tau \boldsymbol{U}^\top (\boldsymbol{w}_{t-1}^{(\mathrm{ES})} - \boldsymbol{w}_*) + \boldsymbol{U}(\boldsymbol{I}_p - (\boldsymbol{I}_p - \boldsymbol{S}_t \boldsymbol{\Gamma}_t)^\tau) \boldsymbol{\Gamma}_t^{-1} \boldsymbol{U}^\top \frac{\boldsymbol{X}_t^\top}{n_t} \boldsymbol{\varepsilon}_t.
\end{aligned}$$

Therefore, for each $j = 1, \dots, p$, we have

$$\boldsymbol{u}_j^\top (\boldsymbol{w}_t^{(\tau)} - \boldsymbol{w}_*) = (1 - s_j \gamma_j)^\tau \boldsymbol{u}_j^\top (\boldsymbol{w}_{t-1}^{(\mathrm{ES})} - \boldsymbol{w}_*) + (1 - (1 - s_j \gamma_j)^\tau) \boldsymbol{u}_j^\top \frac{\boldsymbol{X}_t^\top}{\gamma_j n} \boldsymbol{\varepsilon}_j.$$

Note that By the proof of Theorem 4.2, the solution of generalized ℓ_2 regularization estimator satisfies

$$\boldsymbol{u}_j^\top (\widehat{\boldsymbol{w}}_t^{(\mathrm{GR})} - \boldsymbol{w}_*) = (\gamma_j + \lambda_j)^{-1} \lambda_j \boldsymbol{u}_j^\top (\widehat{\boldsymbol{w}}_{t-1}^{(\mathrm{GR})} - \boldsymbol{w}_*) + (\gamma_j + \lambda_j)^{-1} \gamma_j \boldsymbol{u}_j^\top \frac{\boldsymbol{X}_t^\top}{\gamma_j n} \boldsymbol{\varepsilon}_t.$$

Therefore, if λ_j and s_j satisfy

$$(1 - s_j \gamma_j)^{m_t} = \frac{\lambda_j}{\gamma_j + \lambda_j},$$

namely

$$s_j = \frac{1-(\lambda_j/(\gamma_j+\lambda_j))^{1/m_t}}{\gamma_j}$$

or

$$\lambda_j = \frac{\gamma_j(1-s_j\gamma_j)^{m_t}}{1-(1-s_j\gamma_j)^{m_t}},$$

the early stopping and ℓ_2 regularization output the same estimator, i.e.,

$$\boldsymbol{w}_t^{(\mathrm{ES})} = \boldsymbol{w}_t^{(m_t)} = \widehat{\boldsymbol{w}}_t^{(\mathrm{GR})}.$$

□

Proof of Corollary 5.2. By Theorem 4.2 and Theorem 5.1, if $s_j^{(t)}$ and m_t satisfy

$$\left(1-s_j^{(t)}\gamma_j\right)^{m_t} = \frac{\sigma^2/(\gamma_j n)}{\mathbb{E}\left[e_j^{(t-1)}\right]+\sigma^2/(\gamma_j n)},$$

the ES estimator $\widehat{\boldsymbol{w}}_t^{(\mathrm{ES})}$ equals the optimal generalized ℓ_2 regularization estimator defined in Theorem 4.2. In this case, its estimation error satisfies

$$\mathcal{L}(\widehat{\boldsymbol{w}}_t^{(\mathrm{ES})}) = \sum_{j=1}^{p} \frac{\sigma^2}{\sigma^2/e_j^{(0)}+\gamma_j^{(1)}n_1+\cdots+\gamma_j^{(t)}n_t}.$$

□

D. Proof for Section 6

Proof of Theorem 6.1. For simplicity, we omit the superscript of the GR estimator in this proof. Let $\boldsymbol{U}_t\boldsymbol{\Gamma}_t\boldsymbol{U}_t^\top$ be the eigendecomposition of $\boldsymbol{\Sigma}_t$ and define $e_{j,t_1}^{(t_2)} := ((\boldsymbol{u}_j^{(t_1)})^\top(\widehat{\boldsymbol{w}}_{t_2}-\boldsymbol{w}_*))^2$ as the projected error of $\boldsymbol{w}_{t_2}$ onto the jth eigenvector of $\boldsymbol{w}_{t_1}$. Note that if Assumption 4.2 holds, $\boldsymbol{u}_j^{(t_1)} = \boldsymbol{u}_j^{(t_2)}$ for every $t_1, t_1 \in [T]$ and $e_{j,t_1}^{(t_2)}$ equals to $e_j^{(t_2)}$ defined in Section 4 for each t_1.

By the same derivation of (12), we directly have

$$\mathbb{E}\left[e_{j,t}^{(t)}\right] = \frac{1}{\left(\mathbb{E}\left[e_{j,t}^{(t-1)}\right]\right)^{-1}+(\sigma^2/(\gamma_j n_t))^{-1}} \le \mathbb{E}\left[e_{j,t}^{(t-1)}\right].$$

Therefore, summing them up with respect to j gives

$$\mathcal{L}(\widehat{\boldsymbol{w}}_t) \le \mathcal{L}(\widehat{\boldsymbol{w}}_{t-1}).$$

Obviously, the inequality holds strictly as long as $\sum_{j=1}^p \gamma_j^{(t)} > 0$, i.e., there exists one j such that $\gamma_j > 0$.

We remark that if $e_{j,t}^{(t-1)} = e_{j,t-1}^{(t-1)}$ for every t, we can use mathematical induction to derive the estimation error in Theorem 4.2. However, without Assumption 4.2, we cannot ensure this equation holds true.

□