FIXED DESIGN ANALYSIS OF REGULARIZATION-BASED CONTINUAL LEARNING

Haoran Li [*]
Department of Computer Science
Rice University
Houston, TX
lihr@rice.edu

Jingfeng Wu [*]
Department of Computer Science
Johns Hopkins University
Baltimore, MD
uuujf@jhu.edu

Vladimir Braverman
Department of Computer Science
Rice University
Houston, TX
vb21@rice.edu

ABSTRACT

We consider a continual learning (CL) problem with two linear regression tasks in the fixed design setting, where the feature vectors are assumed fixed and the labels are assumed to be random variables. We consider an ℓ_2-regularized CL algorithm, which computes an Ordinary Least Squares parameter to fit the first dataset, then computes another parameter that fits the second dataset under an ℓ_2-regularization penalizing its deviation from the first parameter, and outputs the second parameter. For this algorithm, we provide tight bounds on the average risk over the two tasks. Our risk bounds reveal a provable trade-off between forgetting and intransigence of the ℓ_2-regularized CL algorithm: with a large regularization parameter, the algorithm output forgets less information about the first task but is intransigent to extract new information from the second task; and vice versa. Our results suggest that catastrophic forgetting could happen for CL with dissimilar tasks (under a precise similarity measurement) and that a well-tuned ℓ_2-regularization can partially mitigate this issue by introducing intransigence.

1 INTRODUCTION

In *continual learning* (CL, also known as lifelong learning), an agent is provided with multiple learning tasks in a *sequential* manner and aims to solve these tasks with *limited long-term memory*. Due to the shortage of long-term memory, a CL problem is fundamentally more challenging than a single-task learning problem, as the CL agent is unable to memorize all data (neither can the agent memorize multiple task-specific solutions respectively) and thus a CL problem cannot be reduced to a single-task (i.e., the combination of all tasks) learning problem (Parisi et al., 2019). On the other hand, without using exceedingly large long-term memory, one may view a CL problem as an online multi-task problem and fits a model with the data provided for each task, sequentially. However, the final model outputted by such an online approach could be poor for CL, as the performance is evaluated with respect to all seen tasks, but the final model tends to forget information from earlier tasks. This issue is known as *catastrophic forgetting* (McCloskey & Cohen, 1989).

One effective category of CL methods mitigates catastrophic forgetting by *regularization*. Specifically, the CL agent stores a copy of the latest model in the long-term memory, and when fitting a new model for an incoming task, a regularization is imposed to prevent the model from deviating too much from the old one (Kirkpatrick et al., 2017; Aljundi et al., 2018; Zenke et al., 2017). The regularization-based methods temper catastrophic forgetting by lowering its flexibility of fitting new data (aka, *intransigence*): with a strong regularization strength, the agent tends to memorize old tasks more and learn the new task less; and vice versa for a weaker regularization strength. When the regularization strength is properly set, the regularization-based methods can learn new tasks well without forgetting old tasks catastrophically, hence achieving good performance for CL. Understanding the *trade-off* between *forgetting* with *intransigence* is the key to understanding the effectiveness of the regularization-based CL methods.

Contributions. This work theoretically investigates the trade-off between forgetting and intransigence induced by a regularization-based CL algorithm. Specifically, we consider a (domain-incremental) CL problem with two linear regression tasks in the fix design setting (see Assumption 1). We consider an ℓ_2-regularized CL algorithm (ℓ_2-RCL), where an ℓ_2-regularization is imposed when learning the second task, penalizing the model deviating from the one learned from the first task (see (3)). We make the following contributions:

---

[*] Equal Contribution

1. We provide sharp risk bounds for ℓ_2-RCL. This implies a provable trade-off between forgetting and intransigence, which is controlled by the regularization parameter: a strong regularization effectively reduces forgetting but unavoidably enlarges intransigence and vice versa.

2. We show that, without regularization, forgetting could become catastrophic when the two tasks are dissimilar (under a precise task similarity measurement). Moreover, for CL with moderately dissimilar tasks, the catastrophic forgetting can be tempered with a well-tuned ℓ_2-regularization by introducing a tolerable intransigence.

3. Finally, we show that there exists a CL problem with extremely dissimilar tasks that cannot be solved by ℓ_2-RCL. In this case, forgetting and intransigence are in strong tension and cannot be reduced simultaneously. This indicates the necessity of considering a more advanced regularization for CL.

Our theoretical results are validated with numerical simulations. All proofs are deferred to the appendix.

2   PRELIMINARIES

Two linear regression tasks.   We set up a 2-task CL problem with two linear regression problems – our results can be easily extended to CL with more than two tasks. Let $\left(\mathbf{x}_t^{(1)}, y_t^{(1)}\right)_{t=1}^n \in \mathbb{R}^d \times \mathbb{R}$ be n pair of data vector and label variable that are drawn independently from a population distribution $\mathcal{D}^{(1)}$, where $d > 0$ refers to the ambient dimension. Similarly, let $\left(\mathbf{x}_t^{(2)}, y_t^{(2)}\right)_{t=1}^n \in \mathbb{R}^d \times \mathbb{R}$ be another n pair of data vector and label variable that are drawn independently from a different population distribution $\mathcal{D}^{(2)}$. For simplicity, we denote

$$\mathbf{X}^{(k)} := \left(\mathbf{x}_1^{(k)}, \ldots, \mathbf{x}_n^{(k)}\right)^\top \in \mathbb{R}^{n \times d}, \quad \mathbf{y}^{(k)} := \left(y_1^{(k)}, \ldots, y_n^{(k)}\right) \in \mathbb{R}^n,$$

for $k = 1, 2$. We use $\mathbf{w} \in \mathbb{R}^d$ to denote a model parameter. Then we define the population risks for the two linear regression tasks by

$$\mathcal{R}_1(\mathbf{w}) := \frac{1}{n}\mathbb{E}_{\mathcal{D}^{(1)}}\|\mathbf{y}^{(1)} - \mathbf{X}^{(1)}\mathbf{w}\|_2^2, \qquad \mathcal{R}_2(\mathbf{w}) := \frac{1}{n}\mathbb{E}_{\mathcal{D}^{(2)}}\|\mathbf{y}^{(2)} - \mathbf{X}^{(2)}\mathbf{w}\|_2^2,$$

respectively.

Continual learning.   The goal of CL is to learn a parameter to minimize the *joint population risk*

$$\mathcal{R}(\mathbf{w}) := \mathcal{R}_1(\mathbf{w}) + \mathcal{R}_2(\mathbf{w}). \tag{1}$$

Unlike other multi-task learning problems, in CL the agent learns the parameter under the constraints that the task-specific datasets can only be accessed sequentially and that its long-term memory is limited. Specifically, a two-task CL problem involves two local learning phases and a memory consolidation phase (Kirkpatrick et al., 2017; Zenke et al., 2017). A CL algorithm begins with the first local learning phase, where the CL agent is provided with the first dataset $\left(\mathbf{X}^{(1)}, \mathbf{y}^{(1)}\right)$ and is allowed to do arbitrary computations with the dataset. After that, the agent is switched to the memory consolidation phase and must delete all of its local data except for a d-dimensional vector, which we call *long-term memory*. Then the agent enters the second local learning phase, where it retains the long-term memory and is provided with the second dataset $\left(\mathbf{X}^{(2)}, \mathbf{y}^{(2)}\right)$, and is allowed to do arbitrary computation with the d-dimensional long-term memory vector and the dataset. The aim of the CL agent is to report a model (parameter) that achieves a small joint population risk (1).

Remark 1 (Memory size). In the above CL setup, we only allow a d-dimensional long-term memory vector. On the one hand, this amount of memory is necessary for storing a d-dimensional model parameter (without additional knowledge). On the other hand, the condition can be relaxed to allow a kd-dimensional long-term memory vector for $k > 1$, which enables more powerful CL algorithms (e.g., EWC (Kirkpatrick et al., 2017), SCP (Kolouri et al., 2020), sketched EWC (Li et al., 2021), etc.). As this paper initiates the theoretical framework for CL, we focus on the simpler setting and leave the general cases as future works.

We now introduce two CL methods, called by us *ordinary continual learning* (OCL) and ℓ_2-*regularized continual learning* (ℓ_2-RCL), respectively.

Ordinary continual learning.   In the first local learning phase, the *ordinary continual learning* (OCL) performs *ordinary least square* (Bartlett et al., 2020) with the first dataset. In the memory consolidation phase, it saves the obtained model parameter (an d-dimensional vector, denoted by $\mathbf{w}^{(1)}$) as its long-term memory. Then in the second local learning phase, it computes a model parameter (denoted by $\mathbf{w}^{(2)}$) that fits the second dataset while minimizing

the ℓ_2-deviation from the previous model parameter (i.e., the long-term memory vector $\mathbf{w}^{(1)}$). Specifically, OCL outputs $\mathbf{w}^{(2)}$ such that

$$\mathbf{w}^{(1)} = \left(\mathbf{X}^{(1)^\top}\mathbf{X}^{(1)}\right)^{-1}\mathbf{X}^{(1)^\top}\mathbf{y}^{(1)}; \qquad \mathbf{w}^{(2)} = \mathbf{w}^{(1)} + \left(\mathbf{X}^{(2)^\top}\mathbf{X}^{(2)}\right)^{-1}\mathbf{X}^{(2)^\top}\left(\mathbf{y}^{(2)} - \mathbf{X}^{(2)}\mathbf{w}^{(1)}\right). \tag{2}$$

It is worth noting that OCL has been considered in Evron et al. (2022). However, their focus is on the optimization behaviors of OCL, while we focus on its statistical behaviors.

Remark 2 (OCL in the interpolation regime). When the number of parameters exceeds the number of samples, i.e., $d > 2n$, then both datasets can be interpolated almost surely (Bartlett et al., 2020). In the interpolation regime, (2) corresponds to

$$\mathbf{w}^{(1)} = \underset{\mathbf{w}:\mathbf{y}^{(1)}=\mathbf{X}^{(1)}\mathbf{w}}{\arg\min} \|\mathbf{w}\|_2^2; \qquad \mathbf{w}^{(2)} = \underset{\mathbf{w}:\mathbf{y}^{(2)}=\mathbf{X}^{(2)}\mathbf{w}}{\arg\min} \|\mathbf{w} - \mathbf{w}^{(1)}\|_2^2.$$

Our results make no assumption on d and can be applied in both underparameterized and overparameterized cases.

ℓ_2-Regularized continual learning. The ℓ_2-regularized continual learning (ℓ_2-RCL) is identical to OCL in the first local learning phase and the memory consolidation phase. In the second local learning phase, however, the agent computes a model parameter that fits the second dataset under an ℓ_2-penalty from deviating from the previous model parameter. Specially, ℓ_2-RCL outputs $\mathbf{w}^{(2)}$ such that

$$\mathbf{w}^{(1)} = \left(\mathbf{X}^{(1)^\top}\mathbf{X}^{(1)}\right)^{-1}\mathbf{X}^{(1)^\top}\mathbf{y}^{(1)}; \qquad \mathbf{w}^{(2)} = \arg\min_{\mathbf{w}} \frac{1}{n}\|\mathbf{y}^{(2)} - \mathbf{X}^{(2)}\mathbf{w}\|_2^2 + \mu\|\mathbf{w} - \mathbf{w}^{(1)}\|_2^2, \tag{3}$$

where $\mu > 0$ is a hyperparameter. The ℓ_2-RCL has been studied by Heckel (2022). However, they focus on the existence of continual learning settings that the ℓ_2-RCL training can provably succeed or fail, while we focus on the success and failure condition of the ℓ_2-RCL algorithm. We also note that the ℓ_2-RCL algorithm is a special case of the EWC algorithm (Kirkpatrick et al., 2017) when the covariance matrix of the first dataset is isotropic (i.e., all eigenvalues are equal). Finally, we note that the ℓ_2-RCL output reduces to OCL when $\mu \to 0$.

A forgetting-intransigence decomposition. A CL algorithm targets to minimize a joint population risk (e.g., (1) in our setting). Additionally, in practice, the performance of a CL algorithm is often measured by the following two metrics: (1) *forgetting*, which refers to the amount of performance degradation with respect to the old tasks after learning a new task, and (2) *intransigence*, which evaluates the (in)ability of the algorithm to fit a new task after having already fitted a sequence of old tasks (Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2018a). Heuristically, forgetting is believed to be in tension with intransigence, and a good CL algorithm achieves a balance between forgetting and intransigence (Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2018a).

Our main results theoretically characterize the trade-off between forgetting and intransigence for the ℓ_2-RCL algorithm. Specifically, let $\mathbf{w}^{(2)}$ and $\mathbf{w}^{(1)}$ be given by (3), then we define forgetting and intransigence by

$$\begin{aligned} \mathcal{F}\left(\mathbf{w}^{(2)}, \mathbf{w}^{(1)}\right) &:= \mathcal{R}_1\left(\mathbf{w}^{(2)}\right) - \mathcal{R}_1\left(\mathbf{w}^{(1)}\right), \\ \mathcal{I}\left(\mathbf{w}^{(2)}, \mathbf{w}^{(1)}\right) &:= \mathcal{R}_2\left(\mathbf{w}^{(2)}\right) - \min \mathcal{R}_2(\cdot) + \mathcal{R}_1\left(\mathbf{w}^{(1)}\right) - \min \mathcal{R}_1(\cdot), \end{aligned} \tag{4}$$

respectively. The following *forgetting-intransigence decomposition* holds by a direct calculation:

$$\mathcal{R}\left(\mathbf{w}^{(2)}\right) - \min \mathcal{R}(\cdot) = \mathcal{F}\left(\mathbf{w}^{(2)}, \mathbf{w}^{(1)}\right) + \mathcal{I}\left(\mathbf{w}^{(2)}, \mathbf{w}^{(1)}\right). \tag{5}$$

This suggests that: for the ℓ_2-RCL output (3) to induce a small CL risk (1), it is necessary that the induced forgetting in (4) is small (or negative), and that the induced intransigence in (4) is also small. In the next section, we will show sharp bounds for both forgetting and intransigence.

Remark 3 (Generalization to CL with T tasks). The forgetting-intransigence decomposition (4) is defined for a CL problem with only two tasks. This can be generalized easily. Specifically, let $\mathbf{w}^{(t)}$ be a sequence of model parameters generated by a CL algorithm learning the t-th task, $t = 1, \ldots, T$. Then we can define

$$\mathcal{F}\left(\mathbf{w}^{(t)}, t=1,\ldots,T\right) := \sum_{t=1}^{T-1} \mathcal{R}_t\left(\mathbf{w}^{(T)}\right) - \mathcal{R}_t\left(\mathbf{w}^{(t)}\right), \quad \mathcal{I}\left(\mathbf{w}^{(t)}, t=1,\ldots,T\right) := \sum_{t=1}^{T} \mathcal{R}_t\left(\mathbf{w}^{(t)}\right) - \min \mathcal{R}_t(\cdot).$$

Similarly, the following *forgetting-intransigence decomposition* holds:

$$\sum_{t=1}^{T} \mathcal{R}_t\left(\mathbf{w}^{(T)}\right) - \min \mathcal{R}_t(\cdot) = \mathcal{F}\left(\mathbf{w}^{(t)}, t=1,\ldots,T\right) + \mathcal{I}\left(\mathbf{w}^{(t)}, t=1,\ldots,T\right).$$

This work focuses on $T = 2$ for the conciseness of a theoretical treatment.

3 MAIN RESULTS

Before presenting the main results, we first review a set of assumptions for our analysis.

3.1 ASSUMPTIONS

Assumption 1 (Fixed design). *Assume that the feature vectors $\left(\mathbf{x}_t^{(1)}\right)_{t=1}^n$ and $\left(\mathbf{x}_t^{(2)}\right)_{t=1}^n$ are fixed and that the labels $\left(y_t^{(1)}\right)_{t=1}^n$ and $\left(y_t^{(2)}\right)_{t=1}^n$ are independent random variables.*

This work focuses on a fixed design setting as described by Assumption 1, where features are fixed but the labels are still random variables (Hsu et al., 2012). In comparison, the work by Evron et al. (2022) assumed that both features and labels are fixed. As will be demonstrated later, even in the fixed design we already show very interesting statistic properties of CL that are unknown before our work. We leave the more interesting, a random design (Hsu et al., 2012) analysis of CL as a future direction.

Assumption 2 (Shared optimal parameter). *Assume that there exists a $\mathbf{w}^* \in \mathbb{R}^d$ such that*

$$\mathbf{w}^* \in \arg\min \mathcal{R}_1(\mathbf{w}), \quad \mathbf{w}^* \in \arg\min \mathcal{R}_2(\mathbf{w}).$$

We assume there is a common optimal parameter for the two tasks. This corresponds to the *domain-incremental CL* (Van de Ven & Tolias, 2019). In comparison, Evron et al. (2022) also assumed such a condition when studying the optimization behavior of CL. However, Heckel (2022); Lin et al. (2023) considered a more general CL setting, where the optimal model parameters might be different for different tasks. As shown later, the domain-incremental setting already reflects a key statistical trade-off between forgetting and intransigence in CL.

The next assumption is standard in linear regression literature, which specifies a well-specified additive label noise.

Assumption 3 (Well-specified noise). *Assume that for the $\mathbf{w}^*$ in Assumption 2, it holds that: for $k = 1, 2$ and $t = 1, \ldots, n$,*

$$\mathbb{E}[y_t^{(k)}] = \mathbf{x}_t^{(k)^\top} \mathbf{w}^*, \quad \sigma^2 := \mathbb{E}\big(y_t^{(k)} - \mathbf{x}_t^{(k)^\top} \mathbf{w}^*\big)^2,$$

where $\sigma^2 > 0$ refers to the variance of the label noise.

For conciseness, we assume that the two tasks have the same noise level. This is not restrictive, and our results can be directly generalized to allow different noise levels as well.

Assumption 4 (Commutable data covariance matrices). *Assume that*

$$\mathbf{H}^{(1)} \mathbf{H}^{(2)} = \mathbf{H}^{(2)} \mathbf{H}^{(1)}, \quad where \ \mathbf{H}^{(1)} := \frac{1}{n} \mathbf{X}^{(1)^\top} \mathbf{X}^{(1)} \ and \ \mathbf{H}^{(2)} := \frac{1}{n} \mathbf{X}^{(2)^\top} \mathbf{X}^{(2)}.$$

Denote the eigenvalues of $\mathbf{H}^{(1)}$ and $\mathbf{H}^{(2)}$ by $\left(\lambda_i^{(1)}\right)_{i=1}^d$ and $\left(\lambda_i^{(2)}\right)_{i=1}^d$, respectively. Denote the corresponding eigenvectors to be $(\mathbf{v}_i)_{i=1}^d$.

In this work, we assume that the data covariance matrices are commutable. This is a common assumption in transfer learning literature, see, e.g., Lei et al. (2021) and Section 4 in Wu et al. (2022). Note that this assumption does not requires $\mathbf{H}^{(1)}$ and $\mathbf{H}^{(2)}$ to be diagonal. Extensions to our results by relaxing this assumption are left as an open technical problem.

Notations. For two functions $f(x) \geq 0$ and $g(x) \geq 0$ defined for $x \in [0, \infty)$, we write $f(x) = \mathcal{O}(g(x))$ if $\lim_{x\to\infty} f(x)/g(x) < c$ for some absolute constant $c > 0$; we write $f(x) = \Omega(g(x))$ if $g(x) = \mathcal{O}(f(x))$; and we write $f(x) = \Theta(g(x))$ if $f(x) = \mathcal{O}(g(x))$ and $g(x) = \mathcal{O}(f(x))$. Moreover, we write $f(x) = o(g(x))$ if $\lim_{x\to\infty} f(x)/g(x) = 0$.

For a matrix $\mathbf{A} := \sum_{i=1}^d a_i \mathbf{v}_i \mathbf{v}_i^\top$, where $(\mathbf{v}_i)_{i\geq 1}$ are the eigenvectors of $\mathbf{H}^{(1)}$ (and $\mathbf{H}^{(2)}$) and $(a_i)_{i\geq 1}$ are non-negative scalars, and an index set $\mathbb{K} \subset \{1, \ldots, d\}$, we define

$$\mathbf{A}_{\mathbb{K}} := \sum_{i \in \mathbb{K}} a_i \mathbf{v}_i \mathbf{v}_i^\top, \quad \mathbf{A}_{\mathbb{K}}^{-1} := \sum_{i \in \mathbb{K}, a_i > 0} \frac{1}{a_i} \mathbf{v}_i \mathbf{v}_i^\top.$$

3.2 RISK BOUNDS FOR JOINT LEARNING

When the two datasets can be accessed simultaneously, one can solve the two-task learning problem by *joint learning* (JL), which outputs $\mathbf{w}_{\texttt{joint}}$ such that

$$\mathbf{w}_{\texttt{joint}} = \underset{\mathbf{w}:\mathbf{y}^{(1)}=\mathbf{X}^{(1)}\mathbf{w},\mathbf{y}^{(2)}=\mathbf{X}^{(2)}\mathbf{w}}{\arg\min} \|\mathbf{w}\|_2^2. \tag{6}$$

Note that JL is not a practical CL method, as it requires either simultaneous access to both datasets or equivalently, a nd-dimensional long-term memory vector for storing the first dataset. As JL uses strictly more information than every CL algorithm, it can serve as an imaginary "upper" baseline method that describes the best possible outcome of solving a CL problem (Farajtabar et al., 2020; Mirzadeh et al., 2020; Saha et al., 2021). We follow this convention and limit ourselves to considering only the CL problems that can be solved efficiently by JL. To this end, the following proposition provides a risk bound for solving two linear regression tasks with JL.

Proposition 1 (A risk bound for JL). *Suppose that Assumptions 1 to 4 hold. Then for the JL output* (6)*, it holds that*

$$\mathbb{E}\left[\mathcal{R}(\mathbf{w}_{\texttt{joint}})\right] - \mathcal{R}(\mathbf{w}^*) = \frac{\sigma^2}{n} \cdot \operatorname{rank}(\mathbf{H}^{(1)} + \mathbf{H}^{(2)}).$$

The effect of JL. According to Proposition 1, for JL to achieve an $o(1)$ excess risk, it is necessary (and also sufficient) to have that

$$\operatorname{rank}(\mathbf{H}^{(1)}) = o(n), \quad \operatorname{rank}(\mathbf{H}^{(2)}) = o(n). \tag{7}$$

As discussed earlier, for a CL problem to be learnable by a valid CL algorithm, it has to be learnable by the JL algorithm. Therefore, in what follows when we analyze CL algorithms, we will only focus on the CL problems where (7) holds.

3.3 RISK BOUNDS FOR CONTINUAL LEARNING

In this part, we present our main results. We first provide general risk bounds (Theorem 1) for ℓ_2-RCL (3) with any regularization parameter $\mu \geq 0$; in particular, they allow $\mu = 0$, therefore, cover OCL (2) as a special case.

Theorem 1 (A risk bound for ℓ_2-RCL/OCL). *Suppose that Assumptions 1 to 4 hold. Then for the ℓ_2-RCL output* (3)*, it holds that*

$$\mathbb{E}\left[\mathcal{R}(\mathbf{w}^{(2)})\right] - \min \mathcal{R}(\cdot) = \mathbb{E}\left[\mathcal{F}(\mathbf{w}^{(2)}, \mathbf{w}^{(1)})\right] + \mathbb{E}\left[\mathcal{I}(\mathbf{w}^{(2)}, \mathbf{w}^{(1)})\right].$$

Moreover, it holds that

$$\mathbb{E}\left[\mathcal{F}(\mathbf{w}^{(2)}, \mathbf{w}^{(1)})\right] = \frac{\sigma^2}{n} \cdot \mathbf{d}_{\texttt{F}},$$

$$\mathbb{E}\left[\mathcal{I}(\mathbf{w}^{(2)}, \mathbf{w}^{(1)})\right] = \texttt{bias} + \frac{\sigma^2}{n} \cdot \mathbf{d}_{\texttt{I}},$$

where

$$\mathbf{d}_{\texttt{F}} := \operatorname{tr}\left\{\mathbf{I}_{\mathbb{K}} \cdot \left(\mathbf{H}^{(1)} - \mathbf{H}^{(2)} - 2\mu\mathbf{I}\right) \cdot \mathbf{H}^{(2)} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-2}\right\},$$

$$\mathbf{d}_{\texttt{I}} := \operatorname{tr}\left\{\left(\mu^2 \cdot \mathbf{H}^{(1)^{-1}} + \mathbf{H}^{(2)}\right) \cdot \mathbf{H}^{(2)} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-2}\right\} + \operatorname{rank}\left\{\mathbf{H}^{(1)}\right\}, \tag{8}$$

$$\texttt{bias} := (\mathbf{w}^*)^\top \cdot \mathbf{I}_{\mathbb{K}^c} \cdot \mu^2 \cdot \mathbf{H}^{(2)} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-2} \cdot \mathbf{w}^*,$$

for index sets $\mathbb{K} := \{i : \lambda_i^{(1)} > 0\}$.

In particular, when the regularization parameter $\mu = 0$, i.e., for the OCL output (2)*, it holds that*

$$\mathbf{d}_{\texttt{F}} = \operatorname{tr}\left\{\mathbf{H}^{(1)}\mathbf{H}^{(2)^{-1}}\right\} - \operatorname{rank}\left\{\mathbf{H}^{(1)}\mathbf{H}^{(2)}\right\}, \quad \mathbf{d}_{\texttt{I}} = \operatorname{rank}\left\{\mathbf{H}^{(2)}\right\} + \operatorname{rank}\left\{\mathbf{H}^{(1)}\right\}, \quad \texttt{bias} = 0. \tag{9}$$

Theorems 1 tightly characterizes the performance of ℓ_2-RCL and OCL methods in the fixed design setting. Specifically, the CL risk is decomposed into a forgetting part and an intransigence part, and sharp bounds on both parts are established. In the following, we discuss the effects of OCL and ℓ_2-RCL in turn.

The effect of OCL. Consider (9) in Theorem 1 that characterizes the forgetting and intransigence induced by OCL. When the CL problem can be solved by JL, i.e., (7) holds, it is clear that $\mathtt{d_I}$ is small, i.e., OCL has a small intransigence. Then the performance of OCL is dominated by its forgetting according to Theorem 1. In particular, for OCL to achieve an $o(1)$ CL excess risk, it is sufficient (and also necessary, as $\mathrm{rank}\{\mathbf{H}^{(1)}\mathbf{H}^{(2)}\} = o(n)$ under (7)) to have

$$\mathtt{d_F} \leq \mathrm{tr}\left\{\mathbf{H}^{(1)}\mathbf{H}^{(2)^{-1}}\right\} = o(n). \tag{10}$$

This requires the small eigenvalue space of $\mathbf{H}^{(2)}$ to be low-dimensional and/or to be aligned with the small eigenvalue space of $\mathbf{H}^{(1)}$. On the other hand, if (10) does not hold, then OCL suffers from constant forgetting, i.e., catastrophic forgetting happens. From this perspective, (10) provides a precise task similarity measurement, such that OCL is able to solve CL problems with similar tasks but is unable to solve CL problems with dissimilar tasks due to catastrophic forgetting.

For understanding the performance of ℓ_2-RCL, the following corollary is useful.

Corollary 1 (Risk upper bounds for ℓ_2-RCL). *Under the same setup of Theorem 1, it holds that*

$$
\begin{aligned}
\mathtt{d_F} &\leq \mathrm{tr}\left\{\mathbf{H}^{(1)} \cdot \mathbf{H}^{(2)} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-2}\right\} \\
&\leq \mathrm{tr}\left\{\mathbf{H}^{(1)} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1}\right\}, \\
\mathtt{d_I} &\leq \mathrm{tr}\left\{\mu^2 \cdot \mathbf{H}^{(1)^{-1}} \cdot \mathbf{H}^{(2)} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-2}\right\} + \mathrm{rank}\{\mathbf{H}^{(2)}\} + \mathrm{rank}\{\mathbf{H}^{(1)}\} \\
&\leq \mathrm{tr}\left\{\mu \cdot \mathbf{H}^{(1)^{-1}}\right\} + \mathrm{rank}\{\mathbf{H}^{(2)}\} + \mathrm{rank}\{\mathbf{H}^{(1)}\}, \\
\mathtt{bias} &\leq \mu \cdot \|\mathbf{w}^*\|_2^2.
\end{aligned} \tag{11}
$$

The effect of ℓ_2-RCL. We now discuss the effect of ℓ_2-RCL based on the forgetting and intransigence bounds in Theorem 1 and their upper bounds in Corollary 1. Firstly, according to (8) (or (11)), the regularization parameter μ adjusts the trade-off between the forgetting and intransigence parts in the risk bound of ℓ_2-RCL. Specifically, a large μ reduces $\mathtt{d_F}$ but promotes $\mathtt{d_I}$ and $\mathtt{bias}$, i.e., the forgetting becomes smaller but the intransigence becomes larger, and vice versa. Our theory on the effect of the regularization parameter in ℓ_2-RCL is consistent with the empirical observation for the regularization-based CL methods (Kirkpatrick et al., 2017; Kolouri et al., 2020; Li et al., 2021).

Secondly, we compare the effect of ℓ_2-RCL with that of OCL. According to Theorem 1, for ℓ_2-RCL to achieve an $o(1)$ CL excess risk, it suffices to choose a regularization parameter μ such that $\mathtt{d_F}, \mathtt{d_I} \leq o(n)$ and that $\mathtt{bias} \leq o(1)$. Consider the upper bounds for these terms in (11), and under the additional assumption that the problem can be solved by JL (i.e., (7) holds), it suffices to require that

$$\mathrm{tr}\left\{\mathbf{H}^{(1)} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1}\right\} \leq o(n), \quad \mu \cdot \mathrm{tr}\left\{\mathbf{H}^{(1)^{-1}}\right\} \leq o(n), \quad \mu \leq o(1), \tag{12}$$

holds for some $\mu \geq 0$. Clearly, (12) is a weaker set of requirements than (10) as the latter corresponds to the former with $\mu = 0$. Moreover, even when the two tasks in CL are dissimilar such that OCL cannot achieve a small risk, i.e., (10) does not hold, ℓ_2-RCL is still possible to achieve a small risk as long as (12) holds for some $\mu > 0$. For example, as long as $\mathrm{tr}\{\mathbf{H}^{(1)}\} = o(n)$ and $\mathrm{tr}\{\mathbf{H}^{(1)}\} \cdot \mathrm{tr}\{\mathbf{H}^{(1)^{-1}}\} = o(n^2)$, (12) can be made true for a well tuned μ. Note that in this example, $\mathbf{H}^{(2)}$ might not be well aligned with $\mathbf{H}^{(1)}$ and (10) might not hold. These suggest that a well-tuned ℓ_2-RCL achieves a trade-off between forgetting and intransigence superior to that achieved by OCL.

Thirdly, it is worth noting that there exist CL problems that can be solved by JL but not by ℓ_2-RCL, even with an optimally tuned regularization parameter. See, e.g., Corollary 3. This demonstrates the limitation of ℓ_2-RCL.

3.4 EXAMPLES AND NUMERICAL SIMULATIONS

In this part, we give specific examples to further validate the above discussions on the trade-off between forgetting and intransigence induced by OCL and ℓ_2-RCL. We show that catastrophic forgetting happens when the two tasks are dissimilar.

When is forgetting catastrophic for OCL? We have shown that (10) is a sufficient and necessary condition that determines when forgetting is catastrophic for OCL. We further demonstrate this with the following examples.

Example 1. *Under the same setting as Theorem 1, additionally, assume that*

$$\sigma^2 = \Theta(1), \quad \|\mathbf{w}^*\|_2 = \mathcal{O}(1), \quad \mathrm{rank}\{\mathbf{H}^{(1)}\} = \mathcal{O}(1), \quad \mathrm{rank}\{\mathbf{H}^{(2)}\} = \mathcal{O}(1).$$

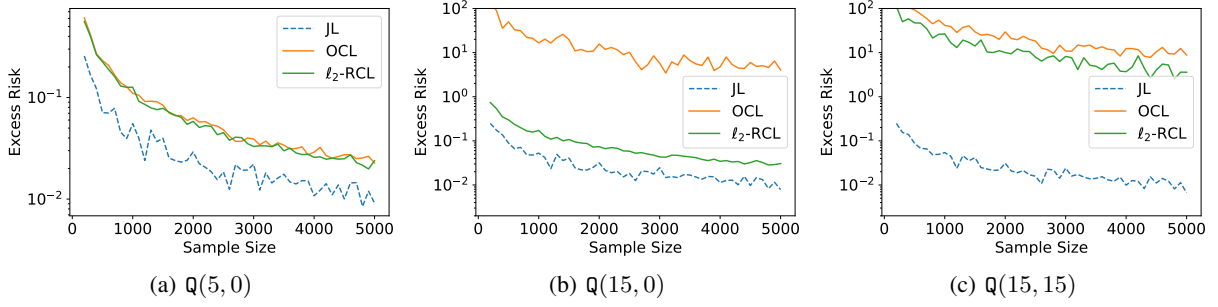*Consider the OCL output (2) (or (3) with $\mu = 0$), then:*

Figure 1: Expected excess risk vs. sample size for JL, OCL and ℓ_2-RCL. For each point in each curve, the x-axis represents the sample size n and the y-axis represents the expected CL excess risk (in the logarithmic scale). The problem instances $\mathtt{Q}(5,0)$, $\mathtt{Q}(15,0)$ and $\mathtt{Q}(15,15)$ are defined by (13). The regularization parameter in the ℓ_2-RCL algorithm is optimally tuned. The dimension of the task is $d = 200$. The expectation of the excess risk is computed by taking an empirical average over 20 independent runs.

1. *If $\|\mathbf{H}^{(1)}\|_2 = \mathcal{O}(1)$ and all non-zero eigenvalues of $\mathbf{H}^{(2)}$ are $\Omega(n^{-1+r})$ for some constant $0 < r \le 1$, then $\mathbb{E}[\mathcal{R}(\mathbf{w}^{(2)})] - \mathcal{R}(\mathbf{w}^*) = \mathcal{O}\left(n^{-r}\right)$. Moreover, it holds that*

$$\mathbb{E}[\mathcal{F}(\mathbf{w}^{(2)}, \mathbf{w}^{(1)})] = \mathcal{O}(n^{-r}), \quad \mathbb{E}[\mathcal{I}(\mathbf{w}^{(2)}, \mathbf{w}^{(1)})] = \mathcal{O}(n^{-1}).$$

2. *If $\lambda_1^{(1)} = 1$ and $\lambda_1^{(2)} = n^{-1}$, then $\mathbb{E}[\mathcal{R}(\mathbf{w}^{(2)})] - \mathcal{R}(\mathbf{w}^*) = \Omega\left(1\right)$. Moreover, it holds that*

$$\mathbb{E}[\mathcal{F}(\mathbf{w}^{(2)}, \mathbf{w}^{(1)})] = \Omega(1), \quad \mathbb{E}[\mathcal{I}(\mathbf{w}^{(2)}, \mathbf{w}^{(1)})] = \mathcal{O}(n^{-1}).$$

In the first problem in Example 1, $\mathbf{H}^{(2)}$ is well conditioned, so it is "similar" to $\mathbf{H}^{(1)}$ according to (10), and the forgetting is small for OCL. In the second problem in Example 1, there exists a direction where $\mathbf{H}^{(2)}$ and $\mathbf{H}^{(1)}$ are highly dissimilar, then the forgetting becomes catastrophic for OCL.

When can ℓ_2-RCL temper catastrophic forgetting? We now present an example where the forgetting is catastrophic for OCL but the issue can be mitigated by ℓ_2-RCL via introducing intransigence.

Example 2. *Under the same setting as Theorem 1, additionally, assume that*

$$\mathbf{w}^* = (0, 1, 0, 0, \dots), \quad \sigma^2 = 1,$$

$$\mathbf{H}^{(1)} = \mathrm{diag}(1, 0, \underbrace{1, \dots, 1}_{\Theta(1) \text{ copies}}, 0, 0, \dots), \quad \mathbf{H}^{(2)} = \mathrm{diag}(n^{-1}, n^{-2/3}, \underbrace{1, \dots, 1}_{\Theta(1) \text{ copies}}, 0, 0, \dots).$$

Then:

- *For the OCL output (2), it holds that $\mathbb{E}[\mathcal{R}(\mathbf{w}^{(2)})] - \mathcal{R}(\mathbf{w}^*) = \Omega\left(1\right)$; moreover,*

$$\mathbb{E}[\mathcal{F}(\mathbf{w}^{(2)}, \mathbf{w}^{(1)})] = \Omega(1), \quad \mathbb{E}[\mathcal{I}(\mathbf{w}^{(2)}, \mathbf{w}^{(1)})] = \mathcal{O}(n^{-1}).$$

- *For the ℓ_2-RCL output (3) with $\mu = n^{-2/3}$, it holds that $\mathbb{E}[\mathcal{R}(\mathbf{w}^{(2)})] - \mathcal{R}(\mathbf{w}^*) = \mathcal{O}\left(n^{-2/3}\right)$; moreover,*

$$\mathbb{E}[\mathcal{F}(\mathbf{w}^{(2)}, \mathbf{w}^{(1)})] = \mathcal{O}(n^{-2/3}), \quad \mathbb{E}[\mathcal{I}(\mathbf{w}^{(2)}, \mathbf{w}^{(1)})] = \mathcal{O}(n^{-2/3}).$$

In Example 2, we see that OCL suffers from a constant CL excess risk due to catastrophic forgetting, while ℓ_2-RCL achieves an $o(1)$ CL excess risk by balancing forgetting and intransigence. This example illustrates that ℓ_2-RCL is able to (partially) mitigate the issue of catastrophic forgetting by introducing intransigence.

When is ℓ_2-RCL not able to temper catastrophic forgetting? We next present a CL problem where the forgetting and intransigence cannot be small simultaneously for ℓ_2-RCL, even with an optimally tuned regularization parameter μ.

Example 3. *Under the same setting as Theorem 1, additionally, assume that*

$$\|\mathbf{w}^*\|_2 \leq 1, \quad \sigma^2 = 1, \quad \mathbf{H}^{(1)} = \mathrm{diag}(1, n^{-2}, 0, 0, \dots), \quad \mathbf{H}^{(2)} = \mathrm{diag}(n^{-1}, n^{-1}, 0, 0, \dots).$$

Then for the ℓ_2-RCL output (3) with any $\mu > 0$, it holds that $\mathbb{E}[\mathcal{R}(\mathbf{w}^{(2)})] - \mathcal{R}(\mathbf{w}^*) = \Omega(1)$. In special cases:

- *if $\mu = \mathcal{O}\left(n^{-1.5}\right)$, then $\mathbb{E}[\mathcal{F}(\mathbf{w}^{(2)}, \mathbf{w}^{(1)})] = \Omega(1)$ and $\mathbb{E}[\mathcal{I}(\mathbf{w}^{(2)}, \mathbf{w}^{(1)})] = \mathcal{O}\left(n^{-1}\right)$;*

- *if $\mu = \Theta\left(n^{-1}\right)$, then $\mathbb{E}[\mathcal{F}(\mathbf{w}^{(2)}, \mathbf{w}^{(1)})] = \Omega(1)$ and $\mathbb{E}[\mathcal{I}(\mathbf{w}^{(2)}, \mathbf{w}^{(1)})] = \Omega(1)$;*

- *if $\mu = \Theta\left(n^{-0.5}\right)$, then $\mathbb{E}[\mathcal{F}(\mathbf{w}^{(2)}, \mathbf{w}^{(1)})] = \mathcal{O}\left(n^{-1}\right)$ and $\mathbb{E}[\mathcal{I}(\mathbf{w}^{(2)}, \mathbf{w}^{(1)})] = \Omega(1)$.*

Example 3 demonstrates the limitation of ℓ_2-RCL: for ℓ_2-RCL on some CL problems, while it is possible to choose a small μ such that the intransigence is small or a large μ such that the forgetting is small, it is not possible to make intransigence and forgetting small simultaneously. We conjecture that such a hard CL problem can be solved with more comprehensive regularization methods that utilize the Hessian information (Kirkpatrick et al., 2017; Kolouri et al., 2020; Li et al., 2021). The limitation of ℓ_2-RCL has also been observed in practical CL problems (Hsu et al., 2018).

Numerical simulations. We design a set of experiments to verify our theory. We construct a set of CL problems where the data covariance matrices $\mathbf{H}^{(1)}$ and $\mathbf{H}^{(2)}$ have different levels of similarity. The design is motivated by Wu et al. (2022). Specifically, we fix $d = 200$. The eigenvalues for $\mathbf{H}^{(1)}$ and $\mathbf{H}^{(2)}$ are given by $\{1/2^{i-1}\}_{i=1}^d$, and we control the similarity between $\mathbf{H}^{(1)}$ and $\mathbf{H}^{(2)}$ by the alignment between their eigenvalues. Mathematically, for two integers $k, l > 0$, we specify a CL problem, $\mathbb{Q}(k, l) = (\mathbf{w}^*, \sigma^2, \mathbf{H}^{(1)}, \mathbf{H}^{(2)})$, by:

$$
\begin{aligned}
\mathbf{H}^{(1)} &= \mathrm{diag}\left(1, \frac{1}{2}, \frac{1}{2^2}, \dots, \frac{1}{2^{k-1}}, \frac{1}{2^k}, \dots, \frac{1}{2^{k+l-1}}, \frac{1}{2^{k+l}}, \dots\right), \\
\mathbf{H}^{(2)} &= \mathrm{diag}\left(\frac{1}{2^{k-1}}, \frac{1}{2^{k-2}}, \dots, 1, \frac{1}{2^{k+l-1}}, , \dots, \frac{1}{2^k}, \frac{1}{2^{k+l}}, \dots\right), \\
\mathbf{w}^* &= \Big(\underbrace{1, \dots, 1}_{k+l+2 \text{ copies}}, \frac{1}{k+l+3}, \frac{1}{k+l+4}, \dots\Big)^\top, \qquad \sigma^2 = 1.
\end{aligned}
\tag{13}
$$

It is clear that ranging from $\mathbb{Q}(5, 0)$ to $\mathbb{Q}(15, 0)$ to $\mathbb{Q}(15, 15)$, $\mathbf{H}^{(1)}$ and $\mathbf{H}^{(2)}$ become more and more dissimilar. We then apply JL, OCL, and ℓ_2-RCL to the three problem instances, and measure their achieved CL excess risks versus the number of samples n. The results are shown in Figure 1. We make the following observations from Figure 1:

- For the problem $\mathbb{Q}(5, 0)$ where the two tasks are very similar, all three methods achieve a vanishing CL excess risk. Moreover, the risk convergence rate of OCL and ℓ_2-RCL matches that of JL, ignoring constant factors. This suggests that CL with similar tasks is easy and can be solved by even OCL.

- For problem $\mathbb{Q}(15, 0)$ where the two tasks are not similar, OCL fails to compete with JL due to catastrophic forgetting. However, ℓ_2-RCL can still match the convergence rate of JL by balancing the forgetting and intransigence.

- For problem $\mathbb{Q}(15, 15)$ where the two tasks are very dissimilar, there is a sharp gap in terms of the convergence rate between JL and OCL/ℓ_2-RCL. This demonstrates the limitation of ℓ_2-RCL for solving very hard CL problems.

In order to understand the different behaviors of ℓ_2-RCL in problem instances $\mathbb{Q}(15, 0)$ and $\mathbb{Q}(15, 15)$, we further compute the (expected) forgetting and intransigence of ℓ_2-RCL in these two problem instances with a varying regularization parameter (μ). The results are shown in Figure 2. From Figure 2, we see that for both problem instances, ℓ_2-RCL is able to reduce forgetting by using a large μ, or reduce intransigence by using a small μ. However, for $\mathbb{Q}(15, 0)$, there is a μ that simultaneously reduces forgetting and intransigence, and for $\mathbb{Q}(15, 15)$, forgetting and intransigence cannot be reduced simultaneously. This explains why ℓ_2-RCL can solve $\mathbb{Q}(15, 0)$ but cannot solve $\mathbb{Q}(15, 15)$.

4 RELATED WORKS

Continual learning in practice. Our theoretical framework for CL is motivated by the algorithmic research for CL (see, e.g., Parisi et al. (2019)). Among them, the most relevant to us is the *regularization based* CL method (Kirkpatrick et al., 2017; Zenke et al., 2017; Aljundi et al., 2018; Liu et al., 2018; Ritter et al., 2018; Kolouri et al., 2020; Li et al.,
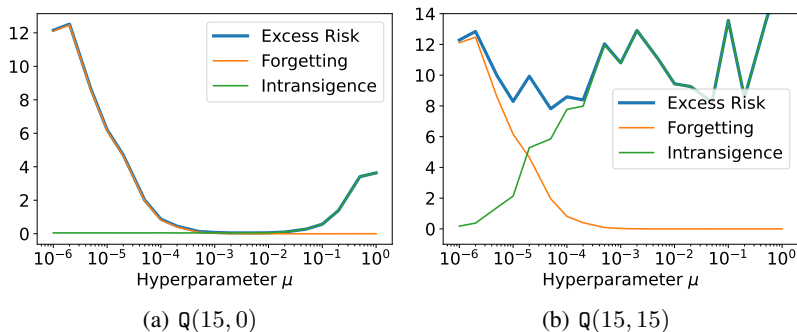
(a) $\mathtt{Q}(15,0)$

(b) $\mathtt{Q}(15,15)$

Figure 2: The trade-off between forgetting and intransigence for ℓ_2-RCL. For each point in each curve, the x-axis represents the sample size n and the y-axis represents the expected CL excess risk or the expected forgetting or the expected intransigence. The problem instances $\mathtt{Q}(15,0)$ and $\mathtt{Q}(15,15)$ are defined by (13). The dimension of the task is $d = 200$. The sample size is $n = 2,000$. The expectation is computed by taking an empirical average over 20 independent runs.

2021), where a regularization term is adopted to prevent the model from deviating from the old one when learning a new task, at the cost of introducing intransigence. Along this line, the research focus is to find the best regularization for achieving a good balance between forgetting and intransigence. For example, the *Elastic Weight Consolidation* (EWC) algorithm (Kirkpatrick et al., 2017) uses a weighted ℓ_2-regularization, where the weight corresponds to the Hessian diagonal from the last model, and the *Sliced Cramer Preservation* (SCP) algorithm (Kolouri et al., 2020) considers a regularization defined by the sliced Cramer distance instead of an ℓ_2-distance. Despite the advances they have made in practice, a statistical continual learning theory is still lacking. Our work fills this gap by establishing a statistical framework for studying the performance of regularization-based CL algorithms (e.g., ℓ_2-RCL).

In addition to the regularization-based CL method, the memory based (Rebuffi et al., 2017; Chaudhry et al., 2019; Shin et al., 2017; Chaudhry et al., 2018b; Farajtabar et al., 2020; Saha et al., 2021) and the architecture based (Rusu et al., 2016; Xu & Zhu, 2018; Mallya & Lazebnik, 2018; Serra et al., 2018; Yoon et al., 2020) CL methods are also popular in practice. These two categories of CL methods are beyond the focus in this paper, but we believe our problem formulation can be migrated to analyze them as well.

Continual learning in theory. Compared to the development of CL in practice, the theoretical results for CL are less rich (Bennani et al., 2020; Doan et al., 2021; Lee et al., 2021; Chen et al., 2022; Evron et al., 2022; Heckel, 2022; Yin et al., 2020; Lin et al., 2023). We now discuss their relationship to our work in turn.

The work by Evron et al. (2022) is most relevant to us. We both consider a domain-incremental CL problem (Van de Ven & Tolias, 2019) with linear regression tasks. However, there are two notable differences. Firstly, the label is assumed to be a random variable in our setting (see Assumption 1), while the entire dataset is considered to be fixed in their setting. So they only studied the optimization aspects of CL, while we focus on the statistical aspects of CL. Secondly, Evron et al. (2022) only considered the OCL method (2), but we study a more interesting RCL algorithm (3) that reveals a richer theoretical picture of CL, e.g., the regularization parameter can adjust the trade-off between forgetting intransigence (see Theorem 1).

The works by Heckel (2022); Lin et al. (2023) considered a task-incremental CL setting (Van de Ven & Tolias, 2019), where the true model parameters are different across tasks; however, they additionally assumed that the data covariance matrices are the same (an identity matrix) for all tasks. In comparison, we focus on a domain-incremental CL setting (Van de Ven & Tolias, 2019), where the true model parameters are shared for all tasks but the data covariance matrices are different across tasks. So our results are not directly comparable to those in Heckel (2022); Lin et al. (2023). In particular, we emphasize that the forgetting-intransigence trade-off shown in our work has not appeared in their works.

The work by Yin et al. (2020) considered CL with a Hessian-aware regularization. However, their main focus is still the optimization behavior. Their statistical results are obtained by standard uniform convergence analysis, which does not reveal the trade-off between forgetting and intransigence as shown in our work.

The remaining papers are not directly related to ours. Specifically, the works by Bennani et al. (2020); Doan et al. (2021) analyzed the generalization error and the forgetting of the *Orthogonal Gradient Descent* (OGD) method (Fara-

jtabar et al., 2020). Lee et al. (2021) studied CL in a multi-head setting. Finally, Chen et al. (2022) showed a negative, memory lower bound for CL.

5 CONCLUSION

We consider an ℓ_2-regularized continual learning with two linear regression tasks in the fixed design setting. We derive sharp risk bounds for the algorithm. We show a provable trade-off between induced forgetting and intransigence, which can be adjusted by the regularization parameter. We show that forgetting could be catastrophic when the tasks are dissimilar and that an ℓ_2-regularization (partially) mitigates the issue by introducing intransigence. We demonstrate the power and limitation of ℓ_2-regularized continual learning with concrete examples.

REFERENCES

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

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Mehdi Abbana Bennani, Thang Doan, and Masashi Sugiyama. Generalisation guarantees for continual learning with orthogonal gradient descent. *arXiv preprint arXiv:2006.11942*, 2020.

Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 532–547, 2018a.

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

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

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. IEEE, 2022.

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

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

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3762–3773. PMLR, 2020.

Reinhard Heckel. Provable continual learning via sketched jacobian approximations. In *International Conference on Artificial Intelligence and Statistics*, pp. 10448–10470. PMLR, 2022.

Daniel Hsu, Sham M Kakade, and Tong Zhang. Random design analysis of ridge regression. In *Conference on learning theory*, pp. 9–1. JMLR Workshop and Conference Proceedings, 2012.

Yen-Chang Hsu, Yen-Cheng Liu, Anita Ramasamy, and Zsolt Kira. Re-evaluating continual learning scenarios: A categorization and case for strong baselines. *arXiv preprint arXiv:1810.12488*, 2018.

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.

Soheil Kolouri, Nicholas A Ketz, Andrea Soltoggio, and Praveen K Pilly. Sliced cramer synaptic consolidation for preserving deeply learned representations. In *International Conference on Learning Representations*, 2020.

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

Qi Lei, Wei Hu, and Jason Lee. Near-optimal linear regression under distribution shift. In *International Conference on Machine Learning*, pp. 6164–6174. PMLR, 2021.

Haoran Li, Aditya Krishnan, Jingfeng Wu, Soheil Kolouri, Praveen K Pilly, and Vladimir Braverman. Lifelong learning with sketched structural regularization. In *Asian conference on machine learning*, pp. 985–1000. PMLR, 2021.

Sen Lin, Peizhong Ju, Yingbin Liang, and Ness Shroff. Theory on forgetting and generalization of continual learning. *arXiv preprint arXiv:2302.05836*, 2023.

Xialei Liu, Marc Masana, Luis Herranz, Joost Van de Weijer, Antonio M Lopez, and Andrew D Bagdanov. Rotate your networks: Better weight consolidation and less catastrophic forgetting. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 2262–2268. IEEE, 2018.

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

Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 7765–7773, 2018.

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

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Razvan Pascanu, and Hassan Ghasemzadeh. Understanding the role of training regimes in continual learning. *Advances in Neural Information Processing Systems*, 33:7308–7320, 2020.

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

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

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

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. *arXiv preprint arXiv:2103.09762*, 2021.

Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *International Conference on Machine Learning*, pp. 4548–4557. PMLR, 2018.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017.

Gido M Van de Ven and Andreas S Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019.

Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham M Kakade. The power and limitation of pretraining-finetuning for linear regression under covariate shift. *The 36th Conference on Neural Information Processing Systems*, 2022.

Ju Xu and Zhanxing Zhu. Reinforced continual learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Dong Yin, Mehrdad Farajtabar, Ang Li, Nir Levine, and Alex Mott. Optimization and generalization of regularization-based continual learning: a loss approximation viewpoint. *arXiv preprint arXiv:2006.10974*, 2020.

Jaehong Yoon, Saehoon Kim, Eunho Yang, and Sung Ju Hwang. Scalable and order-robust continual learning with additive parameter decomposition. In *Eighth International Conference on Learning Representations, ICLR 2020*. ICLR, 2020.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pp. 3987–3995. PMLR, 2017.

A MISSING PROOFS

A.1 PROOF OF THEOREM 1

Proof of Theorem 1. The proof can be done by direct computation.

Computing $\mathbf{w}^{(1)}$. We first compute $\mathbf{w}^{(1)}$. By definition, we have

$$\begin{aligned}
\mathbf{w}^{(1)} &= \left((\mathbf{X}^{(1)})^\top \mathbf{X}^{(1)}\right)^{-1} (\mathbf{X}^{(1)})^\top \mathbf{y} \\
&= \left((\mathbf{X}^{(1)})^\top \mathbf{X}^{(1)}\right)^{-1} (\mathbf{X}^{(1)})^\top \mathbf{X}^{(1)} \mathbf{w}^* + \left((\mathbf{X}^{(1)})^\top \mathbf{X}^{(1)}\right)^{-1} (\mathbf{X}^{(1)})^\top \boldsymbol{\epsilon}^{(1)} \\
&= \mathbf{I}_{\mathbb{K}} \cdot \mathbf{w}^* + \frac{1}{n} \mathbf{H}^{(1)^{-1}} \mathbf{X}^{(1)^\top} \boldsymbol{\epsilon}^{(1)}.
\end{aligned}$$

Therefore

$$\mathbf{w}^{(1)} - \mathbf{w}^* = -\mathbf{I}_{\mathbb{K}^c} \cdot \mathbf{w}^* + \frac{1}{n} \mathbf{H}^{(1)^{-1}} \mathbf{X}^{(1)^\top} \boldsymbol{\epsilon}^{(1)}.$$

Now noticing that

$$\mathbb{E}\boldsymbol{\epsilon}^{(1)} = 0, \quad \mathbb{E}\boldsymbol{\epsilon}^{(1)} \boldsymbol{\epsilon}^{(1)^\top} = \sigma^2 \mathbf{I},$$

so the covariance of $\mathbf{w}^{(1)} - \mathbf{w}^*$ is

$$\begin{aligned}
\mathbb{E}(\mathbf{w}^{(1)} - \mathbf{w}^*)(\mathbf{w}^{(1)} - \mathbf{w}^*)^\top &= \mathbf{I}_{\mathbb{K}^c} \cdot \mathbf{w}^* \mathbf{w}^{*\top} \cdot \mathbf{I}_{\mathbb{K}^c} + \frac{\sigma^2}{n^2} \mathbf{H}^{(1)^{-1}} \mathbf{X}^{(1)^\top} \mathbf{X}^{(1)} \mathbf{H}^{(1)^{-1}} \\
&= \mathbf{I}_{\mathbb{K}^c} \cdot \mathbf{w}^* \mathbf{w}^{*\top} \cdot \mathbf{I}_{\mathbb{K}^c} + \frac{\sigma^2}{n} \cdot \mathbf{H}^{(1)^{-1}}.
\end{aligned} \tag{14}$$

Computing $\mathbf{w}^{(2)}$. Then we compute $\mathbf{w}^{(2)}$, which is the solution of (3). By the first-order optimality condition, we have

$$\frac{1}{n} \mathbf{X}^{(2)^\top} \left(\mathbf{X}^{(2)} \mathbf{w}^{(2)} - \mathbf{y}^{(2)}\right) + \mu(\mathbf{w}^{(2)} - \mathbf{w}^{(1)}) = 0,$$

which implies

$$\begin{aligned}
\mathbf{w}^{(2)} &= \left(\frac{1}{n} \mathbf{X}^{(2)^\top} \mathbf{X}^{(2)} + \mu \mathbf{I}\right)^{-1} \left(\frac{1}{n} \mathbf{X}^{(2)^\top} \mathbf{y}^{(2)} + \mu \mathbf{w}^{(1)}\right) \\
&= \left(\frac{1}{n} \mathbf{X}^{(2)^\top} \mathbf{X}^{(2)} + \mu \mathbf{I}\right)^{-1} \left(\frac{1}{n} \mathbf{X}^{(2)^\top} \mathbf{X}^{(2)} \mathbf{w}^* + \frac{1}{n} \mathbf{X}^{(2)^\top} \boldsymbol{\epsilon}^{(2)} + \mu \mathbf{w}^{(1)}\right) \\
&= \left(\mathbf{H}^{(2)} + \mu \mathbf{I}\right)^{-1} \left(\mathbf{H}^{(2)} \mathbf{w}^* + \frac{1}{n} \mathbf{X}^{(2)^\top} \boldsymbol{\epsilon}^{(2)} + \mu \mathbf{w}^{(1)}\right).
\end{aligned}$$

Then we have

$$\begin{aligned}
\mathbf{w}^{(2)} - \mathbf{w}^* &= \left(\mathbf{H}^{(2)} + \mu \mathbf{I}\right)^{-1} \left(\mathbf{H}^{(2)} \mathbf{w}^* + \frac{1}{n} \mathbf{X}^{(2)^\top} \boldsymbol{\epsilon}^{(2)} + \mu \mathbf{w}^{(1)} - \left(\mathbf{H}^{(2)} + \mu \mathbf{I}\right) \mathbf{w}^*\right) \\
&= \left(\mathbf{H}^{(2)} + \mu \mathbf{I}\right)^{-1} \left(\mu(\mathbf{w}^{(1)} - \mathbf{w}^*) + \frac{1}{n} \mathbf{X}^{(2)^\top} \boldsymbol{\epsilon}^{(2)}\right).
\end{aligned}$$

Similarly, noticing that

$$\mathbb{E}\boldsymbol{\epsilon}^{(2)} = 0, \quad \mathbb{E}\boldsymbol{\epsilon}^{(2)} \boldsymbol{\epsilon}^{(2)^\top} = \sigma^2 \mathbf{I},$$

so the covariance of $\mathbf{w}^{(2)} - \mathbf{w}^*$ is

$$\begin{aligned}
\mathbb{E}(\mathbf{w}^{(2)} - \mathbf{w}^*)(\mathbf{w}^{(2)} - \mathbf{w}^*)^\top &= \mu^2 \cdot \left(\mathbf{H}^{(2)} + \mu \mathbf{I}\right)^{-1} \cdot \mathbb{E}(\mathbf{w}^{(1)} - \mathbf{w}^*)(\mathbf{w}^{(1)} - \mathbf{w}^*)^\top \cdot \left(\mathbf{H}^{(2)} + \mu \mathbf{I}\right)^{-1} \\
&\quad + \frac{\sigma^2}{n^2} \left(\mathbf{H}^{(2)} + \mu \mathbf{I}\right)^{-1} \cdot \mathbf{X}^{(2)^\top} \mathbf{X}^{(2)} \cdot \left(\mathbf{H}^{(2)} + \mu \mathbf{I}\right)^{-1} \\
&= \mu^2 \cdot \left(\mathbf{H}^{(2)} + \mu \mathbf{I}\right)^{-1} \cdot \mathbb{E}(\mathbf{w}^{(1)} - \mathbf{w}^*)(\mathbf{w}^{(1)} - \mathbf{w}^*)^\top \cdot \left(\mathbf{H}^{(2)} + \mu \mathbf{I}\right)^{-1} \\
&\quad + \frac{\sigma^2}{n} \left(\mathbf{H}^{(2)} + \mu \mathbf{I}\right)^{-2} \mathbf{H}^{(2)}.
\end{aligned} \tag{15}$$

13

Risk decomposition. According to the risk definition and the assumptions on the noise, we have

$$
\begin{aligned}
\mathcal{R}_1(\mathbf{w}) &= \frac{1}{n}\mathbb{E}\big\|\mathbf{X}^{(1)}\mathbf{w} - \mathbf{y}^{(1)}\big\|_2^2 \\
&= \frac{1}{n}\mathbb{E}\big\|\mathbf{X}^{(1)}\mathbf{w} - \mathbf{X}^{(1)}\mathbf{w}^* - \boldsymbol{\epsilon}^{(1)}\big\|_2^2 \\
&= (\mathbf{w} - \mathbf{w}^*)^\top \mathbf{H}^{(1)}(\mathbf{w} - \mathbf{w}^*) + \sigma^2 \\
&= \langle \mathbf{H}^{(1)},\ (\mathbf{w} - \mathbf{w}^*)(\mathbf{w} - \mathbf{w}^*)^\top \rangle + \sigma^2.
\end{aligned}
\tag{16}
$$

Similarly, the risk for the second task is

$$
\mathcal{R}_2(\mathbf{w}) = \langle \mathbf{H}^{(2)},\ (\mathbf{w} - \mathbf{w}^*)(\mathbf{w} - \mathbf{w}^*)^\top \rangle + \sigma^2.
\tag{17}
$$

Based on (16) and (17), we can compute the forgetting and intransigence as follows:

$$
\begin{aligned}
\mathcal{F} &:= \mathbb{E}[\mathcal{R}_1(\mathbf{w}^{(2)}) - \mathcal{R}_1(\mathbf{w}^{(1)})] \\
&= \langle \mathbf{H}^{(1)},\ \mathbb{E}(\mathbf{w}^{(2)} - \mathbf{w}^*)(\mathbf{w}^{(2)} - \mathbf{w}^*)^\top \rangle - \langle \mathbf{H}^{(1)},\ \mathbb{E}(\mathbf{w}^{(1)} - \mathbf{w}^*)(\mathbf{w}^{(1)} - \mathbf{w}^*)^\top \rangle,
\end{aligned}
\tag{18}
$$

and

$$
\begin{aligned}
\mathcal{I} &:= \mathbb{E}[\mathcal{R}_2(\mathbf{w}^{(2)})] - \min \mathcal{R}_2 + \mathbb{E}[\mathcal{R}_1(\mathbf{w}^{(1)})] - \min \mathcal{R}_1 \\
&= \langle \mathbf{H}^{(2)},\ \mathbb{E}(\mathbf{w}^{(2)} - \mathbf{w}^*)(\mathbf{w}^{(2)} - \mathbf{w}^*)^\top \rangle + \langle \mathbf{H}^{(1)},\ \mathbb{E}(\mathbf{w}^{(1)} - \mathbf{w}^*)(\mathbf{w}^{(1)} - \mathbf{w}^*)^\top \rangle.
\end{aligned}
\tag{19}
$$

Computing forgetting. We now compute forgetting according to (18), (14), and (15). Firstly, let's use (18), (15) and Assumption 4 to obtain

$$
\begin{aligned}
\mathcal{F} &= \langle \mathbf{H}^{(1)},\ \mathbb{E}(\mathbf{w}^{(2)} - \mathbf{w}^*)(\mathbf{w}^{(2)} - \mathbf{w}^*)^\top \rangle - \langle \mathbf{H}^{(1)},\ \mathbb{E}(\mathbf{w}^{(1)} - \mathbf{w}^*)(\mathbf{w}^{(1)} - \mathbf{w}^*)^\top \rangle \\
&= \langle \mu^2 \cdot \big(\mathbf{H}^{(2)} + \mu\mathbf{I}\big)^{-2}\mathbf{H}^{(1)},\ \mathbb{E}(\mathbf{w}^{(1)} - \mathbf{w}^*)(\mathbf{w}^{(1)} - \mathbf{w}^*)^\top \rangle + \frac{\sigma^2}{n}\langle \mathbf{H}^{(1)},\ \big(\mathbf{H}^{(2)} + \mu\mathbf{I}\big)^{-2}\mathbf{H}^{(2)} \rangle \\
&\quad - \langle \mathbf{H}^{(1)},\ \mathbb{E}(\mathbf{w}^{(1)} - \mathbf{w}^*)(\mathbf{w}^{(1)} - \mathbf{w}^*)^\top \rangle \\
&= \frac{\sigma^2}{n}\langle \mathbf{H}^{(1)},\ \big(\mathbf{H}^{(2)} + \mu\mathbf{I}\big)^{-2}\mathbf{H}^{(2)} \rangle + \langle \mu^2 \cdot \big(\mathbf{H}^{(2)} + \mu\mathbf{I}\big)^{-2}\mathbf{H}^{(1)} - \mathbf{H}^{(1)},\ \mathbb{E}(\mathbf{w}^{(1)} - \mathbf{w}^*)(\mathbf{w}^{(1)} - \mathbf{w}^*)^\top \rangle \\
&= \frac{\sigma^2}{n}\langle \mathbf{H}^{(1)},\ \big(\mathbf{H}^{(2)} + \mu\mathbf{I}\big)^{-2}\mathbf{H}^{(2)} \rangle - \big\langle \big(2\mu\mathbf{H}^{(2)} + \mathbf{H}^{(2)^2}\big) \cdot \big(\mathbf{H}^{(2)} + \mu\mathbf{I}\big)^{-2} \cdot \mathbf{H}^{(1)},\ \mathbb{E}(\mathbf{w}^{(1)} - \mathbf{w}^*)(\mathbf{w}^{(1)} - \mathbf{w}^*)^\top \big\rangle.
\end{aligned}
$$

Next, we bring in (14) to obtain

$$
\begin{aligned}
\mathcal{F} &= \frac{\sigma^2}{n}\langle \mathbf{H}^{(1)},\ \big(\mathbf{H}^{(2)} + \mu\mathbf{I}\big)^{-2}\mathbf{H}^{(2)} \rangle - \frac{\sigma^2}{n}\big\langle \big(2\mu\mathbf{H}^{(2)} + \mathbf{H}^{(2)^2}\big) \cdot \big(\mathbf{H}^{(2)} + \mu\mathbf{I}\big)^{-2} \cdot \mathbf{H}^{(1)},\ \mathbf{H}^{(1)^{-1}} \big\rangle \\
&\quad - \big\langle \big(2\mu\mathbf{H}^{(2)} + \mathbf{H}^{(2)^2}\big) \cdot \big(\mathbf{H}^{(2)} + \mu\mathbf{I}\big)^{-2} \cdot \mathbf{H}^{(1)},\ \mathbf{I}_{\mathbb{K}^c} \cdot \mathbf{w}^*\mathbf{w}^{*\top} \cdot \mathbf{I}_{\mathbb{K}^c} \big\rangle \\
&= \frac{\sigma^2}{n}\langle \mathbf{H}^{(1)},\ \big(\mathbf{H}^{(2)} + \mu\mathbf{I}\big)^{-2}\mathbf{H}^{(2)} \rangle - \frac{\sigma^2}{n}\big\langle \big(2\mu\mathbf{H}^{(2)} + \mathbf{H}^{(2)^2}\big) \cdot \big(\mathbf{H}^{(2)} + \mu\mathbf{I}\big)^{-2} \cdot \mathbf{H}^{(1)},\ \mathbf{H}^{(1)^{-1}} \big\rangle,
\end{aligned}
$$

where the last equality is because $\mathbf{H}^{(1)}\mathbf{I}_{\mathbb{K}^c} = 0$ by definition. We can use Assumption 4 to further simplify $\mathcal{F}$:

$$
\begin{aligned}
\mathcal{F} &= \frac{\sigma^2}{n} \cdot \langle \mathbf{H}^{(1)},\ \big(\mathbf{H}^{(2)} + \mu\mathbf{I}\big)^{-2}\mathbf{H}^{(2)} \rangle - \frac{\sigma^2}{n} \cdot \big\langle \big(2\mu\mathbf{I} + \mathbf{H}^{(2)}\big) \cdot \mathbf{I}_{\mathbb{K}},\ \big(\mathbf{H}^{(2)} + \mu\mathbf{I}\big)^{-2}\mathbf{H}^{(2)} \big\rangle \\
&= \frac{\sigma^2}{n} \cdot \big\langle \mathbf{H}^{(1)} - \big(2\mu\mathbf{I} + \mathbf{H}^{(2)}\big) \cdot \mathbf{I}_{\mathbb{K}},\ \big(\mathbf{H}^{(2)} + \mu\mathbf{I}\big)^{-2}\mathbf{H}^{(2)} \big\rangle \\
&= \frac{\sigma^2}{n} \cdot \big\langle \big(\mathbf{H}^{(1)} - 2\mu\mathbf{I} - \mathbf{H}^{(2)}\big) \cdot \mathbf{I}_{\mathbb{K}},\ \big(\mathbf{H}^{(2)} + \mu\mathbf{I}\big)^{-2}\mathbf{H}^{(2)} \big\rangle.
\end{aligned}
$$

We have completed the proof for the forgetting.

Computing intransigence. We next compute intransigence according to (14), (15) and (19) . We first use (15), (19), and Assumption 4 to obtain:

$$
\mathcal{I} = \langle \mathbf{H}^{(2)},\ \mathbb{E}(\mathbf{w}^{(2)} - \mathbf{w}^*)(\mathbf{w}^{(2)} - \mathbf{w}^*)^\top \rangle + \langle \mathbf{H}^{(1)},\ \mathbb{E}(\mathbf{w}^{(1)} - \mathbf{w}^*)(\mathbf{w}^{(1)} - \mathbf{w}^*)^\top \rangle
$$

$$
\begin{aligned}
&= \langle \mu^2 \cdot \left(\mathbf{H}^{(2)} + \mu \mathbf{I}\right)^{-2} \mathbf{H}^{(2)}, \ \mathbb{E}(\mathbf{w}^{(1)} - \mathbf{w}^*)(\mathbf{w}^{(1)} - \mathbf{w}^*)^\top \rangle + \frac{\sigma^2}{n} \langle \mathbf{H}^{(2)}, \ \left(\mathbf{H}^{(2)} + \mu \mathbf{I}\right)^{-2} \mathbf{H}^{(2)} \rangle \\
&\quad + \langle \mathbf{H}^{(1)}, \ \mathbb{E}(\mathbf{w}^{(1)} - \mathbf{w}^*)(\mathbf{w}^{(1)} - \mathbf{w}^*)^\top \rangle \\
&= \frac{\sigma^2}{n} \cdot \mathrm{tr}\left\{\mathbf{H}^{(2)^2}\left(\mathbf{H}^{(2)} + \mu \mathbf{I}\right)^{-2}\right\} + \langle \mu^2 \cdot \left(\mathbf{H}^{(2)} + \mu \mathbf{I}\right)^{-2} \mathbf{H}^{(2)} + \mathbf{H}^{(1)}, \ \mathbb{E}(\mathbf{w}^{(1)} - \mathbf{w}^*)(\mathbf{w}^{(1)} - \mathbf{w}^*)^\top \rangle.
\end{aligned}
$$

Next, let's insert (14) into the above, then we obtain

$$
\begin{aligned}
\mathcal{I} &= \frac{\sigma^2}{n} \cdot \mathrm{tr}\left\{\mathbf{H}^{(2)^2}\left(\mathbf{H}^{(2)} + \mu \mathbf{I}\right)^{-2}\right\} + \frac{\sigma^2}{n} \cdot \langle \mu^2 \cdot \left(\mathbf{H}^{(2)} + \mu \mathbf{I}\right)^{-2} \mathbf{H}^{(2)} + \mathbf{H}^{(1)}, \ \mathbf{H}^{(1)^{-1}} \rangle \\
&\quad + \langle \mu^2 \cdot \left(\mathbf{H}^{(2)} + \mu \mathbf{I}\right)^{-2} \mathbf{H}^{(2)} + \mathbf{H}^{(1)}, \ \mathbf{I}_{\mathbb{K}^c} \mathbf{w}^* \mathbf{w}^{*\top} \mathbf{I}_{\mathbb{K}^c} \rangle \\
&= \frac{\sigma^2}{n} \cdot \mathrm{tr}\left\{\mathbf{H}^{(2)^2}\left(\mathbf{H}^{(2)} + \mu \mathbf{I}\right)^{-2}\right\} + \frac{\sigma^2}{n} \cdot \mathrm{tr}\left\{\mu^2 \cdot \left(\mathbf{H}^{(2)} + \mu \mathbf{I}\right)^{-2} \mathbf{H}^{(2)} \mathbf{H}^{(1)^{-1}} + \mathbf{I}_{\mathbb{K}}\right\} \\
&\quad + \langle \mu^2 \cdot \left(\mathbf{H}^{(2)} + \mu \mathbf{I}\right)^{-2} \mathbf{H}^{(2)} \mathbf{I}_{\mathbb{K}^c}, \ \mathbf{w}^* \mathbf{w}^{*\top} \rangle \\
&= \frac{\sigma^2}{n} \cdot \left(\mathrm{tr}\left\{\mathbf{H}^{(2)^2}\left(\mathbf{H}^{(2)} + \mu \mathbf{I}\right)^{-2}\right\} + \mathrm{tr}\left\{\mu^2 \cdot \left(\mathbf{H}^{(2)} + \mu \mathbf{I}\right)^{-2} \mathbf{H}^{(2)} \mathbf{H}^{(1)^{-1}}\right\} + \mathrm{rank}\left(\mathbf{H}^{(1)}\right) \right) \\
&\quad + \mathbf{w}^{*\top} \cdot \left(\mathbf{I}_{\mathbb{K}^c} \cdot \mu^2 \mathbf{H}^{(2)}\left(\mathbf{H}^{(2)} + \mu \mathbf{I}\right)^{-2}\right) \mathbf{w}^*.
\end{aligned}
$$

We have completed the proof for the intransigence.

$\qquad\quad \square$

A.2 Proof of Proposition 1

Proof of Proposition 1. Define the joint data by

$$
\mathbf{X} := \begin{pmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{pmatrix} \in \mathbb{R}^{2n \times d}, \quad \mathbf{y} := \begin{pmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \end{pmatrix} \in \mathbb{R}^{2n}.
$$

Define

$$
\mathbf{H} := \mathbf{H}^{(1)} + \mathbf{H}^{(2)}.
$$

Then it holds that

$$
\mathbf{X}^\top \mathbf{X} = \mathbf{H}^{(1)} + \mathbf{H}^{(2)} = \mathbf{H}.
$$

Similarly, define the joint label noise by

$$
\boldsymbol{\epsilon} := \begin{pmatrix} \boldsymbol{\epsilon}^{(1)} \\ \boldsymbol{\epsilon}^{(2)} \end{pmatrix} \in \mathbb{R}^{2n},
$$

then

$$
\mathbb{E}\boldsymbol{\epsilon} = 0, \quad \mathbb{E}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top = \sigma^2 \mathbf{I}.
$$

By the definition of (6), we obtain

$$
\begin{aligned}
\mathbf{w}_{\texttt{joint}} &= \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y} \\
&= \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{X} \mathbf{w}^* + \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \boldsymbol{\epsilon},
\end{aligned}
$$

which implies that

$$
\mathbf{w}_{\texttt{joint}} - \mathbf{w}^* = \left(\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{X} - \mathbf{I}\right) \cdot \mathbf{w}^* + \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \boldsymbol{\epsilon}.
$$

Therefore, it holds that

$$
\begin{aligned}
\mathbb{E}(\mathbf{w}_{\texttt{joint}} - \mathbf{w}^*)(\mathbf{w}_{\texttt{joint}} - \mathbf{w}^*)^\top &= \left(\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{X} - \mathbf{I}\right) \cdot \mathbf{w}^* \mathbf{w}^{*\top} \left(\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{X} - \mathbf{I}\right) + \sigma^2 \cdot \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \\
&= (\mathbf{H}^{-1}\mathbf{H} - \mathbf{I}) \cdot \mathbf{w}^* \mathbf{w}^{*\top} \cdot (\mathbf{H}^{-1}\mathbf{H} - \mathbf{I}) + \frac{\sigma^2}{n} \cdot \mathbf{H}^{-1}. \qquad (20)
\end{aligned}
$$

Moreover, the total risk can be reformulated to

$$
\mathcal{R}(\mathbf{w}) - \min \mathcal{R} = \mathcal{R}_1(\mathbf{w}) - \min \mathcal{R}_1 + \mathcal{R}_2(\mathbf{w}) - \min \mathcal{R}_2
$$

$$= \langle \mathbf{H}^{(1)}, \ (\mathbf{w} - \mathbf{w}^*)(\mathbf{w} - \mathbf{w}^*)^\top \rangle + \langle \mathbf{H}^{(2)}, \ (\mathbf{w} - \mathbf{w}^*)(\mathbf{w} - \mathbf{w}^*)^\top \rangle$$
$$= \langle \mathbf{H}, \ (\mathbf{w} - \mathbf{w}^*)(\mathbf{w} - \mathbf{w}^*)^\top \rangle.$$

Bringing (20) into the above, we obtain

$$\mathbb{E}\mathcal{R}(\mathbf{w}_{\texttt{joint}}) - \min \mathcal{R} = \langle \mathbf{H}, \ \mathbb{E}(\mathbf{w}_{\texttt{joint}} - \mathbf{w}^*)(\mathbf{w}_{\texttt{joint}} - \mathbf{w}^*)^\top \rangle$$
$$= \langle \mathbf{H}, \ .(\mathbf{H}^{-1}\mathbf{H} - \mathbf{I}) \cdot \mathbf{w}^*\mathbf{w}^{*\top} \cdot (\mathbf{H}^{-1}\mathbf{H} - \mathbf{I}) \rangle + \frac{\sigma^2}{n} \cdot \langle \mathbf{H}, \mathbf{H}^{-1} \rangle$$
$$= \frac{\sigma^2}{n} \cdot \mathrm{rank}(\mathbf{H})$$
$$= \frac{\sigma^2}{n} \cdot \mathrm{rank}(\mathbf{H}^{(1)} + \mathbf{H}^{(2)}).$$

We have finished the proof. $\qquad \square$

A.3 PROOF OF EXAMPLES

Proof of Example 1. We examine $\mathbb{E}\left[\mathcal{R}(\mathbf{w}^{(2)}) - \mathcal{R}(\mathbf{w}^*)\right]$ by separately examine the order of $\mathcal{F}$ and $\mathcal{I}$.

1. By Theorem 1 we have

$$\mathcal{F} = \frac{\sigma^2}{n}\Big(\mathrm{tr}\big\{\mathbf{H}^{(1)}\mathbf{H}^{(2)^{-1}}\big\} - \mathrm{rank}\big\{\mathbf{H}^{(1)}\mathbf{H}^{(2)}\big\}\Big)$$
$$= \frac{\sigma^2}{n}(\mathcal{O}\left(n^{1-r}\right) - \mathcal{O}\left(1\right)) = \mathcal{O}\left(n^{-r}\right),$$
$$\mathcal{I} = \frac{\sigma^2}{n}(\mathrm{rank}\big\{\mathbf{H}^{(2)} + \mathrm{rank}\big\{\mathbf{H}^{(2)} = \mathcal{O}\left(n^{-1}\right)\big\}\big\})$$

As a result, $\mathbb{E}\left[\mathcal{R}(\mathbf{w}^{(2)}) - \mathcal{R}(\mathbf{w}^*)\right] = \mathcal{O}\left(n^{-r}\right).$

2. By Theorem 1, we have

$$\mathcal{F} = \frac{\sigma^2}{n}\Big(\mathrm{tr}\big\{\mathbf{H}^{(1)}\mathbf{H}^{(2)^{-1}}\big\} - \mathrm{rank}\big\{\mathbf{H}^{(1)}\mathbf{H}^{(2)}\big\}\Big)$$
$$= \frac{\sigma^2}{n}(\Omega\left(n\right) - \mathcal{O}\left(1\right)) = \Omega\left(1\right),$$
$$\mathcal{I} = \frac{\sigma^2}{n}(\mathrm{rank}\big\{\mathbf{H}^{(2)} + \mathrm{rank}\big\{\mathbf{H}^{(2)} = \mathcal{O}\left(n^{-1}\right)\big\}\big\})$$

As a result, $\mathbb{E}\left[\mathcal{R}(\mathbf{w}^{(2)}) - \mathcal{R}(\mathbf{w}^*)\right] = \Omega\left(1\right).$

$\qquad \square$

Proof of Example 2. We examine $\mathbb{E}\left[\mathcal{R}(\mathbf{w}^{(2)}) - \mathcal{R}(\mathbf{w}^*)\right]$ by separately examine the order of $\mathcal{F}$ and $\mathcal{I}$.

1. **OCL:** This is a special case of the second example in Corollary 1.

2. ℓ_2-**RCL:** By Theorem 1 we have

$$\mathcal{F} = \frac{\sigma^2}{n} \cdot \mathtt{d_F}, \qquad \mathcal{I} = \mathtt{bias} + \frac{\sigma^2}{n} \cdot \mathtt{d_I},$$

where

$$\mathtt{d_F} \le \mathrm{tr}\left\{\mathbf{H}^{(1)} \cdot \mathbf{H}^{(2)} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-2}\right\}$$
$$= \mathcal{O}\left(n^{1/3}\right) + \mathcal{O}\left(1\right) = \mathcal{O}\left(n^{1/3}\right),$$
$$\mathtt{d_I} \le \mathrm{tr}\left\{\mu^2 \cdot \mathbf{H}^{(1)^{-1}} \cdot \mathbf{H}^{(2)} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-2}\right\} + \mathrm{rank}\big\{\mathbf{H}^{(2)}\big\} + \mathrm{rank}\big\{\mathbf{H}^{(1)}\big\}$$
$$= \mathcal{O}\left(1\right) + \mathcal{O}\left(1\right) + \mathcal{O}\left(1\right) = \mathcal{O}\left(1\right),$$
$$\mathtt{bias} \le \mu \cdot \|\mathbf{w}^*\|_2^2 = \mathcal{O}\left(n^{-2/3}\right).$$

As a result, $\mathcal{F} = \mathcal{O}\left(n^{-2/3}\right)$, $\mathcal{I} = \mathcal{O}\left(n^{-2/3}\right)$, and $\mathbb{E}\left[\mathcal{R}(\mathbf{w}^{(2)}) - \mathcal{R}(\mathbf{w}^*)\right] = \mathcal{O}\left(n^{-2/3}\right)$.

$\square$

Proof of Example 3. Similar to the above, we separately examine the order of $\mathcal{F}$ and $\mathcal{I}$. By Theorem 1, the forgetting satisfies

$$\mathcal{F} = \frac{2\sigma^2}{n}\frac{\mu^2}{(1/n+\mu)^2} + \frac{\sigma^2}{n}(n+n^{-1})\frac{(1/n)^2}{(1/n+\mu)^2} - \frac{\sigma^2}{n}\mathrm{rank}(\mathbf{H}^{(1)})$$

so $\mathcal{F} = \Omega\left(1\right)$ for $\mu = \mathcal{O}\left(n^{-1}\right)$, and $\mathcal{F} = \mathcal{O}\left(n^{-1}\right)$ for $\mu = \Theta\left(n^{-0.5}\right)$; and the intransigence satisfies

$$\mathcal{I} \leq \mu\|\mathbf{w}^*\|_2^2 + \frac{\sigma^2}{n}(n+n^{-1})\frac{\mu^2}{(1/n+\mu)^2} + \frac{2\sigma^2}{n}\frac{(1/n)^2}{(1/n+\mu)^2} + \frac{\sigma^2}{n}\mathrm{rank}(\mathbf{H}^{(1)})$$

$$\mathcal{I} \geq \frac{\sigma^2}{n}(n+n^{-1})\frac{\mu^2}{(1/n+\mu)^2} + \frac{2\sigma^2}{n}\frac{(1/n)^2}{(1/n+\mu)^2} + \frac{\sigma^2}{n}\mathrm{rank}(\mathbf{H}^{(1)})$$

so $\mathcal{I} = \Omega\left(1\right)$ for $\mu = \Omega\left(n^{-1}\right)$, and $\mathcal{I} = \mathcal{O}\left(n^{-1}\right)$ for $\mu = \Theta\left(n^{-1.5}\right)$.

We note that for every $\mu > 0$,

$$\mathbb{E}\left[\mathcal{R}(\mathbf{w}^{(2)}) - \mathcal{R}(\mathbf{w}^*)\right] = \mathcal{F} + \mathcal{I}$$

$$\geq \frac{\sigma^2}{n}(n+n^{-1})\frac{\mu^2}{(1/n+\mu)^2} + \frac{\sigma^2}{n}(n+n^{-1})\frac{(1/n)^2}{(1/n+\mu)^2}$$

$$\geq \sigma^2\left(\frac{\mu^2}{(1/n+\mu)^2} + \frac{(1/n)^2}{(1/n+\mu)^2}\right) = \Omega\left(1\right).$$

$\square$

B   EXTENSION OF SECTION 3.3

The main paper discusses the continual learning implications when there exist certain assumptions, including Assumptions 2 and 4. In this section we would like to discuss the case without these assumptions.

B.1   RISK BOUNDS FOR CONTINUAL LEARNING WITHOUT ASSUMPTION 2

Theorem 2 (A risk bound for ℓ_2-RCL/OCL without Assumption 2)**.** *Suppose that Assumptions 1 , 3 and 4 hold. Denote $\mathbf{w}^{(1)^*}, \mathbf{w}^{(2)^*} \in \mathbb{R}^d$ such that*

$$\mathbf{w}^{(1)^*} \in \arg\min \mathcal{R}_1(\mathbf{w}), \quad \mathbf{w}^{(2)^*} \in \arg\min \mathcal{R}_2(\mathbf{w}).$$

Then for the ℓ_2-RCL output (3), it holds that

$$\mathbb{E}\left[\mathcal{R}\left(\mathbf{w}^{(2)}\right)\right] - \min\mathcal{R}(\cdot) = \mathbb{E}\left[\mathcal{F}\left(\mathbf{w}^{(2)}, \mathbf{w}^{(1)}\right)\right] + \mathbb{E}\left[\mathcal{I}\left(\mathbf{w}^{(2)}, \mathbf{w}^{(1)}\right)\right].$$

Moreover, it holds that

$$\mathbb{E}\left[\mathcal{F}\left(\mathbf{w}^{(2)}, \mathbf{w}^{(1)}\right)\right] = \mathtt{bias_F} + \frac{\sigma^2}{n} \cdot \mathtt{d_F},$$

$$\mathbb{E}\left[\mathcal{I}\left(\mathbf{w}^{(2)}, \mathbf{w}^{(1)}\right)\right] = \mathtt{bias_I} + \frac{\sigma^2}{n} \cdot \mathtt{d_I},$$

where

$$\mathtt{bias_F} := (\mathbf{w}^{(1)^*} - \mathbf{w}^{(2)^*})^\top \cdot \mathbf{H}^{(1)} \cdot \mathbf{H}^{(2)^2} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-2} \cdot (\mathbf{w}^{(1)^*} - \mathbf{w}^{(2)^*}),$$

$$\mathtt{bias_I} := (\mathbf{I}_\mathbb{K}\mathbf{w}^{(1)^*} - \mathbf{w}^{(2)^*})^\top \cdot \mu^2 \cdot \mathbf{H}^{(2)} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-2} \cdot (\mathbf{I}_\mathbb{K}\mathbf{w}^{(1)^*} - \mathbf{w}^{(2)^*}),$$

(21)

for index sets $\mathbb{K} := \left\{i : \lambda_i^{(1)} > 0\right\}$, and

$$\mathtt{d_F} := \mathrm{tr}\left\{\mathbf{I}_\mathbb{K} \cdot \left(\mathbf{H}^{(1)} - \mathbf{H}^{(2)} - 2\mu\mathbf{I}\right) \cdot \mathbf{H}^{(2)} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-2}\right\},$$

$$\mathtt{d_I} := \mathrm{tr}\left\{\left(\mu^2 \cdot {\mathbf{H}^{(1)}}^{-1} + \mathbf{H}^{(2)}\right) \cdot \mathbf{H}^{(2)} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-2}\right\} + \mathrm{rank}\{\mathbf{H}^{(1)}\},$$

(22)

are the same as in Theorem 1.

This shows that Assumption 2 only has an effect on the bias term of the bound. The implication of this result is twofold.

Firstly, in terms of the OCL algorithm, to achieve an $o(1)$ CL excess risk, it suffices that

$$\operatorname{tr}\left\{\mathbf{H}^{(1)}\mathbf{H}^{(2)^{-1}}\right\} \le o(n), \quad (\mathbf{w}^{(1)^*} - \mathbf{w}^{(2)^*})^\top \cdot \mathbf{H}^{(1)} \cdot \mathbf{H}^{(2)^2} \cdot \left(\mathbf{H}^{(2)}\right)^{-2} \cdot (\mathbf{w}^{(1)^*} - \mathbf{w}^{(2)^*}) = o(1). \quad (23)$$

This can count as another similarity measurement of OCL in the non-shared-optimal case, and is tougher than the condition in (10). A necessary condition for this to happen is that $\mathbf{w}^{(1)^*} - \mathbf{w}^{(2)^*}$ has $o(1)$ entries for all indices i that satisfy $\lambda_i^{(1)} \ge c, \lambda_i^{(2)} > 0$ for a constant c, which is not the usual case.

Secondly we explain the effect on the ℓ_2-RCL algorithm. According to (21) and (22), the regularizer coefficient μ controls a trade-off between forgetting and intransigence, in similar terms to the non-shared-optimal case that increasing μ reduces forgetting ($\mathtt{bias_F}$ and $\mathtt{d_F}$) while produces intransigence ($\mathtt{bias_I}$ and $\mathtt{d_I}$), and vice versa. Moreover, comparing to the OCL algorithm and the scenario with shared optimal, for the ℓ_2-RCL to achieve an $o(1)$ CL excess risk without Assumption 2, it suffices that $\mathtt{bias_F}, \mathtt{bias_I} = o(1)$ and $\mathtt{d_F}, \mathtt{d_I} = o(n)$. Consider the upper bound of $\mathtt{d_F}, \mathtt{d_I}$ in (11), it suffices to require that

$$\operatorname{tr}\left\{\mathbf{H}^{(1)} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1}\right\} \le o(n), \quad \mu \cdot \operatorname{tr}\left\{\mathbf{H}^{(1)^{-1}}\right\} \le o(n), \quad \operatorname{rank}(\mathbf{H}^{(1)}) + \operatorname{rank}(\mathbf{H}^{(2)}) \le o(n),$$

$$(\mathbf{w}^{(1)^*} - \mathbf{w}^{(2)^*})^\top \cdot \mathbf{H}^{(1)} \cdot \mathbf{H}^{(2)^2} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-2} \cdot (\mathbf{w}^{(1)^*} - \mathbf{w}^{(2)^*}) \le o(1), \quad (24)$$

$$(\mathbf{I}_\mathbb{K}\mathbf{w}^{(1)^*} - \mathbf{w}^{(2)^*})^\top \cdot \mu^2 \cdot \mathbf{H}^{(2)} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-2} \cdot (\mathbf{I}_\mathbb{K}\mathbf{w}^{(1)^*} - \mathbf{w}^{(2)^*})^\top \le o(1),$$

for some $\mu > 0$. The OCL condition in (23) is a special case of this condition when $\mathbf{w}^{(1)^*} = \mathbf{w}^{(2)^*}$.

B.2 Risk Bounds For Continual Learning without Assumption 4

Theorem 3 (A risk bound for ℓ_2-RCL/OCL without Assumption 4). *Suppose that Assumptions 1 to 3 hold. Then for the ℓ_2-RCL output* (3)*, it holds that*

$$\mathbb{E}\left[\mathcal{R}\left(\mathbf{w}^{(2)}\right)\right] - \min \mathcal{R}(\cdot) = \mathbb{E}\left[\mathcal{F}\left(\mathbf{w}^{(2)}, \mathbf{w}^{(1)}\right)\right] + \mathbb{E}\left[\mathcal{I}\left(\mathbf{w}^{(2)}, \mathbf{w}^{(1)}\right)\right].$$

Moreover, it holds that

$$\mathbb{E}\left[\mathcal{F}\left(\mathbf{w}^{(2)}, \mathbf{w}^{(1)}\right)\right] = \mathtt{bias_F} + \frac{\sigma^2}{n} \cdot \mathtt{d_F},$$

$$\mathbb{E}\left[\mathcal{I}\left(\mathbf{w}^{(2)}, \mathbf{w}^{(1)}\right)\right] = \mathtt{bias_I} + \frac{\sigma^2}{n} \cdot \mathtt{d_I},$$

where

$$\mathtt{d_F} := \operatorname{tr}\left\{\left(\mu^2 \cdot \mathbf{H}^{(1)^{-1}} + \mathbf{H}^{(2)}\right) \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1} \cdot \mathbf{H}^{(1)} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1}\right\} - \operatorname{rank}\left\{\mathbf{H}^{(1)}\right\},$$

$$\mathtt{d_I} := \operatorname{tr}\left\{\left(\mu^2 \cdot \mathbf{H}^{(1)^{-1}} + \mathbf{H}^{(2)}\right) \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1} \cdot \mathbf{H}^{(2)} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1}\right\} + \operatorname{rank}\left\{\mathbf{H}^{(1)}\right\},$$

$$\mathtt{bias_F} := (\mathbf{w}^*)^\top \cdot \mathbf{I}_{\mathbb{K}^c} \cdot \mu^2 \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1} \cdot \mathbf{H}^{(1)} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1} \cdot \mathbf{w}^*, \quad (25)$$

$$\mathtt{bias_I} := (\mathbf{w}^*)^\top \cdot \mathbf{I}_{\mathbb{K}^c} \cdot \mu^2 \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1} \cdot \mathbf{H}^{(2)} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1} \cdot \mathbf{w}^*,$$

for index sets $\mathbb{K} := \left\{i : \lambda_i^{(1)} > 0\right\}$.

This shows that by dropping Assumption 4, to achieve an $o(1)$ CL excess risk, it suffices that

$$\operatorname{tr}\left\{\left(\mu^2 \cdot \mathbf{H}^{(1)^{-1}} + \mathbf{H}^{(2)}\right) \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1} \cdot \left(\mathbf{H}^{(1)} + \mathbf{H}^{(2)}\right) \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1}\right\} \le o(n),$$

$$(\mathbf{w}^*)^\top \cdot \mathbf{I}_{\mathbb{K}^c} \cdot \mu^2 \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1} \cdot \mathbf{H}^{(1)} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1} \cdot \mathbf{w}^* \le o(1), \quad \mu \le o(1). \quad (26)$$

We note that (26) and Theorem 3 are extensions of the main result in (11) and Theorem 1.

B.3 Proof of Theorem 2

Proof of Theorem 2. The proof can be done by direct computation and is similar to the proof of Theorem 1.

Risk decomposition. According to the risk definition and the assumptions on the noise, we have

$$
\begin{aligned}
\mathcal{R}_1(\mathbf{w}) &= \frac{1}{n}\mathbb{E}\big\|\mathbf{X}^{(1)}\mathbf{w} - \mathbf{y}^{(1)}\big\|_2^2 \\
&= \frac{1}{n}\mathbb{E}\big\|\mathbf{X}^{(1)}\mathbf{w} - \mathbf{X}^{(1)}\mathbf{w}^* - \boldsymbol{\epsilon}^{(1)}\big\|_2^2 \\
&= (\mathbf{w} - \mathbf{w}^*)^\top \mathbf{H}^{(1)}(\mathbf{w} - \mathbf{w}^{(1)^*}) + \sigma^2 \\
&= \langle \mathbf{H}^{(1)},\ (\mathbf{w} - \mathbf{w}^{(1)^*})(\mathbf{w} - \mathbf{w}^{(1)^*})^\top \rangle + \sigma^2.
\end{aligned}
\tag{27}
$$

Similarly, the risk for the second task is

$$
\mathcal{R}_2(\mathbf{w}) = \langle \mathbf{H}^{(2)},\ (\mathbf{w} - \mathbf{w}^{(2)^*})(\mathbf{w} - \mathbf{w}^{(1)^*})^\top \rangle + \sigma^2.
\tag{28}
$$

Based on (27) and (28), we can compute the forgetting and intransigence as follows:

$$
\begin{aligned}
\mathcal{F} &:= \mathbb{E}[\mathcal{R}_1(\mathbf{w}^{(2)}) - \mathcal{R}_1(\mathbf{w}^{(1)})] \\
&= \langle \mathbf{H}^{(1)},\ \mathbb{E}(\mathbf{w}^{(2)} - \mathbf{w}^{(1)^*})(\mathbf{w}^{(2)} - \mathbf{w}^{(1)^*})^\top \rangle - \langle \mathbf{H}^{(1)},\ \mathbb{E}(\mathbf{w}^{(1)} - \mathbf{w}^{(1)^*})(\mathbf{w}^{(1)} - \mathbf{w}^{(1)^*})^\top \rangle,
\end{aligned}
\tag{29}
$$

and

$$
\begin{aligned}
\mathcal{I} &:= \mathbb{E}[\mathcal{R}_2(\mathbf{w}^{(2)})] - \min \mathcal{R}_2 + \mathbb{E}[\mathcal{R}_1(\mathbf{w}^{(1)})] - \min \mathcal{R}_1 \\
&= \langle \mathbf{H}^{(2)},\ \mathbb{E}(\mathbf{w}^{(2)} - \mathbf{w}^{(2)^*})(\mathbf{w}^{(2)} - \mathbf{w}^{(2)^*})^\top \rangle + \langle \mathbf{H}^{(1)},\ \mathbb{E}(\mathbf{w}^{(1)} - \mathbf{w}^{(1)^*})(\mathbf{w}^{(1)} - \mathbf{w}^{(1)^*})^\top \rangle.
\end{aligned}
\tag{30}
$$

Computing $\mathbf{w}^{(1)}$. We first compute $\mathbf{w}^{(1)}$. By definition, we have

$$
\begin{aligned}
\mathbf{w}^{(1)} &= \big((\mathbf{X}^{(1)})^\top \mathbf{X}^{(1)}\big)^{-1}(\mathbf{X}^{(1)})^\top \mathbf{y} \\
&= \big((\mathbf{X}^{(1)})^\top \mathbf{X}^{(1)}\big)^{-1}(\mathbf{X}^{(1)})^\top \mathbf{X}^{(1)}\mathbf{w}^{(1)^*} + \big((\mathbf{X}^{(1)})^\top \mathbf{X}^{(1)}\big)^{-1}(\mathbf{X}^{(1)})^\top \boldsymbol{\epsilon}^{(1)} \\
&= \mathbf{I}_{\mathbb{K}} \cdot \mathbf{w}^* + \frac{1}{n}\mathbf{H}^{(1)^{-1}}\mathbf{X}^{(1)^\top}\boldsymbol{\epsilon}^{(1)}.
\end{aligned}
$$

Therefore

$$
\mathbf{w}^{(1)} - \mathbf{w}^{(1)^*} = -\mathbf{I}_{\mathbb{K}^c}\cdot \mathbf{w}^{(1)^*} + \frac{1}{n}\mathbf{H}^{(1)^{-1}}\mathbf{X}^{(1)^\top}\boldsymbol{\epsilon}^{(1)}.
$$

Now noticing that

$$
\mathbb{E}\boldsymbol{\epsilon}^{(1)} = 0, \quad \mathbb{E}\boldsymbol{\epsilon}^{(1)}\boldsymbol{\epsilon}^{(1)^\top} = \sigma^2 \mathbf{I},
$$

so the covariance of $\mathbf{w}^{(1)} - \mathbf{w}^{(1)^*}$ is

$$
\begin{aligned}
\mathbb{E}(\mathbf{w}^{(1)} - \mathbf{w}^{(1)^*})(\mathbf{w}^{(1)} - \mathbf{w}^{(1)^*})^\top &= \mathbf{I}_{\mathbb{K}^c}\cdot \mathbf{w}^{(1)^*}\mathbf{w}^{(1)^{*\top}}\cdot \mathbf{I}_{\mathbb{K}^c} + \frac{\sigma^2}{n^2}\mathbf{H}^{(1)^{-1}}\mathbf{X}^{(1)^\top}\mathbf{X}^{(1)}\mathbf{H}^{(1)^{-1}} \\
&= \mathbf{I}_{\mathbb{K}^c}\cdot \mathbf{w}^{(1)^*}\mathbf{w}^{(1)^{*\top}}\cdot \mathbf{I}_{\mathbb{K}^c} + \frac{\sigma^2}{n}\cdot \mathbf{H}^{(1)^{-1}}.
\end{aligned}
\tag{31}
$$

and that

$$
\langle \mathbf{H}^{(1)},\ \mathbb{E}(\mathbf{w}^{(1)} - \mathbf{w}^{(1)^*})(\mathbf{w}^{(1)} - \mathbf{w}^{(1)^*})^\top \rangle = \frac{\sigma^2}{n}\mathrm{rank}(\mathbf{H}^{(1)})
\tag{32}
$$

since by definition $\mathbf{H}^{(1)}\mathbf{I}_{\mathbb{K}^c} = 0$.

Computing $\mathbf{w}^{(2)}$. Then we compute $\mathbf{w}^{(2)}$, which is the solution of (3). By the first-order optimality condition, we have

$$
\frac{1}{n}\mathbf{X}^{(2)^\top}\big(\mathbf{X}^{(2)}\mathbf{w}^{(2)} - \mathbf{y}^{(2)}\big) + \mu(\mathbf{w}^{(2)} - \mathbf{w}^{(1)}) = 0,
$$

which implies

$$
\mathbf{w}^{(2)} = \left(\frac{1}{n}\mathbf{X}^{(2)^\top}\mathbf{X}^{(2)} + \mu\mathbf{I}\right)^{-1}\left(\frac{1}{n}\mathbf{X}^{(2)^\top}\mathbf{y}^{(2)} + \mu\mathbf{w}^{(1)}\right)
$$

$$= \left(\frac{1}{n}\mathbf{X}^{(2)\top}\mathbf{X}^{(2)} + \mu\mathbf{I}\right)^{-1}\left(\frac{1}{n}\mathbf{X}^{(2)\top}\mathbf{X}^{(2)}\mathbf{w}^{(2)*} + \frac{1}{n}\mathbf{X}^{(2)\top}\boldsymbol{\epsilon}^{(2)} + \mu\mathbf{w}^{(1)}\right)$$

$$= \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1}\left(\mathbf{H}^{(2)}\mathbf{w}^{(2)*} + \frac{1}{n}\mathbf{X}^{(2)\top}\boldsymbol{\epsilon}^{(2)} + \mu\mathbf{w}^{(1)}\right).$$

Then we have

$$\mathbf{w}^{(2)} - \mathbf{w}^{(2)*} = \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1}\left(\mathbf{H}^{(2)}\mathbf{w}^{(2)*} + \frac{1}{n}\mathbf{X}^{(2)\top}\boldsymbol{\epsilon}^{(2)} + \mu\mathbf{w}^{(1)} - \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)\mathbf{w}^{(2)*}\right)$$

$$= \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1}\left(\mu(\mathbf{w}^{(1)} - \mathbf{w}^{(2)*}) + \frac{1}{n}\mathbf{X}^{(2)\top}\boldsymbol{\epsilon}^{(2)}\right)$$

$$= \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1}\left(\mu(\mathbf{I}_{\mathbb{K}} \cdot \mathbf{w}^{(1)*} - \mathbf{w}^{(2)*}) + \frac{1}{n}\mu\mathbf{H}^{(1)-1}\mathbf{X}^{(1)\top}\boldsymbol{\epsilon}^{(1)} + \frac{1}{n}\mathbf{X}^{(2)\top}\boldsymbol{\epsilon}^{(2)}\right),$$

and

$$\mathbf{w}^{(2)} - \mathbf{w}^{(1)*} = \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1}\left(\mu(\mathbf{w}^{(1)} - \mathbf{w}^{(2)*}) - \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)(\mathbf{w}^{(1)*} - \mathbf{w}^{(2)*}) + \frac{1}{n}\mathbf{X}^{(2)\top}\boldsymbol{\epsilon}^{(2)}\right)$$

$$= \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1}\left(\mu(\mathbf{w}^{(1)} - \mathbf{w}^{(1)*}) - \mathbf{H}^{(2)}(\mathbf{w}^{(1)*} - \mathbf{w}^{(2)*}) + \frac{1}{n}\mathbf{X}^{(2)\top}\boldsymbol{\epsilon}^{(2)}\right)$$

$$= \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1}\left(-\mu\mathbf{I}_{\mathbb{K}^c} \cdot \mathbf{w}^{(1)*} - \mathbf{H}^{(2)}(\mathbf{w}^{(1)*} - \mathbf{w}^{(2)*}) + \frac{1}{n}\mu\mathbf{H}^{(1)-1}\mathbf{X}^{(1)\top}\boldsymbol{\epsilon}^{(1)} + \frac{1}{n}\mathbf{X}^{(2)\top}\boldsymbol{\epsilon}^{(2)}\right).$$

Similarly, noticing that

$$\mathbb{E}\boldsymbol{\epsilon}^{(1)} = 0, \quad \mathbb{E}\boldsymbol{\epsilon}^{(1)}\boldsymbol{\epsilon}^{(1)\top} = \sigma^2\mathbf{I}, \quad \mathbb{E}\boldsymbol{\epsilon}^{(2)} = 0, \quad \mathbb{E}\boldsymbol{\epsilon}^{(2)}\boldsymbol{\epsilon}^{(2)\top} = \sigma^2\mathbf{I},$$

so

$$\langle\mathbf{H}^{(1)}, \mathbb{E}(\mathbf{w}^{(2)} - \mathbf{w}^{(1)*})(\mathbf{w}^{(2)} - \mathbf{w}^{(1)*})^\top\rangle = (\mathbf{w}^{(1)*} - \mathbf{w}^{(2)*})^\top \cdot \mathbf{H}^{(1)} \cdot \mathbf{H}^{(2)2} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-2} \cdot (\mathbf{w}^{(1)*} - \mathbf{w}^{(2)*})$$

$$+ \frac{\sigma^2}{n}\mu^2\langle\mathbf{H}^{(2)}, \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1} \cdot \mathbf{H}^{(1)-1} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1}\rangle$$

$$+ \frac{\sigma^2}{n}\langle\mathbf{H}^{(2)}, \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1} \cdot \mathbf{H}^{(2)} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1}\rangle$$

$$= \texttt{bias}_\texttt{F} + \frac{\sigma^2}{n}\left(\texttt{d}_\texttt{F} + \mathrm{rank}(\mathbf{H}^{(1)})\right), \tag{33}$$

where the first line is because by definition $\mathbf{H}^{(1)}\mathbf{I}_{\mathbb{K}^c} = 0$, and

$$\langle\mathbf{H}^{(2)}, \mathbb{E}(\mathbf{w}^{(2)} - \mathbf{w}^{(2)*})(\mathbf{w}^{(2)} - \mathbf{w}^{(2)*})^\top\rangle = (\mathbf{I}_{\mathbb{K}}\mathbf{w}^{(1)*} - \mathbf{w}^{(2)*})^\top \cdot \mu^2 \cdot \mathbf{H}^{(2)} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-2} \cdot (\mathbf{I}_{\mathbb{K}}\mathbf{w}^{(1)*} - \mathbf{w}^{(2)*})$$

$$+ \frac{\sigma^2}{n}\mu^2\langle\mathbf{H}^{(2)}, \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1} \cdot \mathbf{H}^{(1)-1} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1}\rangle$$

$$+ \frac{\sigma^2}{n}\langle\mathbf{H}^{(2)}, \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1} \cdot \mathbf{H}^{(2)} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1}\rangle$$

$$= \texttt{bias}_\texttt{I} + \frac{\sigma^2}{n}\left(\texttt{d}_\texttt{I} - \mathrm{rank}(\mathbf{H}^{(1)})\right). \tag{34}$$

Combine (32), (33), (34) and we are done with the proof of forgetting and intransigence. □

B.4 PROOF OF THEOREM 3

Proof of Theorem 3. The proof can be done by direct computation and is similar to the proof of Theorem 1.

We can adopt (14), (15), (16), (17) from the proof of Theorem 1 that

$$\mathbb{E}(\mathbf{w}^{(1)} - \mathbf{w}^*)(\mathbf{w}^{(1)} - \mathbf{w}^*)^\top = \mathbf{I}_{\mathbb{K}^c} \cdot \mathbf{w}^*\mathbf{w}^{*\top} \cdot \mathbf{I}_{\mathbb{K}^c} + \frac{\sigma^2}{n} \cdot \mathbf{H}^{(1)-1},$$

$$\mathbb{E}(\mathbf{w}^{(2)} - \mathbf{w}^*)(\mathbf{w}^{(2)} - \mathbf{w}^*)^\top = \mu^2 \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1} \cdot \mathbb{E}(\mathbf{w}^{(1)} - \mathbf{w}^*)(\mathbf{w}^{(1)} - \mathbf{w}^*)^\top \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1}$$

$$+ \frac{\sigma^2}{n} \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1} \cdot \mathbf{H}^{(2)} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1},$$

$$\mathcal{R}_1(\mathbf{w}) = \langle\mathbf{H}^{(1)}, \ (\mathbf{w} - \mathbf{w}^*)(\mathbf{w} - \mathbf{w}^*)^\top\rangle + \sigma^2,$$

$$\mathcal{R}_2(\mathbf{w}) = \langle\mathbf{H}^{(2)}, \ (\mathbf{w} - \mathbf{w}^*)(\mathbf{w} - \mathbf{w}^*)^\top\rangle + \sigma^2.$$

As a result,

$$\langle\mathbf{H}^{(1)}, \ \mathbb{E}(\mathbf{w}^{(1)} - \mathbf{w}^*)(\mathbf{w}^{(1)} - \mathbf{w}^*)^\top\rangle = \frac{\sigma^2}{n}\mathrm{rank}(\mathbf{H}^{(1)}) \tag{35}$$

since by definition $\mathbf{H}^{(1)}\mathbf{I}_{\mathbb{K}^c} = 0$,

$$
\begin{aligned}
\langle\mathbf{H}^{(1)}, \ \mathbb{E}(\mathbf{w}^{(2)} - \mathbf{w}^*)(\mathbf{w}^{(2)} - \mathbf{w}^*)^\top\rangle &= \mu^2 \cdot \langle\mathbf{H}^{(1)}, \ \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1} \cdot \mathbf{I}_{\mathbb{K}^c} \cdot \mathbf{w}^*{\mathbf{w}^*}^\top \cdot \mathbf{I}_{\mathbb{K}^c} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1}\rangle \\
&\quad + \frac{\sigma^2}{n}\mu^2 \cdot \langle\mathbf{H}^{(1)}, \ \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1} \cdot {\mathbf{H}^{(1)}}^{-1} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1}\rangle \\
&\quad + \frac{\sigma^2}{n} \langle\mathbf{H}^{(1)}, \ \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1} \cdot \mathbf{H}^{(2)} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1}\rangle \\
&= \mathtt{bias_F} + \frac{\sigma^2}{n}\left(\mathtt{d_F} + \mathrm{rank}(\mathbf{H}^{(1)})\right),
\end{aligned}
\tag{36}
$$

and that

$$
\begin{aligned}
\langle\mathbf{H}^{(2)}, \ \mathbb{E}(\mathbf{w}^{(2)} - \mathbf{w}^*)(\mathbf{w}^{(2)} - \mathbf{w}^*)^\top\rangle &= \mu^2 \cdot \langle\mathbf{H}^{(2)}, \ \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1} \cdot \mathbf{I}_{\mathbb{K}^c} \cdot \mathbf{w}^*{\mathbf{w}^*}^\top \cdot \mathbf{I}_{\mathbb{K}^c} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1}\rangle \\
&\quad + \frac{\sigma^2}{n}\mu^2 \cdot \langle\mathbf{H}^{(2)}, \ \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1} \cdot {\mathbf{H}^{(1)}}^{-1} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1}\rangle \\
&\quad + \frac{\sigma^2}{n} \langle\mathbf{H}^{(2)}, \ \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1} \cdot \mathbf{H}^{(2)} \cdot \left(\mathbf{H}^{(2)} + \mu\mathbf{I}\right)^{-1}\rangle \\
&= \mathtt{bias_I} + \frac{\sigma^2}{n}\left(\mathtt{d_I} - \mathrm{rank}(\mathbf{H}^{(1)})\right).
\end{aligned}
\tag{37}
$$

Combine (35), (36), (37) and we are done with the proof of forgetting and intransigence. $\qquad\square$