Proceedings of Machine Learning Research vol 178:1–52, 2022 35th Annual Conference on Learning Theory

# How catastrophic can catastrophic forgetting be in linear regression?

**Itay Evron** ITAY@EVRON.ME
*Department of Electrical and Computer Engineering, Technion, Haifa, Israel*
**Edward Moroshko** EDWARD.MOROSHKO@GMAIL.COM
*Department of Electrical and Computer Engineering, Technion, Haifa, Israel*
**Rachel Ward** RWARD@MATH.UTEXAS.EDU
*Oden Institute for Computational Engineering and Sciences, University of Texas, Austin, TX*
**Nati Srebro** NATI@TTIC.EDU
*Toyota Technological Institute at Chicago, Chicago IL, USA*
**Daniel Soudry** DANIEL.SOUDRY@GMAIL.COM
*Department of Electrical and Computer Engineering, Technion, Haifa, Israel*

**Editors:** Po-Ling Loh and Maxim Raginsky



## Abstract

To better understand catastrophic forgetting, we study fitting an overparameterized linear model to a sequence of tasks with different input distributions. We analyze how much the model forgets the true labels of earlier tasks after training on subsequent tasks, obtaining exact expressions and bounds. We establish connections between continual learning in the linear setting and two other research areas – alternating projections and the Kaczmarz method. In specific settings, we highlight differences between forgetting and convergence to the offline solution as studied in those areas. In particular, when $T$ tasks in $d$ dimensions are presented cyclically for $k$ iterations, we prove an upper bound of $T^2 \min\{1/\sqrt{k}, d/k\}$ on the forgetting. This stands in contrast to the convergence to the offline solution, which can be arbitrarily slow according to existing alternating projection results. We further show that the $T^2$ factor can be lifted when tasks are presented in a random ordering.

## 1. Introduction

Continual learning or lifelong learning is a machine learning setting where data from different tasks are presented sequentially to the learner. The goal is to adapt the model to new tasks while preserving its performance on previously-learned tasks [53, 62, 70]. A key challenge in continual learning is the *catastrophic forgetting* phenomenon [21, 39, 56, 57], wherein adaptation of models to fit to new tasks often (unsurprisingly) leads to degradation in performance on previous tasks.

Despite recent advances in theoretical understanding of continual learning [4, 6, 14, 32, 34], catastrophic forgetting is not *fully* understood even in simple models. Consider sequentially learning from a stream of tasks. One should ask: what are the best and worst-case sequences of tasks? How does the similarity between tasks effect catastrophic forgetting? What can be said analytically about the benefits to revisiting (*i.e.,* replaying) tasks? When does forgetting truly becomes "catastrophic", so that it is impossible to learn *all* tasks sequentially?

In this work, we aim to theoretically characterize the *worst-case* catastrophic forgetting in overparameterized linear regression models.[1] To this end, we sequentially fit a linear model to tasks observed in some ordering. We analyze the forgetting convergence of linear regressors obtained by GD/SGD which is trained *to convergence* on data from the current task.

1. We believe it is necessary to do so before moving to more complex models. Moreover, any linear regression result can be applied to complex models (e.g., deep networks) in the neural kernel regime (NTK), as in [6, 14].

We explain how in this linear setting, continual learning repeatedly *projects* previous solutions onto the solution spaces of newer tasks. Interestingly, we show that, under a realizability assumption, these projections contract the distance to the minimum norm offline solution that solves *all* tasks. We analyze this contraction and the resulting convergence.

A setting similar to ours has been extensively studied in the *alternating projections* literature. There, a vector is iteratively projected onto closed subspaces (or convex sets in general), in order to find a solution in their intersection. For instance, Kayalar and Weinert [30] studied Halperin's cyclic setting [25] and analyzed the convergence of $\left\|(\boldsymbol{P}_2\boldsymbol{P}_1)^n-\boldsymbol{P}_{1\cap 2}\right\|^2$, where $\boldsymbol{P}_1$, $\boldsymbol{P}_2$, and $\boldsymbol{P}_{1\cap 2}$ are orthogonal projections onto subspaces $H_1$, $H_2$, and $H_1\cap H_2$ (respectively). In contrast, our goal is to analyze the projection *residuals* (*i.e.,* the forgetting). That is, we mainly focus on $\left\|(\boldsymbol{I}-\boldsymbol{P}_1)(\boldsymbol{P}_2\boldsymbol{P}_1)^n\right\|^2$, rather than on the convergence to $H_1\cap H_2$ (*i.e.,* an offline solution). Due to this difference, we are able to derive *uniform* data-independent upper bounds on forgetting, even when convergence to the offline solution is arbitrarily slow and "traditional" bounds become trivial.

Moreover, our fitting procedure can also be seen as a Kaczmarz method [29], where one solves a linear equation system $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$ by iteratively solving subsets (*i.e.,* tasks) of it. Here also, typical bounds involve data-dependent properties like the spectrum of $\boldsymbol{A}$, which are trivial in the worst case.

**Our Contributions.** We thoroughly analyze catastrophic forgetting in linear regression optimized by the plain memoryless (S)GD algorithm (*i.e.,* without actively trying to mitigate forgetting), and:

- Identify cases where there is no forgetting.
- Show that without any restrictions, one can construct task sequences where forgetting is maximal and essentially *catastrophic*.
- Connect catastrophic forgetting to a large body of research on alternating projections and the Kaczmarz method. Then, we use this perspective to investigate the *worst-case* forgetting when $T$ tasks in $d$ dimensions are seen *repeatedly* for $k$ iterations, under two different orderings:
  - **Cyclic task orderings.** For $T\geq 3$, we uniformly upper bound the forgetting by $\frac{T^2}{\sqrt{k}}$ and $\frac{T^2 d}{k}$, and lower bound it by $\frac{T^2}{k}$. To the best of our knowledge, our analysis uncovers novel bounds for residuals in the cyclic block Kaczmarz setting [15] as well.
  - **Random task orderings.** We upper bound the expected forgetting by $\frac{d}{k}$, independently of $T$.
- Analyze the effect of similarity, or angles, between two consecutive tasks on the forgetting. We find that after seeing these tasks *once* — intermediate angles are most prone to forgetting, but after *repeating* the tasks — small angles (*i.e.,* nearly-aligned tasks) cause the highest forgetting.

## 2. Problem setting

**Motivating example.** Before we formalize our problem, we give a motivating example to keep in mind. Suppose we are interested to learn a predictor for pedestrian detection in an autonomous car. This detector is required to operate well in $T$ geographically distant environments (*e.g.,* different countries), or "tasks". To train the detector, we need to drive the car around, but can do this only in one environment at a time. We do this until the detector has good performance in that environment. Then we ship the car to another environment, drive it there to train the detector, and so on (potentially revisiting past environments). Notably, while the overall problem remains constant (pedestrian detection), each time we need to solve it for a potentially different landscape (*e.g.,* city,

forest, desert), which can radically change the input distribution. The question we aim to answer is: when would the detector be able to work well in all the environments it visited during training, even though it only observed them sequentially?

**General setting.** We consider fitting a linear model $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$, parameterized by $\boldsymbol{w} \in \mathbb{R}^d$, on a sequence of tasks originating from an arbitrary set of $T$ regression problems $\left\{\left(\boldsymbol{X}_m, \boldsymbol{y}_m\right)\right\}_{m=1}^T$. During $k$ iterations, tasks are seen according to a *task ordering* $\tau\colon \mathbb{N}^+ \to [T]$. That is, the learner is presented with a sequence $\left(\boldsymbol{X}_{\tau(1)}, \boldsymbol{y}_{\tau(1)}\right), \left(\boldsymbol{X}_{\tau(2)}, \boldsymbol{y}_{\tau(2)}\right), \ldots, \left(\boldsymbol{X}_{\tau(k)}, \boldsymbol{y}_{\tau(k)}\right)$. Starting from $\boldsymbol{w}_0 = \boldsymbol{0}_d$, the model is fitted sequentially, yielding a sequence of iterates $\boldsymbol{w}_1, ..., \boldsymbol{w}_k$. At each iteration $t$ we obtain $\boldsymbol{w}_t$ by fitting the *current* task $\left(\boldsymbol{X}_{\tau(t)}, \boldsymbol{y}_{\tau(t)}\right)$, *without* access to other tasks.

**Tasks.** Each *task* $m = 1, 2, \ldots, T$ corresponds to a regression problem $\left(\boldsymbol{X}_m, \boldsymbol{y}_m\right)$, *i.e.,* to a data matrix $\boldsymbol{X}_m \in \mathbb{R}^{n_m \times d}$ with $n_m$ samples in $d$ dimensions and a vector $\boldsymbol{y}_m \in \mathbb{R}^{n_m}$ of the corresponding labels. We focus on the overparametrized regime, where every task $\left(\boldsymbol{X}_m, \boldsymbol{y}_m\right)$ is rank deficient, *i.e.,* $r_m \triangleq \operatorname{rank}\left(\boldsymbol{X}_m\right) < d$, thus admitting infinitely-many linear models that perfectly fit the data.

**Notation.** Denote by $\boldsymbol{A}^+$ the Moore–Penrose inverse of $\boldsymbol{A}$. Denote by $\boldsymbol{P}_m$ the orthogonal projection onto the *null space* of a matrix $\boldsymbol{X}_m$, *i.e.,* $\boldsymbol{P}_m = \boldsymbol{I} - \boldsymbol{X}_m^+ \boldsymbol{X}_m$. Denote the $\ell^2$ norm of a vector by $\|\boldsymbol{v}\|$, and the *spectral* norm of a matrix by $\|\boldsymbol{A}\|$. Denote the $d$-dimensional closed Euclidean ball as $\mathcal{B}^d \triangleq \left\{\boldsymbol{v} \in \mathbb{R}^d \;\middle|\; \|\boldsymbol{v}\| \leq 1\right\}$. To avoid ambiguity, we explicitly exclude zero from the set of natural numbers and use $\mathbb{N}^+ \triangleq \mathbb{N} \backslash \{0\}$. Finally, denote the natural numbers from 1 to $T$ by $[T] \triangleq \{1, \ldots, T\}$.

### 2.1. Task collections

A *task collection* is an (unordered) set of $T$ tasks, *i.e.,* $S = \left\{\left(\boldsymbol{X}_m, \boldsymbol{y}_m\right)\right\}_{m=1}^T$. Throughout the paper, we focus on task collections from the set:

$$\mathcal{S}_T \triangleq \left\{ \left\{\left(\boldsymbol{X}_m, \boldsymbol{y}_m\right)\right\}_{m=1}^T \;\middle|\; \underbrace{\forall m\colon \|\boldsymbol{X}_m\| \leq 1}_{\text{Assumption 1}}, \; \underbrace{\exists \boldsymbol{w} \in \mathcal{B}^d\colon \boldsymbol{y}_m = \boldsymbol{X}_m \boldsymbol{w}, \; \forall m \in [T]}_{\text{Assumption 2}} \right\}.$$

That is, we make the two following assumptions on the task collections.

**Assumption 1 (Bounded data)** *Singular values of all data matrices are bounded (w.l.o.g., by* 1*).*

**Assumption 2 (Realizability)** *Tasks are jointly realizable by a linear predictor with a bounded norm (w.l.o.g., bounded by* 1*).*

Assumption 1 is merely a technical assumption on the data scale, simplifying our presentation. Assumption 2 on joint realizability of tasks is essential to our derivations, especially for Eq. (5). We note that this is a reasonable assumption in a highly overparameterized models like wide neural networks in the NTK regime, or in noiseless settings with an underlying linear model.

### 2.2. Forgetting and catastrophic forgetting

We fit the linear model sequentially on tasks $\tau(1), \ldots, \tau(k)$. Starting from the solution to its preceding task, each task is fitted *without* access to previous tasks. The goal in continual learning is to not *forget* what we learned on previous tasks. For example, in the motivating example given in the beginning of the section, we want the pedestrian detector to still be able to work well in previous environments after we train it in a new environment.

Formally, like in a recent work [14], we measure the forgetting of $\boldsymbol{w}$ on the task presented at iteration $t$ by the squared loss $\mathcal{L}_t(\boldsymbol{w}) \triangleq \left\|\boldsymbol{X}_{\tau(t)}\boldsymbol{w}-\boldsymbol{y}_{\tau(t)}\right\|^2$, and define the forgetting as follows.

**Definition 3 (Forgetting)** *Let $S=\left\{\left(\boldsymbol{X}_m,\boldsymbol{y}_m\right)\right\}_{m=1}^{T}\in\mathcal{S}_T$ be a task collection fitted according to an ordering $\tau$: $\mathbb{N}^+\to[T]$. The forgetting at iteration $k$ is the average loss of already seen tasks, i.e.,*

$$F_{\tau,S}(k)=\frac{1}{k}\sum_{t=1}^{k}\mathcal{L}_t(\boldsymbol{w}_k)=\frac{1}{k}\sum_{t=1}^{k}\left\|\boldsymbol{X}_{\tau(t)}\boldsymbol{w}_k-\boldsymbol{y}_{\tau(t)}\right\|^2 .$$

In words, we say that the model "forgets" the labels of a task seen at iteration $t$ after fitting $k$ tasks, if $\boldsymbol{w}_k$ does not perfectly predict $\boldsymbol{y}_{\tau(t)}$ from $\boldsymbol{X}_{\tau(t)}$, *i.e.,* $\mathcal{L}_t(\boldsymbol{w}_k)>0$. We note in passing that *forgetting* should not be confused with *regret* since they can exhibit different behaviors (see Section 7).

In the following, we capture the worst possible convergence behavior under a given ordering.

**Definition 4 (Worst-case forgetting)** *The worst-case forgetting of a task ordering $\tau$ after $k$ iterations is the maximal forgetting at iteration $k$ on* any *task collection in $\mathcal{S}_T$, i.e.,* $\sup_{S\in\mathcal{S}_T}F_{\tau,S}(k)$.

For specific task orderings, we are able to bound this worst-case forgetting by non-trivial uniform (data-independent) bounds that converge to 0 with the number of iterations $k$. When the worst-case forgetting does not converge to 0, we say that forgetting is *catastrophic*. More formally,

**Definition 5 (Catastrophic forgetting of a task ordering)** *Given an ordering $\tau$ over $T$ tasks, the forgetting is* not *catastrophic if* $\lim_{k\to\infty}(\sup_{S\in\mathcal{S}_T}F_{\tau,S}(k))=0$.

In this paper we focus on the quantities $F_{\tau,S}(k)$ and $\sup_{S\in\mathcal{S}_T}F_{\tau,S}(k)$, aiming to answer the following questions: under what conditions on $\tau$ and $S$ there is no forgetting, *i.e.,* $F_{\tau,S}(k)=0$? What can we say about $\sup_{S\in\mathcal{S}_T}F_{\tau,S}(k)$ for general task orderings? Can forgetting be catastrophic when fitting a finite number of tasks in a cyclic or random ordering?

### 2.3. Fitting procedure

Our ultimate goal is to minimize the forgetting (Definition 3). To this end, in this paper we analyze the following fitting procedure of tasks, corresponding to the simplest continual learning setting. At each iteration $t=1,..,k$, we start from the previous iterate $\boldsymbol{w}_{t-1}$ and run (stochastic) gradient descent *to convergence* so as to minimize the squared loss on the *current* task, *i.e.,* $\mathcal{L}_t(\boldsymbol{w})=\left\|\boldsymbol{X}_{\tau(t)}\boldsymbol{w}-\boldsymbol{y}_{\tau(t)}\right\|^2$, obtaining the new iterate $\boldsymbol{w}_t$.

Since we work in the overparameterized regime, where $r_t<d$ for all $t$, each $\boldsymbol{w}_t$ perfectly fits its corresponding task $\tau(t)$, *i.e.,* $\boldsymbol{X}_{\tau(t)}\boldsymbol{w}_t=\boldsymbol{y}_{\tau(t)}$. In addition, for each task, (S)GD at convergence returns the unique solution which implicitly minimizes the $\ell^2$ distance to the initialization [22, 80], hence we can express $\boldsymbol{w}_t$ as the unique solution to the following optimization problem:

$$\boldsymbol{w}_t=\underset{\boldsymbol{w}}{\operatorname{argmin}}\left\|\boldsymbol{w}-\boldsymbol{w}_{t-1}\right\| \ , \quad \text{s.t. } \boldsymbol{X}_{\tau(t)}\boldsymbol{w}=\boldsymbol{y}_{\tau(t)} \ . \tag{1}$$

**Our iterative update rule.** The solution to the above optimization problem is given by

$$\boldsymbol{w}_t=\boldsymbol{w}_{t-1}+\boldsymbol{X}_{\tau(t)}^{+}(\boldsymbol{y}_{\tau(t)}-\boldsymbol{X}_{\tau(t)}\boldsymbol{w}_{t-1}) \ . \tag{2}$$

When $\boldsymbol{X}_{\tau(t)}$ contains one sample only (*i.e.,* its rank is 1), our update rule is equivalent to those of the Kaczmarz method [29] and the NLMS algorithm [26, 67]. Moreover, it can be seen as an SGD update on the forgetting from Definition 3, *i.e.,* $\boldsymbol{w}_t=\boldsymbol{w}_{t-1}+\frac{1}{\|\boldsymbol{x}_{\tau(t)}\|^2}\left(y_{\tau(t)}-\boldsymbol{x}_{\tau(t)}^{\top}\boldsymbol{w}_{t-1}\right)\boldsymbol{x}_{\tau(t)}$. Given tasks with *many* samples (rank greater than 1), our rule is equivalent to that of the block Kaczmarz method [15]. We discuss these connections in Section 7.

**Minimum norm offline solution.** Under the realizability assumption 2 there might be infinitely-many *offline solutions* that perfectly fit *all* tasks. To facilitate our discussion, we focus on the minimum $\ell^2$-norm offline solution $\boldsymbol{w}^\star$ (often referred to as the *offline solution* for brevity), *i.e.,*

$$\boldsymbol{w}^\star \triangleq \operatorname{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} \tfrac{1}{2}\|\boldsymbol{w}\|^2, \text{ s.t. } \boldsymbol{X}_m\boldsymbol{w} = \boldsymbol{y}_m, \ \forall m \in [T] \ . \tag{3}$$

**Task solution spaces.** Finally, from Eq. (1) we see that at the end of the $t$-th iteration, the iterate $\boldsymbol{w}_t$ must lie in the *solution space* of task $\tau(t)$, which is an affine subspace defined as follows

$$\mathcal{W}_{\tau(t)} \triangleq \left\{\boldsymbol{w} \,\middle|\, \boldsymbol{X}_{\tau(t)}\boldsymbol{w} = \boldsymbol{y}_{\tau(t)}\right\} = \boldsymbol{w}^\star + \operatorname{null}(\boldsymbol{X}_{\tau(t)}) \ . \tag{4}$$

## 3. Forgetting dynamics

To analyze forgetting, we emphasize the projective nature of learning. We rewrite the update rule from Eq. (2) by employing the realizability Assumption 2 and plugging in $\boldsymbol{X}_m\boldsymbol{w}^\star = \boldsymbol{y}_m$ into that equation. Then, we subtract $\boldsymbol{w}^\star$ from both sides, and reveal an equivalent affine update rule, *i.e.,*

$$\boldsymbol{w}_t - \boldsymbol{w}^\star = \left(\boldsymbol{I} - \boldsymbol{X}^+_{\tau(t)}\boldsymbol{X}_{\tau(t)}\right)\boldsymbol{w}_{t-1} + \boldsymbol{X}^+_{\tau(t)}\boldsymbol{X}_{\tau(t)}\boldsymbol{w}^\star - \boldsymbol{w}^\star \triangleq \boldsymbol{P}_{\tau(t)}\left(\boldsymbol{w}_{t-1} - \boldsymbol{w}^\star\right) , \tag{5}$$

where we remind that $\boldsymbol{P}_m \triangleq \boldsymbol{I} - \boldsymbol{X}^+_m\boldsymbol{X}_m$ is the projection operator on the solution space $\mathcal{W}_m$.

**Geometric interpretation.** Using properties of pseudo-inverses and operator norms we get that $\forall m \in [T]: \ \left\|\boldsymbol{X}_m\boldsymbol{u}\right\|^2 = \left\|\boldsymbol{X}_m\boldsymbol{X}^+_m\boldsymbol{X}_m\boldsymbol{u}\right\|^2 \le \left\|\boldsymbol{X}_m\right\|^2\left\|\boldsymbol{X}^+_m\boldsymbol{X}_m\boldsymbol{u}\right\|^2 = \left\|\boldsymbol{X}_m\right\|^2\left\|\left(\boldsymbol{I}-\boldsymbol{P}_m\right)\boldsymbol{u}\right\|^2$. Then, we recall that $\left\|\boldsymbol{X}_m\right\| \le 1$ (Assumption 1), and reveal that the forgetting can be seen as the mean of the squared *residuals* from projecting $\left(\boldsymbol{w}_k - \boldsymbol{w}^\star\right)$ onto previously-seen solution spaces. That is,

$$F_{\tau,S}(k) = \frac{1}{k}\sum\nolimits_{t=1}^{k}\left\|\boldsymbol{X}_{\tau(t)}\left(\boldsymbol{w}_k - \boldsymbol{w}^\star\right)\right\|^2 \le \frac{1}{k}\sum\nolimits_{t=1}^{k}\left\|\left(\boldsymbol{I} - \boldsymbol{P}_{\tau(t)}\right)\left(\boldsymbol{w}_k - \boldsymbol{w}^\star\right)\right\|^2 . \tag{6}$$

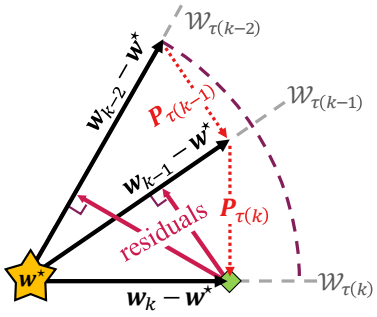



Figure 1: **Projection illustration.** According to Eq. (5), $(\boldsymbol{w}_{k-1} - \boldsymbol{w}^\star)$ is given by projecting $(\boldsymbol{w}_{k-2} - \boldsymbol{w}^\star)$ onto the solution space of the $(k-1)$th task, *i.e.,* $\mathcal{W}_{\tau(k-1)}$, which in this figure is a rank-1 affine subspace in $\mathbb{R}^2$. In turn, $(\boldsymbol{w}_{k-1} - \boldsymbol{w}^\star)$ is projected onto $\mathcal{W}_{\tau(k)}$ to obtain $(\boldsymbol{w}_k - \boldsymbol{w}^\star)$, and so on. Overall, the solution is continually getting closer to $\boldsymbol{w}^\star$. Moreover, the forgetting (magenta) is the mean of the squared residuals from projecting $(\boldsymbol{w}_k - \boldsymbol{w}^\star)$ onto previously seen solution spaces, as can be seen from Eq. (6).

**Learning as contracting.** Recursively, the affine update rule from Eq. (5) provides a closed-form expression for the distance between the iterate and the offline solution (recall $\boldsymbol{w}_0 = \boldsymbol{0}$):

$$\boldsymbol{w}_t - \boldsymbol{w}^\star = \boldsymbol{P}_{\tau(t)}\cdots\boldsymbol{P}_{\tau(1)}\left(\cancel{\boldsymbol{w}_0} - \boldsymbol{w}^\star\right) . \tag{7}$$

Since orthogonal projections are non-expansive operators, it also follows that

$$\forall t \in [T]: \ \|\boldsymbol{w}_t - \boldsymbol{w}^\star\| \le \|\boldsymbol{w}_{t-1} - \boldsymbol{w}^\star\| \le \cdots \le \|\cancel{\boldsymbol{w}_0} - \boldsymbol{w}^\star\| = \|\boldsymbol{w}^\star\| ,$$

hinting at a possible convergence towards the offline solution, as depicted in Figure 1.

Combining Eq. (6) and (7), we express the average forgetting (Definition 3) of an ordering $\tau$ over a task collection $S = \{(\boldsymbol{X}_m, \boldsymbol{y}_m)\}_{m=1}^T$ as follows,

$$F_{\tau,S}(k) = \frac{1}{k}\sum_{t=1}^{k} \left\| \boldsymbol{X}_{\tau(t)} \boldsymbol{P}_{\tau(k)} \cdots \boldsymbol{P}_{\tau(1)} \boldsymbol{w}^\star \right\|^2 \leq \frac{1}{k}\sum_{t=1}^{k} \left\| \left(\boldsymbol{I} - \boldsymbol{P}_{\tau(t)}\right) \boldsymbol{P}_{\tau(k)} \cdots \boldsymbol{P}_{\tau(1)} \boldsymbol{w}^\star \right\|^2 . \quad (8)$$

**Worst-case formulation.** So far, $\|\boldsymbol{X}_m \boldsymbol{X}_m^+ \boldsymbol{X}_m \boldsymbol{u}\|^2 \leq \|\boldsymbol{X}_m^+ \boldsymbol{X}_m \boldsymbol{u}\|^2$ is the only inequality we used (at Eq. (6), relying on Assumption 1). Importantly, this inequality saturates when all non-zero singular values of $\boldsymbol{X}_m, \forall m \in [T]$ are 1. Consequentially, the *worst-case* forgetting in Definition 4 can be simply expressed in terms of $T$ projection matrices $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_T \in \mathbb{R}^{d\times d}$, as

$$\sup_{S\in\mathcal{S}_T} F_{\tau,S}(k) \;=\; \sup_{\boldsymbol{P}_1,\ldots,\boldsymbol{P}_T} \frac{1}{k}\sum\nolimits_{t=1}^{k} \left\| \left(\boldsymbol{I} - \boldsymbol{P}_{\tau(t)}\right) \boldsymbol{P}_{\tau(k)} \cdots \boldsymbol{P}_{\tau(1)} \right\|^2 , \quad (9)$$

where we also used Assumption 2 that $\|\boldsymbol{w}^\star\| \leq 1$.

Throughout this paper, we mainly analyze Eq. (8) and (9) from different perspectives. Multiplying from the left by $\left\{\boldsymbol{I} - \boldsymbol{P}_{\tau(t)}\right\}_{t=1}^k$ is what distinguishes our quantity of interest — the forgetting — from quantities studied in the area of alternating projections.

**Principal angles between two tasks.** Finally, we briefly present *principal angles*, which affect forgetting dynamics, as discussed throughout this paper. Much of the research on alternating projections has focused on establishing notions of angles between subspaces [11, 51]. *Principal angles*, also known as canonical angles, are a popular choice for angles between two linear subspaces [5, 9], having many applications in numerical analysis (e.g., in the generalized eigenvalue problem [19]). These angles geometrically describe a pair of subspaces, by recursively taking the smallest angle between any two vectors in these subspaces that are orthogonal to previously chosen vectors. We elaborate on the definition and the role of these angles in App A.3. There, we visualize these angles and explain that the non-zero principal angles between the row spaces of two tasks, *i.e.,* $\mathrm{range}(\boldsymbol{X}_1^\top), \mathrm{range}(\boldsymbol{X}_2^\top)$, are identical to those between the corresponding solution spaces $\mathcal{W}_1, \mathcal{W}_2$.

**The rest of our paper.** We study forgetting under different task orderings. In Section 4 we consider arbitrary orderings and show when there is provably *no* forgetting, and when forgetting is arbitrarily high, *i.e.,* catastrophic. We analyze cyclic and random orderings in Sections 5 and 6. For both these orderings, we derive convergence guarantees and prove forgetting *cannot* be catastrophic.

## 4. Arbitrary task orderings

**Identity ordering.** In this section we consider arbitrary sequences of tasks, *i.e.,* we do not impose any specific ordering. To this end, we take $k = T$ and an *identity ordering* $\tau$ s.t. $\tau(t) = t$, $\forall t \in \mathbb{N}^+$. To simplify notation, in this section only, we suppress $\tau$ and use $\boldsymbol{X}_{\tau(m)} = \boldsymbol{X}_m$ interchangeably.

### 4.1. No forgetting cases

Consider learning two tasks sequentially: $\left(\boldsymbol{X}_1, \boldsymbol{y}_1\right)$ and then $\left(\boldsymbol{X}_2, \boldsymbol{y}_2\right)$. Right after learning the second task, we have $\mathcal{L}_2(\boldsymbol{w}_2) = \|\boldsymbol{X}_2 \boldsymbol{w}_2 - \boldsymbol{y}_2\|^2 = 0$. Thus, the forgetting from Eq. (8) becomes $F_{\tau,S}(2) = \frac{1}{2}\|\boldsymbol{X}_1 \boldsymbol{P}_2 \boldsymbol{P}_1 \boldsymbol{w}^\star\|^2$. We now derive sufficient and necessary conditions for no forgetting.

**Theorem 6 (No forgetting in two-task collections)** *Let $S = \{(\boldsymbol{X}_1, \boldsymbol{y}_1), (\boldsymbol{X}_2, \boldsymbol{y}_2)\} \in \mathcal{S}_{T=2}$ be a task collection with 2 tasks, fitted under an identity ordering $\tau$, i.e., $(\boldsymbol{X}_1, \boldsymbol{y}_1)$ and then $(\boldsymbol{X}_2, \boldsymbol{y}_2)$. Then the following conditions are equivalent:*

1. *For* any *labeling $\boldsymbol{y}_1, \boldsymbol{y}_2$ (or equivalently, any minimum norm solution $\boldsymbol{w}^\star$), after fitting the second task, the model does not "forget" the first one. That is, $F_{\tau,S}(2) = 0$.*
2. *It holds that $\boldsymbol{X}_1 \boldsymbol{P}_2 \boldsymbol{P}_1 = \boldsymbol{0}_{n_1 \times d}$.*
3. *Each principal angle between the tasks, i.e., $\mathrm{range}(\boldsymbol{X}_1^\top)$ and $\mathrm{range}(\boldsymbol{X}_2^\top)$, is either $0$ or $\pi/2$.*

The proof is given in App. B.1. The exact definition of angles between tasks is given in App. A.3.

For instance, the above conditions hold when $\mathrm{range}(\boldsymbol{X}_1^\top) \subseteq \mathrm{range}(\boldsymbol{X}_2^\top)$ (or vice-versa), *i.e.,* there is no forgetting when tasks have maximum overlap in the row span of inputs. The third condition aligns with the empirical observation in Ramasesh et al. [56], wherein catastrophic forgetting in overparameterized neural networks is small when the tasks are either very similar or very distinct. On the other hand, this seemingly contradicts the conclusions in Doan et al. [14] that similar tasks are *potentially* bad for forgetting. However, their conclusions are based on a loose upper bound on the forgetting. In Section 5.1, we carefully analyze the forgetting dynamics of two tasks seen repeatedly in cycles and show that increasingly-similar tasks *can* be worse for forgetting, but only after multiple cycles. We elaborate on these connections in App B.1.3.

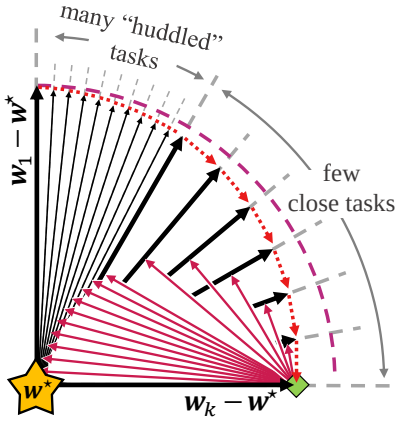
### 4.2. Maximal forgetting cases: can $F_{\tau,S}(k) \to 1$ ?

Now, we present an adversarial task collection that yields arbitrarily high forgetting: at the end of learning, the learner almost completely "forgets" previously-seen tasks. We use this opportunity to build further intuition on two factors causing high forgetting.

Our construction is intuitively based on the geometric interpretation from Eq. (6) and Figure 1, that the forgetting is the *mean* of the squared residuals from projecting $(\boldsymbol{w}_k - \boldsymbol{w}^\star)$ onto the solution spaces of previously-seen tasks, *i.e.,* $F_{\tau,S}(k) \leq \frac{1}{k}\sum_{t=1}^{k} \left\| \left(\boldsymbol{I} - \boldsymbol{P}_{\tau(t)}\right)\left(\boldsymbol{w}_k - \boldsymbol{w}^\star\right) \right\|^2$.

Figure 2: **Illustrating the adversarial construction.** For the discussed residuals to be large, our construction ensures that:

1. **The iterates are kept afar from the $\boldsymbol{w}^\star$.** Since $F_{\tau,S}(k) \leq \|\boldsymbol{w}_k - \boldsymbol{w}^\star\|^2 \leq \|\boldsymbol{w}_t - \boldsymbol{w}^\star\|^2, \forall t \leq k$, it is important to maintain a large $\|\boldsymbol{w}_t - \boldsymbol{w}^\star\|$ in all iterations. We achieve this by using similar *consecutive* tasks.
2. ***Most* solution spaces are orthogonal to the last one.** For the averaged residuals to be large, the last $(\boldsymbol{w}_k - \boldsymbol{w}^\star)$ should be orthogonal to as many previous solution spaces as possible. For this, we "huddle" most of the tasks near the first one, almost orthogonally to the last.




**Theorem 7 (Forgetting can be arbitrarily bad)** *When using the identity ordering (i.e., $\tau(t) = t$), thus seeing each task once, the worst-case forgetting after $k$ iterations is arbitrarily bad, i.e.,*

$$1 - \sup_{S \in \mathcal{S}_{T=k}} F_{\tau,S}(k) \leq \mathcal{O}\left(1/\sqrt{k}\right) .$$

The exact construction details and the proof are given in App B.2.

By now, we understand that under arbitrary task orderings, there exist task sequences where the learner almost *completely* forgets previously-learned expertise and forgetting is indeed *catastrophic*. In the sections to follow, we show that cyclic and random orderings *do not* suffer from this flaw.

## 5. Cyclic task orderings

We again consider collections of $T$ tasks $\left\{\left(\boldsymbol{X}_m, \boldsymbol{y}_m\right)\right\}_{m=1}^{T}$, but now we study the forgetting when tasks are presented in a cyclic ordering $\tau$, *i.e.,* $\tau\left(t\right) = \left(\left(t-1\right) \bmod T\right) + 1, \ \forall t \in \mathbb{N}^{+}$. For example, suppose we want to train a pedestrian detector continuously during different times of the day (morning, noon, evening, and night), so that the task order forms a fixed cycle. Such cyclic settings also arise in search engines, e-commerce, and social networks, where tasks (*i.e.,* distributions) are largely influenced by events that recur either weekly (*e.g.,* weekdays vs. weekends), monthly (*e.g.,* paydays), annually (*e.g.,* holidays), and so on.

Under cyclic orderings, the forgetting from Eq. (8) after $n$ cycles, becomes

$$F_{\tau,S}(k=nT) = \frac{1}{T}\sum_{m=1}^{T}\left\|\boldsymbol{X}_m\left(\boldsymbol{P}_T\cdots\boldsymbol{P}_1\right)^n\boldsymbol{w}^\star\right\|^2 \leq \frac{1}{T}\sum_{m=1}^{T}\left\|\left(\boldsymbol{I}-\boldsymbol{P}_m\right)\left(\boldsymbol{P}_T\cdots\boldsymbol{P}_1\right)^n\right\|^2. \quad (10)$$

### 5.1. Warm up: Exact forgetting analysis with $T = 2$ tasks

Exploiting the connection we made to the field of alternating projections, we first analyze the convergence to the minimum norm offline solution in terms of the *Friedrichs angle* [18] between the tasks, *i.e.,* their minimal *non-zero* principal angle $\theta_F \triangleq \min_{i:\theta_i\neq 0}\theta_i > 0$ (as explained in App A.3).

**Theorem 8 (Convergence to the minimum norm offline solution)** *For any task collection of two distinct tasks fitted in a cyclic ordering $\tau$, the distance from the offline solution after $k = 2n$ iterations ($n$ cycles) is* tightly *upper bounded by $\left\|\boldsymbol{w}_k - \boldsymbol{w}^\star\right\|^2 \leq \left(\cos^2\theta_F\right)^{k-1}\left\|\boldsymbol{w}^\star\right\|^2$, where $\theta_F$ is the Friedrichs angle between the given tasks, as defined above.*

**Proof** Plugging in the cyclic ordering definition into the recursive form of Eq. (7), we obtain $\left\|\boldsymbol{w}_k - \boldsymbol{w}^\star\right\| = \left\|\boldsymbol{P}_{\tau(k)}\cdots\boldsymbol{P}_{\tau(1)}\boldsymbol{w}^\star\right\| = \left\|\left(\boldsymbol{P}_2\boldsymbol{P}_1\right)^n\boldsymbol{w}^\star\right\|$.

A known alternating projection result by Kayalar and Weinert [30] (Theorem 2 therein) states that $\left\|\left(\boldsymbol{P}_2\boldsymbol{P}_1\right)^n - \boldsymbol{P}_{1\cap 2}\right\|^2 = \left(\cos^2\theta_F\right)^{k-1}$, where in our context, $\boldsymbol{P}_{1\cap 2}$ projects onto the null spaces' intersection, *i.e.,* $\mathrm{null}(\boldsymbol{X}_1)\cap\mathrm{null}(\boldsymbol{X}_2)$. Since the *minimum norm* solution $\boldsymbol{w}^\star$ must lie in $\mathrm{range}\left(\boldsymbol{X}_1^\top\right)\cup\mathrm{range}\left(\boldsymbol{X}_2^\top\right)$, then by properties of orthogonal complements we have $\boldsymbol{P}_{1\cap 2}\boldsymbol{w}^\star = \boldsymbol{0}$. Then, we see that $\left\|\boldsymbol{w}_k - \boldsymbol{w}^\star\right\|^2 = \left\|\left(\boldsymbol{P}_2\boldsymbol{P}_1\right)^n\boldsymbol{w}^\star\right\|^2 = \left\|\left(\left(\boldsymbol{P}_2\boldsymbol{P}_1\right)^n - \boldsymbol{P}_{1\cap 2}\right)\boldsymbol{w}^\star\right\|^2$, and conclude:

$$\left\|\boldsymbol{w}_k - \boldsymbol{w}^\star\right\|^2 = \left\|\left(\left(\boldsymbol{P}_2\boldsymbol{P}_1\right)^n - \boldsymbol{P}_{1\cap 2}\right)\boldsymbol{w}^\star\right\|^2 \leq \left\|\left(\boldsymbol{P}_2\boldsymbol{P}_1\right)^n - \boldsymbol{P}_{1\cap 2}\right\|^2\left\|\boldsymbol{w}^\star\right\|^2 = \left(\cos^2\theta_F\right)^{k-1}\left\|\boldsymbol{w}^\star\right\|^2.$$

Clearly, a carefully chosen $\boldsymbol{w}^\star$ (induced by $\boldsymbol{y}_1, \boldsymbol{y}_2$) can saturate the inequality, making it tight. ■

Note that the rate of convergence to $\boldsymbol{w}^\star$ can be arbitrarily slow when the Friedrichs angle $\theta_F \to 0$. Importantly, this means that there can be no data-independent convergence guarantees to $\boldsymbol{w}^\star$. One might think that this implies that the forgetting (*i.e.,* the residuals) is also only trivially bounded, however a careful analysis shows that this is not the case.

In contrast to the above Theorem 8, we now show that the forgetting *is* non-trivially bounded.

**Lemma 9 (Angles' effect on forgetting $T=2$)** *For any task collection $S \in \mathcal{S}_{T=2}$ of two tasks, the forgetting after $k=2n$ iterations (i.e., $n$ cycles) is tightly upper bounded by*

$$F_{\tau,S}(k) \leq \frac{1}{2}\max_i \left\{ \left(\cos^2\theta_i\right)^{k-1}\left(1-\cos^2\theta_i\right) \right\} ,$$

*where $\{\theta_i\}_i \subseteq \left(0, \frac{\pi}{2}\right]$ are the non-zero principal angles between the two tasks in $S$. Moreover, the above inequality saturates when all non-zero singular values of the first task (i.e., of $\boldsymbol{X}_1$) are 1s.*

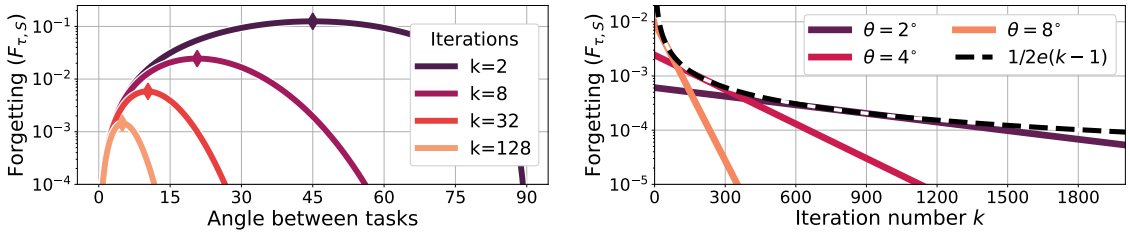
In contrast to $\cos^2\theta_F$ that bounds the distance to the offline solution (see Theorem 8) and can be arbitrarily close to 1, the quantities $\left(\cos^2\theta_i\right)^{k-1}\left(1-\cos^2\theta_i\right)$ in Lemma 9 are upper bounded *uniformly* for any $\theta_i$, which allows deriving a data-independent expression for the worst-case forgetting in the next theorem.

**Theorem 10 (Worst-case forgetting when $T=2$)** *For a cyclic ordering $\tau$ of two tasks, the worst-case forgetting after $k=2n$ iterations (i.e., $n$ cycles), is*

$$\sup_{S\in\mathcal{S}_{T=2}} F_{\tau,S}(k) = \frac{1}{2e\,(k-1)} - \frac{1}{4e\,(k-1)^2} + \mathcal{O}\left(\frac{1}{k^3}\right) .$$

The proofs for both Lemma 9 and Theorem 10 are given in App C.1.

**Demonstration.** Figure 3 demonstrates our analysis for the worst-case forgetting on $T=2$ tasks. We consider a simplistic case where both tasks are of rank $d-1$, *i.e.,* $\mathrm{rank}(\boldsymbol{X}_1)=\mathrm{rank}(\boldsymbol{X}_2)=d-1$, thus having solution spaces of rank 1 with a straightforward single angle $\theta$ between them.




($a$) Effect of task similarity on forgetting.

($b$) A sharp uniform bound for the forgetting.

Figure 3: Demonstration of forgetting (Lemma 9) and worst-case forgetting (Theorem 10) for $T=2$.

Figure 3($a$) demonstrates the analytical effect of task angles, *i.e.,* $(\cos^2\theta)^{k-1}(1-\cos^2\theta)$ from Lemma 9. After one cycle ($k=2$), small and nearly-orthogonal angles induce low forgetting, while intermediate angles are troublesome. However, as $k$ increases, the angle that maximizes forgetting goes to zero. Importantly, we find that the effect of task similarity depends on the number of cycles!

Figure 3($b$) demonstrates the worst-case analysis of Theorem 10. As tasks become *more* similar, *i.e.,* $\theta$ decreases, the initial forgetting is smaller, yet convergence is slower since the contraction is small at each iteration. Conversely, larger angles lead to larger initial forgetting but also to faster convergence due to a more significant contraction.

Our findings after seeing each task *once* (*i.e.,* $k=2$) resemble findings from Lee et al. [34] that intermediate task similarity causes the most forgetting (however, their setup and notion of similarity are different). Like we mentioned in Section 4.1, our analysis contradicts a corollary from Doan et al. [14] implying a higher risk of forgetting when two tasks are *more* aligned. This discrepancy stems from an upper bound in [14] being looser than the tight bounds we derive (see App B.1.3).

### 5.2. Main Result: Worst-case forgetting with $T \geq 3$ tasks

For the general cyclic case, we provide two upper bounds – a dimension-dependent bound, and more importantly, a dimension-independent one. Both follow a power-law *w.r.t.* the iteration number $k$.

**Theorem 11 (Worst-case forgetting when $T \geq 3$)** *For any number of tasks $T \geq 3$ under a cyclic ordering $\tau$, the worst-case forgetting after $k = nT \geq T^2$ iterations (i.e., $n \geq T$ cycles), is*

$$\frac{T^2}{24ek} \;\leq\; \sup_{\substack{S\in\mathcal{S}_{T\geq 3}:\\ \forall m:\,\mathrm{rank}(\boldsymbol{X}_m)\leq r_{\max}}} F_{\tau,S}(k) \;\leq\; \min\left\{\frac{T^2}{\sqrt{k}},\; \frac{T^2\,(d-r_{\max})}{2k}\right\}.$$

*Moreover, if the cyclic operator $\left(\boldsymbol{P}_T\cdots\boldsymbol{P}_1\right)$ from Eq. (10) is symmetric (e.g., in a* back-and-forth *setting where tasks $m$ and $(T-m+1)$ are identical $\forall m\in[T]$), then the worst-case forgetting is* sharply $\sup_{S\in\mathcal{S}_{T\geq 3}} F_{\tau,S}(k) = \Theta\left(T^2/k\right)$.

**Proof sketch for the upper bound.** We briefly portray our proof for the above result (given fully in App C.2). For brevity, denote the cyclic operator as $\boldsymbol{M} \triangleq \boldsymbol{P}_T\ldots\boldsymbol{P}_1$. Our proof revolves around the maximal decrease at the $n$th cycle, *i.e.,* $\Delta_n(\boldsymbol{u}) \triangleq \left\|\boldsymbol{M}^n\boldsymbol{u}\right\|^2 - \left\|\boldsymbol{M}^{n+1}\boldsymbol{u}\right\|^2$. We start by showing that the worst-case forgetting on the *first task* (see Eq. (10)) is upper bounded by the maximal decrease. That is,

$$\forall\boldsymbol{u}\in\mathbb{C}^d:\;\left\|(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{M}^n\boldsymbol{u}\right\|^2 \overset{(*)}{=} \left\|\boldsymbol{M}^n\boldsymbol{u}\right\|^2 - \left\|\boldsymbol{P}_1\boldsymbol{M}^n\boldsymbol{u}\right\|^2 \overset{(**)}{\leq} \left\|\boldsymbol{M}^n\boldsymbol{u}\right\|^2 - \left\|\boldsymbol{M}^{n+1}\boldsymbol{u}\right\|^2 = \Delta_n(\boldsymbol{u}),$$

where (*) stems from the idempotence of $\boldsymbol{P}_1$, and (**) is true since projections are non-expansive operators, meaning $\left\|\boldsymbol{P}_1\boldsymbol{M}^n\boldsymbol{u}\right\|^2 \geq \left\|\boldsymbol{P}_T\cdots\boldsymbol{P}_1\boldsymbol{M}^n\boldsymbol{u}\right\|^2 \triangleq \left\|\boldsymbol{M}^{n+1}\boldsymbol{u}\right\|^2$.

More generally, we show that $\left\|(\boldsymbol{I}-\boldsymbol{P}_m)\boldsymbol{M}^n\boldsymbol{u}\right\|^2 \leq m\Delta_n(\boldsymbol{u})$, yielding the overall bound:

$$\frac{1}{T}\sum_{m=1}^{T}\left\|(\boldsymbol{I}-\boldsymbol{P}_m)\boldsymbol{M}^n\right\|^2 \leq \frac{T-1}{2}\cdot\max_{\boldsymbol{u}:\|\boldsymbol{u}\|_2=1}\Delta_n(\boldsymbol{u})\,.$$

Then, we prove that $\max_{\boldsymbol{u}:\|\boldsymbol{u}\|_2=1}\Delta_n(\boldsymbol{u}) \leq 2\|\boldsymbol{M}^n-\boldsymbol{M}^{n+1}\| \leq 2\sqrt{T/n}$, using telescoping sums on elements of $\{\|\left(\boldsymbol{M}^t-\boldsymbol{M}^{t+1}\right)\boldsymbol{v}\|^2\}_{t=1}^n$. Finally, we prove $\max_{\boldsymbol{u}:\|\boldsymbol{u}\|_2=1}\Delta_n(\boldsymbol{u}) \leq \frac{d-r_{\max}}{n}$ by using telescoping sums on the traces of matrices $\left\{(\boldsymbol{M}^t)^\top\boldsymbol{M}^t-(\boldsymbol{M}^{t+1})^\top\boldsymbol{M}^{t+1}\right\}_{t=1}^n$. ■

## 6. Random task orderings

So far, we saw in Section 4 that arbitrary task orderings provide no convergence guarantees and might forget *catastrophically*. In Section 5 we saw that cyclic orderings *do not* suffer from catastrophic forgetting, since their forgetting converges to zero like a power law. We now analyze *random* task ordering, and show that they also have uniform (data-independent) convergence guarantees.

We consider a random task ordering $\tau$ that matches a uniform probability to any task at any iteration, *i.e.,* $\forall m\in[T]\,,\forall t\in\mathbb{N}^+$: $\Pr\left[\,\tau(t)=m\,\right]=1/T$. Below, we adjust the forgetting definitions in 3 and 4 to the random setting by defining the expected forgetting.

**Definition 12 (Expected forgetting of a task collection)** *After $k$ iterations, the* expected forgetting *on a specific task collection $S\in\mathcal{S}_T$ is defined as*

$$\bar{F}_{\tau,S}\left(k\right) \triangleq \mathop{\mathbb{E}}_{\tau}\left[F_{\tau,S}\left(k\right)\right] = \mathop{\mathbb{E}}_{\tau}\left[\frac{1}{k}\sum_{t=1}^{k}\left\|\boldsymbol{X}_{\tau(t)}\boldsymbol{w}_k-\boldsymbol{y}_{\tau(t)}\right\|^2\right].$$

Our main result in this section is a uniform bound on the expected forgetting under the uniform random task ordering. The proof is given in App D.

**Theorem 13 (Worst-case expected forgetting)** *Under the uniform i.i.d. task ordering $\tau$, the worst-case expected forgetting after $k$ iterations is*

$$\sup_{\substack{S\in\mathcal{S}_T:\\ \frac{1}{T}\sum_{m=1}^{T}\operatorname{rank}(\boldsymbol{X}_m)=r_{\mathrm{avg}}}} \bar{F}_{\tau,S}(k) \le \frac{9\,(d-r_{\mathrm{avg}})}{k}\,.$$

**Demonstration.** The following Figure 4 demonstrates the worst-case forgetting under cyclic task orderings (Theorem 11) and the worse-case expected forgetting under random orderings (Theorem 13). We consider a specific 2-d task collection $S$ consisting of $T = 128$ rank-one tasks. The collection is "adversarial" in the sense that its forgetting meets the *cyclic* setting's lower bound of Theorem 11 (dashed orange).

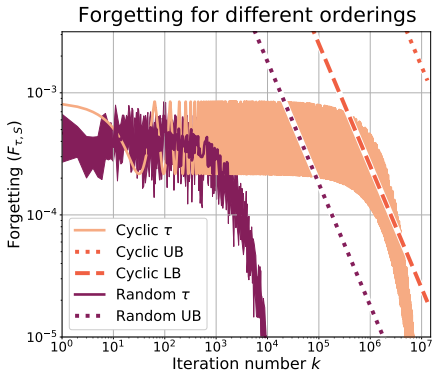



Figure 4: **Demonstrating the bounds from Thm 11 and 13.** The solid orange curve shows the actual forgetting of the cyclic deterministic ordering (the oscillations are formed, naturally, by the task cycles). The purple solid curve shows the expected forgetting of the random ordering (averaged over 5 seeds). The purple band indicates one standard deviation (over the 5 seeds). Also plotted are the corresponding upper bounds of both settings (dotted).
Notice how a random ordering behaves better than a cyclic one, both practically and analytically.

**Remark 14 (Last iterate SGD results)** *In the special case of rank-one tasks, the iterative update rule in Eq. (2) reduces to a* single *SGD step with a $\|\boldsymbol{x}_{\tau(t)}\|^{-2}$ step size, and the above bound becomes $\mathcal{O}(d/k)$. A very recent work [71] derived a* dimension-independent *bound of $\mathcal{O}(\ln k/k)$ for the last iterate of SGD with a constant step size in a similar rank-1 setting. Although our step size is different, we believe our bound's dependence on the dimension (Theorem 13) can probably be lifted.*

**Remark 15 (Average iterate analysis)** *We note that all our results become much tighter when working with the* average iterate $\overline{\boldsymbol{w}}_k \triangleq \frac{1}{k}\sum_{t=1}^{k}\boldsymbol{w}_t$ *instead of the last one $\boldsymbol{w}_k$. Then, it is easier to prove tighter bounds, i.e., $\mathcal{O}(T^2/k)$ for cyclic orderings and $\mathcal{O}(1/k)$ for random ones. App E provides a proof sketch.*

## 7. Related Work

Many practical methods have been proposed over the last decade to address catastrophic forgetting in continual learning. Clearly we cannot cover all methods here, but refer the reader to recent surveys (*e.g.,* [53, 55]) and to App F. In a nutshell, algorithmic approaches to continual learning can be roughly divided into three: *Memory-based replay approaches* (*e.g.,* [59, 66]); *Regularization approaches* (*e.g.,* [7, 31, 35, 36, 79]); and *Parameter isolation approaches* (*e.g.,* [1, 38, 76]).

**Theoretical results for catastrophic forgetting.** We briefly discuss few related theoretical works. Doan et al. [14] analyzed forgetting in linear regression (or, more generally, in the NTK regime). Their derivations largely depend on a matrix that captures principal angles between tasks, very much like we do, but they focus on the smallest angle, *i.e.,* the Dixmier angle [13]. They conclude that increased task similarity leads to more forgetting. In contrast, our $T=2$ analysis reveals the precise role of task similarity at different training stages (see Figure 3). Their work [14] and others [6, 16] also analyzed the OGD algorithm, deriving generalization guarantees and improved variants.

Lee et al. [34] considered a teacher-student two-layer setup with 2 tasks, each having its own last layer. They analyzed a different notion of task similarity (*i.e.,* teacher similarity) and studied how it affects forgetting in student models. They found that intermediate (teacher-)similarity leads to the greatest forgetting. In our linear setup, we find that when two consecutive tasks are shown *once* (*i.e.,* in a single cycle), intermediate principal angles cause higher forgetting (see Figure 3(*a*)). Asanuma et al. [4] considered a linear regression setting, as we do, yet only for 2 tasks having a different teacher. They assumed specific axis-aligned input distributions, and studied forgetting in the infinite-dimension limit, taking into account both the input-space and weight-space similarities. Under a "shared" teacher, their results imply zero forgetting due to their assumptions on inputs.

**Alternating projections (AP).** Throughout our paper, we lay out connections between continual learning and the AP literature (see survey by [20]). Our cyclic setting (Section 5) is a special case of cyclic AP (Halperin [25], Von Neumann [72]). Studies from this field often bound the convergence to the subspace intersection either asymptotically or using notions of generalized angles between subspaces (*e.g.,* [12, 30, 51]). Most results in this area are uninformative at the worst case.

**Kaczmarz method.** Our fitting procedure (described in Section 2.3 and 3) shares a great resemblance with Kaczmarz methods for solving a system $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$. The "basic" method [29] corresponds to tasks of rank $r=1$, while the block variant [15] corresponds to tasks having $r\geq 2$. Traditionally, these methods used a deterministic cyclic ordering, like we do in Section 5. There is also a randomized variant [44, 68, 75], related to our random task ordering in Section 6, which often achieves faster convergence both empirically and theoretically [69]. Studies of these methods often analyze convergence to a feasible set using the spectrum of $\boldsymbol{A}$ (*e.g.,* its condition number; see [24, 46, 52]).

**Our convergence analysis is inherently different.** Most convergence results from the AP and Kaczmarz research areas concentrate on bounding the convergence to the subspace intersection. These results *are* valuable for the continual setting, *e.g.,* they help us bound $\|\boldsymbol{w}_k-\boldsymbol{w}^\star\|$ in Theorem 8, which clearly upper bounds the forgetting. However, most convergence results in these areas depend on the task specifics, precluding any informative *worst-case* analysis. In contrast, we analyze the convergence of the *forgetting* (*i.e.,* projection residuals), allowing us to derive *uniform* worst-case guarantees. In Section 5.1, we demonstrate these differences by comparing the two quantities. Finally, we believe our bounds are novel and can provide a new perspective in these areas, especially in the cyclic block Kaczmarz setting [15, 45], *i.e.,* that *worst-case* last-iterate analyses become feasible when examining the residual convergence instead of the distance from $\boldsymbol{w}^\star$.

**Normalized Least-Mean-Square (NLMS).** The NLMS algorithm is a popular choice for adaptive filtering. In its basic form, it fits one random sample at each iteration using update rules identical to those of our stochastic setting and the randomized Kaczmarz method. There are known convergence guarantees for the single-sample algorithm [67] and its multiple-samples counterpart, the APA [61], but these are proven only under limiting assumptions on the input signals.

**Other optimization methods.** It is interesting to note that when fitting $T$ tasks in a cyclic or random ordering, our update rule in Eq. (2) is equivalent to dual block coordinate descent on the dual of Eq. (3), *i.e.,* $\operatorname{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^N}\left(\frac{1}{2}\left\|\boldsymbol{X}_{1:T}^\top\boldsymbol{\alpha}\right\|^2-\boldsymbol{y}_{1:T}^\top\boldsymbol{\alpha}\right)$, where $\boldsymbol{X}_{1:T}\in\mathbb{R}^{N\times d}$ and $\boldsymbol{y}_{1:T}\in\mathbb{R}^N$ (for $N=\sum_{m=1}^T n_m$) are the concatenation of the data matrices and label vectors from all tasks. The primal variable $\boldsymbol{w}$ and the dual one $\boldsymbol{\alpha}$ are related through $\boldsymbol{w}=\boldsymbol{X}_{1:T}^\top\boldsymbol{\alpha}$. Shalev-Shwartz and Zhang [64] analyzed the stochastic dual coordinate ascent method for minimizing a *regularized* loss, and proved convergence rates for the primal suboptimality. However, when the regularization parameter vanishes, as in our case, their bounds tend to infinity. Others [69] analyzed the convergence of coordinate descent on quadratic functions like the dual above, and derived data-dependent bounds.

The stochastic proximal point algorithm (SPPA) [8, 54, 60] follows an update rule (for a step size $\eta$) of $\boldsymbol{w}_t=\operatorname{argmin}_{\boldsymbol{w}}\left(\frac{1}{2}(\boldsymbol{X}_{\tau(t)}\boldsymbol{w}-\boldsymbol{y}_{\tau(t)})^2+\frac{1}{2\eta}\left\|\boldsymbol{w}-\boldsymbol{w}_{t-1}\right\|^2\right)$, equivalent when $\eta\to\infty$ to our rule in Eq. (1). As far as we know, no uniform convergence-rates were previously proved for SPPA.

Minimizing forgetting over $T$ tasks can also be seen as a finite sum minimization problem [47, 74]. Many algorithms have been proposed for this problem, some of which achieve better convergence rates than ours using additional *memory* (*e.g.,* for storing gradients [3, 28]). We derive forgetting convergence rates for the fitting procedure in Eq. (1) that is equivalent to running (S)GD to convergence for each presented task, which is a natural choice for continual learning settings.

**Forgetting vs. regret.** Our Definition 3 of the forgetting should not be confused with the notion of regret, mainly studied in the context of online learning (*e.g.,* [65]). Specifically, using our notations, the average regret after $k$ iteration is $\frac{1}{k}\sum_{t=1}^k\|\boldsymbol{X}_{\tau(t)}\boldsymbol{w}_{t-1}-\boldsymbol{y}_{\tau(t)}\|^2$. Comparing the two quantities, we note that forgetting quantifies degradation on *previous* tasks, while regret captures the ability to predict *future* tasks. To illustrate the differences, consider a finite sequence of orthogonal tasks. In this case, the forgetting is 0 (as discussed in Section 4.1), but the regret is large. Conversely, given a task sequence like in Figure 2, when $k\to\infty$ the regret vanishes while forgetting goes to 1 (see Theorem 7). Nevertheless, in the cyclic and random settings, both quantities will go to 0 when $k\to\infty$ since we converge to an offline solution. However, their rates of convergence may differ.

## 8. Conclusion

Catastrophic forgetting is not yet fully understood theoretically. Therefore, one must first study it in the simplest model exhibiting this phenomenon — linear regression. In this setting, we provide sharp uniform worst-case bounds. Most of our analysis does not depend on task specifics, but only on the number of tasks $T$, the number of iterations $k$, and the task ordering. On the one hand, we prove that for an arbitrary ordering, forgetting *can* be catastrophic. On the other hand, for cyclic orderings, we prove forgetting *cannot* be catastrophic and that even in the worst-case it vanishes at most as $\frac{T^2}{\sqrt{k}}$ or $\frac{T^2d}{k}$. Lastly, we prove worst-case bounds for random orderings, independent of $T$.

Our bounds complement existing Kaczmarz and alternating projection bounds, which focus on convergence to a feasible set. Unlike ours, their bounds strictly depend on task specifics (*e.g.,* principal angles or condition numbers) and become trivial in worst-case analysis.

There are many intriguing directions for future research. These include extending our results to the non-realizable case, analyzing forgetting in classification tasks, and deriving optimal presentation task orderings. Lastly, one can try to extend our results to more complex models (*e.g.,* deep models) and other training algorithms. We hope a thorough theoretical understanding of catastrophic forgetting, will facilitate the development of new practical methods to alleviate it.

## Acknowledgments

We would like to thank Suriya Gunasekar (Microsoft Research) for her insightful comments and suggestions. We would also like to thank Simeon Reich (Technion), Rafal Zalas (Technion), and Timur Oikhberg (Univ. of Illinois) for their fruitful discussions. R. Ward was partially supported by AFOSR MURI FA9550-19-1-0005, NSF DMS 1952735, NSF HDR1934932, and NSF 2019844. N. Srebro was partially supported by NSF IIS award #1718970 and the NSF-Simons Funded Collaboration on the Mathematics of Deep Learning. D. Soudry was supported by the Israel Science Foundation (Grant No. 1308/18) and the Israel Innovation Authority (the Avatar Consortium).

## References

[1] Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. Expert gate: Lifelong learning with a network of experts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3366–3375, 2017. (cited on p. 11)

[2] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 139–154, 2018. (cited on p. 51)

[3] Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods, 2018. (cited on p. 13)

[4] Haruka Asanuma, Shiro Takagi, Yoshihiro Nagano, Yuki Yoshida, Yasuhiko Igarashi, and Masato Okada. Statistical mechanical analysis of catastrophic forgetting in continual learning with teacher and student networks. *Journal of the Physical Society of Japan*, 90(10):104001, Oct 2021. (cited on p. 1, 12)

[5] Christian Bargetz, Jona Klemenc, Simeon Reich, and Natalia Skorokhod. On angles, projections and iterations. *Linear Algebra and its Applications*, 603:41–56, 2020. (cited on p. 6, 22)

[6] Mehdi Abbana Bennani, Thang Doan, and Masashi Sugiyama. Generalisation guarantees for continual learning with orthogonal gradient descent. *arXiv preprint arXiv:2006.11942*, 2020. (cited on p. 1, 12)

[7] Frederik Benzing. Unifying regularisation methods for continual learning. *AISTATS*, 2022. (cited on p. 11, 51)

[8] Dimitri P. Bertsekas. Incremental proximal methods for large scale convex optimization. *Mathematical Programming*, 129:163–195, 2011. (cited on p. 13)

[9] Ake Björck and Gene H Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of computation*, 27(123):579–594, 1973. (cited on p. 6, 22)

[10] Matthias Delange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Greg Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. (cited on p. 51)

[11] Frank Deutsch. The angle between subspaces of a hilbert space. In *Approximation theory, wavelets and applications*, pages 107–130. Springer, 1995. (cited on p. 6)

[12] Frank Deutsch and Hein Hundal. The rate of convergence for the method of alternating projections, ii. *Journal of Mathematical Analysis and Applications*, 205(2):381–405, 1997. (cited on p. 12)

[13] Jacques Dixmier. Étude sur les variétés et les opérateurs de julia, avec quelques applications. *Bulletin de la Société Mathématique de France*, 77:11–101, 1949. (cited on p. 12, 22, 28)

[14] Thang Doan, Mehdi Abbana Bennani, Bogdan Mazoure, Guillaume Rabusseau, and Pierre Alquier. A theoretical analysis of catastrophic forgetting through the ntk overlap matrix. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 1072–1080, 2021. (cited on p. 1, 4, 7, 9, 12, 28)

[15] Tommy Elfving. Block-iterative methods for consistent and inconsistent linear equations. *Numerische Mathematik*, 35(1):1–12, 1980. (cited on p. 2, 4, 12)

[16] Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3762–3773. PMLR, 2020. (cited on p. 12, 51)

[17] fedja (https://mathoverflow.net/users/1131/fedja). Bounding the decrease after applying a contraction operator $n$ vs $n+1$ times. MathOverflow, 2021. URL:https://mathoverflow.net/q/408834. (cited on p. 39)

[18] Kurt Friedrichs. On certain inequalities and characteristic value problems for analytic functions and for functions of two variables. *Transactions of the American Mathematical Society*, 41(3):321–364, 1937. (cited on p. 8, 22)

[19] Rong Ge, Chi Jin, Praneeth Netrapalli, Aaron Sidford, et al. Efficient algorithms for large-scale generalized eigenvector computation and canonical correlation analysis. In *International Conference on Machine Learning*, pages 2741–2750. PMLR, 2016. (cited on p. 6)

[20] Omer Ginat. The method of alternating projections. *arXiv preprint arXiv:1809.05858*, 2018. (cited on p. 12)

[21] Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013. (cited on p. 1, 51)

[22] Suriya Gunasekar, Jason D. Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *ICML*, 2018. (cited on p. 4)

[23] Shay Hacohen-Gourgy, Luis Pedro García-Pintos, Leigh S Martin, Justin Dressel, and Irfan Siddiqi. Incoherent qubit control using the quantum zeno effect. *Physical review letters*, 120 (2):020505, 2018. (cited on p. 29)

[24] Jamie Haddock and Anna Ma. Greed works: An improved analysis of sampling kaczmarz–motzkin. *SIAM Journal on Mathematics of Data Science*, 3(1):342–368, 2021. (cited on p. 12)

[25] Israel Halperin. The product of projection operators. *Acta Sci. Math.(Szeged)*, 23(1):96–99, 1962. (cited on p. 2, 12)

[26] Simon Haykin. *Adaptive filter theory*. Prentice Hall, 2002. (cited on p. 4)

[27] Stanislaw Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three Factors Influencing Minima in SGD. In *International Conference of Artificial Neural Networks (ICANN)*, 2018. (cited on p. 51)

[28] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013. (cited on p. 13)

[29] S Kaczmarz. Angenaherte auflosung von systemen linearer glei-chungen. *Bull. Int. Acad. Pol. Sic. Let., Cl. Sci. Math. Nat.*, pages 355–357, 1937. (cited on p. 2, 4, 12)

[30] Selahattin Kayalar and Howard L Weinert. Error bounds for the method of alternating projections. *Mathematics of Control, Signals and Systems*, 1(1):43–59, 1988. (cited on p. 2, 8, 12)

[31] 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. (cited on p. 11, 51)

[32] Jeremias Knoblauch, Hisham Husain, and Tom Diethe. Optimal continual learning has perfect memory and is np-hard. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, pages 5327–5337, 2020. (cited on p. 1)

[33] Andrew Knyazev and Merico Argentati. Majorization for changes in angles between subspaces, ritz values, and graph laplacian spectra. *SIAM J. Matrix Analysis Applications*, 29: 15–32, 01 2006. doi: 10.1137/060649070. (cited on p. 22)

[34] Sebastian Lee, Sebastian Goldt, and Andrew Saxe. Continual learning in the teacher-student setup: Impact of task similarity. In *International Conference on Machine Learning*, pages 6109–6119. PMLR, 2021. (cited on p. 1, 9, 12)

[35] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. (cited on p. 11, 51)

[36] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30:6467–6476, 2017. (cited on p. 11, 51)

[37] Ekdeep Singh Lubana, Puja Trivedi, Danai Koutra, and Robert P. Dick. How do quadratic regularizers prevent catastrophic forgetting: The role of interpolation. In *ICML Workshop on Theory and Foundations of Continual Learning*, 2021. (cited on p. 51)

[38] 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*, pages 7765–7773, 2018. (cited on p. 11, 51)

[39] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989. (cited on p. 1)

[40] Carl D Meyer. *Matrix analysis and applied linear algebra*, volume 71. Siam, 2000. (cited on p. 21, 22, 24)

[41] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Razvan Pascanu, and Hassan Ghasemzadeh. Understanding the role of training regimes in continual learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7308–7320. Curran Associates, Inc., 2020. (cited on p. 51)

[42] Seyed Iman Mirzadeh, Arslan Chaudhry, Huiyi Hu, Razvan Pascanu, Dilan Gorur, and Mehrdad Farajtabar. Wide neural networks forget less catastrophically. *ICML*, 2022. (cited on p. 51)

[43] Md Sarowar Morshed, Md Saiful Islam, and Md Noor-E-Alam. Accelerated sampling kaczmarz motzkin algorithm for the linear feasibility problem. *Journal of Global Optimization*, 77 (2):361–382, 2020. (cited on p. 49)

[44] Deanna Needell. Randomized kaczmarz solver for noisy linear systems. *BIT Numerical Mathematics*, 50(2):395–403, 2010. (cited on p. 12)

[45] Deanna Needell and Joel A Tropp. Paved with good intentions: analysis of a randomized block kaczmarz method. *Linear Algebra and its Applications*, 441:199–221, 2014. (cited on p. 12)

[46] Deanna Needell, Rachel Ward, and Nati Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *Advances in neural information processing systems*, 27:1017–1025, 2014. (cited on p. 12)

[47] Geoffrey Negiar, Gideon Dresdner, Alicia Tsai, Laurent El Ghaoui, Francesco Locatello, Robert Freund, and Fabian Pedregosa. Stochastic frank-wolfe for constrained finite-sum minimization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7253–7262, 2020. (cited on p. 13)

[48] Anupan Netyanun and Donald C Solmon. Iterated products of projections in hilbert space. *The American Mathematical Monthly*, 113(7):644–648, 2006. (cited on p. 37)

[49] Cuong V Nguyen, Yingzhen Li, Thang D Bui, and Richard E Turner. Variational continual learning. In *International Conference on Learning Representations*, 2018. (cited on p. 51)

[50] Timur Oikhberg. Products of orthogonal projections. *Proceedings of the American Mathematical Society*, 127(12):3659–3669, 1999. (cited on p. 39)

[51] Izhar Oppenheim. Angle criteria for uniform convergence of averaged projections and cyclic or random products of projections. *Israel Journal of Mathematics*, 223(1):343–362, 2018. (cited on p. 6, 12)

[52] Peter Oswald and Weiqi Zhou. Convergence analysis for kaczmarz-type methods in a hilbert space framework. *Linear Algebra and its Applications*, 478:131–161, 2015. (cited on p. 12)

[53] 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. (cited on p. 1, 11)

[54] Andrei Patrascu and Ion Necoara. Nonasymptotic convergence of stochastic proximal point methods for constrained convex optimization. *Journal of Machine Learning Research*, page 7204–7245, 2017. (cited on p. 13)

[55] Haoxuan Qu, Hossein Rahmani, Li Xu, Bryan Williams, and Jun Liu. Recent advances of continual learning in computer vision: An overview, 2021. (cited on p. 11)

[56] Vinay Venkatesh Ramasesh, Ethan Dyer, and Maithra Raghu. Anatomy of catastrophic forgetting: Hidden representations and task semantics. In *International Conference on Learning Representations*, 2020. (cited on p. 1, 7)

[57] Roger Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285, 1990. (cited on p. 1)

[58] Ice B Risteski and Kostadin G Trencevski. Principal values and principal subspaces of two subspaces of vector spaces with inner product. *Beiträge zur Algebra und Geometrie*, 42(1): 289–300, 2001. (cited on p. 24)

[59] Anthony Robins. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146, 1995. (cited on p. 11, 51)

[60] E. Ryu and S. Boyd. Stochastic proximal iteration: A non-asymptotic improvement upon stochastic gradient descent. *Author website*, 2016. (cited on p. 13)

[61] Sundar G Sankaran and AA Louis Beex. Convergence behavior of affine projection algorithms. *IEEE Transactions on Signal Processing*, 48(4):1086–1096, 2000. (cited on p. 12, 44)

[62] Jeffrey C Schlimmer and Douglas Fisher. A case study of incremental concept induction. In *AAAI*, volume 86, pages 496–501, 1986. (cited on p. 1)

[63] Jonathan Schwarz, Siddhant Jayakumar, Razvan Pascanu, Peter Latham, and Yee Teh. Powerpropagation: A sparsity inducing weight reparameterisation. *Advances in Neural Information Processing Systems*, 34, 2021. (cited on p. 51)

[64] Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss. *J. Mach. Learn. Res.*, 14(1):567–599, 2013. (cited on p. 13)

[65] Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012. (cited on p. 13)

[66] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017. (cited on p. 11, 51)

[67] Dirk TM Slock. On the convergence behavior of the lms and the normalized lms algorithms. *IEEE Transactions on Signal Processing*, 41(9):2811–2825, 1993. (cited on p. 4, 12, 44)

[68] Thomas Strohmer and Roman Vershynin. A randomized kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009. (cited on p. 12)

[69] Ruoyu Sun and Yinyu Ye. Worst-case complexity of cyclic coordinate descent: $O(n^2)$ gap with randomized version. *Mathematical Programming*, pages 1–34, 2019. (cited on p. 12, 13)

[70] Sebastian Thrun and Tom M Mitchell. Lifelong robot learning. *Robotics and autonomous systems*, 15(1-2):25–46, 1995. (cited on p. 1)

[71] Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Last iterate convergence of SGD for least-squares in the interpolation regime. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. (cited on p. 11, 49)

[72] John Von Neumann. On rings of operators. reduction theory. *Annals of Mathematics*, pages 401–485, 1949. (cited on p. 12)

[73] Daphna Weinshall and Dan Amir. Theory of curriculum learning, with convex loss functions. *Journal of Machine Learning Research*, 21(222):1–19, 2020. (cited on p. 51)

[74] Blake E. Woodworth and Nati Srebro. Tight complexity bounds for optimizing composite objectives. In *NIPS*, pages 3639–3647, 2016. (cited on p. 13)

[75] Hua Xiang and Lin Zhang. Randomized iterative methods with alternating projections. *arXiv preprint arXiv:1708.09845*, 2017. (cited on p. 12)

[76] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. In *International Conference on Learning Representations*, 2018. (cited on p. 11, 51)

[77] Eduardo H Zarantonello. Projections on convex sets in hilbert space and spectral theory: Part i. projections on convex sets: Part ii. spectral theory. In *Contributions to nonlinear functional analysis*, pages 237–424. Elsevier, 1971. (cited on p. 21)

[78] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pages 3987–3995. PMLR, 2017. (cited on p. 51)

[79] Chen Zeno, Itay Golan, Elad Hoffer, and Daniel Soudry. Task-Agnostic Continual Learning Using Online Variational Bayes With Fixed-Point Updates. *Neural Computation*, 33(11): 3139–3177, Oct 2021. (cited on p. 11, 51)

[80] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017. (cited on p. 4)

## Appendix A. Preliminary notations and lemmas

### A.1. Additional notations for the appendices

We start by adding several notations to those we defined in Section 2.

1. Like in the main text, $\|\cdot\|$ denotes either the Euclidean $\ell^2$-norm of a vector or the spectral norm of a matrix.
2. We denote the **conjugate transpose** (*i.e.,* Hermitian transpose) of complex vectors by $\boldsymbol{v}^*$.
3. We use the **singular value decomposition** of *real* matrices, *i.e.,*

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \in \mathbb{R}^{N\times d},$$

where $\boldsymbol{U} \in \mathbb{R}^{N\times N}$, $\boldsymbol{V} \in \mathbb{R}^{d\times d}$ are two orthonormal matrices and $\boldsymbol{\Sigma} \in \mathbb{R}^{N\times d}$ has the same rank and dimensions as $\boldsymbol{X}$. We assume w.l.o.g. that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\operatorname{rank}\boldsymbol{X}} \geq \underbrace{0 = \ldots = 0}_{(d-\operatorname{rank}\boldsymbol{X}) \text{ times}}$. In addition, we often decompose $\boldsymbol{V}$ into $\boldsymbol{V} = \big[\ \underbrace{\boldsymbol{V}^r}_{d\times\operatorname{rank}\boldsymbol{X}}\ \big|\ \underbrace{\boldsymbol{V}^\perp}_{d\times(d-\operatorname{rank}\boldsymbol{X})}\ \big]$ to distinguish between the columns of $\boldsymbol{V}$ that span the range of $\boldsymbol{X}$ from the ones that span its orthogonal complement.

### A.2. Useful general properties

Following are several useful inequalities and properties that will facilitate our proofs. We only work with *real* matrices, so we often use the transpose and the Hermitian transpose interchangeably.

**Lemma 16** *For any $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T \in \mathbb{C}^d$, it holds that $\|\boldsymbol{x}_1 + \ldots + \boldsymbol{x}_T\|^2 \leq T\Big(\|\boldsymbol{x}_1\|^2 + \cdots + \|\boldsymbol{x}_T\|^2\Big)$.*

**Proof** Notice that

$$0 \leq \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 = \|\boldsymbol{x}_i\|^2 - 2\mathrm{Re}\left(\boldsymbol{x}_i^*\boldsymbol{x}_j\right) + \|\boldsymbol{x}_j\|^2 \iff 2\mathrm{Re}\left(\boldsymbol{x}_i^*\boldsymbol{x}_j\right) \leq \|\boldsymbol{x}_i\|^2 + \|\boldsymbol{x}_j\|^2 .$$

Now, we see use the Cauchy-Schwarz inequality and show

$$\|\boldsymbol{x}_1 + \cdots + \boldsymbol{x}_T\|^2 = \sum_{i=1}^T \|\boldsymbol{x}_i\|^2 + \sum_{i>j} 2\mathrm{Re}\left(\boldsymbol{x}_i^*\boldsymbol{x}_j\right) \leq \sum_{i=1}^T \|\boldsymbol{x}_i\|^2 + \sum_{i>j}\Big(\|\boldsymbol{x}_i\|^2 + \|\boldsymbol{x}_j\|^2\Big)$$

$$[\text{each appears } T \text{ times}] = T\sum_{i=1}^T \|\boldsymbol{x}_i\|^2 = T\left(\|\boldsymbol{x}_1\|^2 + \cdots + \|\boldsymbol{x}_T\|^2\right) .$$

■

**Property 1 (Spectral norm properties)** *Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{B} \in \mathbb{R}^{n\times p}$. Then, the spectral norm holds the following properties:*

1. ***Definition.*** $\|\boldsymbol{A}\| = \|\boldsymbol{A}\|_2 \triangleq \max_{\boldsymbol{v}\in\mathbb{R}^n\setminus\{0\}} \frac{\|\boldsymbol{A}\boldsymbol{v}\|_2}{\|\boldsymbol{v}\|_2} = \max_{\substack{\boldsymbol{v}\in\mathbb{R}^n:\\ \|\boldsymbol{v}\|_2=1}} \|\boldsymbol{A}\boldsymbol{v}\|_2 = \sigma_1(\boldsymbol{A})$ ;
2. ***Invariance to transposition.*** $\|\boldsymbol{A}\| = \|\boldsymbol{A}^\top\|$ ;
3. ***Triangle inequality.*** $\|\boldsymbol{A}+\boldsymbol{B}\| \le \|\boldsymbol{A}\| + \|\boldsymbol{B}\|$ ;
4. ***Squared norm of matrix sum.*** $\|\boldsymbol{A}_1 + \cdots + \boldsymbol{A}_T\|^2 \le T\sum_{i=1}^T \|\boldsymbol{A}_i\|^2$ ;
5. ***Multiplicative norm inequality.*** $\|\boldsymbol{A}\boldsymbol{B}\| \le \|\boldsymbol{A}\|\,\|\boldsymbol{B}\|$ ;
6. ***Invariance to rotations.*** *Let $\boldsymbol{V}_1 \in \mathbb{R}^{m'\times m}$ and $\boldsymbol{V}_2 \in \mathbb{R}^{n'\times n}$ be matrices with orthonormal* columns *($n' \ge n, m' \ge m$). Then, $\|\boldsymbol{V}_1\boldsymbol{A}\| = \|\boldsymbol{A}\| = \|\boldsymbol{A}\boldsymbol{V}_2^\top\| = \|\boldsymbol{V}_1\boldsymbol{A}\boldsymbol{V}_2^\top\|$.

See Chapter 5.2 in Meyer [40] for the proofs and for more such properties. The squared norm inequality follows immediately from Lemma 16 and the definition of the spectral norm.

**Property 2 (Orthogonal projection properties)** *Let $\boldsymbol{P}$ be a* real *orthogonal-projection linear-operator that projects onto a linear subspace $H \subseteq \mathbb{R}^d$. Let $\boldsymbol{v} \in \mathbb{R}^d$ be an arbitrary vector. Then, $\boldsymbol{P}$ holds the following properties:*

1. ***Geometric definition.*** $\|\boldsymbol{v} - \boldsymbol{P}\boldsymbol{v}\| = \inf_{\boldsymbol{x}\in H} \|\boldsymbol{v}-\boldsymbol{x}\|$ ;
2. ***Symmetry.*** $\boldsymbol{P} = \boldsymbol{P}^\top$ ;
3. ***Idempotence.*** $\boldsymbol{P}^2 = \boldsymbol{P}$ ;
4. *$\boldsymbol{I}-\boldsymbol{P}$ is also a projection operator, projecting onto the subspace orthogonal to $H$. Consequentially, $(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{P} = \boldsymbol{0}$ ;*
5. *Using the above, we get $\|(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{v}\|^2 = \boldsymbol{v}^*(\boldsymbol{I}-\boldsymbol{P})^2\boldsymbol{v} = \boldsymbol{v}^*(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{v} = \|\boldsymbol{v}\|^2 - \|\boldsymbol{P}\boldsymbol{v}\|^2$ ;*
6. ***Contraction.*** *$\|\boldsymbol{P}\boldsymbol{v}\| \le \|\boldsymbol{v}\|$, holding in equality if and only if $\boldsymbol{P}\boldsymbol{v} = \boldsymbol{v}$ ;*
7. ***Singular values.*** *All the singular values of $\boldsymbol{P}$ are in $\{0,1\}$, implying that $\|\boldsymbol{P}\| = 1 \iff \boldsymbol{P} \neq \boldsymbol{0}$.*

See Zarantonello [77] for the proofs and for more properties.

The next corollary stems directly from the properties above (see Chapter 5.13 in Meyer [40]).

**Corollary 17 (Projection onto solution spaces)** *Let $\boldsymbol{X}_m$ be the matrix of task $m$. Then, the projection onto its solution space, or equivalently onto its null space, is given by: $\boldsymbol{P}_m = \boldsymbol{I} - \boldsymbol{X}_m^+\boldsymbol{X}_m = \boldsymbol{V}_m\left(\boldsymbol{I} - \boldsymbol{\Sigma}_m^+\boldsymbol{\Sigma}_m\right)\boldsymbol{V}_m^\top = \boldsymbol{V}_m^\perp\boldsymbol{V}_m^{\perp\top}$.*

### A.3. Principal angles between two tasks

We continue our discussion from Section 3 and present principal angles more thoroughly for completeness. We mostly follow the definitions in [5, 9, 40].

**Definition 18 (Principal angles between two tasks)** *Let $\boldsymbol{X}_1 \in \mathbb{R}^{n_1 \times d}, \boldsymbol{X}_2 \in \mathbb{R}^{n_2 \times d}$ be two data matrices, corresponding to two given tasks. Let $r_{\min} = \min(\mathrm{rank}(\boldsymbol{X}_1), \mathrm{rank}(\boldsymbol{X}_2))$ be their minimal rank. The principal angles $\theta_1, \ldots, \theta_{r_{\min}} \in [0, \pi/2]$ between the two tasks, i.e., the angles between the row spaces of $\boldsymbol{X}_1, \boldsymbol{X}_2$, are recursively defined by*

$$\begin{aligned} \boldsymbol{u}_i, \boldsymbol{v}_i = \operatorname{argmax} & |\boldsymbol{u}_i^* \boldsymbol{v}_i| \\ & \text{s.t. } \boldsymbol{u}_i \in \mathrm{range}(\boldsymbol{X}_1^\top),\ \boldsymbol{v}_i \in \mathrm{range}(\boldsymbol{X}_2^\top),\ \ \|\boldsymbol{u}_i\| = \|\boldsymbol{v}_i\| = 1, \\ & \quad \forall j \in [i-1]\colon\ \boldsymbol{u}_i \perp \boldsymbol{u}_j,\ \boldsymbol{v}_i \perp \boldsymbol{v}_j \\ \theta_i = \arccos & |\boldsymbol{u}_i^* \boldsymbol{v}_i| \ . \end{aligned}$$

Notice that according to our definition, the principal angles hold $\pi/2 \geq \theta_{r_{\min}} \geq \cdots \geq \theta_1 \geq 0$ . Important for our analysis is the fact that unlike the principal *vectors* $\{\boldsymbol{u}_i, \boldsymbol{v}_i\}_i$, the principal *angles* between two subspaces are uniquely defined [9]. Two fundamental principal angles in the field of alternating projections are the minimal principal angle, *i.e.,* the Dixmier angle [13]; and the minimal non-zero principal angle, *i.e.,* the Friedrichs angle [18].

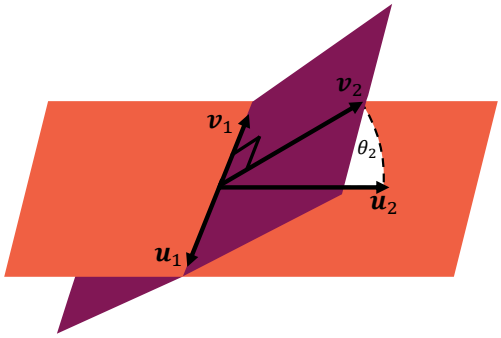



Figure 5: **Principal angles between subspaces.** Since the two hyperplanes share an intersecting direction, $\boldsymbol{u}_1, \boldsymbol{v}_1$ are chosen inside this intersection, meaning that $\theta_1 = \arccos|\boldsymbol{u}_1^* \boldsymbol{v}_1| = \arccos(1) = 0$. In this case, $\theta_1$ is the Dixmier angle. From the remaining directions that are orthogonal to $\boldsymbol{u}_1, \boldsymbol{v}_1$, the recursive definition chooses two unit vectors $\boldsymbol{u}_2, \boldsymbol{v}_2$, forming a non-zero angle $\theta_2 = \arccos|\boldsymbol{u}_2^* \boldsymbol{v}_2|$. In this case, $\theta_2$ is the Friedrichs angle.

**Claim 19 (Principal angles between "equivalent" subspaces)** *Let $(\boldsymbol{X}_1, \boldsymbol{y}_1), (\boldsymbol{X}_2, \boldsymbol{y}_2)$ be two tasks. Then, the sets of* non-zero *principal angles between the following pairs of subspaces, are all the same:*

1. *The data row-spaces, i.e.,* $\mathrm{range}(\boldsymbol{X}_1^\top)$ *and* $\mathrm{range}(\boldsymbol{X}_2^\top)$ *;*
2. *The null spaces, i.e.,* $\mathrm{null}(\boldsymbol{X}_1)$ *and* $\mathrm{null}(\boldsymbol{X}_2)$ *;*
3. *The affine solution spaces, i.e.,* $\mathcal{W}_1 = \boldsymbol{w}^\star + \mathrm{null}(\boldsymbol{X}_1)$ *and* $\mathcal{W}_2 = \boldsymbol{w}^\star + \mathrm{null}(\boldsymbol{X}_1)$ .

**Proof** The equivalence between (1) and (2) is proven by Theorem 2.7 of Knyazev and Argentati [33] (stating that the *non-zero* principal angles between two subspaces are essentially the same as those between their orthogonal complements). Moreover, since the solution spaces $\mathcal{W}_1, \mathcal{W}_2$ are merely affine subspaces of the null spaces themselves, they induce the exact same principal angles.

It *is* possible however that these pairs of subspaces *do* have a different amount of zero principal angles between them, but this generally does not interfere with our analyses in the appendices. ■

We now use principal angles to prove a lemma that will facilitate our proofs for Sections 4 and 5.

**Lemma 20** *Let $\boldsymbol{X}_1, \boldsymbol{X}_2$ be two data matrices of two tasks and let the corresponding orthogonal projections onto their null spaces be $\boldsymbol{P}_1 = \boldsymbol{I} - \boldsymbol{X}_1^+ \boldsymbol{X}_1,\ \boldsymbol{P}_2 = \boldsymbol{I} - \boldsymbol{X}_2^+ \boldsymbol{X}_2$. Then, for any $n \in \mathbb{N}^+$ we have that*

$$\left\|\left(\boldsymbol{I} - \boldsymbol{P}_1\right)\left(\boldsymbol{P}_2 \boldsymbol{P}_1\right)^n\right\|^2 = \max_i \left\{\left(\cos^2 \theta_i\right)^{2n-1}\left(1 - \cos^2 \theta_i\right)\right\} ,$$

*where $\{\theta_i\}_i \subseteq \left(0, \frac{\pi}{2}\right]$ are the non-zero principal angles between the two tasks in $S$, i.e., between $\operatorname{range}(\boldsymbol{X}_1^\top)$ and $\operatorname{range}(\boldsymbol{X}_2^\top)$ or equivalently between $\operatorname{null}(\boldsymbol{X}_1)$ and $\operatorname{null}(\boldsymbol{X}_2)$ or between the affine solution spaces $\mathcal{W}_1$ and $\mathcal{W}_2$.*

**Proof** We use the SVD notation defined in Appendix A.1 and denote by $\boldsymbol{V}_1^\perp$, $\boldsymbol{V}_2^\perp$ the matrices whose orthonormal columns span $\operatorname{null}(\boldsymbol{X}_1)$, $\operatorname{null}(\boldsymbol{X}_2)$ respectively. Thus, we can express the projections as $\boldsymbol{P}_1 = \boldsymbol{V}_1^\perp \boldsymbol{V}_1^{\perp^\top}$ and $\boldsymbol{P}_2 = \boldsymbol{V}_2^\perp \boldsymbol{V}_2^{\perp^\top}$.

Now, we show that

$$\begin{aligned}
\left\|\left(\boldsymbol{I} - \boldsymbol{P}_1\right)\left(\boldsymbol{P}_2 \boldsymbol{P}_1\right)^n\right\|^2 &= \left\|\left(\boldsymbol{I} - \boldsymbol{P}_1\right)\left(\boldsymbol{V}_2^\perp \boldsymbol{V}_2^{\perp^\top} \boldsymbol{V}_1^\perp \boldsymbol{V}_1^{\perp^\top}\right)^n\right\|^2 \\
&= \left\|\left(\boldsymbol{I} - \boldsymbol{P}_1\right) \boldsymbol{V}_2^\perp \boldsymbol{V}_2^{\perp^\top} \boldsymbol{V}_1^\perp \left(\boldsymbol{V}_1^{\perp^\top} \boldsymbol{V}_2^\perp \boldsymbol{V}_2^{\perp^\top} \boldsymbol{V}_1^\perp\right)^{n-1} \boldsymbol{V}_1^{\perp^\top}\right\|^2 \\
\left[\begin{smallmatrix}\text{spectral norm}\\ \text{properties}\end{smallmatrix}\right] &= \left\|\left(\boldsymbol{I} - \boldsymbol{P}_1\right) \boldsymbol{V}_2^\perp \boldsymbol{V}_2^{\perp^\top} \boldsymbol{V}_1^\perp \left(\boldsymbol{V}_1^{\perp^\top} \boldsymbol{V}_2^\perp \boldsymbol{V}_2^{\perp^\top} \boldsymbol{V}_1^\perp\right)^{n-1}\right\|^2 .
\end{aligned}$$

We now notice that the idempotence of $\boldsymbol{P}_1 = \boldsymbol{V}_1^\perp \boldsymbol{V}_1^{\perp^\top}$ and $\boldsymbol{P}_2 = \boldsymbol{V}_2^\perp \boldsymbol{V}_2^{\perp^\top}$ implies

$$\begin{aligned}
\boldsymbol{V}_1^{\perp^\top} \boldsymbol{V}_2^\perp \boldsymbol{V}_2^{\perp^\top} \left(\boldsymbol{I} - \boldsymbol{P}_1\right)^\top \left(\boldsymbol{I} - \boldsymbol{P}_1\right) \boldsymbol{V}_2^\perp \boldsymbol{V}_2^{\perp^\top} \boldsymbol{V}_1^\perp &= \boldsymbol{V}_1^{\perp^\top} \boldsymbol{V}_2^\perp \boldsymbol{V}_2^{\perp^\top} \left(\boldsymbol{I} - \boldsymbol{P}_1\right) \boldsymbol{V}_2^\perp \boldsymbol{V}_2^{\perp^\top} \boldsymbol{V}_1^\perp \\
&= \boldsymbol{V}_1^{\perp^\top} \boldsymbol{V}_2^\perp \boldsymbol{V}_2^{\perp^\top} \boldsymbol{V}_1^\perp - \boldsymbol{V}_1^{\perp^\top} \boldsymbol{V}_2^\perp \boldsymbol{V}_2^{\perp^\top} \boldsymbol{V}_1^\perp \boldsymbol{V}_1^{\perp^\top} \boldsymbol{V}_2^\perp \boldsymbol{V}_2^{\perp^\top} \boldsymbol{V}_1^\perp \\
&= \boldsymbol{V}_1^{\perp^\top} \boldsymbol{V}_2^\perp \boldsymbol{V}_2^{\perp^\top} \boldsymbol{V}_1^\perp - \left(\boldsymbol{V}_1^{\perp^\top} \boldsymbol{V}_2^\perp \boldsymbol{V}_2^{\perp^\top} \boldsymbol{V}_1^\perp\right)^2 .
\end{aligned}$$

Since for the spectral norm of any real matrix $\boldsymbol{A}$ it holds that $\|\boldsymbol{A}\|^2 = \left\|\boldsymbol{A}^\top \boldsymbol{A}\right\|$, we get that

$$\begin{aligned}
&\left\|\left(\boldsymbol{I} - \boldsymbol{P}_1\right) \boldsymbol{V}_2^\perp \boldsymbol{V}_2^{\perp^\top} \boldsymbol{V}_1^\perp \left(\boldsymbol{V}_1^{\perp^\top} \boldsymbol{V}_2^\perp \boldsymbol{V}_2^{\perp^\top} \boldsymbol{V}_1^\perp\right)^{n-1}\right\|^2 \\
&= \left\|\left(\boldsymbol{V}_1^{\perp^\top} \boldsymbol{V}_2^\perp \boldsymbol{V}_2^{\perp^\top} \boldsymbol{V}_1^\perp\right)^{n-1} \boldsymbol{V}_1^{\perp^\top} \boldsymbol{V}_2^\perp \boldsymbol{V}_2^{\perp^\top} \left(\boldsymbol{I} - \boldsymbol{P}_1\right)^2 \boldsymbol{V}_2^\perp \boldsymbol{V}_2^{\perp^\top} \boldsymbol{V}_1^\perp \left(\boldsymbol{V}_1^{\perp^\top} \boldsymbol{V}_2^\perp \boldsymbol{V}_2^{\perp^\top} \boldsymbol{V}_1^\perp\right)^{n-1}\right\| \\
&= \left\|\left(\boldsymbol{V}_1^{\perp^\top} \boldsymbol{V}_2^\perp \boldsymbol{V}_2^{\perp^\top} \boldsymbol{V}_1^\perp\right)^{2n-1} - \left(\boldsymbol{V}_1^{\perp^\top} \boldsymbol{V}_2^\perp \boldsymbol{V}_2^{\perp^\top} \boldsymbol{V}_1^\perp\right)^{2n}\right\| .
\end{aligned}$$

Denote the spectral decomposition of the Gram matrix $\boldsymbol{V}_1^{\perp\top}\boldsymbol{V}_2^{\perp}\boldsymbol{V}_2^{\perp\top}\boldsymbol{V}_1^{\perp} \succeq \mathbf{0}$ as $\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$, with its (non-negative) eigenvalues ordered in a non-ascending order on the diagonal of $\boldsymbol{\Lambda}$ and $\boldsymbol{Q}$ being some orthonormal matrix. The upper bound thus becomes

$$\left\|\left(\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top\right)^{2n-1} - \left(\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top\right)^{2n}\right\| = \left\|\boldsymbol{\Lambda}^{2n-1} - \boldsymbol{\Lambda}^{2n}\right\| \,.$$
$$= \max_i \lambda_i^{2n-1}(1-\lambda_i) = \max_i \left\{\left(\cos^2\theta_i\right)^{2n-1}\left(1-\cos^2\theta_i\right)\right\} \,,$$

where $\{\theta_i\}_i \subseteq \left[0, \frac{\pi}{2}\right]$ are all the principal angles (zeros included) between $\mathrm{null}(\boldsymbol{X}_1)$ and $\mathrm{null}(\boldsymbol{X}_2)$. The last equality stems from a known analysis result (Theorem 2.1 in Risteski and Trencevski [58]) relating the principal angles between $\mathrm{null}(\boldsymbol{X}_1)$ and $\mathrm{null}(\boldsymbol{X}_2)$ to the eigenvalues of the Gram matrix we defined, *i.e.,* $\boldsymbol{V}_1^{\perp\top}\boldsymbol{V}_2^{\perp}\boldsymbol{V}_2^{\perp\top}\boldsymbol{V}_1^{\perp} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$. More formally, the result states that $\forall i$, $\lambda_i = \cos^2\theta_i$. See also Chapter 5.15 in Meyer [40] for more detailed explanations on this relation.

We conclude this proof by using Claim 19 showing that the *non-zero* principal angles between $\mathrm{null}(\boldsymbol{X}_1)$ and $\mathrm{null}(\boldsymbol{X}_2)$ are essentially the same as those between their orthogonal complements, *i.e.,* $\mathrm{range}(\boldsymbol{X}_1^\top)$ and $\mathrm{range}(\boldsymbol{X}_2^\top)$. ■

### A.3.1. AUXILIARY LEMMAS ON FORGETTING

We first state an auxiliary lemma for deriving lower bounds on the forgetting of tasks of rank $d-1$.

**Lemma 21 (Lower bound on forgetting in the $r = d-1$ case)** *Let $S \in \mathcal{S}_T$ be a task collection with $T$ data matrices of rank $r = d-1$. Let $\boldsymbol{v}_1, \dots, \boldsymbol{v}_T$ be the normalized vectors spanning the $T$ rank-*one *solution spaces $\mathcal{W}_1, \dots, \mathcal{W}_T$. Let $\theta_{\tau(i,j)} \in [0, \pi/2]$ be the angle between $\boldsymbol{v}_{\tau(i)}$ and $\boldsymbol{v}_{\tau(j)}$, $\forall i, j \in [k]$, meaning that $\cos^2 \theta_{\tau(i,j)} = \left(\boldsymbol{v}_{\tau(i)}^\top \boldsymbol{v}_{\tau(j)}\right)^2$. Finally, let $\boldsymbol{w}^\star$ be the minimum norm offline solution of $S$. Then, the forgetting on $S$ after $k$ iterations is lower bounded by*

$$F_{\tau,S}(k) \geq \min_{m\in[T]} \left\{\sigma_{\min}^2(\boldsymbol{X}_m)\right\} \left(\boldsymbol{v}_{\tau(1)}^\top \boldsymbol{w}^\star\right)^2 \frac{1}{k} \left(\prod_{i=1}^{k-1} \cos^2 \theta_{\tau(i,i+1)}\right) \sum_{j=1}^{k-1} \left(1 - \cos^2 \theta_{\tau(j,k)}\right) \ ,$$

where $\sigma_{\min}^2(\boldsymbol{X}_m)$ is the smallest squared non-zero singular value of $\boldsymbol{X}_m$.

**Proof** [Proof for Lemma 21] We start from Eq. (8) and perform similar derivations to the ones that lead us to Eq. (9). We have,

$$\begin{aligned}
F_{\tau,S}(k) &= \frac{1}{k} \sum_{t=1}^{k} \left\| \boldsymbol{X}_{\tau(t)} \boldsymbol{P}_{\tau(k)} \cdots \boldsymbol{P}_{\tau(1)} \boldsymbol{w}^\star \right\|^2 \\
[\text{pseudo-inverse}] &= \frac{1}{k} \sum_{t=1}^{k} \left\| \boldsymbol{X}_{\tau(t)} \left(\boldsymbol{X}_{\tau(t)}^+ \boldsymbol{X}_{\tau(t)}\right) \boldsymbol{P}_{\tau(k)} \cdots \boldsymbol{P}_{\tau(1)} \boldsymbol{w}^\star \right\|^2 \\
\left[\begin{smallmatrix}\text{spectral}\\ \text{norm}\\ \text{properties}\end{smallmatrix}\right] &\geq \frac{1}{k} \sum_{t=1}^{k} \sigma_{\min}^2(\boldsymbol{X}_{\tau(t)}) \left\| \left(\boldsymbol{X}_{\tau(t)}^+ \boldsymbol{X}_{\tau(t)}\right) \boldsymbol{P}_{\tau(k)} \cdots \boldsymbol{P}_{\tau(1)} \boldsymbol{w}^\star \right\|^2 \\
&\geq \min_{m\in[T]} \left\{\sigma_{\min}^2(\boldsymbol{X}_m)\right\} \frac{1}{k} \sum_{t=1}^{k} \left\| \left(\boldsymbol{I} - \boldsymbol{P}_{\tau(t)}\right) \boldsymbol{P}_{\tau(k)} \cdots \boldsymbol{P}_{\tau(1)} \boldsymbol{w}^\star \right\|^2 \\
[r{=}d{-}1] &= \min_{m\in[T]} \left\{\sigma_{\min}^2(\boldsymbol{X}_m)\right\} \frac{1}{k} \sum_{t=1}^{k-1} \left\| \left(\boldsymbol{I} - \boldsymbol{v}_{\tau(t)} \boldsymbol{v}_{\tau(t)}^\top\right) \boldsymbol{v}_{\tau(k)} \boldsymbol{v}_{\tau(k)}^\top \cdots \boldsymbol{v}_{\tau(1)} \boldsymbol{v}_{\tau(1)}^\top \boldsymbol{w}^\star \right\|^2 \ ,
\end{aligned}$$

where we used the rank-1 SVD of the data matrices, *i.e.,* $\boldsymbol{X}_m = \boldsymbol{v}_m \boldsymbol{v}_m^\top$.

Now, we denote $\forall i, j : \left|\boldsymbol{v}_{\tau(i)}^\top \boldsymbol{v}_{\tau(j)}\right| = \theta_{i,j}$ and get (notice the signs do not matter in the norm),

$$\begin{aligned}
&= \min_{m\in[T]} \left\{\sigma_{\min}^2(\boldsymbol{X}_m)\right\} \frac{1}{k} \sum_{t=1}^{k-1} \left\| \left(\boldsymbol{I} - \boldsymbol{v}_{\tau(t)} \boldsymbol{v}_{\tau(t)}^\top\right) \left(\prod_{i=1}^{k-1} \cos \theta_{\tau(i,i+1)}\right) \boldsymbol{v}_\perp^{\tau(k)} \boldsymbol{v}_\perp^{\tau(1)^\top} \boldsymbol{w}^\star \right\|^2 \\
&= \min_{m\in[T]} \left\{\sigma_{\min}^2(\boldsymbol{X}_m)\right\} \frac{(\boldsymbol{v}_{\tau(1)}^\top \boldsymbol{w}^\star)^2}{k} \left(\prod_{i=1}^{k-1} \cos^2 \theta_{\tau(i,i+1)}\right) \sum_{j=1}^{k-1} \left\| \left(\boldsymbol{I} - \boldsymbol{v}_\perp^{\tau(j)} \boldsymbol{v}_\perp^{\tau(j)^\top}\right) \boldsymbol{v}_\perp^{\tau(k)} \right\|^2 \\
[\text{idempotence}] &= \min_{m\in[T]} \left\{\sigma_{\min}^2(\boldsymbol{X}_m)\right\} \frac{(\boldsymbol{v}_{\tau(1)}^\top \boldsymbol{w}^\star)^2}{k} \left(\prod_{i=1}^{k-1} \cos^2 \theta_{\tau(i,i+1)}\right) \sum_{j=1}^{k-1} \boldsymbol{v}_\perp^{\tau(k)^\top} \left(\boldsymbol{I} - \boldsymbol{v}_\perp^{\tau(j)} \boldsymbol{v}_\perp^{\tau(j)^\top}\right) \boldsymbol{v}_\perp^{\tau(k)} \\
&= \min_{m\in[T]} \left\{\sigma_{\min}^2(\boldsymbol{X}_m)\right\} \frac{(\boldsymbol{v}_{\tau(1)}^\top \boldsymbol{w}^\star)^2}{k} \left(\prod_{i=1}^{k-1} \cos^2 \theta_{\tau(i,i+1)}\right) \sum_{j=1}^{k-1} \left(1 - \cos^2 \theta_{\tau(j,k)}\right) \ .
\end{aligned}$$

■

## Appendix B. Supplementary material: Arbitrary task orderings (Section 4)

### B.1. No forgetting cases (Section 4.1)

Consider learning two tasks sequentially: $\left(\boldsymbol{X}_1, \boldsymbol{y}_1\right)$ and then $\left(\boldsymbol{X}_2, \boldsymbol{y}_2\right)$. Right after learning the second task, we have $\mathcal{L}_2(\boldsymbol{w}_2) = \|\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{w}_2\|^2 = 0$. Thus, the forgetting from Eq. (8) becomes $F_{\tau,S}(2) = \frac{1}{2}\left\|\boldsymbol{X}_1\boldsymbol{P}_2\boldsymbol{P}_1\boldsymbol{w}^\star\right\|^2$. We now derive sufficient and necessary conditions for no forgetting.

**Recall Theorem 6.** Let $S = \left\{\left(\boldsymbol{X}_1, \boldsymbol{y}_1\right), \left(\boldsymbol{X}_2, \boldsymbol{y}_2\right)\right\} \in \mathcal{S}_{T=2}$ be a task collection with 2 tasks, fitted under an identity ordering $\tau$, *i.e.,* $\left(\boldsymbol{X}_1, \boldsymbol{y}_1\right)$ and then $\left(\boldsymbol{X}_2, \boldsymbol{y}_2\right)$. Then the following conditions are equivalent:

1. For *any* labeling $\boldsymbol{y}_1, \boldsymbol{y}_2$ (or equivalently, any minimum norm solution $\boldsymbol{w}^\star$), after fitting the second task, the model does not "forget" the first one. That is, $F_{\tau,S}(2) = 0$.
2. It holds that $\boldsymbol{X}_1\boldsymbol{P}_2\boldsymbol{P}_1 = \boldsymbol{0}_{n_1 \times d}$.
3. Each principal angle between the tasks, *i.e.,* $\mathrm{range}(\boldsymbol{X}_1^\top)$ and $\mathrm{range}(\boldsymbol{X}_2^\top)$, is either $0$ or $\pi/2$.

#### B.1.1. EXAMPLE: SUFFICIENT CONDITIONS FOR NO FORGETTING

Before we prove the theorem above, we exemplify some of its implications by showing clear and simple sufficient conditions for holding the conditions of the theorem.

(a) $\mathrm{range}\left(\boldsymbol{X}_2^\top\right) \subseteq \mathrm{range}\left(\boldsymbol{X}_1^\top\right)$; or $\mathrm{range}\left(\boldsymbol{X}_2^\top\right) \supseteq \mathrm{range}\left(\boldsymbol{X}_1^\top\right)$; or

(b) $\mathrm{range}\left(\boldsymbol{X}_2^\top\right) \subseteq \mathrm{null}\left(\boldsymbol{X}_1\right)$; or $\mathrm{range}\left(\boldsymbol{X}_2^\top\right) \supseteq \mathrm{null}\left(\boldsymbol{X}_1\right)$.

These conditions can help understand that maximal task (=sample) similarity or dissimilarity can help prevent forgetting in the linear setting.

#### B.1.2. PROVING THE THEOREM

**Proof** Like we explain in Section 4.1, the forgetting after learning the second task is equal to $F_{\tau,S}(2) = \mathcal{L}_1(\boldsymbol{w}_2) = \frac{1}{2}\left\|\boldsymbol{X}_1\boldsymbol{P}_2\boldsymbol{P}_1\boldsymbol{w}^\star\right\|^2$. We notice that $\boldsymbol{P}_1 = \boldsymbol{I} - \boldsymbol{X}_1^+\boldsymbol{X}_1$ and $\boldsymbol{P}_2 = \boldsymbol{I} - \boldsymbol{X}_2^+\boldsymbol{X}_2$, meaning that the labels $\boldsymbol{y}_1, \boldsymbol{y}_2$ do not have any effect on the matrix $\boldsymbol{X}_1\boldsymbol{P}_2\boldsymbol{P}_1$. However, these labels *do* effect the minimum norm solution $\boldsymbol{w}^\star$.

Here, we briefly discuss the relation between $\boldsymbol{w}^\star$ and $\boldsymbol{y}_1, \boldsymbol{y}_2$, so as to facilitate our proof below.

**Relating the minimum norm solution and the labelings.** Recall the constraints between the offline solution and the labels, *i.e.,* $\boldsymbol{X}_1\boldsymbol{w}^\star = \boldsymbol{y}_1$ and $\boldsymbol{X}_2\boldsymbol{w}^\star = \boldsymbol{y}_2$. Also recall that under Assumption 2, we have that $\boldsymbol{w}^\star \in \mathcal{B}^d$. Note that a minimum norm solution *must* lie in the row span of *both* data matrices, *i.e.,* $\boldsymbol{w}^\star \in \mathrm{range}(\boldsymbol{X}_1^\top) \cup \mathrm{range}(\boldsymbol{X}_2^\top)$, since any contributions from the nullspaces will not affect its predictions $\boldsymbol{X}_1\boldsymbol{w}^\star$ and $\boldsymbol{X}_2\boldsymbol{w}^\star$ but *will* increase its norm.

Moreover, notice that the $\boldsymbol{y}_1, \boldsymbol{y}_2$ can yield *any* minimum norm solution that is inside $\mathrm{range}(\boldsymbol{X}_1^\top)$, since we could just choose an arbitrary vector $\boldsymbol{w}^\star \in \mathrm{range}(\boldsymbol{X}_1^\top)$ and set $\boldsymbol{y}_2 = \boldsymbol{X}_2\boldsymbol{w}^\star$ (we do not have restrictions on the labelings).

We are now ready to complete our proof.

**Condition (1) $\iff$ Condition (2).** Clearly, since $F_{\tau,S}(2)=\frac{1}{2}\left\|\boldsymbol{X}_1\boldsymbol{P}_2\boldsymbol{P}_1\boldsymbol{w}^\star\right\|^2$, we have that

$$\underbrace{\boldsymbol{X}_1\boldsymbol{P}_2\boldsymbol{P}_1=\boldsymbol{0}}_{(2)} \implies \underbrace{\forall \boldsymbol{y}_1,\boldsymbol{y}_2,\boldsymbol{w}^\star:\ F_{\tau,S}(2)=\tfrac{1}{2}\left\|\boldsymbol{X}_1\boldsymbol{P}_2\boldsymbol{P}_1\boldsymbol{w}^\star\right\|^2=0}_{(1)}\ .$$

Since $\boldsymbol{P}_1$ is a symmetric operator projecting onto the row span of $\boldsymbol{X}_1$, we have that $\mathrm{range}(\boldsymbol{P}_1)=\mathrm{range}(\boldsymbol{P}_1^\top)=\mathrm{range}(\boldsymbol{X}_1^\top)$. It is readily seen that

$$\left(\forall \boldsymbol{w}^\star\in\mathrm{range}(\boldsymbol{X}_1^\top):\ \left\|\boldsymbol{X}_1\boldsymbol{P}_2\boldsymbol{P}_1\boldsymbol{w}^\star\right\|=0\right) \iff \boldsymbol{X}_1\boldsymbol{P}_2\boldsymbol{P}_1=\boldsymbol{0}\,.$$

We explicitly denote the minimum norm solution that two labelings $\boldsymbol{y}_1,\boldsymbol{y}_2$ induce as $\boldsymbol{w}^\star(\boldsymbol{y}_1,\boldsymbol{y}_2)$. As explained, $\boldsymbol{y}_1,\boldsymbol{y}_2$ can yield *any* minimum solution inside $\mathrm{range}(\boldsymbol{X}_1^\top)$. Assume $\forall \boldsymbol{y}_1,\boldsymbol{y}_2:\ \left\|\boldsymbol{X}_1\boldsymbol{P}_2\boldsymbol{P}_1\boldsymbol{w}^\star(\boldsymbol{y}1,\boldsymbol{y}2)\right\|=0$. Then, it follows that $\forall \boldsymbol{w}^\star\in\mathrm{range}(\boldsymbol{X}_1^\top):\ \left\|\boldsymbol{X}_1\boldsymbol{P}_2\boldsymbol{P}_1\boldsymbol{w}^\star\right\|=0$. In this case we proved that it follows that $\boldsymbol{X}_1\boldsymbol{P}_2\boldsymbol{P}_1=\boldsymbol{0}$.

Overall we showed:

$$\underbrace{\left(\forall \boldsymbol{y}_1,\boldsymbol{y}_2\in\mathrm{range}(\boldsymbol{X}_1^\top):\ \left\|\boldsymbol{X}_1\boldsymbol{P}_2\boldsymbol{P}_1\boldsymbol{w}^\star\right\|=0\right)}_{(1)} \implies \underbrace{\boldsymbol{X}_1\boldsymbol{P}_2\boldsymbol{P}_1=\boldsymbol{0}}_{(2)}\ ,$$

thus completing the proof for (1) $\iff$ (2).

**Condition (2) $\iff$ Condition (3).** First, we notice that

$$\left\|\boldsymbol{X}_1\boldsymbol{P}_2\boldsymbol{P}_1\right\|=\boldsymbol{0} \iff \left\|\left(\boldsymbol{I}-\boldsymbol{P}_1\right)\boldsymbol{P}_2\boldsymbol{P}_1\right\|=\boldsymbol{0}$$

since simple norm properties give us:

$$\begin{aligned}
&\left\|\left(\boldsymbol{I}-\boldsymbol{P}_1\right)\boldsymbol{P}_2\boldsymbol{P}_1\right\| = \left\|\boldsymbol{V}_1\boldsymbol{\Sigma}_1^+\boldsymbol{\Sigma}_1\boldsymbol{V}_1^\top\boldsymbol{P}_2\boldsymbol{P}_1\right\| = \left\|\boldsymbol{\Sigma}_1^+\boldsymbol{\Sigma}_1\boldsymbol{V}_1^\top\boldsymbol{P}_2\boldsymbol{P}_1\right\| \\
&\le \underbrace{\left\|\boldsymbol{\Sigma}_1^+\right\|}_{>0}\left\|\boldsymbol{\Sigma}_1\boldsymbol{V}_1^\top\boldsymbol{P}_2\boldsymbol{P}_1\right\| \propto \left\|\boldsymbol{\Sigma}_1\boldsymbol{V}_1^\top\boldsymbol{P}_2\boldsymbol{P}_1\right\| = \left\|\boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^\top\boldsymbol{P}_2\boldsymbol{P}_1\right\| = \left\|\boldsymbol{X}_1\boldsymbol{P}_2\boldsymbol{P}_1\right\|\ ,
\end{aligned}$$

and

$$\begin{aligned}
&\left\|\boldsymbol{X}_1\boldsymbol{P}_2\boldsymbol{P}_1\right\| = \left\|\boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^\top\boldsymbol{P}_2\boldsymbol{P}_1\right\| = \left\|\boldsymbol{\Sigma}_1\boldsymbol{V}_1^\top\boldsymbol{P}_2\boldsymbol{P}_1\right\| = \left\|\boldsymbol{\Sigma}_1\left(\boldsymbol{\Sigma}_1^+\boldsymbol{\Sigma}_1\right)\boldsymbol{V}_1^\top\boldsymbol{P}_2\boldsymbol{P}_1\right\| \\
&\le \underbrace{\left\|\boldsymbol{\Sigma}_1\right\|}_{>0}\left\|\boldsymbol{\Sigma}_1^+\boldsymbol{\Sigma}_1\boldsymbol{V}_1^\top\boldsymbol{P}_2\boldsymbol{P}_1\right\| \\
&\propto \left\|\boldsymbol{\Sigma}_1^+\boldsymbol{\Sigma}_1\boldsymbol{V}_1^\top\boldsymbol{P}_2\boldsymbol{P}_1\right\| = \left\|\boldsymbol{V}_1\boldsymbol{\Sigma}_1^+\boldsymbol{\Sigma}_1\boldsymbol{V}_1^\top\boldsymbol{P}_2\boldsymbol{P}_1\right\| = \left\|\left(\boldsymbol{I}-\boldsymbol{P}_1\right)\boldsymbol{P}_2\boldsymbol{P}_1\right\|\,.
\end{aligned}$$

Then, we use Lemma 20 and get that

$$\left\|\left(\boldsymbol{I}-\boldsymbol{P}_1\right)\boldsymbol{P}_2\boldsymbol{P}_1\right\| = \max_i\left\{\left(\cos^2\theta_i\right)\left(1-\cos^2\theta_i\right)\right\} = \frac{1}{4}\max_i\left\{\sin^2(2\theta_i)\right\}\ ,$$

where $\{\theta_i\}_i\subseteq\left(0,\frac{\pi}{2}\right]$ are the non-zero principal angles between the two tasks, *i.e.,* between $\mathrm{range}(\boldsymbol{X}_1^\top)$ and $\mathrm{range}(\boldsymbol{X}_2^\top)$. Finally, it is now clear that

$$\left\|\boldsymbol{X}_1\boldsymbol{P}_2\boldsymbol{P}_1\right\|=\boldsymbol{0} \iff \left\|\left(\boldsymbol{I}-\boldsymbol{P}_1\right)\boldsymbol{P}_2\boldsymbol{P}_1\right\|=\frac{1}{4}\max_i\left\{\sin^2(2\theta_i)\right\}=\boldsymbol{0} \iff \forall i:\theta_i\in\left\{0,\frac{\pi}{2}\right\}\ .$$

■

### B.1.3. COMPARISON TO DOAN ET AL. [14]

Doan et al. [14] also studied forgetting in a linear setting where a series of tasks are learned sequentially by SGD on squared loss with ridge penalty. In the special case where only $T = 2$ tasks are given and no regularization is used (*i.e.,* $\lambda = 0$), their expression for forgetting in Theorem 1 [14] is equivalent to our derivation in Eq. (8) up to scaling by $\frac{1}{2}$. However, their subsequent upper bound stated in Corollary 1 is a looser characterization of forgetting, which can be paraphrased in terms of our notation and framework as follows (for convenience, we attach their notations beneath the last equation):

$$\begin{aligned} F_{\tau,S}(2) &= \frac{1}{2}\left\|\boldsymbol{X}_1\boldsymbol{P}_2\boldsymbol{P}_1(\boldsymbol{w}_0-\boldsymbol{w}^\star)\right\|^2 = \frac{1}{2}\left\|\boldsymbol{X}_1\cdot\boldsymbol{X}_1^+\boldsymbol{X}_1\boldsymbol{P}_2\boldsymbol{P}_1(\boldsymbol{w}_0-\boldsymbol{w}^\star)\right\|^2 \\ &= \frac{1}{2}\left\|\boldsymbol{X}_1(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{P}_2\boldsymbol{P}_1(\boldsymbol{w}_0-w^\star)\right\|^2 \\ &\overset{(*)}{=} \frac{1}{2}\|\boldsymbol{X}_1(\boldsymbol{I}-\boldsymbol{P}_1)(\boldsymbol{I}-\boldsymbol{P}_2)\boldsymbol{P}_\perp^{(1)}(\boldsymbol{w}_0-\boldsymbol{w}^\star)\|^2 \\ &\le \frac{1}{2}\|\boldsymbol{X}_1\|^2\big\|\underbrace{(\boldsymbol{I}-\boldsymbol{P}_1)(\boldsymbol{I}-\boldsymbol{P}_2)}_{\Theta^{1\to2}}\big\|^2\big\|\underbrace{\boldsymbol{P}_1(\boldsymbol{w}_0-\boldsymbol{w}^\star)}_{=\boldsymbol{M}_2\tilde{\boldsymbol{y}}_2}\big\|^2, \end{aligned} \tag{11}$$

where (*) follows from plugging in $\boldsymbol{P}_2 = \boldsymbol{P}_2 - \boldsymbol{I} + \boldsymbol{I}$ and $(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{P}_1 = 0$.

Based on the above upper bound, the authors informally argue that higher similarity of the principal components between the source task and target task leads to higher risk of forgetting. In contrast, the sufficient condition (a) in our Appendix B.1.1 shows that there is no forgetting when tasks have maximum overlap in the row spans of their inputs.

Concretely, consider the following two tasks that hold the sufficient conditions of Theorem 6 presented in Appendix B.1.1:

$$\left(\boldsymbol{X}^{(1)} = [1,0,0,0],\ \boldsymbol{y}^{(1)} = \frac{1}{\sqrt{2}}[1]\right),\quad \left(\boldsymbol{X}^{(2)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix},\ \boldsymbol{y}^{(2)} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}\right).$$

Note that $\boldsymbol{w}^\star = \frac{1}{\sqrt{2}}[1,1,0,0]^\top$ is a unit norm linear predictor that realizes both tasks. In this case, there is clearly no forgetting as the first task is also part of second task, *i.e.,* $F_{\tau,S}(2) = 0$. However, we can verify that $\|\boldsymbol{X}_1\|\|\boldsymbol{P}^{(1)}\boldsymbol{P}^{(2)}\| = 1$ and $\|\boldsymbol{P}_1(\boldsymbol{w}_0-\boldsymbol{w}^\star)\| = 1/\sqrt{2}$, thus the upper bound in Eq. (11) evaluates to $F_{\tau,S}(2) \le 1/4$. This demonstrates the weakness of the upper bound in Corollary 1 of [14].

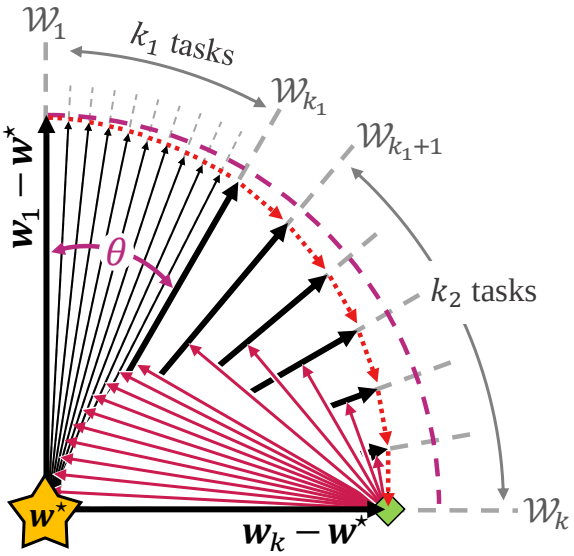
This gap can also be seen from a principal angle perspective. Their so-called overlap matrix, *i.e.,* $\Theta^{1\to2}$, is a diagonal matrix holding the singular values of $\boldsymbol{V}_2^{\perp\top}\boldsymbol{V}_1^\perp$. As they explain in Corollary 1 and we explain in the proof of Lemma 20, these singular values are actually connected to the principal angles between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. Thus, when they use the spectral norm $\left\|\Theta^{1\to2}\right\|$ which equals 1, they are actually using *only* the largest singular values, *i.e.,* the *smallest* principal angle which is called the Dimixer angle [13]. In the above example, this angle, which is the only principal angle, is zero (thus holding our conditions from Theorem 6). In contrast, our analysis in Lemma 9 uses *all* principal angles, revealing more delicate effects of task similarity on forgetting dynamics.

### B.2. Maximal forgetting cases (Section 4.2)

**Recall Theorem 7.** When using the identity ordering (*i.e.,* $\tau(t) = t$), thus seeing each task once, the worst-case forgetting after $k$ iterations is arbitrarily bad, *i.e.,*

$$1 - \sup_{S \in \mathcal{S}_{T=k}} F_{\tau,S}(k) \leq \mathcal{O}\left(1/\sqrt{k}\right) .$$

**Recall Figure 2.** Sequence of tasks where $F_{\tau,S}(k) \to 1$. Each black arrow represents a solution space of a rank 1 task in $\mathbb{R}^2$. There are $k_1$ tasks between $[0, \theta]$ and $k_2$ tasks between $[\theta, \pi/2]$. The green diamond shows the overall contraction after fitting all tasks. The red arrows show projections back onto the solution spaces. The mean squared length of these arrows is the forgetting.




**Proof sketch.** We show that for any $\epsilon > 0$ there exists a task collection $S$ of $T = k = \mathcal{O}(1/\epsilon^2)$ tasks, such that under the identity ordering $\tau$, the forgetting is $F_{\tau,S}(k) > 1 - \epsilon$.

We construct a task sequence as follows (illustrated in the figure in 2 dimensions):

1. For some small angle $\theta \sim \sqrt{\epsilon}$ we define $k_1 \sim 1/\epsilon^2$ tasks uniformly in $[0, \theta]$ ;
2. We add another $k_2 \sim 1/\epsilon$ tasks uniformly in $(\theta, \pi/2]$ .

With this construction we show that on the one hand there is almost no contraction, but on the other hand there is a large forgetting especially on the first $k_1$ tasks, since they are almost orthogonal to the last task and the projection is large.

We note in passing a related phenomenon in quantum physics, known as the *quantum Zeno effect* [23], where the state of a quantum system, described by a vector, can be manipulated by applying infinitesimally-spaced measurements, which act on it as orthogonal projections.

The full proof of Theorem 7 is given below.

**Proof** [Proof of Theorem 7]

We show that for any $\epsilon \in (0,1)$ there exists a task collection $S$ of $k = \mathcal{O}\left(1/\epsilon^2\right)$ tasks such that under the identity ordering, $F_{\tau,S}(k) > 1-\epsilon$. From Lemma 21 we know that for any choice of a task collection $S$ of rank $d-1$,

$$F_{\tau,S}(k) \geq \frac{1}{k}\left(\prod_{m=1}^{k-1}\cos^2\theta_{\tau(m,m+1)}\right)\sum_{m=1}^{k-1}\left(1-\cos^2\theta_{\tau(m,k)}\right) = \frac{1}{k}\left(\prod_{m=1}^{k-1}\cos^2\theta_{m,m+1}\right)\sum_{m=1}^{k-1}\sin^2\theta_{m,k}. \tag{12}$$

We construct a sequence with $T = k = 72/\epsilon^2 + 1$ tasks as following: let $\theta = \theta(\epsilon) \in \left(0, \frac{\pi}{2}\right)$ and assume we have $k_1 = k_1(\epsilon)$ tasks uniformly in $[0,\theta]$ and $k_2 = k_2(\epsilon)$ tasks uniformly in $\left(\theta, \frac{\pi}{2}\right]$. Note that we require $k_1 + k_2 = k$.

Then from Eq. (12) we have

$$\begin{aligned}
F_{\tau,S}(k) &\geq \frac{1}{k_1+k_2}\cos^{2(k_1-1)}\left(\frac{\theta}{k_1-1}\right)\cos^{2k_2}\left(\frac{\frac{\pi}{2}-\theta}{k_2}\right)\left(\sum_{i=1}^{k_1-1}\sin^2\underbrace{\theta_{i,k}}_{\geq \pi/2-\theta} + \underbrace{\sum_{i=k_1}^{k-1}\sin^2\theta_{i,k}}_{\geq 0}\right) \\
&\geq \frac{1}{k_1+k_2}\cos^{2k_1-2}\left(\frac{\theta}{k_1-1}\right)\cos^{2k_2}\left(\frac{\pi-2\theta}{2k_2}\right)\sum_{i=1}^{k_1-1}\sin^2\left(\frac{\pi}{2}-\theta\right) \\
&= \frac{1}{k_1+k_2}\cos^{2k_1-2}\left(\frac{\theta}{k_1-1}\right)\cos^{2k_2}\left(\frac{\pi-2\theta}{2k_2}\right)\left((k_1-1)\sin^2\left(\frac{\pi}{2}-\theta\right)\right) \\
&= \underbrace{\cos^{2k_1-2}\left(\frac{\theta}{k_1-1}\right)}_{\triangleq F_1}\cdot\underbrace{\cos^{2k_2}\left(\frac{\pi-2\theta}{2k_2}\right)}_{\triangleq F_2}\cdot\underbrace{\frac{k_1-1}{k_1+k_2}\cos^2\theta}_{\triangleq F_3}.
\end{aligned}$$

We show that for any $\epsilon \in (0,1)$ we can choose $k_1, k_2, \theta$ such that

$$F_{\tau,S}(k) \geq F_1(k_1,\theta)\cdot F_2(k_2,\theta)\cdot F_3(k_1,k_2,\theta) > 1-\epsilon\,.$$

Specifically, let $k_1 = \frac{72-12\epsilon}{\epsilon^2}+1, \qquad k_2 = \frac{12}{\epsilon}, \qquad \theta = \sqrt{\frac{\epsilon}{6}}$.

Then, we use the inequality $\cos x \geq 1 - x^2/2$, and get

$$\begin{aligned}
F_1(k_1,\theta) &= \cos^{2(k_1-1)}\left(\frac{\theta}{k_1-1}\right) \geq \cos^{2(k_1-1)}\left(\frac{1}{k_1-1}\right) \geq \left(1-\frac{1}{2(k_1-1)^2}\right)^{2k_1-2} \\
&\geq 1-\frac{1}{k_1-1} = 1-\frac{\epsilon^2}{72-12\epsilon} \geq 1-\frac{\epsilon}{3} \\
F_2(k_2,\theta) &= \cos^{2k_2}\left(\frac{\pi-2\theta}{2k_2}\right) \geq \cos^{2k_2}\left(\frac{2}{k_2}\right) \geq \left(1-\frac{2}{k_2^2}\right)^{2k_2} \geq 1-\frac{4}{k_2} = 1-\frac{\epsilon}{3} \\
F_3(k_1,k_2,\theta) &= \frac{k_1-1}{k_1+k_2}\cos^2\theta \geq \frac{\frac{72-12\epsilon}{\epsilon^2}}{\frac{72-12\epsilon}{\epsilon^2}+\frac{12}{\epsilon}+1}\left(1-\frac{\epsilon}{6}\right) = \frac{72-12\epsilon}{72+\epsilon^2}\left(1-\frac{\epsilon}{6}\right) \geq 1-\frac{\epsilon}{3}\,.
\end{aligned}$$

Therefore, we get that $F_{\tau,S}(k) \geq F_1(k_1,\theta)\cdot F_2(k_2,\theta)\cdot F_3(k_1,k_2,\theta) \geq \left(1-\frac{\epsilon}{3}\right)^3 \geq 1-\epsilon\,.$

Finally note that $k = k_1 + k_2 = 72/\epsilon^2 + 1$, which concludes the proof. ■

## Appendix C. Supplementary material: Cyclic task orderings (Section 5)

### C.1. Forgetting with $T = 2$ tasks (Section 5.1)

**Recall Lemma 9.** For any task collection $S \in \mathcal{S}_{T=2}$ of two tasks, the forgetting after $k = 2n$ iterations (*i.e.,* $n$ cycles) is tightly upper bounded by

$$F_{\tau,S}(k) \leq \frac{1}{2} \max_i \left\{ \left(\cos^2 \theta_i\right)^{k-1} \left(1 - \cos^2 \theta_i\right) \right\} ,$$

where $\{\theta_i\}_i \subseteq \left(0, \frac{\pi}{2}\right]$ are the non-zero principal angles between the two tasks in $S$. Moreover, the above inequality saturates when all non-zero singular values of the first task (*i.e.,* of $\boldsymbol{X}_1$) are 1s.

**Proof** [Proof for Lemma 9] We notice that at the end of each cycle we perfectly fit the second task, thus having forgetting only on the first one. In the cyclic case, from Eq. (10) we have

$$F_{\tau,S}(k) \leq \frac{1}{2} \left\| \left(\boldsymbol{I} - \boldsymbol{P}_1\right) \left(\boldsymbol{P}_2 \boldsymbol{P}_1\right)^n \right\|^2 .$$

We now apply Lemma 20 (recall that $k = 2n$) and conclude that:

$$F_{\tau,S}(k) \leq \frac{1}{2} \left\| \left(\boldsymbol{I} - \boldsymbol{P}_1\right) \left(\boldsymbol{P}_2 \boldsymbol{P}_1\right)^n \right\|^2 = \frac{1}{2} \max_i \left\{ \left(\cos^2 \theta_i\right)^{k-1} \left(1 - \cos^2 \theta_i\right) \right\} ,$$

where $\{\theta_i\}_i \subseteq \left(0, \frac{\pi}{2}\right]$ are the non-zero principal angles between the two tasks. Finally, we show that when all non-zero singular values of $\boldsymbol{X}_1$ are 1s it holds that $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_1^+ \boldsymbol{\Sigma}_1$ and we get

$$\begin{aligned}
F_{\tau,S}(k) &= \frac{1}{2} \sum_{m=1}^{T} \left\| \boldsymbol{X}_m \left(\boldsymbol{P}_2 \boldsymbol{P}_1\right)^n \boldsymbol{w}^\star \right\|^2 = \frac{1}{2} \left( \left\| \boldsymbol{X}_1 \left(\boldsymbol{P}_2 \boldsymbol{P}_1\right)^n \boldsymbol{w}^\star \right\|^2 + \underbrace{\left\| \boldsymbol{X}_2 \left(\boldsymbol{P}_2 \boldsymbol{P}_1\right)^n \boldsymbol{w}^\star \right\|^2}_{=0} \right) \\
&= \frac{1}{2} \left\| \boldsymbol{U}_1 \boldsymbol{\Sigma}_1 \boldsymbol{V}_1^\top \left(\boldsymbol{P}_2 \boldsymbol{P}_1\right)^n \boldsymbol{w}^\star \right\|^2 \\
[\text{Prop. 1}] &= \frac{1}{2} \left\| \boldsymbol{\Sigma}_1 \boldsymbol{V}_1^\top \left(\boldsymbol{P}_2 \boldsymbol{P}_1\right)^n \boldsymbol{w}^\star \right\|^2 = \frac{1}{2} \left\| \boldsymbol{\Sigma}_1^+ \boldsymbol{\Sigma}_1 \boldsymbol{V}_1^\top \left(\boldsymbol{P}_2 \boldsymbol{P}_1\right)^n \boldsymbol{w}^\star \right\|^2 \\
&= \frac{1}{2} \left\| \left(\boldsymbol{I} - \boldsymbol{P}_1\right) \left(\boldsymbol{P}_2 \boldsymbol{P}_1\right)^n \boldsymbol{w}^\star \right\|^2 ,
\end{aligned}$$

proving the inequality saturates in this case. ■

**Recall Theorem 10.** For a cyclic ordering $\tau$ of 2 tasks, the worst-case forgetting after $k = 2n \geq 2$ iterations (*i.e.,* $n \geq 1$ cycles), is

$$\sup_{S \in \mathcal{S}_{T=2}} F_{\tau,S}(k) = \frac{1}{2e(k-1)} - \frac{1}{4e(k-1)^2} + \mathcal{O}\left(\frac{1}{k^3}\right) .$$

**Proof** Following our previous lemma, the key to deriving the worst-case bound is to find the maximum of

$$\frac{1}{2}\left(\cos^2\theta\right)^x \left(1 - \cos^2\theta\right) , \quad \forall x \in \mathbb{N} \cup \{0\} .$$

For $x = 0$, the expression above is maximized by $\theta = 0$ and equals $1/2$. Generally, one could show that for any integer $x \in \mathbb{N}^+$, the angle that maximizes the expression holds $\sin^2\theta = \frac{1}{x+1}$. Plugging that solution into the expression, we get:

$$\frac{1}{2}\max_{\theta \in [0,\pi]} \left(\cos^2\theta\right)^x \left(1 - \cos^2\theta\right) = \frac{1}{2}\left(1 - \frac{1}{x+1}\right)^x \frac{1}{x+1} = \frac{1}{2ex} - \frac{1}{4ex^2} + \mathcal{O}\left(\frac{1}{x^3}\right) .$$

Using Lemma 9, we conclude that

$$\sup_{S \in \mathcal{S}_{T=2}} F_{\tau,S}(k) = \frac{1}{2e(k-1)} - \frac{1}{4e(k-1)^2} + \mathcal{O}\left(\frac{1}{k^3}\right) .$$

■

### C.2. Forgetting with $T \geq 3$ tasks (Section 5.2)

**Recall Theorem 11.** For any number of tasks $T \geq 3$ under a cyclic ordering $\tau$, the worst-case forgetting after $k = nT \geq T^2$ iterations (*i.e.,* $n \geq T$ cycles), is

$$\frac{T^2}{24ek} \leq \sup_{\substack{S \in \mathcal{S}_{T\geq 3}: \\ \forall m: \operatorname{rank}(\boldsymbol{X}_m) \leq r_{\max}}} F_{\tau,S}(k) \leq \min\left\{\frac{T^2}{\sqrt{k}}, \frac{T^2\left(d - r_{\max}\right)}{2k}\right\}.$$

Moreover, if the cyclic operator $\left(\boldsymbol{P}_T \cdots \boldsymbol{P}_1\right)$ from Eq. (10) is symmetric (*e.g.,* in a *back-and-forth* setting where tasks $m$ and $(T-m)$ are identical $\forall m \in [T]$), then the worst-case forgetting is *sharply* $\sup_{S \in \mathcal{S}_{T\geq 3}} F_{\tau,S}(k) = \Theta\left(T^2/k\right)$.

#### C.2.1. PROVING THE UPPER BOUND

We prove the upper bound using the two following lemmas. The proofs of the lemmas are given on the following pages.

Generally, the proofs revolve around the quantity $\|\boldsymbol{M}^n\boldsymbol{u}\|_2^2 - \left\|\boldsymbol{M}^{n+1}\boldsymbol{u}\right\|_2^2$, that we bound both with and without using the dimension of the tasks.

**Lemma 22 (Dimension-independent upper bounds)** *Let $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_T \in \mathbb{R}^{d\times d}$ be $T$ orthogonal projection operators forming a cyclic operator $\boldsymbol{M} = \boldsymbol{P}_T \cdots \boldsymbol{P}_1 \in \mathbb{R}^{d\times d}$. Then:*

***(Lemma 22a)*** *For any $\boldsymbol{v} \in \mathbb{C}^d$, $m \in [T-1]$, it holds that*

$$\|\left(\boldsymbol{I} - \boldsymbol{P}_m\right)\boldsymbol{v}\|_2^2 \leq m\left(\|\boldsymbol{v}\|_2^2 - \|\boldsymbol{M}\boldsymbol{v}\|_2^2\right);$$

***(Lemma 22b)*** *For any $\boldsymbol{v} \in \mathbb{C}^d$, $m \in [T-1]$, it holds that $\|\left(\boldsymbol{I} - \boldsymbol{M}\right)\boldsymbol{v}\|_2^2 \leq T\left(\|\boldsymbol{v}\|_2^2 - \|\boldsymbol{M}\boldsymbol{v}\|_2^2\right)$;*

***(Lemma 22c)*** *For any vector $\boldsymbol{u} \in \mathbb{C}^d$ holding $\|\boldsymbol{u}\| \leq 1$, after $n \geq 1$ cycles, it holds that*

$$\|\boldsymbol{M}^n\boldsymbol{u}\|_2^2 - \left\|\boldsymbol{M}^{n+1}\boldsymbol{u}\right\|_2^2 \leq 2\left\|\boldsymbol{M}^n - \boldsymbol{M}^{n+1}\right\|;$$

***(Lemma 22d)*** *The forgetting on task $m \in [T-1]$ after $n \geq 1$ cycles is upper bounded by*

$$\|\left(\boldsymbol{I} - \boldsymbol{P}_m\right)\boldsymbol{M}^n\|^2 \leq 2m\left\|\boldsymbol{M}^n - \boldsymbol{M}^{n+1}\right\|;$$

***(Lemma 22e)*** *For any number of cycles $n \geq 1$, it holds that $\left\|\boldsymbol{M}^{n-1} - \boldsymbol{M}^n\right\| \leq \sqrt{T/n}$;*
*Moreover, when $\boldsymbol{M}$ is* symmetric*, we have $\left\|\boldsymbol{M}^{n-1} - \boldsymbol{M}^n\right\| \leq 1/e(n-1)$.*

**Lemma 23 (Rank-dependent upper bound)** *For any vector $\boldsymbol{u} \in \mathbb{C}^d$ holding $\|\boldsymbol{u}\| \leq 1$, and for any non-expansive operator $\boldsymbol{M} \in \mathbb{R}^{d\times d}$, i.e., $\|\boldsymbol{M}\|_2 \leq 1$, it holds that*

$$\|\boldsymbol{M}^n\boldsymbol{u}\|_2^2 - \left\|\boldsymbol{M}^{n+1}\boldsymbol{u}\right\|_2^2 \leq \frac{\operatorname{rank}\left(\boldsymbol{M}\right)}{n} \leq \frac{d}{n}.$$

**Proof** [Proof for Theorem 11 – upper bound]

First, we remind the reader that the forgetting after $n$ cycles is upper bounded by

$$\frac{1}{T}\left\|\begin{bmatrix} I-\boldsymbol{P}_1 \\ \vdots \\ I-\boldsymbol{P}_{T-1} \end{bmatrix}(\boldsymbol{P}_T\cdots\boldsymbol{P}_1)^n\right\|_2^2 \leq \frac{1}{T}\sum_{m=1}^{T-1}\left\|(\boldsymbol{I}-\boldsymbol{P}_m)(\boldsymbol{P}_T\cdots\boldsymbol{P}_1)^n\right\|_2^2 . \tag{13}$$

The dimension-independent bound directly follows from **(Lemma 22d)** and **(Lemma 22e)**:

$$\frac{1}{T}\sum_{m=1}^{T-1}\left\|(\boldsymbol{I}-\boldsymbol{P}_m)(\boldsymbol{P}_T\cdots\boldsymbol{P}_1)^n\right\|_2^2 \leq \underbrace{\frac{1}{T}\sum_{m=1}^{T-1}2m}_{=(T-1)}\sqrt{\frac{T}{n}} = (T-1)\sqrt{\frac{T}{n}} \leq T\sqrt{\frac{T}{n}} = \frac{T^2}{\sqrt{k}} .$$

**A symmetric cyclic operator $\boldsymbol{M}=(\boldsymbol{P}_T\cdots\boldsymbol{P}_1)$.** Consider a symmetric cyclic operator appearing for instance in back-and-forth settings where $\forall m\in[T]$ we have $\mathcal{W}_m=\mathcal{W}_{T-m}$. Then, **(Lemma 22e)** gives a tighter bound, which in turn (assuming $n\geq 2$) yields an overall bound of

$$\frac{1}{T}\sum_{m=1}^{T-1}\left\|(\boldsymbol{I}-\boldsymbol{P}_m)(\boldsymbol{P}_T\cdots\boldsymbol{P}_1)^n\right\|_2^2 \leq \frac{T-1}{e(n-1)} \leq \frac{T}{n} = \frac{T^2}{k} .$$

Finally, since we prove the theorem's $T^2/24ek$ lower bound in Appendix C.2.2 using a back-and-forth task collection, *i.e.,* using a symmetric cyclic operator, we get that in these back-and-forth settings we have a sharp worst-case behavior of $\Theta(T^2/k)$

Finally, we notice that **(Lemma 22a)** implies that

$$\left\|(\boldsymbol{I}-\boldsymbol{P}_m)\boldsymbol{M}^n\right\|_2^2 \triangleq \max_{\boldsymbol{u}:\|\boldsymbol{u}\|_2=1}\left\|(\boldsymbol{I}-\boldsymbol{P}_m)\boldsymbol{M}^n\boldsymbol{u}\right\|_2^2 \leq m\cdot\max_{\boldsymbol{u}:\|\boldsymbol{u}\|_2=1}\left(\left\|\boldsymbol{M}^n\boldsymbol{u}\right\|^2-\left\|\boldsymbol{M}^{n+1}\boldsymbol{u}\right\|^2\right),$$

and by applying Lemma 23 on $\boldsymbol{M}\triangleq\boldsymbol{P}_T\cdots\boldsymbol{P}_1$, we conclude the dimension-dependent bound:

$$\frac{1}{T}\sum_{m=1}^{T-1}\left\|(\boldsymbol{I}-\boldsymbol{P}_m)(\boldsymbol{P}_T\cdots\boldsymbol{P}_1)^n\right\|_2^2 \leq \underbrace{\frac{1}{T}\sum_{m=1}^{T-1}m}_{=(T-1)/2}\frac{\operatorname{rank}(\boldsymbol{M})}{n} \leq \frac{T(d-r_{\max})}{2n} = \frac{T^2(d-r_{\max})}{2k} ,$$

where we used our notation of $r_{\max}\triangleq\max_{m\in[T]}\operatorname{rank}\boldsymbol{X}_m=\max_{m\in[T]}(d-\operatorname{rank}\boldsymbol{P}_m)$ and the fact that $\forall m\in[T]:\operatorname{rank}(\boldsymbol{M})=\operatorname{rank}(\boldsymbol{P}_T\cdots\boldsymbol{P}_1)\leq\min_{m\in[T]}\operatorname{rank}\boldsymbol{P}_m$. ■

Before we prove the lemmas above, we state an auxiliary claim (and a corollary).

**Claim 24** *For any $m \in [T]$, it holds that*

$$\boldsymbol{I} = (\boldsymbol{I} - \boldsymbol{P}_1) + \sum_{\ell=1}^{m-1} (\boldsymbol{I} - \boldsymbol{P}_{\ell+1}) (\boldsymbol{P}_\ell \cdots \boldsymbol{P}_1) + \boldsymbol{P}_m \cdots \boldsymbol{P}_1 \ .$$

**Proof** Recursively, we show that

$$\begin{aligned}
&\boldsymbol{I} - \boldsymbol{P}_m \cdots \boldsymbol{P}_1 \\
&= (\boldsymbol{I} - \boldsymbol{P}_1) + (\boldsymbol{P}_1 - \boldsymbol{P}_2\boldsymbol{P}_1) + (\boldsymbol{P}_2\boldsymbol{P}_1 - \boldsymbol{P}_3\boldsymbol{P}_2\boldsymbol{P}_1) + \cdots + (\boldsymbol{P}_{m-1} \cdots \boldsymbol{P}_1 - \boldsymbol{P}_m \cdots \boldsymbol{P}_1) \\
&= (\boldsymbol{I} - \boldsymbol{P}_1) + (\boldsymbol{I} - \boldsymbol{P}_2) \boldsymbol{P}_1 + (\boldsymbol{I} - \boldsymbol{P}_3) (\boldsymbol{P}_2\boldsymbol{P}_1) + \cdots + (\boldsymbol{I} - \boldsymbol{P}_m) (\boldsymbol{P}_{m-1} \cdots \boldsymbol{P}_1) \\
&= (\boldsymbol{I} - \boldsymbol{P}_1) + \sum_{\ell=1}^{m-1} (\boldsymbol{I} - \boldsymbol{P}_{\ell+1}) (\boldsymbol{P}_\ell \cdots \boldsymbol{P}_1) \ ,
\end{aligned}$$

which proves our claim. ■

**Corollary 25** *For any $m \in [T]$, it holds that,*

$$(\boldsymbol{I} - \boldsymbol{P}_m) = (\boldsymbol{I} - \boldsymbol{P}_m) (\boldsymbol{I} - \boldsymbol{P}_1) + \sum_{\ell=1}^{m-1} (\boldsymbol{I} - \boldsymbol{P}_m) (\boldsymbol{I} - \boldsymbol{P}_{\ell+1}) (\boldsymbol{P}_\ell \cdots \boldsymbol{P}_1) \ .$$

**Proof** We use Claim 24 and get that

$$\begin{aligned}
&(\boldsymbol{I} - \boldsymbol{P}_m) = (\boldsymbol{I} - \boldsymbol{P}_m) \boldsymbol{I} \\
&= (\boldsymbol{I} - \boldsymbol{P}_m) \left[ (\boldsymbol{I} - \boldsymbol{P}_1) + \sum_{\ell=1}^{m-1} (\boldsymbol{I} - \boldsymbol{P}_{\ell+1}) (\boldsymbol{P}_\ell \cdots \boldsymbol{P}_1) + \boldsymbol{P}_m \cdots \boldsymbol{P}_1 \right] \\
&= (\boldsymbol{I} - \boldsymbol{P}_m) \left[ (\boldsymbol{I} - \boldsymbol{P}_1) + \sum_{\ell=1}^{m-1} (\boldsymbol{I} - \boldsymbol{P}_{\ell+1}) (\boldsymbol{P}_\ell \cdots \boldsymbol{P}_1) \right] + \underbrace{(\boldsymbol{I} - \boldsymbol{P}_m) \boldsymbol{P}_m}_{=\boldsymbol{0}, \text{ by Prop. 2}} \boldsymbol{P}_{m-1} \cdots \boldsymbol{P}_1 \ .
\end{aligned}$$

■

**Proof for Lemma 22** . We now prove our dimension-independent upper bounds, step by step, by proving all the statements of Lemma 22.

**Proof** [Proof for **(Lemma 22a)**] For $m = 1$, one can show that:

$$\|(\boldsymbol{I} - \boldsymbol{P}_1)\,\boldsymbol{v}\|_2^2 \overset{\text{idempotence}}{=} \|\boldsymbol{v}\|_2^2 - \|\boldsymbol{P}_1\boldsymbol{v}\|_2^2 \overset{\substack{\text{projections}\\ \text{contract}}}{\leq} 1 \cdot \left(\|\boldsymbol{v}\|_2^2 - \|(\boldsymbol{P}_T \cdots \boldsymbol{P}_1)\,\boldsymbol{v}\|_2^2\right) .$$

Then, for $m = 2, 3, \ldots, T-1$, we apply Corollary 25 and obtain

$$\begin{aligned}
&\|(\boldsymbol{I} - \boldsymbol{P}_m)\,\boldsymbol{v}\|_2^2 \\
[\text{Corollary 25}] &= \left\|(\boldsymbol{I} - \boldsymbol{P}_m)\,(\boldsymbol{I} - \boldsymbol{P}_1)\,\boldsymbol{v} + \sum_{\ell=1}^{m-1} (\boldsymbol{I} - \boldsymbol{P}_m)\,(\boldsymbol{I} - \boldsymbol{P}_{\ell+1})\,(\boldsymbol{P}_\ell \cdots \boldsymbol{P}_1)\,\boldsymbol{v}\right\|^2 \\
[\text{Lemma 16}] &\leq m\left(\|(\boldsymbol{I} - \boldsymbol{P}_m)\,(\boldsymbol{I} - \boldsymbol{P}_1)\,\boldsymbol{v}\|^2 + \sum_{\ell=1}^{m-1} \|(\boldsymbol{I} - \boldsymbol{P}_m)\,(\boldsymbol{I} - \boldsymbol{P}_{\ell+1})\,(\boldsymbol{P}_\ell \cdots \boldsymbol{P}_1)\,\boldsymbol{v}\|^2\right) \\
\left[\substack{\text{projections}\\ \text{contract}}\right] &\leq m\left(\|(\boldsymbol{I} - \boldsymbol{P}_1)\,\boldsymbol{v}\|^2 + \sum_{\ell=1}^{m-1} \|(\boldsymbol{I} - \boldsymbol{P}_{\ell+1})\,(\boldsymbol{P}_\ell \cdots \boldsymbol{P}_1)\,\boldsymbol{v}\|^2\right) \\
[\text{idempotence}] &= m\left(\|\boldsymbol{v}\|^2 - \|\boldsymbol{P}_1\boldsymbol{v}\|^2 + \sum_{\ell=1}^{m-1} \left(\|(\boldsymbol{P}_\ell \cdots \boldsymbol{P}_1)\,\boldsymbol{v}\|^2 - \|(\boldsymbol{P}_{\ell+1} \cdots \boldsymbol{P}_1)\,\boldsymbol{v}\|^2\right)\right) \\
[\text{telescoping}] &= m\left(\|\boldsymbol{v}\|^2 - \|(\boldsymbol{P}_m \cdots \boldsymbol{P}_1)\,\boldsymbol{v}\|^2\right) \leq m\left(\|\boldsymbol{v}\|^2 - \|(\boldsymbol{P}_T \cdots \boldsymbol{P}_1)\,\boldsymbol{v}\|^2\right) ,
\end{aligned}$$

which completes our proof. ■

**Remark.** We note in passing that this dependence on $m$ can be further improved. Using similar techniques, one can also prove that $\|(\boldsymbol{I} - \boldsymbol{P}_m)\,\boldsymbol{M}\boldsymbol{v}\|_2^2 \leq (T - m)\left(\|\boldsymbol{v}\|^2 - \|\boldsymbol{M}\boldsymbol{v}\|^2\right)$. This in turn can help tighten the upper bound in Theorem 11 by a multiplicative factor of 2, but yields a slightly less elegant expression.

**Proof** [Proof for **(Lemma 22b)**] Notice that Claim 24 (with $m = T$) implies that

$$\boldsymbol{I} - \boldsymbol{P}_T \cdots \boldsymbol{P}_1 = (\boldsymbol{I} - \boldsymbol{P}_1) + \sum_{\ell=1}^{T-1} (\boldsymbol{I} - \boldsymbol{P}_{\ell+1}) (\boldsymbol{P}_\ell \cdots \boldsymbol{P}_1) \ .$$

Then, we prove our lemma:

$$\begin{aligned}
\|\boldsymbol{v} - \boldsymbol{P}_T \cdots \boldsymbol{P}_1 \boldsymbol{v}\|^2 &= \left\| (\boldsymbol{I} - \boldsymbol{P}_1) \boldsymbol{v} + \sum_{\ell=1}^{T-1} (\boldsymbol{I} - \boldsymbol{P}_{\ell+1}) (\boldsymbol{P}_\ell \cdots \boldsymbol{P}_1) \boldsymbol{v} \right\|^2 \\
[\text{Lemma 16}] &\leq T \left( \|(\boldsymbol{I} - \boldsymbol{P}_1) \boldsymbol{v}\|^2 + \sum_{\ell=1}^{T-1} \|(\boldsymbol{I} - \boldsymbol{P}_{\ell+1}) (\boldsymbol{P}_\ell \cdots \boldsymbol{P}_1) \boldsymbol{v}\|^2 \right) \\
[\text{idempotence}] &= T \left( \|\boldsymbol{v}\|^2 - \|\boldsymbol{P}_1 \boldsymbol{v}\|^2 + \sum_{\ell=1}^{T-1} \left( \|(\boldsymbol{P}_\ell \cdots \boldsymbol{P}_1) \boldsymbol{v}\|^2 - \|(\boldsymbol{P}_{\ell+1} \cdots \boldsymbol{P}_1) \boldsymbol{v}\|^2 \right) \right) \\
[\text{telescoping}] &= T \left( \|\boldsymbol{v}\|^2 - \|(\boldsymbol{P}_T \cdots \boldsymbol{P}_1) \boldsymbol{v}\|^2 \right) \ .
\end{aligned}$$

■

**Remark.** We note in passing that after completing our proof above, we found that a similar proof was already presented in Netyanun and Solmon [48] while discussing Kakutani's lemma. We still brought our proof here for the sake of completeness. Moreover, they showed that for a non-expansive *self-adjoint positive semi-definite* operator $\boldsymbol{M}$, the factor $T$ can be alleviated from the inequality. That is, $\|\boldsymbol{v} - \boldsymbol{P}_T \cdots \boldsymbol{P}_1 \boldsymbol{v}\|^2 \leq \mathcal{X} \left( \|\boldsymbol{v}\|^2 - \|(\boldsymbol{P}_T \cdots \boldsymbol{P}_1) \boldsymbol{v}\|^2 \right)$. But clearly this does not suit our general cyclic operators which are not necessarily self-adjoint. Indeed, our proof for **(Lemma 22e)** yields a similar conclusion for *symmetric* cyclic operators.

**Proof** [Proof for **(Lemma 22c)**] Let $\boldsymbol{u} \in \mathbb{C}^d$. First, we notice that since $\boldsymbol{M}$ is a non-expansive operator, we have that $\|\boldsymbol{M}^n \boldsymbol{u}\|_2^2 \geq \left\|\boldsymbol{M}^{n+1} \boldsymbol{u}\right\|_2^2$. In turn, this means

$$0 \leq \|\boldsymbol{M}^n \boldsymbol{u}\|_2^2 - \left\|\boldsymbol{M}^{n+1} \boldsymbol{u}\right\|_2^2 = \boldsymbol{u}^* \Big( \underbrace{\left(\boldsymbol{M}^n\right)^* \boldsymbol{M}^n - \left(\boldsymbol{M}^{n+1}\right)^* \boldsymbol{M}^{n+1}}_{\succeq \mathbf{0} \text{ because it is clearly symmetric as well}} \Big) \boldsymbol{u} \ .$$

Then, we use the fact that for any positive semi-definite matrix $\boldsymbol{A} \succeq \mathbf{0}$, we have that $\boldsymbol{u}^* \boldsymbol{A} \boldsymbol{u} \leq \lambda_1(\boldsymbol{A}) = \|\boldsymbol{A}\|$, and get

$$\begin{aligned}
\|\boldsymbol{M}^n \boldsymbol{u}\|_2^2 - \left\|\boldsymbol{M}^{n+1} \boldsymbol{u}\right\|_2^2 &\leq \left\| \left(\boldsymbol{M}^n\right)^* \boldsymbol{M}^n - \left(\boldsymbol{M}^{n+1}\right)^* \boldsymbol{M}^{n+1} \right\| \\
[\boldsymbol{A}^*\boldsymbol{A} - \boldsymbol{B}^*\boldsymbol{B} = (\boldsymbol{A}^* - \boldsymbol{B}^*)\boldsymbol{A} + \boldsymbol{B}^*(\boldsymbol{A} - \boldsymbol{B})] &= \left\| \left(\boldsymbol{M}^n\right)^* \left(\boldsymbol{M}^n - \boldsymbol{M}^{n+1}\right) + \left( \left(\boldsymbol{M}^n\right)^* - \left(\boldsymbol{M}^{n+1}\right)^* \right) \boldsymbol{M}^n \right\| \\
[\text{triangle inequality}] &\leq \left\| \left(\boldsymbol{M}^n\right)^* \left(\boldsymbol{M}^n - \boldsymbol{M}^{n+1}\right) \right\| + \left\| \left( \left(\boldsymbol{M}^n\right)^* - \left(\boldsymbol{M}^{n+1}\right)^* \right) \boldsymbol{M}^n \right\| \\
[\ \|\boldsymbol{A}^*\| = \|\boldsymbol{A}\|\ ] &\leq 2 \left\| \left(\boldsymbol{M}^n\right)^* \left(\boldsymbol{M}^n - \boldsymbol{M}^{n+1}\right) \right\| \\
[\text{contraction}] &\leq 2 \left\| \boldsymbol{M}^n - \boldsymbol{M}^{n+1} \right\| \ .
\end{aligned}$$

■

**Proof** [Proof for (Lemma 22d)] This lemma is a direct corollary of (Lemma 22a) and (Lemma 22c).

First, we set $\boldsymbol{v} = \boldsymbol{M}^n \boldsymbol{u}$ and apply (Lemma 22a):

$$\left\|\left(\boldsymbol{I}-\boldsymbol{P}_m\right)\boldsymbol{M}^n\right\|^2 = \max_{\boldsymbol{u}:\|\boldsymbol{u}\|_2=1} \left\|\left(\boldsymbol{I}-\boldsymbol{P}_m\right)\boldsymbol{M}^n\boldsymbol{u}\right\|^2 \leq m \cdot \max_{\boldsymbol{u}:\|\boldsymbol{u}\|_2=1} \left(\left\|\boldsymbol{M}^n\boldsymbol{u}\right\|_2^2 - \left\|\boldsymbol{M}^{n+1}\boldsymbol{u}\right\|_2^2\right) .$$

Then, we apply (Lemma 22c) to get:

$$\left\|\left(\boldsymbol{I}-\boldsymbol{P}_m\right)\boldsymbol{M}^n\right\|^2 \leq m \cdot \max_{\boldsymbol{u}:\|\boldsymbol{u}\|_2=1} \left(\left\|\boldsymbol{M}^n\boldsymbol{u}\right\|_2^2 - \left\|\boldsymbol{M}^{n+1}\boldsymbol{u}\right\|_2^2\right) \leq 2m \left\|\boldsymbol{M}^n - \boldsymbol{M}^{n+1}\right\| .$$

■

**Proof** [Proof for (Lemma 22e)] We use (Lemma 22b) to show that

$$\begin{aligned}\sum_{t=0}^{n-1} \left\|\left(\boldsymbol{M}^t - \boldsymbol{M}^{t+1}\right)\boldsymbol{v}\right\|^2 = \sum_{t=0}^{n-1} \left\|\left(\boldsymbol{I}-\boldsymbol{M}\right)\boldsymbol{M}^t\boldsymbol{v}\right\|^2 &\leq T \sum_{t=0}^{n-1} \left(\left\|\boldsymbol{M}^t\boldsymbol{v}\right\|^2 - \left\|\boldsymbol{M}^{t+1}\boldsymbol{v}\right\|^2\right) \\ \text{[telescoping]} &= T\left(\|\boldsymbol{v}\|^2 - \left\|\boldsymbol{M}^n\boldsymbol{v}\right\|^2\right) \leq T\|\boldsymbol{v}\|^2 .\end{aligned}$$

Since $\left\|\left(\boldsymbol{M}^t - \boldsymbol{M}^{t+1}\right)\boldsymbol{v}\right\|^2 = \left\|\boldsymbol{M}\left(\left(\boldsymbol{M}^{t-1} - \boldsymbol{M}^t\right)\boldsymbol{v}\right)\right\|^2$ and since the cyclic operator $\boldsymbol{M}$ is a contraction operator, we get that the series $\left\{\left\|\left(\boldsymbol{M}^t - \boldsymbol{M}^{t+1}\right)\boldsymbol{v}\right\|^2\right\}_t$ is monotonic non-increasing. This in turn means that

$$\begin{aligned}\forall \boldsymbol{v} \in \mathbb{C}^d : \left\|\left(\boldsymbol{M}^{n-1} - \boldsymbol{M}^n\right)\boldsymbol{v}\right\|^2 &= \min_{t=0,\ldots,n-1} \left\|\left(\boldsymbol{M}^t - \boldsymbol{M}^{t+1}\right)\boldsymbol{v}\right\|^2 \\ &\leq \frac{1}{n}\sum_{t=0}^{n-1} \left\|\left(\boldsymbol{M}^t - \boldsymbol{M}^{t+1}\right)\boldsymbol{v}\right\|^2 \leq \frac{T}{n}\|\boldsymbol{v}\|^2 ,\end{aligned}$$

which finally implies that $\left\|\boldsymbol{M}^{n-1} - \boldsymbol{M}^n\right\|^2 \leq T/n$ as required.

**A symmetric cyclic operator $\boldsymbol{M} = (\boldsymbol{P}_T \cdots \boldsymbol{P}_1)$.** When $\boldsymbol{M}$ is Hermitian matrix, its spectral decomposition $\boldsymbol{M} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^*$ reveals a tighter bound:

$$\left\|\boldsymbol{M}^{n-1} - \boldsymbol{M}^n\right\| = \left\|\boldsymbol{Q}(\boldsymbol{\Lambda}^{n-1} - \boldsymbol{\Lambda}^n)\boldsymbol{Q}^*\right\| = \left\|\boldsymbol{\Lambda}^{n-1} - \boldsymbol{\Lambda}^n\right\| = \max_i \left(\lambda_i^{n-1}\left(1-\lambda_i\right)\right) \leq \frac{1}{e(n-1)} .$$

■

**Proof** [Proof for Lemma 23] We follow a proof by fedja [17].

Define the series of matrices $\{\boldsymbol{B}_n\}$, where $\boldsymbol{B}_n = \boldsymbol{M}^{n^\top}\boldsymbol{M}^n - \boldsymbol{M}^{n+1^\top}\boldsymbol{M}^{n+1}$. Notice that it holds that $\boldsymbol{B}_n \succeq 0$ (since it is symmetric and $\forall \boldsymbol{v} : \boldsymbol{v}^*\boldsymbol{B}_n\boldsymbol{v} = \|\boldsymbol{M}^n\boldsymbol{v}\|_2^2 - \left\|\boldsymbol{M}^{n+1}\boldsymbol{v}\right\|_2^2 \geq 0$).

Furthermore, Von Neumann's trace inequality and the fact that $\boldsymbol{M}$ is a contraction operator, imply that

$$\operatorname{tr}\left(\boldsymbol{B}_{n+1}\right) = \operatorname{tr}\left(\boldsymbol{M}^\top\boldsymbol{B}_n\boldsymbol{M}\right) = \operatorname{tr}\left(\underbrace{\boldsymbol{M}\boldsymbol{M}^\top}_{\succeq 0}\underbrace{\boldsymbol{B}_n}_{\succeq 0}\right) \leq \left\|\boldsymbol{M}\boldsymbol{M}^\top\right\|_2 \operatorname{tr}\left(\boldsymbol{B}_n\right) \leq \operatorname{tr}\left(\boldsymbol{B}_n\right) \ ,$$

meaning that the sequence $\{\operatorname{tr}\left(\boldsymbol{B}_n\right)\}_n$ is monotonically non-increasing. Hence, we get that

$$\begin{aligned}
n\operatorname{tr}\left(\boldsymbol{B}_n\right) &\leq \sum_{\ell=1}^{n}\operatorname{tr}\left(\boldsymbol{B}_\ell\right) = \sum_{\ell=1}^{n}\operatorname{tr}\left(\boldsymbol{M}^{\ell^\top}\boldsymbol{M}^\ell - \boldsymbol{M}^{\ell+1^\top}\boldsymbol{M}^{\ell+1}\right) \\
&= \sum_{\ell=1}^{n}\left(\operatorname{tr}\left(\boldsymbol{M}^{\ell^\top}\boldsymbol{M}^\ell\right) - \operatorname{tr}\left(\boldsymbol{M}^{\ell+1^\top}\boldsymbol{M}^{\ell+1}\right)\right) \\
[\text{telescoping}] &= \operatorname{tr}\left(\boldsymbol{M}^\top\boldsymbol{M}\right) - \underbrace{\operatorname{tr}\left(\boldsymbol{M}^{n+1^\top}\boldsymbol{M}^{n+1}\right)}_{\geq 0} \leq \operatorname{tr}\left(\boldsymbol{M}^\top\boldsymbol{M}\right) \overset{\text{contraction}}{\leq} \operatorname{rank}\left(\boldsymbol{M}\right) \\
\implies \operatorname{tr}\left(\boldsymbol{B}_n\right) &\leq \frac{\operatorname{rank}\left(\boldsymbol{M}\right)}{n} \ .
\end{aligned}$$

We are now ready to conclude this lemma,

$$\begin{aligned}
\max_{\boldsymbol{v}:\|\boldsymbol{v}\|_2=1}\left(\|\boldsymbol{M}^n\boldsymbol{v}\|_2^2 - \left\|\boldsymbol{M}^{n+1}\boldsymbol{v}\right\|_2^2\right) &= \max_{\boldsymbol{v}:\|\boldsymbol{v}\|_2=1}\boldsymbol{v}^*\left(\boldsymbol{M}^{n^\top}\boldsymbol{M}^n - \boldsymbol{M}^{n+1^\top}\boldsymbol{M}^{n+1}\right)\boldsymbol{v} \\
&= \max_{\boldsymbol{v}:\|\boldsymbol{v}\|_2=1}\boldsymbol{v}^*\underbrace{\boldsymbol{B}_n}_{\succeq 0}\boldsymbol{v} = \underbrace{\lambda_1\left(\boldsymbol{B}_n\right)}_{\in\mathbb{R}_+} \leq \operatorname{tr}\left(\boldsymbol{B}_n\right) \leq \frac{\operatorname{rank}\left(\boldsymbol{M}\right)}{n} \ .
\end{aligned}$$

■

**Remarks.** For the sake of completeness, we briefly discuss our result above. First, this result is useful for our derivations since our cyclic operator $(\boldsymbol{P}_T \cdots \boldsymbol{P}_1)$ is essentially a contraction operator. In fact, being a product of projections, our cyclic operator has great expressiveness. It was shown by Oikhberg [50] that *any* contraction operator can be decomposed into a product of sufficiently, though sometimes infinitely, many projections. This explains our interest in analyzing "general" contraction operators.

Finally, we should mention that the bound derived here is sharp (up to a constant). For instance, for the Toeplitz operator $\boldsymbol{M} = \begin{bmatrix} & & 1 \\ & \iddots & \\ 1-\epsilon & & \end{bmatrix}$, a proper choice of $\epsilon$ can yield a rate of $d/en$ for the quantity we bound in the lemma above. However, such an operator cannot be expressed as a product of a *finite* number of projection operators. One can approximate this operator arbitrarily-well by replacing the 1s with $\beta \in (0,1)$. However, we were not able to use this approximation to improve our lower bound in Theorem 11, since $\beta < 1$ implies a geometric contraction of *all* elements at every cycle and requires a very large number of constructing projections (*i.e.,* tasks).

### C.2.2. PROVING THE LOWER BOUND

**Proof** [Proof for Theorem 11 – lower bound]

To prove the lower bound, we show a construction of a task collection $S \in \mathcal{S}_T$ with $T$ tasks whose forgetting is

$$\frac{T^2}{24ek} \leq F_{\tau,S}(k) \,.$$

We construct $S$ using data matrices of rank $r = d-1$ whose non-zero singular values all equal to 1. Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_T$ be the normalized vectors spanning the $T$ rank-*one* solution spaces $\mathcal{W}_1, \ldots, \mathcal{W}_T$. We also choose to generate labels using a *unit norm* offline solution, *i.e.,* $\|\boldsymbol{w}^\star\| = 1$. We spread $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_T$ on a 2-dimensional hyperplane such that the tasks "go" back-and-forth, *i.e.,* switch direction at the middle task (see illustration in Figure 6).

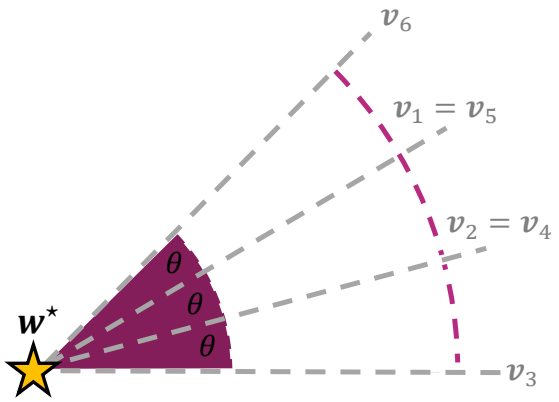
Using Lemma 21 we have

$$F_{\tau,S}\left(k\right) \geq \frac{1}{k}\left(\prod_{i=1}^{k-1}\cos^2\theta_{\tau(i,i+1)}\right)\sum_{m=1}^{k-1}\left(1-\cos^2\theta_{\tau(m,k)}\right)$$

$$= \frac{1}{k}\left(\prod_{i=1}^{k-1}\cos^2\theta_{\tau(i,i+1)}\right)\frac{k}{T}\sum_{m=1}^{T}\left(1-\cos^2\theta_{m,T}\right) = \frac{1}{T}\left(\prod_{i=1}^{k-1}\cos^2\theta_{\tau(i,i+1)}\right)^{k-1}\sum_{m=1}^{T}\left(1-\cos^2\theta_{m,T}\right).$$

We consider separately the case of even $T$ and the case of odd $T$.

Figure 6: Construction of $T = 6$ "back-and-forth" tasks for the lower bound proof in Theorem 11.



**Even $T$:** In this case, we set $\forall i \in [T-1]:\ \theta_{i,i+1} = \begin{cases} \theta & i < \frac{T}{2} \\ -\theta & i \geq \frac{T}{2} \end{cases}$, thus obtaining

$$\frac{1}{T}\left(\prod_{i=1}^{k-1}\cos^2\theta_{\tau(i,i+1)}\right)^{k-1}\sum_{m=1}^{T}\left(1-\cos^2\theta_{m,T}\right) = \frac{1}{T}\left(\cos^2\theta\right)^{k-1}\sum_{m=1}^{T}\left(1-\cos^2\left(\left(\frac{T}{2}-m\right)\theta\right)\right)\ .$$

We choose $\theta = \sqrt{1/k-1}$ and get

$$\begin{aligned}&\frac{1}{T}\left(\cos^2\theta\right)^{k-1}\sum_{m=1}^{T}\left(1-\cos^2\left(\left(\frac{T}{2}-m\right)\theta\right)\right)\\ &= \frac{1}{T}\left(\cos^2\sqrt{\frac{1}{k-1}}\right)^{k-1}\sum_{m=1}^{T}\left(1-\cos^2\left(\left(\frac{T}{2}-m\right)\sqrt{\frac{1}{k-1}}\right)\right).\end{aligned}$$

Next, using the inequality $1-\cos^2 x \geq x^2 - \frac{1}{3}x^4$, we get

$$\begin{aligned}&\frac{1}{T}\left(\cos^2\theta\right)^{k-1}\sum_{m=1}^{T}\left(1-\cos^2\left(\left(\frac{T}{2}-m\right)\theta\right)\right)\\ &\geq \frac{1}{T}\left(\cos^2\sqrt{\frac{1}{k-1}}\right)^{k-1}\sum_{m=1}^{T}\left(\left(\frac{T}{2}-m\right)^2\frac{1}{k-1}-\frac{1}{3}\left(\frac{T}{2}-m\right)^4\frac{1}{(k-1)^2}\right)\\ &\geq \frac{1}{T}\frac{1}{k-1}\left(\cos^2\sqrt{\frac{1}{k-1}}\right)^{k-1}\sum_{m=1}^{T}\left(\left(\frac{T}{2}-m\right)^2-\left(\frac{T}{2}-m\right)^4\frac{1}{3\,(k-1)}\right)\ .\end{aligned}$$

By using $\forall x \geq 1:\ \frac{1}{x}\left(\cos^2\sqrt{1/x}\right)^x \geq 1/e(x+1)$ we get

$$\begin{aligned}&\frac{1}{T}\left(\cos^2\theta\right)^{k-1}\sum_{m=1}^{T}\left(1-\cos^2\left(\left(\frac{T}{2}-m\right)\theta\right)\right)\\ &\geq \frac{1}{Tek}\sum_{m=1}^{T}\left(\left(\frac{T}{2}-m\right)^2-\left(\frac{T}{2}-m\right)^4\frac{1}{3\,(k-1)}\right)\\ &= \frac{1}{Tek}\left(\frac{1}{12}T\left(T^2+2\right)-\frac{T\left(3T^4+20T^2-8\right)}{240\cdot 3(k-1)}\right) = \frac{1}{12ek}\left(T^2+2-\frac{3T^4+20T^2-8}{60\,(k-1)}\right)\ .\end{aligned}$$

For $k \geq T^2$ we get

$$\frac{1}{T}\left(\cos^2\theta\right)^{k-1}\sum_{m=1}^{T}\left(1-\cos^2\left(\left(\frac{T}{2}-m\right)\theta\right)\right) \geq \frac{1}{12ek}\left(T^2+2-\frac{3T^4+20T^2-8}{60\,(T^2-1)}\right)\ . \quad (14)$$

Note that for $T \geq 4$ we have $\frac{1}{12ek}\left(T^2+2-\frac{3T^4+20T^2-8}{60(T^2-1)}\right) \geq \frac{T^2}{13ek}$.

Therefore for even $T \geq 4$ we have

$$F_{\tau,S}(k) \geq \frac{1}{T}\left(\cos^2\theta\right)^{k-1}\sum_{m=1}^{T}\left(1-\cos^2\left(\left(\frac{T}{2}-m\right)\theta\right)\right) \geq \frac{T^2}{13ek}\,.$$

**Odd $T$:** In this case we set $\forall i \in [T-1]: \ \theta_{i,i+1} = \begin{cases} \theta & i < \frac{T-1}{2} \\ -\frac{\theta}{2} & i = \frac{T-1}{2} \text{ or } i = \frac{T+1}{2} \\ -\theta & i > \frac{T+1}{2} \end{cases}$, and since $\cos(\theta/2) \geq \cos(\theta)$ (for small $\theta$, as we choose later) we have

$$\frac{1}{T}\left(\prod_{i=1}^{k-1}\cos^2\theta_{\tau(i,i+1)}\right)^{k-1}\sum_{m=1}^{T}\left(1-\cos^2\theta_{m,T}\right) \geq \frac{1}{T}\left(\cos^2\theta\right)^{k-1}\sum_{m=1}^{T}\left(1-\cos^2\theta_{m,T}\right) .$$

Next,

$$\sum_{m=1}^{T}\left(1-\cos^2\theta_{m,T}\right) = 1-\cos^2\left(\left(\frac{T-2}{2}\right)\theta\right)+\sum_{m=1}^{T-1}\left(1-\cos^2\left(\left(\frac{T-1}{2}-m\right)\theta\right)\right) .$$

Therefore we get

$$\begin{aligned}
&\frac{1}{T}\left(\cos^2\theta\right)^{k-1}\sum_{m=1}^{T}\left(1-\cos^2\theta_{m,T}\right) \\
=&\frac{1}{T}\left(\cos^2\theta\right)^{k-1}\left(1-\cos^2\left(\left(\frac{T-2}{2}\right)\theta\right)\right) \\
&+\frac{1}{T}\left(\cos^2\theta\right)^{k-1}\sum_{m=1}^{T-1}\left(1-\cos^2\left(\left(\frac{T-1}{2}-m\right)\theta\right)\right) \\
=&\frac{1}{T}\left(\cos^2\theta\right)^{k-1}\left(1-\cos^2\left(\left(\frac{T-2}{2}\right)\theta\right)\right) \\
&+\frac{T-1}{T}\frac{1}{T-1}\left(\cos^2\theta\right)^{k-1}\sum_{m=1}^{T-1}\left(1-\cos^2\left(\left(\frac{T-1}{2}-m\right)\theta\right)\right) . \qquad (15)
\end{aligned}$$

We choose again $\theta = \sqrt{1/k-1}$. For the first term in Eq. (15) we have

$$\begin{aligned}
&\frac{1}{T}\left(\cos^2\theta\right)^{k-1}\left(1-\cos^2\left(\left(\frac{T-2}{2}\right)\theta\right)\right) \\
&= \frac{1}{T}\left(\cos^2\sqrt{\frac{1}{k-1}}\right)^{k-1}\left(1-\cos^2\left(\left(\frac{T-2}{2}\right)\sqrt{\frac{1}{k-1}}\right)\right) \\
&\geq \frac{1}{T}\left(\cos^2\sqrt{\frac{1}{k-1}}\right)^{k-1}\left(1-\cos^2\left(\frac{1}{2}\sqrt{\frac{1}{k-1}}\right)\right) \geq \frac{1}{4Tek} .
\end{aligned}$$

For the second term in Eq. (15) we use Eq. (14) and get

$$\begin{aligned}
&\frac{T-1}{T}\frac{1}{T-1}\left(\cos^2\theta\right)^{k-1}\sum_{m=1}^{T-1}\left(1-\cos^2\left(\left(\frac{T-1}{2}-m\right)\theta\right)\right) \\
&\geq \frac{T-1}{T}\frac{1}{12ek}\left((T-1)^2+2-\frac{3(T-1)^4+20(T-1)^2-8}{60\left((T-1)^2-1\right)}\right) .
\end{aligned}$$

Therefore in Eq. (15) we have

$$\begin{aligned}
&\frac{1}{T}\left(\cos^2\theta\right)^{k-1}\sum_{m=1}^{T}\left(1-\cos^2\theta_{m,T}\right)\\
&\geq\frac{1}{4Tek}+\frac{T-1}{T}\frac{1}{12ek}\left(\left(T-1\right)^2+2-\frac{3\left(T-1\right)^4+20\left(T-1\right)^2-8}{60\left(\left(T-1\right)^2-1\right)}\right)\\
&=\frac{1}{ek}\left(\frac{1}{4T}+\frac{T-1}{12T}\left(\left(T-1\right)^2+2-\frac{3\left(T-1\right)^4+20\left(T-1\right)^2-8}{60\left(\left(T-1\right)^2-1\right)}\right)\right)\\
[T\geq 3]&\geq\frac{T^2}{24ek}\,,
\end{aligned}$$

and thus for odd $T \geq 3$ we have $F_{\tau,S}(k) \geq {}^{T^2}/_{24ek}$, which concludes the proof.

**Regarding the rank.** Notice that the lower bound we derived used a construction consisting of tasks of rank $d-1$. On the other hand, the lower bound in Theorem 11 should be correct even when $r_{\max} < d-1$. In such a case, we can always use an identical task collection using $T$ tasks spread on a 2-d hyperplane, and add to the tasks' data matrices any number of directions which are orthogonal to that hyperplane, but appear identically on *all* tasks. Then, any expression of the form $\left\|\left(\boldsymbol{I}-\boldsymbol{P}_m\right)\left(\boldsymbol{P}_T\cdots\boldsymbol{P}_1\right)^n\right\|$ will disregard these added directions, since they are shared across all tasks (hence they do not appear in $\boldsymbol{I}-\boldsymbol{P}_m$). The remaining hyperplane's behavior will remain unchanged as in the analysis above. Thus our analysis above is valid for any task rank. ■

## Appendix D. Supplementary material: Random task orderings (Section 6)

**Recall Theorem 13.** Under the uniform i.i.d. task ordering $\tau$, the worst-case expected forgetting after $k$ iterations is

$$\sup_{\substack{S\in\mathcal{S}_T:\\ \frac{1}{T}\sum_{m=1}^{T}\operatorname{rank}(\boldsymbol{X}_m)=r_{\text{avg}}}} \bar{F}_{\tau,S}\left(k\right) \leq \frac{9\left(d-r_{\text{avg}}\right)}{k} \ .$$

**Proof** [Proof for Theorem 13] Let $S=\{\boldsymbol{P}_1,\ldots,\boldsymbol{P}_T\}$ be a given collection of $T$ arbitrary projection matrices sampled i.i.d. at each iteration from a uniform distribution.

As in Eq. (8), we first express the expected forgetting on a task collection $S\in\mathcal{S}_T$, defined in Definition 12, in terms of the projection matrices induced by the $T$ data matrices in $S$. That is, we bound the expected forgetting by

$$\bar{F}_{\tau,S}\left(k\right) \triangleq \mathop{\mathbb{E}}_{\tau}\left[\frac{1}{k}\sum_{t=1}^{k}\left\|\boldsymbol{X}_{\tau(t)}\boldsymbol{w}_k-\boldsymbol{y}_{\tau(t)}\right\|^2\right] \leq \mathop{\mathbb{E}}_{\tau}\left[\frac{1}{k}\sum_{t=1}^{k}\left\|\left(\boldsymbol{I}-\boldsymbol{P}_{\tau(t)}\right)\boldsymbol{P}_{\tau(k)}\cdots\boldsymbol{P}_{\tau(1)}\right\|^2\right] . \quad (16)$$

### D.1. Detour: Rephrasing and bounding a more natural expression

#### D.1.1. REPHRASING

Before we will bound the upper bound above, we start by bounding a slightly different quantity which is easier to work with and can be seen as a straightforward MSE loss over $T$ tasks (as studied in NLMS papers, *e.g.,* [61, 67]). Instead of averaging the forgetting (*i.e.,* residuals) over *previously-seen* tasks, we average over *all* tasks in the collection $S$. That is, we start by bounding.

$$\mathop{\mathbb{E}}_{\tau}\left[\frac{1}{T}\sum_{m=1}^{T}\left\|\left(\boldsymbol{I}-\boldsymbol{P}_m\right)\boldsymbol{P}_{\tau(k)}\cdots\boldsymbol{P}_{\tau(1)}\right\|_2^2\right] = \mathop{\mathbb{E}}_{\tau,\boldsymbol{P}}\left\|\left(\boldsymbol{I}-\boldsymbol{P}\right)\boldsymbol{P}_{\tau(k)}\cdots\boldsymbol{P}_{\tau(1)}\right\|_2^2 , \quad (17)$$

where $\mathbb{E}_{\boldsymbol{P}}$ means averaging over all $T$ projection matrices w.r.t. a uniform distribution.

To further ease reading, we slightly abuse notation such that instead of explicitly sampling a task ordering $\tau$, we simply uniformly sample $k+1$ orthogonal projection operators, $\boldsymbol{P}_1,\boldsymbol{P}_2,\ldots,\boldsymbol{P}_k$ and $\boldsymbol{P}$ from some i.i.d. uniform distribution (*e.g.,* uniformly from a finite set of $T$ such operators). Finally, we obtain the following expression:

$$\mathbb{E}_{\boldsymbol{P}_1,\ldots,\boldsymbol{P}_k,\boldsymbol{P}}\left\|\left(\boldsymbol{I}-\boldsymbol{P}\right)\boldsymbol{P}_k\cdots\boldsymbol{P}_1\right\|_2^2 . \quad (18)$$

Next, we will bound Eq. (18) which is equivalent to Eq. (17). Then, we will use Eq. (17) to bound the right hand side of Eq. (16) and conclude an upper bound for the expected forgetting, as required.

### D.1.2. BOUNDING EQ. (18)

We begin by using a known inequality between the spectral norm and the Frobenius norm, that is, $\|\boldsymbol{A}\|_2^2 = \lambda_1\left(\boldsymbol{A}^*\boldsymbol{A}\right) \leq \mathrm{tr}\left(\boldsymbol{A}^*\boldsymbol{A}\right) = \|\boldsymbol{A}\|_F^2$, and present an easier, but perhaps looser, surrogate quantity $f(k)$ which we will bound:

$$\underset{\boldsymbol{P}_1,\ldots,\boldsymbol{P}_k,\boldsymbol{P}}{\mathbb{E}} \left\|\left(\boldsymbol{I}-\boldsymbol{P}\right)\boldsymbol{P}_k\cdots\boldsymbol{P}_1\right\|_2^2 \leq \underset{\boldsymbol{P}_1,\ldots,\boldsymbol{P}_k,\boldsymbol{P}}{\mathbb{E}} \mathrm{tr}\left[\boldsymbol{P}_1\cdots\boldsymbol{P}_k\left(\boldsymbol{I}-\boldsymbol{P}\right)\boldsymbol{P}_k\cdots\boldsymbol{P}_1\right] \triangleq f(k) \ . \quad (19)$$

Key to our following derivations, is the fact that all the projection matrices including $\boldsymbol{P}$ are identically distributed, allowing us to freely change variables. This property, together with the linearity of the expectation and trace operators, facilitates our derivations.

We now notice that the bound we got is *non-increasing* in $k$, *i.e.,*

$$\begin{aligned}
f(k-1)-f(k) &= \underset{\boldsymbol{P}_1,\ldots,\boldsymbol{P}_{k-1},\boldsymbol{P}}{\mathbb{E}} \mathrm{tr}\left[\boldsymbol{P}_1\cdots\boldsymbol{P}_{k-1}\left(\boldsymbol{I}-\boldsymbol{P}\right)\boldsymbol{P}_{k-1}\cdots\boldsymbol{P}_1\right] - \\
&\qquad \underset{\boldsymbol{P}_1,\ldots,\boldsymbol{P}_k,\boldsymbol{P}}{\mathbb{E}} \mathrm{tr}\left[\boldsymbol{P}_1\cdots\boldsymbol{P}_k\left(\boldsymbol{I}-\boldsymbol{P}\right)\boldsymbol{P}_k\cdots\boldsymbol{P}_1\right] \\
&\overset{\left[\substack{\text{identically}\\ \text{distributed}}\right]}{=} \underset{\boldsymbol{P}_2,\ldots,\boldsymbol{P}_{k+1}}{\mathbb{E}} \mathrm{tr}\left[\boldsymbol{P}_2\cdots\boldsymbol{P}_k\left(\boldsymbol{I}-\boldsymbol{P}_{k+1}\right)\boldsymbol{P}_k\cdots\boldsymbol{P}_2\right] - \\
&\qquad \underset{\boldsymbol{P}_1,\ldots,\boldsymbol{P}_{k+1}}{\mathbb{E}} \mathrm{tr}\left[\boldsymbol{P}_1\cdots\boldsymbol{P}_k\left(\boldsymbol{I}-\boldsymbol{P}_{k+1}\right)\boldsymbol{P}_k\cdots\boldsymbol{P}_1\right] \\
&[\text{linearity}] = \underset{\boldsymbol{P}_1,\ldots,\boldsymbol{P}_{k+1}}{\mathbb{E}} \left[\mathrm{tr}\left[\underbrace{\boldsymbol{P}_2\cdots\boldsymbol{P}_k\left(\boldsymbol{I}-\boldsymbol{P}_{k+1}\right)\boldsymbol{P}_k\cdots\boldsymbol{P}_2}_{\triangleq\boldsymbol{M}}\right] - \right. \\
&\qquad \left. \mathrm{tr}\left[\boldsymbol{P}_1\underbrace{\boldsymbol{P}_2\cdots\boldsymbol{P}_k\left(\boldsymbol{I}-\boldsymbol{P}_{k+1}\right)\boldsymbol{P}_k\cdots\boldsymbol{P}_2}_{\triangleq\boldsymbol{M}}\boldsymbol{P}_1\right]\right] \\
&= \underset{\boldsymbol{P}_1,\ldots,\boldsymbol{P}_k,\boldsymbol{P}}{\mathbb{E}} \left[\underbrace{\mathrm{tr}\left[\boldsymbol{M}\right]-\mathrm{tr}\left[\boldsymbol{P}_1\boldsymbol{M}\boldsymbol{P}_1\right]}_{\geq 0}\right] \geq 0 \ ,
\end{aligned}$$

where in the last inequality we (again) used Von Neumann's trace inequality to show that

$$\mathrm{tr}\left(\boldsymbol{P}_1\boldsymbol{M}\boldsymbol{P}_1\right) = \mathrm{tr}\left(\boldsymbol{P}_1\boldsymbol{P}_1\boldsymbol{M}\right) = \mathrm{tr}\left(\underbrace{\boldsymbol{P}_1}_{\succeq 0}\underbrace{\boldsymbol{M}}_{\succeq 0}\right) \leq \left\|\boldsymbol{P}_1\right\|_2 \mathrm{tr}\left(\boldsymbol{M}\right) = \mathrm{tr}\left(\boldsymbol{M}\right) \ .$$

Since the bounds are non-increasing, we bound $f(k)$ at iteration k, using $f(1), \ldots, f(k)$:

$$
\begin{aligned}
f(k) &= \min_{t\in[k]} f(t) \le \frac{1}{k}\sum_{t=1}^{k} f(t) = \frac{1}{k}\sum_{t=1}^{k} \mathop{\mathbb{E}}_{\boldsymbol{P}_1,\ldots,\boldsymbol{P}_t,\boldsymbol{P}} \operatorname{tr}\left[\boldsymbol{P}_1\cdots\boldsymbol{P}_t\left(\boldsymbol{I}-\boldsymbol{P}\right)\boldsymbol{P}_t\cdots\boldsymbol{P}_1\right] \\
&= \frac{1}{k}\sum_{t=1}^{k}\left(\mathop{\mathbb{E}}_{\boldsymbol{P}_1,\ldots,\boldsymbol{P}_t} \operatorname{tr}\left[\boldsymbol{P}_1\cdots\boldsymbol{P}_t\cdots\boldsymbol{P}_1\right] - \mathop{\mathbb{E}}_{\boldsymbol{P}_1,\ldots,\boldsymbol{P}_t,\boldsymbol{P}} \operatorname{tr}\left[\boldsymbol{P}_1\cdots\boldsymbol{P}_t\boldsymbol{P}\boldsymbol{P}_t\cdots\boldsymbol{P}_1\right]\right) \\
\left[\begin{smallmatrix}\text{identically}\\ \text{distributed}\end{smallmatrix}\right] &= \frac{1}{k}\sum_{t=1}^{k}\left(\mathop{\mathbb{E}}_{\boldsymbol{P}_1,\ldots,\boldsymbol{P}_t} \operatorname{tr}\left[\boldsymbol{P}_1\cdots\boldsymbol{P}_t\cdots\boldsymbol{P}_1\right] - \mathop{\mathbb{E}}_{\boldsymbol{P}_1,\ldots,\boldsymbol{P}_t,\boldsymbol{P}_{t+1}} \operatorname{tr}\left[\boldsymbol{P}_1\cdots\boldsymbol{P}_{t+1}\cdots\boldsymbol{P}_1\right]\right) \\
[\text{telescoping}] &= \frac{1}{k}\mathop{\mathbb{E}}_{\boldsymbol{P}_1}\operatorname{tr}\left[\boldsymbol{P}_1\right] - \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{P}_1,\ldots\boldsymbol{P}_{k+1}} \operatorname{tr}\left[\boldsymbol{P}_1\cdots\boldsymbol{P}_{k+1}\cdots\boldsymbol{P}_1\right]}_{\ge 0,\text{ because } \boldsymbol{P}_1\cdots\boldsymbol{P}_{k+1}\cdots\boldsymbol{P}_1\succeq\boldsymbol{0}} \\
&\le \frac{1}{k}\mathop{\mathbb{E}}_{\boldsymbol{P}}\operatorname{rank}\left(\boldsymbol{P}\right) = \frac{1}{k}\cdot\frac{1}{T}\sum_{m=1}^{T}\left(d-\operatorname{rank}\left(\boldsymbol{X}_m\right)\right) \triangleq \frac{d-r_{\text{avg}}}{k}\,,
\end{aligned}
$$

and conclude that

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\tau}\left[\frac{1}{T}\sum_{m=1}^{T}\left\|\left(\boldsymbol{I}-\boldsymbol{P}_m\right)\boldsymbol{P}_{\tau(k)}\cdots\boldsymbol{P}_{\tau(1)}\right\|_2^2\right] &\overset{17\&18}{=} \mathop{\mathbb{E}}_{\boldsymbol{P}_1,\ldots,\boldsymbol{P}_k,\boldsymbol{P}}\left\|\left(\boldsymbol{I}-\boldsymbol{P}\right)\boldsymbol{P}_k\cdots\boldsymbol{P}_1\right\|_2^2 \\
&\overset{19}{\le} f(k) \le \frac{d-r_{\text{avg}}}{k}\,.
\end{aligned}
$$

### D.2. Back to the expected forgetting

Recall that our initial goal in Eq. (16) was to bound $\mathbb{E}_\tau\left[\frac{1}{k}\sum_{t=1}^{k}\left\|\left(\boldsymbol{I}-\boldsymbol{P}_{\tau(t)}\right)\boldsymbol{P}_{\tau(k)}\cdots\boldsymbol{P}_{\tau(1)}\right\|^2\right]$. Instead, we started by bounding the slightly different quantities from Eq. (17) and (18), *i.e.,* we showed that $\mathbb{E}_{\tau,\boldsymbol{P}}\left\|\left(\boldsymbol{I}-\boldsymbol{P}\right)\boldsymbol{P}_{\tau(k)}\cdots\boldsymbol{P}_{\tau(1)}\right\|_2^2 \leq \frac{d-r_{\text{avg}}}{k}$. We now use the bound we proved to bound the original quantity of interest.

$$
\begin{aligned}
&\mathbb{E}_\tau\left[\left\|\left(\boldsymbol{I}-\boldsymbol{P}_{\tau(t)}\right)\boldsymbol{P}_{\tau(k)}\cdots\boldsymbol{P}_{\tau(t)}\cdots\boldsymbol{P}_{\tau(1)}\right\|_2^2\right]\\
&\quad=\mathbb{E}_\tau\left[\left\|\left(\boldsymbol{I}-\boldsymbol{P}_{\tau(t)}\right)\boldsymbol{P}_{\tau(k)}\cdots\boldsymbol{P}_{\tau(t+1)}\left(\boldsymbol{P}_{\tau(t)}+\boldsymbol{I}-\boldsymbol{I}\right)\boldsymbol{P}_{\tau(t-1)}\cdots\boldsymbol{P}_{\tau(1)}\right\|_2^2\right]\\
[\text{Prop. 1}]&\leq 2\Bigg(\mathbb{E}_\tau\left[\left\|\left(\boldsymbol{I}-\boldsymbol{P}_{\tau(t)}\right)\boldsymbol{P}_{\tau(k)}\cdots\boldsymbol{P}_{\tau(t+1)}\left(\boldsymbol{I}-\boldsymbol{P}_{\tau(t)}\right)\boldsymbol{P}_{\tau(t-1)}\cdots\boldsymbol{P}_1\right\|_2^2\right]+\\
&\qquad+\mathbb{E}_\tau\left\|\left(\boldsymbol{I}-\boldsymbol{P}_{\tau(t)}\right)\boldsymbol{P}_{\tau(k)}\cdots\boldsymbol{P}_{\tau(t+1)}\boldsymbol{P}_{\tau(t-1)}\cdots\boldsymbol{P}_1\right\|_2^2\Bigg)\\
[\text{i.i.d.}]&=2\Bigg(\mathbb{E}_\tau\left\|\left(\boldsymbol{I}-\boldsymbol{P}_{\tau(t)}\right)\boldsymbol{P}_{\tau(k)}\cdots\boldsymbol{P}_{\tau(t+1)}\left(\boldsymbol{I}-\boldsymbol{P}_{\tau(t)}\right)\boldsymbol{P}_{\tau(t-1)}\cdots\boldsymbol{P}_1\right\|_2^2+\\
&\qquad+\underbrace{\mathbb{E}_{\tau,\boldsymbol{P}}\left\|\left(\boldsymbol{I}-\boldsymbol{P}\right)\boldsymbol{P}_{\tau(k-1)}\cdots\boldsymbol{P}_1\right\|_2^2}_{\leq(d-r_{\text{avg}})/(k-1)}\Bigg)\\
[\text{norms}]&\leq 2\left(\mathbb{E}_\tau\left[\left\|\left(\boldsymbol{I}-\boldsymbol{P}_{\tau(t)}\right)\boldsymbol{P}_{\tau(k)}\cdots\boldsymbol{P}_{\tau(t+1)}\right\|_2^2\left\|\left(\boldsymbol{I}-\boldsymbol{P}_{\tau(t)}\right)\boldsymbol{P}_{\tau(t-1)}\cdots\boldsymbol{P}_1\right\|_2^2\right]+\frac{d-r_{\text{avg}}}{k-1}\right)\\
&\quad=2\left(\mathbb{E}_{\tau,\boldsymbol{P}}\left[\left\|\left(\boldsymbol{I}-\boldsymbol{P}\right)\boldsymbol{P}_{\tau(k)}\cdots\boldsymbol{P}_{\tau(t+1)}\right\|_2^2\left\|\left(\boldsymbol{I}-\boldsymbol{P}\right)\boldsymbol{P}_{\tau(t-1)}\cdots\boldsymbol{P}_1\right\|_2^2\right]+\frac{d-r_{\text{avg}}}{k-1}\right)\,.
\end{aligned}
$$

Recall that both the spectral norms above are upper bounded by 1. We upper bound the norm consisting of less projections by 1 and keep the other one.

Assume $t > {}^{(k+1)}/_2$, and thus $t-1 > k-t$. We get, $\mathbb{E}_{\tau,\boldsymbol{P}}\left\|\left(\boldsymbol{I}-\boldsymbol{P}\right)\boldsymbol{P}_{\tau(t-1)}\cdots\boldsymbol{P}_1\right\|_2^2 \leq \frac{d-r_{\text{avg}}}{t-1}$. Similarly, when $t \leq {}^{(k+1)}/_2$ and $t-1 \leq k-t$, we get $\mathbb{E}_{\tau,\boldsymbol{P}}\left\|\left(\boldsymbol{I}-\boldsymbol{P}\right)\boldsymbol{P}_{\tau(k)}\cdots\boldsymbol{P}_{\tau(t+1)}\right\|_2^2 \leq \frac{d-r_{\text{avg}}}{k-t}$.

Overall, we get a final bound of

$$
\begin{aligned}
&\frac{1}{k}\sum_{t=1}^{k-1}\mathbb{E}_{\tau}\left[\left\|\left(\boldsymbol{I}-\boldsymbol{P}_{\tau(t)}\right)\boldsymbol{P}_{\tau(k)}\cdots\boldsymbol{P}_{\tau(t)}\cdots\boldsymbol{P}_{\tau(1)}\right\|_2^2\right]\\
&\leq\frac{2}{k}\left(\left(k-1\right)\frac{d-r_{\text{avg}}}{k-1}+\sum_{t=1}^{\lfloor(k+1)/2\rfloor}\frac{d-r_{\text{avg}}}{k-t}+\sum_{t=\lfloor(k+1)/2\rfloor+1}^{k-1}\frac{d-r_{\text{avg}}}{t-1}\right)\\
&=\frac{2(d-r_{\text{avg}})}{k}\left(1+\sum_{t=k-\lfloor(k+1)/2\rfloor}^{k-1}1/t+\sum_{t=\lfloor(k+1)/2\rfloor}^{k-2}1/t\right)\leq\frac{2(d-r_{\text{avg}})}{k}\left(1+2\sum_{t=\lfloor k/2\rfloor}^{k-1}1/t\right)\\
&\overset{[k\geq 2]}{\leq}\frac{2(d-r_{\text{avg}})}{k}\left(1+2+2\int_{\lfloor k/2\rfloor}^{k-1}\frac{1}{t}dt\right)\leq\frac{2(d-r_{\text{avg}})}{k}\left(3+2\left(\ln\left(k-1\right)-\ln\left(\lfloor k/2\rfloor\right)\right)\right)\\
&=\frac{2(d-r_{\text{avg}})}{k}\left(3+2\ln\left(\frac{k-1}{\lfloor k/2\rfloor}\right)\right)\leq\frac{2(d-r_{\text{avg}})}{k}\underbrace{\left(3+2\ln\left(2\right)\right)}_{\leq 4.5}\leq\frac{9(d-r_{\text{avg}})}{k}\,.
\end{aligned}
$$

■

## Appendix E. Extension to the average iterate (Remark 15)

Here we briefly demonstrate how our results from Sections 5 and 6 can be readily improved by considering the average iterate $\overline{\boldsymbol{w}}_k \triangleq \frac{1}{k}\sum_{t=1}^{k} \boldsymbol{w}_t$ instead of the last iterate $\boldsymbol{w}_k$. The average iterate is known to be easier to analyze, and yields generally stronger results when analyzing the SGD algorithm [71] and the Kaczmard method [43]. Of course, working with the average iterate requires maintaining a running mean parameter vector, which is generally more memory consuming.

### E.1. Proof sketch for the cyclic setting

As a proof of concept, we prove better bounds for an "easier" iterate that averages only iterates at the end of cycles, *i.e.,* for $k=nT$ we define $\overline{\boldsymbol{w}}_n = \frac{1}{n}\sum_{n'=1}^{n} \boldsymbol{w}_{Tn'}$. For the first task, we get

$$\begin{aligned}
\left\|\boldsymbol{X}_1\left(\overline{\boldsymbol{w}}_n - \boldsymbol{w}^\star\right)\right\|^2 &= \left\|\boldsymbol{X}_1\Big(\frac{1}{n}\sum_{n'=1}^{n} \boldsymbol{w}_{Tn'} - \boldsymbol{w}^\star\Big)\right\|^2 = \frac{1}{n^2}\left\|\boldsymbol{X}_1 \sum_{n'=1}^{n} \left(\boldsymbol{P}_T \cdots \boldsymbol{P}_1\right)^{n'} \left(\cancel{\boldsymbol{w}_0} - \boldsymbol{w}^\star\right)\right\|^2 \\
&\le \frac{1}{n^2}\left\|\sum_{n'=1}^{n} \left(\boldsymbol{I} - \boldsymbol{P}_1\right)\left(\boldsymbol{P}_T \cdots \boldsymbol{P}_1\right)^{n'} \boldsymbol{w}^\star\right\|^2 \\
[\text{Lemma 16}] &\le \frac{1}{n}\sum_{n'=1}^{n}\left\|\left(\boldsymbol{I} - \boldsymbol{P}_1\right)\left(\boldsymbol{P}_T \cdots \boldsymbol{P}_1\right)^{n'} \boldsymbol{w}^\star\right\|^2 \\
&= \frac{1}{n}\sum_{n'=1}^{n}\left(\left\|\left(\boldsymbol{P}_T \cdots \boldsymbol{P}_1\right)^{n'} \boldsymbol{w}^\star\right\|^2 - \left\|\boldsymbol{P}_1\left(\boldsymbol{P}_T \cdots \boldsymbol{P}_1\right)^{n'} \boldsymbol{w}^\star\right\|^2\right) \\
&\le \frac{1}{n}\sum_{n'=1}^{n}\left(\left\|\left(\boldsymbol{P}_T \cdots \boldsymbol{P}_1\right)^{n'} \boldsymbol{w}^\star\right\|^2 - \left\|\left(\boldsymbol{P}_T \cdots \boldsymbol{P}_1\right)^{n'+1} \boldsymbol{w}^\star\right\|^2\right) \\
&= \frac{1}{n}\Big(\underbrace{\left\|\left(\boldsymbol{P}_T \cdots \boldsymbol{P}_1\right)\boldsymbol{w}^\star\right\|^2}_{\le 1} - \underbrace{\left\|\left(\boldsymbol{P}_T \cdots \boldsymbol{P}_1\right)^{n+1}\boldsymbol{w}^\star\right\|^2}_{\ge 0}\Big) \le \frac{1}{n}\,.
\end{aligned}$$

Similarly, for a general task $m \in [T]$, we have

$$\begin{aligned}
\left\|\boldsymbol{X}_m\left(\overline{\boldsymbol{w}}_n - \boldsymbol{w}^\star\right)\right\|^2 &\le \frac{1}{n}\sum_{n'=1}^{n}\left\|\left(\boldsymbol{I} - \boldsymbol{P}_m\right)\left(\boldsymbol{P}_T \cdots \boldsymbol{P}_1\right)^{n'} \boldsymbol{w}^\star\right\|^2 \\
&= \frac{1}{n}\sum_{n'=1}^{n}\left(\left\|\left(\boldsymbol{P}_T \cdots \boldsymbol{P}_1\right)^{n'} \boldsymbol{w}^\star\right\|^2 - \left\|\boldsymbol{P}_m\left(\boldsymbol{P}_T \cdots \boldsymbol{P}_1\right)^{n'} \boldsymbol{w}^\star\right\|^2\right) \\
[\textbf{(Lemma 22a)}] &\le \frac{m}{n}\sum_{n'=1}^{n}\left(\left\|\left(\boldsymbol{P}_T \cdots \boldsymbol{P}_1\right)^{n'} \boldsymbol{w}^\star\right\|^2 - \left\|\left(\boldsymbol{P}_T \cdots \boldsymbol{P}_1\right)^{n'+1} \boldsymbol{w}^\star\right\|^2\right) \\
&= \frac{m}{n}\left(\left\|\left(\boldsymbol{P}_T \cdots \boldsymbol{P}_1\right)\boldsymbol{w}^\star\right\|^2 - \left\|\left(\boldsymbol{P}_T \cdots \boldsymbol{P}_1\right)^{n+1}\boldsymbol{w}^\star\right\|^2\right) \le \frac{m}{n}\,.
\end{aligned}$$

Overall, we get the following bound on the forgetting in the cyclic setting:

$$F_{\tau,S}(\overline{\boldsymbol{w}}_n) = \frac{1}{T}\sum_{m=1}^{T}\left\|\boldsymbol{X}_m\left(\overline{\boldsymbol{w}}_n - \boldsymbol{w}^\star\right)\right\|^2 \le \frac{1}{T}\sum_{m=1}^{T-1}\frac{m}{n} = \frac{T-1}{2n} \le \frac{T^2}{2k}\,.$$

### E.2. Proof sketch for the random setting

For this case, we will demonstrate how one can easily bound a similar expected forgetting to the one in Eq. (18), by $\frac{1}{k}$ instead of $\frac{d-r_{\text{avg}}}{k}$.

Plugging in the average iterate we get,

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{P}_1,\ldots,\boldsymbol{P}_k,\boldsymbol{P}}\left\|(\boldsymbol{I}-\boldsymbol{P})\left(\overline{\boldsymbol{w}}_k-\boldsymbol{w}^\star\right)\right\|_2^2 \\
&= \mathbb{E}_{\boldsymbol{P}_1,\ldots,\boldsymbol{P}_k,\boldsymbol{P}}\left\|(\boldsymbol{I}-\boldsymbol{P})\left(\boldsymbol{w}^\star-\frac{1}{k}\sum_{t=1}^k\boldsymbol{w}_t\right)\right\|_2^2 \\
&= \frac{1}{k^2}\mathbb{E}_{\boldsymbol{P}_1,\ldots,\boldsymbol{P}_k,\boldsymbol{P}}\left\|(\boldsymbol{I}-\boldsymbol{P})\sum_{t=1}^k\left(\boldsymbol{w}_t-\boldsymbol{w}^\star\right)\right\|_2^2 \\
[\text{Lemma 16}] &\leq \frac{1}{k}\sum_{t=1}^k\mathbb{E}_{\boldsymbol{P}_1,\ldots,\boldsymbol{P}_k,\boldsymbol{P}}\left\|(\boldsymbol{I}-\boldsymbol{P})\left(\boldsymbol{w}_t-\boldsymbol{w}^\star\right)\right\|_2^2 \\
[\text{Lemma 7}] &= \frac{1}{k}\sum_{t=1}^k\mathbb{E}_{\boldsymbol{P}_1,\ldots,\boldsymbol{P}_k,\boldsymbol{P}}\left\|(\boldsymbol{I}-\boldsymbol{P})\,\boldsymbol{P}_t\cdots\boldsymbol{P}_1\boldsymbol{w}^\star\right\|_2^2 \\
[\text{Prop. 2}] &= \frac{1}{k}\sum_{t=1}^k\mathbb{E}_{\boldsymbol{P}_1,\ldots,\boldsymbol{P}_k,\boldsymbol{P}}\left[\left\|\boldsymbol{P}_t\cdots\boldsymbol{P}_1\boldsymbol{w}^\star\right\|_2^2-\left\|\boldsymbol{P}\boldsymbol{P}_t\cdots\boldsymbol{P}_1\boldsymbol{w}^\star\right\|_2^2\right] \\
[\text{i.i.d.}] &= \frac{1}{k}\sum_{t=1}^k\mathbb{E}_{\boldsymbol{P}_1,\ldots,\boldsymbol{P}_k}\left[\left\|\boldsymbol{P}_t\cdots\boldsymbol{P}_1\boldsymbol{w}^\star\right\|_2^2-\left\|\boldsymbol{P}_{t+1}\cdots\boldsymbol{P}_1\boldsymbol{w}^\star\right\|_2^2\right] \\
[\text{telescoping}] &= \frac{1}{k}\mathbb{E}_{\boldsymbol{P}_1,\ldots,\boldsymbol{P}_k}\left[\left\|\boldsymbol{P}_1\boldsymbol{w}^\star\right\|_2^2-\left\|\boldsymbol{P}_{k+1}\cdots\boldsymbol{P}_1\boldsymbol{w}^\star\right\|_2^2\right] \leq \frac{1}{k}\,.
\end{aligned}
$$

## Appendix F. Additional related works

### F.1. Continual learning

**Practical methods for continual learning.** Like we explained in Section 7, algorithmic approaches for preventing catastrophic forgetting roughly partition into three categories:

1. *Memory-based replay approaches* actively repeat examples from previous tasks to avoid forgetting. Some store examples from observed tasks (*e.g.,* Robins [59]), while others replay *generated* synthetic data from previous tasks to the main model (*e.g.,* Shin et al. [66]).

2. *Regularization approaches* limit the plasticity of the learned model in order to enhance its stability. One can perform regularization in the *parameter space*, to protect parameters that are important to previous tasks (*e.g.,* Kirkpatrick et al. [31], Nguyen et al. [49], Zeno et al. [79]), or in the *function space*, to protect the *outputs* of previous tasks (*e.g.,* Farajtabar et al. [16], Li and Hoiem [35], Lopez-Paz and Ranzato [36]). Few recent works (Benzing [7], Lubana et al. [37]) focus on obtaining a better understanding of popular quadratic regularization techniques like EWC [31], MAS [2], and SI [78].

3. *Parameter isolation approaches* allocate different subsets of parameters to different tasks. This can be done by expanding the model upon seeing a new task (*e.g.,* Yoon et al. [76]), or by compressing the existing architecture after each task, to free parameters for future tasks (*e.g.,* Mallya and Lazebnik [38], Schwarz et al. [63]).

**Understanding customary training techniques** Some recent papers also test the effect of common deep learning training techniques that change the optimization dynamics and/or have an implicit regularizing implications. For instance, Mirzadeh et al. [41] study how different training regimes affect the loss landscape geometry, thus influencing the overall degree of forgetting in continual learning settings. Goodfellow et al. [21] were the first to suggest that training with the *dropout* technique can be beneficial to remedy catastrophic forgetting in continual learning settings. Delange et al. [10] also show that dropout is fruitful in many continual learning methods (some of which are mentioned above). They point out that it mainly improves the initial performance on learned tasks (by mitigating overfitting), but leads to an increased amount of forgetting when learning later tasks.

Optimization hyperparameters are known to change the geometry of the loss landscape and affect generalization (see for instance Jastrzebski et al. [27]). Mirzadeh et al. [41] point out that most empirical papers in this field use small batch sizes and SGD with learning rate decay. They also discuss how a large learning rate increases the plasticity of deep models, thus having an ill-effect on forgetting.

Finally, Mirzadeh et al. [42] pointed out that architectural decisions also have implications on catastrophic forgetting, and specifically observe that wide neural networks tend to forget less catastrophically.

### F.2. Wider scope

Finally, we note that similar questions to those we ask here and that are asked generally in the continual learning paradigm, are often asked on other fields, such as multi-task learning, meta learning, transfer learning, curriculum learning [73], and online learning. Despite the similarities, there are differences though. For example, in multi-task learning, the data from all tasks are simultaneously

available, in transfer learning the goal to adapt models for a new target tasks and preserving performance on old source task is not a priority, and in online learning, typically, training data for previously seen tasks are assumed to be available in sequentially adapting to data from new tasks. However, a more detailed review of these related areas is beyond the scope of this work.