From Continual Learning to SGD and Back: Better Rates for Continual Linear Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

We theoretically study the common continual learning setup where an overparameterized model is sequentially fitted to a set of jointly realizable tasks. We analyze the forgetting—loss on previously seen tasks—after k iterations. For continual linear models, we prove that fitting a task is equivalent to a *single* stochastic gradient descent (SGD) step on a modified objective. We develop novel last-iterate SGD upper bounds in the realizable least squares setup, which we then leverage to derive new results for continual learning. Focusing on random orderings over T tasks, we establish *universal* forgetting rates, whereas existing rates depend on the problem dimensionality or complexity. Specifically, in continual regression with replacement, we improve the best existing rate from $\mathcal{O}((d-\bar{r})/k)$ to $\mathcal{O}(\min(1/\sqrt[4]{k}, \sqrt{d-\bar{r}}/k, \sqrt{T\bar{r}}/k))$, where d is the dimensionality and $\bar{r}$ the average task rank. Furthermore, we establish the first rate for random task orderings *without* replacement. The obtained rate of $\mathcal{O}(\min(1/\sqrt[4]{T},\ (d-r)/T))$ proves for the first time that randomization alone—with no task repetition—can prevent catastrophic forgetting in sufficiently long task sequences. Finally, we prove a matching $\mathcal{O}(1/\sqrt[4]{k})$ forgetting rate for continual linear *classification* on separable data. Our universal rates apply for broader projection methods, such as block Kaczmarz and POCS, illuminating their loss convergence under i.i.d. and one-pass orderings.

1 Introduction

In continual learning (CL), tasks are presented sequentially, one at a time. The goal is for the learner to adapt to the current task—*e.g.,* by fine-tuning using gradient-based algorithms—while retaining knowledge from previous tasks. A central challenge in this setting is termed *catastrophic forgetting*, where expertise from earlier tasks is lost when adapting to newer ones. Forgetting is influenced by factors such as task similarity and overparameterization [20], and is also related to trade-offs like the plasticity-stability dilemma [46]. CL is becoming increasingly important with the rise of foundation models, where retraining is prohibitively expensive and data from prior tasks is often unavailable, *e.g.,* due to privacy or data retention constraints.

Previous work has shown, both analytically [9, 15, 16, 31, 34] and empirically [26, 39], that forgetting diminishes over time when task ordering is cyclic or random. Different orderings can be explored from multiple perspectives: as a strategy to mitigate forgetting (*e.g.,* by actively ordering an agent's learning environments); as a naturally occurring phenomenon, such as periodic trends in e-commerce; or as a means to model and analyze popular CL benchmarks, such as randomly split datasets.

Our work focuses on a widely studied analytical setting—realizable continual linear regression,[1] where T tasks are learned sequentially over k iterations in a uniform *random ordering*. Evron et al. [15] established that the worst-case expected forgetting lies between $\Omega\left(1/k\right)$ and $\mathcal{O}\left((d-\bar{r})/k\right)$, where d is the problem dimensionality, and $\bar{r}$ the average rank of individual data matrices. This raises a fundamental question, critical in highly overparameterized regimes: *Does worst-case forgetting necessarily scale with dimensionality, and if so, is the dependence indeed linear?*

To this end, we bridge continual learning and the literature on last-iterate stochastic gradient descent (SGD) analysis. We revisit an established connection between continual linear regression and the Kaczmarz method for solving systems of linear equations [15, 32]. Given rank-1 tasks, this method is known to perform a *normalized* stochastic gradient step on the least squares objective, fully minimizing the current task's loss and implying a "stepwise-optimal" step size. Deepening this connection, we prove that even for *general* data ranks, learning a task in continual linear regression and performing an update in the Kaczmarz method, are both equivalent to a *single* SGD step on a modified objective with a constant, stepwise-optimal step size.

Motivated by this, we prove convergence rates for the last iterate of fixed-step size SGD that, crucially, hold for a broad range of step sizes not covered by prior work [e.g., 4, 18, 62, 67, 69, 73]. Curiously, prior results either hold only for the average iterate [e.g., 1], or for small step sizes, bounded away from the stepwise-optimal step size crucial for the continual learning setup [e.g., 67]. We overcome this challenge through a careful combination of analysis techniques for SGD [62, 65], further tightening the analysis to accommodate a wider range of step sizes, including the stepwise-optimal one.

Applying our last-iterate analysis to continual regression, we tighten the existing forgetting rate and establish the first dimension-*independent* rate (see Table 1). Furthermore, we provide the first rate for random task orderings *without* replacement, proving that task repetition is not obligatory to guarantee convergence when $k = T \to \infty$, thus highlighting the effect of randomization as compared to repetition. Our results also yield novel rates for the closely related Kaczmarz and NLMS methods.

Finally, we prove a matching rate for the squared loss of the broader Projection Onto Convex Sets framework [22]. This extends our results to continual linear *classification* on separable data—previously linked to projection algorithms [16]—and provides this setting's first universal rate, independent of the problem's "complexity".

Summary of Contributions. To summarize, our main contributions in this paper are:

- We establish a new reduction from continual linear regression to SGD with a particular choice of a "stepwise-optimal" step size, generalizing ideas from prior work that only applied to rank-1 tasks, to tasks of arbitrary rank. This facilitates last-iterate analysis for studying forgetting.
- We provide novel last-iterate convergence analysis for SGD in a realizable least squares setup. To the best of our knowledge, this is the first analysis that provides nontrivial rates for large step sizes, which are crucial to the reduction to continual learning scenarios.
- Our main contribution, building on these techniques, is a set of improved rates of forgetting in continual linear regression, including the first *universal rates*, independent of the problem dimensionality or complexity, as well as the first rates for *without-replacement* orderings, indicating task repetition is not mandatory to diminish forgetting. See Table 1 for a summary.
- We further relate and extend our results to other settings, including continual linear classification, the block Kaczmarz method, and the Projection Onto Convex Sets framework (POCS).

2 Main Setting: Continual Linear Regression

We mainly investigate the fundamental continual linear regression setting, as studied in many theoretical papers. This setting is easy-to-analyze, yet often sheds light on important CL phenomena.[1]

Notation. Vectors and matrices are in boldface. $\|\cdot\|$ denotes Euclidean, spectral, or operator norms of vectors, matrices, or linear operators. $\mathbf{X}^+$ denotes the Moore-Penrose inverse. Finally, $[n] \triangleq 1, ..., n$.

[1] Despite its simplicity, the continual linear regression setting is insightful enough to investigate different factors and aspects in continual settings, *e.g.,* task similarity [15, 27, 42], task recurrence [15, 34], overparameterization [20], and algorithms [13, 55]. We follow prior theoretical work assuming tasks are jointly realizable [e.g., 15, 20, 34]. Other notable work alternatively allow label noise, at the cost of assuming either i.i.d. features [2, 19, 42] or commutative data matrices [40, 41, 72]—while our analysis facilitates *any* data matrices.

Formally, the learner is given a collection of T linear regression tasks, $(\mathbf{X}_1, \mathbf{y}_1), \ldots, (\mathbf{X}_T, \mathbf{y}_T)$, where $\mathbf{X}_m \in \mathbb{R}^{n_m \times d}, \mathbf{y}_m \in \mathbb{R}^{n_m}$. For k iterations, tasks are learned according to a *task ordering* $\tau : [k] \to [T]$. We analyze random orderings, previously studied in continual linear models [e.g., 15, 16, 31].

Definition 2.1 (Random Task Ordering)**.** A random ordering selects tasks uniformly at random from the task collection $[T]$, *i.e.,* $\tau(1), \ldots, \tau(k) \sim \text{Unif}([T])$, with or without replacement.

We are now ready to define the learning scheme we study, which, at each iteration, naively minimizes the sum of squared errors for the current regression task.[2]

Scheme 1 Continual Linear Regression (to Convergence)

Initialize $\mathbf{w}_0 = \mathbf{0}_d$
For each iteration $t = 1, \ldots, k$:
  $\mathbf{w}_t \leftarrow$ Start from $\mathbf{w}_{t-1}$ and minimize the current task's loss $\mathcal{L}_{\tau(t)}(\mathbf{w}) \triangleq \frac{1}{2} \left\| \mathbf{X}_{\tau(t)} \mathbf{w} - \mathbf{y}_{\tau(t)} \right\|^2$
    with (S)GD to convergence
Output $\mathbf{w}_k$

This scheme was previously linked to the Kaczmarz method and, in a special case, to normalized SGD [15]. In Section 3, we explain and develop these connections to enable novel analysis.

Our main assumption is the existence of *offline solutions* that perfectly solve all T tasks jointly, as assumed in much of the theoretical CL literature [e.g., 15, 16, 20, 31]. This assumption simplifies the analysis[1] and rules out cases where forgetting previous tasks is *beneficial*, as new tasks may directly contradict them. Finally, this assumption is reasonable in highly overparameterized models and is thus linked to the linear dynamics of deep networks in the neural tangent kernel (NTK) regime [29].

Assumption 2.2 (Joint Linear Realizability)**.** We assume the set of offline solutions that solve *all* tasks is nonempty. That is, $\mathcal{W}_\star \triangleq \left\{ \mathbf{w} \in \mathbb{R}^d \;\middle|\; \mathbf{X}_m \mathbf{w} = \mathbf{y}_m, \ \forall m \in [T] \right\} \neq \varnothing$.

To facilitate the results and discussions in our paper, we focus on the offline solution with minimal norm, often associated with good generalization capabilities.

Definition 2.3 (Minimum-Norm Offline Solution)**.** We denote, $\mathbf{w}_\star \triangleq \operatorname{argmin}_{\mathbf{w} \in \mathcal{W}_\star} \|\mathbf{w}\|$.

Commonly in continual learning setups, the model performance on past tasks degrades, sometimes significantly, even in linear models [15]. Our goal is to bound this degradation, *i.e.,* "forgetting". Following common definitions [e.g., 13, 16], we define forgetting as the average increase in the loss of the *last* iterate on previous tasks.

Definition 2.4 (Forgetting)**.** Let $\mathbf{w}_1, \ldots, \mathbf{w}_k$ be the iterates of Scheme 1 under a task ordering τ. The forgetting at iteration k is the average increase in the loss of previously seen tasks. In our realizable setting, the forgetting becomes an *in-sample* loss. Formally,

$$F_\tau(k) = \frac{1}{k} \sum_{t=1}^{k} \left(\mathcal{L}_{\tau(t)}(\mathbf{w}_k) - \underbrace{\mathcal{L}_{\tau(t)}(\mathbf{w}_t)}_{=0} \right) = \frac{1}{2k} \sum_{t=1}^{k} \left\| \mathbf{X}_{\tau(t)} \mathbf{w}_k - \mathbf{y}_{\tau(t)} \right\|^2 .$$

Under arbitrary orderings, Evron et al. [15] showed forgetting can be "catastrophic" in the sense that $\lim_{k \to \infty} \mathbb{E}[F_\tau(k)] > 0$. However, as we show, this *cannot* happen under the random ordering.

Remark 2.5 (Forgetting vs. Regret). While regret and forgetting are related, they can differ significantly [15]. Regret is a key quantity in online learning, defined as $\frac{1}{2k} \sum_{t=1}^{k} \|\mathbf{X}_{\tau(t)} \mathbf{w}_{t-1} - \mathbf{y}_{\tau(t)}\|^2$ in our setting. That is, it measures the suboptimality of each iterate on the *consecutive* task. In contrast, forgetting evaluates an iterate's performance across *earlier* tasks.

We further define the training loss to easily discuss links to *other* fields, such as Kaczmarz.

Definition 2.6 (Training Loss)**.** The training loss of any vector $\mathbf{w} \in \mathbb{R}^d$ is given by,

$$\mathcal{L}(\mathbf{w}) = \frac{1}{T} \sum_{m=1}^{T} \mathcal{L}_m(\mathbf{w}) = \frac{1}{2T} \sum_{m=1}^{T} \|\mathbf{X}_m \mathbf{w} - \mathbf{y}_m\|^2 .$$

[2]This objective is natural for regression; our analysis also extends to the *mean* squared error (refining our R).

We bound both the forgetting and the loss, leveraging a key property—expected (in-sample) forgetting can be upper bounded using expected training loss across all tasks. Specifically, Lemma B.1 (in App. B) states that $\mathbb{E}_\tau[F_\tau(k)] \le 2\mathbb{E}_\tau\left[\mathcal{L}\left(\mathbf{w}_{k-1}\right)\right] + \frac{\|\mathbf{w}_\star\|^2 R^2}{k}$ in orderings with replacement, where $R \triangleq \max_{m\in[T]} \|\mathbf{X}_m\|$ is the data "radius" and the dependence of $\mathbf{w}_{k-1}$ on $\tau_1, \dots, \tau_{k-1}$ is implicit. Without-replacement orderings yield a related but more refined bound. The additive $\frac{\|\mathbf{w}_\star\|^2 R^2}{k}$ term is negligible compared to other terms in our bounds.

3 Reductions: From Continual Linear Regression to Kaczmarz to SGD

Previous work established connections between continual linear regression and the Kaczmarz method [15]. We revisit these connections pedagogically to ensure our paper is self-contained. Importantly, this leads to novel links between CL, the Kaczmarz method, and SGD on special functions (Schemes 1,2,3), allowing us to improve the rates for continual and Kaczmarz methods by analyzing the last iterate of SGD instead.

Scheme 2 The Block Kaczmarz Method	**Scheme 3** SGD with $\eta = 1$ on special $\{f_m\}_m$
Input: Jointly realizable $(\mathbf{X}_m, \mathbf{y}_m), \forall m \in [T]$	**Input:** $f_m(\mathbf{w}) = \frac{1}{2}\left\|\mathbf{X}_m^+\left(\mathbf{X}_m\mathbf{w} - \mathbf{y}_m\right)\right\|^2, \forall m \in [T]$
Initialize $\mathbf{w}_0 = \mathbf{0}_d$	Initialize $\mathbf{w}_0 = \mathbf{0}_d$
For each iteration $t = 1, \dots, k$:	For each iteration $t = 1, \dots, k$:
$\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} - \mathbf{X}_{\tau(t)}^+\big(\mathbf{X}_{\tau(t)}\mathbf{w}_{t-1} - \mathbf{y}_{\tau(t)}\big)$	$\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} - \nabla_{\mathbf{w}} f_{\tau(t)}\big(\mathbf{w}_{t-1}\big)$

3.1 Revisit: Continual Linear Regression and the Kaczmarz Method

The (block) Kaczmarz method in Scheme 2 [14, 32] is a classical iterative method for solving a linear system $\mathbf{X}\mathbf{w} = \mathbf{y}$, easily mapped to our learning problem by stacking tasks in blocks, *i.e.,*

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_T \end{pmatrix} \in \mathbb{R}^{N\times d}, \quad \mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_T \end{pmatrix} \in \mathbb{R}^{N}, \quad \text{where } N = \sum_{m=1}^{T} n_m.$$

In each iteration, the Kaczmarz method (Scheme 2) perfectly solves the current block, *i.e.,* $\mathbf{X}_{\tau(t)}\mathbf{w}_t = \mathbf{y}_{\tau(t)}$ (to see that, recall that $\mathbf{X}_{\tau(t)}^+$ denotes the Moore-Penrose pseudo-inverse of $\mathbf{X}_{\tau(t)}^+$).

The continual Scheme 1 also minimizes the current loss *to convergence*, *i.e.,* until it is perfectly solved (in the realizable case). In fact, Evron et al. [15] identified the following reduction.

Reduction 1 (Continual Regression $\Rightarrow$ Block Kaczmarz)**.** *In the realizable case (Assumption 2.2) under any ordering τ, continual linear regression learned to convergence*[3] *is equivalent to the block Kaczmarz method. That is, the iterates $\mathbf{w}_0, \dots, \mathbf{w}_k$ of Schemes 1 and 2 coincide.*

3.2 New Reduction: Kaczmarz Method and Stepwise-Optimal Stochastic Gradient Descent

Rank-1 data. It is known that when each task contains *just one* row, each update in the Kaczmarz method corresponds to a gradient step on with a specific "normalizing" step size [51]. That is, since in rank-1 we have $\mathcal{L}_{\tau(t)}(\mathbf{w}) = \frac{1}{2}\left\|\mathbf{x}_{\tau(t)}^\top\mathbf{w} - y_{\tau(t)}\right\|^2$, the Kaczmarz updates hold

$$\mathbf{w}_t = \mathbf{w}_{t-1} - \frac{1}{\|\mathbf{x}_{\tau(t)}\|^2}\big(\mathbf{x}_{\tau(t)}^\top\mathbf{w}_{t-1} - y_{\tau(t)}\big)\mathbf{x}_{\tau(t)} = \mathbf{w}_{t-1} - \frac{1}{\|\mathbf{x}_{\tau(t)}\|^2}\nabla_{\mathbf{w}}\mathcal{L}_{\tau(t)}(\mathbf{w}_{t-1})\,. \tag{1}$$

What about *higher* data ranks? We now establish a more general reduction from the *block* Kaczmarz method—at *any* rank—to SGD (in Section 6, we similarly connect SGD and the broader Projection Onto Convex Sets framework, extending our results to continual linear *classification*).

Reduction 2 (Block Kaczmarz $\Rightarrow$ SGD)**.** *In the realizable case (Assumption 2.2) under any ordering τ, the block Kaczmarz method is equivalent to SGD with a step size of $\eta = 1$, applied w.r.t. a convex, 1-smooth least squares objective: $\left\{f_m(\mathbf{w}) \triangleq \frac{1}{2}\left\|\mathbf{X}_m^+\left(\mathbf{X}_m\mathbf{w} - \mathbf{y}_m\right)\right\|^2\right\}_{m=1}^{T}$. That is, the iterates $\mathbf{w}_0, \dots, \mathbf{w}_k$ of Schemes 2 and 3 coincide.*

[3] The learner minimizes $\mathcal{L}_{\tau(t)}$ with (S)GD to convergence; the pseudo-inverse is *not* computed explicitly.

The reduction follows from the next lemma, which states the convexity and smoothness of f_m and expresses the gradient $\nabla_{\mathbf{w}} f_{\tau(t)}$, subsequently substituted into $\left(\mathbf{w}_{t-1} - \nabla_{\mathbf{w}} f_{\tau(t)}(\mathbf{w}_{t-1})\right)$ to complete the proof. The proof for the lemma is given in App. B.

Lemma 3.1 (Properties of the Modified Objective)**.** *Consider any realizable task collection such that* $\mathbf{X}_m \mathbf{w}_\star = \mathbf{y}_m, \forall m \in [T]$. *Define* $f_m(\mathbf{w}) = \frac{1}{2}\left\|\mathbf{X}_m^+ \left(\mathbf{X}_m \mathbf{w} - \mathbf{y}_m\right)\right\|^2$. *Then,* $\forall m \in [T], \mathbf{w} \in \mathbb{R}^d$

(i) ***Upper bound:*** $\mathcal{L}_m(\mathbf{w}) \le R^2 f_m(\mathbf{w}) \triangleq \max_{m' \in [T]} \|\mathbf{X}_{m'}\|^2 f_m(\mathbf{w})$.

(ii) ***Gradient:*** $\nabla_{\mathbf{w}} f_m(\mathbf{w}) = \mathbf{X}_m^+ \mathbf{X}_m \mathbf{w} - \mathbf{X}_m^+ \mathbf{y}_m$.

(iii) ***Convexity and Smoothness:*** f_m *is convex and* 1*-smooth.*

4 Rates for Random-Order Continual Linear Regression and Kaczmarz

This section focuses on improving the best upper bound known in prior continual learning literature on random orderings, summarized in Table 1. Specifically, for with-replacement random orderings, Evron et al. [15] proved a forgetting rate of $\mathbb{E}_\tau\left[F_\tau(k)\right] = \mathcal{O}\left(\frac{d-\bar{r}}{k}\right)$ where $\bar{r} \triangleq \frac{1}{T}\sum_m \operatorname{rank}(\mathbf{X}_m)$. Notably, this rate depends on the problem dimensionality d, raising concerns when generalizing insights from linear models to deep networks, which are often heavily overparameterized (*e.g.,* in the NTK regime). Encouragingly, that paper only provided (implicitly) a $1/k$ lower bound for the worst-case forgetting, calling for further research into narrowing this gap.

We tighten the existing rate's problem-dependent term from $\left(d - \bar{r}\right)$ to $\min\left(\sqrt{d-\bar{r}}, \sqrt{T\bar{r}}\right)$ and also prove a problem-*independent* rate of $1/\sqrt[4]{k}$. Finally, we provide the first rates for *without*-replacement orderings, isolating the effect of randomness versus repetition.

Table 1: **Forgetting and Loss Rates in Continual Linear Regression (and Block Kaczmarz).** Upper bounds apply to any T realizable tasks (or blocks). Lower bounds indicate *worst cases*, *i.e.,* specific constructions. Random ordering bounds apply to the *expected* forgetting (or loss).
We omit mild constant multiplicative factors and an unavoidable $\|\mathbf{w}_\star\|^2 R^2$ term. Finally, $a \wedge b \triangleq \min(a, b)$.
Recall: $k =$ iterations; $d =$ dimensionality; $\bar{r}, r_{\max} =$ average and maximum data matrix ranks.

Paper / Ordering	**Bound**	**Random with Replacement**	**Random w/o Replacement**	**Cyclic**
Evron et al. [15]	Upper	$\frac{d-\bar{r}}{k}$	—	$\frac{T^2}{\sqrt{k}} \wedge \frac{T^2(d-r_{\max})}{k}$
Kong et al. [34]	Upper	—	—	$\frac{T^3}{k}$
Ours (2025)	Upper	$\frac{1}{\sqrt[4]{k}} \wedge \frac{\sqrt{d-\bar{r}}}{k} \wedge \frac{\sqrt{T\bar{r}}}{k}$	$\frac{1}{\sqrt[4]{T}} \wedge \frac{d-\bar{r}}{T}$	—
Evron et al. [15]	Lower	$\frac{1}{k}$ (*)	$\frac{1}{T}$ (*)	$\frac{T^2}{k}$

(*) They did not explicitly provide such lower bounds, but the $T=2$ tasks construction from their proof of Theorem 10, can yield a $\Theta(1/k)$ random behavior by cloning those 2 tasks $\lfloor T/2 \rfloor$ times for any general T.

4.1 A Parameter-Dependent $\mathcal{O}(1/k)$ Rate

Here, we present a tighter $\sqrt{d-\bar{r}}$ term and a term depending only on the rank and number of tasks.

Theorem 4.1 (Parameter-Dependent Forgetting Rate for Random With Replacement)**.** *Under a random ordering with replacement over T jointly realizable tasks, the expected loss and forgetting of Schemes 1, 2 after $k \ge 3$ iterations are bounded as,*

$$\mathbb{E}_\tau[\mathcal{L}(\mathbf{w}_k)] \le \frac{\min\left(\sqrt{d-\bar{r}}, \sqrt{T\bar{r}}\right)\|\mathbf{w}_\star\|^2 R^2}{2e(k-1)}, \quad \mathbb{E}_\tau\left[F_\tau(k)\right] \le \frac{3\min\left(\sqrt{d-\bar{r}}, \sqrt{T\bar{r}}\right)\|\mathbf{w}_\star\|^2 R^2}{2\left(k-2\right)},$$

where $\bar{r} \triangleq \frac{1}{T}\sum_{m\in[T]} \operatorname{rank}(\mathbf{X}_m)$. *(Recall that* $R \triangleq \max_{m\in[T]} \|\mathbf{X}_m\|$.*)*

The $\|\mathbf{w}_\star\|^2 R^2$ scaling term is generally unavoidable. Our proof, given in App. C, is related to a recent work [23] that characterized the weak error (resembling our loss) by analyzing a linear map. Unlike ours, the polynomial rates they derive involve matrix properties related to the condition number.

4.2 A Universal $\mathcal{O}(1/\sqrt[4]{k})$ Rate

Next, we present a forgetting rate *independent* on the dimensionality, rank, and number of tasks. This is crucial in highly overparameterized regimes which are connected to deep neural networks.

Theorem 4.2 (Universal Forgetting Rate for With-Replacement Random Ordering)**.** *Under a random ordering with replacement over T jointly realizable tasks, the expected loss and forgetting of Schemes 1, 2 after $k \geq 2$ iterations are bounded as,*

$$\mathbb{E}_\tau \left[\mathcal{L}\left(\mathbf{w}_k\right)\right] \leq \frac{2\left\|\mathbf{w}_\star\right\|^2 R^2}{\sqrt[4]{k}}\,, \qquad \mathbb{E}_\tau \left[F_\tau(k)\right] \leq \frac{5\left\|\mathbf{w}_\star\right\|^2 R^2}{\sqrt[4]{k-1}}\,.$$

We prove this result in App. D.1 by leveraging the connections between CL and SGD. Specifically, in Section 3 we showed that continual linear regression is equivalent to SGD with a step size of *exactly* 1 on a related least squares objective that bounds the original continual learning loss. Our result then follows from our novel last-iterate SGD bounds that, crucially, apply even to that specific step size. To ease readability, here we focused on a CL perspective, deferring last-iterate analysis to Section 5.

4.3 Random Task Orderings Without Replacement

Evron et al. [15] suggested defining forgetting as "catastrophic" only when $\lim_{k\to\infty} \mathbb{E}\left[F_\tau(k)\right] > 0$. They presented such an adversarial case with a deterministic task ordering where $k = T \to \infty$, and showed that task recurrence, under cyclic or random orderings, mitigates forgetting. So far, in random orderings, it was hard to isolate the effect of randomness from that of repetitions. It was thus unclear whether catastrophic forgetting can be alleviated by randomly permuting the tasks. Below, we provide the first result demonstrating that no recurrence is needed under random orderings.

Theorem 4.3 (Forgetting Rates for Without-Replacement Random Ordering)**.** *Under a random ordering without replacement over T jointly realizable tasks, the expected loss and forgetting of Schemes 1, 2 after $k \in \{2, \dots, T\}$ iterations are both bounded as,*

$$\mathbb{E}_\tau \left[\mathcal{L}\left(\mathbf{w}_k\right)\right],\, \mathbb{E}_\tau \left[F_\tau(k)\right] \leq \min\left(\frac{7}{\sqrt[4]{k-1}}, \frac{d - \bar{r} + 1}{k-1}\right) \left\|\mathbf{w}_\star\right\|^2 R^2.$$

The proof of the dimensionality-dependent term is similar to the one of the with-replacement case, given in Section D.1.2 of Evron et al. [15], but requires a more careful upper bound on the (in-sample) forgetting. The proof of the dimensionality-independent term again relies on last-iterate analysis, as presented in App. E.2. Both proofs are given in App. D.2.

App. A discusses connections between our result above and related areas like shuffle SGD.

5 Last-Iterate SGD Bounds for Linear Regression

In this self-contained section, we derive last-iterate guarantees for SGD in the realizable stochastic least squares setup. Motivated by the connection with continual regression discussed in Section 3, we focus on regression problems that are β-smooth individually, and obtain upper bounds for the last SGD iterate that apply for a significantly wider range of step sizes compared to prior art [67]. Notably, this is the first time convergence of SGD in this setup is established for a range of step sizes completely independent of the optimization horizon. Table 2 in App. A compares our bounds with related work and classical results in the field.

Recent work has analyzed SGD in the realizable least squares setting, with most assuming a somewhat more general noise model [4, 18, 67, 68, 69, 73]. These studies are primarily motivated by connections to deep networks in the overparameterized regime [45], where models are expressive enough to perfectly fit the training data. With the exception of Varre et al. [67], most of these works focus on non-fixed step sizes and/or provide guarantees for the average iterate. See App. A for further discussion. Here, we study a stochastic, jointly realizable least squares problem, as defined next.

Setup 1. Let $\mathcal{I}$ be an index set, and $\mathcal{D}$ a distribution over $\mathcal{I}$. We consider the optimization objective:

$$\min_{\mathbf{w}\in\mathbb{R}^d}\left\{\bar{f}(\mathbf{w}) \triangleq \mathbb{E}_{i\sim\mathcal{D}} f(\mathbf{w};i) \triangleq \mathbb{E}_{i\sim\mathcal{D}}\left[\tfrac{1}{2}\left\|\mathbf{A}_i\mathbf{w}-\mathbf{b}_i\right\|^2\right]\right\},$$

where $\mathbf{A}_i \in \mathbb{R}^{n_i\times d}, \mathbf{b}_i \in \mathbb{R}^{n_i}$, $\forall i \in \mathcal{I}$. We specifically focus on β-*smooth* functions, *i.e.,* $\left\|\mathbf{A}_i^\top\mathbf{A}_i\right\| \le \beta, \forall i \in \mathcal{I}$, under a *realizable* assumption, *i.e.,* $\exists \mathbf{w}_\star \in \mathbb{R}^d : \bar{f}(\mathbf{w}_\star) = 0$.

Our main result establishes last-iterate guarantees for with-replacement SGD, defined next. Given an initialization $\mathbf{w}_0 \in \mathbb{R}^d$ and step-size $\eta > 0$:

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta\nabla f(\mathbf{w}_t; i_t), \quad i_t \sim \mathcal{D}. \tag{2}$$

Below, we state our theorem and then provide an overview of the analysis.

Theorem 5.1 (Last-Iterate Bound for Realizable Regression With Replacement)**.** *Consider the β-smooth, realizable Setup 1. Then, for any initialization $\mathbf{w}_0 \in \mathbb{R}^d$, with-replacement SGD (Eq.* (2)*) with step size $\eta < 2/\beta$, holds:*

$$\mathbb{E}\bar{f}(\mathbf{w}_T) \le \frac{eD^2}{2\eta(2-\eta\beta)T^{1-\eta\beta(1-\eta\beta/4)}}, \quad \forall T \ge 1,$$

where $D \triangleq \|\mathbf{w}_0 - \mathbf{w}_\star\|$. In particular, for $\eta = \frac{1}{\beta}$, $\mathbb{E}\bar{f}(\mathbf{w}_T) \le \frac{e\beta D^2}{2\sqrt[4]{T}}$.

The important feature of Theorem 5.1 is the $(2-\eta\beta)$ factor in the denominator, replacing the common $(1-\eta\beta)$ of the standard analysis. This difference makes our theorem applicable to the continual regression setting which requires setting $\eta = 1/\beta$ (Reduction 2). In addition, for $\eta = 1/(\beta\log T)$, we recover the near-optimal rate obtained by Varre et al. [67], *i.e.,* $\mathbb{E}\bar{f}(\mathbf{w}_T) = \mathcal{O}\left(\frac{\beta D^2\log T}{T}\right)$.

Extension to Without-Replacement SGD. In App. E.2, we extend Theorem 5.1 to the setting of SGD without replacement. The proof leverages algorithmic stability for SGD [6, 25, 61], focusing on a variant tailored to without-replacement sampling [35, 63]. In particular, we establish a new bound for this variant in the smooth and realizable regime, which has not appeared in prior work.

Analysis Overview. Here, we briefly outline the proof of Theorem 5.1, which follows immediately by combining the two lemmas below (while noting that $\eta < 2/\beta \Rightarrow e^{\eta\beta(1-\eta\beta/4)} \le e$). The first step of the proof is to establish a regret bound for SGD when applied to $f(\mathbf{w}; i_1)\dots f(\mathbf{w}; i_T)$, holding for any step size $\eta < 2/\beta$. This already departs from the standard $\eta < 1/\beta$ mandated by standard analysis. All proofs for this section are given in App. E.1.

Lemma 5.2 (Gradient Descent Regret Bound for Smooth Optimization)**.** *Consider the β-smooth, realizable Setup 1, and let $T \ge 1$, $(i_0,\dots,i_T) \in \mathcal{I}^{T+1}$ be an arbitrary sequence of indices in $\mathcal{I}$, and $\mathbf{w}_0 \in \mathbb{R}^d$ be an arbitrary initialization. Then, the gradient descent iterates given by $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta\nabla f(\mathbf{w}_t; i_t)$ for a step size $\eta < 2/\beta$, hold:*

$$\sum_{t=0}^{T} f(\mathbf{w}_t; i_t) \le \frac{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{2\eta(2-\eta\beta)}.$$

The second and main step of the analysis is to relate the loss of the last SGD iterate to the regret of the algorithm. For this, we carefully adapt an existing approach for last-iterate convergence in the non-smooth case [62]. The result, given below, is slightly more general to accommodate without-replacement sampling, addressed in the next section.

Lemma 5.3. *Consider the β-smooth, realizable Setup 1. Let $T \ge 1$. Assume $\mathcal{P}$ is a distribution over $\mathcal{I}^{T+1}$ such that for every $0 \le t \le \tau_1 \le \tau_2 \le T$, the following holds: For any $i_0,\dots i_{t-1} \in \mathcal{I}^t, i \in \mathcal{I}$, $\Pr(i_{\tau_1} = i | i_0,\dots,i_{t-1}) = \Pr(i_{\tau_2} = i | i_0,\dots,i_{t-1})$. Then, for any initialization $\mathbf{w}_0 \in \mathbb{R}^d$, with-replacement SGD (Eq.* (2)*) with step-size $\eta < 2/\beta$, holds:*

$$\mathbb{E}f(\mathbf{w}_T, i_T) \le (eT)^{\eta\beta(1-\eta\beta/4)}\mathbb{E}\left[\tfrac{1}{T+1}\sum\nolimits_{t=0}^{T} f(\mathbf{w}_t; i_t)\right],$$

where the expectation is taken with respect to $i_0,\dots,i_T$ sampled from $\mathcal{P}$.

6 Extensions

6.1 A Universal $\mathcal{O}(1/\sqrt[4]{k})$ Rate for General Projections Onto Convex Sets

Projections Onto Convex Sets (POCS) is a classical method that iteratively projects onto closed convex sets to find a point in their intersection [7, 22]. Formally,

Scheme 4 Projections onto Convex Sets (POCS)

Input: A set of T closed convex sets $\mathcal{C}_1, \ldots, \mathcal{C}_T$; an initial $\mathbf{w}_0 \in \mathbb{R}^d$; an ordering $\tau : [k] \to [T]$
For each iteration $t = 1, \ldots, k$:
$\mathbf{w}_t \leftarrow \mathbf{\Pi}_{\tau(t)}(\mathbf{w}_{t-1}) \triangleq \operatorname{argmin}_{\mathbf{w}\in\mathcal{C}_{\tau(t)}} \|\mathbf{w} - \mathbf{w}_{t-1}\|$

Generalizing Reduction 2 (Kazcmarz⇒SGD), we note that POCS algorithms also implicitly perform stepwise-optimal SGD w.r.t. a convex, 1-smooth least squares objective. This has been partially observed in the POCS literature [e.g., 49]. All proofs for this section are given in App. F.

Reduction 3 (POCS ⇒ SGD)**.** *Consider T arbitrary (nonempty) closed convex sets $\mathcal{C}_1, \ldots, \mathcal{C}_T$, initial point $\mathbf{w}_0 \in \mathbb{R}^d$, and ordering τ. Define $f_m(\mathbf{w}) = \frac{1}{2}\|\mathbf{w} - \mathbf{\Pi}_m(\mathbf{w})\|^2, \forall m \in [T]$. Then,*

(i) f_m is convex and 1-smooth.

(ii) The POCS update is equivalent to an SGD step: $\mathbf{w}_t = \mathbf{\Pi}_{\tau(t)}(\mathbf{w}_{t-1}) = \mathbf{w}_{t-1} - \nabla_{\mathbf{w}} f_{\tau(t)}(\mathbf{w}_{t-1})$.

We can now employ our analysis from Section 5 to yield a universal rate.

Theorem 6.1 (Universal POCS Rate)**.** *Consider the conditions of Reduction 3 and assume a nonempty intersection $\mathcal{C}_\star = \bigcap_{m=1}^T \mathcal{C}_m \neq \varnothing$. Then, under a random ordering with or without replacement, the expected "residual" of Scheme 4 after $\forall k \geq 1$ iterations ($k \in [T]$ without replacement) is,*

$$\mathbb{E}_\tau\left[\frac{1}{2T}\sum_{m=1}^T \|\mathbf{w}_k - \mathbf{\Pi}_m(\mathbf{w}_k)\|^2\right] = \mathbb{E}_\tau\left[\frac{1}{2T}\sum_{m=1}^T \operatorname{dist}^2(\mathbf{w}_k, \mathcal{C}_m)\right] \leq \frac{7}{\sqrt[4]{k}} \min_{\mathbf{w}\in\mathcal{C}_\star} \|\mathbf{w}_0 - \mathbf{w}\|^2 .$$

To the best of our knowledge, this is the first universal rate in the POCS literature, independent of problem parameters such as its regularity or complexity, as demonstrated next in Section 6.2. Universal rates are only achievable when analyzing individual distances, *i.e.,* $f_m(w) = \operatorname{dist}^2(\mathbf{w}, \mathcal{C}_m) = \|\mathbf{w} - \mathbf{\Pi}_m(\mathbf{w})\|^2$, rather than the distance to the intersection, *i.e.,* $\operatorname{dist}^2(\mathbf{w}, \mathcal{C}_\star)$. In machine learning, the squared distance from individual sets is linked to important losses like MSE in regression or squared hinge loss in classification [15, 16], naturally leading to our next continual model.

6.2 A Universal $\mathcal{O}(1/\sqrt[4]{k})$ Rate for Random Orderings in Continual Linear Classification

Regularization methods are commonly used to prevent forgetting in CL [see 33]. Evron et al. [16] studied a regularized linear model for continual classification. They considered $T \geq 2$ jointly separable, binary classification tasks, defined by datasets $S_1, \ldots, S_T$ consisting of vectors $\mathbf{x} \in \mathbb{R}^d$ and their labels $y \in \{-1, +1\}$. They proved that a weakly-regularized scheme implicitly applies sequential max-margin projections. That is, in the limit as $\lambda \to 0$, the iterates of the two following schemes align in direction, enabling the study of continual classification through projection algorithms.

Scheme 5 Regularized Continual Classification

Initialize $\mathbf{w}_0^{(\lambda)} = \mathbf{0}_d$
For each iteration $t = 1, \ldots, k$:
$\mathbf{w}_t^{(\lambda)} \leftarrow \operatorname{argmin}_{\mathbf{w}\in\mathbb{R}^d} \sum_{(\mathbf{x},y)\in S_t} e^{-y\mathbf{w}^\top\mathbf{x}} + \frac{\lambda}{2}\left\|\mathbf{w} - \mathbf{w}_{t-1}^{(\lambda)}\right\|^2$

Scheme 6 Sequential Max-Margin Projections

Initialize $\mathbf{w}_0 = \mathbf{0}_d$
For each iteration $t = 1, \ldots, k$:
$\mathbf{w}_t \leftarrow \mathbf{\Pi}_{\tau(t)}(\mathbf{w}_{t-1}) \triangleq \operatorname*{argmin}_{\mathbf{w}\in\mathcal{C}_{\tau(t)}} \|\mathbf{w} - \mathbf{w}_{t-1}\|$ where $\mathcal{C}_m \triangleq \left\{\mathbf{w}\in\mathbb{R}^d \mid y\mathbf{w}^\top\mathbf{x} \geq 1,\ \forall(\mathbf{x}, y)\in S_m\right\}$

They studied forgetting under several orderings, using an equivalent of our Definition 2.4:

$$F_\tau(k) = \frac{1}{k}\sum\nolimits_{t=1}^{k} \big(\mathcal{L}_{\tau(t)}(\mathbf{w}_k) - \mathcal{L}_{\tau(t)}(\mathbf{w}_t)\big) \le \frac{R^2}{2k}\sum\nolimits_{t=1}^{k} \left\|\mathbf{w}_{k-1} - \mathbf{\Pi}_{\tau(t)}(\mathbf{w}_{k-1})\right\|^2 .$$

Our POCS rate (Theorem 6.1) combined with SGD stability arguments gives the following.

Theorem 6.2. *Under a random ordering, with or without replacement, over T jointly separable tasks, the expected forgetting of the weakly-regularized Scheme 5 (at $\lambda \to 0$) after $k \ge 1$ iterations is,*

$$\mathbb{E}_\tau\big[F_\tau(k)\big] \le \frac{7\left\|\mathbf{w}_\star\right\|^2 R^2}{\sqrt[4]{k}}, \quad \textit{where } \mathbf{w}_\star \in \operatorname{argmin}_{\mathbf{w}\in\mathcal{C}_1\cap\cdots\cap\mathcal{C}_T} \left\|\mathbf{w}_0 - \mathbf{w}\right\|^2 .$$

As shown in Table 3 of App. F, our rate is universal while the previous one depends on $\left\|\mathbf{w}_\star\right\|^2 R^2$, often seen as the "complexity" of classification problems. For example, after $k = 4T\left\|\mathbf{w}_\star\right\|^2 R^2$ iterations, the existing (normalized) rate is e^{-1}, while ours is potentially much smaller: $\frac{5}{T^{1/4}\sqrt{\|\mathbf{w}_\star\| R}}$.

7 Discussion

Our work established a reduction from continual linear regression to the (block) Kaczmarz method and then to "stepwise-optimal" SGD. This enabled the development of analytic tools for last-iterate SGD schemes, leading to significantly improved and even universal rates for random orderings in continual learning and the Kaczmarz method. Our main results are summarized in Tables 1 and 2.

Much of the related work has been covered throughout the paper. A further discussion of related work can be found in App. A. Here, we briefly highlight additional aspects of our work.

Random Continual Benchmarks. Many popular continual benchmarks in deep learning implicitly assume a random ordering, such as the permuted MNIST benchmark [33]. Our paper shows that in sufficiently long task sequences, random ordering is enough to prevent catastrophic forgetting, and the training loss goes to zero, even in the worst case. In accordance with our results, Lesort et al. [39] examined a random CL benchmark—in which a subset of *classes* is randomly sampled in each task—and observed that forgetting diminishes as more tasks are sampled, even while training with standard SGD (without any modifications to mitigate forgetting). This suggests that random orderings may contaminate continual learning benchmarks, making it harder to isolate the algorithmic effects being tested. Furthermore, real-world tasks often change gradually, not adhering to random orderings. Such "gradually evolving" datasets might be more challenging and relevant as continual benchmarks.

Connections to the Kaczmarz Method. In Section 3.1 we revisited known connections between continual regression and the Kaczmarz method [15]. We broadened this connection in Section 3.2, bridging the *block* Kaczmarz method and "stepwise-optimal" SGD, thus applying our novel SGD bounds to the Kaczmarz method. Using Kaczmarz terminology, given a system $\mathbf{A}\mathbf{x} = \mathbf{b}$ consisting of T blocks of an average rank $\bar{r}$ where $\mathbf{A}_m \in \mathbb{R}^{n_m \times d}, \mathbf{b}_m \in \mathbb{R}^{n_m}$, our rates from Section 4 can be summarized as $\mathbb{E}_\tau\big[\frac{1}{2T}\sum_{m=1}^{T}\left\|\mathbf{A}_m\mathbf{x}_k - \mathbf{b}_m\right\|^2\big] = \mathcal{O}\big(\min\big(k^{-1/4}, \frac{1}{k}\sqrt{d-\bar{r}}, \frac{1}{k}\sqrt{T\bar{r}}\big)\big)$ for random orderings with replacement and $\mathcal{O}\big(\min\big(k^{-1/4}, \frac{1}{k}(d-\bar{r})\big)\big)$ without replacement. Note that we bounded the *loss*, rather than the "error" $\left\|\mathbf{w}_k - \mathbf{w}_\star\right\|^2$, thus enabling the derivation of rates independent of quantities like the condition number that can make convergence arbitrarily slow.

Non-uniform Sampling. The seminal work of Strohmer and Vershynin [66] proposed a Kaczmarz variant that samples rows with probability proportional to their squared norm. Our approach accommodates non-uniform sampling, including norm based ones. This may tighten Theorem 4.2, replacing the dependence on the maximum row norm R with the average one. In the block version, both uniform and non-uniform variants exist [21, 50]. There also, our approach should yield better bounds when weighting *blocks* by to their norms, alleviating the dependence on the *maximum* norm.

Future Work. We narrowed the gap between existing lower and upper worst-case bounds for random orderings in continual linear regression (see Table 1). However, a considerable gap remains between $\Omega(1/k)$ and $\mathcal{O}(1/k^{1/4})$. Generally, we conjecture that the last-iterate SGD rates can be improved beyond those in Theorem 5.1, and that Theorem E.4 can be extended to the multi-epoch setup. Following our reductions (Section 3), improved rates for "stepwise-optimal" SGD rates would immediately refine the bounds for continual linear regression and classification.

References

[1] F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). *Advances in neural information processing systems*, 26, 2013.

[2] M. Banayeeanzade, M. Soltanolkotabi, and M. Rostami. Theoretical insights into overparameterized models in multi-task and replay-based continual learning. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

[3] H. H. Bauschke, P. L. Combettes, et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

[4] R. Berthier, F. Bach, and P. Gaillard. Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model. *Advances in Neural Information Processing Systems*, 33:2576–2586, 2020.

[5] L. Bottou. Curiously fast convergence of some stochastic gradient descent algorithms. In *Proceedings of the symposium on learning and data science, Paris*, volume 8, pages 2624–2633. Citeseer, 2009.

[6] O. Bousquet and A. Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

[7] S. Boyd, J. Dattorro, et al. Alternating projections. *EE392o, Stanford University*, 2003.

[8] S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends®in Machine Learning*, 8(3-4):231–357, 2015.

[9] X. Cai and J. Diakonikolas. Last iterate convergence of incremental methods and applications in continual learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

[10] X. Cai, C. Y. Lin, and J. Diakonikolas. Empirical risk minimization with shuffled sgd: a primal-dual perspective and improved bounds. *arXiv preprint arXiv:2306.12498*, 2023.

[11] J. Cha, J. Lee, and C. Yun. Tighter lower bounds for shuffling sgd: Random permutations and beyond. In *International Conference on Machine Learning*, pages 3855–3912. PMLR, 2023.

[12] C. M. De Sa. Random reshuffling is not always better. *Advances in Neural Information Processing Systems*, 33:5957–5967, 2020.

[13] T. Doan, M. Abbana Bennani, B. Mazoure, G. Rabusseau, and P. 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.

[14] T. Elfving. Block-iterative methods for consistent and inconsistent linear equations. *Numerische Mathematik*, 35(1):1–12, 1980.

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

[16] I. Evron, E. Moroshko, G. Buzaglo, M. Khriesh, B. Marjieh, N. Srebro, and D. Soudry. Continual learning in linear classification on separable data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 9440–9484. PMLR, 23–29 Jul 2023.

[17] F. Facchinei and J.-S. Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.

[18] R. Ge, S. M. Kakade, R. Kidambi, and P. Netrapalli. The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares. *Advances in Neural Information Processing Systems*, 32, 2019.

[19] D. Goldfarb and P. Hand. Analysis of catastrophic forgetting for random orthogonal transformation tasks in the overparameterized regime. In *International Conference on Artificial Intelligence and Statistics*, pages 2975–2993. PMLR, 2023.

[20] D. Goldfarb, I. Evron, N. Weinberger, D. Soudry, and P. Hand. The joint effect of task similarity and overparameterization on catastrophic forgetting - an analytical model. In *The Twelfth International Conference on Learning Representations*, 2024.

[21] R. M. Gower and P. Richtárik. Randomized iterative methods for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1660–1690, 2015.

[22] L. Gubin, B. T. Polyak, and E. Raik. The method of projections for finding the common point of convex sets. *USSR Computational Mathematics and Mathematical Physics*, 7(6):1–24, 1967.

[23] X. Guo, J. Lin, and D.-X. Zhou. Rates of convergence of randomized kaczmarz algorithms in hilbert spaces. *Applied and Computational Harmonic Analysis*, 61:288–318, 2022.

[24] D. Han and J. Xie. A simple linear convergence analysis of the reshuffling kaczmarz method. *arXiv preprint arXiv:2410.01140*, 2024.

[25] M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234. PMLR, 2016.

[26] H. Hemati, L. Pellegrini, X. Duan, Z. Zhao, F. Xia, M. Masana, B. Tscheschner, E. Veas, Y. Zheng, S. Zhao, et al. Continual learning in the presence of repetition. In *CVPR Workshop on Continual Learning in Computer Vision*, 2024.

[27] N. Hiratani. Disentangling and mitigating the impact of task similarity for continual learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[28] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2012.

[29] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[30] P. Jain, D. Nagaraj, and P. Netrapalli. Making the last iterate of sgd information theoretically optimal. In *Conference on Learning Theory*, pages 1752–1755. PMLR, 2019.

[31] H. Jung, H. Cho, and C. Yun. Convergence and implicit bias of gradient descent on continual linear classification. In *The Thirteenth International Conference on Learning Representations*, 2025.

[32] S. Kaczmarz. Angenaherte auflosung von systemen linearer glei-chungen. *Bull. Int. Acad. Pol. Sic. Let., Cl. Sci. Math. Nat.*, pages 355–357, 1937.

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

[34] M. Kong, W. Swartworth, H. Jeong, D. Needell, and R. Ward. Nearly optimal bounds for cyclic forgetting. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[35] T. Koren, R. Livni, Y. Mansour, and U. Sherman. Benign underfitting of stochastic gradient descent. *Advances in Neural Information Processing Systems*, 35:19605–19617, 2022.

[36] Z. Lai and L.-H. Lim. Recht-ré noncommutative arithmetic-geometric mean conjecture is false. In *International Conference on Machine Learning*, pages 5608–5617. PMLR, 2020.

[37] G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.

[38] Y. Lei and Y. Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5809–5819. PMLR, 13–18 Jul 2020.

[39] T. Lesort, O. Ostapenko, P. Rodríguez, D. Misra, M. R. Arefin, L. Charlin, and I. Rish. Challenging common assumptions about catastrophic forgetting and knowledge accumulation. In *Conference on Lifelong Learning Agents*, pages 43–65. PMLR, 2023.

[40] H. Li, J. Wu, and V. Braverman. Fixed design analysis of regularization-based continual learning. In S. Chandar, R. Pascanu, H. Sedghi, and D. Precup, editors, *Proceedings of The 2nd Conference on Lifelong Learning Agents*, volume 232 of *Proceedings of Machine Learning Research*, pages 513–533. PMLR, 22–25 Aug 2023.

[41] H. Li, J. Wu, and V. Braverman. Memory-statistics tradeoff in continual learning with structural regularization. *arXiv preprint arXiv:2504.04039*, 2025.

[42] S. Lin, P. Ju, Y. Liang, and N. Shroff. Theory on forgetting and generalization of continual learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 21078–21100. PMLR, 23–29 Jul 2023.

[43] Z. Liu and Z. Zhou. Revisiting the last-iterate convergence of stochastic gradient methods. In *The Twelfth International Conference on Learning Representations*, 2024.

[44] Z. Liu and Z. Zhou. On the last-iterate convergence of shuffling gradient methods. In *Forty-first International Conference on Machine Learning*, 2024.

[45] S. Ma, R. Bassily, and M. Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pages 3325–3334. PMLR, 2018.

[46] M. Mermillod, A. Bugaiska, and P. Bonin. The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects, 2013.

[47] K. Mishchenko, A. Khaled Ragab Bayoumi, and P. Richtárik. Random reshuffling: Simple analysis with vast improvements. *Advances in Neural Information Processing Systems*, 33, 2020.

[48] D. Nagaraj, P. Jain, and P. Netrapalli. Sgd without replacement: Sharper rates for general smooth convex functions. In *International Conference on Machine Learning*, pages 4703–4711. PMLR, 2019.

[49] A. Nedić. Random projection algorithms for convex set intersection problems. In *49th IEEE Conference on Decision and Control (CDC)*, pages 7655–7660. IEEE, 2010.

[50] D. Needell and J. A. Tropp. Paved with good intentions: analysis of a randomized block kaczmarz method. *Linear Algebra and its Applications*, 441:199–221, 2014.

[51] D. Needell, R. Ward, and N. Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *Advances in neural information processing systems*, 27: 1017–1025, 2014.

[52] Y. Nesterov. Introductory lectures on convex programming volume i: Basic course. *Lecture notes*, 3(4):5, 1998.

[53] F. Orabona. Last iterate of sgd converges (even in unbounded domains), 2020. *Accessed: May*, 2020. URL `https://parameterfree.com/2020/08/07/last-iterate-of-sgd-converges-even-in-unbounded-domains`.

[54] P. Oswald and W. Zhou. Convergence analysis for kaczmarz-type methods in a hilbert space framework. *Linear Algebra and its Applications*, 478:131–161, 2015.

[55] L. Peng, P. Giampouras, and R. Vidal. The ideal continual learner: An agent that never forgets. In *International Conference on Machine Learning*, 2023.

[56] S. Rajput, A. Gupta, and D. Papailiopoulos. Closing the convergence gap of SGD without replacement. In *International Conference on Machine Learning*, pages 7964–7973. PMLR, 2020.

[57] B. Recht and C. Ré. Beneath the valley of the noncommutative arithmetic-geometric mean inequality: conjectures, case-studies, and consequences. In *Conference on Learning Theory (COLT)*, 2012.

[58] B. Recht and C. Ré. Toward a noncommutative arithmetic-geometric mean inequality: conjectures, case-studies, and consequences. In *Conference on Learning Theory*, pages 11–1. JMLR Workshop and Conference Proceedings, 2012.

[59] I. Safran and O. Shamir. How good is SGD with random shuffling? In *Conference on Learning Theory*, pages 3250–3284. PMLR, 2020.

[60] S. G. Sankaran and A. L. Beex. Convergence behavior of affine projection algorithms. *IEEE Transactions on Signal Processing*, 48(4):1086–1096, 2000.

[61] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

[62] O. Shamir and T. Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 71–79. PMLR, 2013.

[63] U. Sherman, T. Koren, and Y. Mansour. Optimal rates for random order online optimization. *Advances in Neural Information Processing Systems*, 34:2097–2108, 2021.

[64] D. T. Slock. On the convergence behavior of the lms and the normalized lms algorithms. *IEEE Transactions on Signal Processing*, 41(9):2811–2825, 1993.

[65] N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low noise and fast rates. *Advances in neural information processing systems*, 23, 2010.

[66] T. Strohmer and R. Vershynin. A randomized kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009.

[67] A. V. Varre, L. Pillaud-Vivien, and N. Flammarion. Last iterate convergence of SGD for least-squares in the interpolation regime. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[68] S. Vaswani, F. Bach, and M. Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd international conference on artificial intelligence and statistics*, pages 1195–1204. PMLR, 2019.

[69] J. Wu, D. Zou, V. Braverman, Q. Gu, and S. Kakade. Last iterate risk bounds of sgd with decaying stepsize for overparameterized linear regression. In *International Conference on Machine Learning*, pages 24280–24314. PMLR, 2022.

[70] C. Yun, S. Sra, and A. Jadbabaie. Open problem: Can single-shuffle sgd be better than reshuffling sgd and gd? In *Conference on Learning Theory*, pages 4653–4658. PMLR, 2021.

[71] M. Zamani and F. Glineur. Exact convergence rate of the last iterate in subgradient methods. *arXiv preprint arXiv:2307.11134*, 2023.

[72] X. Zhao, H. Wang, W. Huang, and W. Lin. A statistical theory of regularization-based continual learning. In *Forty-first International Conference on Machine Learning*, 2024.

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

A Related Work

Most of the related work is already discussed in the main body of the paper. Here, we elaborate on several interesting connections that remain open.

Last-iterate Guarantees for SGD. For the general (non-realizable) smooth stochastic setup, the recent work of Liu and Zhou [43] was the first (and only, to our knowledge) to provide upper bounds on the convergence rate of the last SGD iterate. While their bounds are applicable in the realizable setting, they require non-constant step sizes to obtain non-trivial convergence, and are therefore not useful for our purposes (see Table 2). Our analysis technique in Section 5 borrows from the work of Shamir and Zhang [62, also mentioned in Table 2] which, in fact, belongs to the comparatively-richer line of work on the non-smooth setting [30, 43, 62, 71]. Notably, SGD in a stochastic non-realizable (either smooth or non-smooth) setup requires uniformly bounded noise assumptions, and generally cannot accommodate a constant step size independent of the optimization horizon.

Table 2: **State-of-the-art Loss Bounds for Fixed-Step-Size SGD.** We consider stochastic convex optimization with an objective $\bar{f}(\mathbf{w}) \triangleq \mathbb{E}_\xi f(\mathbf{w};\xi)$, where $f(\cdot;\xi)$ is β-smooth almost surely, $\sigma^2 \geq \mathbb{E}\|\nabla f(\mathbf{w};\xi) - \nabla \bar{f}(\mathbf{w})\|^2$, $\sigma_\star^2 \triangleq \mathbb{E}\|\nabla f(\mathbf{w}_\star;\xi) - \nabla \bar{f}(\mathbf{w}_\star)\|^2$, and $G > 0$ is such that $\|\nabla f(\mathbf{w};\xi)\| \leq G$ for any $\mathbf{w}$ and ξ. Dependence on constant numerical factors and the distance to an optimal solution is suppressed.

Setting	Reference	Bound at Iteration T	Last Iterate Guarantee	Convergence for $\eta = 1/\beta$
Stochastic	(∗) Shamir and Zhang [62]	$\frac{1}{\eta T} + \eta G^2 \log T$	✓	✗
Deterministic Smooth ($\sigma = 0$)	Nesterov [52]	$\frac{1}{(2-\eta\beta)\eta T}$	✓	✓
Stochastic Smooth	Lan [37]	$\frac{1}{\eta T} + \eta\sigma^2$	✗	✗
	Liu and Zhou [43]	$\frac{1}{\eta T} + \eta\sigma^2 \log T$	✓	✗
Stochastic Smooth Realizable ($\sigma_\star = 0$)	Srebro et al. [65]	$\frac{1}{(1-\eta\beta)\eta T}$	✗	✗
Stochastic Regression Realizable ($\sigma_\star = 0$)	Bach and Moulines [1]	$\frac{1}{\eta T}$	✗	✓
	Varre et al. [67]	$\frac{1}{(1-2\eta\beta\log T)\eta T}$	✓	✗
	Ours (2025)	$\frac{1}{(2-\eta\beta)\eta T^{1-\eta\beta(1-\eta\beta/4)}}$	✓	✓

(∗) Shamir and Zhang [62] consider bounded domains; Liu and Zhou [43], Orabona [53] obtain similar bounds for the unconstrained case. For non-fixed step sizes Jain et al. [30] obtain minimax optimal bounds without log factors.

Our analysis for SGD *without*-replacement is related to a long line of work primarily focused on the average iterate convergence rates [e.g., 10, 11, 47, 48, 56, 58, 59]. For the non-strongly convex case, near-optimal bounds (for the average iterate) have been established for the general smooth case [47, 48]. In a subsequent work, Cai et al. [10] refined the dependence on problem parameters for the smooth realizable case (among others). Guarantees for the *last* iterate have only been established recently by Liu and Zhou [44] and Cai and Diakonikolas [9]. However, their bounds decay with the number of epochs rather than the number of iterations and apply only to non-constant step sizes, making them inapplicable to our setting. Specifically, in a realizable β-smooth setup, after J without-replacement SGD epochs over a finite sum of size n, Cai et al. [10], Mishchenko et al. [47] obtained an $O(\beta/J)$ bound for the average iterate with step size $\eta = 1/(\beta n)$; and Cai and

Diakonikolas [9], Liu and Zhou [44] derived similar bounds for the last iterate up to logarithmic factors.

Another line of work related to ours studies algorithmic stability [6, 61] of gradient methods, which is the main technique we use in the proof of Theorem E.4. Our approach is similar in nature to that of Koren et al. [35], Nagaraj et al. [48], Sherman et al. [63] and primarily builds on Sherman et al. [63], who were the first to formally introduce the notion of without-replacement stability. For with-replacement SGD, [25] discussed its algorithmic stability under smooth loss functions. Later, [38], improved this bound in the realizable loss case. The case we consider—*i.e.,* the stability of without-replacement SGD under smooth and realizable loss functions—is not covered in the existing literature.

With versus Without Replacement in Kaczmarz Methods. Our results in Section 4 establish universal bounds for random orderings, both with and without replacement. Both the with- and without-replacement variants converge linearly towards the minimum-norm solution $\mathbf{w}_\star$ [21, 24], but as we explained in Section 7, the rates can be arbitrarily slow. Recht and Ré [57] formulated a noncommutative analog of the arithmetic-geometric mean inequality that, if true, could have shown that without-replacement orderings lead to faster loss convergence than with-replacement orderings in Kaczmarz methods, and consequently in continual linear regression. Years later, Lai and Lim [36] proved that this inequality does not hold in general [see also 12]. Moreover, as in other areas, empirical studies found that row shuffling followed by cyclic orderings performs as well as i.i.d. orderings [54]. This naturally connects to interesting observations and open questions regarding various forms of shuffled SGD [5, 70]. Our rates are similar for both with- and without-replacement orderings (up to small constants), meaning they do not indicate a clear advantage for either. However, we believe they are far from tight, leaving interesting open questions in this direction.

Connections to Normalized Least Mean Squares. The NLMS algorithm is a classical adaptive filtering method. In its simplest version [64], the method perfectly fits a single—usually noisy—random sample at a time, using the same update rule as the Kaczmarz method (and thus, as our continual Scheme 1 in a rank-1 case). There also exists a more complex version of this method, which uses more samples per update [60]. Both papers give strong $\mathcal{O}(1/k)$ MSE rates in the noiseless setting (matching our realizable setting). However, they assume a very limited data model, where the sampled vectors are either orthogonal or identical up-to-scaling. Under such conditions, Evron et al. [15] showed that there is no forgetting (of previously learned tasks), implying that the MSE decays as the number of tasks still unseen at time k.

B Auxiliary Proofs

Lemma B.1 (Bounding Forgetting Using the Training Loss)**.** *In a realizable setting (Assumption 2.2), the iterates of Scheme 1 under a random task ordering* τ *(with or without replacement) hold* $\forall k \geq 1$*,*

$$\mathbb{E}_\tau[F_\tau(k)] = \mathbb{E}_\tau\left[\frac{1}{2k}\sum_{t=1}^{k}\left\|\mathbf{X}_{\tau(t)}\mathbf{w}_k - \mathbf{y}_{\tau(t)}\right\|^2\right] \leq \mathbb{E}_\tau\left\|\mathbf{X}_{\tau(k)}\mathbf{w}_{k-1} - \mathbf{y}_{\tau(k)}\right\|^2 + \frac{\left\|\mathbf{w}_\star\right\|^2 R^2}{k},$$

where $R \triangleq \max_{m\in[T]}\|\mathbf{X}_m\|$ *is the "radius" of the data. Notice that the dependence of* $\mathbf{w}_{k-1}$ *on* $\tau_1, \ldots, \tau_{k-1}$ *is implicit. Particularly, in an ordering with replacement, we get,*

$$\mathbb{E}_\tau[F_\tau(k)] \leq \mathbb{E}_\tau\left[\frac{1}{T}\sum_{m=1}^{T}\|\mathbf{X}_m\mathbf{w}_{k-1} - \mathbf{y}_m\|^2\right] + \frac{\left\|\mathbf{w}_\star\right\|^2 R^2}{k} = 2\mathbb{E}_\tau\left[\mathcal{L}\left(\mathbf{w}_{k-1}\right)\right] + \frac{\left\|\mathbf{w}_\star\right\|^2 R^2}{k}.$$

Proof. As discussed in Section 3.1, Scheme 2 governs the updates of the iterates $\mathbf{w}_t \in \mathbb{R}^d$. Under Assumption 2.2, we define the orthogonal projection as $\mathbf{P}_{\tau(t)} \triangleq \mathbf{I}_d - \mathbf{X}^+_{\tau(t)}\mathbf{X}_{\tau(t)}$, revealing a recursive form:

$$\begin{aligned}
\mathbf{w}_t &= \mathbf{X}^+_{\tau(t)}\mathbf{y}_{\tau(t)} + \left(\mathbf{I}_d - \mathbf{X}^+_{\tau(t)}\mathbf{X}_{\tau(t)}\right)\mathbf{w}_{t-1} \\
[\text{Assumption 2.2}] &= \mathbf{X}^+_{\tau(t)}\mathbf{X}_{\tau(t)}\mathbf{w}_\star + \left(\mathbf{I}_d - \mathbf{X}^+_{\tau(t)}\mathbf{X}_{\tau(t)}\right)\mathbf{w}_{t-1} = (\mathbf{I}_d - \mathbf{P}_{\tau(t)})\mathbf{w}_\star + \mathbf{P}_{\tau(t)}\mathbf{w}_{t-1} \\
\mathbf{w}_t - \mathbf{w}_\star &= \mathbf{P}_{\tau(t)}\left(\mathbf{w}_{t-1} - \mathbf{w}_\star\right) && (3) \\
\mathbf{w}_t - \mathbf{w}_\star &= \mathbf{P}_{\tau(t)}\cdots\mathbf{P}_{\tau(1)}\left(\mathbf{w}_0 - \mathbf{w}_\star\right). && (4)
\end{aligned}$$

We show that,

$$\begin{aligned}
\mathbb{E}_\tau\left[F_\tau(k)\right] &= \frac{1}{2k}\sum_{t=1}^{k}\mathbb{E}_\tau\left\|\mathbf{X}_{\tau(t)}\mathbf{w}_k - \mathbf{y}_{\tau(t)}\right\|^2 = \frac{1}{2k}\sum_{t=1}^{k}\mathbb{E}_\tau\left\|\mathbf{X}_{\tau(t)}\left(\mathbf{w}_k - \mathbf{w}_\star\right)\right\|^2 \\
&= \frac{1}{2k}\sum_{t=1}^{k}\mathbb{E}_\tau\left\|\mathbf{X}_{\tau(t)}\mathbf{P}_{\tau(k)}\cdots\mathbf{P}_{\tau(t+1)}\mathbf{P}_{\tau(t)}\left(\mathbf{w}_{t-1} - \mathbf{w}_\star\right)\right\|^2 \\
&= \frac{1}{2k}\sum_{t=1}^{k}\mathbb{E}_\tau\left\|\mathbf{X}_{\tau(t)}\mathbf{P}_{\tau(k)}\cdots\mathbf{P}_{\tau(t+1)}\left(\mathbf{I} - \mathbf{P}_{\tau(t)} - \mathbf{I}\right)\left(\mathbf{w}_{t-1} - \mathbf{w}_\star\right)\right\|^2 \\
[\text{Jensen}] &\leq \frac{1}{k}\sum_{t=1}^{k}\Bigg(\mathbb{E}_\tau\underbrace{\left\|\mathbf{X}_{\tau(t)}\mathbf{P}_{\tau(k)}\cdots\mathbf{P}_{\tau(t+1)}\left(\mathbf{I} - \mathbf{P}_{\tau(t)}\right)\left(\mathbf{w}_{t-1} - \mathbf{w}_\star\right)\right\|^2}_{\leq R^2\left\|\left(\mathbf{I}-\mathbf{P}_{\tau(t)}\right)\left(\mathbf{w}_{t-1}-\mathbf{w}_\star\right)\right\|^2,\text{ since projections contract}} + \\
&\qquad\qquad \mathbb{E}_\tau\left\|\mathbf{X}_{\tau(t)}\mathbf{P}_{\tau(k)}\cdots\mathbf{P}_{\tau(t+1)}\left(\mathbf{w}_{t-1} - \mathbf{w}_\star\right)\right\|^2\Bigg) \\
&\leq \frac{1}{k}\sum_{t=1}^{k}\Bigg(R^2\mathbb{E}_\tau\left\|\left(\mathbf{I} - \mathbf{P}_{\tau(t)}\right)\left(\mathbf{w}_{t-1} - \mathbf{w}_\star\right)\right\|^2 + \\
&\qquad\qquad \mathbb{E}_\tau\left\|\mathbf{X}_{\tau(t)}\mathbf{P}_{\tau(k)}\cdots\mathbf{P}_{\tau(t+1)}\mathbf{P}_{\tau(t-1)}\cdots\mathbf{P}_{\tau(1)}\left(\mathbf{w}_0 - \mathbf{w}_\star\right)\right\|^2\Bigg).
\end{aligned}$$

For the first term, we employ the Pythagorean theorem for orthogonal projections to get a telescoping sum and show that

$$
\begin{aligned}
&\frac{R^2}{k}\sum_{t=1}^{k}\mathbb{E}_\tau\left\|\left(\mathbf{I}-\mathbf{P}_{\tau(t)}\right)\left(\mathbf{w}_{t-1}-\mathbf{w}_\star\right)\right\|^2\\
&=\frac{R^2}{k}\sum_{t=1}^{k}\left(\mathbb{E}_\tau\left\|\mathbf{w}_{t-1}-\mathbf{w}_\star\right\|^2-\mathbb{E}_\tau\left\|\mathbf{P}_{\tau(t)}\left(\mathbf{w}_{t-1}-\mathbf{w}_\star\right)\right\|^2\right)\\
&=\frac{R^2}{k}\sum_{t=1}^{k}\left(\mathbb{E}_\tau\left\|\mathbf{w}_{t-1}-\mathbf{w}_\star\right\|^2-\mathbb{E}_\tau\left\|\mathbf{w}_{t}-\mathbf{w}_\star\right\|^2\right)\\
&=\frac{R^2}{k}\Big(\underbrace{\mathbb{E}_\tau\left\|\mathbf{w}_{0}-\mathbf{w}_\star\right\|^2}_{=\left\|\mathbf{w}_\star\right\|^2}-\underbrace{\mathbb{E}_\tau\left\|\mathbf{w}_{k}-\mathbf{w}_\star\right\|^2}_{\geq 0}\Big)\leq\frac{\left\|\mathbf{w}_\star\right\|^2R^2}{k}\,.
\end{aligned}
$$

For the second term, we use the exchangeability of τ which applies with or without replacement,

$$
\begin{aligned}
&\mathbb{E}_\tau\left\|\mathbf{X}_{\tau(t)}\mathbf{P}_{\tau(k)}\cdots\mathbf{P}_{\tau(t+1)}\mathbf{P}_{\tau(t-1)}\cdots\mathbf{P}_{\tau(1)}\left(\mathbf{w}_0-\mathbf{w}_\star\right)\right\|^2\\
&=\mathbb{E}_\tau\left\|\mathbf{X}_{\tau(k)}\mathbf{P}_{\tau(k-1)}\cdots\mathbf{P}_{\tau(1)}\left(\mathbf{w}_0-\mathbf{w}_\star\right)\right\|^2=\mathbb{E}_\tau\left\|\mathbf{X}_{\tau(k)}\left(\mathbf{w}_{k-1}-\mathbf{w}_\star\right)\right\|^2\,.
\end{aligned}
$$

Combining the two, we get

$$
\mathbb{E}_\tau\left[F_\tau\left(k\right)\right]\leq\mathbb{E}_\tau\left\|\mathbf{X}_{\tau(k)}\mathbf{w}_{k-1}-\mathbf{y}_{\tau(k)}\right\|^2+\frac{\left\|\mathbf{w}_\star\right\|^2R^2}{k}\,,
$$

which completes the first part of the proof.

For the second part, simply notice that in an i.i.d. setting, the index $\tau(k)\sim\text{Unif}\left([T]\right)$ is independent of earlier indices (which yielded $\mathbf{w}_{k-1}$), and thus

$$
\mathbb{E}_\tau\left\|\mathbf{X}_{\tau(k)}\mathbf{w}_{k-1}-\mathbf{y}_{\tau(k)}\right\|^2=\mathbb{E}_\tau\left[\frac{1}{T}\sum_{m=1}^{T}\left\|\mathbf{X}_m\mathbf{w}_{k-1}-\mathbf{y}_m\right\|^2\right]\,.
$$

■

Proposition B.2 (Bounding The Training Loss Using Forgetting in Without-Replacement Orderings)**.** *Under a random ordering τ without replacement, the iterates of Scheme 1 (continual regression) satisfy $\forall k\in[T]$:*

$$
\mathbb{E}_\tau\left[\mathcal{L}\left(\mathbf{w}_k\right)\right]=\frac{k}{T}\mathbb{E}_\tau\left[F_\tau\left(k\right)\right]+\frac{T-k}{2T}\mathbb{E}_\tau\left\|\mathbf{X}_{\tau(k+1)}\mathbf{w}_k-\mathbf{y}_{\tau(k+1)}\right\|^2\,.
$$

Similarly, the iterates of Scheme 4 (POCS) satisfy:

$$
\mathbb{E}_\tau\left[\mathcal{L}\left(\mathbf{w}_k\right)\right]=\frac{k}{T}\mathbb{E}_\tau\left[F_\tau\left(k\right)\right]+\frac{T-k}{2T}\mathbb{E}_\tau\left\|\mathbf{w}_k-\mathbf{\Pi}_{\tau(k+1)}\left(\mathbf{w}_k\right)\right\|^2\,,
$$

where in such a POCS setting, the loss and forgetting are defined as:

$$
\mathcal{L}\left(\mathbf{w}_k\right)=\frac{1}{2T}\sum_{m=1}^{T}\left\|\mathbf{w}_k-\mathbf{\Pi}_m\left(\mathbf{w}_k\right)\right\|^2,\quad F_\tau\left(k\right)=\frac{1}{2k}\sum_{t=1}^{k}\left\|\mathbf{w}_k-\mathbf{\Pi}_{\tau(t)}\left(\mathbf{w}_k\right)\right\|^2\,.
$$

Proof. We first prove the claim in the continual regression setting. If $k = T$ then $\mathbb{E}_\tau [\mathcal{L}(\mathbf{w}_k)] = \mathbb{E}_\tau [F_\tau(k)]$, and the claim follows. For $k < T$, we have:

$$\begin{aligned}
\mathbb{E}_\tau [\mathcal{L}(\mathbf{w}_k)] &= \frac{1}{2T}\sum_{m=1}^{T} \mathbb{E}_\tau \|\mathbf{X}_m \mathbf{w}_k - \mathbf{y}_m\|^2 \\
[\text{without replacement}] &= \frac{1}{2T}\sum_{t=1}^{T} \mathbb{E}_\tau \left\|\mathbf{X}_{\tau(t)} \mathbf{w}_k - \mathbf{y}_{\tau(t)}\right\|^2 \\
&= \frac{1}{2T}\sum_{t=1}^{k} \mathbb{E}_\tau \left\|\mathbf{X}_{\tau(t)} \mathbf{w}_k - \mathbf{y}_{\tau(t)}\right\|^2 + \frac{1}{2T}\sum_{t=k+1}^{T} \mathbb{E}_\tau \left\|\mathbf{X}_{\tau(t)} \mathbf{w}_k - \mathbf{y}_{\tau(t)}\right\|^2 \\
&= \frac{k}{T}\mathbb{E}_\tau [F_\tau(k)] + \frac{1}{2T}\sum_{t=k+1}^{T} \mathbb{E}_\tau \left\|\mathbf{X}_{\tau(t)} \mathbf{w}_k - \mathbf{y}_{\tau(t)}\right\|^2 \\
[\text{exchangeability}] &= \frac{k}{T}\mathbb{E}_\tau [F_\tau(k)] + \frac{T-k}{2T}\mathbb{E}_\tau \left\|\mathbf{X}_{\tau(k+1)} \mathbf{w}_k - \mathbf{y}_{\tau(k+1)}\right\|^2 .
\end{aligned}$$

For the POCS case, simply replace $\|\mathbf{X}_m \mathbf{w}_k - \mathbf{y}_m\|^2$ with $\|\mathbf{w}_k - \mathbf{\Pi}_m(\mathbf{w}_k)\|^2$. ■

Recall Lemma 3.1. Consider any realizable task collection such that $\mathbf{X}_m \mathbf{w}_\star = \mathbf{y}_m, \forall m \in [T]$. Define $f_m(\mathbf{w}) = \frac{1}{2}\|\mathbf{X}_m^+ \mathbf{X}_m(\mathbf{w} - \mathbf{w}_\star)\|^2$. Then, $\forall m \in [T], \mathbf{w} \in \mathbb{R}^d$

(i) **Upper bound:** $\mathcal{L}_m(\mathbf{w}) \le R^2 f_m(\mathbf{w}) \triangleq \max_{m' \in [T]} \|\mathbf{X}_{m'}\|^2 f_m$.

(ii) **Gradient:** $\nabla_{\mathbf{w}} f_m(\mathbf{w}) = \mathbf{X}_m^+ \mathbf{X}_m(\mathbf{w} - \mathbf{w}_\star) = \mathbf{X}_m^+ \mathbf{X}_m \mathbf{w} - \mathbf{X}_m^+ \mathbf{y}_m$.

(iii) **Convexity and Smoothness:** f_m is convex and 1-smooth.

Proof. First, we use the realizability and simple norm inequalities to obtain,

$$\mathcal{L}_m(\mathbf{w}) = \tfrac{1}{2}\|\mathbf{X}_m \mathbf{w} - \mathbf{y}_m\|^2 = \tfrac{1}{2}\|\mathbf{X}_m(\mathbf{w} - \mathbf{w}_\star)\|^2 \le \tfrac{\|\mathbf{X}_m\|^2}{2}\left\|\mathbf{X}_m^+ \mathbf{X}_m(\mathbf{w} - \mathbf{w}_\star)\right\|^2 \le R^2 f(\mathbf{w}) .$$

Since $\mathbf{X}_m^+ \mathbf{X}_m$ is an orthogonal projection operator—and thus symmetric and idempotent—we get,

$$\nabla_{\mathbf{w}} f_m(\mathbf{w}) = \left(\mathbf{X}_m^+ \mathbf{X}_m\right)^\top \mathbf{X}_m^+ \mathbf{X}_m(\mathbf{w} - \mathbf{w}_\star) = \mathbf{X}_m^+ \mathbf{X}_m(\mathbf{w} - \mathbf{w}_\star) = \mathbf{X}_m^+ \mathbf{X}_m \mathbf{w} - \mathbf{X}_m^+ \mathbf{y}_m .$$

Then, the above and the fact that projection operators are non-expansive imply that $\forall \mathbf{w}, \mathbf{z} \in \mathbb{R}^d$,

$$\|\nabla_{\mathbf{w}} f_m(\mathbf{w}) - \nabla_{\mathbf{z}} f_m(\mathbf{z})\| = \left\|\mathbf{X}_m^+ \mathbf{X}_m(\mathbf{w} - \mathbf{w}_\star - \mathbf{z} + \mathbf{w}_\star)\right\| = \left\|\mathbf{X}_m^+ \mathbf{X}_m(\mathbf{w} - \mathbf{z})\right\| \le \|\mathbf{w} - \mathbf{z}\| .$$

Finally, the convexity of f_m is immediate since $\nabla^2_{\mathbf{w}} f_m(\mathbf{w}) = \mathbf{X}_m^+ \mathbf{X}_m \succeq \mathbf{0}$. ■

C Proofs for Section 4.1: A Parameter-Dependent $\mathcal{O}(1/k)$ Rate

Recall Theorem 4.1. Under a random ordering with replacement over T jointly realizable tasks, the expected loss and forgetting of Schemes 1, 2 after $k \geq 3$ iterations are upper bounded as,

$$\mathbb{E}_\tau\left[\mathcal{L}\left(\mathbf{w}_k\right)\right] = \mathbb{E}_\tau\left[\frac{1}{2T}\sum_{m=1}^{T}\left\|\mathbf{X}_m\mathbf{w}_k - \mathbf{y}_m\right\|^2\right] \leq \frac{\min\left(\sqrt{d-\bar{r}}, \sqrt{T\bar{r}}\right)}{2e(k-1)}\|\mathbf{w}_\star\|^2 R^2$$
$$\mathbb{E}_\tau\left[F_\tau(k)\right] = \mathbb{E}_\tau\left[\frac{1}{2k}\sum_{t=1}^{k}\left\|\mathbf{X}_{\tau(t)}\mathbf{w}_t - \mathbf{y}_{\tau(t)}\right\|^2\right] \leq \frac{3\min\left(\sqrt{d-\bar{r}}, \sqrt{T\bar{r}}\right)}{2\left(k-2\right)}\|\mathbf{w}_\star\|^2 R^2\,,$$

where $\bar{r} \triangleq \frac{1}{T}\sum_{m\in[T]}\operatorname{rank}(\mathbf{X}_m)$. (Recall that $R \triangleq \max_{m\in[T]}\|\mathbf{X}_m\|$.)

Here, we prove the main result, followed by the necessary auxiliary corollaries and lemmas in App. C.1.

Proof Idea. We rewrite the Kaczmarz update (Scheme 2) in a recursive form of the differences, *i.e.,* $\mathbf{w}_t - \mathbf{w}_\star = \mathbf{P}_{\tau(t)}\left(\mathbf{w}_{t-1} - \mathbf{w}_\star\right)$, with a suitable projection matrix $\mathbf{P}_{\tau(t)}$. We define the linear map $Q\left[\mathbf{A}\right] = \frac{1}{T}\sum_{m=1}^{T}\mathbf{P}_m\mathbf{A}\mathbf{P}_m$ to capture the evolution of the difference's second moments, enabling sharp analysis of the expected loss in terms of Q. Using properties of Q, norm inequalities, and the spectral mapping theorem, we establish a fast $\mathcal{O}\left(1/k\right)$ rate with explicit dependence on T, d, and $\bar{r}$.

Proof. We analyze the randomized block Kaczmarz algorithm for solving the linear system $\mathbf{X}\mathbf{w} = \mathbf{y}$, where the matrix and vector are partitioned into blocks as follows:

$$\mathbf{X} = \begin{pmatrix}\mathbf{X}_1\\ \vdots \\ \mathbf{X}_T\end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix}\mathbf{y}_1\\ \vdots \\ \mathbf{y}_T\end{pmatrix}.$$

By defining $\mathbf{z}_t = \mathbf{w}_t - \mathbf{w}_\star$ and exploiting the recursive form of Eq. (3) from the proof of Lemma B.1, we obtain $\mathbf{z}_t = \mathbf{P}_{\tau(t)}\mathbf{z}_{t-1}$. Note that $\mathbf{z}_0 = \mathbf{0}_d - \mathbf{w}_\star = -\mathbf{w}_\star$.

Now, define the linear map $Q : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ as

$$Q\left[\mathbf{A}\right] = \mathop{\mathbb{E}}_{m\sim\text{Unif}([T])}\left[\mathbf{P}_m\mathbf{A}\mathbf{P}_m\right] = \frac{1}{T}\sum_{m=1}^{T}\mathbf{P}_m\mathbf{A}\mathbf{P}_m. \tag{5}$$

This map plays a central role in our analysis and has been studied in similar forms in prior work [23]. Note that $\mathbf{P}_m$ is an orthogonal projection, which implies that it is symmetric and idempotent. Thus,

$$\mathop{\mathbb{E}}_{\tau}\big[\mathbf{z}_{t+1}\mathbf{z}_{t+1}^\top\big] = \mathop{\mathbb{E}}_{m,\tau}\big[\mathbf{P}_m\mathbf{z}_t\mathbf{z}_t^\top\mathbf{P}_m^\top\big] = \mathop{\mathbb{E}}_{m,\tau}\big[\mathbf{P}_m\mathbf{z}_t\mathbf{z}_t^\top\mathbf{P}_m\big] = \mathop{\mathbb{E}}_{m}\left[\mathbf{P}_m\mathop{\mathbb{E}}_{\tau}\left[\mathbf{z}_t\mathbf{z}_t^\top\right]\mathbf{P}_m\right] = Q\left[\mathop{\mathbb{E}}_{\tau}\left[\mathbf{z}_t\mathbf{z}_t^\top\right]\right].$$

It follows that

$$\mathop{\mathbb{E}}_{\tau}\left[\mathbf{z}_t\mathbf{z}_t^\top\right] = Q^t\left[\mathop{\mathbb{E}}_{\tau}\left[\mathbf{z}_0\mathbf{z}_0^\top\right]\right] = Q^t\left[\mathbf{z}_0\mathbf{z}_0^\top\right] = Q^t\left[\left(\mathbf{w}_0 - \mathbf{w}_\star\right)\left(\mathbf{w}_0 - \mathbf{w}_\star\right)^\top\right] = Q^t\left[\mathbf{w}_\star\mathbf{w}_\star^\top\right]\,,$$

where Q^t denotes t applications of Q. The map Q captures the evolution of the error's second-moment under Kaczmarz updates, offering a tractable approach to analyzing the algorithm's convergence. The expected loss at step t is given by

$$\begin{aligned}
\mathbb{E}_\tau\left[\mathcal{L}\left(\mathbf{w}_t\right)\right] &= \mathop{\mathbb{E}}_{\tau}\left[\frac{1}{2T}\sum_{i=1}^{T}\left\|\mathbf{X}_i\mathbf{w}_t - \mathbf{y}_i\right\|^2\right] = \mathop{\mathbb{E}}_{\tau}\left[\frac{1}{2T}\sum_{i=1}^{T}\left\|\mathbf{X}_i\left(\mathbf{w}_t - \mathbf{w}_\star\right)\right\|^2\right]\\
&= \mathop{\mathbb{E}}_{\tau}\left[\frac{1}{2T}\sum_{i=1}^{T}\left\|\mathbf{X}_i\mathbf{z}_t\right\|^2\right] = \mathop{\mathbb{E}}_{\tau}\left[\frac{1}{2T}\left\|\mathbf{X}\mathbf{z}_t\right\|^2\right]\\
&= \mathop{\mathbb{E}}_{\tau}\left[\frac{1}{2T}\mathbf{z}_t^\top\mathbf{X}^\top\mathbf{X}\mathbf{z}_t\right] = \mathop{\mathbb{E}}_{\tau}\left[\operatorname{tr}\left(\frac{1}{2T}\mathbf{X}^\top\mathbf{X}\mathbf{z}_t\mathbf{z}_t^\top\right)\right]\\
&= \operatorname{tr}\left(\frac{1}{2T}\mathbf{X}^\top\mathbf{X}\mathop{\mathbb{E}}_{\tau}\left[\mathbf{z}_t\mathbf{z}_t^\top\right]\right) = \operatorname{tr}\left(\frac{1}{2T}\mathbf{X}^\top\mathbf{X}Q^t\left[\mathbf{w}_\star\mathbf{w}_\star^\top\right]\right).
\end{aligned}$$

We are now ready to derive the final bound. From Lemma C.7, we have

$$\frac{1}{R^2 T}\mathbf{X}^\top \mathbf{X} \preccurlyeq \mathbf{X}^{+}\mathbf{X} - Q\left[\mathbf{X}^{+}\mathbf{X}\right].$$

Additionally, by Corollary C.5, $Q^k\left[\mathbf{w}_\star \mathbf{w}_\star^\top\right]$ is symmetric and positive semidefinite (PSD). We also note that $\frac{1}{T}\mathbf{X}^\top\mathbf{X}$ is symmetric PSD. The key insight from Lemma C.7, combined with the trace product inequality (Lemma C.6), is that it allows the expected loss to be expressed using a polynomial in Q. This reformulation simplifies the convergence analysis by reducing it to examining the spectral properties of Q. Invoking the trace product inequality, we obtain:

$$\begin{aligned}
\mathbb{E}_\tau\left[\mathcal{L}\left(\mathbf{w}_k\right)\right] &= \operatorname{tr}\left(\frac{1}{2T}\mathbf{X}^\top\mathbf{X}Q^t\left[\mathbf{w}_\star\mathbf{w}_\star^\top\right]\right) \le \frac{R^2}{2}\operatorname{tr}\left(\left(\mathbf{X}^{+}\mathbf{X} - Q\left[\mathbf{X}^{+}\mathbf{X}\right]\right)Q^k\left[\mathbf{w}_\star\mathbf{w}_\star^\top\right]\right) \\
\left[\text{Lemma C.8}\right] &= \frac{R^2}{2}\operatorname{tr}\left(Q^k\left[\mathbf{X}^{+}\mathbf{X} - Q\left[\mathbf{X}^{+}\mathbf{X}\right]\right]\mathbf{w}_\star\mathbf{w}_\star^\top\right) = \frac{R^2}{2}\mathbf{w}_\star^\top Q^k\left[\mathbf{X}^{+}\mathbf{X} - Q\left[\mathbf{X}^{+}\mathbf{X}\right]\right]\mathbf{w}_\star \\
&\le \frac{\left\|\mathbf{w}_\star\right\|^2 R^2}{2}\left\|Q^k\left[\mathbf{X}^{+}\mathbf{X} - Q\left[\mathbf{X}^{+}\mathbf{X}\right]\right]\right\|_2 = \frac{\left\|\mathbf{w}_\star\right\|^2 R^2}{2}\left\|\left(Q^k\left(I-Q\right)\right)\left[\mathbf{X}^{+}\mathbf{X}\right]\right\|_2 \\
&= \frac{\left\|\mathbf{w}_\star\right\|^2 R^2}{2}\left\|\left(Q^{k-1}\left(I-Q\right)\right)Q\left[\mathbf{X}^{+}\mathbf{X}\right]\right\|_2 \\
&\le \frac{\left\|\mathbf{w}_\star\right\|^2 R^2}{2}\left\|\left(Q^{k-1}\left(I-Q\right)\right)Q\left[\mathbf{X}^{+}\mathbf{X}\right]\right\|_F \\
\left[\text{operator norm}\right] &\le \frac{\left\|\mathbf{w}_\star\right\|^2 R^2}{2}\left\|Q^{k-1}\left(I-Q\right)\right\|\cdot\left\|Q\left[\mathbf{X}^{+}\mathbf{X}\right]\right\|_F \\
\left[\substack{\text{Lemmas}\\ \text{C.11, C.12}}\right] &\le \frac{\left\|\mathbf{w}_\star\right\|^2 R^2}{2e\left(k-1\right)}\min\left(\sqrt{T\bar{r}}, \sqrt{d-\bar{r}}\right).
\end{aligned}$$

To clarify, the operator norm of a linear map H is defined as $\|H\| = \sup_{\mathbf{A}\in\mathbb{R}^{d\times d}, \|\mathbf{A}\|_F=1}\|H\left[\mathbf{A}\right]\|_F$. The reason for switching from the spectral norm to the Frobenius norm is to enable the use of the spectral mapping theorem to bound the operator norm of $Q^{k-1}\left(I-Q\right)$, applicable only for inner-product-based norms. We complete the proof by bounding the forgetting using the training loss (Lemma B.1). That is,

$$\begin{aligned}
\mathbb{E}_\tau[F_\tau(k)] = \mathbb{E}_\tau\left[\frac{1}{2k}\sum_{t=1}^{k}\left\|\mathbf{X}_{\tau(t)}\mathbf{w}_t - \mathbf{y}_{\tau(t)}\right\|^2\right] &\le 2\mathbb{E}_\tau\left[\mathcal{L}\left(\mathbf{w}_{k-1}\right)\right] + \frac{\left\|\mathbf{w}_\star\right\|^2 R^2}{k} \\
&\le \frac{3\left\|\mathbf{w}_\star\right\|^2 R^2}{2\left(k-2\right)}\min\left(\sqrt{T\bar{r}}, \sqrt{d-\bar{r}}\right).
\end{aligned}$$

■

C.1 Key Properties and Auxiliary Lemmas

Definition C.1 (Positive Map)**.** A positive map $H : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ is a linear map that maps PSD matrices to PSD matrices. Formally, if $\mathbf{0} \preccurlyeq \mathbf{A} \in \mathbb{R}^{d\times d}$, then $\mathbf{0} \preccurlyeq H[\mathbf{A}]$.

Definition C.2 (Symmetric Map)**.** A symmetric map $H : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ is a linear map that maps symmetric matrices to symmetric matrices. Formally, if $\mathbf{A} = \mathbf{A}^\top \in \mathbb{R}^{d\times d}$, then $H[\mathbf{A}] = H[\mathbf{A}]^\top$.

Corollary C.3. *Q, defined in Eq.* (5)*, is a positive map.*

Proof. Let $\mathbf{0} \preccurlyeq \mathbf{A} \in \mathbb{R}^{d\times d}$. Then, for all $i \in [T]$, $\mathbf{0} \preccurlyeq \mathbf{P}_i\mathbf{A}\mathbf{P}_i$. Meaning $Q[\mathbf{A}]$ is PSD as a convex combination of PSD matrices. ■

Corollary C.4. *Q is a symmetric map. Moreover, for all $\mathbf{A} \in \mathbb{R}^{d\times d}$, it satisfies $Q[\mathbf{A}]^\top = Q\left[\mathbf{A}^\top\right]$.*

Proof. Let $\mathbf{A} \in \mathbb{R}^{d\times d}$. Then,

$$Q[\mathbf{A}]^\top = \frac{1}{T}\sum_{i=1}^{T}(\mathbf{P}_i\mathbf{A}\mathbf{P}_i)^\top = \frac{1}{T}\sum_{i=1}^{T}\mathbf{P}_i^\top\mathbf{A}^\top\mathbf{P}_i^\top = \frac{1}{T}\sum_{i=1}^{T}\mathbf{P}_i\mathbf{A}^\top\mathbf{P}_i = Q\left[\mathbf{A}^\top\right].$$

■

Corollary C.5. *For $n \in \mathbb{N}^+$, the iterated application of the map Q, denoted Q^n, is a positive symmetric map.*

Proof. For $n = 1$, given by Corollaries C.3 and C.4. For $n > 1$, this follows trivially by induction. ■

Lemma C.6 (Trace Product Inequality)**.** *Let $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{d\times d}$ be symmetric PSD matrices such that $\mathbf{A} \preccurlyeq \mathbf{B}$. Then,* $\operatorname{tr}(\mathbf{AC}) \le \operatorname{tr}(\mathbf{BC})$.

Proof. Since $\mathbf{0} \preccurlyeq \mathbf{C} = \mathbf{C}^\top$, it has a square symmetric PSD root $\mathbf{C}^{1/2}$. Given that $\mathbf{A}, \mathbf{B}$ are symmetric and $\mathbf{A} \preccurlyeq \mathbf{B}$, it follows that $\mathbf{C}^{1/2}\mathbf{A}\mathbf{C}^{1/2} \preccurlyeq \mathbf{C}^{1/2}\mathbf{B}\mathbf{C}^{1/2}$ [from 28, Theorem 7.7.2.a]. Applying the cyclic property of the trace and using the fact that for symmetric matrices ordered in the Löwner sense, their traces are also ordered [28, Corollary 7.7.4.d], we obtain

$$\operatorname{tr}(\mathbf{AC}) = \operatorname{tr}\left(\mathbf{A}\mathbf{C}^{1/2}\mathbf{C}^{1/2}\right) = \operatorname{tr}\left(\mathbf{C}^{1/2}\mathbf{A}\mathbf{C}^{1/2}\right) \le \operatorname{tr}\left(\mathbf{C}^{1/2}\mathbf{B}\mathbf{C}^{1/2}\right) = \operatorname{tr}(\mathbf{BC}).$$

■

Lemma C.7. *Let $R = \max_{i\in[T]}\|\mathbf{X}_i\|$. Then, $\frac{1}{R^2T}\mathbf{X}^\top\mathbf{X} \preccurlyeq \mathbf{X}^+\mathbf{X} - Q[\mathbf{X}^+\mathbf{X}]$*

Proof. We perform SVD on each $\mathbf{X}_i = \mathbf{U}_i\mathbf{\Sigma}_i\mathbf{V}_i^\top$. Then,

$$\frac{1}{R^2T}\mathbf{X}^\top\mathbf{X} = \frac{1}{R^2T}\sum_{i=1}^{T}\mathbf{X}_i^\top\mathbf{X}_i = \frac{1}{R^2T}\sum_{i=1}^{T}\mathbf{V}_i\mathbf{\Sigma}_i^2\mathbf{V}_i^\top$$

On the other hand:

$$\begin{aligned}
\mathbf{X}^+\mathbf{X} - Q\left[\mathbf{X}^+\mathbf{X}\right] &= \mathbf{X}^+\mathbf{X} - \frac{1}{T}\sum_{i=1}^{T}\left(\mathbf{I} - \mathbf{X}_i^+\mathbf{X}_i\right)\mathbf{X}^+\mathbf{X}\left(\mathbf{I} - \mathbf{X}_i^+\mathbf{X}_i\right) \\
&= \mathbf{X}^+\mathbf{X} - \frac{1}{T}\sum_{i=1}^{T}\mathbf{X}^+\mathbf{X} - \mathbf{X}_i^+\mathbf{X}_i\mathbf{X}^+\mathbf{X} - \mathbf{X}^+\mathbf{X}\mathbf{X}_i^+\mathbf{X}_i + \mathbf{X}_i^+\mathbf{X}_i\mathbf{X}^+\mathbf{X}\mathbf{X}_i^+\mathbf{X}_i \\
\left[\substack{\operatorname{Im}\left(\mathbf{X}_i^+\mathbf{X}_i\right) \\ \subseteq\operatorname{Im}\left(\mathbf{X}^+\mathbf{X}\right)}\right] &= -\frac{1}{T}\sum_{i=1}^{T} -\mathbf{X}_i^+\mathbf{X}_i - \mathbf{X}_i^+\mathbf{X}_i + \mathbf{X}_i^+\mathbf{X}_i = \frac{1}{T}\sum_{i=1}^{T}\mathbf{X}_i^+\mathbf{X}_i = \frac{1}{T}\sum_{i=1}^{T}\mathbf{V}_i\mathbf{\Sigma}_i^+\mathbf{\Sigma}_i\mathbf{V}_i^\top.
\end{aligned}$$

Now consider the difference:

$$\left(\mathbf{X}^+\mathbf{X} - Q\left[\mathbf{X}^+\mathbf{X}\right]\right) - \frac{1}{R^2T}\mathbf{X}^\top\mathbf{X} = \frac{1}{T}\sum_{i=1}^{T}\mathbf{V}_i\left(\mathbf{\Sigma}_i^+\mathbf{\Sigma}_i - \frac{1}{R^2}\mathbf{\Sigma}_i^2\right)\mathbf{V}_i^\top.$$

We know that $\frac{1}{R}(\mathbf{\Sigma}_i)_{j,j} \in [0,1]$. We analyze two cases for each diagonal entry:

- If $(\mathbf{\Sigma}_i)_{j,j} = 0$, then $\left(\mathbf{\Sigma}_i^{+}\mathbf{\Sigma}_i - \frac{1}{R^2}\mathbf{\Sigma}_i^2\right)_{j,j} = 0$.
- Otherwise, $\left(\mathbf{\Sigma}_i^{+}\mathbf{\Sigma}_i\right)_{j,j} = 1$, and $\frac{1}{R^2}\left(\mathbf{\Sigma}_i^2\right)_{j,j} \leq 1$, which gives $\left(\mathbf{\Sigma}_i^{+}\mathbf{\Sigma}_i - \frac{1}{R^2}\mathbf{\Sigma}_i^2\right)_{j,j} \geq 0$.

Thus,

$$\mathbf{0} \preccurlyeq \mathbf{V}_i \left(\mathbf{\Sigma}_i^{+}\mathbf{\Sigma}_i - \frac{1}{R^2}\mathbf{\Sigma}_i^2\right)\mathbf{V}_i^{\top}\,.$$

Averaging over all i, we get:

$$\mathbf{0} = \frac{1}{T}\sum_{i=1}^{T}\mathbf{0} \preccurlyeq \frac{1}{T}\sum_{i=1}^{T}\mathbf{V}_i\left(\mathbf{\Sigma}_i^{+}\mathbf{\Sigma}_i - \frac{1}{R^2}\mathbf{\Sigma}_i^2\right)\mathbf{V}_i^{\top} = \left(\mathbf{X}^{+}\mathbf{X} - Q\left[\mathbf{X}^{+}\mathbf{X}\right]\right) - \frac{1}{R^2T}\mathbf{X}^{\top}\mathbf{X}$$

$$\frac{1}{R^2T}\mathbf{X}^{\top}\mathbf{X} \preccurlyeq \mathbf{X}^{+}\mathbf{X} - Q\left[\mathbf{X}^{+}\mathbf{X}\right]\,.$$

■

Lemma C.8. *Let* $\mathbf{A},\mathbf{B} \in \mathbb{R}^{d\times d}$ *and* $n \in \mathbb{N}^{+}$. *Then,* $\operatorname{tr}\left(\mathbf{A}Q^{n}\left[\mathbf{B}\right]\right) = \operatorname{tr}\left(Q^{n}\left[\mathbf{A}\right]\mathbf{B}\right)$.

Proof. From the definition of Q (Eq. (5)),

$$\begin{aligned}
\operatorname{tr}\left(\mathbf{A}Q^{n}\left[\mathbf{B}\right]\right) &= \operatorname{tr}\left(\mathbf{A}\frac{1}{T^n}\sum_{j_1,\dots,j_n=1}^{T}\mathbf{P}_{j_1}\cdots\mathbf{P}_{j_n}\mathbf{B}\mathbf{P}_{j_n}\cdots\mathbf{P}_{j_1}\right)\\
\left[\text{linearity}\right] &= \frac{1}{T^n}\sum_{j_1,\dots,j_n=1}^{T}\operatorname{tr}\left(\mathbf{A}\mathbf{P}_{j_1}\cdots\mathbf{P}_{j_n}\mathbf{B}\mathbf{P}_{j_n}\cdots\mathbf{P}_{j_1}\right)\\
\left[\text{cyclic property}\right] &= \frac{1}{T^n}\sum_{j_1,\dots,j_n=1}^{T}\operatorname{tr}\left(\mathbf{P}_{j_n}\cdots\mathbf{P}_{j_1}\mathbf{A}\mathbf{P}_{j_1}\cdots\mathbf{P}_{j_n}\mathbf{B}\right)\\
\left[\text{linearity}\right] &= \operatorname{tr}\left(\left(\frac{1}{T^n}\sum_{j_1\dots,j_n=1}^{T}\mathbf{P}_{j_n}\cdots\mathbf{P}_{j_1}\mathbf{A}\mathbf{P}_{j_1}\cdots\mathbf{P}_{j_n}\right)\mathbf{B}\right)\\
&= \operatorname{tr}\left(Q^{n}\left[\mathbf{A}\right]\mathbf{B}\right)\,.
\end{aligned}$$

■

Proposition C.9. *Q is self adjoint.*

Proof. Let $\mathbf{A},\mathbf{B} \in \mathbb{R}^{d\times d}$. Then,

$$\begin{aligned}
\left\langle Q\left[\mathbf{A}\right],\mathbf{B}\right\rangle &= \operatorname{tr}\left(Q\left[\mathbf{A}\right]^{\top}\mathbf{B}\right) = \operatorname{tr}\left(\mathbf{B}^{\top}Q\left[\mathbf{A}\right]\right)\\
\left[\text{Lemma C.8}\right] &= \operatorname{tr}\left(Q\left[\mathbf{B}^{\top}\right]\mathbf{A}\right)\\
\left[\text{Corollary C.4}\right] &= \operatorname{tr}\left(Q\left[\mathbf{B}\right]^{\top}\mathbf{A}\right) = \operatorname{tr}\left(\mathbf{A}^{\top}Q\left[\mathbf{B}\right]\right) = \left\langle\mathbf{A},Q\left[\mathbf{B}\right]\right\rangle\,.
\end{aligned}$$

■

Proposition C.10. *The spectrum of Q is contained in the interval* $[0,1]$.

Proof. Let $\mathbf{A} \in \mathbb{R}^{d\times d}$. Then, by definition,

$$\begin{aligned}
\left\langle Q\left[\mathbf{A}\right],\mathbf{A}\right\rangle &= \operatorname{tr}\left(Q\left[\mathbf{A}\right]^{\top}\mathbf{A}\right) = \frac{1}{T}\sum_{i=1}^{T}\operatorname{tr}\left(\mathbf{P}_i\mathbf{A}^{\top}\mathbf{P}_i\mathbf{A}\right)\\
\begin{bmatrix}\text{idempotence,}\\ \text{cyclic property}\end{bmatrix} &= \frac{1}{T}\sum_{i=1}^{T}\operatorname{tr}\left(\mathbf{P}_i\mathbf{A}^{\top}\mathbf{P}_i\mathbf{P}_i\mathbf{A}\mathbf{P}_i\right) = \frac{1}{T}\sum_{i=1}^{T}\left\|\mathbf{P}_i\mathbf{A}\mathbf{P}_i\right\|_F^2 \geq 0\,.
\end{aligned}$$

Since each $\mathbf{P}_i$ is an orthogonal projection, its spectral norm satisfies $\|\mathbf{P}_i\|_2 = 1$. Applying the operator inequality $\|\mathbf{XY}\|_F \leq \|\mathbf{X}\|_2 \|\mathbf{Y}\|_F$ twice, we obtain

$$\frac{1}{T}\sum_{i=1}^{T} \|\mathbf{P}_i \mathbf{A} \mathbf{P}_i\|_F^2 \leq \|\mathbf{P}_i\|_2^4 \|\mathbf{A}\|_F^2 = \|\mathbf{A}\|_F^2 .$$

Thus, for any $\mathbf{A} \in \mathbb{R}^{d\times d}$,

$$0 \leq \langle Q[\mathbf{A}], \mathbf{A}\rangle \leq \|\mathbf{A}\|_F^2 .$$

From the Rayleigh quotient characterization of eigenvalues, this implies that every eigenvalue λ of Q satisfies $0 \leq \lambda \leq 1$, *i.e.*, $\sigma(Q) \subset [0,1]$. ■

Lemma C.11. $\|Q^n (I - Q)\| \leq \frac{1}{en}$, *for* $n \in \mathbb{N}^+$.

Proof. By Proposition C.9, Q is self adjoint. Thus, we can apply the spectral mapping theorem to the polynomial $x \mapsto x^n(1-x)$. The eigenvalues of $Q^n(I-Q)$ are of the form $\lambda^n(1-\lambda)$, where λ is an eigenvalue of Q. From Proposition C.10, we know that $\lambda \in [0,1]$. Using an algebraic property of $\lambda^n(1-\lambda)$ for $\lambda \in [0,1]$, we conclude that $\lambda^n(1-\lambda) \in \left[0, \frac{1}{en}\right]$.
Therefore, $\|Q^n(I-Q)\| \leq \frac{1}{en}$. ■

Lemma C.12. $\|Q[\mathbf{X}^+\mathbf{X}]\|_F \leq \min\left(\sqrt{T\bar{r}}, \sqrt{d-\bar{r}}\right)$.

Proof. We first bound $\|Q[\mathbf{X}^+\mathbf{X}]\|_F$ using the operator norm bound on Q (Proposition C.10):

$$\left\|Q\left[\mathbf{X}^+\mathbf{X}\right]\right\|_F \leq \underbrace{\|Q\|}_{\leq 1} \cdot \left\|\mathbf{X}^+\mathbf{X}\right\|_F \leq \left\|\mathbf{X}^+\mathbf{X}\right\|_F = \sqrt{\operatorname{rank}\left(\mathbf{X}^+\mathbf{X}\right)} = \sqrt{T\bar{r}} .$$

Next, we use a pseudo-inverse property—that $\mathbf{X}^+\mathbf{X} \preccurlyeq \mathbf{I}$—and the positivity of Q to show,

$$\begin{aligned}
\mathbf{0} &\preccurlyeq Q\left[\mathbf{I} - \mathbf{X}^+\mathbf{X}\right] \\
Q\left[\mathbf{X}^+\mathbf{X}\right] &\preccurlyeq Q[\mathbf{I}] \\
\left\|Q\left[\mathbf{X}^+\mathbf{X}\right]\right\|_F &\leq \|Q[\mathbf{I}]\|_F = \left\|\frac{1}{T}\sum_{i=1}^{T}\mathbf{P}_i\right\|_F \leq \frac{1}{T}\sum_{i=1}^{T}\|\mathbf{P}_i\|_F \\
&= \frac{1}{T}\sum_{i=1}^{T}\sqrt{\operatorname{rank}(\mathbf{P}_i)} = \frac{1}{T}\sum_{i=1}^{T}\sqrt{d - \operatorname{rank}(\mathbf{X}_i)} \\
[\text{Jensen (concave)}] &\leq \sqrt{d-\bar{r}} .
\end{aligned}$$

■

D Proofs of Universal Continual Regression Rates (Sections 4.2 and 4.3)

The proofs in this appendix focus on the properties of forgetting and loss, "translating" them into the language of last-iterate SGD. We then apply our last-iterate results, proved in App. E.

D.1 Proof of Theorem 4.2: A Universal $\mathcal{O}(1/\sqrt[4]{k})$ Rate

Recall Theorem 4.2. Under a random ordering with replacement over T jointly realizable tasks, the expected loss and forgetting of Schemes 1, 2 after $k \geq 2$ iterations are bounded as,

$$\mathbb{E}_\tau\left[\mathcal{L}\left(\mathbf{w}_k\right)\right] = \mathbb{E}_\tau\left[\frac{1}{2T}\sum_{m=1}^{T}\left\|\mathbf{X}_m\mathbf{w}_k - \mathbf{y}_m\right\|^2\right] \leq \frac{2}{\sqrt[4]{k}}\left\|\mathbf{w}_\star\right\|^2 R^2\,,$$

$$\mathbb{E}_\tau\left[F_\tau(k)\right] = \mathbb{E}_\tau\left[\frac{1}{2k}\sum_{t=1}^{k}\left\|\mathbf{X}_{\tau(t)}\mathbf{w}_k - \mathbf{y}_{\tau(t)}\right\|^2\right] \leq \frac{5}{\sqrt[4]{k-1}}\left\|\mathbf{w}_\star\right\|^2 R^2\,.$$

Proof. Let τ be a random with-replacement ordering, and $\mathbf{w}_0, \ldots, \mathbf{w}_k$ be the corresponding iterates produced by the continual Scheme 1 (or the equivalent Kaczmarz Scheme 2). By Reduction 2, these are exactly the (stochastic) gradient descent iterates produced given an initialization $\mathbf{w}_0$ and a step size of $\eta = 1$, on the loss sequence $f_{\tau(1)}, \ldots, f_{\tau(k)}$, where we defined:

$$f_m(\mathbf{w}) \triangleq \frac{1}{2}\left\|\mathbf{X}_m^+\mathbf{X}_m(\mathbf{w}-\mathbf{w}_\star)\right\|^2.$$

Furthermore, Lemma 3.1 states that for all $\mathbf{w} \in \mathbb{R}^d$,

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2T}\sum_{m=1}^{T}\left\|\mathbf{X}_m\mathbf{w} - \mathbf{y}_m\right\|^2 = \mathbb{E}_{m\sim\text{Unif}([T])}\mathcal{L}_m(\mathbf{w}) \leq R^2\mathbb{E}_{m\sim\text{Unif}([T])}f_m(\mathbf{w})\,.$$

Therefore, establishing last iterate convergence of with-replacement SGD (Eq. (2)) on the objective function

$$\bar{f}(\mathbf{w}) \triangleq \mathbb{E}_{m\sim[T]}f_m(\mathbf{w})\,,$$

will imply the desired result. Indeed, again by Lemma 3.1, $f_m(\cdot)$ is 1-smooth for all $m \in [T]$. Hence, plugging in $\mathbf{A} = \mathbf{X}_m^+\mathbf{X}_m \Rightarrow \|\mathbf{A}\| = 1 = \beta$ into Theorem 5.1, SGD with $\eta = 1$ guarantees that after $k \geq 1$ gradient steps:

$$\mathbb{E}\bar{f}(\mathbf{w}_k) \leq \frac{e\left\|\mathbf{w}_0 - \mathbf{w}_\star\right\|^2}{2\sqrt[4]{k}} \leq \frac{2\left\|\mathbf{w}_0 - \mathbf{w}_\star\right\|^2}{\sqrt[4]{k}}\,,$$

and therefore $\mathbb{E}\mathcal{L}(\mathbf{w}_k) \leq \frac{2R^2\|\mathbf{w}_0-\mathbf{w}_\star\|^2}{\sqrt[4]{k}}$, which proves the first claim. The second claim follows immediately from Lemma B.1, and we are done. ∎

D.2 Proving Theorem 4.3: Main Result for Without Replacement Orderings

Recall Theorem 4.3. Under a random ordering without replacement over T jointly realizable tasks, the expected loss and forgetting of Schemes 1, 2 after $k \in \{2, \ldots, T\}$ iterations are both bounded as,

$$\mathbb{E}\left[\mathcal{L}\left(\mathbf{w}_k\right)\right], \mathbb{E}\left[F_\tau(k)\right] \leq \min\left(\frac{7}{\sqrt[4]{k-1}}, \frac{d-\bar{r}+1}{k-1}\right)\left\|\mathbf{w}_\star\right\|^2 R^2 .$$

Proof. From Lemmas 3.1 and B.1, we have

$$\mathop{\mathbb{E}}_{\tau}\left[F_\tau\left(k\right)\right] \leq \mathop{\mathbb{E}}_{\tau}\left\|\mathbf{X}_{\tau(k)}\mathbf{w}_{k-1}-\mathbf{y}_{\tau(k)}\right\|^2 + \frac{\left\|\mathbf{w}_\star\right\|^2 R^2}{k} \leq 2R^2 \mathop{\mathbb{E}}_{\tau} f_{\tau(k)}(\mathbf{w}_{k-1}) + \frac{\left\|\mathbf{w}_\star\right\|^2 R^2}{k} .$$

Combining with Proposition B.2, we get,

$$\begin{aligned}
\mathbb{E}_\tau\left[\mathcal{L}\left(\mathbf{w}_k\right)\right] &= \frac{k}{T}\mathbb{E}_\tau\left[F_\tau\left(k\right)\right] + \frac{T-k}{2T}\mathbb{E}_\tau\left\|\mathbf{X}_{\tau(k+1)}\mathbf{w}_k-\mathbf{y}_{\tau(t)}\right\|^2 \\
&\leq \frac{k}{T}\left(2R^2 \mathop{\mathbb{E}}_{\tau} f_{\tau(k)}(\mathbf{w}_{k-1}) + \frac{\left\|\mathbf{w}_\star\right\|^2 R^2}{k}\right) + \frac{T-k}{2T}\mathop{\mathbb{E}}_{\tau} f_{\tau(k+1)}(\mathbf{w}_k) \\
[k \leq T] &\leq R^2\left(\frac{2k}{T}\mathop{\mathbb{E}}_{\tau} f_{\tau(k)}(\mathbf{w}_{k-1}) + \frac{T-k}{T}\mathop{\mathbb{E}}_{\tau} f_{\tau(k+1)}(\mathbf{w}_k)\right) + \frac{\left\|\mathbf{w}_\star\right\|^2 R^2}{k} .
\end{aligned}$$

Thus, to bound both the expected forgetting and loss, we need to bound expressions like $\mathbb{E}_\tau f_{\tau(k+1)}(\mathbf{w}_k)$.

We first prove the *dimension dependent* term. Note that,

$$2\mathop{\mathbb{E}}_{\tau} f_{\tau(k)}(\mathbf{w}_{k-1}) = \mathop{\mathbb{E}}_{\tau}\left\|\mathbf{X}_{\tau(k)}^{+}\mathbf{X}_{\tau(k)}\left(\mathbf{w}_{k-1}-\mathbf{w}_\star\right)\right\|^2 \triangleq \mathop{\mathbb{E}}_{\tau}\left\|\left(\mathbf{I}-\mathbf{P}_{\tau(k)}\right)\left(\mathbf{w}_{k-1}-\mathbf{w}_\star\right)\right\|^2 .$$

Recall that from Eq. (4) in the proof of Lemma B.1, we have

$$\left(\mathbf{w}_{k-1}-\mathbf{w}_\star\right) = \mathbf{P}_{\tau(k-1)}\cdots\mathbf{P}_{\tau(1)}\left(\mathbf{w}_0-\mathbf{w}_\star\right) = -\mathbf{P}_{\tau(k-1)}\cdots\mathbf{P}_{\tau(1)}\mathbf{w}_\star .$$

Thus, we obtain

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\tau}\left\|\left(\mathbf{I}-\mathbf{P}_{\tau(k)}\right)\left(\mathbf{w}_{k-1}-\mathbf{w}_\star\right)\right\|^2 = \mathop{\mathbb{E}}_{\tau}\left\|\left(\mathbf{I}-\mathbf{P}_{\tau(k)}\right)\mathbf{P}_{\tau(k-1)}\cdots\mathbf{P}_{\tau(1)}\mathbf{w}_\star\right\|^2 \\
&\leq \mathop{\mathbb{E}}_{\tau}\left\|\left(\mathbf{I}-\mathbf{P}_{\tau(k)}\right)\mathbf{P}_{\tau(k-1)}\cdots\mathbf{P}_{\tau(1)}\right\|_2^2 \cdot \left\|\mathbf{w}_\star\right\|^2 \leq \left\|\mathbf{w}_\star\right\|^2 \mathop{\mathbb{E}}_{\tau}\left\|\left(\mathbf{I}-\mathbf{P}_{\tau(k)}\right)\mathbf{P}_{\tau(k-1)}\cdots\mathbf{P}_{\tau(1)}\right\|_F^2 \\
&= \left\|\mathbf{w}_\star\right\|^2 \mathop{\mathbb{E}}_{\tau}\operatorname{tr}\left(\mathbf{P}_{\tau(1)}\cdots\mathbf{P}_{\tau(k-1)}\left(\mathbf{I}-\mathbf{P}_{\tau(k)}\right)\mathbf{P}_{\tau(k-1)}\cdots\mathbf{P}_{\tau(1)}\right) .
\end{aligned}$$

By exchangeability,

$$\begin{aligned}
&\operatorname{tr}\left(\mathbf{P}_{\tau(1)}\cdots\mathbf{P}_{\tau(t-1)}\left(\mathbf{I}-\mathbf{P}_{\tau(t)}\right)\mathbf{P}_{\tau(t-1)}\cdots\mathbf{P}_{\tau(1)}\right) \\
&= \operatorname{tr}\left(\mathbf{P}_{\tau(t)}\cdots\mathbf{P}_{\tau(2)}\left(\mathbf{I}-\mathbf{P}_{\tau(1)}\right)\mathbf{P}_{\tau(2)}\cdots\mathbf{P}_{\tau(t)}\right) .
\end{aligned}$$

Let us define $a_t = \operatorname{tr}\left(\mathbf{P}_{\tau(t)}\cdots\mathbf{P}_{\tau(2)}\left(\mathbf{I}-\mathbf{P}_{\tau(1)}\right)\mathbf{P}_{\tau(2)}\cdots\mathbf{P}_{\tau(t)}\right)$. Then, we have

$$\begin{aligned}
a_{t+1} &= \operatorname{tr}\left(\mathbf{P}_{\tau(t+1)}\cdots\mathbf{P}_{\tau(2)}\left(\mathbf{I}-\mathbf{P}_{\tau(1)}\right)\mathbf{P}_{\tau(2)}\cdots\mathbf{P}_{\tau(t+1)}\right) \\
[\text{cyclic property of trace}] &= \operatorname{tr}\left(\mathbf{P}_{\tau(t+1)}^2\mathbf{P}_{\tau(t)}\cdots\mathbf{P}_{\tau(2)}\left(\mathbf{I}-\mathbf{P}_{\tau(1)}\right)\mathbf{P}_{\tau(2)}\cdots\mathbf{P}_{\tau(t)}\right) \\
[\text{Von Neumann's trace inequality}] &\leq \underbrace{\left\|\mathbf{P}_{\tau(t+1)}^2\right\|_2}_{=1}\operatorname{tr}\left(\mathbf{P}_{\tau(t)}\cdots\mathbf{P}_{\tau(2)}\left(\mathbf{I}-\mathbf{P}_{\tau(1)}\right)\mathbf{P}_{\tau(2)}\cdots\mathbf{P}_{\tau(t)}\right) = a_t ,
\end{aligned}$$

showing $(a_t)_t$ is a non-increasing sequence. Thus, for all $k \geq 2$,

$$2\mathop{\mathbb{E}}_{\tau} f_{\tau(k)}(\mathbf{w}_{k-1}) = \mathop{\mathbb{E}}_{\tau}\left\|\left(\mathbf{I}-\mathbf{P}_{\tau(k)}\right)\mathbf{w}_{k-1}\right\|^2 \leq \left\|\mathbf{w}_\star\right\|^2 \mathop{\mathbb{E}}_{\tau} a_k \leq \tfrac{\left\|\mathbf{w}_\star\right\|^2}{k-1}\sum_{t=2}^{k}\mathop{\mathbb{E}}_{\tau} a_t$$

$$
\begin{aligned}
&= \tfrac{\|\mathbf{w}_\star\|^2}{k-1} \sum_{t=2}^{k} \mathop{\mathbb{E}}_{\tau} \left[\operatorname{tr}\left(\mathbf{P}_{\tau(t)} \cdots \mathbf{P}_{\tau(2)} \cdots \mathbf{P}_{\tau(t)}\right) - \operatorname{tr}\left(\mathbf{P}_{\tau(t)} \cdots \mathbf{P}_{\tau(1)} \cdots \mathbf{P}_{\tau(t)}\right)\right] \\
\left[\text{exchangeability}\right] &= \tfrac{\|\mathbf{w}_\star\|^2}{k-1} \sum_{t=2}^{k} \mathop{\mathbb{E}}_{\tau} \left[\operatorname{tr}\left(\mathbf{P}_{\tau(t-1)} \cdots \mathbf{P}_{\tau(1)} \cdots \mathbf{P}_{\tau(t-1)}\right) - \operatorname{tr}\left(\mathbf{P}_{\tau(t)} \cdots \mathbf{P}_{\tau(1)} \cdots \mathbf{P}_{\tau(t)}\right)\right] \\
\left[\text{telescoping}\right] &= \tfrac{\|\mathbf{w}_\star\|^2}{k-1} \mathop{\mathbb{E}}_{\tau} \left[\operatorname{tr}\left(\mathbf{P}_{\tau(1)}\right) - \operatorname{tr}\left(\mathbf{P}_{\tau(k)} \cdots \mathbf{P}_{\tau(1)} \cdots \mathbf{P}_{\tau(k)}\right)\right] \\
&\le \tfrac{\|\mathbf{w}_\star\|^2}{k-1} \mathop{\mathbb{E}}_{\tau} \left[\operatorname{tr}\left(\mathbf{P}_{\tau(1)}\right)\right] = \frac{\|\mathbf{w}_\star\|^2 (d-\bar{r})}{k-1}.
\end{aligned}
$$

For the second, *parameter independent* term, note that from Lemma 3.1, $f_m(\cdot)$ is 1-smooth for all $m \in [T]$, and recall that the iterates $\mathbf{w}_t$ follow the SGD dynamics with $\eta = 1$ (Reduction 2). Hence, by Lemma E.5, without-replacement SGD with $\beta = \eta = 1$ guarantees that after $k \ge 1$ gradient steps:

$$
\mathop{\mathbb{E}}_{\tau} f_{\tau(k)}(\mathbf{w}_{k-1}) \le \frac{e \cdot \|\mathbf{w}_\star\|^2}{\sqrt[4]{k-1}}.
$$

Plugging in the (monotonic decreasing) bounds that we just derived in the inequalities from the beginning of this proof, we get

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\tau} \left[F_\tau(k)\right] &\le 2R^2 \mathop{\mathbb{E}}_{\tau} f_{\tau(k)}(\mathbf{w}_{k-1}) + \frac{\|\mathbf{w}_\star\|^2 R^2}{k} \\
&\le R^2 \min\left(\frac{2e \|\mathbf{w}_\star\|^2}{\sqrt[4]{k-1}}, \frac{\|\mathbf{w}_\star\|^2 (d-\bar{r})}{k-1}\right) + \frac{\|\mathbf{w}_\star\|^2 R^2}{k} \\
&\le \min\left(\frac{7}{\sqrt[4]{k-1}}, \frac{d-\bar{r}+1}{k-1}\right) \|\mathbf{w}_\star\|^2 R^2,
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}_\tau \left[\mathcal{L}(\mathbf{w}_k)\right] &\le R^2 \left(\frac{k}{T} 2 \mathop{\mathbb{E}}_{\tau} f_{\tau(k)}(\mathbf{w}_{k-1}) + \frac{T-k}{2T} 2 \mathop{\mathbb{E}}_{\tau} f_{\tau(k+1)}(\mathbf{w}_k)\right) + \frac{\|\mathbf{w}_\star\|^2 R^2}{k} \\
&\le \left(\frac{k}{T} + \frac{T-k}{2T}\right) \min\left(\frac{2e}{\sqrt[4]{k-1}}, \frac{d-\bar{r}}{k-1}\right) \|\mathbf{w}_\star\|^2 R^2 + \frac{\|\mathbf{w}_\star\|^2 R^2}{k} \\
&= \frac{T+k}{2T} \min\left(\frac{2e}{\sqrt[4]{k-1}}, \frac{d-\bar{r}}{k-1}\right) \|\mathbf{w}_\star\|^2 R^2 + \frac{\|\mathbf{w}_\star\|^2 R^2}{k} \\
\left[k \le T\right] &\le \min\left(\frac{7}{\sqrt[4]{k-1}}, \frac{d-\bar{r}+1}{k-1}\right) \|\mathbf{w}_\star\|^2 R^2.
\end{aligned}
$$

■

E Proofs of Last-Iterate SGD Bounds (Section 5)

In this section we provide proofs and full technical details of our upper bounds for least squares SGD. We begin by recording a few elementary well-known facts, which can be found in e.g., Bubeck [8]. We provide proof for completeness.

Lemma E.1 (Fundamental regret inequality for gradient descent). *Let $\mathbf{w}_0 \in \mathbb{R}^d, \eta > 0$, and suppose $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{g}_t$ for all t, where $\mathbf{g}_0, \ldots, \mathbf{g}_T \in \mathbb{R}^d$ are arbitrary vectors. Then for any $\tilde{\mathbf{w}} \in \mathbb{R}^d$ it holds that:*

$$\sum_{t=0}^{T} \mathbf{g}_t^\top (\mathbf{w}_t - \tilde{\mathbf{w}}) \leq \frac{\|\mathbf{w}_0 - \tilde{\mathbf{w}}\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=0}^{T} \|\mathbf{g}_t\|^2.$$

Proof. Observe,

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \tilde{\mathbf{w}}\|^2 &= \|\mathbf{w}_t - \tilde{\mathbf{w}}\|^2 - 2\eta \mathbf{g}_t^\top (\mathbf{w}_t - \tilde{\mathbf{w}}) + \eta^2 \|\mathbf{g}_t\|^2 \\
\iff \mathbf{g}_t^\top (\mathbf{w}_t - \tilde{\mathbf{w}}) &= \frac{1}{2\eta} \left(\|\mathbf{w}_t - \tilde{\mathbf{w}}\|^2 - \|\mathbf{w}_{t+1} - \tilde{\mathbf{w}}\|^2 \right) + \frac{\eta}{2} \|\mathbf{g}_t\|^2 .
\end{aligned}$$

Summing the above over $t = 0, \ldots, T$ and telescoping the sum leads to,

$$\begin{aligned}
\sum_{t=0}^{T} \mathbf{g}_t^\top (\mathbf{w}_t - \tilde{\mathbf{w}}) &= \frac{1}{2\eta} \left(\|\mathbf{w}_0 - \tilde{\mathbf{w}}\|^2 - \|\mathbf{w}_{T+1} - \tilde{\mathbf{w}}\|^2 \right) + \frac{\eta}{2} \sum_{t=0}^{T} \|\mathbf{g}_t\|^2 \\
&\leq \frac{\|\mathbf{w}_0 - \tilde{\mathbf{w}}\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=0}^{T} \|\mathbf{g}_t\|^2 ,
\end{aligned}$$

which completes the proof. ∎

Lemma E.2 (Descent lemma). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be β-smooth for $\beta > 0$, and suppose $\min_{\mathbf{w}} f(\mathbf{w}) \in \mathbb{R}$ is attained. Then, for any $\eta > 0$, $\mathbf{w} \in \mathbb{R}^d$, we have for $\mathbf{w}^+ = \mathbf{w} - \eta \nabla f(\mathbf{w})$:*

$$f(\mathbf{w}^+) \leq f(\mathbf{w}) - \eta \left(1 - \frac{\eta\beta}{2} \right) \|\nabla f(\mathbf{w})\|^2.$$

Furthermore, for any $\mathbf{w}_\star \in \operatorname{argmin}_{\mathbf{w}} f(\mathbf{w})$, it holds that:

$$\|\nabla f(\mathbf{w})\|^2 \leq 2\beta \left(f(\mathbf{w}) - f(\mathbf{w}_\star) \right).$$

Proof. Observe, by β-smoothness:

$$\begin{aligned}
f(\mathbf{w}^+) &\leq f(\mathbf{w}) + \nabla f(\mathbf{w}) \cdot (\mathbf{w}^+ - \mathbf{w}) + \frac{\beta}{2} \|\mathbf{w}^+ - \mathbf{w}\|^2 \\
&= f(\mathbf{w}) - \eta \nabla f(\mathbf{w}) \cdot \nabla f(\mathbf{w}) + \frac{\beta}{2} \eta^2 \|\nabla f(\mathbf{w})\|^2 \\
&= f(\mathbf{w}) - \eta \left(1 - \frac{\eta\beta}{2} \right) \|\nabla f(\mathbf{w})\|^2,
\end{aligned}$$

which proves the first claim. For the second claim, apply the above inequality with $\eta = 1/\beta$, which gives

$$\begin{aligned}
f(\mathbf{w}^+) &\leq f(\mathbf{w}) - \frac{1}{2\beta} \|\nabla f(\mathbf{w})\|^2 \\
\iff \|\nabla f(\mathbf{w})\|^2 &\leq 2\beta \left(f(\mathbf{w}) - f(\mathbf{w}^+) \right).
\end{aligned}$$

The second claim now follows by using the fact that $f(\mathbf{w}_\star) \leq f(\mathbf{w}^+)$. ∎

E.1 Proofs for With Replacement Orderings

As discussed in the main text, our results hold for a wider range of step sizes compared to the classical SGD bounds in the smooth realizable setting. This is enabled due to the following lemma.

Lemma E.3. *Assume that $f(\mathbf{w}) = \frac{1}{2}\left\|\mathbf{A}\mathbf{w}-\mathbf{b}\right\|^2$ for some matrix $\mathbf{A}$ and vector $\mathbf{b}$, and let $\mathbf{w}_\star \in \mathbb{R}^d$ be such that $f(\mathbf{w}_\star) = 0$. Then, we have:*

$$2f(\mathbf{w}) = \nabla f(\mathbf{w})^\top(\mathbf{w}-\mathbf{w}_\star)\,,$$

and for any $\mathbf{z} \in \mathbb{R}^d$ and $\gamma > 0$:

$$(2-\gamma)f(\mathbf{w}) - \frac{1}{\gamma}f(\mathbf{z}) \leq \nabla f(\mathbf{w})^\top(\mathbf{w}-\mathbf{z})\,.$$

Proof. For any $\mathbf{w} \in \mathbb{R}^d$, since $\mathbf{A}\mathbf{w}_\star = \mathbf{b}$ and $f(\mathbf{w}) = \frac{1}{2}\|\mathbf{A}(\mathbf{w}-\mathbf{w}_\star)\|^2$, we have:

$$\begin{aligned}
\nabla f(\mathbf{w})^\top(\mathbf{w}-\mathbf{z}) &= \left\langle \mathbf{A}^\top\mathbf{A}(\mathbf{w}-\mathbf{w}_\star), \mathbf{w}-\mathbf{z}\right\rangle \\
&= \left\langle \mathbf{A}^\top\mathbf{A}(\mathbf{w}-\mathbf{w}_\star), \mathbf{w}-\mathbf{w}_\star\right\rangle - \left\langle \mathbf{A}^\top\mathbf{A}(\mathbf{w}-\mathbf{w}_\star), \mathbf{z}-\mathbf{w}_\star\right\rangle \\
&= \left\langle \mathbf{A}\mathbf{w}-\mathbf{b}, \mathbf{A}\mathbf{w}-\mathbf{b}\right\rangle - \left\langle \mathbf{A}\mathbf{w}-\mathbf{b}, \mathbf{A}\mathbf{z}-\mathbf{b}\right\rangle \\
&= 2f(\mathbf{w}) - \left\langle \mathbf{A}\mathbf{w}-\mathbf{b}, \mathbf{A}\mathbf{z}-\mathbf{b}\right\rangle.
\end{aligned}$$

Plugging in $\mathbf{z} = \mathbf{w}_\star$, the second term vanishes (since $\mathbf{A}\mathbf{w}_\star - \mathbf{b} = \mathbf{b}-\mathbf{b} = \mathbf{0}$) and the first claim follows. For the second claim, note that by Young's inequality:

$$\begin{aligned}
\nabla f(\mathbf{w})^\top(\mathbf{w}-\mathbf{z}) &= 2f(\mathbf{w}) - \left\langle \mathbf{A}\mathbf{w}-\mathbf{b}, \mathbf{A}\mathbf{z}-\mathbf{b}\right\rangle \\
&\geq 2f(\mathbf{w}) - \frac{\gamma}{2}\left\|\mathbf{A}\mathbf{w}-\mathbf{b}\right\|^2 - \frac{1}{2\gamma}\left\|\mathbf{A}\mathbf{z}-\mathbf{b}\right\|^2 = (2-\gamma)f(\mathbf{w}) - \frac{1}{\gamma}f(\mathbf{z})\,.
\end{aligned}$$

■

Recall Lemma 5.2. Consider the β-smooth, realizable Setup 1, and let $T \geq 1$, $(i_0, \ldots, i_T) \in \mathcal{I}^{T+1}$ be an arbitrary sequence of indices in $\mathcal{I}$, and $\mathbf{w}_0 \in \mathbb{R}^d$ be an arbitrary initialization. Then, the gradient descent iterates given by $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta\nabla f(\mathbf{w}_t; i_t)$ for a step size $\eta < 2/\beta$, hold:

$$\sum_{t=0}^{T} f\left(\mathbf{w}_t; i_t\right) \leq \frac{\left\|\mathbf{w}_0-\mathbf{w}_\star\right\|^2}{2\eta(2-\eta\beta)}\,.$$

Proof. Denote $f_t(\mathbf{w}) \triangleq f(\mathbf{w}; i_t)$, and observe by Lemma E.1;

$$\begin{aligned}
\sum_{t=0}^{T}\langle\nabla f_t(\mathbf{w}_t), \mathbf{w}_t-\mathbf{w}_\star\rangle &\leq \frac{\left\|\mathbf{w}_0-\mathbf{w}_\star\right\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=0}^{T}\left\|\nabla f_t(\mathbf{w}_t)\right\|^2 \\
&\leq \frac{\left\|\mathbf{w}_0-\mathbf{w}_\star\right\|^2}{2\eta} + \eta\beta\sum_{t=0}^{T} f_t(\mathbf{w}_t) - f_t(\mathbf{w}_\star) = \frac{\left\|\mathbf{w}_0-\mathbf{w}_\star\right\|^2}{2\eta} + \eta\beta\sum_{t=0}^{T} f_t(\mathbf{w}_t)\,,
\end{aligned}$$

where the second inequality follows from Lemma E.2. On the other hand, by Lemma E.3,

$$\sum_{t=0}^{T}\langle\nabla f_t(\mathbf{w}_t), \mathbf{w}_t-\mathbf{w}_\star\rangle = \sum_{t=0}^{T} 2f_t(\mathbf{w}_t)\,.$$

Combining the two displays above, it follows that

$$(2-\eta\beta)\sum_{t=0}^{T} f_t(\mathbf{w}_t) \leq \frac{\left\|\mathbf{w}_0-\mathbf{w}_\star\right\|^2}{2\eta}\,,$$

and the result follows after dividing by $(2-\eta\beta)$. ■

Recall Lemma 5.3. Consider the β-smooth, realizable Setup 1. Let $T \geq 1$. Assume $\mathcal{P}$ is a distribution over $\mathcal{I}^{T+1}$ such that for every $0 \leq t \leq \tau_1 \leq \tau_2 \leq T$, the following holds: For any $i_0, \ldots i_{t-1} \in \mathcal{I}^t, i \in \mathcal{I}$, $\Pr(i_{\tau_1} = i | i_0, \ldots, i_{t-1}) = \Pr(i_{\tau_2} = i | i_0, \ldots, i_{t-1})$. Then, for any initialization $\mathbf{w}_0 \in \mathbb{R}^d$, with-replacement SGD (Eq. (2)) with step-size $\eta < 2/\beta$, holds:

$$\mathbb{E} f(\mathbf{w}_T, i_T) \leq (eT)^{\eta\beta(1-\eta\beta/4)} \mathbb{E}\left[\tfrac{1}{T+1}\sum\nolimits_{t=0}^{T} f(\mathbf{w}_t; i_t)\right],$$

where the expectation is taken with respect to $i_0, \ldots, i_T$ sampled from $\mathcal{P}$.

Proof. Denote $f_t(\mathbf{w}) \triangleq f(\mathbf{w}; i_t)$, $\mathbf{g}_t \triangleq \nabla f_t(\mathbf{w}_t)$, and observe that by Lemma E.1, $\forall \mathbf{z} \in \mathbb{R}^d$, $t \leq T$ (w.p. 1):

$$\begin{aligned}
\sum_{t=T-k}^{T} \langle \mathbf{g}_t, \mathbf{w}_t - \mathbf{z}\rangle &\leq \frac{\|\mathbf{w}_{T-k} - \mathbf{z}\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=T-k}^{T} \|\mathbf{g}_t\|^2 \\
\left[\text{Descent Lemma E.2}\right] &\leq \frac{\|\mathbf{w}_{T-k} - \mathbf{z}\|^2}{2\eta} + \eta\beta \sum_{t=T-k}^{T} f_t(\mathbf{w}_t) - f_t(\mathbf{w}_\star) \\
&= \frac{\|\mathbf{w}_{T-k} - \mathbf{z}\|^2}{2\eta} + \eta\beta \sum_{t=T-k}^{T} f_t(\mathbf{w}_t) - f_t(\mathbf{z}) + f_t(\mathbf{z}) - f_t(\mathbf{w}_\star)\,.
\end{aligned}$$

By Lemma E.3, this implies for any $\gamma > 0$:

$$\begin{aligned}
&\sum_{t=T-k}^{T} (2 - \gamma - \eta\beta) f_t(\mathbf{w}_t) - \left(\frac{1}{\gamma} - \eta\beta\right) f_t(\mathbf{z}) \\
&= \sum_{t=T-k}^{T} \left((2-\gamma) f_t(\mathbf{w}_t) - \frac{1}{\gamma} f_t(\mathbf{z})\right) + \eta\beta \sum_{t=T-k}^{T} f_t(\mathbf{z}) - f_t(\mathbf{w}_t) \\
\left[\text{Lemma E.3}\right] &\leq \sum_{t=T-k}^{T} \langle \mathbf{g}_t, \mathbf{w}_t - \mathbf{z}\rangle + \eta\beta \sum_{t=T-k}^{T} f_t(\mathbf{z}) - f_t(\mathbf{w}_t) \\
\left[\text{above}\right] &\leq \frac{\|\mathbf{w}_{T-k} - \mathbf{z}\|^2}{2\eta} + \eta\beta \sum_{t=T-k}^{T} f_t(\mathbf{z}) - \underbrace{f_t(\mathbf{w}_\star)}_{=0} \\
\implies (2 - \gamma - \eta\beta) \sum_{t=T-k}^{T} f_t(\mathbf{w}_t) &\leq \frac{\|\mathbf{w}_{T-k} - \mathbf{z}\|^2}{2\eta} + \frac{1}{\gamma} \sum_{t=T-k}^{T} f_t(\mathbf{z})\,.
\end{aligned}$$

Now, set $\mathbf{z} = \mathbf{w}_{T-k}$ and take expectations to obtain:

$$\begin{aligned}
(2 - \gamma - \eta\beta) \sum_{t=T-k}^{T} \mathbb{E} f_t(\mathbf{w}_t) &\leq 0 + \frac{1}{\gamma} \sum_{t=T-k}^{T} \mathbb{E} f_t(\mathbf{w}_{T-k}) \\
\frac{1}{k+1} \sum_{t=T-k}^{T} \mathbb{E} f_t(\mathbf{w}_t) &\leq \frac{1}{(k+1)\gamma(2 - \gamma - \eta\beta)} \sum_{t=T-k}^{T} \mathbb{E} f_t(\mathbf{w}_{T-k})\,.
\end{aligned}$$

Defining $S_k \triangleq \frac{1}{k+1} \sum_{t=T-k}^{T} f_t(\mathbf{w}_t)$, implies that

$$(k+1) S_k - k S_{k-1} = \sum_{t=T-k}^{T} f_t(\mathbf{w}_t) - \sum_{t=T-k+1}^{T} f_t(\mathbf{w}_t) = f_{T-k}(\mathbf{w}_{T-k})\,,$$

and by the assumption on the distribution $\mathcal{P}$ it follows that $\mathbb{E} f_{T-k}(\mathbf{w}_{T-k}) = \mathbb{E} f_t(\mathbf{w}_{T-k})$ for any $t \geq T - k$.

Thus, combined with our previous display,

$$\begin{aligned}\mathbb{E}S_k &\le \frac{1}{(k+1)\gamma(2-\gamma-\eta\beta)}\sum_{t=T-k}^{T}\mathbb{E}f_t(\mathbf{w}_{T-k})\\ &= \frac{1}{(k+1)\gamma(2-\gamma-\eta\beta)}\sum_{t=T-k}^{T}\Big((k+1)\mathbb{E}S_k - k\mathbb{E}S_{k-1}\Big)\\ &= \frac{1}{\gamma(2-\gamma-\eta\beta)}\Big((k+1)\mathbb{E}S_k - k\mathbb{E}S_{k-1}\Big)\,.\end{aligned}$$

Rearranging, denoting $c \triangleq \gamma(2-\gamma-\eta\beta)$, and requiring $c \in (0,1)$, we get

$$\begin{aligned}\frac{k}{c}\mathbb{E}S_{k-1} &\le \left(\frac{k+1}{c}-1\right)\mathbb{E}S_k\\ \iff \mathbb{E}S_{k-1} &\le \frac{k+1-c}{k}\mathbb{E}S_k\\ \implies \mathbb{E}f_T(\mathbf{w}_T) = \mathbb{E}S_0 &\le \prod_{k=1}^{T}\left(1+\frac{1-c}{k}\right)\mathbb{E}S_T\\ {\scriptstyle\left[1+x\le e^x,\forall x\ge 0\right]} &\le \exp\left(\sum_{k=1}^{T}\frac{1-c}{k}\right)\mathbb{E}S_T\\ &= \exp\left((1-c)\sum_{k=1}^{T}\frac{1}{k}\right)\cdot\mathbb{E}S_T \le \exp\big((1-c)\left(1+\log T\right)\big)\mathbb{E}S_T\\ &= (eT)^{1-c}\cdot\mathbb{E}\left[\frac{1}{T+1}\sum_{t=0}^{T}f_t(\mathbf{w}_t)\right]\,. \qquad (6)\end{aligned}$$

Now, getting the "best" rate requires maximizing $c = \gamma(2-\gamma-\eta\beta)$. To this end, we choose $\gamma = 1-\frac{\eta\beta}{2}$, which implies $c = \left(1-\frac{\eta\beta}{2}\right)^2$ (under the $\eta < \frac{2}{\beta}$ condition, we now have both $\gamma > 0$ and $c \in (0,1)$ as required above). Then, $1-c = \eta\beta\left(1-\frac{\eta\beta}{4}\right)$, and we finally get the required

$$\mathbb{E}f_T(\mathbf{w}_T) = (eT)^{\eta\beta\left(1-\frac{\eta\beta}{4}\right)}\cdot\frac{1}{T+1}\sum_{t=0}^{T}f_t(\mathbf{w}_t)\,.$$

■

E.2 Extending the SGD Bounds to Without Replacement Orderings

Here, we extend Theorem 5.1 to a *without*-replacement setting. Specifically, we consider gradient descent under a random *permutation* of the T tasks. That is, for some initialization $\mathbf{w}_0 \in \mathbb{R}^d$, step size $\eta > 0$, and $\pi_t \sim \mathrm{Unif}(\mathcal{I})$ sampled without replacement,

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta \nabla f(\mathbf{w}_t; \pi_t)\,, \tag{7}$$

where $f(\mathbf{w}; i) \triangleq \frac{1}{2}\left\|\mathbf{A}_i\mathbf{w} - \mathbf{b}_i\right\|^2$ as defined in Setup 1. Our main result is given below.

Theorem E.4. *Last-Iterate Bound for Realizable Regression Without Replacement Consider the β-smooth, realizable Setup 1. Define for all $T \geq 2$, $\hat{f}_{0:T}(\mathbf{w}) \triangleq \frac{1}{T+1}\sum_{t=0}^{T} f(\mathbf{w}; \pi_t)$. Then, without-replacement SGD (Eq. (7)) with step-size $\eta < 2/\beta$, holds:*

$$\mathbb{E}_\pi \hat{f}_{0:T}(\mathbf{w}_T) \leq \frac{eD^2}{\eta(2-\eta\beta)T^{1-\eta\beta(1-\eta\beta/4)}} + \frac{4\beta^2\eta D^2}{T}\,, \quad \forall T = 2, \ldots, n-1\,,$$

where $D \triangleq \|\mathbf{w}_0 - \mathbf{w}_\star\|$. In particular, for $\eta = \frac{1}{\beta \log T}$ yields $\frac{14\beta D^2 \log T}{T}$ and $\eta = \frac{1}{\beta}$ yields $\frac{7\beta D^2}{\sqrt[4]{T}}$.

The proof, given next, is based on the algorithmic stability of SGD [6, 25, 61], and more specifically, on a variant of stability, suitable for without replacement sampling [35, 63].

The proof of our theorem follows by a combination of two lemmas. The first, stated below, establishes a bound on the expected "next sample" loss and follows immediately by combining Lemmas 5.2 and 5.3 (notice that $\eta < \frac{2}{\beta} \implies \exp\left(\eta\beta\left(1 - \frac{\eta\beta}{4}\right)\right) \mapsto \exp\left(z\left(1 - \frac{z}{4}\right)\right)$ for $z \in (0, 2)$, which is monotonic increasing and upper bounded by e).

Lemma E.5. *For any step-size $\eta < 2/\beta$ and initialization $\mathbf{w}_0 \in \mathbb{R}^d$, without-replacement SGD Eq.* (7) *satisfies, for all $1 \leq T \leq n-1$:*

$$\mathbb{E}_\pi f(\mathbf{w}_T; \pi_T) \leq e^{\eta\beta\left(1-\frac{\eta\beta}{4}\right)} T^{\eta\beta\left(1-\frac{\eta\beta}{4}\right)} \mathbb{E}_\pi\left[\frac{1}{T+1}\sum_{t=0}^{T} f(\mathbf{w}_t; \pi_t)\right] \leq \frac{e \cdot \left\|\mathbf{w}_0 - \mathbf{w}_\star\right\|^2}{2\eta(2-\eta\beta)T^{1-\eta\beta\left(1-\frac{\eta\beta}{4}\right)}}\,.$$

Next, we consider the "empirical loss" objective. Given any permutation $\pi \in \mathcal{I} \leftrightarrow \mathcal{I}$, define:

$$\hat{f}_{0:t}(\mathbf{w}) \triangleq \frac{1}{t+1}\sum_{i=0}^{t} f(\mathbf{w}; \pi_i).$$

In the without-replacement setup, our optimization objective is the expected empirical loss $\mathbb{E}_\pi \hat{f}_{0:t}(\mathbf{w})$, which, when $t = n$, satisfies $\mathbb{E}_\pi \hat{f}_{0:t}(\mathbf{w}) = \mathbb{E}_\pi \bar{f}(\mathbf{w})$. Our second lemma (given next) bounds the expected empirical loss w.r.t. the next sample loss. This is the crux of extending our with-replacement upper bound to the without-replacement setup.

Lemma E.6. *For without-replacement SGD Eq.* (7) *with step size $\eta \leq 2/\beta$, for all $1 \leq T \leq n$, we have that the following holds:*

$$\mathbb{E}_\pi \hat{f}_{0:T}(\mathbf{w}_T) \leq 2\mathbb{E}_\pi f(\mathbf{w}_T; \pi_T) + \frac{4\beta^2\eta\left\|\mathbf{w}_0 - \mathbf{w}_\star\right\|^2}{T+1}.$$

The proof of Lemma E.6 builds on an algorithmic stability argument similar to that given in Lei and Ying [38], combined with the without-replacement stability framework proposed by Sherman et al. [63]. Before turning to the proof given in the next subsection, we quickly prove Theorem E.4.

Proof of Theorem E.4. By Lemmas E.5 and E.6,

$$\mathbb{E}_\pi \hat{f}_{0:T}(\mathbf{w}_T) \leq 2\mathbb{E}_\pi f(\mathbf{w}_T; \pi_T) + \tfrac{4\beta^2\eta\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{T+1} \leq \tfrac{e \cdot \|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{\eta(2-\eta\beta)} T^{\eta\beta\left(1-\frac{\eta\beta}{4}\right)-1} + \tfrac{4\beta^2\eta\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{T+1}.$$

The result for $\eta = \frac{1}{\beta}$ is straightforward. To see the result for $\eta = \frac{1}{\beta \log T}$, notice that in this case,

$$\tfrac{eD^2 T^{\eta\beta(1-\eta\beta/4)-1}}{\eta(2-\eta\beta)} = \frac{e\beta D^2 \log T}{T\left(2 - \frac{1}{\log T}\right)} T^{\frac{1}{\log T}\left(1 - \frac{1}{4\log T}\right)} = \tfrac{\beta D^2 \log T}{T} \frac{\exp\left(2 - \frac{1}{4\log T}\right)}{2 - \frac{1}{\log T}} \leq \frac{10\beta D^2 \log T}{T}.$$

■

E.2.1 Proving Lemma E.6

Notation. We first add a few definitions central to our analysis. Given a permutation $\pi \in \mathcal{I} \leftrightarrow \mathcal{I}$, denote:

$$\pi(j \leftrightarrow k) \triangleq \pi \text{ after swapping the } j^{\text{th}} \text{ and } k^{\text{th}} \text{ coordinates},$$
$$\mathbf{w}_\tau^\pi \triangleq \text{The iterate of SGD on step } \tau \text{ when run on permutation } \pi.$$

Most commonly, we will use the following special case of the above:

$$\mathbf{w}_\tau^{\pi(i\leftrightarrow t)} \triangleq \text{The iterate of SGD on step } \tau \text{ when run on } \pi(i \leftrightarrow t).$$

When clear from context, we omit π from the superscript and simply write $\mathbf{w}_\tau^{(i\leftrightarrow t)}$. Concretely, these definitions imply $\mathbf{w}_0^{(i\leftrightarrow t)} \triangleq \mathbf{w}_0$, and $\forall i, t, \tau \in \mathcal{I}$,

$$\mathbf{w}_{\tau+1}^{(i\leftrightarrow t)} = \mathbf{w}_\tau^{(i\leftrightarrow t)} - \eta \nabla f\left(\mathbf{w}_\tau^{(i\leftrightarrow t)}; \pi(i \leftrightarrow t)_\tau\right).$$

We have the following important relation, to be used later in the proof.

Lemma E.7. *For all $i, t, \tau \in \mathcal{I}, i \leq \tau \leq t$, we have:*

$$\mathbb{E}_\pi f(\mathbf{w}_\tau; \pi_i) = \mathbb{E}_\pi f(\mathbf{w}_\tau^{(i\leftrightarrow t)}; \pi(i \leftrightarrow t)_i).$$

Proof. The proof follows from observing that the random variables $f(\mathbf{w}_\tau; \pi_i)$ and $f(\mathbf{w}_\tau^{(i\leftrightarrow t)}; \pi(i \leftrightarrow t)_i)$ are distributed identically (the indices π_i, π_t are exchangeable). Formally, let $\Pi(\mathcal{I}) \triangleq \{\pi \in \mathcal{I} \leftrightarrow \mathcal{I}\}$ be the set of all permutations over $\mathcal{I}$, and observe

$$\mathbb{E}_\pi f(\mathbf{w}_\tau^{(i\leftrightarrow t)}; \pi(i \leftrightarrow t)_i) = \frac{1}{|\Pi(\mathcal{I})|} \sum_{\pi \in \Pi(\mathcal{I})} f(\mathbf{w}_\tau^{\pi(i\leftrightarrow t)}; \pi(i \leftrightarrow t)_i).$$

On the other hand,

$$\mathbb{E}_\pi f(\mathbf{w}_\tau, \pi_i) = \frac{1}{|\Pi(\mathcal{I})|} \sum_{\pi \in \Pi(\mathcal{I})} f(\mathbf{w}_\tau^\pi; \pi_i).$$

Hence, since there is a one-to-one correspondence between π and $\pi(\tau \leftrightarrow i)$, in particular,

$$\{\pi \mid \pi \in \Pi(\mathcal{I})\} = \{\pi(i \leftrightarrow t) \mid \pi \in \Pi(\mathcal{I})\},$$

the result follows. ■

Our next lemma, originally given in Sherman et al. [63, Lemma 2 therein], can be thought of as a without-replacement version of the well known stability $\iff$ generalization argument of the with-replacement sampling case [25, 61].

Lemma E.8. *The iterates of without-replacement SGD Eq.* (7)*, satisfy for all t:*

$$\mathbb{E}_\pi\left[f(\mathbf{w}_t;\pi_t)-\hat{f}_{0:t-1}(\mathbf{w}_t)\right]=\frac{1}{t}\sum_{i=0}^{t-1}\mathbb{E}_\pi\left[f(\mathbf{w}_t;\pi_t)-f(\mathbf{w}_t^{(i\leftrightarrow t)};\pi_t)\right]$$

Proof. We have, by definition of $\hat{f}_{0:t-1}$ and Lemma E.7:

$$\begin{aligned}\mathbb{E}_\pi\left[\hat{f}_{0:t-1}(\mathbf{w}_t)\right]&=\frac{1}{t}\sum_{i=0}^{t-1}\mathbb{E}_\pi\left[f(\mathbf{w}_t;\pi_i)\right]\\&=\frac{1}{t}\sum_{i=0}^{t-1}\mathbb{E}_\pi\left[f(\mathbf{w}_t^{(i\leftrightarrow t)};\pi(i\leftrightarrow t)_i)\right]=\frac{1}{t}\sum_{i=0}^{t-1}\mathbb{E}_\pi\left[f(\mathbf{w}_t^{(i\leftrightarrow t)};\pi_t)\right],\end{aligned}$$

where the last equality is immediate since by definition, $\pi(i\leftrightarrow t)_i=\pi_t$. The claim now follows by linearity of expectation. ∎

We are now ready to prove our main lemma. We note that the proof shares some features with that of the with-replacement case (Lemma F.2).

Proof of Lemma E.6. We prove the theorem for every t. Any β-smooth realizable function $h:\mathbb{R}^d\to\mathbb{R}_{\geq 0}$ holds that

$$\begin{aligned}|h(\tilde{\mathbf{w}})-h(\mathbf{w})|&\leq\left|\nabla h(\mathbf{w})^\top(\tilde{\mathbf{w}}-\mathbf{w})\right|+\frac{\beta}{2}\|\tilde{\mathbf{w}}-\mathbf{w}\|^2\\ \left[\text{Young's ineq.}\right]&\leq\frac{1}{2\beta}\|\nabla h(\mathbf{w})\|^2+\frac{\beta}{2}\|\tilde{\mathbf{w}}-\mathbf{w}\|^2+\frac{\beta}{2}\|\tilde{\mathbf{w}}-\mathbf{w}\|^2\\&\leq h(\mathbf{w})+\beta\|\tilde{\mathbf{w}}-\mathbf{w}\|^2\,.\end{aligned}\tag{8}$$

Hence, by Lemma E.8,

$$\begin{aligned}\left|\mathbb{E}_\pi\left[f(\mathbf{w}_t;\pi_t)-\hat{f}_{0:t-1}(\mathbf{w}_t)\right]\right|&=\left|\frac{1}{t}\sum_{i=0}^{t-1}\mathbb{E}_\pi\left[f(\mathbf{w}_t;\pi_t)-f(\mathbf{w}_t^{(i\leftrightarrow t)};\pi_t)\right]\right|\\ [\text{Jensen}]&\leq\frac{1}{t}\sum_{i=0}^{t-1}\mathbb{E}_\pi\left|f(\mathbf{w}_t;\pi_t)-f(\mathbf{w}_t^{(i\leftrightarrow t)};\pi_t)\right|\\ [\text{Eq. (8)}]&\leq\frac{1}{t}\sum_{i=0}^{t-1}\mathbb{E}_\pi\left[f(\mathbf{w}_t;\pi_t)+\beta\left\|\mathbf{w}_t^{(i\leftrightarrow t)}-\mathbf{w}_t\right\|^2\right]\\&=\mathbb{E}_\pi f(\mathbf{w}_t,\pi_t)+\frac{\beta}{t}\sum_{i=0}^{t-1}\mathbb{E}_\pi\left\|\mathbf{w}_t^{(i\leftrightarrow t)}-\mathbf{w}_t\right\|^2.\end{aligned}\tag{9}$$

Next, we bound $\left\|\mathbf{w}_t^{(i\leftrightarrow t)}-\mathbf{w}_t\right\|^2$. For any $0\leq\tau\leq t-1$, we denote $f_\tau\triangleq f(\cdot;\pi_\tau)$, and $f_\tau^{(i\leftrightarrow t)}\triangleq f(\cdot;\pi(i\leftrightarrow t)_\tau)$. Observe that for any τ such that $\tau\neq i$, we have $f_\tau=f_\tau^{(i\leftrightarrow t)}$, thus, by the non-expansiveness of gradient steps in the convex and β-smooth regime when $\eta\leq 2/\beta$ (see Lemma 3.6 in 25):

$$\begin{aligned}\tau\leq i&\implies\left\|\mathbf{w}_\tau^{(i\leftrightarrow t)}-\mathbf{w}_\tau\right\|=0,\\ i<\tau&\implies\left\|\mathbf{w}_{\tau+1}^{(i\leftrightarrow t)}-\mathbf{w}_{\tau+1}\right\|^2\leq\left\|\mathbf{w}_{i+1}^{(i\leftrightarrow t)}-\mathbf{w}_{i+1}\right\|^2.\end{aligned}$$

Further,

$$\begin{aligned}
\left\|\mathbf{w}_{i+1}^{(i\leftrightarrow t)}-\mathbf{w}_{i+1}\right\|^2 &= \left\|\mathbf{w}_{i}^{(i\leftrightarrow t)}-\eta\nabla f_i^{(i\leftrightarrow t)}(\mathbf{w}_{i}^{(i\leftrightarrow t)})-(\mathbf{w}_i-\eta\nabla f_i(\mathbf{w}_i))\right\|^2 \\
\left[\mathbf{w}_{i}^{(i\leftrightarrow t)}=\mathbf{w}_i\right] &= \eta^2\left\|\nabla f_i^{(i\leftrightarrow t)}(\mathbf{w}_{i}^{(i\leftrightarrow t)})-\nabla f_i(\mathbf{w}_i)\right\|^2 \\
[\text{Jensen}] &\le 2\eta^2\left\|\nabla f_i^{(i\leftrightarrow t)}(\mathbf{w}_{i}^{(i\leftrightarrow t)})\right\|^2+2\eta^2\left\|\nabla f_i(\mathbf{w}_i)\right\|^2 \\
&\le 4\beta\eta^2 f_i^{(i\leftrightarrow t)}(\mathbf{w}_{i}^{(i\leftrightarrow t)})+4\beta\eta^2 f_i(\mathbf{w}_i)\,,
\end{aligned}$$

and by Lemma E.7 $\mathbb{E}f_i(\mathbf{w}_i)=\mathbb{E}f_i^{(i\leftrightarrow t)}(\mathbf{w}_i^{(i\leftrightarrow t)})$. Hence,

$$\mathbb{E}\left\|\mathbf{w}_{t}^{(i\leftrightarrow t)}-\mathbf{w}_{t}\right\|^2\le\mathbb{E}\left\|\mathbf{w}_{i+1}^{(i\leftrightarrow t)}-\mathbf{w}_{i+1}\right\|^2\le 8\beta\eta^2\mathbb{E}f_i(\mathbf{w}_i)\,.$$

Now,

$$\frac{\beta}{t}\sum_{i=0}^{t-1}\mathbb{E}_\pi\left\|\mathbf{w}_{t}^{(i\leftrightarrow t)}-\mathbf{w}_{t}\right\|^2\le\left(8\beta^2\eta^2\right)\mathbb{E}\left[\frac{1}{t}\sum_{i=0}^{t-1}f_i(\mathbf{w}_i)\right],$$

which, when combined with Eq. (9) yields:

$$\left|\mathbb{E}_\pi\left[f(\mathbf{w}_t;\pi_t)-\hat{f}_{0:t-1}(\mathbf{w}_t)\right]\right|\le\mathbb{E}_\pi f(\mathbf{w}_t;\pi_t)+\left(8\beta^2\eta^2\right)\mathbb{E}\left[\frac{1}{t}\sum_{i=0}^{t-1}f_i(\mathbf{w}_i)\right].$$

Finally, by the regret bound given in Lemma 5.2, $\sum_{i=0}^{t-1}f_i(\mathbf{w}_i)\le\frac{\|\mathbf{w}_0-\mathbf{w}_\star\|^2}{2\eta(2-\eta\beta)}$, and therefore,

$$\begin{aligned}
&\left|\mathbb{E}_\pi\left[f(\mathbf{w}_t;\pi_t)-\hat{f}_{0:t-1}(\mathbf{w}_t)\right]\right|\le\mathbb{E}_\pi f(\mathbf{w}_t;\pi_t)+\frac{4\beta^2\eta\left\|\mathbf{w}_0-\mathbf{w}_\star\right\|^2}{(2-\eta\beta)t} \\
&\qquad\implies\mathbb{E}\hat{f}_{0:t-1}(\mathbf{w}_t)\le 2\mathbb{E}_\pi f(\mathbf{w}_t;\pi_t)+\frac{4\beta^2\eta\left\|\mathbf{w}_0-\mathbf{w}_\star\right\|^2}{(2-\eta\beta)t}.
\end{aligned}$$

Finally, since $\hat{f}_{0:t}=\frac{t}{t+1}\hat{f}_{0:t-1}+\frac{1}{t+1}f_t$, we obtain

$$\mathbb{E}\hat{f}_{0:t}(\mathbf{w}_t)=\frac{t}{t+1}\mathbb{E}\hat{f}_{0:t-1}(\mathbf{w}_t)+\frac{1}{t+1}\mathbb{E}f_t(\mathbf{w}_t)\le\frac{2t+1}{t+1}\mathbb{E}_\pi f(\mathbf{w}_t;\pi_t)+\frac{4\beta^2\eta\left\|\mathbf{w}_0-\mathbf{w}_\star\right\|^2}{(2-\eta\beta)(t+1)},$$

which completes the proof. ■

F Supplementary Material for the Extension Section (Section 6)

Table 3: **Forgetting Rates in Weakly-Regularized Continual Linear Classification on Separable Data.** In *all* cells, we omit mild constant multiplicative factors and normalize by an unavoidable $\|\mathbf{w}_\star\|^2 R^2$ term.

Paper / Ordering	Random with Replacement	Random w/o Replacement	Cyclic
Evron et al. [16]	$\exp\left(-\frac{k}{4T\|\mathbf{w}_\star\|^2 R^2}\right)$	—	$\frac{T^2}{\sqrt{k}} \wedge \exp\left(-\frac{k}{16T^2\|\mathbf{w}_\star\|^2 R^2}\right)$
Ours (2025)	$\frac{1}{\sqrt[4]{k}}$	$\frac{1}{\sqrt[4]{T}}$	—

Recall Reduction 3. Consider T arbitrary (nonempty) closed convex sets $\mathcal{C}_1, \ldots, \mathcal{C}_T$, initial point $\mathbf{w}_0 \in \mathbb{R}^d$, and ordering τ. Define $f_m(\mathbf{w}) = \frac{1}{2}\|\mathbf{w} - \mathbf{\Pi}_m(\mathbf{w})\|^2, \forall m \in [T]$. Then,

(i) f_m is convex and 1-smooth.

(ii) The POCS update is equivalent to an SGD step: $\mathbf{w}_t = \mathbf{\Pi}_{\tau(t)}(\mathbf{w}_{t-1}) = \mathbf{w}_{t-1} - \nabla_{\mathbf{w}} f_{\tau(t)}(\mathbf{w}_{t-1})$.

Proof. First, by Theorem 1.5.5 in Facchinei and Pang [17], f_m is continuously differentiable and for every $\mathbf{w} \in \mathbb{R}^d, m \in [T]$, $\nabla f_m(\mathbf{w}) = \mathbf{w} - \mathbf{\Pi}_m(\mathbf{w})$. Plugging in $\nabla f_{\tau(t)}(\mathbf{w}_{t-1})$ into an appropriate SGD step, we get

$$\mathbf{w}_t = \mathbf{w}_{t-1} - \nabla_{\mathbf{w}} f_{\tau(t)}(\mathbf{w}_{t-1}) = \mathbf{w}_{t-1} - \big(\mathbf{w}_{t-1} - \mathbf{\Pi}_{\tau(t)}(\mathbf{w}_{t-1})\big) = \mathbf{\Pi}_{\tau(t)}(\mathbf{w}_{t-1}),$$

and the second part of the lemma follows. In addition, $\forall \mathbf{x}, \mathbf{w} \in \mathbb{R}^d$, we prove convexity by using a projection inequality (also from Theorem 1.5.5 in 17). That is,

$$\begin{aligned}
& f_m(\mathbf{x}) - f_m(\mathbf{w}) - \langle \nabla f_m(\mathbf{w}), \mathbf{x} - \mathbf{w} \rangle \\
&= \frac{1}{2}\|\mathbf{x} - \mathbf{\Pi}_m(\mathbf{x})\|^2 - \frac{1}{2}\|\mathbf{w} - \mathbf{\Pi}_m(\mathbf{w})\|^2 - \langle \mathbf{w} - \mathbf{\Pi}_m(\mathbf{w}), \mathbf{x} - \mathbf{w} \rangle \\
&= \frac{1}{2}\|\mathbf{x} - \mathbf{\Pi}_m(\mathbf{x})\|^2 - \frac{1}{2}\|\mathbf{w} - \mathbf{\Pi}_m(\mathbf{w})\|^2 - \langle \mathbf{w} - \mathbf{\Pi}_m(\mathbf{w}), \mathbf{x} - \mathbf{\Pi}_m(\mathbf{x}) \rangle \\
&\qquad + \langle \mathbf{w} - \mathbf{\Pi}_m(\mathbf{w}), \mathbf{\Pi}_m(\mathbf{w}) - \mathbf{\Pi}_m(\mathbf{x}) \rangle + \langle \mathbf{w} - \mathbf{\Pi}_m(\mathbf{w}), \mathbf{w} - \mathbf{\Pi}_m(\mathbf{w}) \rangle \\
&\geq \frac{1}{2}\|\mathbf{x} - \mathbf{\Pi}_m(\mathbf{x})\|^2 - \frac{1}{2}\|\mathbf{w} - \mathbf{\Pi}_m(\mathbf{w})\|^2 - \langle \mathbf{w} - \mathbf{\Pi}_m(\mathbf{w}), \mathbf{x} - \mathbf{\Pi}_m(\mathbf{x}) \rangle + 0 + \|\mathbf{w} - \mathbf{\Pi}_m(\mathbf{w})\|^2 \\
&= \frac{1}{2}\|\mathbf{x} - \mathbf{\Pi}_m(\mathbf{x}) - \mathbf{w} + \mathbf{\Pi}_m(\mathbf{w})\|^2 \geq 0\,.
\end{aligned}$$

For the 1-smoothness,

$$\begin{aligned}
\|\nabla f_m(\mathbf{x}) - \nabla f_m(\mathbf{w})\| &= \|\mathbf{x} - \mathbf{\Pi}_m(\mathbf{x}) - (\mathbf{w} - \mathbf{\Pi}_m(\mathbf{w}))\| \\
&= \|(\mathbf{I} - \mathbf{\Pi}_m)(\mathbf{x}) - (\mathbf{I} - \mathbf{\Pi}_m)(\mathbf{w})\| \leq \|\mathbf{x} - \mathbf{w}\|\ ,
\end{aligned}$$

where we used the non-expansiveness of $\mathbf{I} - \mathbf{\Pi}_m$ [Propositions 4.2, 4.8 in 3]. ■

Lemma F.1. *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a nonempty closed and convex set, and $f(\mathbf{w}) = \frac{1}{2}\left\|\mathbf{w} - \mathbf{\Pi}_{\mathcal{K}}(\mathbf{w})\right\|^2$. Then, we have for any $\mathbf{z} \in \mathbb{R}^d$ and $\gamma > 0$*

$$(2-\gamma)f(\mathbf{w}) - \frac{1}{\gamma}f(\mathbf{z}) \leq \nabla f(\mathbf{w})^\top(\mathbf{w} - \mathbf{z})\,.$$

In addition, for any $\mathbf{u} \in \mathcal{K}$ we have

$$2f(\mathbf{w}) \leq \nabla f(\mathbf{w})^\top(\mathbf{w} - \mathbf{u})\,.$$

Proof. We have $\nabla f(\mathbf{w}) = \mathbf{w} - \mathbf{\Pi}_{\mathcal{K}}(\mathbf{w})$. Hence, by Theorem 1.5.5 in Facchinei and Pang [17],

$$\begin{aligned}
\langle \nabla f(\mathbf{w}), \mathbf{w} - \mathbf{z}\rangle &= \langle \mathbf{w} - \mathbf{\Pi}_{\mathcal{K}}(\mathbf{w}), \mathbf{w} - \mathbf{z}\rangle \\
&= \langle \mathbf{w} - \mathbf{\Pi}_{\mathcal{K}}(\mathbf{w}), \mathbf{w} - \mathbf{\Pi}_{\mathcal{K}}(\mathbf{w})\rangle + \langle \mathbf{w} - \mathbf{\Pi}_{\mathcal{K}}(\mathbf{w}), \mathbf{\Pi}_{\mathcal{K}}(\mathbf{w}) - \mathbf{z}\rangle \\
&= 2f(\mathbf{w}) + \langle \mathbf{w} - \mathbf{\Pi}_{\mathcal{K}}(\mathbf{w}), \mathbf{\Pi}_{\mathcal{K}}(\mathbf{w}) - \mathbf{\Pi}_{\mathcal{K}}(\mathbf{z})\rangle - \langle \mathbf{w} - \mathbf{\Pi}_{\mathcal{K}}(\mathbf{w}), \mathbf{z} - \mathbf{\Pi}_{\mathcal{K}}(\mathbf{z})\rangle \\
&\geq 2f(\mathbf{w}) - \langle \mathbf{w} - \mathbf{\Pi}_{\mathcal{K}}(\mathbf{w}), \mathbf{z} - \mathbf{\Pi}_{\mathcal{K}}(\mathbf{z})\rangle\ .
\end{aligned}$$

Plugging in $\mathbf{z} = \mathbf{u}$, the second term vanishes (since $\mathbf{u} - \mathbf{\Pi}_{\mathcal{K}}(\mathbf{u}) = \mathbf{0}$) and the second claim follows.

For the first claim, note that by Young's inequality:

$$\begin{aligned}
\langle \nabla f(\mathbf{w}), \mathbf{w} - \mathbf{z}\rangle &= 2f(\mathbf{w}) - \langle \mathbf{w} - \mathbf{\Pi}_{\mathcal{K}}(\mathbf{w}), \mathbf{z} - \mathbf{\Pi}_{\mathcal{K}}(\mathbf{z})\rangle \\
&\geq 2f(\mathbf{w}) - \frac{\gamma}{2}\left\|\mathbf{w} - \mathbf{\Pi}_{\mathcal{K}}(\mathbf{w})\right\|^2 - \frac{1}{2\gamma}\left\|\mathbf{z} - \mathbf{\Pi}_{\mathcal{K}}(\mathbf{z})\right\|^2 \\
&= 2f(\mathbf{w}) - \gamma f(\mathbf{w}) - \frac{1}{\gamma}f(\mathbf{z})\,.
\end{aligned}$$

■

Recall Theorem 6.1. Consider the same conditions of Reduction 3 and assume a nonempty set intersection $\mathcal{C}_\star = \bigcap_{m=1}^{T} \mathcal{C}_m \neq \varnothing$. Then, under a random ordering with or without replacement, the expected "residual" of Scheme 4 after $\forall k \geq 1$ iterations (without replacement: $k \in [T]$) is bounded as,

$$\mathbb{E}_\tau\Big[\frac{1}{2T}\sum\nolimits_{m=1}^{T} \left\|\mathbf{w}_k - \mathbf{\Pi}_m(\mathbf{w}_k)\right\|^2\Big] = \mathbb{E}_\tau\Big[\frac{1}{2T}\sum\nolimits_{m=1}^{T} \operatorname{dist}^2(\mathbf{w}_k, \mathcal{C}_m)\Big] \leq \frac{7}{\sqrt[4]{k}} \min_{\mathbf{w}\in\mathcal{C}_\star} \left\|\mathbf{w}_0 - \mathbf{w}\right\|^2 .$$

Proof. The proof largely follows the same steps of Theorems 4.2 and 4.3. Let τ be any random ordering, $\mathbf{w}_0 \in \mathbb{R}^d$ an initialization, and $\mathbf{w}_1, \ldots, \mathbf{w}_k$ be the corresponding iterates produced by Scheme 4. By Reduction 3, these are exactly the (stochastic) gradient descent iterates produced when initializing at $\mathbf{w}_0$ and using a step size of $\eta = 1$, on the 1-smooth loss sequence $f_{\tau(1)}, \ldots, f_{\tau(k)}$ defined by:

$$f_m(\mathbf{w}) \triangleq \frac{1}{2}\left\|\mathbf{w} - \mathbf{\Pi}_m(\mathbf{w})\right\|^2 .$$

Proceeding, we denote the objective function:

$$\bar{f}(\mathbf{w}) \triangleq \mathbb{E}_{m\sim\text{Unif}([T])} f_m(\mathbf{w}) = \frac{1}{2T}\sum_{m=1}^{T} \left\|\mathbf{w} - \mathbf{\Pi}_m(\mathbf{w})\right\|^2 .$$

Now, for a **with-replacement** ordering τ, invoke Theorem 5.1, except we use Lemma F.1 in the proof instead of Lemma E.3, to obtain:

$$\mathbb{E}_\tau \bar{f}(\mathbf{w}_k) \leq \frac{e}{2\sqrt[4]{k}} \min_{\mathbf{w}\in\mathcal{C}_\star} \left\|\mathbf{w}_0 - \mathbf{w}\right\|^2 , \qquad (\tau \text{ with-replacement})$$

which completes the proof for the with-replacement case.

For a **without-replacement** ordering τ, invoke Theorem E.4 (with $\eta = 1/\beta$), except again we use Lemma F.1 in the proof instead of Lemma E.3, to obtain:

$$\mathbb{E}_\tau \hat{f}_{0:k-1}(\mathbf{w}_k) \triangleq \mathbb{E}_\tau\Big[\frac{1}{k}\sum_{t=0}^{k-1} f(\mathbf{w}_k)\Big] \leq \frac{7}{\sqrt[4]{k}} \min_{\mathbf{w}\in\mathcal{C}_\star} \left\|\mathbf{w}_0 - \mathbf{w}\right\|^2 . \qquad (\tau \text{ without-replacement})$$

Similarly, by Lemma E.5,

$$\mathbb{E}_\tau f_{\tau(k+1)}(\mathbf{w}_k) \triangleq \mathbb{E}_\tau \tfrac{1}{2}\left\|\mathbf{w}_k - \mathbf{\Pi}_{\tau(k+1)}(\mathbf{w}_k)\right\|^2 \leq \frac{e}{2\sqrt[4]{k}} \min_{\mathbf{w}\in\mathcal{C}_\star} \left\|\mathbf{w}_0 - \mathbf{w}\right\|^2 . \qquad (\tau \text{ without-replacement})$$

Combining the last two displays with Proposition B.2, we now obtain:

$$\begin{aligned}
\mathbb{E}_\tau \bar{f}(\mathbf{w}_k) &\triangleq \mathbb{E}_\tau\Big[\frac{1}{2T}\sum_{m=1}^{T} \left\|\mathbf{w}_k - \mathbf{\Pi}_m(\mathbf{w}_k)\right\|^2\Big] \qquad (\tau \text{ without-replacement}) \\
&= \frac{k}{T}\mathbb{E}_\tau \hat{f}_{0:k-1}(\mathbf{w}_k) + \frac{T-k}{2T}\mathbb{E}_\tau \left\|\mathbf{w}_k - \mathbf{\Pi}_{\tau(k+1)}(\mathbf{w}_k)\right\|^2 \\
&\leq \left(\frac{7k}{T} + \frac{\frac{e}{2}(T-k)}{T}\right)\frac{1}{\sqrt[4]{k}} \min_{\mathbf{w}\in\mathcal{C}_\star} \left\|\mathbf{w}_0 - \mathbf{w}\right\|^2 \leq \frac{7}{\sqrt[4]{k}} \min_{\mathbf{w}\in\mathcal{C}_\star} \left\|\mathbf{w}_0 - \mathbf{w}\right\|^2 ,
\end{aligned}$$

which proves the without-replacement case and thus completes the proof. ■

Recall Theorem 6.2. Under a random ordering, with or without replacement, over T jointly separable tasks, the expected forgetting of the weakly-regularized Scheme 5 (at $\lambda \to 0$) after $k \geq 1$ iterations (without replacement: $k \in [T]$) is bounded as

$$\mathbb{E}_\tau\big[F_\tau(k)\big] \leq \frac{7\left\|\mathbf{w}_\star\right\|^2 R^2}{\sqrt[4]{k}}, \quad \text{where } \mathbf{w}_\star \triangleq \min_{\mathbf{w}\in\mathcal{C}_1\cap\cdots\cap\mathcal{C}_T} \left\|\mathbf{w}_0 - \mathbf{w}\right\|^2 .$$

Proof. We adopt the same notation as used above:

$$f_m(\mathbf{w}) \triangleq \frac{1}{2}\left\|\mathbf{w} - \mathbf{\Pi}_m(\mathbf{w}))\right\|^2$$

$$\bar{f}(\mathbf{w}) \triangleq \mathbb{E}_{m\sim\text{Unif}([T])} f_m(\mathbf{w}) = \frac{1}{2T}\sum_{m=1}^{T}\left\|\mathbf{w} - \mathbf{\Pi}_m(\mathbf{w})\right\|^2 .$$

For τ sampled **with replacement**, by Lemma F.2 (given below) and the with-replacement result (inside the proof) of Theorem 6.1, we have

$$\begin{aligned}\mathbb{E}_\tau[F_\tau(k)] = \mathbb{E}\hat{f}_{0:k-1}(\mathbf{w}_k) &\leq 2\mathbb{E}\bar{f}(\mathbf{w}_k) + \frac{4\left\|\mathbf{w}_0 - \mathbf{w}_\star\right\|^2}{k} \\ &\leq \left(\frac{e}{\sqrt[4]{k}} + \frac{4}{k}\right)\left\|\mathbf{w}_0 - \mathbf{w}_\star\right\|^2 \leq \frac{7\left\|\mathbf{w}_0 - \mathbf{w}_\star\right\|^2}{\sqrt[4]{k}}.\end{aligned}$$

For τ sampled **without replacement**, as argued in Theorem 6.1, by Lemma E.5:

$$\mathbb{E}_\tau f_{\tau(k+1)}(\mathbf{w}_k) \leq \frac{\frac{e}{2}\left\|\mathbf{w}_0 - \mathbf{w}_\star\right\|^2}{\sqrt[4]{k}},$$

and thus by Lemma E.6,

$$\mathbb{E}_\tau[F_\tau(k)] = \mathbb{E}\hat{f}_{0:k-1}(\mathbf{w}_k) \leq \left(\frac{e}{\sqrt[4]{k}} + \frac{4}{k}\right)\left\|\mathbf{w}_0 - \mathbf{w}_\star\right\|^2 \leq \frac{7\left\|\mathbf{w}_0 - \mathbf{w}_\star\right\|^2}{\sqrt[4]{k}} .$$

which completes the proof. ■

Lemma F.2. *Consider with-replacement SGD Eq.* (2) *with step size $\eta \leq 2/\beta$, and define, for every $0 \leq T$, $\hat{f}_{0:T}(\mathbf{w}) \triangleq \frac{1}{T+1}\sum_{t=0}^{T} f(\mathbf{w}; i_t)$. For all $1 \leq T$, the following holds:*

$$\mathbb{E}\hat{f}_{0:T-1}(\mathbf{w}_T) \leq 2\mathbb{E}\bar{f}(\mathbf{w}_T) + \frac{4\beta^2\eta\left\|\mathbf{w}_0 - \mathbf{w}_\star\right\|^2}{T} .$$

Proof. Our proof here mostly follows the proof of Lemma E.6. Recall that from Eq. (8), any β-smooth realizable function $h : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ holds that $|h(\tilde{\mathbf{w}}) - h(\mathbf{w})| \leq h(\mathbf{w}) + \beta\left\|\tilde{\mathbf{w}} - \mathbf{w}\right\|^2$. Denote $f_t \triangleq f(\cdot; i_t)$ for all $t \in \{0, ..., T\}$. Now, by the standard stability $\iff$ generalization argument [25, 61], and denoting by $\mathbf{w}_\tau^{(i)}$ the SGD iterate after τ steps on the training set where the i^{th} example was resampled as j_i:

$$\begin{aligned}\left|\mathbb{E}\left[\bar{f}(\mathbf{w}_T) - \hat{f}_{0:T-1}(\mathbf{w}_T)\right]\right| &= \left|\frac{1}{T}\sum_{i=0}^{T-1}\mathbb{E}_{j_i\sim\mathcal{D}}\left[f(\mathbf{w}_T; j_i) - f(\mathbf{w}_T^{(i)}; j_i)\right]\right| \\ \text{[Jensen; Eq. (8)]} &\leq \frac{1}{T}\sum_{i=0}^{T-1}\mathbb{E}\left[f(\mathbf{w}_T; j_i) + \beta\left\|\mathbf{w}_T^{(i)} - \mathbf{w}_T\right\|^2\right] \\ &= \mathbb{E}\bar{f}(\mathbf{w}_T) + \frac{\beta}{T}\sum_{i=0}^{T-1}\mathbb{E}\left\|\mathbf{w}_T^{(i)} - \mathbf{w}_T\right\|^2 .\end{aligned}$$

Next, we bound $\left\|\mathbf{w}_T^{(i)} - \mathbf{w}_T\right\|^2$. By the non-expansiveness of gradient steps in the convex and β-smooth regime when $\eta \leq 2/\beta$ [see Lemma 3.6 in 25]:

$$\tau \leq i \implies \left\|\mathbf{w}_\tau^{(i)} - \mathbf{w}_\tau\right\| = 0,$$

$$i < \tau \implies \left\|\mathbf{w}_{\tau+1}^{(i)} - \mathbf{w}_{\tau+1}\right\|^2 \le \left\|\mathbf{w}_{i+1}^{(i)} - \mathbf{w}_{i+1}\right\|^2 .$$

874 Further,

$$\begin{aligned}
\left\|\mathbf{w}_{i+1}^{(i)} - \mathbf{w}_{i+1}\right\|^2 &= \left\|\mathbf{w}_{i}^{(i)} - \eta\nabla f_{j_i}(\mathbf{w}_i^{(i)}) - (\mathbf{w}_i - \eta\nabla f_i(\mathbf{w}_i))\right\|^2 \\
{\scriptstyle [\mathbf{w}_i^{(i)}=\mathbf{w}_i]} &= \eta^2 \left\|\nabla f_{j_i}(\mathbf{w}_i^{(i)}) - \nabla f_i(\mathbf{w}_i)\right\|^2 \\
{\scriptstyle [\text{Jensen}]} &\le 2\eta^2 \left\|\nabla f_{j_i}(\mathbf{w}_i^{(i)})\right\|^2 + 2\eta^2 \left\|\nabla f_i(\mathbf{w}_i)\right\|^2 \\
{\scriptstyle \left[\substack{\text{smoothness,} \\ \text{non-negativity}}\right]} &\le 4\beta\eta^2 f_{j_i}(\mathbf{w}_i^{(i)}) + 4\beta\eta^2 f_i(\mathbf{w}_i) .
\end{aligned}$$

875 Therefore,

$$\mathbb{E}\left\|\mathbf{w}_{T}^{(i)} - \mathbf{w}_{T}\right\|^2 \le \mathbb{E}\left\|\mathbf{w}_{i+1}^{(i)} - \mathbf{w}_{i+1}\right\|^2 \le 4\beta\eta^2\mathbb{E} f_{j_i}(\mathbf{w}_i^{(i)}) + 4\beta\eta^2\mathbb{E} f_i(\mathbf{w}_i) = 8\beta\eta^2\mathbb{E} f_i(\mathbf{w}_i) .$$

876 Now,

$$\frac{\beta}{T}\sum_{i=0}^{T-1}\mathbb{E}\left\|\mathbf{w}_{T}^{(i)} - \mathbf{w}_{T}\right\|^2 \le 12\beta^2\eta^2\,\mathbb{E}\left[\frac{1}{T}\sum_{i=0}^{T-1} f_i(\mathbf{w}_i)\right] .$$

877 Summarizing, we have shown that:

$$\begin{aligned}
\left|\mathbb{E}\left[\bar{f}(\mathbf{w}_T) - \hat{f}_{0:T-1}(\mathbf{w}_T)\right]\right| &\le \mathbb{E}\bar{f}(\mathbf{w}_T) + \frac{\beta}{T}\sum_{i=0}^{T-1}\mathbb{E}\left\|\mathbf{w}_{T}^{(i)} - \mathbf{w}_{T}\right\|^2 \\
&\le \mathbb{E}\bar{f}(\mathbf{w}_T) + 8\beta^2\eta^2\,\mathbb{E}\left[\frac{1}{T}\sum_{i=0}^{T-1} f_i(\mathbf{w}_i)\right] .
\end{aligned}$$

878 Finally, by the regret bound given in Lemma 5.2, *i.e.,* $\sum_{i=0}^{T-1} f_i(\mathbf{w}_i) \le \frac{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{2\eta(2-\eta\beta)}$, we have

$$\left|\mathbb{E}\left[\bar{f}(\mathbf{w}_T) - \hat{f}_{0:T-1}(\mathbf{w}_T)\right]\right| \le \mathbb{E}\bar{f}(\mathbf{w}_T) + \frac{4\beta^2\eta\left\|\mathbf{w}_0 - \mathbf{w}_\star\right\|^2}{(2-\eta\beta)T} .$$

879 and the result follows. ■