Proceedings of Machine Learning Research vol 313:1–50, 2026 37th International Conference on Algorithmic Learning Theory

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

Itay Evron[*] ITAY@EVRON.ME
Meta

Ran Levinstein[*] RANLEVINSTEIN@GMAIL.COM
Department of Computer Science, Technion

Matan Schliserman[*] SCHLISERMAN@MAIL.TAU.AC.IL
Uri Sherman[*] URISHERMAN@MAIL.TAU.AC.IL
Blavatnik School of Computer Science and AI, Tel Aviv University

Tomer Koren TKOREN@TAUEX.TAU.AC.IL
Blavatnik School of Computer Science and AI, Tel Aviv University, and Google Research

Daniel Soudry DANIEL.SOUDRY@GMAIL.COM
Department of Electrical and Computing Engineering, Technion

Nathan Srebro NATI@TTIC.EDU
Toyota Technological Institute at Chicago

Editors: Matus Telgarsky and Jonathan Ullman

Abstract

We study the common continual learning setup where an overparameterized model is sequentially fitted to a set of jointly realizable tasks. We analyze forgetting, defined as the 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 and leverage them to derive new results for continual learning. Focusing on random orderings over T tasks, we establish *universal* forgetting rates, whereas existing rates depend on problem dimensionality or complexity and become prohibitive in highly overparameterized regimes. 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 resulting rate $\mathcal{O}(\min(1/\sqrt[4]{T}, (d-\bar{r})/T))$ shows that randomization alone, without task repetition, prevents 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 extend to broader methods, such as block Kaczmarz and POCS, illuminating their loss convergence under i.i.d. and single-pass orderings.

Keywords: Continual learning, Lifelong learning, Last iterate, SGD, Forgetting, Task ordering

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 (Goldfarb et al., 2024), and is also related to trade-offs like the plasticity-stability dilemma (Mermillod et al., 2013). CL is becoming increasingly

[*] Equal contribution. Equal-contributing authors are listed in alphabetical order, followed by senior authors.

© 2026 I. Evron, R. Levinstein, M. Schliserman, U. Sherman, T. Koren, D. Soudry & N. Srebro.

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 (e.g., Evron et al., 2022, 2023; Kong et al., 2023; Jung et al., 2025; Cai and Diakonikolas, 2025) and empirically (Lesort et al., 2023; Hemati et al., 2024), that forgetting diminishes when tasks are ordered randomly or cyclically. Task 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 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. (2022) 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 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 (Kaczmarz, 1937; Evron et al., 2022). Given rank-1 tasks, each update of these methods is known to reduce to a normalized stochastic gradient step, fully minimizing the current task's least squares objective using a "stepwise-optimal" step size. Extending to general rank, we prove that learning an *entire* task in continual linear regression is equivalent to a *single* SGD step on a modified objective with a fixed, 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., Ge et al., 2019; Berthier et al., 2020; Zou et al., 2021; Wu et al., 2022). Specifically, prior results either hold only for the average iterate (e.g., Bach and Moulines, 2013) or small step sizes bounded away from the stepwise-optimal step size crucial for our continual setup (e.g., Varre et al., 2021). We overcome this challenge by carefully refining analysis techniques for SGD (Srebro et al., 2010b; Shamir and Zhang, 2013) 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 related Kaczmarz and NLMS methods.

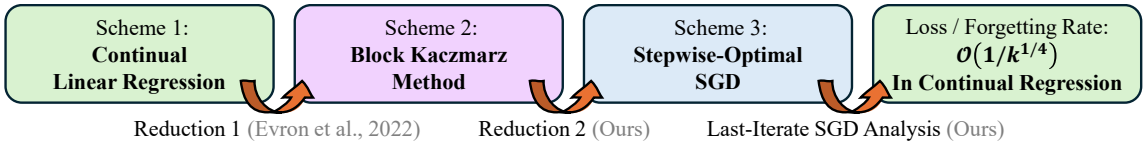

Figure 1: Analysis Flow Leading to Our Improved Regression Rates—From CL to SGD and Back.

1. While simple, continual linear regression captures key factors in CL, *e.g.,* task similarity (Hiratani, 2024; Tsipory et al., 2025), task recurrence (Evron et al., 2022), overparameterization (Goldfarb and Hand, 2023), and algorithmic effects (Doan et al., 2021; Peng et al., 2023). We follow prior work analyzing continual *optimization* dynamics under the assumption that training data across tasks are jointly realizable (Evron et al., 2022, 2023). In contrast, *statistical* formulations allow label noise but assume i.i.d. features (Lin et al., 2023; Banayeeanzade et al., 2025) or commutative covariances (Li et al., 2023; Zhao et al., 2024)—while our analysis applies to arbitrary data matrices.

Finally, by proving a matching rate for the squared loss of the broader Projection Onto Convex Sets framework (Gubin et al., 1967), we extend our results to continual linear *classification* on separable data, providing 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 new reductions from continual linear models to SGD with a rather large, "stepwise-optimal" step size, generalizing results from prior work—limited to rank-1 tasks—to arbitrary rank. This enables last-iterate analysis for studying forgetting.
- We provide novel last-iterate SGD analysis for a realizable least squares setup, yielding the first informative rates for fixed step sizes large enough to support the reductions to continual learning.
- Our main results are improved loss and forgetting rates in both continual linear regression and classification (see Tables 1 and 2, respectively), which (i) are dimensionality-independent and hold even in highly overparameterized regimes, previously uncovered by existing rates; and (ii) extend to without-replacement orderings, revealing that task repetition is not required to mitigate forgetting.

2. Main Setting: Continual Linear Regression

We focus primarily on the fundamental continual linear regression setting, widely-studied in theoretical work. This setting is easy-to-analyze, yet often sheds light on important CL phenomena.[1]

Notation. Boldface denotes vectors and matrices. We use $\|\cdot\|$ for Euclidean, spectral, or operator norms. $\mathbf{X}^+$ denotes the Moore–Penrose inverse. Finally, we define $[n] \triangleq \{1, \ldots, n\}$.

Formally, we are given a collection of T linear regression tasks, $(\mathbf{X}_1, \mathbf{y}_1), \ldots, (\mathbf{X}_T, \mathbf{y}_T)$, where $\mathbf{X}_m \in \mathbb{R}^{n_m \times d}, \mathbf{y}_m \in \mathbb{R}^{n_m}$. Over k iterations, tasks are learned under a *task ordering* $\tau : [k] \to [T]$, and we focus on random orderings studied in, *e.g.,* Evron et al. (2022, 2023); Jung et al. (2025).

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

We study a direct learning scheme which minimizes the sum of squared errors for the current regression task,[2] without mitigating forgetting algorithmically (*e.g.,* with replay). This scheme (i) illuminates "raw" continual dynamics of gradient-based algorithms, and (ii) roughly captures linear dynamics of deep networks in the neural tangent kernel regime (Jacot et al., 2018).

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[3]
Output $\mathbf{w}_k$

This scheme was previously linked to the Kaczmarz method and, in a special case, to normalized SGD (Evron et al., 2022). In Section 3, we develop these connections to enable novel analysis.

2. This objective is natural for regression; our analysis also extends to the *mean* squared error (refining our R).
3. Learning to convergence facilitates the analysis, but other analytical choices exist (see Jung et al., 2025).

Our main assumption is the existence of *offline solutions* that fit the training data of all T tasks jointly, as assumed in much of the theoretical CL literature (e.g., Evron et al., 2022, 2023; Kong et al., 2023; Goldfarb et al., 2024; Jung et al., 2025). 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 very reasonable in highly overparameterized models, *e.g.,* deep networks in the neural tangent kernel (NTK) regime (Jacot et al., 2018).

Assumption 1 (Joint Linear Realizability of Training Data) *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 (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 (Evron et al., 2022). Our goal is to bound this degradation, *i.e.,* "forgetting". Following common definitions (e.g., Doan et al., 2021; Evron et al., 2023), we define forgetting as the average increase in the loss of the *last* iterate on previous tasks.

Definition 3 (Forgetting of Training Data) *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 training 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. (2022) showed forgetting can be "catastrophic" in the sense that $\lim_{k \to \infty} \mathbb{E}\left[F_\tau(k)\right] > 0$. However, as we show, this *cannot* happen under random orderings.

Remark 4 (Forgetting vs. Regret) *While regret and forgetting are related, they can differ significantly (Evron et al., 2022). Regret is a key quantity in online learning, defined in our setting as* $\frac{1}{2k} \sum_{t=1}^{k} \|\mathbf{X}_{\tau(t)} \mathbf{w}_{t-1} - \mathbf{y}_{\tau(t)}\|^2$*. 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 average training loss to easily connect with *other* fields, such as Kaczmarz.

Definition 5 (Training Loss) *The training loss of any vector* $\mathbf{w} \in \mathbb{R}^d$ *is given by,*

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

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 $\mathbb{E}_\tau[F_\tau(k)] \leq 2\mathbb{E}_\tau\left[\mathcal{L}\left(\mathbf{w}_{k-1}\right)\right] + \frac{\|\mathbf{w}_\star\|^2 R^2}{k}$ 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 $\tau_1, \ldots, \tau_{k-1}$ 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

Prior work has drawn connections between continual linear regression and the Kaczmarz method (Evron et al., 2022), which we revisit pedagogically to keep the paper self-contained. Importantly, this leads us to a novel reduction between the (block) Kaczmarz method and SGD on special functions (Schemes 2 and 3). As illustrated in Figure 1, this analytical flow allows us to improve the rates for continual and Kaczmarz methods by analyzing the last iterate of SGD instead.

Scheme 2 The Block Kaczmarz Method

Input: Jointly realizable $(\mathbf{X}_m, \mathbf{y}_m), \forall m \in [T]$
Initialize $\mathbf{w}_0 = \mathbf{0}_d$
For each iteration $t = 1, \ldots, k$:
$\quad \mathbf{w}_t \leftarrow \mathbf{w}_{t-1} - \mathbf{X}^+_{\tau(t)}(\mathbf{X}_{\tau(t)}\mathbf{w}_{t-1} - \mathbf{y}_{\tau(t)})$

Scheme 3 SGD with $\eta = 1$ on special $\{f_m\}_m$

Input: $f_m(\mathbf{w}) = \frac{1}{2}\|\mathbf{X}^+_m(\mathbf{X}_m\mathbf{w} - \mathbf{y}_m)\|^2, \forall m \in [T]$
Initialize $\mathbf{w}_0 = \mathbf{0}_d$
For each iteration $t = 1, \ldots, k$:
$\quad \mathbf{w}_t \leftarrow \mathbf{w}_{t-1} - \nabla_{\mathbf{w}} f_{\tau(t)}(\mathbf{w}_{t-1})$

3.1. Revisit: Continual Linear Regression and the Kaczmarz Method

The (block) Kaczmarz method in Scheme 2 (Kaczmarz, 1937; Elfving, 1980) 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. (2022) identified the following reduction.

Reduction 1 (Continual Regression ⇒ Block Kaczmarz) *In the realizable case (Assumption 1) under any ordering τ, continual linear regression learned to convergence*[4] *is equivalent to the block Kaczmarz method. That is, the iterates $\mathbf{w}_0, \ldots, \mathbf{w}_k$ 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 (Needell et al., 2014). That is, since in rank-1 we have $\mathcal{L}_{\tau(t)}(\mathbf{w}) = \frac{1}{2}\left\|\mathbf{x}^\top_{\tau(t)}\mathbf{w} - y_{\tau(t)}\right\|^2$, Kaczmarz updates hold

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

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*).

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

Reduction 2 (Block Kaczmarz ⇒ SGD) *In the realizable case (Assumption 1) 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 $\mathbf{w}_0, \ldots, \mathbf{w}_k$ of Schemes 2 and 3 coincide.*

Intuitively, the $\mathbf{X}_m^+$ term in the modified objectives $\{f_m\}$ generalizes the normalizing step size from the rank-1 case, fitting all directions in the current block precisely with the same step size $\eta = 1$.

The reduction above is key to our analysis flow (Figure 1) as it reveals that continual linear regression can be analyzed directly via SGD analysis. It follows from substituting the gradient from the next lemma into $(\mathbf{w}_{t-1} - \nabla_{\mathbf{w}} f_{\tau(t)}(\mathbf{w}_{t-1}))$ in Scheme 3. The lemma is proved in App. B.

Lemma 6 (Properties of the Modified Objective) *Consider any realizable task collection s.t. $\mathbf{X}_m \mathbf{w}_\star = \mathbf{y}_m, \forall m \in [T]$. 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: $\mathcal{L}_m(\mathbf{w}) \leq R^2 f_m(\mathbf{w}) \triangleq \max_{m' \in [T]} \left\| \mathbf{X}_{m'} \right\|^2 f_m(\mathbf{w})$.

(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 improves the best known upper bound: for random orderings with replacement, Evron et al. (2022) 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, their rate depends on the dimensionality d, challenging the transfer of insights from linear models to highly overparameterized deep networks (*e.g.,* via the NTK regime). Encouragingly, they only provided a worst-case lower bound of $1/k$, calling for further research to narrow this gap.

We tighten the existing problem-dependent rate from $\left(d - \bar{r} \right)$ to $\min \left(\sqrt{d - \bar{r}}, \sqrt{T \bar{r}} \right)$, and 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. See summary in the table below.

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. (2022)	Upper	$\frac{d - \bar{r}}{k}$	—	$\frac{T^2}{\sqrt{k}} \wedge \frac{T^2 (d - r_{\max})}{k}$
Kong et al. (2023)	Upper	—	—	$\frac{T^3}{k}$
Ours	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. (2022)	Lower	$\frac{1}{k}$ $(\star)$	$\frac{1}{T}$ $(\star)$	$\frac{T^2}{k}$

$(\star)$ They did not explicitly provide such lower bounds, but the 2-task 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 7 (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 $k \geq 3$ iterations are bounded as,*

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

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

Our proof, given in App. C, is related to a recent work (Guo et al., 2022) that characterizes the weak error (similar to our loss) by analyzing a linear map. Unlike ours, the polynomial rates they derive involve matrix properties related to the condition number.

Proof Idea. We rewrite the Kaczmarz update (Scheme 2) in a recursive form of the differences, *i.e.*, $\mathbf{w}_t - \mathbf{w}_\star = \mathbf{P}_{\tau(t)}(\mathbf{w}_{t-1} - \mathbf{w}_\star)$ for a suitable projection matrix $\mathbf{P}_{\tau(t)}$. We define the linear map $Q[\mathbf{A}] = \frac{1}{T}\sum_{m=1}^{T} \mathbf{P}_m \mathbf{A} \mathbf{P}_m$ 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 $\mathcal{O}(1/k)$ rate with explicit dependence on T, d, and $\bar{r}$.

Remark 8 (The $\|\mathbf{w}_\star\|^2 R^2$ Scaling Term) *All the rates we derive contain a multiplicative factor of $\|\mathbf{w}_\star\|^2 R^2$, a generally unavoidable scaling term in linear regression. Prior work on continual learning has either normalized it away implicitly—e.g., by assuming $\|\mathbf{w}_\star\|^2, R \leq 1$ (Evron et al., 2022)—or included it explicitly, as we do (Evron et al., 2023; Lin et al., 2023). The rate in Theorem 7 involves additional problem parameters, i.e., T, d, and $\bar{r}$, whereas the rate in Theorem 9 below is "universal" in the sense that it does not depend on any such parameter.*

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 independence is crucial in highly overparameterized regimes, as encountered in deep neural networks.

Theorem 9 (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,*

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

We prove this result in App. D.1 by leveraging the connections between CL and SGD. Specifically, Section 3 showed that continual linear regression is equivalent to SGD with step size *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 this specific step size. To ease readability, we keep a CL perspective here and defer last-iterate analysis to Section 5.

4.3. Random Task Orderings Without Replacement

Evron et al. (2022) suggested that forgetting is 'catastrophic' only when $\lim_{k\to\infty} \mathbb{E}\left[F_\tau(k)\right] > 0$, and presented such an adversarial case with a deterministic task ordering where $k = T \to \infty$. In contrast, they showed that cyclic or random task orderings mitigate forgetting, perhaps due to task repetition. So far, under random orderings, it has been difficult to disentangle the effect of randomness from that of repetition—*i.e.,* whether their remedying impact arises from random permutation or repeated exposure. Below, we provide the first result demonstrating that randomly permuting tasks is sufficient to alleviate catastrophic forgetting.

Theorem 10 (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 $k \in \{2, \ldots, T\}$ iterations are both bounded as,*

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

The proof is given in App. D.2. The dimensionality-dependent term parallels the with-replacement case in App. D.1.2 of Evron et al. (2022), but requires a refined upper bound on in-sample forgetting. The dimensionality-independent term again relies on last-iterate analysis, as presented in App. E.2.

In App. A, we discuss connections between our result above and 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 (Varre et al., 2021). 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 3 in App. A compares our bounds with related work and classical results in the field.

Recent work has analyzed SGD in *realizable* (possibly noisy) least squares settings (Ge et al., 2019; Vaswani et al., 2019; Berthier et al., 2020; Zou et al., 2021; Varre et al., 2021; Wu et al., 2022). Realizable settings are primarily motivated by connections to deep networks in the overparameterized regime (Ma et al., 2018), where models are expressive enough to perfectly fit the training data. With the exception of Varre et al. (2021), most of these works focus on non-fixed step sizes and/or provide guarantees for the average iterate (see App. A for discussion). Similarly, here we study the following stochastic, jointly realizable least squares problem.

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

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

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, that is, $\left\|\mathbf{A}_i^\top\mathbf{A}_i\right\| \leq \beta, \forall i \in \mathcal{I}$, 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 11 (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:*

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

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

An important part of Theorem 11 is the $(2-\eta\beta)$ factor in the denominator, replacing a $(1-\eta\beta)$ common in 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. (2021), *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 11 to SGD without replacement. The proof leverages algorithmic stability for SGD (Bousquet and Elisseeff, 2002; Shalev-Shwartz et al., 2010; Hardt et al., 2016), focusing on a variant tailored to without-replacement sampling (Sherman et al., 2021; Koren et al., 2022). 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 11, which follows immediately by combining the two lemmas below (while noting that $\eta < 2/\beta \Rightarrow e^{\eta\beta(1-\eta\beta/4)} \leq e$). The first step of the proof is to establish a regret bound for SGD when applied to $f(\mathbf{w}; i_1) \ldots f(\mathbf{w}; i_T)$, 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 12 (Gradient Descent Regret Bound for Smooth Optimization) *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:*

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

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 (Shamir and Zhang, 2013). The result, given below, is slightly more general to accommodate without-replacement sampling, addressed in the next section.

Lemma 13 *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:*

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

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

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 (Gubin et al., 1967; Boyd et al., 2003). 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.[5] 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 14 (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 bounded as,*

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

To the best of our knowledge, this is the first universal rate in the POCS literature, independent of problem parameters such as regularity or complexity, as demonstrated 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, squared distances from individual sets relate not only to MSE, but also to losses such as the squared hinge loss for classification (Evron et al., 2023), naturally leading to our next continual model.

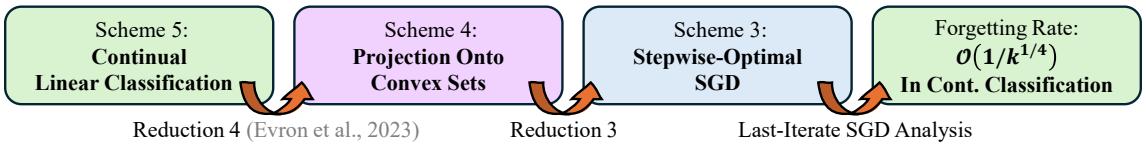

Figure 2: Analysis Flow Leading to Our Improved Classification Rates.

5. This has been *partially* observed in the POCS literature (e.g., Nedić, 2010).

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 Kirkpatrick et al., 2017; Aljundi et al., 2018; Li et al., 2023). Evron et al. (2023) studied a weakly-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 $y \in \{-1, +1\}$.

Scheme 5 Weakly-Regularized Continual Linear Classification (for $\lambda \to 0$)

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

Specifically, Evron et al. (2023) proved that learning an entire (separable) classification task in this continual scheme implicitly applies projection onto convex sets. More formally,

Reduction 4 (Continual Classification $\Rightarrow$ POCS) *Given jointly separable tasks, the continual iterates of Scheme 5 in the limit as $\lambda \to 0$, align in direction of the sequential projections of Scheme 4 onto the convex sets defined as $\mathcal{C}_m \triangleq \left\{ \mathbf{w} \in \mathbb{R}^d \mid y\mathbf{w}^\top \mathbf{x} \geq 1,\ \forall (\mathbf{x}, y) \in S_m \right\}$, $\forall m \in [T]$.*

Importantly, this reduction enables the study of continual classification through projection algorithms. In particular, Evron et al. (2023) studied forgetting using an equivalent of our Definition 3:

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

As illustrated in Figure 2, we derive the following bound by combining their reduction with our POCS rate (Theorem 14) and SGD stability arguments.

Theorem 15 *Under a random ordering, with or without replacement, over T jointly separable tasks, the expected forgetting of the weakly-regularized Scheme 5 after $k \geq 1$ iterations is,*

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

As shown in Table 2, 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: $\frac{5}{T^{1/4}\sqrt{\|\mathbf{w}_\star\| R}}$.

Table 2: **Forgetting Rates in Weakly-Regularized Continual Linear Classification on Separable Data.** All cells omit mild multiplicative constants and normalize by an unavoidable $\left\| \mathbf{w}_\star \right\|^2 R^2$ term.

Paper / Ordering	Random with Replacement	Random w/o Replacement	Cyclic
Evron et al. (2023)	$\exp\left(-\frac{k}{4T\|\mathbf{w}_\star\|^2 R^2}\right)$	—	$\frac{T^2}{\sqrt{k}} \wedge \exp\left(-\frac{k}{16T^2\|\mathbf{w}_\star\|^2 R^2}\right)$
Ours	$\frac{1}{\sqrt[4]{k}}$	$\frac{1}{\sqrt[4]{T}}$	—

7. Discussion

Our work established reductions from continual linear regression and classification 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. Our main results are summarized in Tables 1, 2 and 3.

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 (Kirkpatrick et al., 2017). 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. (2023) 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 (Evron et al., 2022). 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 $\mathbb{E}_\tau\left[\frac{1}{2T}\sum_{m=1}^{T}\left\|\mathbf{A}_m\mathbf{x}_k - \mathbf{b}_m\right\|^2\right] = \mathcal{O}\big(\min\left(k^{-1/4}, \frac{1}{k}\sqrt{d-\bar{r}}, \frac{1}{k}\sqrt{T\bar{r}}\right)\big)$ for random orderings with replacement and $\mathcal{O}\big(\min\left(k^{-1/4}, \frac{1}{k}\left(d-\bar{r}\right)\right)\big)$ without replacement. Note that we bounded the *loss*, rather than the "error" $\|\mathbf{w}_k - \mathbf{w}_\star\|^2$, 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 (2009) proposed a Kaczmarz variant that samples rows with probability proportional to their squared norm. Our approach also accommodates non-uniform sampling, including norm-based ones. Specifically, Claim G.1 tightens Theorem 9 by employing norm-weighted *block* sampling, thereby replacing the dependence on the maximum row norm R with the average $\bar{R}$.

From Training to Generalization Results. The rates derived in our paper apply to the forgetting and loss on the training sets (Definitions 3 and 5). These rates can be extended to the generalization loss at the cost of an additive $\mathcal{O}(1/\sqrt{N})$ term (where N is the number of samples per task) via uniform-convergence arguments—or even an additive $\mathcal{O}(1/N)$ term using the more refined Rademacher bounds for linear models from Srebro et al. (2010a).

Future and Follow-up 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 $\mathcal{O}(1/k^{1/4})$, largely stemming from our proof technique in Lemma 13. Our argument follows the approach of Shamir and Zhang (2013), originally developed to control the last SGD iterate in the convex, non-smooth regime. When instantiated in our (smooth,

realizable) setting, this approach introduces an exponential dependence on $\eta\beta$. Because of this exponential sensitivity, employing coarse inequalities (*e.g.,* bounding a negative term by 0) can be costly: even small constant-factor losses may effectively change the power of T appearing in the final bound.

Following our reductions (Sections 3 and 6), improved rates for "stepwise-optimal" SGD rates would immediately refine the bounds for continual linear regression and classification. Indeed, a follow-up work by Attia et al. (2025) departs from Shamir and Zhang (2013) and builds on a different technique proposed by Zamani and Glineur (2023). Using a more refined analysis, they establish an improved rate for SGD and, combined with our reductions, obtain a tighter upper bound of $\mathcal{O}(1/\sqrt{k})$ for continual linear regression (and classification). Finding the exact worst-case complexity between $\Omega(1/k)$ and $\mathcal{O}(1/\sqrt{k})$ remains an open question.

Finally, we note that our analytical flow relies on Reduction 1, which in turn assumes that the continual Scheme 1 learns each task *to convergence*. Nonetheless, a concurrent follow-up by Levinstein et al. (2025) follows a similar overall analytical flow, despite studying more practical continual learning schemes that do not learn to convergence—namely, those using ℓ_2 regularization or finite step budgets. Similarly to us, they reduce learning an entire task to a single gradient step and apply last-iterate SGD analysis to obtain convergence rates.

Acknowledgments

We thank Edgar Dobriban (University of Pennsylvania) and Amit Attia (Tel Aviv University) for fruitful discussions.

The research of DS was funded by the European Union (ERC, A-B-C-Deep, 101039436). Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency (ERCEA). Neither the European Union nor the granting authority can be held responsible for them. DS also acknowledges the support of the Schmidt Career Advancement Chair in AI.

The research of TK has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 101078075). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. This work received additional support from the Israel Science Foundation (ISF, grant number 3174/23), and a grant from the Tel Aviv University Center for AI and Data Science (TAD).

NS was supported in part by the Simons Foundation and the NSF-TRIPOD Institute on Data Economics and Learning (IDEAL).

References

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 139–154, 2018. (cited on p. 11)

Amit Attia, Matan Schliserman, Uri Sherman, and Tomer Koren. Fast last-iterate convergence of sgd in the smooth interpolation regime. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*, 2025. (cited on p. 13)

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). *Advances in neural information processing systems*, 26, 2013. (cited on p. 2, 20)

Mohammadamin Banayeeanzade, Mahdi Soltanolkotabi, and Mohammad Rostami. Theoretical insights into overparameterized models in multi-task and replay-based continual learning. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. (cited on p. 2)

Heinz H Bauschke, Patrick L Combettes, et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011. (cited on p. 45)

Raphaël Berthier, Francis Bach, and Pierre Gaillard. Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model. *Advances in Neural Information Processing Systems*, 33:2576–2586, 2020. (cited on p. 2, 8)

Léon 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. (cited on p. 21)

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002. (cited on p. 9, 21, 40)

Stephen Boyd, Jon Dattorro, et al. Alternating projections. *EE392o, Stanford University*, 2003. (cited on p. 10)

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends®in Machine Learning*, 8(3-4):231–357, 2015. (cited on p. 35)

Xufeng Cai and Jelena Diakonikolas. Last iterate convergence of incremental methods and applications in continual learning. In *The Thirteenth International Conference on Learning Representations*, 2025. (cited on p. 2, 21)

Xufeng Cai, Cheuk Yin Lin, and Jelena Diakonikolas. Tighter convergence bounds for shuffled SGD via primal-dual perspective. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. (cited on p. 21)

Jaeyoung Cha, Jaewook Lee, and Chulhee Yun. Tighter lower bounds for shuffling sgd: Random permutations and beyond. In *International Conference on Machine Learning*, pages 3855–3912. PMLR, 2023. (cited on p. 21)

E.K.P. Chong and S.H. Zak. *An Introduction to Optimization*. Wiley-Interscience Series in Discrete Mathematics and Optimi. Wiley, 2004. ISBN 9780471654001. (cited on p. 22)

Christopher M De Sa. Random reshuffling is not always better. *Advances in Neural Information Processing Systems*, 33:5957–5967, 2020. (cited on p. 21)

Thang Doan, Mehdi Abbana Bennani, Bogdan Mazoure, Guillaume Rabusseau, and Pierre Alquier. A theoretical analysis of catastrophic forgetting through the ntk overlap matrix. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 1072–1080, 2021. (cited on p. 2, 4)

Tommy Elfving. Block-iterative methods for consistent and inconsistent linear equations. *Numerische Mathematik*, 35(1):1–12, 1980. (cited on p. 5)

Itay Evron, Edward Moroshko, Rachel Ward, Nathan Srebro, and Daniel Soudry. How catastrophic can catastrophic forgetting be in linear regression? In *Conference on Learning Theory (COLT)*, pages 4028–4079. PMLR, 2022. (cited on p. 2, 3, 4, 5, 6, 7, 8, 12, 22)

Itay Evron, Edward Moroshko, Gon Buzaglo, Maroun Khriesh, Badea Marjieh, Nathan Srebro, and Daniel 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. (cited on p. 2, 3, 4, 7, 10, 11, 22)

Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003. (cited on p. 45, 46)

Rong Ge, Sham M Kakade, Rahul Kidambi, and Praneeth Netrapalli. The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares. *Advances in Neural Information Processing Systems*, 32, 2019. (cited on p. 2, 8)

Daniel Goldfarb and Paul 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. (cited on p. 2)

Daniel Goldfarb, Itay Evron, Nir Weinberger, Daniel Soudry, and Paul Hand. The joint effect of task similarity and overparameterization on catastrophic forgetting - an analytical model. In *The Twelfth International Conference on Learning Representations*, 2024. (cited on p. 1, 4)

Robert M Gower and Peter Richtárik. Randomized iterative methods for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1660–1690, 2015. (cited on p. 21)

LG Gubin, Boris T Polyak, and EV Raik. The method of projections for finding the common point of convex sets. *USSR Computational Mathematics and Mathematical Physics*, 7(6):1–24, 1967. (cited on p. 3, 10)

Xin Guo, Junhong Lin, and Ding-Xuan Zhou. Rates of convergence of randomized kaczmarz algorithms in hilbert spaces. *Applied and Computational Harmonic Analysis*, 61:288–318, 2022. (cited on p. 7, 26)

Deren Han and Jiaxin Xie. A simple linear convergence analysis of the reshuffling kaczmarz method. *arXiv preprint arXiv:2410.01140*, 2024. (cited on p. 21)

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234. PMLR, 2016. (cited on p. 9, 21, 40, 42, 43, 48, 49)

Hamed Hemati, Lorenzo Pellegrini, Xiaotian Duan, Zixuan Zhao, Fangfang Xia, Marc Masana, Benedikt Tscheschner, Eduardo Veas, Yuxiang Zheng, Shiji Zhao, et al. Continual learning in the presence of repetition. In *CVPR Workshop on Continual Learning in Computer Vision*, 2024. (cited on p. 2)

Naoki Hiratani. Disentangling and mitigating the impact of task similarity for continual learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. (cited on p. 2)

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2012. (cited on p. 28)

Arthur Jacot, Franck Gabriel, and Clement 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. (cited on p. 3, 4, 22)

Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Making the last iterate of sgd information theoretically optimal. In *Conference on Learning Theory*, pages 1752–1755. PMLR, 2019. (cited on p. 20)

Hyunji Jung, Hanseul Cho, and Chulhee Yun. Convergence and implicit bias of gradient descent on continual linear classification. In *The Thirteenth International Conference on Learning Representations*, 2025. (cited on p. 2, 3, 4)

S Kaczmarz. Angenäherte auflösung von systemen linearer gleichungen. *Bull. Int. Acad. Pol. Sic. Let., Cl. Sci. Math. Nat.*, pages 355–357, 1937. (cited on p. 2, 5)

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. (cited on p. 11, 12)

Mark Kong, William Swartworth, Halyun Jeong, Deanna Needell, and Rachel Ward. Nearly optimal bounds for cyclic forgetting. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. (cited on p. 2, 4, 6)

Tomer Koren, Roi Livni, Yishay Mansour, and Uri Sherman. Benign underfitting of stochastic gradient descent. *Advances in Neural Information Processing Systems*, 35:19605–19617, 2022. (cited on p. 9, 21, 40)

Zehua Lai and Lek-Heng Lim. Recht-ré noncommutative arithmetic-geometric mean conjecture is false. In *International Conference on Machine Learning*, pages 5608–5617. PMLR, 2020. (cited on p. 21)

Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012. (cited on p. 20)

Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In Hal Daumé III and Aarti 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. (cited on p. 21, 41)

Timothée Lesort, Oleksiy Ostapenko, Pau Rodríguez, Diganta Misra, Md Rifat Arefin, Laurent Charlin, and Irina Rish. Challenging common assumptions about catastrophic forgetting and knowledge accumulation. In *Conference on Lifelong Learning Agents*, pages 43–65. PMLR, 2023. (cited on p. 2, 12)

Ran Levinstein, Amit Attia, Matan Schliserman, Uri Sherman, Tomer Koren, Daniel Soudry, and Itay Evron. Optimal rates in continual linear regression via increasing regularization. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*, 2025. (cited on p. 13)

Haoran Li, Jingfeng Wu, and Vladimir Braverman. Fixed design analysis of regularization-based continual learning. In Sarath Chandar, Razvan Pascanu, Hanie Sedghi, and Doina 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. (cited on p. 2, 11)

Sen Lin, Peizhong Ju, Yingbin Liang, and Ness Shroff. Theory on forgetting and generalization of continual learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 21078–21100. PMLR, 23–29 Jul 2023. (cited on p. 2, 7)

Zijian Liu and Zhengyuan Zhou. Revisiting the last-iterate convergence of stochastic gradient methods. In *The Twelfth International Conference on Learning Representations*, 2024a. (cited on p. 20)

Zijian Liu and Zhengyuan Zhou. On the last-iterate convergence of shuffling gradient methods. In *Forty-first International Conference on Machine Learning*, 2024b. (cited on p. 21)

Siyuan Ma, Raef Bassily, and Mikhail 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. (cited on p. 8)

Martial Mermillod, Aurélia Bugaiska, and Patrick Bonin. The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects, 2013. (cited on p. 1)

Konstantin Mishchenko, Ahmed Khaled Ragab Bayoumi, and Peter Richtárik. Random reshuffling: Simple analysis with vast improvements. *Advances in Neural Information Processing Systems*, 33, 2020. (cited on p. 21)

Dheeraj Nagaraj, Prateek Jain, and Praneeth Netrapalli. Sgd without replacement: Sharper rates for general smooth convex functions. In *International Conference on Machine Learning*, pages 4703–4711. PMLR, 2019. (cited on p. 21)

Angelia Nedić. Random projection algorithms for convex set intersection problems. In *49th IEEE Conference on Decision and Control (CDC)*, pages 7655–7660. IEEE, 2010. (cited on p. 10)

Deanna Needell, Rachel Ward, and Nati Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *Advances in neural information processing systems*, 27: 1017–1025, 2014. (cited on p. 5)

Yurii Nesterov. Introductory lectures on convex programming volume i: Basic course. *Lecture notes*, 3(4):5, 1998. (cited on p. 20)

Francesco 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. (cited on p. 20)

Peter Oswald and Weiqi Zhou. Convergence analysis for kaczmarz-type methods in a hilbert space framework. *Linear Algebra and its Applications*, 478:131–161, 2015. (cited on p. 21)

Liangzu Peng, Paris Giampouras, and René Vidal. The ideal continual learner: An agent that never forgets. In *International Conference on Machine Learning*, 2023. (cited on p. 2)

Shashank Rajput, Anant Gupta, and Dimitris Papailiopoulos. Closing the convergence gap of SGD without replacement. In *International Conference on Machine Learning*, pages 7964–7973. PMLR, 2020. (cited on p. 21)

Benjamin Recht and Christopher Ré. Beneath the valley of the noncommutative arithmetic-geometric mean inequality: conjectures, case-studies, and consequences. In *Conference on Learning Theory (COLT)*, 2012a. (cited on p. 21)

Benjamin Recht and Christopher 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, 2012b. (cited on p. 21)

Itay Safran and Ohad Shamir. How good is SGD with random shuffling? In *Conference on Learning Theory*, pages 3250–3284. PMLR, 2020. (cited on p. 21)

Sundar G Sankaran and AA Louis Beex. Convergence behavior of affine projection algorithms. *IEEE Transactions on Signal Processing*, 48(4):1086–1096, 2000. (cited on p. 21)

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010. (cited on p. 9, 21, 40, 42, 48)

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 71–79. PMLR, 2013. (cited on p. 2, 9, 12, 13, 20)

Uri Sherman, Tomer Koren, and Yishay Mansour. Optimal rates for random order online optimization. *Advances in Neural Information Processing Systems*, 34:2097–2108, 2021. (cited on p. 9, 21, 40, 41, 42)

Dirk TM Slock. On the convergence behavior of the lms and the normalized lms algorithms. *IEEE Transactions on Signal Processing*, 41(9):2811–2825, 1993. (cited on p. 21)

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Optimistic rates for learning with a smooth loss. *arXiv preprint arXiv:1009.3896*, 2010a. (cited on p. 12)

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. *Advances in neural information processing systems*, 23, 2010b. (cited on p. 2, 20)

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

Matan Tsipory, Ran Levinstein, Itay Evron, Mark Kong, Deanna Needell, and Daniel Soudry. Are greedy task orderings better than random in continual linear regression? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=8JdPqAMpi4. (cited on p. 2)

Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Last iterate convergence of SGD for least-squares in the interpolation regime. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. (cited on p. 2, 8, 9, 20)

Sharan Vaswani, Francis Bach, and Mark 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. (cited on p. 8)

Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham 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. (cited on p. 2, 8)

Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Open problem: Can single-shuffle sgd be better than reshuffling sgd and gd? In *Conference on Learning Theory*, pages 4653–4658. PMLR, 2021. (cited on p. 21)

Moslem Zamani and François Glineur. Exact convergence rate of the last iterate in subgradient methods. *arXiv preprint arXiv:2307.11134*, 2023. (cited on p. 13, 20)

Xuyang Zhao, Huiyuan Wang, Weiran Huang, and Wei Lin. A statistical theory of regularization-based continual learning. In *Forty-first International Conference on Machine Learning*, 2024. (cited on p. 2)

Huiping Zhuang, Zhiping Lin, and Kar-Ann Toh. Blockwise recursive moore–penrose inverse for network learning. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(5):3237–3250, 2021. (cited on p. 22)

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Benign overfitting of constant-stepsize sgd for linear regression. In *Conference on Learning Theory*, pages 4633–4635. PMLR, 2021. (cited on p. 2, 8)

Appendix 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 (2024a) 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 3). Our analysis technique in Section 5 borrows from the work of Shamir and Zhang (2013, also mentioned in Table 3) which, in fact, belongs to the comparatively-richer line of work on the non-smooth setting (Shamir and Zhang, 2013; Jain et al., 2019; Zamani and Glineur, 2023; Liu and Zhou, 2024a). 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 3: **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 (2013)	$\frac{1}{\eta T} + \eta G^2 \log T$	✓	✗
Deterministic Smooth ($\sigma = 0$)	Nesterov (1998)	$\frac{1}{(2-\eta\beta)\eta T}$	✓	✓
Stochastic Smooth	Lan (2012)	$\frac{1}{\eta T} + \eta\sigma^2$	✗	✗
	Liu and Zhou (2024a)	$\frac{1}{\eta T} + \eta\sigma^2 \log T$	✓	✗
Stochastic Smooth Realizable ($\sigma_\star = 0$)	Srebro et al. (2010b)	$\frac{1}{(1-\eta\beta)\eta T}$	✗	✗
Stochastic Regression Realizable ($\sigma_\star = 0$)	Bach and Moulines (2013)	$\frac{1}{\eta T}$	✗	✓
	Varre et al. (2021)	$\frac{1}{(1-2\eta\beta\log T)\eta T}$	✓	✗
	Ours	$\frac{1}{(2-\eta\beta)\eta T^{1-\eta\beta(1-\eta\beta/4)}}$	✓	✓

(⋆) They consider bounded domains (Shamir and Zhang, 2013); Orabona (2020); Liu and Zhou (2024a) obtain similar bounds for the unconstrained case. For non-fixed step sizes, Jain et al. (2019) 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., Recht and Ré, 2012b; Nagaraj et al., 2019; Safran and Shamir, 2020; Rajput et al., 2020; Mishchenko et al., 2020; Cha et al., 2023; Cai et al., 2024). For the non-strongly convex case, near-optimal bounds (for the average iterate) have been established for the general smooth case (Nagaraj et al., 2019; Mishchenko et al., 2020). In a subsequent work, Cai et al. (2024) 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 (2024b) and Cai and Diakonikolas (2025). 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, Mishchenko et al. (2020); Cai et al. (2024) obtained an $O(\beta/J)$ bound for the average iterate with step size $\eta = 1/(\beta n)$; and Liu and Zhou (2024b); Cai and Diakonikolas (2025) derived similar bounds for the last iterate up to logarithmic factors.

Another line of work related to ours studies algorithmic stability (Bousquet and Elisseeff, 2002; Shalev-Shwartz et al., 2010) 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 Nagaraj et al. (2019); Sherman et al. (2021); Koren et al. (2022) and primarily builds on Sherman et al. (2021), who were the first to formally introduce the notion of without-replacement stability. For with-replacement SGD, Hardt et al. (2016) discussed its algorithmic stability under smooth loss functions. Later, Lei and Ying (2020), 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 $\mathbf{w}_\star$ (Gower and Richtárik, 2015; Han and Xie, 2024), but as we explained in Section 7, the rates can be arbitrarily slow. Recht and Ré (2012a) 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 (2020) proved that this inequality does not hold in general (see also De Sa, 2020). Moreover, as in other areas, empirical studies found that row shuffling followed by cyclic orderings performs as well as i.i.d. orderings (Oswald and Zhou, 2015). This naturally connects to interesting observations and open questions regarding various forms of shuffled SGD (Bottou, 2009; Yun et al., 2021). 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 (Slock, 1993), 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 (Sankaran and Beex, 2000). 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. (2022) 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.

Alternative Continual Schemes That Do Not Forget. Schemes such as Recursive Least Squares (RLS; see Chong and Zak 2004, Chapter 12.2) and its block variant BRMP (Zhuang et al., 2021) provide analytical alternatives to the block Kaczmarz method (Scheme 2). These methods effectively avoid forgetting by maintaining an $\mathcal{O}(d^2)$ matrix. See also Proposition 5.5 in Evron et al. (2023).

Importantly, we study a continual scheme (Scheme 1) that closely characterizes common training practices. In particular, training with (S)GD to convergence coincides with the analytical updates of the (block) Kaczmarz method (Scheme 2), making the latter illustrative of most gradient-based continual learning approaches. Understanding continual gradient-based algorithms in linear models is especially relevant given the linear dynamics of deep neural networks in the NTK regime (Jacot et al., 2018). These regimes are typically highly overparameterized, yet they remain covered by our analysis and by the dimensionality-independent rates we derive for the naturally forgetful but *memoryless* Scheme 1.

Appendix B. Auxiliary Proofs

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

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

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,

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

Proof. As discussed in Section 3.1, Scheme 2 governs the updates of the iterates $\mathbf{w}_t \in \mathbb{R}^d$. Under Assumption 1, 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:

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

We show that,

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

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

$$\begin{aligned}
&\frac{R^2}{k}\sum_{t=1}^{k}\mathbb{E}_\tau\left\|\left(\mathbf{I}-\mathbf{P}_{\tau(t)}\right)\left(\mathbf{w}_{t-1}-\mathbf{w}_\star\right)\right\|^2\\
&=\frac{R^2}{k}\sum_{t=1}^{k}\left(\mathbb{E}_\tau\left\|\mathbf{w}_{t-1}-\mathbf{w}_\star\right\|^2-\mathbb{E}_\tau\left\|\mathbf{P}_{\tau(t)}\left(\mathbf{w}_{t-1}-\mathbf{w}_\star\right)\right\|^2\right)\\
&=\frac{R^2}{k}\sum_{t=1}^{k}\left(\mathbb{E}_\tau\left\|\mathbf{w}_{t-1}-\mathbf{w}_\star\right\|^2-\mathbb{E}_\tau\left\|\mathbf{w}_{t}-\mathbf{w}_\star\right\|^2\right)\\
&=\frac{R^2}{k}\left(\underbrace{\mathbb{E}_\tau\left\|\mathbf{w}_0-\mathbf{w}_\star\right\|^2}_{=\|\mathbf{w}_\star\|^2}-\underbrace{\mathbb{E}_\tau\left\|\mathbf{w}_k-\mathbf{w}_\star\right\|^2}_{\geq 0}\right)\leq\frac{\|\mathbf{w}_\star\|^2R^2}{k}\,.
\end{aligned}$$

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

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

which completes the first part of the proof.

For the second part, simply notice that in an i.i.d. setting, the index $\tau(k)\sim\mathrm{Unif}\left([T]\right)$ 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 $\mathbb{E}_\tau\left[\mathcal{L}\left(\mathbf{w}_k\right)\right] = \mathbb{E}_\tau\left[F_\tau\left(k\right)\right]$, and the claim follows. For $k < T$, we have:

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

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

Recall Lemma 6. Consider any realizable task collection such that $\mathbf{X}_m\mathbf{w}_\star = \mathbf{y}_m, \forall m \in [T]$. Define $f_m(\mathbf{w}) = \frac{1}{2}\left\|\mathbf{X}_m^+\mathbf{X}_m\left(\mathbf{w} - \mathbf{w}_\star\right)\right\|^2$. Then, $\forall m \in [T], \mathbf{w} \in \mathbb{R}^d$

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

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

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

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

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

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 $\forall \mathbf{w}, \mathbf{z} \in \mathbb{R}^d$,

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

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

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

Recall Theorem 7. 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 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 (Guo et al., 2022). Note that $\mathbf{P}_m$ is an orthogonal projection, *i.e.,* symmetric and idempotent. Thus,

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

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

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

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:

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

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,

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

■

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 Horn and Johnson, 2012, 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 (Horn and Johnson, 2012, 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,* $\frac{1}{R^2T}\mathbf{X}^\top\mathbf{X} \preccurlyeq \mathbf{X}^+\mathbf{X} - Q\left[\mathbf{X}^+\mathbf{X}\right]$

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:

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

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

- If $\left(\mathbf{\Sigma}_i\right)_{j,j} = 0$, 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 $\frac{1}{R^2}\left(\mathbf{\Sigma}_i^2\right)_{j,j} \le 1$, which gives $\left(\mathbf{\Sigma}_i^+\mathbf{\Sigma}_i - \frac{1}{R^2}\mathbf{\Sigma}_i^2\right)_{j,j} \ge 0$.

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:

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

■

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)),

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

■

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,

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

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

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

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

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

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 $\|Q^n\left(I-Q\right)\| \leq \frac{1}{en}$, *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 $x \mapsto x^n\left(1-x\right)$. The eigenvalues of $Q^n\left(I-Q\right)$ are of the form $\lambda^n\left(1-\lambda\right)$, where λ is an eigenvalue of Q. From Proposition C.10, we know that $\lambda \in [0,1]$. Using an algebraic property of $\lambda^n\left(1-\lambda\right)$ for $\lambda \in [0,1]$, we conclude that $\lambda^n\left(1-\lambda\right) \in \left[0, \frac{1}{en}\right]$.
Therefore, $\|Q^n\left(I-Q\right)\| \leq \frac{1}{en}$. ■

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

Proof. We first bound $\|Q\left[\mathbf{X}^+\mathbf{X}\right]\|_F$ 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,

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

■

Appendix 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 9: A Universal $\mathcal{O}(1/\sqrt[4]{k})$ Rate

Recall Theorem 9. 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 6 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 6, $f_m(\cdot)$ 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 11, 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 10: Main Result for Without Replacement Orderings

Recall Theorem 10. 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,

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

Proof. From Lemmas 6 and B.1, we have

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

Combining with Proposition B.2, we get,

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

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

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

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

$$a_{t+1} = \operatorname{tr}\left(\mathbf{P}_{\tau(t+1)}\cdots\mathbf{P}_{\tau(2)}\left(\mathbf{I}-\mathbf{P}_{\tau(1)}\right)\mathbf{P}_{\tau(2)}\cdots\mathbf{P}_{\tau(t+1)}\right)$$

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

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

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

For the **second**, *parameter independent* term, note that from Lemma 6, $f_m(\cdot)$ is 1-smooth for all $m \in [T]$, and recall that the iterates $\mathbf{w}_t$ follow 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

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

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

■

Appendix 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 (2015). 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 $f(\mathbf{w}) = \frac{1}{2}\|\mathbf{A}\mathbf{w} - \mathbf{b}\|^2$ 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:

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

■

Recall Lemma 12. 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:

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

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

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

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

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

Combining the two displays above, it follows that

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

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

Recall Lemma 13. 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 $i_0, \dots i_{t-1} \in \mathcal{I}^t, i \in \mathcal{I}$, $\Pr(i_{\tau_1} = i | i_0, \dots, i_{t-1}) = \Pr(i_{\tau_2} = i | i_0, \dots, i_{t-1})$. Then, for any initialization $\mathbf{w}_0 \in \mathbb{R}^d$, with-replacement SGD (Eq. (2)) with step-size $\eta < 2/\beta$, holds:

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

where the expectation is taken with respect to $i_0, \dots, i_T$ 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):

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

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

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

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

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

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,

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

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

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

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 $c = \big(1-\frac{\eta\beta}{2}\big)^2$ (under the $\eta < \frac{2}{\beta}$ condition, we now have both $\gamma > 0$ and $c \in (0,1)$ as required above). Then, $1-c = \eta\beta\big(1-\frac{\eta\beta}{4}\big)$, and we finally get the required

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

■

E.2. Extending the SGD Bounds to Without Replacement Orderings

Here, we extend Theorem 11 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 $\pi_t \sim \text{Unif}(\mathcal{I})$ 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 (Bousquet and Elisseeff, 2002; Shalev-Shwartz et al., 2010; Hardt et al., 2016), and more specifically, on a variant of stability, suitable for without replacement sampling (Sherman et al., 2021; Koren et al., 2022).

The proof of our theorem follows by a combination of Lemmas E.5 and E.6. The first, stated below, establishes a bound on the expected "next sample" loss and follows immediately by combining Lemmas 12 and 13 (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 $1 \leq T \leq n-1$:*

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

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 (2020), combined with the without-replacement stability framework proposed by Sherman et al. (2021). 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,

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

■

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. (2021, 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 (Shalev-Shwartz et al., 2010; Hardt et al., 2016).

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

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

Hence, by Lemma E.8,

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

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 Hardt et al., 2016):

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

Further,

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

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,

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

Now,

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

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

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

Finally, by the regret bound given in Lemma 12, $\sum_{i=0}^{t-1}f_i(\mathbf{w}_i)\leq\frac{\|\mathbf{w}_0-\mathbf{w}_\star\|^2}{2\eta(2-\eta\beta)}$, and therefore,

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

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

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

which completes the proof. ■

Appendix F. Supplementary Material for the Extension Section (Section 6)

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 $f_m(\mathbf{w}) = \frac{1}{2}\left\|\mathbf{w} - \mathbf{\Pi}_m(\mathbf{w})\right\|^2, \forall m \in [T]$. 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 (2003), 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

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

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 Facchinei and Pang, 2003). 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 Bauschke et al., 2011). ■

Lemma F.1 *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a nonempty closed and convex set, and $f(\mathbf{w}) = \frac{1}{2}\|\mathbf{w} - \mathbf{\Pi}_{\mathcal{K}}(\mathbf{w})\|^2$. 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 already established that $\nabla f(\mathbf{w}) = \mathbf{w} - \mathbf{\Pi}_{\mathcal{K}}(\mathbf{w})$. Combining this with simple algebra, we obtain,

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

By Theorem 1.5.5 (b) in Facchinei and Pang (2003), we have that

$$\langle \mathbf{w} - \mathbf{\Pi}_{\mathcal{K}}(\mathbf{w}), \mathbf{\Pi}_{\mathcal{K}}(\mathbf{w}) - \mathbf{\Pi}_{\mathcal{K}}(\mathbf{z})\rangle \geq 0\,,$$

and finally we get,

$$\langle \nabla f(\mathbf{w}), \mathbf{w} - \mathbf{z}\rangle \geq 2f(\mathbf{w}) - \langle \mathbf{w} - \mathbf{\Pi}_{\mathcal{K}}(\mathbf{w}), \mathbf{z} - \mathbf{\Pi}_{\mathcal{K}}(\mathbf{z})\rangle\ .$$

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:

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

■

Recall Theorem 14. 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,

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

Proof. The proof largely follows the same steps of Theorems 9 and 10. 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:

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

Proceeding, we denote the objective function:

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

Now, for a **with-replacement** ordering τ, invoke Theorem 11, except we use Lemma F.1 in the proof instead of Lemma E.3, to obtain:

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

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:

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

Similarly, by Lemma E.5,

$$\mathbb{E}_\tau f_{\tau(k+1)}(\mathbf{w}_k) \triangleq \mathbb{E}_\tau \tfrac{1}{2}\left\|\mathbf{w}_k - \mathbf{\Pi}_{\tau(k+1)}(\mathbf{w}_k)\right\|^2 \leq \frac{e}{2\sqrt[4]{k}} \min_{\mathbf{w}\in\mathcal{C}_\star} \|\mathbf{w}_0 - \mathbf{w}\|^2 .$$

$$(\tau \text{ without-replacement})$$

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

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

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

Recall Theorem 15. 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

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

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 14, 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 14, 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 (Shalev-Shwartz et al., 2010; Hardt et al., 2016), and denoting by $\mathbf{w}_\tau^{(i)}$ the SGD iterate after τ steps on the training set where the ith example was resampled as j_i:

$$\left|\mathbb{E}\left[\bar{f}(\mathbf{w}_T) - \hat{f}_{0:T-1}(\mathbf{w}_T)\right]\right| = \left|\frac{1}{T}\sum_{i=0}^{T-1}\mathbb{E}_{j_i\sim\mathcal{D}}\left[f(\mathbf{w}_T; j_i) - f(\mathbf{w}_T^{(i)}; j_i)\right]\right|$$

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

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 Hardt et al., 2016):

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

Further,

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

Therefore,

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

Now,

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

Summarizing, we have shown that:

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

Finally, by the regret bound given in Lemma 12, *i.e.,* $\sum_{i=0}^{T-1} f_i(\mathbf{w}_i) \leq \frac{\|\mathbf{w}_0-\mathbf{w}_\star\|^2}{2\eta(2-\eta\beta)}$, we have

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

and the result follows. ■

Appendix G. Supplementary Material for the Discussion Section (Section 7)

Claim G.1 (Average-Norm Universal 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,*

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

where $\bar{R} = \sum_{m=1}^{M} \|\mathbf{X}_m\|^2 / T$.

Proof sketch. Taking the non-worst case bound from Lemma 6, we have $\mathcal{L}_m(w) \leq \alpha_m f_m(w)$ for $\alpha_m = \|\mathbf{X}_m\|^2$. Then in the proof of Theorem 9, $\mathcal{L}(w) \leq \frac{A}{2T} \sum \frac{\alpha_m}{A} f_m(w)$, where $A = \sum \alpha_m$, and we may apply Theorem 11 (which supports arbitrary distributions $\mathcal{D}$, see Setup 1) with the distribution given by $\Pr_{\mathcal{D}}(i) = \alpha_i / A$. Finally, we have $\bar{R} = A/T$.