Continual Learning in Linear Classification on Separable Data

Itay Evron [1] **Edward Moroshko** [1] **Gon Buzaglo** [1] **Maroun Khriesh** [1] **Badea Marjieh** [1]
Nathan Srebro [2] **Daniel Soudry** [1]

Abstract

We analyze continual learning on a sequence of separable linear classification tasks with binary labels. We show theoretically that learning with weak regularization reduces to solving a sequential max-margin problem, corresponding to a special case of the Projection Onto Convex Sets (POCS) framework. We then develop upper bounds on the forgetting and other quantities of interest under various settings with recurring tasks, including cyclic and random orderings of tasks. We discuss several practical implications to popular training practices like regularization scheduling and weighting. We point out several theoretical differences between our continual classification setting and a recently studied continual regression setting.

1. Introduction

Continual learning deals with learning settings where distributions, or tasks, change over time, breaking traditional i.i.d. assumptions. While models trained sequentially are expected to accumulate knowledge and improve over time, practically they suffer from *catastrophic forgetting* (McCloskey & Cohen, 1989; Goodfellow et al., 2013), *i.e.,* their performance on previously seen tasks deteriorates over time.

Much research in continual learning has focused on heuristic approaches to remedying forgetting. Recent approaches achieve impressive empirical performance, but often require storing examples from previous tasks (*e.g.,* Robins (1995); Rolnick et al. (2019)), iteratively expanding the learned models (*e.g.,* Yoon et al. (2018)), or lessening the plasticity of these models and harming their performance on new tasks (*e.g.,* Kirkpatrick et al. (2017)).

[1]Department of Electrical and Computer Engineering, Technion, Haifa, Israel [2]Toyota Technological Institute at Chicago, Chicago IL, USA. Correspondence to: Itay Evron <itay@evron.me>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

We theoretically study the continual learning of a linear classification model on separable data with binary classes. Even though this is a fundamental setup to consider, there are still very few analytic results on it, since most of the continual learning theory thus far has focused on regression settings (*e.g.,* Bennani et al. (2020); Doan et al. (2021); Asanuma et al. (2021); Lee et al. (2021); Evron et al. (2022); Goldfarb & Hand (2023); Li et al. (2023)). Even in the broader deep learning scope, theoreticians often start from very simple models and use them to gain insight into phenomena arising in more practical models (*e.g.,* Belkin et al. (2018); Woodworth et al. (2020)).

Our paper reveals a surprising algorithmic bias of linear classifiers trained continually to minimize the exponential loss on separable data with weak regularization. Specifically, we prove that the weights converge in the same direction as the iterates of a Sequential Max-Margin scheme. This creates a bridge between the popular regularization methods for continual learning (Kirkpatrick et al., 2017; Zenke et al., 2017) and the well-studied POCS framework – Projections Onto Convex Sets (also known as the convex feasibility problem or successive projections).

Our results complement those of a recent paper (Evron et al., 2022) that analyzed the worst-case performance of continual linear regression. That paper showed that continually learning linear regression tasks with vanilla SGD, implicitly performs sequential projections onto closed *subspaces*, and connected that regime to the area of Alternating Projections (Von Neumann, 1949; Halperin, 1962). In our paper, we draw comparisons between our continual *classification* setting and their continual *regression* setting (summarized in App. A). We point out inherent differences between these two settings, emphasizing the need for a proper and thorough analytical understanding dedicated to continual *classification* settings.

Our Contributions Our analysis reveals the following:

- Explicit regularization methods for continual learning of linear classification models are linked to the framework of Projections Onto Convex Sets (POCS).
- Each learned task brings the learner *closer* to an "offline" feasible solution solving *all* tasks. However, there exist

task sequences for which the learner stays arbitrarily far from offline feasibility, even after infinitely many tasks.

- When tasks recur cyclically or randomly, the learner converges to an offline solution with linear rates.
- If we converge to an offline solution, it may not be the minimum-norm solution (in contrast to continual regression), but it still needs to be 2-optimal (minimal).
- Scheduling the regularization strength endangers convergence to an offline solution and optimality guarantees.
- Using popular regularization weighting schemes based on Fisher-information matrices, does not necessarily prevent forgetting (in contrast to continual regression).
- Early stopping (without regularization) does *not* yield the same solutions as weak regularization (unlike in stationary settings with a single task).

2. Setting

We consider $T \geq 2$ binary classification tasks. Each task $m = 1, \ldots, T$ is defined by a dataset S_m consisting of tuples of D-dimensional samples and their binary labels, *i.e.,* each tuple is $(\mathbf{x}, y) \in \mathbb{R}^D \times \{-1, +1\}$, for a finite D.

Notation. Throughout the paper, we denote the (isotropic) Euclidean norm of vectors by $\|\mathbf{v}\|$, and the weighted norm by $\|\mathbf{v}\|_{\mathbf{B}} \triangleq \mathbf{v}^\top \mathbf{B} \mathbf{v}$, for some $\mathbf{B} \succ \mathbf{0}_{D \times D}$. We denote the set of natural numbers starting from 1 by $\mathbb{N}^+ \triangleq \mathbb{N} \setminus \{0\}$ and the natural numbers from 1 to n by $[n]$. We define the distance of a vector $\mathbf{w} \in \mathbb{R}^D$ from a closed set $\mathcal{C} \subseteq \mathbb{R}^D$ as $d(\mathbf{w}, \mathcal{C}) \triangleq \min_{\mathbf{v} \in \mathcal{C}} \|\mathbf{w} - \mathbf{v}\|$. Finally, we denote the maximal norm of any data point by $R \triangleq \max_{m \in [T]} \max_{(\mathbf{x}, y) \in S_m} \|\mathbf{x}\|$.

Our main assumption in this paper is that the tasks are jointly-separable, *i.e.,* they can be perfectly fitted simultaneously, as in "offline" non-continual settings. This can be formally stated as follows.

Assumption 2.1 (Separability)**.** Each task $m \in [T]$ is separable, *i.e.,* it has a *non-empty* feasible set defined as

$$\mathcal{W}_m \triangleq \left\{ \mathbf{w} \in \mathbb{R}^D \mid y \mathbf{w}^\top \mathbf{x} \geq 1, \ \forall (\mathbf{x}, y) \in S_m \right\} .$$

Moreover, the T tasks are *jointly*-separable — there exists a non-empty offline feasible set:

$$\mathcal{W}^\star \triangleq \mathcal{W}_1 \cap \cdots \cap \mathcal{W}_T \neq \emptyset .$$

A similar assumption was made in the continual regression setting (Evron et al., 2022). It is a reasonable assumption in overparameterized regimes, where feasible offline solutions often *do* exist. Practically, this is commonly the case in modern deep networks. Theoretically, with sufficient overparameterization (Du et al., 2019) or high enough margin (Ji & Telgarsky, 2020), it is often easy to converge to a zero-loss solution.

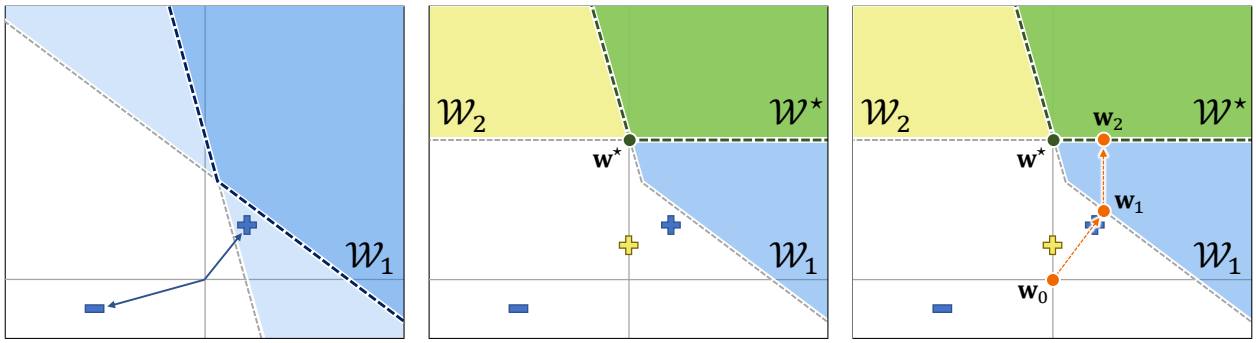
To facilitate our results and discussions, we specifically define the minimum-norm offline solution. This solution is traditionally linked to good generalization.

Definition 2.2 (Minimum-norm offline solution)**.** We denote the offline solution with the minimal norm by

$$\mathbf{w}^\star \triangleq \operatorname{argmin}_{\mathbf{w} \in \mathcal{W}^\star} \|\mathbf{w}\| .$$

Figures 1a and 1b illustrate our definitions. Notice how both the feasible set $\mathcal{W}_m$ of each task and the offline feasible set $\mathcal{W}^\star$ are closed, convex, and affine polyhedral cones.

Figure 1. Illustrating our setting from Section 2 and the Sequential Max-Margin dynamics from Scheme 2 in Section 3.



(a) A 2-dimensional task with two samples (one positive and one negative). Each sample $(\mathbf{x}, y)$ induces a constraint to a *half-space* $y\mathbf{w}^\top \mathbf{x} \geq 1$. The task's feasible set $\mathcal{W}_1$ is defined as the intersection of these two halfspaces, and is thus an affine polyhedral cone.

(b) Two tasks, the first having two samples and the second having a single sample. The intersection of $\mathcal{W}_1$ and $\mathcal{W}_2$ defines the *offline* feasible set $\mathcal{W}^\star$, in which all samples of all tasks are correctly classified with a margin of at least 1. Notice the min-norm offline solution $\mathbf{w}^\star \in \mathcal{W}^\star$.

(c) Learning the 1st task projects $\mathbf{w}_0 = \mathbf{0}_D$ onto $\mathcal{W}_1$. The obtained iterate $\mathbf{w}_1 = \mathbf{P}_1(\mathbf{w}_0)$ is the max-margin solution of the 1st task. Then, learning the 2nd task, projects $\mathbf{w}_1$ onto $\mathcal{W}_2$ to obtain (in this case) an offline solution $\mathbf{w}_2 = \mathbf{P}_2(\mathbf{w}_1) \in \mathcal{W}^\star \subset \mathcal{W}_2$. Notably, $\mathbf{w}_2$ is *not* the min-norm $\mathbf{w}^\star$ (Section 4.3).

3. Algorithmic Bias in Regularization Methods

Regularization methods are highly influential in continual learning (Kirkpatrick et al., 2017; Zenke et al., 2017; Aljundi et al., 2018). In this section, we propose a novel analysis for such methods on separable datasets in the spirit of theoretical work on algorithmic biases outside the scope of continual learning. Concretely, we discover that weakly-regularized models, trained sequentially to minimize the exponential loss,[1] converge in direction to the iterates of a sequential projection scheme.

Specifically, we study Scheme 1, in which the learner sequentially sees one task (out of T) at a time, for k iterations ($k > T$ implies repetitions). Starting from $\mathbf{w}_0^{(\lambda)} = \mathbf{0}_D$, at each iteration $t \in [k]$, the learner minimizes the exponential loss of the current task's dataset S_t, while biasing towards the previous task's solution $\mathbf{w}_{t-1}^{(\lambda)}$ using the Euclidean norm. The regularization strength is determined by a sequence of (possibly constant) positive scalars $\lambda_1, ..., \lambda_k > 0$. The norms are possibly weighted by a sequence of positive-definite matrices $\mathbf{B}_1, ..., \mathbf{B}_k \succ \mathbf{0}_{D\times D}$.

Scheme 1 Regularized Continual Learning

Initialization: $\mathbf{w}_0^{(\lambda)} = \mathbf{0}_D$
Iterative update for each task $t \in [k]$**:**

$$\mathbf{w}_t^{(\lambda)} = \underset{\mathbf{w}\in\mathbb{R}^D}{\operatorname{argmin}} \sum_{(\mathbf{x},y)\in S_t} e^{-y\mathbf{w}^\top\mathbf{x}} + \frac{\lambda_t}{2}\left\|\mathbf{w} - \mathbf{w}_{t-1}^{(\lambda)}\right\|_{\mathbf{B}_t}^2 \quad (1)$$

To clarify, each of the k iterations corresponds to learning a *whole* task (to convergence), and not to performing a single gradient step.

As a start, we first focus on *constant* strengths $\lambda > 0$ and on "vanilla" L2 regularization, *i.e.,* $\mathbf{B}_1 = ... = \mathbf{B}_k = \mathbf{I}$. Vanilla L2 regularization has recently been shown to be competitive with more popular norm-weighting schemes like EWC (Lubana et al., 2022; Smith et al., 2023). We address the more complicated setups in Section 5.

For an arbitrary λ, the objective in Eq. (1) is hard to analyze, even for the first task, where the regularization term reduces to the traditional unbiased term $\|\mathbf{w}\|^2$. Previous works were still able to perform non-trivial analysis by examining the weakly-regularized case, *i.e.,* in the limit of $\lambda \to 0$ (*e.g.,* for linear models (Rosset et al., 2004) or homogeneous neural networks (Wei et al., 2019)).

Under our continual setting, we also take this approach and analyze the weakly-regularized model. This enables us to gain analytical insights into regularization methods for continual learning. We establish an equivalence between the weakly-regularized Scheme 1 and the following Sequential Max-Margin Scheme 2, and propose novel perspectives and techniques for analyzing regularization methods.

Scheme 2 Sequential Max-Margin

Initialization: $\mathbf{w}_0 = \mathbf{0}_D$
Iterative update for each task $t \in [k]$**:**

$$\mathbf{w}_t = \mathbf{P}_t(\mathbf{w}_{t-1}) \triangleq \underset{\mathbf{w}\in\mathbb{R}^D}{\operatorname{argmin}} \|\mathbf{w} - \mathbf{w}_{t-1}\|^2 \quad (2)$$
$$\text{s.t.} \quad y\mathbf{w}^\top\mathbf{x} \geq 1,\ \forall(\mathbf{x},y)\in S_t$$

We are now ready to state our fundamental result, showing that the regularized continual iterates $(\mathbf{w}_t^{(\lambda)})$ of Scheme 1 converge in direction to the Sequential Max-Margin iterates $(\mathbf{w}_t) = (\mathbf{P}_t(\mathbf{w}_{t-1}))$ of Scheme 2, obtained by successive projections onto closed convex sets, as depicted in Figure 1c.

Theorem 3.1 (Weakly-regularized Continual Learning converges to Sequential Max-Margin)**.** *Let $\lambda_t = \lambda > 0$, and $\mathbf{B}_t = \mathbf{I}$, $\forall t \in [k]$. Then, for almost all separable datasets,*[2] *in the limit of $\lambda \to 0$, it holds that $\mathbf{w}_t^{(\lambda)} \to \ln\left(\frac{1}{\lambda}\right)\mathbf{w}_t$ with a residual of $\|\mathbf{w}_t^{(\lambda)} - \ln\left(\frac{1}{\lambda}\right)\mathbf{w}_t\| = \mathcal{O}\left(t\ln\ln\left(\frac{1}{\lambda}\right)\right)$. As a result, at any iteration $t = o\left(\frac{\ln(1/\lambda)}{\ln\ln(1/\lambda)}\right)$, we get*

$$\lim_{\lambda\to 0} \frac{\mathbf{w}_t^{(\lambda)}}{\|\mathbf{w}_t^{(\lambda)}\|} = \frac{\mathbf{w}_t}{\|\mathbf{w}_t\|}.$$

In Appendix B, we prove this theorem. There, we also discuss the limitations of our analysis (Appendix B.3) and identify aspects where it can be improved.

Remark 3.2 (Important differences from existing analysis). Existing works (*e.g.,* Rosset et al. (2004)) analyzed weakly-regularized models with an *unbiased* regularizer $\|\mathbf{w}\|_p^p$. This allowed them to concentrate on the limit *margin* which implies convergence in direction. However, we show that to analyze the *continual* regularizer in Eq. (1), one must also take into account the *scale* of the solutions $(\mathbf{w}_t^{(\lambda)})$. Therefore, we analyze both the scale *and* the direction of weakly-regularized solutions, requiring more refined techniques.

4. Sequential Max-Margin Projections

Now, we turn to exploit the connection we have established between the weakly-regularized continual Scheme 1 and the Sequential Max-Margin Scheme 2. Using tools from existing literature on Projections Onto Convex Sets (POCS), we gain valuable insights into the dynamics of training models continually. Specifically, we derive optimality guarantees and convergence bounds in several interesting settings.

[1] It should be possible to extend our results to other losses with exponential tails, *e.g.,* cross-entropy, as has been done in previous theoretical works on stationary settings (Soudry et al., 2018).

[2] This holds w.p. 1 for separable datasets (Assumption 2.1) sampled from *any* absolutely continuous distribution. It also holds even when the datasets are separable but not *jointly*-separable.

4.1. Quantities of Interest

We have three quantities of interest. While the first two are widely used in the POCS literature, the latter is specific to our continual classification setting.

Definition 4.1 (Quantities of interest)**.** Let $\mathbf{w}_t \in \mathbb{R}^D$ be the tth iterate, obtained while continually learning T jointly-separable tasks. The following quantities are of interest:

1. **Distance to the offline feasible set:** $d(\mathbf{w}_t, \mathcal{W}^\star)$
2. **Maximum dist. to *any* feasible set:** $\max_{m\in[T]} d(\mathbf{w}_t, \mathcal{W}_m)$
3. **Forgetting:** We define the forgetting of a previously-seen task $m \in [t]$ as the maximal squared hinge loss on *any* sample $(\mathbf{x}, y)$ of the task. More formally,

$$F_m(\mathbf{w}_t) \triangleq \max_{(\mathbf{x},y)\in S_m} \left(\max\left\{0,\, 1 - y\mathbf{w}_t^\top \mathbf{x}\right\}\right)^2 .$$

Throughout our paper, we analyze both the maximal and the average forgetting, *i.e.,* $\max_{m\in[t]} F_m(\mathbf{w}_t)$ and $\frac{1}{t}\sum_{m=1}^{t} F_m(\mathbf{w}_t)$.

Explaining our forgetting. Previous works on continual linear *regression* defined forgetting using the MSE on previously-seen tasks, *i.e.,* $\forall m < t\colon\ F_m(\mathbf{w}_t) = \frac{1}{|S_m|}\sum_{(\mathbf{x},y)\in S_m} \left(\mathbf{w}_t^\top \mathbf{x} - y\right)^2$ (*e.g.,* Doan et al. (2021); Evron et al. (2022)). Then, lower forgetting implies better *training* loss on previous tasks.

In continual linear *classification*, hinge losses capture similar properties. First, per our definitions, immediately after learning the mth task, the forgetting on it is 0, since $\forall \mathbf{v} \in \mathbb{R}^D : \mathbf{P}_m(\mathbf{v}) \in \mathcal{W}_m$ and $\forall \mathbf{u} \in \mathcal{W}_m : F_m(\mathbf{u}) = 0$. Moreover, a lower forgetting implies better training margins and generalization performance on previous tasks. Since the squared hinge loss is a surrogate for the 0-1 loss, $F_m(\mathbf{w}_t) < 1$ implies no prediction errors on the mth task.

Remark 4.2 (Forgetting vs. Regret). **Forgetting** is different from the **regret** used to analyze online learning algorithms (Crammer et al., 2006; Shalev-Shwartz et al., 2012; Hoi et al., 2021). Forgetting quantifies the degradation on previous tasks in hindsight, while regret cumulatively captures the ability to predict future datapoints (or tasks).

A favorable property of the quantities we defined, is that they bound each other.

Lemma 4.3 (Connecting quantities)**.** *Recall our definition of* $R \triangleq \max_{m\in[T]} \max_{(\mathbf{x},y)\in S_m} \|\mathbf{x}\|$. *The quantities of Def. 4.1 are related as follows,* $\forall \mathbf{w} \in \mathbb{R}^D$, $m \in [T]$*:*

$$F_m(\mathbf{w}) \le d^2(\mathbf{w}, \mathcal{W}_m) \max_{(\mathbf{x},y)\in S_m} \|\mathbf{x}\|^2 \le d^2(\mathbf{w}, \mathcal{W}^\star) R^2 .$$

Moreover, for the Sequential Max-Margin iterates $(\mathbf{w}_t)$ *of Scheme 2, all quantities are upper bounded by the "problem complexity", i.e.,* $d^2(\mathbf{w}_t, \mathcal{W}^\star) R^2 \le \|\mathbf{w}^\star\|^2 R^2$.

The proof is given in Appendix C.

Remark 4.4 (Problem complexity). Many of our bounds use $\|\mathbf{w}^\star\|^2 R^2$, which can be seen as the problem complexity due to its links to sample complexities (Novikoff, 1962). Moreover, Assumption 2.1, implies that $\|\mathbf{w}^\star\|^2 R^2 \ge 1$.

The following is a known useful result from the POCS literature (*e.g.,* Lemma 3 in Gubin et al. (1967)).

Lemma 4.5 (Monotonicity of distances to offline feasibility)**.** *Distances from the Sequential Max-Margin iterates of Scheme 2 to the offline feasible set are non-increasing, i.e.,*

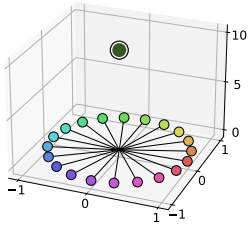
$$d(\mathbf{w}_t, \mathcal{W}^\star) \le d(\mathbf{w}_{t-1}, \mathcal{W}^\star), \ \ \forall t \in [k] .$$

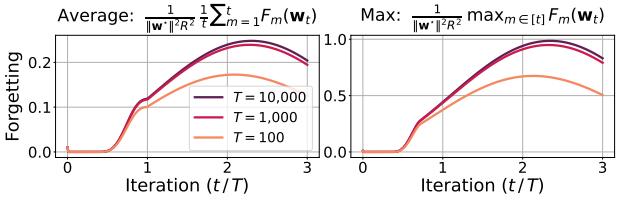
Question An immediate question arises from Lemma 4.5: when learning infinite jointly-separable tasks ($k = T \to \infty$), *must we converge to the offline feasible set* $\mathcal{W}^\star$*?* Next, we answer this question in the negative.

4.2. Adversarial Construction: Maximal Forgetting

Example 1 (Adversarial construction). We present a construction of task sequences that seemingly exhibit arbitrarily bad continual performance. Even after seeing $k = T \to \infty$ jointly-separable tasks, the learner stays afar from the offline feasible set $\mathcal{W}^\star$, *and* the forgetting of previously-seen tasks is maximal. The learner *fails* to successfully accumulate experience. See further details in Appendix C.1.

(a) Our construction with $T = 20$ tasks in $D = 3$. Datapoints (each defining a task) have a norm of $R = 1$ and are spread uniformly on a plane, slightly elevated above the xy-plane. It holds that $\|\mathbf{w}^\star\| = 10$. As $T \to \infty$, angles between consecutive tasks and applied projections, get smaller.





(b) The average and maximum forgetting for the adversarial construction for an increasing number of tasks T. Recall that the maximum forgetting lower bounds the distance to the offline feasible set (Lemma 4.3). Notably, after learning $T \to \infty$ jointly separable tasks, the quantities do *not* decay but seemingly become arbitrarily bad (close to $\|\mathbf{w}^\star\|^2 R^2$) at some point of learning.

Figure 2. Illustrations of our adversarial construction.

Remark 4.6 (Order of limits). Our paper analyzes the continual learning of T tasks for k iterations, possibly taking $k \to \infty$ (*e.g.,* by repeating tasks). We take $\lambda \to 0$ *after* fixing the number of iterations k. As a result, limit iterates hold

$$\frac{\mathbf{w}_\infty}{\|\mathbf{w}_\infty\|} \triangleq \lim_{k\to\infty} \frac{\mathbf{w}_k}{\|\mathbf{w}_k\|} = \lim_{k\to\infty} \lim_{\lambda\to 0} \frac{\mathbf{w}_k^{(\lambda)}}{\|\mathbf{w}_k^{(\lambda)}\|} .$$

4.3. Convergence to the Minimum-Norm Solution

When solving feasibility problems for classification (*e.g.,* in hard-margin SVM), the minimum-norm solution is also the max-margin solution. In turn, max-margin solutions are theoretically linked to better generalization performance.

In realizable continual linear regression (or generally, in alternating projections onto closed *subspaces*), it is known that if iterates converge to an offline (or globally feasible) solution, then that solution must be the closest to $\mathbf{w}_0$, *i.e.,* have a minimum norm (Evron et al., 2022; Halperin, 1962).

In contrast, in separable continual linear classification settings like ours (or generally, in projections onto closed convex sets), there is no such guarantee and we can converge to a suboptimal offline solution, as depicted in Figure 1c.

Nevertheless, the following optimality guarantee *does* hold.

Theorem 4.7 (Optimality guarantee)**.** *Any iterate $\mathbf{w}_t$ obtained by Scheme 2 holds $\|\mathbf{w}_t\| \leq 2\|\mathbf{w}^\star\|$, where $\mathbf{w}^\star$ is the minimum-norm offline solution (Def. 2.2).*

If additionally, $\mathbf{w}_t$ is "offline"-feasible, i.e., $\mathbf{w}_t \in \mathcal{W}^\star$, then

$$\|\mathbf{w}^\star\| \leq \|\mathbf{w}_t\| \leq 2\|\mathbf{w}^\star\| \ .$$

The proof is given in Appendix C.2. Further comparisons to continual regression are drawn in Appendix A.

4.4. Recurring Tasks

Naturally, in many practical continual problems, certain concepts and experiences recur at different tasks (*e.g.,* environments of an autonomous vehicle, levels of a computer game, trends of a search engine, etc.).

Several recent papers have observed empirically that task repetitions mitigate catastrophic forgetting in continual learning (Stojanov et al., 2019; Cossu et al., 2022), even when training is performed with vanilla SGD, without any forgetting-preventing method (Lesort et al., 2022).

In this section, we analytically study the influence of repetitions. To accomplish this, we leverage the connection that we have established between continual learning and successive projection algorithms for convex feasibility problems. Importantly, we *do not* propose repetitions as a training method but rather aim to understand their effects on continual learning from a projection perspective.

The results of this section are summarized in Table 1.

Recurring tasks vs. Batch learning The recurring tasks setting should not be confused with standard batch training (*i.e.,* by regarding each task as a batch). While in batch training, a *single* gradient-descent step is made for each batch, in our continual learning setting (Scheme 1) each task is solved completely (to separation).

The following is a key lemma in our paper. Much of the research on POCS has focused on defining and applying linear regularity conditions, under which iterates $(\mathbf{w}_t)$ converge linearly (like c^t for some $c \in [0,1)$) to the feasible sets' intersection $\mathcal{W}^\star$ (*e.g.,* Bauschke & Borwein (1993)). In realizable continual linear *regression*, Evron et al. (2022) showed that no such regularity holds, and while their forgetting, *i.e.,* $F_m(\mathbf{w}_t) \triangleq d^2(\mathbf{w}_t, \mathcal{W}_m)$, $\forall m \in [T]$, is upper bounded *universally*, no universal bounds can be derived for $d^2(\mathbf{w}_t, \mathcal{W}^\star)$ even when $\|\mathbf{w}^\star\|$ and R are bounded. In contrast, we prove that in separable continual linear *classification*, linear regularity *does* hold and nontrivial bounds *can* be derived for $d^2(\mathbf{w}_t, \mathcal{W}^\star)$.

Lemma 4.8 (Linear regularity of Sequential Max-Margin)**.** *At the tth iteration, the distance to the offline feasible set is tied to the distance to the farthest feasible set of any* specific *task. Specifically, it holds that $\forall t \in \mathbb{N}^+$,*

$$d^2(\mathbf{w}_t, \mathcal{W}^\star) \leq 4\|\mathbf{w}^\star\|^2 R^2 \max_{m \in [T]} d^2\big(\mathbf{w}_t, \mathcal{W}_m\big) \ .$$

The proofs for this section are given in Appendix D.

So far we assumed that at iteration t, the learner solves a task, *i.e.,* a dataset S_t (out of T possible datasets), corresponding to a projection $\mathbf{P}_t$. To facilitate our next results for cases where tasks recur, we now define task ordering functions.

Definition 4.9 (Task ordering)**.** A task ordering is a function

$$\tau : \mathbb{N}^+ \to [T]$$

that maps an iteration to a learning task.

4.4.1. Cyclic Ordering

Mathematically, when analyzing successive projections (like in our Scheme 2), it is common to first study a cyclic setting, where the projections form a clearer (cyclic) operator. This setting has been at the center of focus in many theoretical papers (*e.g.,* Agmon (1954); Halperin (1962); Deutsch & Hundal (2006a); Borwein et al. (2014); Evron et al. (2022)).

Practically, in continual learning, cyclic orderings are indeed less flexible. However, we believe that they *do* emerge naturally in real-world scenarios. For instance, virtual assistants support the recurring daily routines of their customers and should be able to continue learning from new experiences.

Definition 4.10 (Cyclic task ordering)**.** A cyclic ordering over T tasks is defined as

$$\tau(t) \triangleq 1 + ((t-1) \bmod T) \ , \quad \forall t \in \mathbb{N}^+ \ ,$$

and induces a cyclic operator, in the sense that

$$\mathbf{w}_{nT} = \big(\mathbf{P}_{\tau(nT)} \circ \cdots \circ \mathbf{P}_{\tau(1)}\big)(\mathbf{w}_0) = (\mathbf{P}_T \circ \cdots \circ \mathbf{P}_1)^n(\mathbf{w}_0).$$

We illustrate such orderings in the following Figure 3.

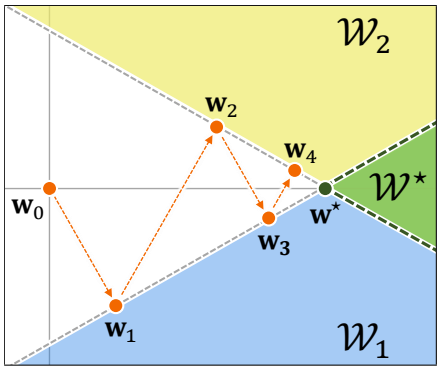


Figure 3. A cyclic setting with two tasks. Each time we solve a task, we project the previous iterate onto its corresponding convex set.

Using the linear regularity from Lemma 4.8, we prove linear convergence results for cyclic orderings. Our proofs use tools and results from existing POCS literature.

Lemma 4.11 (Limit guarantees for cyclic orderings)**.** *Under a cyclic ordering (and the separability assumption 2.1), the iterates converge to a 2-optimal $\mathbf{w}_\infty \in \mathcal{W}^\star$. That is,*

$$\lim_{k\to\infty} d(\mathbf{w}_k, \mathcal{W}^\star) = 0, \quad \|\mathbf{w}^\star\| \le \|\mathbf{w}_\infty\| \le 2\,\|\mathbf{w}^\star\| \ .$$

The proofs for this section are given in Appendix D.3.

Theorem 4.12 (Linear rates for cyclic orderings)**.** *For $T \ge 2$ jointly-separable tasks learned cyclically, after $k = nT$ iterations (n cycles), our quantities of interest (Def. 4.1) converge linearly as*

$$\underbrace{\max_{m\in[T]} F_m(\mathbf{w}_k)}_{\text{Maximum forgetting}} \le \underbrace{\max_{m\in[T]} d^2(\mathbf{w}_k, \mathcal{W}_m)}_{\text{Max. dist. to any feasible set}} R^2 \le$$

$$\le \underbrace{d^2(\mathbf{w}_k, \mathcal{W}^\star)}_{\substack{\text{Dist. to}\\ \text{offline feasible set}}} R^2 \le g(k)\,\|\mathbf{w}^\star\|^2 R^2 \ ,$$

where

$$g(k) \triangleq \begin{cases} \exp\left(-\frac{k}{4\|\mathbf{w}^\star\|^2 R^2}\right) & T = 2 \\ 4\exp\left(-\frac{k}{16T^2\|\mathbf{w}^\star\|^2 R^2}\right) & T \ge 3 \end{cases}$$

and $\|\mathbf{w}^\star\|^2 R^2 \ge 1$ is the problem complexity (Rem. 4.4).

Detour: Universal rates for *general* cyclic Projections onto Convex Sets (POCS) settings. We take a brief detour to derive universal bounds for general POCS settings, where problems do not necessarily hold regularity conditions like the ones in our Lemma 4.8. While cyclic POCS settings have been studied for decades (Agmon, 1954; Deutsch & Hundal, 2006a;b; 2008), most previous works either focused on settings with regularity assumptions or yielded problem-dependent rates that can be arbitrarily bad.

Our result below extends the universal results of a recent work that focused on closed *subspaces* only (Evron et al., 2022); and is also related to a recent work that considered closed *convex sets* as well, but dismissed the number of sets T as $\mathcal{O}(1)$ (Reich & Zalas, 2023).

Proposition 4.13 (Universal rates for general cyclic POCS)**.** *Let $\mathcal{W}_1, \ldots, \mathcal{W}_T$ be closed convex sets with nonempty intersection $\mathcal{W}^\star$. Let $\mathbf{w}_k = (\mathbf{P}_T \circ \cdots \circ \mathbf{P}_1)^n(\mathbf{w}_0)$ be the iterate after $k = nT$ iterations (n cycles) of cyclic projections onto these convex sets. Then, the maximal distance to any (specific) convex set, is upper bounded* universally *as,*

$$\textit{For } T = 2\textit{:} \quad \max_{m\in[T]} d^2(\mathbf{w}_k, \mathcal{W}_m) \le \frac{1}{k+1}\, d^2(\mathbf{w}_0, \mathcal{W}^\star) \ ,$$

$$\textit{For } T \ge 3\textit{:} \quad \max_{m\in[T]} d^2(\mathbf{w}_k, \mathcal{W}_m) \le \frac{2T^2}{\sqrt{k}}\, d^2(\mathbf{w}_0, \mathcal{W}^\star) \ .$$

Clearly, the above yields universal rates for the forgetting as well (see Lemma 4.3), but these are essentially worse than the linear rates we got (in the presence of regularity).

4.4.2. Random Ordering

In this section, we consider a uniform i.i.d. task ordering over a set of T tasks. Random orderings are considered more realistic than cyclic ones (*e.g.,* driverless taxis will likely encounter recurring routes/environments randomly, rather than in a certain cycle). Mathematically, they also require other analytical tools and often offer different convergence guarantees. Much like cyclic orderings, random orderings have been studied in many related areas (*e.g.,* Nedić (2010); Needell & Tropp (2014); Evron et al. (2022)).

Table 1. Summary of our upper bounds for recurring orderings (Section 4.4). Upper bounds of random orderings apply to the *expectations*.

Ordering	Iterate type	$\frac{1}{\|\mathbf{w}^\star\|^2 R^2}\max_{m\in[T]} F_m(\mathbf{w}_k)$	$\frac{1}{\|\mathbf{w}^\star\|^2}\max_{m\in[T]} d^2(\mathbf{w}_k, \mathcal{W}_m)$	$\frac{1}{\|\mathbf{w}^\star\|^2} d^2(\mathbf{w}_k, \mathcal{W}^\star)$
Cyclic	Last ($T = 2$)	$\min\left\{\frac{1}{k+1}, \exp\left(-\frac{k}{4\|\mathbf{w}^\star\|^2 R^2}\right)\right\}$		$\exp\left(-\frac{k}{4\|\mathbf{w}^\star\|^2 R^2}\right)$
	Last ($T \ge 3$)	$\min\left\{\frac{2T^2}{\sqrt{k}}, 4\exp\left(-\frac{k}{16T^2\|\mathbf{w}^\star\|^2 R^2}\right)\right\}$		$\exp\left(-\frac{k}{16T^2\|\mathbf{w}^\star\|^2 R^2}\right)$
	Average ($T \ge 2$)	T^2/k		—
Random (i.i.d.)	Last	$\exp\left(-\frac{k}{4T\|\mathbf{w}^\star\|^2 R^2}\right)$		$\exp\left(-\frac{k}{4T\|\mathbf{w}^\star\|^2 R^2}\right)$
	Average	T/k		—

Definition 4.14 (Random task ordering)**.** A random (uniform) ordering over T tasks is defined as

$$\Pr\left(\tau(t)=m\right)=\Pr\left(\tau(t')=m'\right),\ \ \forall t,t'\in\mathbb{N}^{+},\ m,m'\in[T].$$

Theorem 4.15 (Linear rates for random orderings)**.** *For T jointly-separable tasks learned in a random ordering, our quantities of interest (Def. 4.1) converge linearly as*

$$\mathop{\mathbb{E}}_{\tau}\Big[\underbrace{\max_{m\in[T]} F_m(\mathbf{w}_k)}_{\text{Maximum forgetting}}\Big] \le \mathop{\mathbb{E}}_{\tau}\Big[\underbrace{\max_{m\in[T]} d^2(\mathbf{w}_k,\mathcal{W}_m)}_{\text{Max. dist. to any feasible set}}\Big]R^2 \le$$

$$\le \mathop{\mathbb{E}}_{\tau}\Big[\underbrace{d^2(\mathbf{w}_k,\mathcal{W}^\star)}_{\substack{\text{Dist. to}\\ \text{offline feasible set}}}\Big]R^2 \le \exp\Big(-\tfrac{k}{4T\|\mathbf{w}^\star\|^2R^2}\Big)\,\|\mathbf{w}^\star\|^2R^2.$$

where $\|\mathbf{w}^\star\|^2R^2\ge 1$ is the problem complexity (Rem. 4.4).

To show this theorem, we prove a property of *our* convex sets (*i.e.,* polyhedral cones; see Figure 1a) and apply a result from the (random) POCS literature (Nedić, 2010). The proofs for this section are given in Appendix D.4.

Lemma 4.16 (Limit guarantees of random orderings)**.** *Under a random ordering (and assumption 2.1), the iterates converge* almost surely *to $\mathcal{W}^\star$, such that*

$$\lim_{k\to\infty} d(\mathbf{w}_k,\mathcal{W}^\star)=0,\quad \|\mathbf{w}^\star\|\le\|\mathbf{w}_\infty\|\le 2\,\|\mathbf{w}^\star\|\ .$$

Remark 4.17 (Beyond uniform distributions). The result from Nedić (2010) that we used to prove Theorem 4.15, is applicable to any i.i.d. distribution with a nonzero $p_{\min}\triangleq\min_{m\in[T]}\Pr\left(\tau\left(\cdot\right)=m\right)>0$. In such cases, the rate from Theorem 4.15 changes to $\exp\Big(-\frac{p_{\min}k}{4\|\mathbf{w}^\star\|^2R^2}\Big)$.

4.4.3. Average Iterate Analysis

So far, we analyzed the convergence of the last iterate $\mathbf{w}_k$. Next, we analyze the average iterate $\overline{\mathbf{w}}_k\triangleq\frac{1}{k}\sum_{t=1}^{k}\mathbf{w}_t$. Such analysis often allows for stronger bounds (*e.g.,* in SGD (Shamir & Zhang, 2013), Kaczmarz methods (Morshed et al., 2022), and continual regression (Evron et al., 2022)).

Proposition 4.18 (Universal rates for the average iterate)**.** *After n cycles under a cyclic ordering ($k=nT$) we have*

$$\underbrace{\max_{m\in[T]} F_m(\overline{\mathbf{w}}_k)}_{\text{Maximum forgetting}} \le \underbrace{\max_{m\in[T]} d^2(\overline{\mathbf{w}}_k,\mathcal{W}_m)}_{\text{Max. dist. to any feasible set}} R^2 \le \frac{T^2}{k}\|\mathbf{w}^\star\|^2R^2$$

and after k iterations under a random ordering we have

$$\underbrace{\mathop{\mathbb{E}}_{\tau}\Big[\tfrac{1}{T}\sum_{m=1}^{T}F_m(\overline{\mathbf{w}}_k)\Big]}_{\text{Average forgetting}} \le \underbrace{\mathop{\mathbb{E}}_{\tau}\Big[\tfrac{1}{T}\sum_{m=1}^{T}d^2(\overline{\mathbf{w}}_k,\mathcal{W}_m)\Big]}_{\text{Avg. distance to feasible sets}}R^2 \le \tfrac{\|\mathbf{w}^\star\|^2R^2}{k}$$

(implying a $\frac{T}{k}\|\mathbf{w}^\star\|^2R^2$ bound on the expected maximum *forgetting and* maximum *distance to any feasible set).*

The proof is given in Appendix D.5.

5. Extensions

5.1. Regularization Strength Scheduling

Up to this point, we assumed that the regularization strengths in Scheme 1 are constant, *i.e.,* $\lambda_1=\ldots=\lambda_k\triangleq\lambda$. Alternatively, these strengths can *vary*, as done in practice in stationary settings (Lewkowycz & Gur-Ari, 2020). Related work in continual learning (Mirzadeh et al., 2020) used a learning-rate decay scheme, which can be seen as a form of varying regularization. We now aim to understand the bias and implications of varying regularization strengths.

Given $\lambda>0$, we parameterize the regularization strengths as $\lambda_t\triangleq\lambda_t(\lambda)>0$ for arbitrary functions $\lambda_t:\mathbb{R}_{>0}\to\mathbb{R}_{>0}$ holding that $\lim_{\lambda\to0}\lambda_t(\lambda)=0$ and $\lim_{\lambda\to0}\frac{\ln\lambda_{t-1}}{\ln\lambda_t}<\infty$ is well-defined $\forall t\in[k]$. We show that the weakly-regularized Scheme 1 (with $\lambda_1,\ldots,\lambda_k$) converges to the scheme below.

Scheme 3 Scheduled Sequential Max-Margin

Initialization: $\mathbf{w}_0=\mathbf{0}_D$

Iterative update $\forall t\in[k]$**:** $\mathbf{w}_t=\mathbf{P}_t\Big(\big(\lim_{\lambda\to0}\frac{\ln\lambda_{t-1}}{\ln\lambda_t}\big)\mathbf{w}_{t-1}\Big)$

Theorem 5.1 (Weakly-regularized models with scheduling)**.** *For almost all separable datasets, when scheduling the regularization strength as described above, it holds that*

$$\lim_{\lambda\to0}\frac{\mathbf{w}_t^{(\lambda)}}{\|\mathbf{w}_t^{(\lambda)}\|}=\frac{\mathbf{w}_t}{\|\mathbf{w}_t\|},\quad\forall t\in[k]\,.$$

The proofs for this section are given in Appendix E & E.1.

Specifically, using a **double exponential scheduling** rule,

$$\lambda_t=\lambda_{t-1}^{p}\,(=\lambda_1^{p^{t-1}}=\lambda^{p^{t-1}})\,,\quad\text{for some } p>0\,,\qquad(3)$$

the update rule in Scheme 3 becomes $\mathbf{w}_t=\mathbf{P}_t\big(\frac{1}{p}\mathbf{w}_{t-1}\big)$. We illustrate this in Figure 4a.

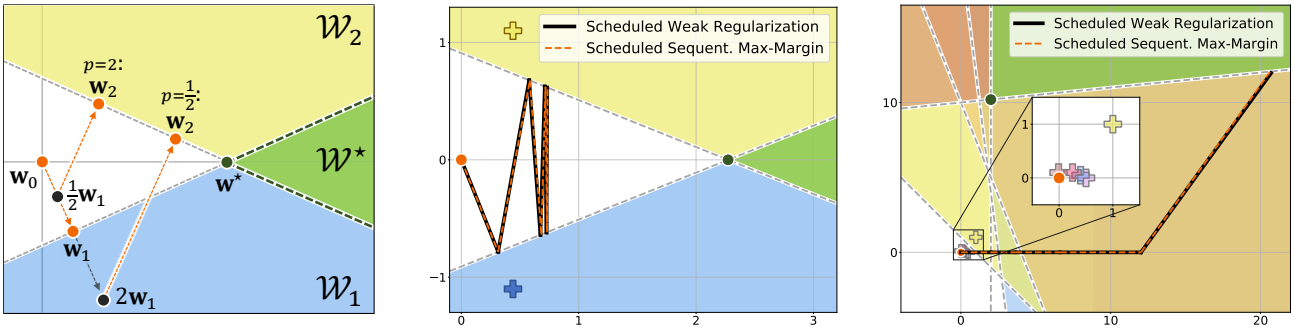
In this case, in contrast to the $p=1$ case, even when tasks recur we may not converge near the min-norm solution or the feasibility set, as demonstrated in the next examples.

Example 2 ($p>1\implies$ Possibly $\mathbf{w}_\infty\notin\mathcal{W}^\star$). In Figure 4b, we cycle between two tasks using Schemes 1 and 3 with $p=2$. The strengths *decay* like $\lambda,\lambda^2,\lambda^4,\lambda^8$, and so on. Note that both schemes agree. Unlike the constant λ case (Thm. 4.12), we do *not* converge to the offline set $\mathcal{W}^\star$.

More generally, we prove that when $p>1$, the limit distance from the offline feasible set $\mathcal{W}^\star$ can be arbitrarily bad.

Proposition 5.2. *There exists a construction of two jointly-separable tasks with $R=1$, in which the iterates of the cyclic ordering do not converge to $\mathcal{W}^\star$, for any $p>1$. Specifically, for any $p>1$, $\|\mathbf{w}^\star\|>1$, it holds that*

$$\lim_{k\to\infty}\frac{d(\mathbf{w}_k,\mathcal{W}^\star)}{\|\mathbf{w}^\star\|}=\frac{\|\mathbf{w}^\star\|^2\,(p-1)}{2+\|\mathbf{w}^\star\|^2\,(p-1)}\sqrt{1-\tfrac{1}{\|\mathbf{w}^\star\|^2}}\,.$$



(a) Two iterations of the double exponential scheduler with two values $p = 1/2, 2$. $p > 1$ shrinks $\mathbf{w}_1$ before projecting it onto $\mathcal{W}_2$, while $p < 1$ inflates it.

(b) $p > 1 \implies$ Might not converge to $\mathcal{W}^\star$. Here, $p{=}2$, $\lambda_1{=}\lambda{=}10^{-32}$, $\lambda_6{=}10^{-1024}$. Black lines indicate changes in the *directions* of the weakly-regularized solutions (their actual scale is much larger).

(c) $p > 1 \implies$ Possibly $\|\mathbf{w}_t\| > 2\|\mathbf{w}^\star\|$ (the guarantee from Thm. 4.7 does not hold). Here, $p = 0.55$, $\lambda_1 = \lambda = 10^{-4096}$ and the rest $\lambda_2, \lambda_3, \lambda_4, \lambda_5$ are $10^{-2253}, 10^{-1239}, 10^{-681}, 10^{-375}$.

Figure 4. Illustrations and experiments for the double exponential regularization scheduling (Eq. (3) in Section 5.1).

Example 3 ($p < 1 \implies$ No optimality guarantees on $\mathbf{w}_t$). In Figure 4c, we run both Schemes 1 and 3 with $p = 0.55$ for one "pass" over the tasks. The strengths *increase* like $\lambda, \lambda^{0.55}, \lambda^{0.303}, \lambda^{0.166}$, and so on (however, our analysis still requires that $\lambda \to 0$; see Remark E.1). Note that the iterates of both schemes agree. Unlike the guarantees for the constant λ case (Theorem 4.7), here $\|\mathbf{w}_5\| > 2\|\mathbf{w}^\star\|$.

Roughly speaking, adversarial placements of examples (*i.e.*, feasible sets) can make the iterates $(\mathbf{w}_t)$ grow like $\mathcal{O}(p^{-t})$, even in an almost orthogonal direction to the min-norm $\mathbf{w}^\star$.

Remark 5.3. We analyzed the double exponential rule (3) since "milder" rules (*e.g.*, $\lambda_t{=}c^t\lambda$ for $c{=}\mathcal{O}(1)$) do not influence Scheme 3, thus not affecting Scheme 1 when $\lambda{\to}0$. Even a rule like $\lambda_t{=}\lambda^t$ implies $\lim_{\lambda\to 0}\frac{\ln\lambda_{t-1}}{\ln\lambda_t}{=}1{-}\frac{1}{t}$ and becomes insignificant as t increases.

5.2. Weighted Regularization

So far, we analyzed *unweighted* regularizers in Scheme 1. Most regularization methods for continual learning employ *weighted* norms (*e.g.*, Kirkpatrick et al. (2017); Zenke et al. (2017); Aljundi et al. (2018)). Now, we focus on such norms and link them to Sequential Max-Margin when $\lambda \to 0$.

Given $\mathbf{B}_1, \ldots, \mathbf{B}_k \succ \mathbf{0}_{D\times D}$, we define the following scheme.

Scheme 4 Weighted Sequential Max-Margin

Initialization: $\mathbf{w}_0 = \mathbf{0}_D$

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

$$\mathbf{w}_t = \mathbf{P}_t(\mathbf{w}_{t-1}) \triangleq \operatorname{argmin}_{\mathbf{w}} \|\mathbf{w} - \mathbf{w}_{t-1}\|^2_{\mathbf{B}_t} \quad \text{s.t.} \quad y\mathbf{w}^\top\mathbf{x} \geq 1,\ \forall(\mathbf{x}, y) \in S_t$$

Theorem 5.4 (Weak weighted regularization)**.** *For almost all separable datasets and bounded weighting schemes (holding* $0{<}\mu{\leq}\sigma_{\min}(\mathbf{B}_t){\leq}\sigma_{\max}(\mathbf{B}_t){\leq}M{<}\infty$*), in the limit of* $\lambda \to 0$*, Schemes 1 and 4 coincide. That is,* $\forall t{\in}[k]$*, it holds that* $\lim_{\lambda\to 0}\frac{\mathbf{w}_t^{(\lambda)}}{\|\mathbf{w}_t^{(\lambda)}\|} = \frac{\mathbf{w}_t}{\|\mathbf{w}_t\|}$.

The proofs for this section are given in Appendix E & E.2.

Practically, most regularization methods use weighting matrices based on Fisher information (FI) (Benzing, 2022). We present a novel result showing that in continual *regression*, such weighting schemes *prevent* forgetting (thus forming an "ideal continual learner" as defined by Peng et al. (2023)).

Proposition 5.5. *Using a Fisher-information-based weighting scheme of* $\mathbf{B}_t = \sum_{i=1}^{t-1}\sum_{\mathbf{x}\in S_i}\mathbf{x}\mathbf{x}^\top$*, there is no forgetting in (realizable) continual linear*[3] *regression.*

In contrast, below we show a simple example where such weighting schemes, even when using the full FI matrices, do not prevent forgetting in continual linear *classification*.

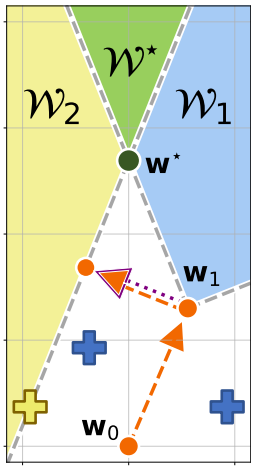


Figure 5. A 2-dimensional setting with solely two orthogonal examples $\mathbf{x}_1 \perp \mathbf{x}_2$ in the 1st task. Here, the iterate $\mathbf{w}_1$ holds $\mathbf{w}_1^\top\mathbf{x}_1{=}\mathbf{w}_1^\top\mathbf{x}_2{=}1$, implying a FI matrix of $\mathbf{B}_2{\propto}\sum_{i\in[2]}\mathbf{x}_i\mathbf{x}_i^\top{=}\mathbf{I}$. Then, FI weighting (dotted) reverts to vanilla L2 regularization (dashed) which does not prevent forgetting. In contrast, the FI matrices in high dimensions are often non-invertible. Then, the learner can move freely in directions orthogonal to previous data ($\|\cdot\|_{\mathbf{B}_t}$ is no longer a "proper" norm), thus avoiding forgetting. See further examples in App. E.2.1.

An interesting question arises: Are there general weighting schemes that can prevent forgetting in Scheme 4? We find this to be an exciting direction for future research.

[3] A similar guarantee also applies more broadly, for linear networks of any depth and non-linear networks in the NTK regime.

6. Related Work

Many papers have investigated various aspects of catastrophic forgetting, including when it occurs (Evron et al., 2022), strategies to avoid it (Peng et al., 2023), the influence of task similarity (Lee et al., 2021), its impact on transferability (Chen et al., 2023), and other related factors. A sound understanding of forgetting can potentially advance the continual learning field significantly.

Throughout our paper, we discussed many connections to other works from many fields. Interestingly, our Sequential Max-Margin (SMM) scheme can be seen as a *hard* variant of the Adaptive SVM algorithm (Yang et al., 2007), which was previously used to *practically* tackle domain adaptation and transfer learning. (*e.g.,* in Li (2007); Pentina et al. (2015); Tang et al. (2022)). To the best of our knowledge, our paper is the first to highlight the connection between regularization methods for continual learning and Adaptive SVM.

For the special case where each task has one datapoint, the SMM scheme corresponds to the online Passive-Aggressive algorithm (Crammer et al., 2006). If additionally, we consider only two tasks learned in a cyclic ordering, the regret after training task t is related (but not identical) to the forgetting (see Remark 4.2). Consequently, our forgetting bound in Prop. 4.13 matches the regret bound in their Theorem 2.

7. Discussion and Future Work

7.1. Early Stopping

When training a *single task*, it is well known that explicit regularization is related to early stopping. In linear classification with an exponential loss, Rosset et al. (2004) showed that the regularization path converges (in the $\lambda \to 0$ limit) to the hard-margin SVM, which is also the limit of the optimization path (Soudry et al., 2018). In regression, Ali et al. (2019) connected early stopping to ridge regression.

Contrastingly, when learning a *sequence of tasks*, explicit regularization and early stopping can behave differently. We first discuss a case where *both* methods lead to SMM. More generally, we demonstrate that solutions may differ.

Consider a task sequence with 1 sample $(\mathbf{x}_t, y_t)$ per task.[4] We minimize the exponential loss of each task t with gradient flow (GF) until $L_t(\mathbf{w}) \triangleq \exp(-y_t \mathbf{x}_t^\top \mathbf{w}) = \epsilon$ for a fixed $\epsilon < 1$, yielding a predictor $\mathbf{w}_t^{(\epsilon)}$. Denote the "normalized" predictor by $\hat{\mathbf{w}}_t^{(\epsilon)} = \mathbf{w}_t^{(\epsilon)} / \ln(1/\epsilon)$. Since GF stays in the data span, we have $\mathbf{w}_t^{(\epsilon)} = \mathbf{w}_{t-1}^{(\epsilon)} + \alpha_t y_t \mathbf{x}_t$ for some $\alpha_t > 0$. The early stopping implies $y_t \mathbf{x}_t^\top \mathbf{w}_t^{(\epsilon)} = \ln(1/\epsilon)$, and thus $y_t \mathbf{x}_t^\top \hat{\mathbf{w}}_t^{(\epsilon)} = 1$. Hence, $\hat{\mathbf{w}}_t^{(\epsilon)}$ holds the KKT conditions of the SMM, *i.e.,* $\hat{\mathbf{w}}_t^{(\epsilon)} = \operatorname{argmin}_{\mathbf{w}} \left\| \mathbf{w} - \hat{\mathbf{w}}_{t-1}^{(\epsilon)} \right\|^2$ s.t. $y_t \mathbf{x}_t^\top \mathbf{w} \geq 1$ (where $\hat{\alpha}_t \triangleq (\alpha_t / \ln(1/\epsilon)) > 0$ is the dual variable).[5] In the general case (*i.e.,* more than one sample per task), early stopping might *not* agree with SMM, as depicted below.

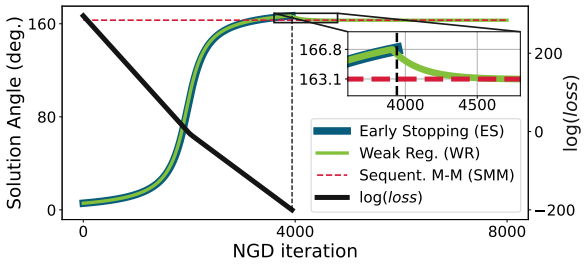


Figure 6. Comparing the weakly-regularized (WR) and early stopping (ES) schemes. We train 2 tasks: $S_1 = \{(10, 1)\}$, $S_2 = \{(-10, 1), (-15, 0.5)\}$ (where all points are labeled $+1$). We plot the angle of different predictors ($\angle(\mathbf{w}, \left[\begin{smallmatrix}1\\0\end{smallmatrix}\right])$) while training on the 2nd task. For instance, the Sequential Max-Margin (SMM) solution for that task is $163.1°$. For ES, we train both tasks with normalized GD (NGD) until their loss is $\epsilon = e^{-200}$. Importantly, ES *stops* when the loss of the 2nd task, shown in the secondary y-axis, is e^{-200} and the angle is $166.8°$. For WR, we set $\lambda = e^{-200}$ and solve Eq. (1) by running NGD to convergence. For the 2nd task, we initialize the NGD by $\mathbf{w}_1^{(\lambda)}$. Unlike ES, the WR solution *does* converge to $163.1°$ like SMM, as guaranteed by Theorem 3.1. Further details in App. F.

7.2. Future Work

There are several interesting avenues for future work, *e.g.,* extending our results to non-separable data (perhaps in the spirit of Yang et al. (2007)), multiclass classification with cross-entropy loss (similarly to Appendix 4.1 in Soudry et al. (2018)), or non-linear models. One can also try to derive forgetting bounds for weighted regularization schemes and look for optimal weighting matrices. Another challenging but rewarding avenue is to extend our analysis to *finite* regularization strengths (λ_t). Finally, it is interesting to understand the exact algorithmic bias for early stopping in continual learning and its relation to explicit regularization.

Acknowledgements

We thank Lior Alon (MIT) for the 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 Schmidt Career Advancement Chair in AI. NS was partially supported by the Simons Foundation and NSF-IIS/CCF awards.

[4] Equivalently, 2 samples $(\mathbf{x}_t, y_t)$ and $(-\mathbf{x}_t, -y_t)$.

[5] If iteration t begins and we already have $L_t(\mathbf{w}_{t-1}^{(\epsilon)}) < \epsilon$ and $y_t \mathbf{x}_t^\top \hat{\mathbf{w}}_{t-1}^{(\epsilon)} \geq 1$, both ES and SMM will not change their solutions.

References

Agmon, S. The relaxation method for linear inequalities. *Canadian Journal of Mathematics*, 6:382–392, 1954. (cited on p. 5, 6)

Ali, A., Kolter, J. Z., and Tibshirani, R. J. A continuous-time view of early stopping for least squares regression. In *AISTATS*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1370–1378. PMLR, 2019. (cited on p. 9)

Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., and Tuytelaars, T. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 139–154, 2018. (cited on p. 3, 8)

Asanuma, H., Takagi, S., Nagano, Y., Yoshida, Y., Igarashi, Y., and Okada, M. Statistical mechanical analysis of catastrophic forgetting in continual learning with teacher and student networks. *Journal of the Physical Society of Japan*, 90(10):104001, Oct 2021. (cited on p. 1)

Bauschke, H. H. Projection algorithms: results and open problems. In *Studies in Computational Mathematics*, volume 8, pp. 11–22. Elsevier, 2001. (cited on p. 29)

Bauschke, H. H. and Borwein, J. M. On the convergence of von neumann's alternating projection algorithm for two sets. *Set-Valued Analysis*, 1(2):185–212, 1993. (cited on p. 5, 27, 29)

Bauschke, H. H., Combettes, P. L., et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011. (cited on p. 26)

Belkin, M., Ma, S., and Mandal, S. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pp. 541–549. PMLR, 2018. (cited on p. 1)

Bennani, M. A., Doan, T., and Sugiyama, M. Generalisation guarantees for continual learning with orthogonal gradient descent. *arXiv preprint arXiv:2006.11942*, 2020. (cited on p. 1)

Benzing, F. Unifying regularisation methods for continual learning. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022. (cited on p. 8)

Borwein, J. M., Li, G., and Yao, L. Analysis of the convergence rate for the cyclic projection algorithm applied to basic semialgebraic convex sets. *SIAM Journal on Optimization*, 24(1):498–527, 2014. (cited on p. 5)

Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004. (cited on p. 14)

Chen, J., Nguyen, T., Gorur, D., and Chaudhry, A. Is forgetting less a good inductive bias for forward transfer? In *The Eleventh International Conference on Learning Representations*, 2023. (cited on p. 9)

Cossu, A., Graffieti, G., Pellegrini, L., Maltoni, D., Bacciu, D., Carta, A., and Lomonaco, V. Is class-incremental enough for continual learning? *Frontiers in Artificial Intelligence*, 5, 2022. (cited on p. 5)

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., and Singer, Y. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7(19):551–585, 2006. (cited on p. 4, 9)

Deutsch, F. and Hundal, H. The rate of convergence for the cyclic projections algorithm i: Angles between convex sets. *Journal of Approximation Theory*, 142(1):36–55, 2006a. (cited on p. 5, 6, 28, 30)

Deutsch, F. and Hundal, H. The rate of convergence for the cyclic projections algorithm ii: norms of nonlinear operators. *Journal of Approximation Theory*, 142(1):56–82, 2006b. (cited on p. 6, 26)

Deutsch, F. and Hundal, H. The rate of convergence for the cyclic projections algorithm iii: Regularity of convex sets. *Journal of Approximation Theory*, 155(2):155–184, 2008. (cited on p. 6, 27, 29)

Doan, T., Abbana Bennani, M., Mazoure, B., Rabusseau, G., and Alquier, P. A theoretical analysis of catastrophic forgetting through the ntk overlap matrix. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pp. 1072–1080, 2021. (cited on p. 1, 4)

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pp. 1675–1685. PMLR, 2019. (cited on p. 2)

Evron, I., Moroshko, E., Ward, R., Srebro, N., and Soudry, D. How catastrophic can catastrophic forgetting be in linear regression? In *Conference on Learning Theory (COLT)*, pp. 4028–4079. PMLR, 2022. (cited on p. 1, 2, 4, 5, 6, 7, 9, 13, 30, 31, 37, 43)

Goldfarb, D. and Hand, P. Analysis of catastrophic forgetting for random orthogonal transformation tasks in the overparameterized regime. In *International Conference on Artificial Intelligence and Statistics*, pp. 2975–2993. PMLR, 2023. (cited on p. 1)

Goodfellow, I. J., Mirza, M., Xiao, D., Courville, A., and Bengio, Y. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013. (cited on p. 1)

Gubin, L., Polyak, B. T., and Raik, E. 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. 4, 27, 28)

Halperin, I. The product of projection operators. *Acta Sci. Math.(Szeged)*, 23(1):96–99, 1962. (cited on p. 1, 5, 42)

Hoi, S. C., Sahoo, D., Lu, J., and Zhao, P. Online learning: A comprehensive survey. *Neurocomputing*, 459:249–289, 2021. ISSN 0925-2312. (cited on p. 4)

Ji, Z. and Telgarsky, M. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. In *International Conference on Learning Representations*, 2020. (cited on p. 2)

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. (cited on p. 1, 3, 8)

Lee, S., Goldt, S., and Saxe, A. Continual learning in the teacher-student setup: Impact of task similarity. In *International Conference on Machine Learning*, pp. 6109–6119. PMLR, 2021. (cited on p. 1, 9)

Lesort, T., Ostapenko, O., Misra, D., Arefin, M. R., Rodríguez, P., Charlin, L., and Rish, I. Scaling the number of tasks in continual learning. *arXiv preprint arXiv:2207.04543*, 2022. (cited on p. 5)

Lewkowycz, A. and Gur-Ari, G. On the training dynamics of deep networks with l_2 regularization. *Advances in Neural Information Processing Systems*, 33:4790–4799, 2020. (cited on p. 7)

Li, H., Wu, J., and Braverman, V. Fixed design analysis of regularization-based continual learning. *arXiv preprint arXiv:2303.10263*, 2023. (cited on p. 1)

Li, X. *Regularized adaptation: Theory, algorithms and applications*. Citeseer, 2007. (cited on p. 9)

Lubana, E. S., Trivedi, P., Koutra, D., and Dick, R. How do quadratic regularizers prevent catastrophic forgetting: The role of interpolation. In *Conference on Lifelong Learning Agents*, pp. 819–837. PMLR, 2022. (cited on p. 3)

McCloskey, M. and Cohen, N. J. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989. (cited on p. 1)

Mirzadeh, S. I., Farajtabar, M., Pascanu, R., and Ghasemzadeh, H. Understanding the role of training regimes in continual learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 7308–7320. Curran Associates, Inc., 2020. (cited on p. 7)

Morshed, M. S., Islam, M. S., and Noor-E-Alam, M. Sampling kaczmarz-motzkin method for linear feasibility problems: generalization and acceleration. *Mathematical Programming*, 194(1-2):719–779, 2022. (cited on p. 7)

Nedić, A. Random projection algorithms for convex set intersection problems. In *49th IEEE Conference on Decision and Control (CDC)*, pp. 7655–7660. IEEE, 2010. (cited on p. 6, 7, 27, 35, 37)

Needell, D. and Tropp, J. A. Paved with good intentions: analysis of a randomized block kaczmarz method. *Linear Algebra and its Applications*, 441:199–221, 2014. (cited on p. 6)

Novikoff, A. B. On convergence proofs on perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*, volume 12, pp. 615–622, New York, NY, USA, 1962. Polytechnic Institute of Brooklyn. (cited on p. 4)

Peng, L., Giampouras, P., and Vidal, R. The ideal continual learner: An agent that never forgets. In *International Conference on Machine Learning*, 2023. (cited on p. 8, 9)

Pentina, A., Sharmanska, V., and Lampert, C. H. Curriculum learning of multiple tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5492–5500, 2015. (cited on p. 9)

Reich, S. and Zalas, R. Polynomial estimates for the method of cyclic projections in hilbert spaces. *Numerical Algorithms*, pp. 1–26, 2023. (cited on p. 6)

Robins, A. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146, 1995. (cited on p. 1)

Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T., and Wayne, G. Experience replay for continual learning. *Advances in Neural Information Processing Systems*, 32, 2019. (cited on p. 1)

Rosset, S., Zhu, J., and Hastie, T. J. Margin maximizing loss functions. In *Advances in neural information processing systems*, pp. 1237–1244, 2004. (cited on p. 3, 9, 19)

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

Shamir, O. and Zhang, T. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International conference on machine learning*, pp. 71–79. PMLR, 2013. (cited on p. 7)

Sidford, A. Ms&e 213/cs 269o: Chapter 3-convexity, 2020. (cited on p. 14)

Smith, J. S., Tian, J., Hsu, Y.-C., and Kira, Z. A closer look at rehearsal-free continual learning. In *CVPR Workshop on Continual Learning in Computer Vision*, 2023. (cited on p. 3)

Soudry, D., Hoffer, E., Shpigel Nacson, M., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *JMLR*, 2018. (cited on p. 3, 9, 14, 15, 16, 19, 22)

Stojanov, S., Mishra, S., Thai, N. A., Dhanda, N., Humayun, A., Yu, C., Smith, L. B., and Rehg, J. M. Incremental object learning from contiguous views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8777–8786, 2019. (cited on p. 5)

Tang, J., Lin, K.-Y., and Li, L. Using domain adaptation for incremental svm classification of drift data. *Mathematics*, 10(19):3579, 2022. (cited on p. 9)

Von Neumann, J. On rings of operators. reduction theory. *Annals of Mathematics*, pp. 401–485, 1949. (cited on p. 1)

Wei, C., Lee, J. D., Liu, Q., and Ma, T. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. *Advances in Neural Information Processing Systems*, 32, 2019. (cited on p. 3, 19)

Woodworth, B., Gunasekar, S., Lee, J. D., Moroshko, E., Savarese, P., Golan, I., Soudry, D., and Srebro, N. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pp. 3635–3673. PMLR, 2020. (cited on p. 1)

Yang, J., Yan, R., and Hauptmann, A. G. Adapting svm classifiers to data with shifted distributions. In *Seventh IEEE international conference on data mining workshops (ICDMW 2007)*, pp. 69–76. IEEE, 2007. (cited on p. 9)

Yoon, J., Yang, E., Lee, J., and Hwang, S. J. Lifelong learning with dynamically expandable networks. In *International Conference on Learning Representations*, 2018. (cited on p. 1)

Zenke, F., Poole, B., and Ganguli, S. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pp. 3987–3995. PMLR, 2017. (cited on p. 1, 3, 8)

A. Comparison: Continual Linear Classification vs. Continual Linear Regression

	CONTINUAL LINEAR CLASSIFICATION (OURS)	CONTINUAL LINEAR REGRESSION (EVRON ET AL., 2022)
FUNDAMENTAL MAPPING (PROVEN ALGORITHMIC BIAS)	Continual linear classification with weakly-regularized exponential losses converges to a Sequential Max-Margin scheme (Theorem 3.1)	Continual linear regression with unregularized losses, solved sequentially with vanilla (S)GD, implicitly performs successive projections (their Eq. (5))
FEASIBLE SETS	Closed, convex, and affine polyhedral cones (Assumption 2.1)	Closed affine subspaces (their Eq. (4))
PROJECTION OPERATORS	Projection onto Convex Sets (POCS); Challenging to analyze (Eq. (2))	Affine projections; Easy to analyze (their Eq. (5))
OPTIMALITY GUARANTEES	$\|\mathbf{w}_t\| \leq 2\,\|\mathbf{w}^\star\|$ (Lemma 4.7)	$\|\mathbf{w}_t\| \leq \|\mathbf{w}^\star\|$ $\mathbf{w}_t \in \mathcal{W}^\star \implies \mathbf{w}_t = \mathbf{w}^\star$
ADVERSARIAL CONSTRUCTION	Arbitrarily bad quantities $1 = \frac{1}{\|\mathbf{w}^\star\|^2 R^2} \lim_{T\to\infty} \max_{m\in[T]} F_m(\mathbf{w}_t) \leq \frac{1}{\|\mathbf{w}^\star\|^2} \lim_{T\to\infty} d^2\left(\mathbf{w}_t, \mathcal{W}^\star\right) \leq 1$ (our Section 4.2, their Section 4.2)	
FISHER-BASED REGULARIZATION	Does not prevent forgetting (Figure 5, Appendix E.2.1)	Prevents forgetting (our Prop. 5.5)
LINEAR REGULARITY	Holds $d^2(\mathbf{w}_t, \mathcal{W}^\star) \leq 4\|\mathbf{w}\|^2 R^2 \max_{m\in[T]} d^2(\mathbf{w}_t, \mathcal{W}_m)$ (Lemma 4.8)	Does not hold
NONTRIVIAL BOUNDS ON DISTANCE TO THE FEASIBLE SET $d^2\left(\mathbf{w}_t, \mathcal{W}^\star\right)$ DEPENDING ON THE PROBLEM COMPLEXITY $\|\mathbf{w}^\star\|^2 R^2$	Possible due to regularity; Exponential decrease (linear convergence) (Table 1, Theorems 4.12 and 4.15)	Impossible (*e.g.,* their Section 5.1)
NONTRIVIAL BOUNDS ON FORGETTING DEPENDING ON THE PROBLEM COMPLEXITY $\|\mathbf{w}^\star\|^2 R^2$		Possible universal bounds (see below)
UNIVERSAL BOUNDS ON FORGETTING UNDER CYCLIC ORDERINGS (INDEPENDENT OF $\|\mathbf{w}^\star\|^2 R^2$)	$\frac{2T^2}{\sqrt{k}}$ (Proposition 4.13)	$\min\left(\frac{T^2}{\sqrt{k}}, \frac{T^2 D}{2k}\right)$ (their Theorem 11)
UNIVERSAL BOUNDS ON FORGETTING UNDER RANDOM ORDERINGS (INDEPENDENT OF $\|\mathbf{w}^\star\|^2 R^2$)	Unclear if possible	$\frac{9D}{k}$ (their Theorem 13)

B. Proofs for Algorithmic Bias in Regularization Methods (Section 3)

Remark B.1 (Simplification). For ease of notation, throughout our appendices, we redefine the samples so as to subsume their labels. That is, we handle only *positive* labels by redefining each $y\mathbf{x} \longmapsto \mathbf{x}$. This notation is common in theoretical papers (*e.g.,* Soudry et al. (2018)).

B.1. Auxiliary Lemmas

We first present two auxiliary results that we need for our main proof.

Lemma B.2. *Let $f : \mathbb{R}^D \to \mathbb{R}$ be a μ-strongly convex objective function (holding $\nabla^2 f(\mathbf{w}) \succeq \mu \mathbf{I}$ for some $\mu > 0$). Then, for any $\mathbf{w} \in \mathbb{R}^D$, the (Euclidean) distance between $\mathbf{w}$ and $\mathbf{w}^\star$ (the minimizer of the objective f), can be upper bounded by:*

$$\|\mathbf{w} - \mathbf{w}^\star\| \leq \frac{1}{\mu} \|\nabla f(\mathbf{w})\| \ .$$

Proof. This lemma is a known convex optimization result (*e.g.,* see Lemma 10 in Sidford (2020)). We prove it here for the sake of completeness.

We make use of the following property of strongly convex functions.

> **Property (9.8) from Boyd & Vandenberghe (2004).** Let $f : \mathbb{R}^D \to \mathbb{R}$ be a μ-strongly convex objective function (holding $\nabla^2 f(\mathbf{w}) \succeq \mu \mathbf{I}$ for some $\mu > 0$). Then, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$, we have
>
> $$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|^2 \ .$$

Using the above property, we prove our lemma, which is a slightly stronger result than Property (9.11) in Boyd & Vandenberghe (2004). Our proof follows the one by Sidford (2020), and is brought here for completeness.

First, we set $\mathbf{x} = \mathbf{w}^\star, \mathbf{y} = \mathbf{w}$ (from our lemma). Since $\mathbf{w}^\star$ is a (global) minimizer, its gradient is zero, *i.e.,* $\nabla f(\mathbf{w}^\star) = \mathbf{0}_D$. We thus get that

$$\begin{aligned} f(\mathbf{w}) &\geq f(\mathbf{w}^\star) + \underbrace{\nabla f(\mathbf{w}^\star)^\top}_{=\mathbf{0}_D} (\mathbf{w} - \mathbf{w}^\star) + \frac{\mu}{2} \|\mathbf{w} - \mathbf{w}^\star\|^2 \\ f(\mathbf{w}) - f(\mathbf{w}^\star) &\geq \frac{\mu}{2} \|\mathbf{w} - \mathbf{w}^\star\|^2 \ . \end{aligned}$$

Using the optimality of $\mathbf{w}^\star$, we get that

$$\begin{aligned} \forall \mathbf{w} \in \mathbb{R}^D : \ f(\mathbf{w}^\star) = \min_{\mathbf{y}} f(\mathbf{y}) &\geq \min_{\mathbf{y}} \left(f(\mathbf{w}) + \nabla f(\mathbf{w})^\top (\mathbf{y} - \mathbf{w}) + \frac{\mu}{2} \|\mathbf{y} - \mathbf{w}\|^2 \right) \\ &= f(\mathbf{w}) + \min_{\mathbf{y}} \left(\nabla f(\mathbf{w})^\top (\mathbf{y} - \mathbf{w}) + \frac{\mu}{2} \|\mathbf{y} - \mathbf{w}\|^2 \right) \ . \end{aligned}$$

Then, by plugging in the minimizer of the right term, we get

$$\begin{aligned} \forall \mathbf{w} \in \mathbb{R}^D : \ f(\mathbf{w}^\star) &\geq f(\mathbf{w}) + \nabla f(\mathbf{w})^\top \left(\mathbf{w} - \frac{1}{\mu} \nabla f(\mathbf{w}) - \mathbf{w} \right) + \frac{\mu}{2} \left\| \mathbf{w} - \frac{1}{\mu} \nabla f(\mathbf{w}) - \mathbf{w} \right\|^2 \\ &= f(\mathbf{w}) - \frac{1}{\mu} \nabla f(\mathbf{w})^\top \nabla f(\mathbf{w}) + \frac{\mu}{2} \left\| \frac{1}{\mu} \nabla f(\mathbf{w}) \right\|^2 = f(\mathbf{w}) - \frac{1}{\mu} \|\nabla f(\mathbf{w})\|^2 + \frac{1}{2\mu} \|\nabla f(\mathbf{w})\|^2 \\ &= f(\mathbf{w}) - \frac{1}{2\mu} \|\nabla f(\mathbf{w})\|^2 \\ \frac{1}{2\mu} \|\nabla f(\mathbf{w})\|^2 &\geq f(\mathbf{w}) - f(\mathbf{w}^\star) \ . \end{aligned}$$

Overall, we showed that $\forall \mathbf{w} \in \mathbb{R}^D$, it holds that

$$\frac{1}{2\mu} \|\nabla f(\mathbf{w})\|^2 \geq f(\mathbf{w}) - f(\mathbf{w}^\star) \geq \frac{\mu}{2} \|\mathbf{w} - \mathbf{w}^\star\|^2 \quad \Longrightarrow \quad \frac{1}{\mu^2} \|\nabla f(\mathbf{w})\|^2 \geq \|\mathbf{w} - \mathbf{w}^\star\|^2 \ ,$$

as required. □

Lemma B.3. *Let $c_1, \dots, c_k \in \mathbb{R}_{\geq 0}$ and let $\mathbf{B}_1, \dots, \mathbf{B}_k \succ \mathbf{0}_{D\times D}$. Consider sequentially solving k separable tasks $(S_t = (\mathbf{X}_t, y_t = \mathbf{1}))_{t\in[k]}$ (recall the simplification in Remark B.1) using the following iterative update rule:*

$$\begin{aligned} &\mathbf{w}_0 = \mathbf{0}_D \\ \forall t \in [k]: \quad &\mathbf{w}_t = \operatorname{argmin}_{\mathbf{w}} \|\mathbf{w} - c_t \mathbf{w}_{t-1}\|^2_{\mathbf{B}_t} \quad \textit{s.t.} \quad \mathbf{w}^\top \mathbf{x} \geq 1, \forall \mathbf{x} \in \mathbf{X}_t \,. \end{aligned} \tag{4}$$

Then, for almost all datasets sampled from k absolutely continuous distributions, when $\mathbf{w}_t \neq c_t \mathbf{w}_{t-1}$, the unique dual solution $\boldsymbol{\alpha}_t \in \mathbb{R}^{|S_t|}_{\geq 0}$ satisfying the KKT conditions of Eq. (4), holds:

$$\mathbf{B}_t(\mathbf{w}_t - c_t \mathbf{w}_{t-1}) = \sum_{\mathbf{x}\in S_t} \mathbf{x}\alpha_t(\mathbf{x}), \quad \textit{AND} \quad \Big(\forall \mathbf{x} \in S_t: \ \big(\alpha_t(\mathbf{x}) > 0 \wedge \mathbf{w}_t^\top \mathbf{x} = 1\big) \ \textit{OR} \ \big(\alpha_t(\mathbf{x}) = 0 \wedge \mathbf{w}_t^\top \mathbf{x} > 1\big)\Big).$$

(That is, there is no support vector for which $\alpha_t(\mathbf{x}) = 0$.)

Proof for Lemma B.3. The proof follows the techniques of the proof of Lemma 12 in Appendix B of Soudry et al. (2018).

Here, we focus only on task sequences where $\mathbf{w}_t = \mathbf{P}_t(c_t \mathbf{w}_{t-1}) \neq c_t \mathbf{w}_{t-1}$, $\forall t \in [k]$ (we can always reduce to such cases by simply removing from the sequence any task t for which $w_t = \mathbf{w}_{t-1}$). Thus, $\mathbf{w}_t$ must lie on the boundary of the tth feasible set, as can be seen from the unique best approximation (or nearest point) property. Consequently, the support set $\hat{\mathcal{S}}_t \triangleq \{\mathbf{x} \in \mathbf{X}_t \mid \mathbf{w}_t^\top \mathbf{x} = 1\}$ is nonempty.

For almost all datasets (all except measure zero), no more than D datapoints will be on the same hyperplane (*e.g.,* the $\mathbf{w}_t^\top \mathbf{x} = 1$ hyperplane). Therefore, for any task $t \in [k]$ there can be at most D support vectors, *i.e.,* $|\hat{\mathcal{S}}_t| \leq D$. Also, for almost all datasets, any set of at most D vectors (*e.g.,* $\hat{\mathcal{S}}_t$) is linearly independent.

Let $\boldsymbol{\alpha}_t \in \mathbb{R}^{|S_t|}_{\geq 0}$ be the dual solution satisfying the KKT conditions of Eq. (4) in the tth task. We denote the matrix whose columns are the support vectors by $\mathbf{X}_{\hat{\mathcal{S}}_t} \in \mathbb{R}^{D\times|\hat{\mathcal{S}}_t|}$ and the dual solution, restricted to its components corresponding to support vectors, by $\boldsymbol{\alpha}_{\hat{\mathcal{S}}_t}$. Due to the complementary slackness, datapoints outside of the support set (*i.e.,* $\mathbf{w}_t^\top \mathbf{x} > 1$), must have a corresponding zero dual variable $\alpha_t(\mathbf{x}) = 0$. We now recall that $\mathbf{B}_t$ is invertible, and rewrite the stationarity condition as,

$$\begin{aligned} \mathbf{B}_t(\mathbf{w}_t - c_t \mathbf{w}_{t-1}) &= \sum_{\mathbf{x}\in S_t} \mathbf{x}\alpha_t(\mathbf{x}) = \sum_{\mathbf{x}\in \hat{\mathcal{S}}_t} \mathbf{x}\alpha_t(\mathbf{x}) + \cancel{\sum_{\mathbf{x}\in S_t:\, \mathbf{w}_t^\top \mathbf{x}>1} \mathbf{x} \underbrace{\alpha_t(\mathbf{x})}_{=0}} = \mathbf{X}_{\hat{\mathcal{S}}_t} \boldsymbol{\alpha}_{\hat{\mathcal{S}}_t} \\ \mathbf{w}_t - c_t \mathbf{w}_{t-1} &= \mathbf{B}_t^{-1} \mathbf{X}_{\hat{\mathcal{S}}_t} \boldsymbol{\alpha}_{\hat{\mathcal{S}}_t} \,. \end{aligned} \tag{5}$$

Multiplying from the left by the full row rank matrix $\mathbf{X}_{\hat{\mathcal{S}}_t}^\top$, we get on the one hand that $\boldsymbol{\alpha}_{\hat{\mathcal{S}}_t}$ is uniquely defined as,

$$\begin{aligned} \underbrace{\mathbf{X}_{\hat{\mathcal{S}}_t}^\top \mathbf{w}_t}_{=\mathbf{1}} - c_t \mathbf{X}_{\hat{\mathcal{S}}_t}^\top \mathbf{w}_{t-1} &= \mathbf{X}_{\hat{\mathcal{S}}_t}^\top \mathbf{B}_t^{-1} \mathbf{X}_{\hat{\mathcal{S}}_t} \boldsymbol{\alpha}_{\hat{\mathcal{S}}_t} \\ \boldsymbol{\alpha}_{\hat{\mathcal{S}}_t} &= \left(\mathbf{X}_{\hat{\mathcal{S}}_t}^\top \mathbf{B}_t^{-1} \mathbf{X}_{\hat{\mathcal{S}}_t}\right)^{-1} \mathbf{1} - c_t \left(\mathbf{X}_{\hat{\mathcal{S}}_t}^\top \mathbf{B}_t^{-1} \mathbf{X}_{\hat{\mathcal{S}}_t}\right)^{-1} \mathbf{X}_{\hat{\mathcal{S}}_t}^\top \mathbf{w}_{t-1} \,, \end{aligned} \tag{6}$$

where $\mathbf{X}_{\hat{\mathcal{S}}_t}^\top \mathbf{B}_t^{-1} \mathbf{X}_{\hat{\mathcal{S}}_t}$ is invertible because $\mathbf{B}_t, \mathbf{B}_t^{-1}$ are positive definite invertible matrices and $\mathbf{X}_{\hat{\mathcal{S}}_t}$ is full column rank as explained above (to see this, notice that $\mathbf{B}_t^{-1}$ has a Cholesky decomposition such that $\sqrt{\mathbf{B}_t^{-1}}$ preserves the full column rank of $\mathbf{X}_{\hat{\mathcal{S}}_t}$). On the other hand, by substituting Eq. (6) back into the condition in Eq. (5), we get that $\forall t \in [k]$,

$$\begin{aligned} \mathbf{w}_t &= \mathbf{B}_t^{-1} \mathbf{X}_{\hat{\mathcal{S}}_t} \boldsymbol{\alpha}_{\hat{\mathcal{S}}_t} + c_t \mathbf{w}_{t-1} \\ &= \mathbf{B}_t^{-1} \mathbf{X}_{\hat{\mathcal{S}}_t} \left(\left(\mathbf{X}_{\hat{\mathcal{S}}_t}^\top \mathbf{B}_t^{-1} \mathbf{X}_{\hat{\mathcal{S}}_t}\right)^{-1} \mathbf{1} - c_t \left(\mathbf{X}_{\hat{\mathcal{S}}_t}^\top \mathbf{B}_t^{-1} \mathbf{X}_{\hat{\mathcal{S}}_t}\right)^{-1} \mathbf{X}_{\hat{\mathcal{S}}_t}^\top \mathbf{w}_{t-1} \right) + c_t \mathbf{w}_{t-1} \\ &= \mathbf{B}_t^{-1} \mathbf{X}_{\hat{\mathcal{S}}_t} \left(\mathbf{X}_{\hat{\mathcal{S}}_t}^\top \mathbf{B}_t^{-1} \mathbf{X}_{\hat{\mathcal{S}}_t}\right)^{-1} \mathbf{1} + c_t \left(\mathbf{I} - \mathbf{B}_t^{-1} \mathbf{X}_{\hat{\mathcal{S}}_t} \left(\mathbf{X}_{\hat{\mathcal{S}}_t}^\top \mathbf{B}_t^{-1} \mathbf{X}_{\hat{\mathcal{S}}_t}\right)^{-1} \mathbf{X}_{\hat{\mathcal{S}}_t}^\top \right) \mathbf{w}_{t-1} \,. \end{aligned} \tag{7}$$

Importantly, the above implies that, recursively, $\mathbf{w}_{t-1}$ is a rational function in the components of $\mathbf{X}_1, \ldots, \mathbf{X}_{t-1}$ (where $c_1, \ldots, c_t$ and $\mathbf{B}_1, \ldots, \mathbf{B}_t$ are *given*). Clearly, the same holds for $\boldsymbol{\alpha}_{\hat{\mathcal{S}}_t}$ (Eq. (6)) which entirely depends on $\mathbf{X}_t, \mathbf{B}_t$, and $\mathbf{w}_{t-1}$. Hence, its entries can be expressed as $(\boldsymbol{\alpha}_{\hat{\mathcal{S}}_t})_n = p_n^{(t)}(\mathbf{X}_1, \ldots, \mathbf{X}_t)/q_n^{(t)}(\mathbf{X}_1, \ldots, \mathbf{X}_t)$ for some polynomials $p_n^{(t)}, q_n^{(t)}$. Now, similarly to Soudry et al. (2018), we notice that $(\boldsymbol{\alpha}_{\hat{\mathcal{S}}_t})_n = 0$ only if $p_n^{(t)}(\mathbf{X}_1, \ldots, \mathbf{X}_t) = 0$, *i.e.,* the components of $\mathbf{X}_1, \ldots, \mathbf{X}_t$ must constitute a root of the polynomial $p_n^{(t)}$. However, the roots of any polynomial have measure zero, unless that polynomial is the zero polynomial, *i.e.,* $p_n^{(t)}(\mathbf{X}_1, \ldots, \mathbf{X}_t) = 0,\ \forall \mathbf{X}_1, \ldots, \mathbf{X}_t$.

Our goal now. To prove that our polynomials $(p_n^{(t)})_t$ cannot be zero polynomials, it is sufficient to construct a specific task sequence for which they are not zero. Then, we will be able to conclude that the event in which $\boldsymbol{\alpha}_{\hat{\mathcal{S}}_t}$ has a zero entry, is measure zero.

For the sake of readability, we divide our proof into two cases.

1. **When $\mathbf{B}_1 = \cdots = \mathbf{B}_k = \mathbf{I}_{D\times D}$**: The expressions from Eq. (6) and Eq. (7) become, $\forall t \in [k]$,

$$\begin{aligned}
\boldsymbol{\alpha}_{\hat{\mathcal{S}}_t} &= \left(\mathbf{X}_{\hat{\mathcal{S}}_t}^\top \mathbf{X}_{\hat{\mathcal{S}}_t}\right)^{-1} \mathbf{1} - c_t \left(\mathbf{X}_{\hat{\mathcal{S}}_t}^\top \mathbf{X}_{\hat{\mathcal{S}}_t}\right)^{-1} \mathbf{X}_{\hat{\mathcal{S}}_t}^\top \mathbf{w}_{t-1} \\
\mathbf{w}_t &= \mathbf{X}_{\hat{\mathcal{S}}_t} \left(\mathbf{X}_{\hat{\mathcal{S}}_t}^\top \mathbf{X}_{\hat{\mathcal{S}}_t}\right)^{-1} \mathbf{1} + c_t \left(\mathbf{I} - \mathbf{X}_{\hat{\mathcal{S}}_t} \left(\mathbf{X}_{\hat{\mathcal{S}}_t}^\top \mathbf{X}_{\hat{\mathcal{S}}_t}\right)^{-1} \mathbf{X}_{\hat{\mathcal{S}}_t}^\top\right) \mathbf{w}_{t-1} \,.
\end{aligned} \tag{8}$$

Detailed construction of the task sequence. We are given an arbitrary dimensionality D. Let $\mathbf{e}_i$ be the ith standard unit vector in $\mathbb{R}^D$. Let $N_1, \ldots, N_k \in [D]$ be an arbitrary sequence of the support sets' sizes.

We define the following datasets:

$$\underbrace{\mathbf{X}_t}_{D\times N_t} = \beta_t \begin{bmatrix} \mathbf{e}_1 & \cdots & \mathbf{e}_{N_t} \end{bmatrix}, \qquad \forall t \in [k]$$

for some sequence (β_t) of strictly positive numbers which we will define later. Under this construction (after choosing appropriate (β_t) that will ensure projections at every iteration), the support set of each of the datasets will be identical to the dataset itself, hence we use $\mathbf{X}_t = \mathbf{X}_{\hat{\mathcal{S}}_t}$ and $\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{\hat{\mathcal{S}}_t}$ interchangeably for the rest of the construction.

Notice that under our construction it holds that

$$\underbrace{\mathbf{X}_t^\top \mathbf{X}_t}_{N_t\times N_t} = \beta_t^2 \mathbf{I}_{N_t\times N_t}, \qquad \underbrace{\mathbf{X}_t \mathbf{X}_t^\top}_{D\times D} = \beta_t^2 \begin{bmatrix} \mathbf{I}_{N_t\times N_t} & \\ & \mathbf{0}_{(D-N_t)\times(D-N_t)} \end{bmatrix}, \qquad \mathbf{X}_t \mathbf{1}_{N_t} = \beta_t \begin{bmatrix} \mathbf{1}_{N_t} \\ \mathbf{0}_{D-N_t} \end{bmatrix} .$$

Using the expression in Eq. (8) and recalling that $\mathbf{w}_0 = \mathbf{0}_D$, we have:

$$\begin{aligned}
\mathbf{w}_t &= \mathbf{X}_t \left(\mathbf{X}_t^\top \mathbf{X}_t\right)^{-1} \mathbf{1} + c_t \left(\mathbf{I} - \mathbf{X}_t \left(\mathbf{X}_t^\top \mathbf{X}_t\right)^{-1} \mathbf{X}_t^\top\right) \mathbf{w}_{t-1} = \frac{1}{\beta_t^2} \mathbf{X}_t \mathbf{1}_{N_t} + c_t \underbrace{\left(\mathbf{I} - \frac{1}{\beta_t^2} \mathbf{X}_t \mathbf{X}_t^\top\right)}_{\triangleq \mathbf{M}_t} \mathbf{w}_{t-1} \\
&= \frac{1}{\beta_t^2} \mathbf{X}_t \mathbf{1}_{N_t} + c_t \mathbf{M}_t \mathbf{w}_{t-1} = \frac{1}{\beta_t^2} \mathbf{X}_t \mathbf{1}_{N_t} + c_t \mathbf{M}_t \left(\frac{1}{\beta_{t-1}^2} \mathbf{X}_{t-1} \mathbf{1}_{N_{t-1}} + c_{t-1} \mathbf{M}_{t-1} \mathbf{w}_{t-2}\right) \\
&= \frac{1}{\beta_t^2} \mathbf{X}_t \mathbf{1}_{N_t} + \frac{1}{\beta_{t-1}^2} c_t \mathbf{M}_t \mathbf{X}_{t-1} \mathbf{1}_{N_{t-1}} + c_t c_{t-1} \mathbf{M}_t \mathbf{M}_{t-1} \mathbf{w}_{t-2} \\
[\text{recursively}] &= \frac{1}{\beta_t^2} \mathbf{X}_t \mathbf{1}_{N_t} + \sum_{i=1}^{t-1} c_t \cdots c_{i+1} \frac{1}{\beta_i^2} \mathbf{M}_t \cdots \mathbf{M}_{i+1} \mathbf{X}_i \mathbf{1}_{N_i} \,.
\end{aligned}$$

Plugging in the diagonal $\mathbf{M}_i = \mathbf{I}_{D\times D} - \frac{1}{\beta_i^2}\mathbf{X}_i\mathbf{X}_i^\top = \begin{bmatrix} \mathbf{0}_{N_i\times N_i} & \\ & \mathbf{I}_{(D-N_i)\times(D-N_i)} \end{bmatrix}$, we get

$$\begin{aligned}
\mathbf{w}_t &= \frac{1}{\beta_t^2}\mathbf{X}_t\mathbf{1}_{N_t} + \sum_{i=1}^{t-1}\left(\frac{1}{\beta_i^2}\prod_{j=i+1}^{t} c_j \begin{bmatrix} \mathbf{0}_{N_j\times N_j} & \\ & \mathbf{I}_{(D-N_j)\times(D-N_j)} \end{bmatrix} \cdot \underbrace{\mathbf{X}_i\mathbf{1}_{N_i}}_{=\beta_i\begin{bmatrix}\mathbf{1}_{N_i}\\ \mathbf{0}_{D-N_i}\end{bmatrix}}\right) \\
&= \frac{1}{\beta_t^2}\mathbf{X}_t\mathbf{1}_{N_t} + \sum_{i=1}^{t-1}\left(\left(\frac{1}{\beta_i}\prod_{j=i+1}^{t} c_j\right)\begin{bmatrix} \mathbb{I}\left[1\le N_i\right]\prod_{j=i+1}^{t}\mathbb{I}\left[1>N_j\right] \\ \vdots \\ \mathbb{I}\left[D\le N_i\right]\prod_{j=i+1}^{t}\mathbb{I}\left[D>N_j\right]\end{bmatrix}\right) \\
&= \frac{1}{\beta_t}\begin{bmatrix}\mathbf{1}_{N_t}\\ \mathbf{0}_{D-N_t}\end{bmatrix} + \sum_{i=1}^{t-1}\left(\left(\frac{1}{\beta_i}\prod_{j=i+1}^{t} c_j\right)\begin{bmatrix} \mathbb{I}\left[N_i\ge 1>\max\left(N_{i+1},\dots,N_t\right)\right] \\ \vdots \\ \mathbb{I}\left[N_i\ge D>\max\left(N_{i+1},\dots,N_t\right)\right]\end{bmatrix}\right),
\end{aligned}$$

and since multiplying by $\mathbf{X}_t^\top$ from the left "trims" the last $D-N_t$ rows of any vector, we get:

$$\begin{aligned}
\underbrace{\mathbf{X}_t^\top\mathbf{w}_{t-1}}_{N_t\times 1} &= \mathbf{X}_t^\top\left(\frac{1}{\beta_{t-1}}\begin{bmatrix}\mathbf{1}_{N_{t-1}}\\ \mathbf{0}_{D-N_{t-1}}\end{bmatrix} + \sum_{i=1}^{t-2}\left(\left(\frac{1}{\beta_i}\prod_{j=i+1}^{t-1} c_j\right)\begin{bmatrix} \mathbb{I}\left[N_i\ge 1>\max\left(N_{i+1},\dots,N_{t-1}\right)\right] \\ \vdots \\ \mathbb{I}\left[N_i\ge D>\max\left(N_{i+1},\dots,N_{t-1}\right)\right]\end{bmatrix}\right)\right) \\
&= \beta_t\left(\frac{1}{\beta_{t-1}}\begin{bmatrix}\mathbf{1}_{\min(N_{t-1},N_t)}\\ \mathbf{0}_{\max(N_t-N_{t-1},0)}\end{bmatrix} + \sum_{i=1}^{t-2}\left(\left(\frac{1}{\beta_i}\prod_{j=i+1}^{t-1} c_j\right)\begin{bmatrix} \mathbb{I}\left[N_i\ge 1>\max\left(N_{i+1},\dots,N_{t-1}\right)\right] \\ \vdots \\ \mathbb{I}\left[N_i\ge N_t>\max\left(N_{i+1},\dots,N_{t-1}\right)\right]\end{bmatrix}\right)\right)
\end{aligned}$$

We then plug in the above into $\boldsymbol{\alpha}_t$ again (using Eq. (6)):

$$\begin{aligned}
\underbrace{\boldsymbol{\alpha}_t}_{N_t\times 1} &= \underbrace{\left(\mathbf{X}_t^\top\mathbf{X}_t\right)^{-1}}_{N_t\times N_t}\underbrace{\left(\mathbf{1}_{N_t} - c_t\mathbf{X}_t^\top\mathbf{w}_{t-1}\right)}_{N_t\times 1} \\
&= \frac{1}{\beta_t^2}\left(\mathbf{1}_{N_t} - c_t\beta_t\left(\frac{1}{\beta_{t-1}}\begin{bmatrix}\mathbf{1}_{\min(N_{t-1},N_t)}\\ \mathbf{0}_{\max(N_t-N_{t-1},0)}\end{bmatrix} + \sum_{i=1}^{t-2}\left(\frac{1}{\beta_i}\prod_{j=i+1}^{t-1} c_j\right)\begin{bmatrix} \mathbb{I}\left[N_i\ge 1>\max\left(N_{i+1},\dots,N_{t-1}\right)\right] \\ \vdots \\ \mathbb{I}\left[N_i\ge N_t>\max\left(N_{i+1},\dots,N_{t-1}\right)\right]\end{bmatrix}\right)\right)
\end{aligned}$$

Then we get the following elementwise formula $\forall n\in[N_t]$:

$$\beta_t^2\left(\boldsymbol{\alpha}_t\right)_n = 1 - \underbrace{c_t\beta_t\frac{1}{\beta_{t-1}}}_{\ge 0}\underbrace{\mathbb{I}\left[n\le N_{t-1}\right]}_{\le 1} - \underbrace{c_t\beta_t}_{\ge 0}\sum_{i=1}^{t-2}\underbrace{\mathbb{I}\left[N_i\ge n>\max\left(N_{i+1},\dots,N_{t-1}\right)\right]}_{\le 1}\underbrace{\frac{1}{\beta_i}\prod_{j=i+1}^{t-1} c_j}_{\ge 0}.$$

Denoting $C\triangleq\max\left\{1,\ \max_{t\in[k]}c_t\right\}$, we lower bound the above as:

$$\beta_t^2\left(\boldsymbol{\alpha}_t\right)_n \ge 1 - c_t\beta_t\frac{1}{\beta_{t-1}} - c_t\beta_t\sum_{i=1}^{t-2}\frac{1}{\beta_i}\prod_{j=i+1}^{t-1} c_j = 1 - \beta_t\sum_{i=1}^{t-1}\frac{1}{\beta_i}\prod_{j=i+1}^{t} c_j \ge 1 - \beta_t\sum_{i=1}^{t-1}\frac{1}{\beta_i}C^{t-i}$$

$$[C\ge 1]\ge 1 - \beta_t C^t\sum_{i=1}^{t-1}\frac{1}{\beta_i} \ge 1 - (t-1)\underbrace{\beta_t C^t\max_{i\in[t-1]}\frac{1}{\beta_i}}_{>0} > 1 - t\beta_t C^t\max_{i\in[t-1]}\frac{1}{\beta_i} \overset{\text{require}}{\ge} 0$$

$$\frac{\min_{i\in[t-1]}\beta_i}{tC^t}\ge\beta_t$$

Thus, to hold the above, we can choose (β_t) to be a decreasing sequence as follows:

$$\beta_1 = 1,\quad \beta_t = \frac{\min_{i\in[t-1]}\beta_i}{tC^t} = \frac{1}{tC^t}\beta_{t-1} = \frac{1}{t!C^{2+3+\cdots+t}} = \frac{1}{t!C^{(t+2)(t-1)/2}}.$$

Summary of the first case. We showed a construction where at *any* iteration $t \in [k]$, *all* N_t entries of the corresponding dual vector $\boldsymbol{\alpha}_{\hat{\mathcal{S}}_t} (= \boldsymbol{\alpha}_t)$ are strictly positive. As explained above, this implies that when $\mathbf{w}_t \neq c_t \mathbf{w}_{t-1}$, the polynomial $p_n^{(t)}$ is *not* a zero polynomial $\forall t \in [k], \forall n \in N_t$ and thus becomes zero only in a finite number of measure zero roots. Using the union bound on the countable number of iterations k and all possible choices of support sizes $N_1, ..., N_k \in [D]$, we still remain with a measure zero event.

2. **General $\mathbf{B}_1, \ldots, \mathbf{B}_t$:** We can extend the techniques above and define the polynomials over the choices of $(\mathbf{B}_t)$ as well, *i.e.,* $(\boldsymbol{\alpha}_{\hat{\mathcal{S}}_t})_n = p_n^{(t)}(\mathbf{X}_1, \ldots, \mathbf{X}_t, \mathbf{B}_1, \ldots, \mathbf{B}_t)/q_n^{(t)}(\mathbf{X}_1, \ldots, \mathbf{X}_t, \mathbf{B}_1, \ldots, \mathbf{B}_t)$ for some polynomials $p_n^{(t)}, q_n^{(t)}$. Then, our first case above, where $\mathbf{B}_t = \mathbf{I}, \forall t$, shows that the roots of these updated polynomials are of measure zero. Again, employing the union bound over the countable number of iterations k and all possible choices of support sizes $N_1, ..., N_k \in [D]$, shows that $\forall t \in [k]: \ \mathbf{w}_t \neq c_t \mathbf{w}_{t-1} \Longrightarrow \boldsymbol{\alpha}_{\hat{\mathcal{S}}_t} \succ \mathbf{0}$ (elementwise) almost surely.

□

After showing that either $\boldsymbol{\alpha}_t = \mathbf{0}$ (when $\mathbf{w}_t = c_t \mathbf{w}_{t-1}$) or $\boldsymbol{\alpha}_{\hat{\mathcal{S}}_t} \succ \mathbf{0}$ (elementwise), we can conclude the following.

Corollary B.4. *Under the conditions and iterative process in Lemma B.3, when $\mathbf{w}_t \neq c_t \mathbf{w}_{t-1}$, there almost surely exists a finite $\tilde{\mathbf{w}}_t$ such that*

$$\sum\nolimits_{\mathbf{x} \in \hat{\mathcal{S}}_t} \mathbf{x} \exp\left(-\tilde{\mathbf{w}}_t^\top \mathbf{x}\right) = \mathbf{B}_t(\mathbf{w}_t - c_t \mathbf{w}_{t-1}) = \sum\nolimits_{\mathbf{x} \in \hat{\mathcal{S}}_t} \mathbf{x} \underbrace{\alpha_t(\mathbf{x})}_{>0} .$$

Proof. The support $\hat{\mathcal{S}}_t$ is linearly independent a.s., so we simply require $\mathbf{X}_{\hat{\mathcal{S}}_t}^\top \tilde{\mathbf{w}}_t = -\ln \boldsymbol{\alpha}_{\hat{\mathcal{S}}_t}$ for a full row rank $\mathbf{X}_{\hat{\mathcal{S}}_t}^\top$. □

B.2. Main Result

We are now ready to prove Theorem 3.1. In Remark 3.2, we explained why existing analysis techniques (*e.g.,* from Rosset et al. (2004); Wei et al. (2019)) are less suitable for our setting, in which the scale of iterates is also of great importance (and not only their direction). More related tools are the ones that were used in Soudry et al. (2018), which analyzed the convergence of the gradient descent iterates under *unregularized* problems with exponential losses to the max-margin solution as the number of gradient steps $t \to \infty$. Here, however, we analyze the convergence of the (unique) minimizer of the *regularized* problem to a max-margin solution, as the regularization strength $\lambda \to 0$. While there are some technical similarities, there are also differences and challenges of a different nature.

Recall Theorem 3.1. Let $\lambda_t = \lambda > 0$, and $\mathbf{B}_t = \mathbf{I}$, $\forall t \in [k]$. Then, for almost all separable datasets,[2] in the limit of $\lambda \to 0$, it holds that $\mathbf{w}_t^{(\lambda)} \to \ln\left(\frac{1}{\lambda}\right)\mathbf{w}_t$ with a residual of $\|\mathbf{w}_t^{(\lambda)} - \ln\left(\frac{1}{\lambda}\right)\mathbf{w}_t\| = \mathcal{O}\left(t \ln\ln\left(\frac{1}{\lambda}\right)\right)$. As a result, at any iteration $t = o\left(\frac{\ln(1/\lambda)}{\ln\ln(1/\lambda)}\right)$, we get

$$\lim_{\lambda\to 0} \frac{\mathbf{w}_t^{(\lambda)}}{\|\mathbf{w}_t^{(\lambda)}\|} = \frac{\mathbf{w}_t}{\|\mathbf{w}_t\|}\,.$$

Proof. We will prove by induction on $t \geq 0$ that the scale of the residual $\mathbf{r}_t^{(\lambda)} \triangleq \mathbf{w}_t^{(\lambda)} - \ln\left(\frac{1}{\lambda}\right)\mathbf{w}_t$ at each iteration is $\mathcal{O}\left(t \ln\ln\left(\frac{1}{\lambda}\right)\right)$; and that consequently (since $\ln\left(\frac{1}{\lambda}\right)\mathbf{w}_t$ grows faster), the iterates are either identical (*i.e.,* $\mathbf{w}_t^{(\lambda)} = \mathbf{w}_t$) or converge in the same direction when $\lambda \to 0$, *i.e.,* $\lim_{\lambda\to 0} \frac{\mathbf{w}_t^{(\lambda)}}{\|\mathbf{w}_t^{(\lambda)}\|} = \frac{\mathbf{w}_t}{\|\mathbf{w}_t\|}$.

For $t = 0$: By the conditions of the theorem, it trivially holds that $\mathbf{w}_0^{(\lambda)} = \mathbf{w}_0 = \mathbf{0}_D$ and $\mathbf{r}_0^{(\lambda)} = \mathbf{w}_0^{(\lambda)} - \ln\left(\frac{1}{\lambda}\right)\mathbf{w}_0 = \mathbf{0}_D$.

For $t \geq 1$: The solved optimization problem (recall Remark B.1) is:

$$\mathbf{w}_t^{(\lambda)} = \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^D} \mathcal{L}_\lambda(\mathbf{w}) \triangleq \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^D} \sum_{\mathbf{x}\in S_t} e^{-\mathbf{w}^\top\mathbf{x}} + \frac{\lambda}{2}\left\|\mathbf{w} - \mathbf{w}_{t-1}^{(\lambda)}\right\|^2.$$

Proof's idea. Notice that the objective above is λ-strongly convex, since its Hessian matrix is $\nabla^2\mathcal{L}_\lambda(\mathbf{w}) = \lambda\mathbf{I} + \sum_{\mathbf{x}} e^{-\mathbf{w}^\top\mathbf{x}}\mathbf{x}\mathbf{x}^\top \succeq \lambda\mathbf{I} \succ \mathbf{0}$. We are going to define an $\mathcal{O}(1)$ vector $\tilde{\mathbf{w}}_t$ and a sign $s_t \in \{-1, +1\}$, and employ the triangle inequality and Lemma B.2 to show that

$$\begin{aligned}
\left\|\mathbf{r}_t^{(\lambda)}\right\| &\triangleq \left\|\mathbf{w}_t^{(\lambda)} - \ln\left(\tfrac{1}{\lambda}\right)\mathbf{w}_t\right\| \\
&= \left\|\mathbf{w}_t^{(\lambda)} - \left(\left(\ln\left(\tfrac{1}{\lambda}\right) + s_t \ln\ln\left(\tfrac{1}{\lambda}\right)\right)\mathbf{w}_t + \tilde{\mathbf{w}}_t\right) + s_t \ln\ln\left(\tfrac{1}{\lambda}\right)\mathbf{w}_t + \tilde{\mathbf{w}}_t\right\| \\
&\leq \underbrace{\frac{1}{\lambda}\left\|\nabla\mathcal{L}_\lambda\left(\left(\ln\left(\tfrac{1}{\lambda}\right) + s_t \ln\ln\left(\tfrac{1}{\lambda}\right)\right)\mathbf{w}_t + \tilde{\mathbf{w}}_t\right)\right\|}_{=\mathcal{O}\left((t-1)\ln\ln\left(\frac{1}{\lambda}\right)\right)} + \ln\ln\left(\tfrac{1}{\lambda}\right)\underbrace{\|\mathbf{w}_t\|}_{=\mathcal{O}(1)} + \underbrace{\|\tilde{\mathbf{w}}_t\|}_{=\mathcal{O}(1)} = \mathcal{O}\left(t\ln\ln\left(\tfrac{1}{\lambda}\right)\right)\,.
\end{aligned} \tag{9}$$

Then, since $\left\|\mathbf{w}_t^{(\lambda)} - \ln\left(\frac{1}{\lambda}\right)\mathbf{w}_t\right\| = \mathcal{O}\left(t\ln\ln\left(\frac{1}{\lambda}\right)\right)$ it will immediately follow that the weakly-regularized solution $\mathbf{w}_t^{(\lambda)}$ converges in direction to the Sequential Max-Margin solution $\mathbf{w}_t$.

Back to the proof. First, we compute the gradient of $\mathcal{L}_\lambda$, normalized by λ:

$$\frac{1}{\lambda}\nabla\mathcal{L}_\lambda(\mathbf{w}) = \frac{1}{\lambda}\left(\sum_{\mathbf{x}\in S_t} -\mathbf{x}\exp\left(-\mathbf{w}^\top\mathbf{x}\right) + \lambda\mathbf{w} - \lambda\mathbf{w}_{t-1}^{(\lambda)}\right) = -\frac{1}{\lambda}\sum_{\mathbf{x}\in S_t}\mathbf{x}\exp\left(-\mathbf{w}^\top\mathbf{x}\right) + \mathbf{w} - \mathbf{w}_{t-1}^{(\lambda)}\,.$$

Then, we plug in $\mathbf{w} = \left(\ln\left(\frac{1}{\lambda}\right) + s_t\ln\ln\left(\frac{1}{\lambda}\right)\right)\mathbf{w}_t + \tilde{\mathbf{w}}_t = \ln\left(\frac{1}{\lambda}\ln^{s_t}\left(\frac{1}{\lambda}\right)\right)\mathbf{w}_t + \tilde{\mathbf{w}}_t$, for some sign $s_t \in \{-1,+1\}$ and a vector $\tilde{\mathbf{w}}_t$ with a norm independent of λ (both will be defined below).

$$\begin{aligned}\frac{1}{\lambda}\nabla\mathcal{L}_\lambda\Big(\left(\ln\left(\tfrac{1}{\lambda}\right) + s_t\ln\ln\left(\tfrac{1}{\lambda}\right)\right)\mathbf{w}_t + \tilde{\mathbf{w}}_t\Big) &= -\frac{1}{\lambda}\sum_{\mathbf{x}\in S_t}\mathbf{x}e^{\ln\left(\lambda\ln^{-s_t}\left(\frac{1}{\lambda}\right)\right)\mathbf{w}_t^\top\mathbf{x}}e^{-\tilde{\mathbf{w}}_t^\top\mathbf{x}} + \ln\left(\tfrac{1}{\lambda}\ln^{s_t}\left(\tfrac{1}{\lambda}\right)\right)\mathbf{w}_t + \tilde{\mathbf{w}}_t - \mathbf{w}_{t-1}^{(\lambda)}\\ &= -\frac{1}{\lambda}\sum_{\mathbf{x}\in S_t}\mathbf{x}\left(\lambda\ln^{-s_t}\left(\tfrac{1}{\lambda}\right)\right)^{\mathbf{w}_t^\top\mathbf{x}}e^{-\tilde{\mathbf{w}}_t^\top\mathbf{x}} + \ln\left(\tfrac{1}{\lambda}\ln^{s_t}\left(\tfrac{1}{\lambda}\right)\right)\mathbf{w}_t + \tilde{\mathbf{w}}_t - \mathbf{w}_{t-1}^{(\lambda)}\,.\end{aligned}$$

Now, denoting the set of support vectors by $\hat{\mathcal{S}}_t \triangleq \left\{\mathbf{x}\in S_t \mid \mathbf{w}_t^\top\mathbf{x} = 1\right\}$ (which might be empty for $t\geq 2$), and using the inductive assumption that $\mathbf{w}_{t-1}^{(\lambda)} = \mathbf{r}_{t-1}^{(\lambda)} + \ln\left(\frac{1}{\lambda}\right)\mathbf{w}_{t-1}$ (where $\left\|\mathbf{r}_{t-1}^{(\lambda)}\right\| = \mathcal{O}\left((t-1)\ln\ln\left(\frac{1}{\lambda}\right)\right)$) , we get

$$\begin{aligned}&\frac{1}{\lambda}\nabla\mathcal{L}_\lambda\Big(\left(\ln\left(\tfrac{1}{\lambda}\right) + s_t\ln\ln\left(\tfrac{1}{\lambda}\right)\right)\mathbf{w}_t + \tilde{\mathbf{w}}_t\Big)\\ &= -\frac{1}{\lambda}\left(\lambda\ln^{-s_t}\left(\tfrac{1}{\lambda}\right)\right)^1\sum_{\mathbf{x}\in\hat{\mathcal{S}}_t}\mathbf{x}e^{-\tilde{\mathbf{w}}_t^\top\mathbf{x}} - \frac{1}{\lambda}\sum_{\mathbf{x}\notin\hat{\mathcal{S}}_t}\mathbf{x}\left(\lambda\ln^{-s_t}\left(\tfrac{1}{\lambda}\right)\right)^{\mathbf{w}_t^\top\mathbf{x}}e^{-\tilde{\mathbf{w}}_t^\top\mathbf{x}} + \ln\left(\tfrac{1}{\lambda}\right)\mathbf{w}_t - \ln\left(\tfrac{1}{\lambda}\right)\mathbf{w}_{t-1} +\\ &\qquad + \ln\ln^{s_t}\left(\tfrac{1}{\lambda}\right)\mathbf{w}_t + \tilde{\mathbf{w}}_t - \mathbf{r}_{t-1}^{(\lambda)}\\ &= -\ln^{-s_t}\left(\tfrac{1}{\lambda}\right)\sum_{\mathbf{x}\in\hat{\mathcal{S}}_t}\mathbf{x}e^{-\tilde{\mathbf{w}}_t^\top\mathbf{x}} - \sum_{\mathbf{x}\notin\hat{\mathcal{S}}_t}\mathbf{x}\lambda^{\mathbf{w}_t^\top\mathbf{x}-1}\left(\ln\left(\tfrac{1}{\lambda}\right)\right)^{-s_t\mathbf{w}_t^\top\mathbf{x}}e^{-\tilde{\mathbf{w}}_t^\top\mathbf{x}} + \ln\left(\tfrac{1}{\lambda}\right)\left(\mathbf{w}_t - \mathbf{w}_{t-1}\right) +\\ &\qquad + \ln\ln^{s_t}\left(\tfrac{1}{\lambda}\right)\mathbf{w}_t + \tilde{\mathbf{w}}_t - \mathbf{r}_{t-1}^{(\lambda)}\,.\end{aligned}$$

By the triangle inequality and since $\left\|\ln\ln^{s_t}\left(\frac{1}{\lambda}\right)\right\| = \left\|s_t\ln\ln\left(\frac{1}{\lambda}\right)\right\| = \overbrace{|s_t|}^{=1}\ln\ln\left(\frac{1}{\lambda}\right)$, we have,

$$\begin{aligned}&\left\|\frac{1}{\lambda}\nabla\mathcal{L}_\lambda\Big(\left(\ln\left(\tfrac{1}{\lambda}\right) + s_t\ln\ln\left(\tfrac{1}{\lambda}\right)\right)\mathbf{w}_t + \tilde{\mathbf{w}}_t\Big)\right\|\\ &\leq \underbrace{\left\|\ln\left(\tfrac{1}{\lambda}\right)\left(\mathbf{w}_t - \mathbf{w}_{t-1}\right) - \ln^{-s_t}\left(\tfrac{1}{\lambda}\right)\sum_{\mathbf{x}\in\hat{\mathcal{S}}_t}\mathbf{x}e^{-\tilde{\mathbf{w}}_t^\top\mathbf{x}}\right\|}_{\triangleq\mathbf{a}_1(\lambda)} + \underbrace{\left\|\sum_{\mathbf{x}\notin\hat{\mathcal{S}}_t}\mathbf{x}\lambda^{\mathbf{w}_t^\top\mathbf{x}-1}\left(\ln\left(\tfrac{1}{\lambda}\right)\right)^{-s_t\mathbf{w}_t^\top\mathbf{x}}e^{-\tilde{\mathbf{w}}_t^\top\mathbf{x}}\right\|}_{\triangleq\mathbf{a}_2(\lambda)} +\\ &\qquad + \ln\ln\left(\tfrac{1}{\lambda}\right)\left\|\mathbf{w}_t\right\| + \left\|\tilde{\mathbf{w}}_t\right\| + \left\|\mathbf{r}_{t-1}^{(\lambda)}\right\|\end{aligned}$$

We now distinguish between two cases in which the behavior of the sequential max-margin differs greatly (see Eq. (2)).

1. **When $\mathbf{w}_t \triangleq \mathbf{P}_t(\mathbf{w}_{t-1}) \neq \mathbf{w}_{t-1}$ (and necessarily $|\hat{\mathcal{S}}_t| \geq 1$):** We choose $s_t = -1$ and $\mathbf{a}_1(\lambda)$ becomes:

$$\mathbf{a}_1(\lambda) = \left\|\ln\left(\tfrac{1}{\lambda}\right)(\mathbf{w}_t - \mathbf{w}_{t-1}) - \ln^{-s_t}\left(\tfrac{1}{\lambda}\right)\sum\nolimits_{\mathbf{x}\in\hat{\mathcal{S}}_t}\mathbf{x}e^{-\tilde{\mathbf{w}}_t^\top\mathbf{x}}\right\| = \ln\left(\tfrac{1}{\lambda}\right)\left\|(\mathbf{w}_t - \mathbf{w}_{t-1}) - \sum\nolimits_{\mathbf{x}\in\hat{\mathcal{S}}_t}\mathbf{x}e^{-\tilde{\mathbf{w}}_t^\top\mathbf{x}}\right\|.$$

   We thus wish to choose $\tilde{\mathbf{w}}_t$ so as to *zero* $\mathbf{a}_1(\lambda)$. Combined with the KKT conditions of Eq. (2), we require

$$\sum_{\mathbf{x}\in\hat{\mathcal{S}}_t}\mathbf{x}e^{-\tilde{\mathbf{w}}_t^\top\mathbf{x}} = \mathbf{w}_t - \mathbf{w}_{t-1} \overset{\text{KKT}}{=} \sum_{\mathbf{x}\in\hat{\mathcal{S}}_t}\mathbf{x}\alpha(\mathbf{x}),$$

   where $\boldsymbol{\alpha} \in \mathbb{R}_{>0}^{|S_t|}$ is the dual solution of the sequential max-margin problem (Eq. (2)). In Lemma B.3 and Corollary B.4 (applied with $\bar{\mathbf{B}} = \mathbf{I}$ and $\mathbf{v} = \mathbf{w}_{t-1}$) we show that such a vector $\tilde{\mathbf{w}}_t$ almost surely exists.

   Furthermore, since $\lim_{\lambda\to 0}\left(\lambda^{c-1}\ln^c(1/\lambda)\right) = 0,\ \forall c > 1$, it holds that $\mathbf{a}_2(\lambda)$ becomes

$$\mathbf{a}_2(\lambda) = \left\|\sum_{\mathbf{x}\notin\hat{\mathcal{S}}_t}\underbrace{\mathbf{x}e^{-\tilde{\mathbf{w}}_t^\top\mathbf{x}}}_{=\mathcal{O}(1)}\underbrace{\lambda^{\overbrace{\mathbf{w}_t^\top\mathbf{x}-1}^{>0}}\left(\ln(1/\lambda)\right)^{\mathbf{w}_t^\top\mathbf{x}}}_{\to 0}\right\| \xrightarrow{\lambda\to 0} 0.$$

   In conclusion, we can choose $\tilde{\mathbf{w}}_t$ and s_t such that $\|\tilde{\mathbf{w}}_t\| = \mathcal{O}(1)$, $\mathbf{a}_1(\lambda) = 0$, and $\mathbf{a}_2(\lambda) \to 0$.

2. **When $\mathbf{w}_t \triangleq \mathbf{P}_t(\mathbf{w}_{t-1}) = \mathbf{w}_{t-1}$ (and possibly $\hat{\mathcal{S}}_t = \emptyset$):** We choose $\tilde{\mathbf{w}}_t = \mathbf{0}_D$ and $s_t = 1$. It follows that

$$\mathbf{a}_1(\lambda) = \left\|\ln^{-s_t}\left(\tfrac{1}{\lambda}\right)\sum_{\mathbf{x}\in\hat{\mathcal{S}}_t}\mathbf{x}\right\| = \underbrace{\ln^{-1}\left(\tfrac{1}{\lambda}\right)}_{\to 0}\underbrace{\left\|\sum\nolimits_{\mathbf{x}\in\hat{\mathcal{S}}_t}\mathbf{x}\right\|}_{=\mathcal{O}(1)} \xrightarrow{\lambda\to 0} 0,$$

$$\mathbf{a}_2(\lambda) = \left\|\sum_{\mathbf{x}\notin\hat{\mathcal{S}}_t}\underbrace{\mathbf{x}}_{=\mathcal{O}(1)}\underbrace{\lambda^{\overbrace{\mathbf{w}_t^\top\mathbf{x}-1}^{>0}}}_{\to 0}\underbrace{\left(\ln\left(\tfrac{1}{\lambda}\right)\right)^{\overbrace{-\mathbf{w}_t^\top\mathbf{x}}^{<-1}}}_{\to 0}\right\| \xrightarrow{\lambda\to 0} 0.$$

As explained in our proof's idea (9), we now use the λ-strong convexity of our objective and Lemma B.2 to bound the distance to the optimum by

$$\begin{aligned}
\left\|\mathbf{r}_t^{(\lambda)}\right\| &\triangleq \left\|\mathbf{w}_t^{(\lambda)} - \ln\left(\tfrac{1}{\lambda}\right)\mathbf{w}_t\right\| = \left\|\mathbf{w}_t^{(\lambda)} - \left(\left(\ln\left(\tfrac{1}{\lambda}\right) + s_t\ln\ln\left(\tfrac{1}{\lambda}\right)\right)\mathbf{w}_t + \tilde{\mathbf{w}}_t\right) + s_t\ln\ln\left(\tfrac{1}{\lambda}\right)\mathbf{w}_t + \tilde{\mathbf{w}}_t\right\| \\
[\text{triangle ineq.}] &\leq \left\|\mathbf{w}_t^{(\lambda)} - \left(\left(\ln\left(\tfrac{1}{\lambda}\right) + s_t\ln\ln\left(\tfrac{1}{\lambda}\right)\right)\mathbf{w}_t + \tilde{\mathbf{w}}_t\right)\right\| + \left\|s_t\ln\ln\left(\tfrac{1}{\lambda}\right)\mathbf{w}_t\right\| + \left\|\tilde{\mathbf{w}}_t\right\| \\
[\text{Lemma B.2}] &\leq \frac{1}{\lambda}\left\|\nabla\mathcal{L}_\lambda\left(\left(\ln\left(\tfrac{1}{\lambda}\right) + s_t\ln\ln\left(\tfrac{1}{\lambda}\right)\right)\mathbf{w}_t + \tilde{\mathbf{w}}_t\right)\right\| + \ln\ln\left(\tfrac{1}{\lambda}\right)\left\|\mathbf{w}_t\right\| + \left\|\tilde{\mathbf{w}}_t\right\| \\
[\text{all the above}] &\leq \underbrace{\mathbf{a}_1(\lambda) + \mathbf{a}_2(\lambda)}_{\to 0} + 2\ln\ln\left(\tfrac{1}{\lambda}\right)\underbrace{\left\|\mathbf{w}_t\right\|}_{=\mathcal{O}(1)} + \underbrace{2\left\|\tilde{\mathbf{w}}_t\right\|}_{=\mathcal{O}(1)} + \underbrace{\left\|\mathbf{r}_{t-1}^{(\lambda)}\right\|}_{=\mathcal{O}\left((t-1)\ln\ln\left(\frac{1}{\lambda}\right)\right)}.
\end{aligned}$$

Finally, we conclude that $\left\|\mathbf{r}_t^{(\lambda)}\right\| = \mathcal{O}\left(t\ln\ln\left(\tfrac{1}{\lambda}\right)\right)$. □

B.3. Limitations of our Analysis

We proved that Lemma B.3 and Corollary B.4, which are pivotal for our proofs of Theorems 3.1, 5.1, and 5.4, hold for almost all datasets and weighting schemes. We took the approach of Soudry et al. (2018) who analyzed a simpler non-continual single-task case and derived similar results from the perspective of the roots of some polynomials. Below, we discuss the limitations of this analytical approach in our case.

1. **Task recurrence.** In our proof of Lemma B.3, we employed a construction where tasks do not recur. When tasks recur, the constructed polynomials have a higher-order dependence on the elements of $\mathbf{X}_1, \ldots, \mathbf{X}_T$ (T is the number of possible tasks, in contrast to the number of iterations k). Then, the analysis becomes more subtle due to some additional constraints on the number of support vectors at each iteration (N_t) (*e.g.,* when the same task is seen at iterations t and $(t+1)$, it must hold that $N_t = N_{t+1}$). Finding a construction for a general sequence of support sizes (N_t) is thus more challenging because not any sequence is attainable. Without such a general construction, it remains possible that task recurrence leads to a collapse into the measure-zero scenarios where the lemma does not hold.

   On the other hand, we *are* able to prove that the lemma holds for some cases where tasks recur. For instance, we can design constructions that allow task recurrence, as long as the support sizes are fixed and hold $N_t = N \leq D/2$, $\forall t \in [k]$. Specifically, we can construct a task sequence of k iterations over $T \leq k$ tasks (using a task ordering τ as in Def. 4.9). We construct the datasets so as to form a "cyclic" sequence with two types of tasks. That is, the columns of $\mathbf{X}_{\tau(t)} = \mathbf{X}_{\hat{\mathcal{S}}_{\tau(t)}} \in \mathbb{R}^{D\times N}$ are $\mathbf{x}_n^{\tau(t)} = \begin{cases} +\cos(\theta)\,\mathbf{e}_n + \sin(\theta)\,\mathbf{e}_{D-n} & \tau(t) = \tau(k) \\ -\cos(\theta)\,\mathbf{e}_n + \sin(\theta)\,\mathbf{e}_{D-n} & \tau(t) \neq \tau(k) \end{cases}$.
   Under this construction, which is easy to analyze (as in Appendix E.1), one can show that at the kth iteration, if $\mathbf{w}_k \neq c_k \mathbf{w}_{k-1}$ then $\boldsymbol{\alpha}_{\hat{\mathcal{S}}_k} = \gamma \mathbf{1}_N \succ \mathbf{0}$, for some $\gamma > 0$. This suffices for showing that the polynomials are nonzero (when $N_t = N \leq D/2$, $\forall t \in [k]$, for an arbitrary length k, with any form of recurrence of the task seen at iteration k), and conclude that recurrence does not necessarily collapse to measure zero events where the lemma does not hold.

2. **Weighting schemes.** We initially proved our lemma for the isotropic weighting scheme where $\mathbf{B}_t = \mathbf{I}$ (notably, this corresponds exactly to the case in our main result in Theorem 3.1). Subsequently, we used these isotropic weighting schemes to establish that our lemma applies to almost all weighting schemes as well. However, common weighting schemes, such as Fisher-information-based schemes, rely on the data observed in previously encountered tasks. Again, this makes the construction of a general task sequence where $\boldsymbol{\alpha}_{\hat{\mathcal{S}}_t} \succ \mathbf{0}$ more complicated. Thus, it is possible that such weighting schemes will collapse into the measure zero cases where the lemma does not hold.

We hypothesize that the reservations we expressed above are merely limitations of the analytical tools we utilized. Various simulations we conducted demonstrated an agreement between the weakly-regularized iterates and the Sequential Max-Margin iterates. Closing these gaps in our analysis will likely require an alternative analytical approach.

B.3.1. Example of a Measure Zero Case

Finally, we briefly demonstrate a measure zero case where Lemma B.3 and Corollary B.4 do not hold.

Let the first task be $\mathbf{X}_1 = [\mathbf{e}_1] \in \mathbb{R}^{D\times 1}$. Then, $\mathbf{w}_1 = \mathbf{e}_1$.

The second task is $\mathbf{X}_2 = [\mathbf{e}_1, \mathbf{e}_2] \in \mathbb{R}^{D\times 2}$. Then,

$$\begin{aligned} \mathbf{w}_2 &= \arg\min \|\mathbf{w} - \mathbf{w}_1\| \text{ s.t. } \left(\mathbf{e}_1^\top \mathbf{w} \geq 1\right) \wedge \left(\mathbf{e}_2^\top \mathbf{w} \geq 1\right) \\ &= \mathbf{e}_1 + \mathbf{e}_2 \,. \end{aligned}$$

Both $\mathbf{e}_1, \mathbf{e}_2$ are support vectors of the 2nd task, but from the stationarity condition (plug in $\mathbf{B}_2 = \mathbf{I}$, $c_2 = 1$ into Eq. (5)), it holds that,

$$\alpha_2(\mathbf{e}_1)\,\mathbf{e}_1 + \alpha_2(\mathbf{e}_2)\,\mathbf{e}_2 = \mathbf{w}_2 - \mathbf{w}_1 = \mathbf{e}_2 \,,$$

thus requiring that $\alpha_2(\mathbf{e}_1) = 0$. Crucially, this prevents the existence of a finite $\tilde{\mathbf{w}}_2$ that holds $\exp\left(-\tilde{\mathbf{w}}_2^\top \mathbf{e}_1\right) = \alpha_2(\mathbf{e}_1)$.

C. Proofs for the Sequential Max-Margin Projections Scheme (Section 4)

Recall Lemma 4.3. Recall our definition of $R \triangleq \max_{m\in[T]} \max_{(\mathbf{x},y)\in S_m} \|\mathbf{x}\|$. The quantities of Def. 4.1 are related as follows:

$$\forall \mathbf{w}\in\mathbb{R}^D,\ m\in[T]:\quad F_m(\mathbf{w}) \le d^2(\mathbf{w},\mathcal{W}_m) \max_{(\mathbf{x},y)\in S_m} \|\mathbf{x}\|^2 \le d^2(\mathbf{w},\mathcal{W}^\star)R^2\,.$$

Moreover, for the Sequential Max-Margin iterates $(\mathbf{w}_t)$ of Scheme 2, all quantities are upper bounded by the "problem complexity", *i.e.,* $d^2(\mathbf{w}_t,\mathcal{W}^\star)R^2 \le \|\mathbf{w}^\star\|^2 R^2$.

Proof. In our proof, we use the simplifying mapping from Remark B.1 ($y\mathbf{x} \longmapsto \mathbf{x}$).

Using simple algebra and the Cauchy-Schwarz inequality, we have $\forall m \in [T]$, $\mathbf{x} \in S_m$, $\mathbf{w} \in \mathbb{R}^D$,

$$\begin{aligned}1-\mathbf{w}^\top\mathbf{x} &= 1-\underbrace{\left(\mathbf{P}_m\left(\mathbf{w}\right)\right)^\top\mathbf{x}}_{\ge 1\ \text{(see below)}} + \left(\mathbf{P}_m\left(\mathbf{w}\right)\right)^\top\mathbf{x} - \mathbf{w}^\top\mathbf{x} \le \left(\mathbf{P}_m\left(\mathbf{w}\right)-\mathbf{w}\right)^\top\mathbf{x}\\ &[\text{Cauchy–Schwarz}] \le \left\|\mathbf{w}-\mathbf{P}_m\left(\mathbf{w}\right)\right\|\|\mathbf{x}\| = d(\mathbf{w},\mathcal{W}_m)\|\mathbf{x}\|\,,\end{aligned}$$

where the inequality in the underbrace stems from the fact that $\mathbf{P}_m(\mathbf{w}) \in \mathcal{W}_m$, and that $\mathcal{W}_m$ is defined (Assumption 2.1) as the set of solutions with a zero hinge loss over the samples in S_m (such as $\mathbf{x}$).

Since $\mathbf{w}^\star$ is the intersection of all $\mathcal{W}_m$, we get that $d(\mathbf{w},\mathcal{W}_m) \le d(\mathbf{w},\mathcal{W}^\star)$.

Overall, it follows that

$$F_m(\mathbf{w}) \triangleq \max_{(\mathbf{x},y)\in S_m}\left(\max\left\{0,\,1-\mathbf{w}^\top\mathbf{x}\right\}\right)^2 \le \max_{\mathbf{x}\in S_m}\Big(d(\mathbf{w},\mathcal{W}_m)\|\mathbf{x}\|\Big)^2 = d^2(\mathbf{w},\mathcal{W}_m)\max_{\mathbf{x}\in S_m}\|\mathbf{x}\|^2 \le d^2(\mathbf{w},\mathcal{W}^\star)R^2.$$

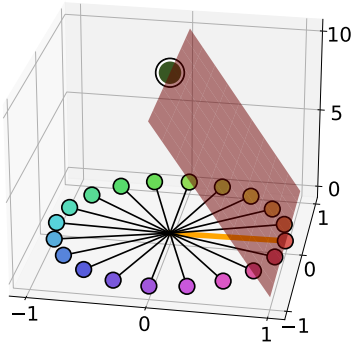
Finally, due to the monotonicity from Lemma 4.5, the iterates of Scheme 2 hold that,

$$d^2(\mathbf{w}_t,\mathcal{W}^\star) \le d^2(\mathbf{w}_0,\mathcal{W}^\star) = \|\mathbf{w}^\star\|^2\,,\quad \forall t\in\mathbb{N}^+\,.$$

□

C.1. Adversarial Construction: Additional Discussion and Illustrations (Section 4.2)

Figure 7. Elaboration on the construction in Figure 2a. We always use $D = 3$ dimensions. Here, we demonstrate our construction with only $T = 20$ tasks. Each task consists of a single datapoint $\mathbf{x}_m$ which is positively labeled (*i.e.,* $y_m = +1$) and has a norm of $\|\mathbf{x}_m\| = 1 = R$. Datapoints are uniformly spread on a plane slightly elevated above the xy-plane. We also plot the plane induced by a specific datapoint $\mathbf{x}_m$ (highlighted in orange), which is the boundary of its feasible set $\mathcal{W}_m \triangleq \{\mathbf{w} \in \mathbb{R}^D \mid \mathbf{w}^\top \mathbf{x}_m \geq 1\}$. All such planes intersect at $\mathbf{w}^\star$. The elevation of the datapoints above the xy-plane determines the magnitude of $\|\mathbf{w}^\star\|$ (lower elevation implies a worse minimum margin and a larger $\|\mathbf{w}^\star\|$). In our experiment, we set $\|\mathbf{w}^\star\| = 10$. As the number of tasks $T \to \infty$, the uniform angles between consecutive tasks and the applied projections become smaller.



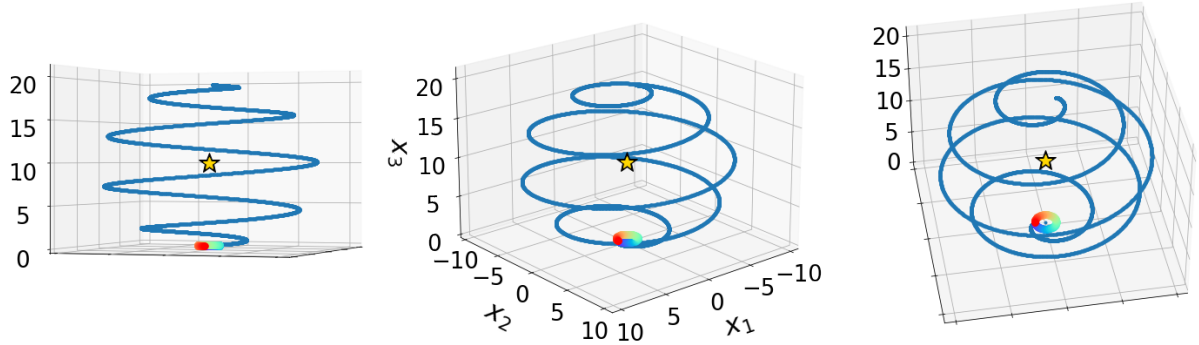


Figure 8. Plotting the path of the iterates $(\mathbf{w}_t)$ during several cycles on $T = 1,000$ tasks (forming a "denser" circle than the one in Figure 2a). Notice how the learner "misses" the minimum norm solution $\mathbf{w}^\star = (0, 0, 10)$, which is indicated by a star. Instead, the iterates seem to converge near $2\mathbf{w}^\star$, still holding the guarantee in Theorem 4.7. Moreover, since at the end of learning it approximately holds that $\|\mathbf{w}_t - \mathbf{w}^\star\| \approx \|2\mathbf{w}^\star - \mathbf{w}^\star\| = 2\|\mathbf{w}^\star\| = 20$, then by the monotonicity that we prove in Corollary D.1, *i.e.,* that $\|\mathbf{w}_t - \mathbf{w}^\star\| \leq \|\mathbf{w}_{t'} - \mathbf{w}^\star\| \leq \|\mathbf{w}_0 - \mathbf{w}^\star\|$, $\forall t' \leq t$, we understand that the iterates remain at a distance of approximately 10 throughout the entire learning process, thus approximately residing on a sphere centered at $\mathbf{w}^\star$ with a radius of $\|\mathbf{w}^\star\|$.

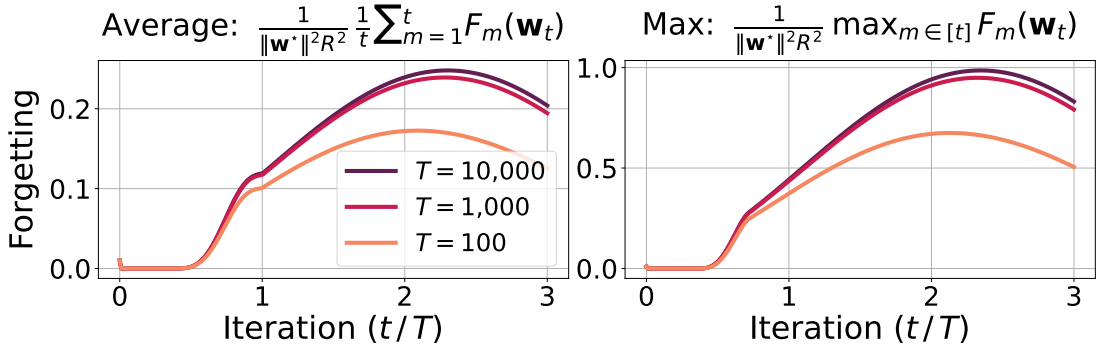


Figure 9. This is a larger version of Figure 2b. The average and maximum forgetting for the adversarial construction. In this experiment, we trained on 3 cycles of the same T tasks, gradually increasing T. Notably, after learning $T \to \infty$ jointly separable tasks, the quantities do *not* decay. Recall that the maximum forgetting lower bounds the distance to the offline feasible set, which is upper bounded by $\|\mathbf{w}^\star\|^2 R^2$ (Lemma 4.3). Hence, the quantities of interest appear to become *arbitrarily bad* at some point during the 3rd cycle of learning, in the sense that they reach $\|\mathbf{w}^\star\|^2 R^2$.

C.2. Convergence to the Minimum-Norm Solution (Section 4.3)

Recall Theorem 4.7. Any iterate $\mathbf{w}_t$ obtained by Scheme 2 holds $\|\mathbf{w}_t\| \leq 2\|\mathbf{w}^\star\|$, where $\mathbf{w}^\star$ is the minimum-norm offline solution (Def. 2.2).

If additionally, $\mathbf{w}_t$ is "offline"-feasible, *i.e.,* $\mathbf{w}_t \in \mathcal{W}^\star$, then

$$\|\mathbf{w}^\star\| \leq \|\mathbf{w}_t\| \leq 2\|\mathbf{w}^\star\| \ .$$

Proof. To prove the first part of the theorem, we use the triangle inequality, the monotonicity from Corollary D.1 (below), and the fact that $\mathbf{w}_0 = \mathbf{0}_d$, to show that

$$\forall t \in \mathbb{N}^+ : \ \ \|\mathbf{w}_t\| = \|\mathbf{w}_t - \mathbf{w}_0\| = \|\mathbf{w}_t - \mathbf{w}^\star + \mathbf{w}^\star - \mathbf{w}_0\| \leq \|\mathbf{w}_0 - \mathbf{w}^\star\| + \|\mathbf{w}_t - \mathbf{w}^\star\| \overset{D.1}{\leq} 2\|\mathbf{w}_0 - \mathbf{w}^\star\| = 2\|\mathbf{w}^\star\| \ .$$

The second part of the theorem follows immediately from the minimality of $\mathbf{w}^\star \in \mathcal{W}^\star$ (Def. 2.2). □

D. Proofs for Recurring Tasks (Section 4.4)

D.1. General Properties

We start by stating a few general properties of projections and norms that will help us throughout the appendices.

Property 1 (Projection properties). Let $\mathbf{P} : \mathbb{R}^D \to \mathbb{R}^D$ be a projection operator onto a nonempty closed convex set $\mathcal{C} \subseteq \mathbb{R}^D$. Then, $\mathbf{P}$ holds the following properties:

1. **Geometric definition.** $\mathbf{P}(\mathbf{v}) = \operatorname{argmin}_{\mathbf{w}\in\mathcal{C}} \|\mathbf{v}-\mathbf{w}\|$ for any $\mathbf{v} \in \mathbb{R}^D$;
2. **Idempotence.** $\mathbf{P}^2 \triangleq \mathbf{P} \circ \mathbf{P} = \mathbf{P}$;
3. **Contraction (Non-expansiveness).** For all $\mathbf{v}, \mathbf{u} \in \mathbb{R}^D$, it holds that $\|\mathbf{P}(\mathbf{v}) - \mathbf{P}(\mathbf{u})\| \leq \|\mathbf{v}-\mathbf{u}\|$. Consequently, when $\mathbf{0}_D \in \mathcal{C}$, it holds that $\|\mathbf{P}(\mathbf{v})\| \leq \|\mathbf{v}\|$ (see Fact 1.9 in Deutsch & Hundal (2006b)) ;
4. **The operator $\mathbf{I} - \mathbf{P}$.** Let $\mathbf{I}(\mathbf{v}) = \mathbf{v}$ be the identity operator and define the operator $(\mathbf{I}-\mathbf{P})(\mathbf{v}) = \mathbf{v} - \mathbf{P}(\mathbf{v})$. Then, for all $\mathbf{v}, \mathbf{u} \in \mathbb{R}^D$, it holds that $\|(\mathbf{I}-\mathbf{P})(\mathbf{v}) - (\mathbf{I}-\mathbf{P})(\mathbf{u})\|^2 \leq \|\mathbf{v}-\mathbf{u}\|^2 - \|\mathbf{P}(\mathbf{v}) - \mathbf{P}(\mathbf{u})\|^2$. Consequently, when $\mathbf{0}_D \in \mathcal{C}$, it holds that $\|\mathbf{v} - \mathbf{P}(\mathbf{v})\|^2 \leq \|\mathbf{v}\|^2 - \|\mathbf{P}(\mathbf{v})\|^2 \leq \|\mathbf{v}\|^2$ for any $\mathbf{v} \in \mathbb{R}^D$ (see Propositions 4.2 and 4.8 in Bauschke et al. (2011) and Fact 1.7 in Deutsch & Hundal (2006b)).

As a result of the non-expansiveness property above, we get the following monotonicity result (stronger than 4.5).

Corollary D.1. *Let $\mathbf{w} \in \mathcal{W}^\star$ be an arbitrary offline solution and $\mathbf{P}_t$ be the projection onto the feasible set of the tth task. The following monotonicity holds for the iterates $(\mathbf{w}_t)$ of Scheme 2:*

$$\|\mathbf{w}_t - \underbrace{\mathbf{w}}_{\in\mathcal{W}_t}\| = \|\mathbf{P}_t(\mathbf{w}_{t-1}) - \mathbf{P}_t(\mathbf{w})\| \leq \|\mathbf{w}_{t-1} - \mathbf{w}\|, \quad \forall t \in \mathbb{N}^+ .$$

Specifically, this holds for the minimum-norm solution $\mathbf{w}^\star$: $$\|\mathbf{w}_t - \mathbf{w}^\star\| \leq \|\mathbf{w}_{t-1} - \mathbf{w}^\star\|, \quad \forall t \in \mathbb{N}^+ .$$

Next, we state a known property of squared Euclidean norms, stemming from their convexity and from Jensen's inequality.

Claim D.2. For any m vectors $\mathbf{v}_1, \ldots, \mathbf{v}_m \in \mathbb{R}^D$, it holds that $\|\mathbf{v}_1 + \ldots + \mathbf{v}_m\|^2 \leq m\Big(\|\mathbf{v}_1\|^2 + \cdots + \|\mathbf{v}_m\|^2\Big)$.

Lemma D.3. *The residual from task m, defined by $\|\mathbf{v} - \mathbf{P}_m(\mathbf{v})\|^2 = \|(\mathbf{I}-\mathbf{P}_m)(\mathbf{v})\|^2$, is a convex function in $\mathbf{v} \in \mathbb{R}^D$.*

Proof. Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^D$, than for every $\alpha \in [0,1]$ it holds that

$$\begin{aligned}
\|(\mathbf{I}-\mathbf{P}_m)(\alpha\mathbf{u} + (1-\alpha)\mathbf{v})\|^2 &= \|\alpha\mathbf{u} + (1-\alpha)\mathbf{v} - \mathbf{P}_m(\alpha\mathbf{u} + (1-\alpha)\mathbf{v})\|^2 \\
[\text{projection properties}] &\triangleq \min_{\mathbf{w}\in\mathcal{W}_m} \|\alpha\mathbf{u} + (1-\alpha)\mathbf{v} - \mathbf{w}\|^2 \leq \Big\|\alpha\mathbf{u} + (1-\alpha)\mathbf{v} - \underbrace{(\alpha\overbrace{\mathbf{P}_m(\mathbf{u})}^{\in\mathcal{W}_m} + (1-\alpha)\overbrace{\mathbf{P}_m(\mathbf{v})}^{\in\mathcal{W}_m})}_{\in\mathcal{W}_m,\ \text{due to its convexity}}\Big\|^2 \\
&= \|\alpha(\mathbf{u} - \mathbf{P}_m(\mathbf{u})) + (1-\alpha)(\mathbf{v} - \mathbf{P}_m(\mathbf{v}))\|^2 \\
[\text{squared norm is convex}] &\leq \alpha\|\mathbf{u} - \mathbf{P}_m(\mathbf{u})\|^2 + (1-\alpha)\|\mathbf{v} - \mathbf{P}_m(\mathbf{v})\|^2 .
\end{aligned}$$

□

D.2. Linear Regularity of Classification Tasks

We start by proving an important property of our setting – that the feasible set has a nonempty interior.

Lemma D.4. *Recall our definition of* $R \triangleq \max_{m\in[T]} \max_{(\mathbf{x},y)\in S_m} \|\mathbf{x}\|$. *Let* $\mathbf{w}^\star$ *be the minimum-norm solution defined in Def. 2.2. Then, the vector* $2\mathbf{w}^\star$ *has a* feasible *ball of radius* $\frac{1}{R}$ *around it. More formally,*

$$\mathcal{B}\left(2\mathbf{w}^\star, 1/R\right) \triangleq \left\{\mathbf{w} \in \mathbb{R}^D \;\middle|\; \|\mathbf{w} - 2\mathbf{w}^\star\| \leq 1/R\right\} \subset \mathcal{W}^\star \,.$$

Proof. By the definition of feasibility, $\forall m \in [T]\,,\, \forall (\mathbf{x}, y) \in S_m\colon\; y\mathbf{w}^{\star\top}\mathbf{x} \geq 1$. Let $\mathbf{w}$ be an arbitrary vector in $\mathcal{B}\left(2\mathbf{w}^\star, 1/R\right)$. Clearly, $\mathbf{w}$ can be instead denoted as $\mathbf{w} = 2\mathbf{w}^\star + \mathbf{v}$ for some $\mathbf{v}$ such that $\|\mathbf{v}\| \leq 1/R$. We then conclude that $\mathbf{w}$ is feasible:

$$\forall m \in [T]\,,\, \forall (\mathbf{x}, y) \in S_m\colon\; y\mathbf{w}^\top\mathbf{x} = 2\underbrace{y\mathbf{w}^{\star\top}\mathbf{x}}_{\geq 1} + y\mathbf{v}^\top\mathbf{x} \geq 2 - \left|\mathbf{v}^\top\mathbf{x}\right| \geq 2 - \underbrace{\|\mathbf{v}\|}_{\leq 1/R}\underbrace{\|\mathbf{x}\|}_{\leq R} \geq 1 \implies \mathbf{w} \in \mathcal{W}^\star \,.$$

□

We are now ready to prove the linear regularity, using Lemma D.4 and techniques from Gubin et al. (1967) (in their proof of Lemma 5) and Nedić (2010) (in their proof of Proposition 8). To be more exact, we prove *bounded* linear regularity (as defined in Bauschke & Borwein (1993); Deutsch & Hundal (2008)), which suffices for our needs in this paper.

Recall Lemma 4.8. At the tth iteration, the distance to the offline feasible set is tied to the distance to the farthest feasible set of any *specific* task. Specifically, it holds that $\forall t \in \mathbb{N}^+$,

$$d^2(\mathbf{w}_t, \mathcal{W}^\star) \leq 4\|\mathbf{w}^\star\|^2 R^2 \max_{m\in[T]} d^2\big(\mathbf{w}_t, \mathcal{W}_m\big) \,.$$

Proof. We start by showing (bounded) regularity for an arbitrary $\mathbf{w} \in \mathbb{R}^D$ (not necessarily an iterate). We define $\varepsilon \triangleq \max_{m\in[T]} d\left(\mathbf{w}, \mathcal{W}_m\right)$ and consider a convex combination between $\mathbf{w}$ and $2\mathbf{w}^\star$:

$$\begin{aligned}\mathbf{y} &= \frac{1/R}{\varepsilon + 1/R}\mathbf{w} + \frac{\varepsilon}{\varepsilon + 1/R}2\mathbf{w}^\star = \frac{1/R}{\varepsilon + 1/R}\mathbf{w} + \frac{\varepsilon}{\varepsilon + 1/R}2\mathbf{w}^\star - \frac{1/R}{\varepsilon + 1/R}\mathbf{P}_m(\mathbf{w}) + \frac{1/R}{\varepsilon + 1/R}\mathbf{P}_m(\mathbf{w}) \\ &\triangleq \frac{\varepsilon}{\varepsilon + 1/R}\mathbf{z}_m + \frac{1/R}{\varepsilon + 1/R}\mathbf{P}_m(\mathbf{w}) \,,\end{aligned}$$

where we denoted $\mathbf{z}_m \triangleq 2\mathbf{w}^\star + \frac{1/R}{\varepsilon}\left(\mathbf{w} - \mathbf{P}_m(\mathbf{w})\right)$, for any $m \in [T]$.

We notice that $\mathbf{z}_m \in \mathcal{B}\left(2\mathbf{w}^\star, 1/R\right) \overset{\text{D.4}}{\subset} \mathcal{W}^\star \subseteq \mathcal{W}_m$, since $\|\mathbf{z}_m - 2\mathbf{w}^\star\| = \frac{1/R}{\varepsilon}\underbrace{\|\mathbf{w} - \mathbf{P}_m(\mathbf{w})\|}_{=d(\mathbf{w},\mathcal{W}_m)\leq\varepsilon} \leq 1/R$. Therefore, $\mathbf{y}$ is a convex combination of $\mathbf{z}_m, \mathbf{P}_m\left(\mathbf{w}\right) \in \mathcal{W}_m$ and is therefore also contained in $\mathcal{W}_m$. Since this is true $\forall m \in [T]$, we get that $\mathbf{y} \in \mathcal{W}^\star$. Then, a bounded linear regularity property follows, since

$$d(\mathbf{w}, \mathcal{W}^\star) \leq \|\mathbf{w} - \mathbf{y}\| = \overbrace{\frac{\varepsilon}{\varepsilon + 1/R}}^{\leq R\varepsilon}\|\mathbf{w} - 2\mathbf{w}^\star\| \leq \|\mathbf{w} - 2\mathbf{w}^\star\| R\varepsilon = \|\mathbf{w} - 2\mathbf{w}^\star\| R \max_{m\in[T]} d\big(\mathbf{w}, \mathcal{W}_m\big) \,.$$

When $\mathbf{w} = \mathbf{w}_t$ is an iterate of the sequential Scheme 2, we can use the monotonicity from Corollary D.1 (notice that $2\mathbf{w}^\star \in \mathcal{W}^\star$) to show that $\|\mathbf{w}_t - 2\mathbf{w}^\star\| \leq \|\mathbf{w}_0 - 2\mathbf{w}^\star\| = \|2\mathbf{w}^\star\| = 2\|\mathbf{w}^\star\|$, and finally, conclude that

$$d(\mathbf{w}_t, \mathcal{W}^\star) \leq 2\|\mathbf{w}^\star\| R \max_{m\in[T]} d\big(\mathbf{w}_t, \mathcal{W}_m\big) \,.$$

□

D.3. Proofs for Cyclic Orderings (Section 4.4.1)

Recall Lemma 4.11. Under a cyclic ordering (and the separability assumption 2.1), the iterates converge to a 2-optimal $\mathbf{w}_\infty \in \mathcal{W}^\star$. That is,

$$\lim_{k\to\infty} d(\mathbf{w}_k, \mathcal{W}^\star) = 0, \quad \|\mathbf{w}^\star\| \leq \|\mathbf{w}_\infty\| \leq 2\,\|\mathbf{w}^\star\| \ .$$

Proof. First, notice that we work in a finite-dimensional Euclidean space $\mathbb{R}^D$ and that from Assumption 2.1 we have that the offline feasibility set, *i.e.,* the convex sets' intersection $\mathcal{W}^\star = \mathcal{W}_1 \cap \cdots \cap \mathcal{W}_T$, is nonempty. Under these conditions, the (weak and strong) convergence of the cyclic iterates $(\mathbf{w}_t)$ to a vector $\mathbf{w}_\infty \in \mathcal{W}^\star$ can be deduced from the rates we derive (independently of the lemma here) in Proposition 4.13. The second part of our corollary stems directly from our Theorem 4.7.

Similar optimality guarantees for cyclic settings have also been proved in previous papers (*e.g.,* Theorem 1 in Gubin et al. (1967) and Theorem 2.7 in Deutsch & Hundal (2006a)). □

Recall Theorem 4.12. For $T \geq 2$ jointly-separable tasks learned cyclically, after $k = nT$ iterations (n cycles), our quantities of interest (Def. 4.1) converge linearly as

$$\underbrace{\max_{m\in[T]} F_m(\mathbf{w}_k)}_{\text{Maximum forgetting}} \leq \underbrace{\max_{m\in[T]} d^2(\mathbf{w}_k, \mathcal{W}_m)}_{\substack{\text{Maximum distance}\\ \text{to any feasible set}}} R^2 \leq \underbrace{d^2(\mathbf{w}_k, \mathcal{W}^\star)}_{\substack{\text{Dist. to}\\ \text{offline feasible set}}} R^2 \leq g(k) \left\|\mathbf{w}^\star\right\|^2 R^2 \,,$$

where $g(k) \triangleq \begin{cases} \exp\left(-\frac{k}{4\|\mathbf{w}^\star\|^2 R^2}\right) & T = 2 \\ 4\exp\left(-\frac{k}{16T^2\|\mathbf{w}^\star\|^2 R^2}\right) & T \geq 3 \end{cases}$ and $\|\mathbf{w}^\star\|^2 R^2 \geq 1$ is the problem complexity (Rem. 4.4).

Proof. Here, we exploit the regularity of our setting, in which the convex sets are defined by the intersections of halfspaces.

1. **Cyclic orderings of $T = 2$ tasks.**

   Given the regularity of our problems (Lemma 4.8), the required upper bound on $d^2(\mathbf{w}_k, \mathcal{W}^\star)$ is given from existing results on projection algorithms with two closed convex sets (specifically from Theorem 3.12 and Corollary 3.14 in Bauschke & Borwein (1993) and Theorem 2.10 in Bauschke (2001)). However, the indexing in some of these previous works may be confusing. Therefore, for the sake of completeness, we prove it here as well.

   Assume w.l.o.g. that $2 \mid k$. Define the projection operator $\mathbf{P}_{1\cap 2}$ onto the intersection $\mathcal{W}^\star = \mathcal{W}_1 \cap \mathcal{W}_2$. It follows that,

$$\begin{aligned}
\frac{1}{4\|\mathbf{w}^\star\|^2 R^2} d^2(\mathbf{w}_k, \mathcal{W}^\star) &\overset{\text{Lemma 4.8}}{\leq} \max_{m\in\{1,2\}} d^2(\mathbf{w}_k, \mathcal{W}_m) \overset{\text{cyclic}}{=} \|\mathbf{w}_k - \mathbf{P}_1(\mathbf{w}_k)\|^2 \\
&= \|\mathbf{w}_k - \mathbf{P}_1(\mathbf{w}_k) - \mathbf{P}_{1\cap 2}(\mathbf{w}_k) + \mathbf{P}_{1\cap 2}(\mathbf{w}_k)\|^2 \\
&= \|\mathbf{w}_k - \mathbf{P}_1(\mathbf{w}_k) - \mathbf{P}_{1\cap 2}(\mathbf{w}_k) + \mathbf{P}_1(\mathbf{P}_{1\cap 2}(\mathbf{w}_k))\|^2 \\
&= \|(\mathbf{I} - \mathbf{P}_1)(\mathbf{w}_k) - (\mathbf{I} - \mathbf{P}_1)(\mathbf{P}_{1\cap 2}(\mathbf{w}_k))\|^2 \\
[\text{Prop. 1}] &\leq \|\mathbf{w}_k - \mathbf{P}_{1\cap 2}(\mathbf{w}_k)\|^2 - \|\mathbf{P}_1(\mathbf{w}_k) - \mathbf{P}_1(\mathbf{P}_{1\cap 2}(\mathbf{w}_k))\|^2 \\
&= \underbrace{\|\mathbf{w}_k - \mathbf{P}_{1\cap 2}(\mathbf{w}_k)\|^2}_{=d^2(\mathbf{w}_k, \mathcal{W}^\star)} - \underbrace{\|\mathbf{w}_{k+1} - \mathbf{P}_{1\cap 2}(\mathbf{w}_k)\|^2}_{\geq d^2(\mathbf{w}_{k+1}, \mathcal{W}^\star)} \leq d^2(\mathbf{w}_k, \mathcal{W}^\star) - d^2(\mathbf{w}_{k+1}, \mathcal{W}^\star)
\end{aligned}$$

$$d^2(\mathbf{w}_{k+1}, \mathcal{W}^\star) \leq \left(1 - \frac{1}{4\|\mathbf{w}^\star\|^2 R^2}\right) d^2(\mathbf{w}_k, \mathcal{W}^\star) \leq \cdots \leq \left(1 - \frac{1}{4\|\mathbf{w}^\star\|^2 R^2}\right)^{k+1} \underbrace{d^2(\mathbf{w}_0, \mathcal{W}^\star)}_{\triangleq \|\mathbf{w}^\star\|^2} .$$

   Finally, we conclude the $T = 2$ case by using the ordering property of our quantities of interest (Lemma 4.3), and using the algebraic identity stating that $\forall z \in (0,1)\,, k > 0 :\ \ (1-z)^k \leq \exp(-kz)$.

   Remark D.5. A similar upper bound on $d^2(\mathbf{w}_k, \mathcal{W}^\star)$, with a slightly worse rate, could be deduced from the stochastic upper bound in Theorem 4.15, by using the fact that for $T = 2$ tasks, random orderings lead to exactly the same projections as cyclic ones, but slower (the idempotence of the projections implies that applying the same projection many consecutive times in a random ordering has the same effect as applying it once like in a cyclic ordering).

2. **Cyclic orderings of $T \geq 3$ tasks.**

   Given the regularity of our problems (Lemma 4.8), the required upper bound is an almost immediate corollary from Theorem 3.15 in Deutsch & Hundal (2008)), stating that in our case we have

$$\|\mathbf{w}_k - \mathbf{w}_\infty\|^2 \leq \left(1 - \frac{1}{4T \cdot 4\left\|\mathbf{w}^\star\right\|^2 R^2}\right)^{k/T} \|\mathbf{w}_0 - \mathbf{w}_\infty\|^2 = \left(1 - \frac{1}{16T\left\|\mathbf{w}^\star\right\|^2 R^2}\right)^{k/T} \|\mathbf{w}_\infty\|^2 .$$

   We use Lemma 4.11 ($\mathbf{w}_\infty \in \mathcal{W}^\star, \|\mathbf{w}_\infty\| \leq 2\left\|\mathbf{w}^\star\right\|$) and the aforementioned algebraic identity, and get

$$d^2(\mathbf{w}_k, \mathcal{W}^\star) \leq \|\mathbf{w}_k - \mathbf{w}_\infty\|^2 \leq 4\left\|\mathbf{w}^\star\right\|^2 \exp\left(-\frac{k}{16T^2\left\|\mathbf{w}^\star\right\|^2 R^2}\right) .$$

□

D.3.1. Detour: Universal bounds for general cyclic Projections Onto Convex Sets settings (POCS)

First, we establish a lemma equivalent to Lemma 22 in Evron et al. (2022) for our POCS setting. They used *linear* projections solely while we analyze more general projection operators. Thus, while we *closely* follow their statements and proofs, we are required to perform several adjustments to capture the wider family of convex operators.

Lemma D.6 (Cyclic-case auxiliary bounds)**.** *Let $\mathbf{Q}_1, \dots, \mathbf{Q}_T : \mathbb{R}^D \to \mathbb{R}^D$ be T projection operators onto the nonempty closed convex sets $\mathcal{C}_1, \dots, \mathcal{C}_T$ (respectively) such that $\mathbf{0}_D \in \mathcal{C}_1 \cap \dots \cap \mathcal{C}_T$. Let $\mathbf{M} = \mathbf{Q}_T \circ \dots \circ \mathbf{Q}_1 : \mathbb{R}^D \to \mathbb{R}^D$ be the cyclic operator formed by these projections. Moreover, let $\mathbf{v} \in \mathbb{R}^D$ be an arbitrary vector. Then:*

- ***(Lemma D.6a)*** *For any $m \in [T-1]$, it holds that* $$d^2(\mathbf{v}, \mathcal{C}_m) = \|\mathbf{v} - \mathbf{Q}_m(\mathbf{v})\|^2 \le m\left(\|\mathbf{v}\|^2 - \|\mathbf{M}(\mathbf{v})\|^2\right) ;$$
- ***(Lemma D.6b)*** *It holds that* $$\|\mathbf{v} - \mathbf{M}(\mathbf{v})\|^2 \le T\left(\|\mathbf{v}\|^2 - \|\mathbf{M}(\mathbf{v})\|^2\right) ;$$
- ***(Lemma D.6c)*** *After $n \ge 1$ cycles it holds that* $$\|\mathbf{M}^n(\mathbf{v})\|^2 - \left\|\mathbf{M}^{n+1}(\mathbf{v})\right\|^2 \le 2\left\|\mathbf{M}^n(\mathbf{v}) - \mathbf{M}^{n+1}(\mathbf{v})\right\| \|\mathbf{v}\| ;$$
- ***(Lemma D.6d)*** *For any $m \in [T-1]$, after $n \ge 1$ cycles it holds that* $$d^2(\mathbf{M}^n(\mathbf{v}), \mathcal{C}_m) = \|\mathbf{M}^n(\mathbf{v}) - \mathbf{Q}_m(\mathbf{M}^n(\mathbf{v}))\|^2 \le 2m\left\|\mathbf{M}^n(\mathbf{v}) - \mathbf{M}^{n+1}(\mathbf{v})\right\| \|\mathbf{v}\| ;$$
- ***(Lemma D.6e)*** *For any number of cycles $n \ge 1$, it holds that* $$\left\|\mathbf{M}^{n-1}(\mathbf{v}) - \mathbf{M}^n(\mathbf{v})\right\| \le \sqrt{T/n}\,\|\mathbf{v}\| .$$

We will prove this lemma after stating a few additional results including our universal bounds.

Remark D.7 (Translates of convex sets). For simplicity, many of our results (*e.g.,* Lemma D.6 above) are derived using nonempty closed convex sets having $\mathbf{0}_D$ in their intersection. We wish to apply these results to our feasible sets $\mathcal{W}_1, \dots, \mathcal{W}_T$ from Assumption 2.1, but their nonempty intersection $\mathcal{W}^\star$ does not contain $\mathbf{0}_D$. As a remedy, it is common to translate the entire space by some feasible solution (*e.g.,* see Deutsch & Hundal (2006a)).

We take this approach and translate our space by $-\mathbf{w}^\star$ (where $\mathbf{w}^\star \in \mathcal{W}$ is the minimum-norm offline solution defined in Def. 2.2). In turn, the resulting feasible sets $C_m \triangleq \mathcal{W}_m - \mathbf{w}^\star$, $\forall m \in [T]$, now have an equivalent intersection $\mathcal{W}^\star - \mathbf{w}^\star$ that *does* contain $\mathbf{0}_D$. Importantly, this simple translation does not change the properties of the space we work in or the ones of any projection that we apply. In particular, we can apply either "type" of these projections interchangeably, in the sense that:

$$\mathbf{P}_m(\mathbf{w}) \triangleq \mathbf{P}_{\mathcal{W}_m}(\mathbf{w}) = \mathbf{P}_{\mathcal{C}_m}(\mathbf{w} - \mathbf{w}^\star) + \mathbf{w}^\star \triangleq \mathbf{Q}_m(\mathbf{w} - \mathbf{w}^\star) + \mathbf{w}^\star .$$

This also holds recursively, *e.g.,* $\mathbf{w}_2 = \mathbf{P}_2(\mathbf{P}_1(\mathbf{w}_0)) = \mathbf{P}_2(\mathbf{Q}_1(\mathbf{w}_0 - \mathbf{w}^\star) + \mathbf{w}^\star) = \mathbf{Q}_2(\mathbf{Q}_1(\mathbf{w}_0 - \mathbf{w}^\star)) + \mathbf{w}^\star$. Moreover, our translation preserves all distances, such that for instance

$$\forall \mathbf{w}\colon \ \|\mathbf{w} - \mathbf{P}_m(\mathbf{w})\| = d(\mathbf{w}, \mathcal{W}_m) = d(\mathbf{w} - \mathbf{w}^\star, \mathcal{C}_m) = \|\mathbf{w} - \mathbf{w}^\star - \mathbf{Q}_m(\mathbf{w} - \mathbf{w}^\star)\| .$$

We will also need the following lemma to prove the universal result for the cyclic $T = 2$ case.

Lemma D.8. *Let $\mathcal{C}_1, \mathcal{C}_2$ be two closed convex subsets of $\mathbb{R}^D$ with a nonempty intersection, and let $(\mathbf{w}_t)$ be a sequence of iterates induced by cyclic projections onto $\mathcal{C}_1, \mathcal{C}_2$ starting from an arbitrary $\mathbf{w}_0 \in \mathbb{R}^D$. Then, the projection "residuals" are monotonically decreasing. That is, For any iteration $t \in \mathbb{N}^+$ we have*

$$\|\mathbf{w}_{t+1} - \mathbf{w}_t\| \le \|\mathbf{w}_t - \mathbf{w}_{t-1}\| .$$

Proof. Assume w.l.o.g. that $2 \mid t$. By the definition of projections,

$$\|\mathbf{w}_{t+1} - \mathbf{w}_t\| = \|\mathbf{P}_1(\mathbf{w}_t) - \mathbf{w}_t\| = \min_{\mathbf{w} \in \mathcal{W}_1} \|\mathbf{w} - \underbrace{\mathbf{w}_t}_{\in \mathcal{W}_2}\| \le \|\underbrace{\mathbf{w}_{t-1}}_{\in \mathcal{W}_1} - \mathbf{w}_t\| .$$

□

Now, we recall and prove our main result for this section (using the above lemmas). Afterward, we will prove Lemma D.6. Our result here generalizes important parts of Theorems 10 and 11 in Evron et al. (2022) and extends these results from projections onto closed subspaces to projections onto general convex sets.

Recall Proposition 4.13. Let $\mathcal{W}_1, \dots, \mathcal{W}_T$ be closed convex sets with nonempty intersection $\mathcal{W}^\star$. Let $\mathbf{w}_k = (\mathbf{P}_T \circ \cdots \circ \mathbf{P}_1)^n(\mathbf{w}_0)$ be the iterate after $k = nT$ iterations (n cycles) of cyclic projections onto these convex sets. Then, the maximal distance to any (specific) convex set, is upper bounded *universally* as,

$$\text{For } T = 2: \quad \max_{m \in [T]} d^2(\mathbf{w}_k, \mathcal{W}_m) \leq \frac{1}{k+1} d^2(\mathbf{w}_0, \mathcal{W}^\star) \ ,$$

$$\text{For } T \geq 3: \quad \max_{m \in [T]} d^2(\mathbf{w}_k, \mathcal{W}_m) \leq \frac{2T^2}{\sqrt{k}} d^2(\mathbf{w}_0, \mathcal{W}^\star) \ .$$

Proof. We divide our proof into two parts.

1. **Cyclic two tasks** ($T = 2$) The proposition deals with cases where $k = nT = 2n$ (thus $2 \mid k$). We get

$$\max\left\{d\left(\mathbf{w}_k, \mathcal{W}_1\right), d\big(\underbrace{\mathbf{w}_k}_{\in \mathcal{W}_2}, \mathcal{W}_2\big)\right\} = d\left(\mathbf{w}_k, \mathcal{W}_1\right) = \min_{\mathbf{w} \in \mathcal{W}_1} \|\mathbf{w} - \mathbf{w}_k\| \triangleq \|\mathbf{w}_{k+1} - \mathbf{w}_k\| \ .$$

We thus focus on $\|\mathbf{w}_{k+1} - \mathbf{w}_k\|$ and bound it using all of the above, using ideas from Evron et al. (2022):

$$\begin{aligned}
\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2 &\overset{\text{Lemma D.8}}{\leq} \frac{1}{k+1} \sum_{t=0}^{k} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2 = \frac{1}{k+1} \sum_{t=0}^{k} \|\mathbf{w}_t - \mathbf{w}_{t+1} - \mathbf{w}^\star + \mathbf{w}^\star\|^2 \\
&= \frac{1}{k+1} \sum_{t=0}^{k} \|(\mathbf{I} - \mathbf{P}_{t+1})(\mathbf{w}_t) - (\mathbf{I} - \mathbf{P}_{t+1})(\mathbf{w}^\star)\|^2 \\
[\text{Prop. 1}] \quad &\leq \frac{1}{k+1} \sum_{t=0}^{k} \left(\|\mathbf{w}_t - \mathbf{w}^\star\|^2 - \|\mathbf{P}_{t+1}\mathbf{w}_t - \mathbf{P}_{t+1}\mathbf{w}^\star\|^2\right) \\
&= \frac{1}{k+1} \sum_{t=0}^{k} \left(\|\mathbf{w}_t - \mathbf{w}^\star\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}^\star\|^2\right) \\
[\text{telescoping}] \quad &= \frac{1}{k+1} \left(\|\mathbf{w}_0 - \mathbf{w}^\star\|^2 - \|\mathbf{w}_{k+1} - \mathbf{w}^\star\|^2\right) \leq \frac{1}{k+1} \|\mathbf{w}_0 - \mathbf{w}^\star\|^2 \triangleq \frac{1}{k+1} d^2(\mathbf{w}_0, \mathcal{W}^\star) \ .
\end{aligned}$$

2. **Cyclic $T \geq 3$ tasks**

Using Lemma D.6 and the translations from Remark D.7, it follows that

$$\begin{aligned}
d^2(\mathbf{w}_k, \mathcal{W}_m) \overset{\text{D.7}}{=} d^2(\mathbf{w}_k - \mathbf{w}^\star, \mathcal{C}_m) &= \|(\mathbf{w}_k - \mathbf{w}^\star) - \mathbf{Q}_m(\mathbf{w}_k - \mathbf{w}^\star)\|^2 \\
[\text{Remark D.7}] &= \|\mathbf{M}^n(\mathbf{w}_0 - \mathbf{w}^\star) - \mathbf{Q}_m(\mathbf{M}^n(\mathbf{w}_0 - \mathbf{w}^\star))\|^2 \\
[\textbf{(Lemma D.6d)}] &\leq 2m \left\|\mathbf{M}^n(\mathbf{w}_0 - \mathbf{w}^\star) - \mathbf{M}^{n+1}(\mathbf{w}_0 - \mathbf{w}^\star)\right\| \|\mathbf{w}_0 - \mathbf{w}^\star\| \\
[\textbf{(Lemma D.6e)}] &\leq 2m \sqrt{\frac{T}{n+1}} \|\mathbf{w}_0 - \mathbf{w}^\star\|^2 \\
&\leq 2T \sqrt{\frac{T}{n}} \|\mathbf{w}_0 - \mathbf{w}^\star\|^2 = \frac{2T^2}{\sqrt{k}} \|\mathbf{w}_0 - \mathbf{w}^\star\|^2 \triangleq \frac{2T^2}{\sqrt{k}} d^2(\mathbf{w}_0, \mathcal{W}^\star) \ .
\end{aligned}$$

□

Proof for Lemma D.6

Proof for (Lemma D.6a). For $m = 1$, using the fact that $\mathbf{0}_D \in \mathcal{C}_1 \cap \cdots \cap \mathcal{C}_T$, Prop. 1 immediately gives

$$\|\mathbf{v} - \mathbf{Q}_m(\mathbf{v})\|^2 = \|\mathbf{v} - \mathbf{Q}_1(\mathbf{v})\|^2 \leq \|\mathbf{v}\|^2 - \|\mathbf{Q}_1(\mathbf{v})\|^2 \overset{\text{contraction}}{\leq} 1 \cdot \left(\|\mathbf{v}\|^2 - \|\mathbf{Q}_T \circ \cdots \circ \mathbf{Q}_1(\mathbf{v})\|^2\right) .$$

Now we will prove the case when $m = 2, 3, \ldots, T-1$.

1. First, let $\mathbf{I}(\mathbf{v}) = \mathbf{v}$ be the identity operator and define the ("part-cyclic") operator $\mathbf{M}_\ell \triangleq \mathbf{Q}_\ell \circ \cdots \circ \mathbf{Q}_1$, $\forall \ell \in [T]$. We show recursively that (for any $m \geq 1$)

$$\begin{aligned}
\mathbf{I} &= (\mathbf{I} - \mathbf{Q}_1) + (\mathbf{Q}_1 - \mathbf{M}_2) + (\mathbf{M}_2 - \mathbf{M}_3) + \cdots + (\mathbf{M}_{m-1} - \mathbf{M}_m) + \mathbf{M}_m \\
&= (\mathbf{I} - \mathbf{Q}_1) + (\mathbf{Q}_1 - \mathbf{Q}_2 \circ \mathbf{M}_1) + (\mathbf{M}_2 - \mathbf{Q}_3 \circ \mathbf{M}_2) + \cdots + (\mathbf{M}_{m-1} - \mathbf{Q}_m \circ \mathbf{M}_{m-1}) + \mathbf{M}_m \\
&= (\mathbf{I} - \mathbf{Q}_1) + (\mathbf{I} - \mathbf{Q}_2) \circ \mathbf{M}_1 + (\mathbf{I} - \mathbf{Q}_3) \circ \mathbf{M}_2 + \cdots + (\mathbf{I} - \mathbf{Q}_m) \circ \mathbf{M}_{m-1} + \mathbf{M}_m \\
&= (\mathbf{I} - \mathbf{Q}_1) + \mathbf{M}_m + \sum_{\ell=1}^{m-1} (\mathbf{I} - \mathbf{Q}_{\ell+1}) \circ \mathbf{M}_\ell .
\end{aligned}$$

   Equivalently, using similar steps we have $\mathbf{Q}_m = (\mathbf{Q}_m - \mathbf{Q}_m \circ \mathbf{Q}_1) + \mathbf{Q}_m \circ \mathbf{M}_m + \sum_{\ell=1}^{m-1} (\mathbf{Q}_m - \mathbf{Q}_m \circ \mathbf{Q}_{\ell+1}) \circ \mathbf{M}_\ell$.

2. Subtracting both of the equations above, we get

$$\begin{aligned}
(\mathbf{I} - \mathbf{Q}_m) &= (\mathbf{I} - \mathbf{Q}_1) - (\mathbf{Q}_m - \mathbf{Q}_m \circ \mathbf{Q}_1) + \mathbf{M}_m - \mathbf{Q}_m \circ \mathbf{M}_m + \sum_{\ell=1}^{m-1} (\mathbf{I} - \mathbf{Q}_{\ell+1}) \circ \mathbf{M}_\ell - (\mathbf{Q}_m - \mathbf{Q}_m \circ \mathbf{Q}_{\ell+1}) \circ \mathbf{M}_\ell \\
&= (\mathbf{I} - \mathbf{Q}_m) - (\mathbf{I} - \mathbf{Q}_m) \circ \mathbf{Q}_1 + (\mathbf{I} - \mathbf{Q}_m) \circ \mathbf{M}_m + \sum_{\ell=1}^{m-1} (\mathbf{I} - \mathbf{Q}_m) \circ \mathbf{M}_\ell - (\mathbf{Q}_{\ell+1} - \mathbf{Q}_m \circ \mathbf{Q}_{\ell+1}) \circ \mathbf{M}_\ell \\
&= (\mathbf{I} - \mathbf{Q}_m) - (\mathbf{I} - \mathbf{Q}_m) \circ \mathbf{Q}_1 + \cancel{(\mathbf{I} - \mathbf{Q}_m) \circ \mathbf{Q}_m \circ \mathbf{M}_{m-1}} + \sum_{\ell=1}^{m-1} (\mathbf{I} - \mathbf{Q}_m) \circ \mathbf{M}_\ell - (\mathbf{I} - \mathbf{Q}_m) \circ \mathbf{Q}_{\ell+1} \circ \mathbf{M}_\ell ,
\end{aligned}$$

   where the third term is canceled since $(\mathbf{I} - \mathbf{Q}_m) \circ \mathbf{Q}_m = \mathbf{Q}_m - \mathbf{Q}_m^2 = \mathbf{0}$ due to the idempotence of projection operators.

3. Finally, we use the above and Claim D.2 to show that

$$\begin{aligned}
&\|\mathbf{v} - \mathbf{Q}_m(\mathbf{v})\|^2 = \|(\mathbf{I} - \mathbf{Q}_m)(\mathbf{v})\|^2 \\
&= \left\| ((\mathbf{I} - \mathbf{Q}_m) - (\mathbf{I} - \mathbf{Q}_m) \circ \mathbf{Q}_1)(\mathbf{v}) + \sum\nolimits_{\ell=1}^{m-1} \left((\mathbf{I} - \mathbf{Q}_m) \circ \mathbf{M}_\ell - (\mathbf{I} - \mathbf{Q}_m) \circ \mathbf{Q}_{\ell+1} \circ \mathbf{M}_\ell\right)(\mathbf{v}) \right\|^2 \\
&\leq m \left(\|(\mathbf{I} - \mathbf{Q}_m)(\mathbf{v}) - (\mathbf{I} - \mathbf{Q}_m)(\mathbf{Q}_1(\mathbf{v}))\|^2 + \sum_{\ell=1}^{m-1} \|(\mathbf{I} - \mathbf{Q}_m)(\mathbf{M}_\ell(\mathbf{v})) - (\mathbf{I} - \mathbf{Q}_m)((\mathbf{Q}_{\ell+1} \circ \mathbf{M}_\ell)(\mathbf{v}))\|^2 \right) .
\end{aligned}$$

   And since according to Prop. 1 we have $\forall \mathbf{v}, \mathbf{u} \in \mathbb{R}^D \colon \|(\mathbf{I} - \mathbf{Q}_m)(\mathbf{v}) - (\mathbf{I} - \mathbf{Q}_m)(\mathbf{u})\|^2 \leq \|\mathbf{v} - \mathbf{u}\|^2$, we get

$$\begin{aligned}
\|\mathbf{v} - \mathbf{Q}_m(\mathbf{v})\|^2 &\leq m \left(\|\mathbf{v} - \mathbf{Q}_1(\mathbf{v})\|^2 + \sum_{\ell=1}^{m-1} \|\mathbf{M}_\ell(\mathbf{v}) - \mathbf{Q}_{\ell+1}(\mathbf{M}_\ell(\mathbf{v}))\|^2 \right) \\
[\text{Prop. 1}] &\leq m \left(\|\mathbf{v}\|^2 - \|\underbrace{\mathbf{Q}_1(\mathbf{v})}_{=\mathbf{M}_1(\mathbf{v})}\|^2 + \sum_{\ell=1}^{m-1} \left(\|\mathbf{M}_\ell(\mathbf{v})\|^2 - \|\underbrace{\mathbf{Q}_{\ell+1}(\mathbf{M}_\ell(\mathbf{v}))}_{=\mathbf{M}_{\ell+1}(\mathbf{v})}\|^2 \right) \right) \\
[\text{telescoping}] &= m \left(\|\mathbf{v}\|^2 - \|\mathbf{M}_m(\mathbf{v})\|^2 \right) = m \left(\|\mathbf{v}\|^2 - \|(\mathbf{Q}_m \circ \cdots \circ \mathbf{Q}_1)(\mathbf{v})\|^2 \right) \\
[\text{contraction}] &\leq m \left(\|\mathbf{v}\|^2 - \|(\mathbf{Q}_T \circ \cdots \circ \mathbf{Q}_1)(\mathbf{v})\|^2 \right) .
\end{aligned}$$

□

Proof for ***(Lemma D.6b)***. Above we defined the identity operator $\mathbf{I}(\mathbf{v}) = \mathbf{v}$ and the operator $\mathbf{M}_\ell \triangleq \mathbf{Q}_\ell \circ \cdots \circ \mathbf{Q}_1$ and showed recursively that $\mathbf{I} = (\mathbf{I} - \mathbf{Q}_1) + \mathbf{M}_m + \sum_{\ell=1}^{m-1} (\mathbf{I} - \mathbf{Q}_{\ell+1}) \circ \mathbf{M}_\ell$. Clearly, this also yields:

$$\mathbf{I} - \mathbf{M} = \mathbf{I} - \mathbf{M}_T = (\mathbf{I} - \mathbf{Q}_1) + \sum_{\ell=1}^{T-1} (\mathbf{I} - \mathbf{Q}_{\ell+1}) \circ \mathbf{M}_\ell \ .$$

Then, we prove our lemma:

$$\begin{aligned}
\|\mathbf{v} - \mathbf{Q}_T \circ \cdots \circ \mathbf{Q}_1(\mathbf{v})\|^2 = \|\mathbf{v} - \mathbf{M}(\mathbf{v})\|^2 &= \left\| \left((\mathbf{I} - \mathbf{Q}_1) + \sum\nolimits_{\ell=1}^{T-1} (\mathbf{I} - \mathbf{Q}_{\ell+1}) \circ \mathbf{M}_\ell \right) (\mathbf{v}) \right\|^2 \\
&= \left\| (\mathbf{I} - \mathbf{Q}_1)(\mathbf{v}) + \sum\nolimits_{\ell=1}^{T-1} (\mathbf{I} - \mathbf{Q}_{\ell+1}) (\mathbf{M}_\ell(\mathbf{v})) \right\|^2 \\
[\text{Claim D.2}] &\leq T \left(\|(\mathbf{I} - \mathbf{Q}_1)(\mathbf{v})\|^2 + \sum\nolimits_{\ell=1}^{T-1} \|(\mathbf{I} - \mathbf{Q}_{\ell+1}) (\mathbf{M}_\ell(\mathbf{v}))\|^2 \right) \\
[\text{Prop. 1}] &\leq T \left(\|\mathbf{v}\|^2 - \|\underbrace{\mathbf{Q}_1(\mathbf{v})}_{=\mathbf{M}_1(\mathbf{v})}\|^2 + \sum\nolimits_{\ell=1}^{T-1} \left(\|\mathbf{M}_\ell(\mathbf{v})\|^2 - \|\underbrace{\mathbf{Q}_{\ell+1}(\mathbf{M}_\ell(\mathbf{v}))}_{=\mathbf{M}_{\ell+1}(\mathbf{v})}\|^2 \right) \right) \\
[\text{telescoping}] &= T \left(\|\mathbf{v}\|^2 - \|\mathbf{M}_T(\mathbf{v})\|^2 \right) = T \left(\|\mathbf{v}\|^2 - \|(\mathbf{Q}_T \circ \cdots \circ \mathbf{Q}_1)(\mathbf{v})\|^2 \right) \ .
\end{aligned}$$

□

Proof for ***(Lemma D.6c)***. We focus on the cyclic operator $\mathbf{M} = \mathbf{Q}_T \circ \cdots \circ \mathbf{Q}_1$ and remind that since each $\mathbf{Q}_\ell$ is a non-expansive operator, their composition $\mathbf{M}$ is also a non-expansive operator. Hence, $\|\mathbf{M}^n(\mathbf{u})\|_2^2 - \left\|\mathbf{M}^{n+1}(\mathbf{u})\right\|_2^2 \geq 0$.

Then, we show that

$$\begin{aligned}
\|\mathbf{M}^n(\mathbf{v})\|_2^2 - \left\|\mathbf{M}^{n+1}(\mathbf{v})\right\|_2^2 &= \underbrace{(\mathbf{M}^n(\mathbf{v}))^\top \mathbf{M}^n(\mathbf{v}) - \left(\mathbf{M}^{n+1}(\mathbf{v})\right)^\top \mathbf{M}^{n+1}(\mathbf{v})}_{\geq 0} \\
[\forall z \geq 0\colon z = \|z\|] &= \|(\mathbf{M}^n(\mathbf{v}))^\top \mathbf{M}^n(\mathbf{v}) - \left(\mathbf{M}^{n+1}(\mathbf{v})\right)^\top \mathbf{M}^{n+1}(\mathbf{v}) + \underbrace{(\mathbf{M}^n(\mathbf{v}))^\top \mathbf{M}^{n+1}(\mathbf{v}) - (\mathbf{M}^n(\mathbf{v}))^\top \mathbf{M}^{n+1}(\mathbf{v})}_{=0}\| \\
&= \left\| (\mathbf{M}^n(\mathbf{v}))^\top \mathbf{M}^n(\mathbf{v}) - (\mathbf{M}^n(\mathbf{v}))^\top \mathbf{M}^{n+1}(\mathbf{v}) + (\mathbf{M}^n(\mathbf{v}))^\top \mathbf{M}^{n+1}(\mathbf{v}) - \left(\mathbf{M}^{n+1}(\mathbf{v})\right)^\top \mathbf{M}^{n+1}(\mathbf{v}) \right\| \\
&= \left\| (\mathbf{M}^n(\mathbf{v}))^\top \left(\mathbf{M}^n(\mathbf{v}) - \mathbf{M}^{n+1}(\mathbf{v})\right) + \left(\mathbf{M}^n(\mathbf{v}) - \mathbf{M}^{n+1}(\mathbf{v})\right)^\top \mathbf{M}^{n+1}(\mathbf{v}) \right\| \\
[\text{triangle inequality}] &\leq \left\| (\mathbf{M}^n(\mathbf{v}))^\top \left(\mathbf{M}^n(\mathbf{v}) - \mathbf{M}^{n+1}(\mathbf{v})\right) \right\| + \left\| \left(\mathbf{M}^n(\mathbf{v}) - \mathbf{M}^{n+1}(\mathbf{v})\right)^\top \mathbf{M}^{n+1}(\mathbf{v}) \right\| \\
[\text{Cauchy-Schwarz}] &\leq \underbrace{\|\mathbf{M}^n(\mathbf{v})\|}_{\leq \|\mathbf{v}\|} \left\|\mathbf{M}^n(\mathbf{v}) - \mathbf{M}^{n+1}(\mathbf{v})\right\| + \left\|\mathbf{M}^n(\mathbf{v}) - \mathbf{M}^{n+1}(\mathbf{v})\right\| \underbrace{\left\|\mathbf{M}^{n+1}(\mathbf{v})\right\|}_{\leq \|\mathbf{v}\|} \\
[\text{contraction}] &\leq 2 \left\|\mathbf{M}^n(\mathbf{v}) - \mathbf{M}^{n+1}(\mathbf{v})\right\| \|\mathbf{v}\| \ .
\end{aligned}$$

□

Proof for ***(Lemma D.6d)***. This stems directly from **(Lemma D.6a)** and **(Lemma D.6c)**:

$$\|\mathbf{M}^n(\mathbf{v}) - \mathbf{Q}_m(\mathbf{M}^n(\mathbf{v}))\|^2 \leq m \left(\|\mathbf{M}^n(\mathbf{v})\|^2 - \left\|\mathbf{M}^{n+1}(\mathbf{v})\right\|^2 \right) \leq 2m \left\|\mathbf{M}^n(\mathbf{v}) - \mathbf{M}^{n+1}(\mathbf{v})\right\| \|\mathbf{v}\| \ .$$

□

Proof for ***(Lemma D.6e)***. First, we use **(Lemma D.6b)** to show that

$$\sum_{t=0}^{n-1} \left\| \left(\mathbf{M}^t - \mathbf{M}^{t+1}\right)(\mathbf{v}) \right\|^2 = \sum_{t=0}^{n-1} \left\| \mathbf{M}^t(\mathbf{v}) - \mathbf{M}\left(\mathbf{M}^t(\mathbf{v})\right) \right\|^2$$
$$\leq T \sum_{t=0}^{n-1} \left(\left\| \mathbf{M}^t(\mathbf{v}) \right\|^2 - \left\| \mathbf{M}^{t+1}(\mathbf{v}) \right\|^2 \right) = T \left(\|\mathbf{v}\|^2 - \|\mathbf{M}^n \mathbf{v}\|^2 \right) \leq T \|\mathbf{v}\|^2 \ .$$

Then, we use the non-expansiveness property (Prop. 1) again to show that

$$\left\| \left(\mathbf{M}^t - \mathbf{M}^{t+1}\right)(\mathbf{v}) \right\|^2 = \left\| \mathbf{M}\left(\mathbf{M}^{t-1}(\mathbf{v})\right) - \mathbf{M}\left(\mathbf{M}^t(\mathbf{v})\right) \right\|^2 \leq \left\| \mathbf{M}^{t-1}(\mathbf{v}) - \mathbf{M}^t(\mathbf{v}) \right\|^2 = \left\| \left(\mathbf{M}^{t-1} - \mathbf{M}^t\right)(\mathbf{v}) \right\|^2$$

and conclude that the series $\left(\left\| \left(\mathbf{M}^t - \mathbf{M}^{t+1}\right) \mathbf{v} \right\|^2 \right)_t$ is monotonically non-increasing.

All of the above means that $\forall \mathbf{v} \in \mathbb{R}^D$, it holds that

$$\left\| \left(\mathbf{M}^{n-1} - \mathbf{M}^n\right)(\mathbf{v}) \right\|^2 = \min_{t=0,\ldots,n-1} \left\| \left(\mathbf{M}^t - \mathbf{M}^{t+1}\right)(\mathbf{v}) \right\|^2 \leq \frac{1}{n} \sum_{t=0}^{n-1} \left\| \left(\mathbf{M}^t - \mathbf{M}^{t+1}\right)(\mathbf{v}) \right\|^2 \leq \frac{T}{n} \|\mathbf{v}\|^2 \ .$$

□

D.4. Proofs for Random Orderings (Section 4.4.2)

Here, we will make use of the following result from the POCS literature, proven in Nedić (2010).

Proposition 8 from Nedić (2010). Let $\mathcal{W}_1, \ldots, \mathcal{W}_T$ be closed convex sets with a non-empty intersection $\mathcal{W}^\star$. Starting from an arbitrary deterministic $\mathbf{w}_0$, let $\mathbf{w}_1, \ldots, \mathbf{w}_k$ be k iterates obtained by iteratively projecting onto the T sets according to an ordering τ sampled from an arbitrary i.i.d. distribution $p(\tau(\ell) = m) = p(\tau(\ell') = m),\ \forall \ell, \ell' \in [k]\,,\ \forall m \in [T]$, with all non-zero probabilities, *i.e.,* $p_{\min} \triangleq \min_{m\in[T]} p(\tau(\cdot) = m) > 0$. Assume that $\mathcal{W}^\star$ has a non-empty interior, *i.e.,* $\mathcal{B}(\overline{\mathbf{w}}, \delta) \subseteq \mathcal{W}^\star$ for some $\overline{\mathbf{w}}$. Then we have

$$\mathop{\mathbb{E}}_{\tau}\left[d^2(\mathbf{w}_k, \mathcal{W}^\star)\right] \leq \left(1 - p_{\min}\tfrac{\delta^2}{\|\mathbf{w}_0 - \overline{\mathbf{w}}\|^2}\right)^k d^2\left(\mathbf{w}_0, \mathcal{W}^\star\right)\,.$$

Recall Theorem 4.15. For T jointly-separable tasks learned in a random ordering, our quantities of interest (Def. 4.1) converge linearly as

$$\mathop{\mathbb{E}}_{\tau}\Big[\underbrace{\max_{m\in[T]} F_m(\mathbf{w}_k)}_{\text{Maximum forgetting}}\Big] \leq \mathop{\mathbb{E}}_{\tau}\Big[\underbrace{\max_{m\in[T]} d^2(\mathbf{w}_k, \mathcal{W}_m)}_{\substack{\text{Maximum distance}\\ \text{to any feasible set}}}\Big]R^2 \leq \mathop{\mathbb{E}}_{\tau}\Big[\underbrace{d^2(\mathbf{w}_k, \mathcal{W}^\star)}_{\substack{\text{Dist. to}\\ \text{offline feasible set}}}\Big]R^2 \leq \exp\Big(-\tfrac{k}{4T\|\mathbf{w}^\star\|^2R^2}\Big)\,\|\mathbf{w}^\star\|^2R^2\,,$$

where $\|\mathbf{w}^\star\|^2R^2 \geq 1$ is the problem complexity (Rem. 4.4).

Proof. We start by applying the above Proposition 8 from Nedić (2010) with the non-empty interior from Lemma D.4 and our specific initialization and "uniform" distribution, that is,

$$\overline{\mathbf{w}} = 2\mathbf{w}^\star,\ \delta = \tfrac{1}{R},\ p_{\min} = \tfrac{1}{T},\ \mathbf{w}_0 = \mathbf{0}_D,\ d(\mathbf{w}_0, \mathcal{W}^\star) = \|\mathbf{w}^\star\|\,.$$

We get

$$\mathop{\mathbb{E}}_{\tau}\left[d^2(\mathbf{w}_k, \mathcal{W}^\star)\right] \leq \left(1 - \frac{1}{4T\,\|\mathbf{w}^\star\|^2\,R^2}\right)^k \underbrace{d^2\left(\mathbf{w}_0, \mathcal{W}^\star\right)}_{=\|\mathbf{w}^\star\|^2} \leq \|\mathbf{w}^\star\|^2 \exp\left(-\frac{k}{4T\|\mathbf{w}^\star\|^2R^2}\right)\,,$$

where we used the algebraic identity $\forall z \in (0,1)\,, k \geq 1:\ \ (1-z)^k \leq \exp\left(-kz\right)$.

We complete our proof by noticing that according to Lemma 4.3 and Theorem 4.7, we have (for any "instantiation" of τ and the sequence $(\mathbf{w}_t)$ it induces),

$$0 \leq \max_{m\in[T]} F_m\left(\mathbf{w}_k\right) \leq \max_{m\in[T]} d^2(\mathbf{w}_k, \mathcal{W}_m)\,R^2 \leq d^2(\mathbf{w}_k, \mathcal{W}^\star)\,R^2 < \infty\,, \quad \forall k \in \mathbb{N}^+\,,$$

and so this order must hold for the expectations as well, and the theorem follows immediately.

□

Recall Lemma 4.16. Under a random i.i.d. ordering (and the separability assumption 2.1), the iterates converge *almost surely* to $\mathcal{W}^\star$, such that

$$\lim_{k\to\infty} d(\mathbf{w}_k, \mathcal{W}^\star) = 0, \quad \|\mathbf{w}^\star\| \leq \|\mathbf{w}_\infty\| \leq 2\,\|\mathbf{w}^\star\|\,.$$

Proof. We start by proving the almost sure convergence. From Lemma 4.5, we know that $\left(d\left(\mathbf{w}_k, \mathcal{W}^\star\right)\right)_k$ is (non-negative and) pointwise monotonically decreasing. Then, by the monotone convergence theorem, it has a pointwise limit $d\left(\mathbf{w}_\infty, \mathcal{W}^\star\right)$ such that $\mathbb{E}\left[d\left(\mathbf{w}_\infty, \mathcal{W}^\star\right)\right] = \lim_{k\to\infty}\mathbb{E}\left[d\left(\mathbf{w}_k, \mathcal{W}^\star\right)\right]$. According to the rates from Theorem 4.15, we have that $\lim_{k\to\infty}\mathbb{E}\left[d\left(\mathbf{w}_k, \mathcal{W}^\star\right)\right] = 0$. In turn, this means that $\mathbb{E}\left[d\left(\mathbf{w}_\infty, \mathcal{W}^\star\right)\right] = 0$. Overall, we get that the limit $d\left(\mathbf{w}_\infty, \mathcal{W}^\star\right)$ is pointwise non-negative with a zero mean, so it must be equal to zero with probability 1.

Finally, from the optimality guarantees of Theorem 4.7, we get that the limit $\mathbf{w}_\infty$, which is almost surely in $\mathcal{W}^\star$, must be 2-optimal. □

D.5. Proofs for the Average Iterate (Section 4.4.3)

Recall Proposition 4.18. After n cycles under a cyclic ordering ($k = nT$) we have

$$\underbrace{\max_{m\in[T]} F_m(\overline{\mathbf{w}}_k)}_{\text{Maximum forgetting}} \le \underbrace{\max_{m\in[T]} d^2(\overline{\mathbf{w}}_k, \mathcal{W}_m)}_{\text{Max. dist. to any feasible set}} R^2 \le \frac{T^2}{k}\|\mathbf{w}^\star\|^2 R^2$$

and after k iterations under a random ordering we have

$$\mathop{\mathbb{E}}_{\tau}\Big[\underbrace{\frac{1}{T}\sum_{m=1}^{T} F_m(\overline{\mathbf{w}}_k)}_{\text{Average forgetting}}\Big] \le \mathop{\mathbb{E}}_{\tau}\Big[\underbrace{\frac{1}{T}\sum_{m=1}^{T} d^2(\overline{\mathbf{w}}_k, \mathcal{W}_m)}_{\text{Avg. distance to feasible sets}}\Big] R^2 \le \frac{1}{k}\|\mathbf{w}^\star\|^2 R^2$$

(implying a $\frac{T}{k}\|\mathbf{w}^\star\|^2 R^2$ bound on the expected *maximum* forgetting and *maximum* distance to any feasible set).

Proof. We split our proof into two (one part for each ordering type).

D.5.1. Cyclic ordering

For task m, we exploit the averaging over all the iterates by combining it with the convexity of $\|\mathbf{v} - \mathbf{Q}_m(\mathbf{v})\|^2$, proved in Lemma D.3. For simplicity, we assume that $T \mid k$. Recall Remark D.7 on the translations. We have that,

$$\begin{aligned}
d^2\left(\overline{\mathbf{w}}_k, \mathcal{W}_m\right) &\overset{\text{D.7}}{=} d^2(\overline{\mathbf{w}}_k - \mathbf{w}^\star, \mathcal{C}_m) = \|(\mathbf{I} - \mathbf{Q}_m)\left(\overline{\mathbf{w}}_k - \mathbf{w}^\star\right)\| = \left\|(\mathbf{I} - \mathbf{Q}_m)\left(\frac{1}{k}\sum\nolimits_{t=1}^{k}\left(\mathbf{w}_t - \mathbf{w}^\star\right)\right)\right\| \\
\left[\substack{\text{Lemma D.3}\\ \text{Jensen's inequality}}\right] &\le \frac{1}{k}\sum_{t=1}^{k}\|(\mathbf{I} - \mathbf{Q}_m)\left(\mathbf{w}_t - \mathbf{w}^\star\right)\| = \frac{1}{k}\sum_{t=1}^{k}\left\|(\mathbf{I} - \mathbf{Q}_m)\left(\mathbf{Q}_{t \bmod T}\circ\cdots\circ\mathbf{Q}_1\right)\left(\mathbf{Q}_T\circ\cdots\circ\mathbf{Q}_1\right)^{\lfloor t/T\rfloor}\left(\mathbf{w}_0 - \mathbf{w}^\star\right)\right\|^2 \\
\left[\text{assuming } T|k\right] &= \frac{1}{k}\sum_{n=1}^{\lfloor k/T\rfloor}\sum_{t=1}^{T}\|(\mathbf{I} - \mathbf{Q}_m)\left(\mathbf{Q}_t\circ\cdots\circ\mathbf{Q}_1\right)\left(\mathbf{Q}_T\circ\cdots\circ\mathbf{Q}_1\right)^{n}\left(\mathbf{w}_0 - \mathbf{w}^\star\right)\|^2 \\
&= \frac{1}{k}\sum_{t=1}^{T}\sum_{n=1}^{\lfloor k/T\rfloor}\|(\mathbf{I} - \mathbf{Q}_m)\left(\mathbf{Q}_t\circ\cdots\circ\mathbf{Q}_1\right)\left(\mathbf{Q}_T\circ\cdots\circ\mathbf{Q}_1\right)^{n}\left(\mathbf{w}_0 - \mathbf{w}^\star\right)\|^2 \\
&= \frac{1}{k}\sum_{t=1}^{T}\sum_{n=1}^{\lfloor k/T\rfloor}\left\|(\mathbf{I} - \mathbf{Q}_m)\left(\mathbf{Q}_t\circ\cdots\circ\mathbf{Q}_1\circ\mathbf{Q}_T\circ\cdots\circ\mathbf{Q}_{t+1}\right)^{n}\Big(\underbrace{\mathbf{Q}_t\circ\cdots\circ\mathbf{Q}_1\left(\mathbf{w}_0 - \mathbf{w}^\star\right)}_{\triangleq\mathbf{u}_t}\Big)\right\|^2 \\
&= \frac{1}{k}\sum_{t=1}^{T}\sum_{n=1}^{\lfloor k/T\rfloor}\|(\mathbf{I} - \mathbf{Q}_m)\left(\mathbf{Q}_t\circ\cdots\circ\mathbf{Q}_1\circ\mathbf{Q}_T\circ\cdots\circ\mathbf{Q}_{t+1}\right)^{n}\left(\mathbf{u}_t\right)\|^2 \\
\left[\textbf{(Lemma D.6a)}\right] &\le \frac{T}{k}\sum_{t=1}^{T}\sum_{n=1}^{\lfloor k/T\rfloor}\|\left(\mathbf{Q}_t\circ\cdots\circ\mathbf{Q}_1\circ\mathbf{Q}_T\circ\cdots\circ\mathbf{Q}_{t+1}\right)^{n}\left(\mathbf{u}_t\right)\|^2 - \left\|\left(\mathbf{Q}_t\circ\cdots\circ\mathbf{Q}_1\circ\mathbf{Q}_T\circ\cdots\circ\mathbf{Q}_{t+1}\right)^{n+1}\left(\mathbf{u}_t\right)\right\|^2 \\
\left[\text{telescoping}\right] &= \frac{T}{k}\sum_{t=1}^{T}\|\left(\mathbf{Q}_t\circ\cdots\circ\mathbf{Q}_1\circ\mathbf{Q}_T\circ\cdots\circ\mathbf{Q}_{t+1}\right)\left(\mathbf{u}_t\right)\|^2 - \left\|\left(\mathbf{Q}_t\cdots\mathbf{Q}_1\circ\mathbf{Q}_T\circ\cdots\circ\mathbf{Q}_{t+1}\right)^{\lfloor k/T\rfloor+1}\left(\mathbf{u}_t\right)\right\|^2 \\
&\le \frac{T}{k}\sum_{t=1}^{T}\|\left(\mathbf{Q}_t\circ\cdots\circ\mathbf{Q}_1\circ\mathbf{Q}_T\circ\cdots\circ\mathbf{Q}_{t+1}\right)\left(\mathbf{u}_t\right)\|^2 = \frac{T}{k}\sum_{t=1}^{T}\|\left(\mathbf{Q}_t\circ\cdots\circ\mathbf{Q}_1\circ\mathbf{Q}_T\circ\cdots\circ\mathbf{Q}_1\right)\left(\mathbf{w}_0 - \mathbf{w}^\star\right)\|^2 \\
\left[\text{contraction}\right] &\le \frac{T}{k}\sum_{t=1}^{T}\|\mathbf{w}_0 - \mathbf{w}^\star\|^2 = \frac{T^2}{k}\|\mathbf{w}^\star\|^2 .
\end{aligned}$$

Overall, we bounded the distance in the cyclic setting: $d^2\left(\overline{\mathbf{w}}_k, \mathcal{W}_m\right) \le \frac{T}{n}\|\mathbf{w}^\star\|^2 = \frac{T^2}{k}\|\mathbf{w}^\star\|^2, \ \forall m\in[T]$.

Then, given Lemma 4.3, we conclude that $F_m(\overline{\mathbf{w}}_k) \le d^2(\overline{\mathbf{w}}_k, \mathcal{W}_m)R^2 \le \frac{T^2}{k}\|\mathbf{w}^\star\|^2 R^2, \ \forall m\in[T]$.

D.5.2. Random ordering

Our random ordering result is related to Proposition 6 and Equations (12) and (13) in Nedić (2010) which analyzed the expected average and maximum distance to any feasible set. Even more strongly related is Remark 15 in Evron et al. (2022) which analyzed the expected *average* forgetting in continual linear regression. Importantly, their proof in Appendix E.2 does not exploit the linearity of the projection operators and works for general projections as well. For completeness only, we closely follow their proof and present our own proof below.

$$
\begin{aligned}
\frac{1}{R^2}\mathop{\mathbb{E}}_{\tau}\Big[\underbrace{\frac{1}{T}\sum_{m=1}^{T}F_m(\overline{\mathbf{w}}_k)}_{\text{Average forgetting}}\Big] &\overset{\text{Lemma 4.3}}{\leq} \mathop{\mathbb{E}}_{\tau}\Big[\frac{1}{T}\sum_{m=1}^{T}d^2(\overline{\mathbf{w}}_k,\mathcal{W}_m)\Big] = \mathop{\mathbb{E}}_{\tau}\Big[\frac{1}{T}\sum_{m=1}^{T}\|(\mathbf{I}-\mathbf{P}_m)\,\overline{\mathbf{w}}_k\|^2\Big] \\
[\text{i.i.d. uniform}] &= \mathbb{E}_{\mathbf{P}_1,\ldots,\mathbf{P}_k,\mathbf{P}}\Big\|(\mathbf{I}-\mathbf{P})\left(\overline{\mathbf{w}}_k\right)\Big\|_2^2 \\
&= \mathbb{E}_{\mathbf{P}_1,\ldots,\mathbf{P}_k,\mathbf{P}}\Big\|(\mathbf{I}-\mathbf{P})\left(\frac{1}{k}\sum_{t=1}^{k}\mathbf{w}_t\right)\Big\|_2^2 \\
[\text{Lemma D.3, Jensen's inequality}] &\leq \frac{1}{k}\sum_{t=1}^{k}\mathbb{E}_{\mathbf{P}_1,\ldots,\mathbf{P}_t,\mathbf{P}}\Big\|(\mathbf{I}-\mathbf{P})\left(\mathbf{w}_t\right)\Big\|_2^2 \\
&= \frac{1}{k}\sum_{t=1}^{k}\mathbb{E}_{\mathbf{P}_1,\ldots,\mathbf{P}_t,\mathbf{P}}\Big[\Big\|(\mathbf{I}-\mathbf{P})\left(\mathbf{w}_t\right) - \underbrace{(\mathbf{I}-\mathbf{P})\left(\mathbf{w}^\star\right)}_{=\mathbf{0}}\Big\|_2^2\Big] \\
[\text{Prop. 1}] &\leq \frac{1}{k}\sum_{t=1}^{k}\mathbb{E}_{\mathbf{P}_1,\ldots,\mathbf{P}_t,\mathbf{P}}\Big[\Big\|\mathbf{w}_t-\mathbf{w}^\star\Big\|_2^2 - \Big\|\mathbf{P}(\mathbf{w}_t) - \underbrace{\mathbf{P}(\mathbf{w}^\star)}_{=\mathbf{w}^\star}\Big\|_2^2\Big] \\
[\text{i.i.d.}] &= \frac{1}{k}\sum_{t=1}^{k}\mathbb{E}_{\mathbf{P}_1,\ldots,\mathbf{P}_{t+1}}\Big[\Big\|\mathbf{w}_t-\mathbf{w}^\star\Big\|_2^2 - \Big\|\underbrace{\mathbf{P}_{t+1}(\mathbf{w}_t)}_{=\mathbf{w}_{t+1}} - \mathbf{w}^\star\Big\|_2^2\Big] \\
[\text{telescoping}] &= \frac{1}{k}\mathbb{E}_{\mathbf{P}_1,\ldots,\mathbf{P}_k}\Big[\underbrace{\Big\|\mathbf{w}_1-\mathbf{w}^\star\Big\|_2^2}_{\substack{\leq\|\mathbf{w}_0-\mathbf{w}^\star\|_2^2,\\ \text{by Corollary D.1}}} - \underbrace{\Big\|\mathbf{w}_{k+1}-\mathbf{w}^\star\Big\|_2^2}_{\geq 0}\Big] \leq \frac{1}{k}\|\mathbf{w}^\star\|^2 \ .
\end{aligned}
$$

□

E. Proof for the Extended Settings (Section 5)

Our next proof follows and generalizes our proof of Theorem 3.1 in Appendix B. We have split the statements and proofs of these two cases for clarity and to make sure that our fundamental Theorem 3.1 stands on its own.

For ease of readability, we mark λ_t and $\mathbf{B}_t$ in red and blue, respectively.

We provide here a unified proof for Theorems 5.1 and 5.4. To this end, we also define a unified scheme.

Scheme 5 Scheduled and Weighted Sequential Max-Margin

Initialization: $\mathbf{w}_0 = \mathbf{0}_D$
Iterative update for each task $t \in [k]$**:**

$$\mathbf{w}_t = \operatorname{argmin}_{\mathbf{w}} \|\mathbf{w} - c_t \mathbf{w}_{t-1}\|^2_{\mathbf{B}_t}, \qquad c_t \triangleq \lim_{\lambda\to 0} \frac{\ln \lambda_{t-1}}{\ln \lambda_t} \tag{10}$$
$$\text{s.t.} \quad y\mathbf{w}^\top \mathbf{x} \geq 1, \ \forall (\mathbf{x}, y) \in S_t$$

As we explain in the main body of the paper, we assume that $\forall t \in [k]$, the singular values of $\mathbf{B}_t$ are bounded, *i.e.,* it holds that $0 < \mu_t \leq \sigma_{\min}(\mathbf{B}_t) \leq \sigma_{\max}(\mathbf{B}_t) \leq M_t < \infty$ for some finite $\mu_t, M_t \in \mathbb{R}_{>0}$ (independent of λ). Moreover, given $\lambda > 0$, we parameterize the regularization strengths as $\lambda_t \triangleq \lambda_t(\lambda) > 0$ for arbitrary functions $\lambda_t : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ holding that $\lim_{\lambda\to 0} \lambda_t(\lambda) = 0$ and $\lim\limits_{\lambda\to 0} \frac{\ln \lambda_{t-1}}{\ln \lambda_t} < \infty$ is well-defined $\forall t \in [k]$.

Proof for Theorems 5.1 and 5.4. We will prove by induction on $t \geq 0$ that the scale of the residual $\mathbf{r}_t^{(\lambda)} \triangleq \mathbf{w}_t^{(\lambda)} - \ln\left(\frac{1}{\lambda_t}\right)\mathbf{w}_t$ at each iteration is $\left\|\mathbf{r}_t^{(\lambda)}\right\| = \mathcal{O}\left(\sum_{t'=1}^{t}\left(\prod_{n=t'}^{t}\frac{M_n}{\mu_n}\right)\ln\ln\left(\frac{1}{\lambda_{t'}}\right)\right)$; and that consequently (since $\ln\left(\frac{1}{\lambda}\right)\mathbf{w}_t$ grows faster; see Remark E.1), the iterates are either identical (*i.e.,* $\mathbf{w}_t^{(\lambda)} = \mathbf{w}_t$) or converge in the same direction when $\lambda \to 0$, *i.e.,* $\lim_{\lambda\to 0} \frac{\mathbf{w}_t^{(\lambda)}}{\|\mathbf{w}_t^{(\lambda)}\|} = \frac{\mathbf{w}_t}{\|\mathbf{w}_t\|}$.

For $t = 0$**:** By the conditions of the theorem, it trivially holds that $\mathbf{w}_0^{(\lambda)} = \mathbf{w}_0 = \mathbf{0}_D$ and $\mathbf{r}_0^{(\lambda)} = \mathbf{0}_D$.

For $t \geq 1$**:** The solved optimization problem (recall Remark B.1) is:

$$\mathbf{w}_t^{(\lambda)} = \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^D} \mathcal{L}_\lambda(\mathbf{w}) \triangleq \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^D} \sum_{\mathbf{x}\in S_t} e^{-\mathbf{w}^\top\mathbf{x}} + \frac{\lambda_t}{2}\left\|\sqrt{\mathbf{B}_t}\mathbf{w} - \sqrt{\mathbf{B}_t}\mathbf{w}_{t-1}^{(\lambda)}\right\|^2.$$

We follow the same ideas as in our proof for Theorem 3.1 in Appendix B (with some adjustments since, for instance, our objective here is no longer λ-strongly convex, but rather $\lambda_t\mu_t$-strongly convex, since its Hessian matrix holds $\nabla^2\mathcal{L}_\lambda(\mathbf{w}) \succeq \lambda_t\mathbf{B}_t \succeq \lambda_t\mu_t\mathbf{I} \succ \mathbf{0}$).

First, we compute the gradient of $\mathcal{L}_\lambda$, normalized by λ_t:

$$\frac{1}{\lambda_t}\nabla\mathcal{L}_\lambda(\mathbf{w}) = -\frac{1}{\lambda_t}\sum_{\mathbf{x}\in S_t}\mathbf{x}\exp\left(-\mathbf{w}^\top\mathbf{x}\right) + \mathbf{B}_t\mathbf{w} - \mathbf{B}_t\mathbf{w}_{t-1}^{(\lambda)}.$$

Then, we plug in $\mathbf{w} = \left(\ln\left(\frac{1}{\lambda_t}\right) + s_t\ln\ln\left(\frac{1}{\lambda_t}\right)\right)\mathbf{w}_t + \tilde{\mathbf{w}}_t = \ln\left(\frac{1}{\lambda_t}\ln^{s_t}\left(\frac{1}{\lambda_t}\right)\right)\mathbf{w}_t + \tilde{\mathbf{w}}_t$, for some sign $s_t \in \{-1, +1\}$ and a vector $\tilde{\mathbf{w}}_t$ with a norm independent of λ (both will be defined below).

$$\begin{aligned}
&\frac{1}{\lambda_t}\nabla\mathcal{L}_\lambda\left(\left(\ln\left(\frac{1}{\lambda_t}\right) + s_t\ln\ln\left(\frac{1}{\lambda_t}\right)\right)\mathbf{w}_t + \tilde{\mathbf{w}}_t\right)\\
&= -\frac{1}{\lambda_t}\sum_{\mathbf{x}\in S_t}\mathbf{x}e^{\ln\left(\lambda_t\ln^{-s_t}\left(\frac{1}{\lambda_t}\right)\right)\mathbf{w}_t^\top\mathbf{x}}e^{-\tilde{\mathbf{w}}_t^\top\mathbf{x}} + \ln\left(\frac{1}{\lambda_t}\ln^{s_t}\left(\frac{1}{\lambda_t}\right)\right)\mathbf{B}_t\mathbf{w}_t + \mathbf{B}_t\tilde{\mathbf{w}}_t - \mathbf{B}_t\mathbf{w}_{t-1}^{(\lambda)}\\
&= -\frac{1}{\lambda_t}\sum_{\mathbf{x}\in S_t}\mathbf{x}\left(\lambda_t\ln^{-s_t}\left(\frac{1}{\lambda_t}\right)\right)^{\mathbf{w}_t^\top\mathbf{x}}e^{-\tilde{\mathbf{w}}_t^\top\mathbf{x}} + \ln\left(\frac{1}{\lambda_t}\ln^{s_t}\left(\frac{1}{\lambda_t}\right)\right)\mathbf{B}_t\mathbf{w}_t + \mathbf{B}_t\tilde{\mathbf{w}}_t - \mathbf{B}_t\mathbf{w}_{t-1}^{(\lambda)}.
\end{aligned}$$

Now, denoting the set of support vectors by $\hat{\mathcal{S}}_t \triangleq \left\{\mathbf{x} \in S_t \mid \mathbf{w}_t^\top \mathbf{x} = 1\right\}$ (which might be empty for $t \geq 2$), and using the inductive assumption that $\mathbf{w}_{t-1}^{(\lambda)} = \mathbf{r}_{t-1}^{(\lambda)} + \ln\left(\frac{1}{\lambda_{t-1}}\right)\mathbf{w}_{t-1}$ (where $\left\|\mathbf{r}_{t-1}^{(\lambda)}\right\| = \mathcal{O}\left(\sum_{t'=1}^{t-1}\left(\prod_{n=t'}^{t-1}\frac{M_n}{\mu_n}\right)\ln\ln\left(\frac{1}{\lambda_{t'}}\right)\right)$), we get

$$
\begin{aligned}
&\frac{1}{\lambda_t}\nabla\mathcal{L}_\lambda\left(\left(\ln\left(\tfrac{1}{\lambda_t}\right) + s_t \ln\ln\left(\tfrac{1}{\lambda_t}\right)\right)\mathbf{w}_t + \tilde{\mathbf{w}}_t\right) \\
&= -\frac{1}{\lambda_t}\left(\lambda_t \ln^{-s_t}\left(\tfrac{1}{\lambda_t}\right)\right)^1 \sum_{\mathbf{x}\in\hat{\mathcal{S}}_t} \mathbf{x} e^{-\tilde{\mathbf{w}}_t^\top \mathbf{x}} - \frac{1}{\lambda_t}\sum_{\mathbf{x}\notin\hat{\mathcal{S}}_t} \mathbf{x}\left(\lambda_t \ln^{-s_t}\left(\tfrac{1}{\lambda_t}\right)\right)^{\mathbf{w}_t^\top \mathbf{x}} e^{-\tilde{\mathbf{w}}_t^\top \mathbf{x}} + \\
&\qquad + \ln\left(\tfrac{1}{\lambda_t}\right)\mathbf{B}_t\mathbf{w}_t - \ln\left(\tfrac{1}{\lambda_{t-1}}\right)\mathbf{B}_t\mathbf{w}_{t-1} + \ln\ln^{s_t}\left(\tfrac{1}{\lambda_t}\right)\mathbf{B}_t\mathbf{w}_t + \mathbf{B}_t\tilde{\mathbf{w}}_t - \mathbf{B}_t\mathbf{r}_{t-1}^{(\lambda)} \\
&= -\ln^{-s_t}\left(\tfrac{1}{\lambda_t}\right)\sum_{\mathbf{x}\in\hat{\mathcal{S}}_t} \mathbf{x} e^{-\tilde{\mathbf{w}}_t^\top \mathbf{x}} - \sum_{\mathbf{x}\notin\hat{\mathcal{S}}_t} \mathbf{x}\lambda_t^{\mathbf{w}_t^\top \mathbf{x}-1}\left(\ln\left(\tfrac{1}{\lambda_t}\right)\right)^{-s_t\mathbf{w}_t^\top \mathbf{x}} e^{-\tilde{\mathbf{w}}_t^\top \mathbf{x}} + \\
&\qquad + \ln\left(\tfrac{1}{\lambda_t}\right)\mathbf{B}_t\left(\mathbf{w}_t - \tfrac{\ln\lambda_{t-1}}{\ln\lambda_t}\mathbf{w}_{t-1}\right) + \ln\ln^{s_t}\left(\tfrac{1}{\lambda_t}\right)\mathbf{B}_t\mathbf{w}_t + \mathbf{B}_t\tilde{\mathbf{w}}_t - \mathbf{B}_t\mathbf{r}_{t-1}^{(\lambda)}\,.
\end{aligned}
$$

By the triangle inequality and since $\left\|\ln\ln^{s_t}\left(\frac{1}{\lambda_t}\right)\right\| = \left\|s_t \ln\ln\left(\frac{1}{\lambda_t}\right)\right\| = \overbrace{|s_t|}^{=1}\ \overbrace{\ln\ln\left(1/\lambda_t\right)}^{>0,\text{ when }\lambda_t<1/e}$, we have,

$$
\begin{aligned}
&\left\|\frac{1}{\lambda_t}\nabla\mathcal{L}_\lambda\left(\ln\left(\tfrac{1}{\lambda_t}\right) + s_t\ln\ln\left(\tfrac{1}{\lambda_t}\right)\mathbf{w}_t + \tilde{\mathbf{w}}_t\right)\right\| \\
&\leq \underbrace{\left\|\ln\left(\tfrac{1}{\lambda_t}\right)\mathbf{B}_t\left(\mathbf{w}_t - \tfrac{\ln\lambda_{t-1}}{\ln\lambda_t}\mathbf{w}_{t-1}\right) - \ln^{-s_t}\left(\tfrac{1}{\lambda_t}\right)\sum_{\mathbf{x}\in\hat{\mathcal{S}}_t}\mathbf{x}e^{-\tilde{\mathbf{w}}_t^\top \mathbf{x}}\right\|}_{\triangleq \mathbf{a}_1(\lambda_t)} \\
&\qquad + \underbrace{\left\|\sum_{\mathbf{x}\notin\hat{\mathcal{S}}_t}\mathbf{x}\lambda_t^{\mathbf{w}_t^\top \mathbf{x}-1}\left(\ln\left(\tfrac{1}{\lambda_t}\right)\right)^{-s_t\mathbf{w}_t^\top \mathbf{x}} e^{-\tilde{\mathbf{w}}_t^\top \mathbf{x}}\right\|}_{\triangleq \mathbf{a}_2(\lambda_t)} + \ln\ln\left(\tfrac{1}{\lambda_t}\right)\left\|\mathbf{B}_t\mathbf{w}_t\right\| + \left\|\mathbf{B}_t\tilde{\mathbf{w}}_t\right\| + \left\|\mathbf{B}_t\mathbf{r}_{t-1}^{(\lambda)}\right\| \\
&\leq \mathbf{a}_1(\lambda_t) + \mathbf{a}_2(\lambda_t) + \ln\ln\left(\tfrac{1}{\lambda_t}\right)M_t\left\|\mathbf{w}_t\right\| + M_t\left\|\tilde{\mathbf{w}}_t\right\| + M_t\left\|\mathbf{r}_{t-1}^{(\lambda)}\right\|\,.
\end{aligned}
$$

Here also, we distinguish between two different behaviors of the SMM solution $\mathbf{w}_t$ (recall that $c_t \triangleq \lim_{\lambda\to 0} \frac{\ln \lambda_{t-1}}{\ln \lambda_t}$):

1. **When $\mathbf{w}_t \triangleq \mathbf{P}_t(c_t\mathbf{w}_{t-1}) \neq c_t\mathbf{w}_{t-1}$ (and necessarily $|\hat{\mathcal{S}}_t| \geq 1$):** We choose $s_t = -1$ and $\mathbf{a}_1(\lambda)$ becomes:

$$\mathbf{a}_1(\lambda_t) = \left\|\ln\left(\tfrac{1}{\lambda_t}\right)\mathbf{B}_t\left(\mathbf{w}_t - c_t\mathbf{w}_{t-1}\right) - \ln^{-s_t}\left(\tfrac{1}{\lambda_t}\right)\sum_{\mathbf{x}\in\hat{\mathcal{S}}_t}\mathbf{x}e^{-\tilde{\mathbf{w}}_t^\top\mathbf{x}}\right\| = \ln\left(\tfrac{1}{\lambda_t}\right)\left\|\mathbf{B}_t\left(\mathbf{w}_t - c_t\mathbf{w}_{t-1}\right) - \sum_{\mathbf{x}\in\hat{\mathcal{S}}_t}\mathbf{x}e^{-\tilde{\mathbf{w}}_t^\top\mathbf{x}}\right\|.$$

We thus wish to choose $\tilde{\mathbf{w}}_t$ so as to *zero* $\mathbf{a}_1(\lambda_t)$. That is, according to the KKT conditions of Eq. (10), we have

$$\sum\nolimits_{\mathbf{x}\in\hat{\mathcal{S}}_t}\mathbf{x}e^{-\tilde{\mathbf{w}}_t^\top\mathbf{x}} = \mathbf{B}_t\left(\mathbf{w}_t - c_t\mathbf{w}_{t-1}\right) \triangleq \sum\nolimits_{\mathbf{x}\in\hat{\mathcal{S}}_t}\mathbf{x}\alpha(\mathbf{x})\,,$$

where $\boldsymbol{\alpha} \in \mathbb{R}_{\geq 0}^{|S_t|}$ is the dual solution of Eq. (10). From Lemma B.3 and Corollary B.4 (applied with our $\mathbf{B}_t$ and c_t here), we know that there almost surely exists such a vector $\tilde{\mathbf{w}}_t$ whose norm is $\mathcal{O}(1)$ (independent of λ).

Furthermore, since $\lim_{\lambda\to 0}\left(\lambda^{c-1}\ln^c\left(1/\lambda\right)\right) = 0,\ \forall c > 1$, it holds that $\mathbf{a}_2(\lambda_t)$ becomes

$$\mathbf{a}_2(\lambda_t) = \left\|\sum_{\mathbf{x}\notin\hat{\mathcal{S}}_t}\underbrace{\mathbf{x}e^{-\tilde{\mathbf{w}}_t^\top\mathbf{x}}}_{=\mathcal{O}(1)}\underbrace{\lambda_t\overbrace{\mathbf{w}_t^\top\mathbf{x}-1}^{>0}\left(\ln\left(1/\lambda_t\right)\right)^{\mathbf{w}_t^\top\mathbf{x}}}_{\to 0}\right\| \xrightarrow{\lambda\to 0} 0\,.$$

In conclusion, we can choose $\tilde{\mathbf{w}}_t$ and s_t such that $\|\tilde{\mathbf{w}}_t\| = \mathcal{O}(1)$, $\mathbf{a}_1(\lambda_t) = 0$, and $\mathbf{a}_2(\lambda_t) \to 0$.

2. **When $\mathbf{w}_t \triangleq \mathbf{P}_t(c_t\mathbf{w}_{t-1}) = c_t\mathbf{w}_{t-1}$ (and possibly $\hat{\mathcal{S}}_t = \emptyset$):** We choose $\tilde{\mathbf{w}}_t = \mathbf{0}_D$ and $s_t = 1$. It follows that

$$\mathbf{a}_1(\lambda_t) = \left\|\ln^{-s_t}\left(\tfrac{1}{\lambda_t}\right)\sum_{\mathbf{x}\in\hat{\mathcal{S}}_t}\mathbf{x}\right\| = \underbrace{\ln^{-1}\left(\tfrac{1}{\lambda_t}\right)}_{\to 0}\underbrace{\left\|\sum\nolimits_{\mathbf{x}\in\hat{\mathcal{S}}_t}\mathbf{x}\right\|}_{=\mathcal{O}(1)} \xrightarrow{\lambda\to 0} 0\,,$$

$$\mathbf{a}_2(\lambda_t) = \left\|\sum_{\mathbf{x}\notin\hat{\mathcal{S}}_t}\underbrace{\mathbf{x}}_{=\mathcal{O}(1)}\underbrace{\lambda_t\overbrace{\mathbf{w}_t^\top\mathbf{x}-1}^{>0}}_{\to 0}\underbrace{\left(\ln\left(\tfrac{1}{\lambda_t}\right)\right)^{\overbrace{-\mathbf{w}_t^\top\mathbf{x}}^{<-1}}}_{\to 0}\right\| \xrightarrow{\lambda\to 0} 0\,.$$

Finally, we use the $\lambda_t\mu_t$-strong convexity of our objective and Lemma B.2 to bound the distance to the optimum by

$$\begin{aligned}
\left\|\mathbf{r}_t^{(\lambda)}\right\| &\triangleq \left\|\mathbf{w}_t^{(\lambda)} - \ln\left(\tfrac{1}{\lambda_t}\right)\mathbf{w}_t\right\| = \left\|\mathbf{w}_t^{(\lambda)} - \left(\left(\ln\left(\tfrac{1}{\lambda_t}\right) + s_t\ln\ln\left(\tfrac{1}{\lambda_t}\right)\right)\mathbf{w}_t + \tilde{\mathbf{w}}_t\right) + s_t\ln\ln\left(\tfrac{1}{\lambda_t}\right)\mathbf{w}_t + \tilde{\mathbf{w}}_t\right\| \\
[\text{triangle ineq.}] &\leq \left\|\mathbf{w}_t^{(\lambda)} - \left(\left(\ln\left(\tfrac{1}{\lambda_t}\right) + s_t\ln\ln\left(\tfrac{1}{\lambda_t}\right)\right)\mathbf{w}_t + \tilde{\mathbf{w}}_t\right)\right\| + \left\|s_t\ln\ln\left(\tfrac{1}{\lambda_t}\right)\mathbf{w}_t\right\| + \left\|\tilde{\mathbf{w}}_t\right\| \\
[\text{Lemma B.2}] &\leq \frac{1}{\lambda_t\mu_t}\left\|\nabla\mathcal{L}_\lambda\left(\left(\ln\left(\tfrac{1}{\lambda_t}\right) + s_t\ln\ln\left(\tfrac{1}{\lambda_t}\right)\right)\mathbf{w}_t + \tilde{\mathbf{w}}_t\right)\right\| + \ln\ln\left(\tfrac{1}{\lambda_t}\right)\left\|\mathbf{w}_t\right\| + \left\|\tilde{\mathbf{w}}_t\right\| \\
[\text{the above}] &\leq \frac{1}{\mu_t}\left(\mathbf{a}_1(\lambda_t) + \mathbf{a}_2(\lambda_t) + \ln\ln\left(\tfrac{1}{\lambda_t}\right)M_t\left\|\mathbf{w}_t\right\| + M_t\left\|\tilde{\mathbf{w}}_t\right\| + M_t\left\|\mathbf{r}_{t-1}^{(\lambda)}\right\|\right) + \ln\ln\left(\tfrac{1}{\lambda_t}\right)\left\|\mathbf{w}_t\right\| + \left\|\tilde{\mathbf{w}}_t\right\| \\
&= \underbrace{\frac{1}{\mu_t}\mathbf{a}_1(\lambda_t) + \frac{1}{\mu_t}\mathbf{a}_2(\lambda_t)}_{\to 0,\ \text{since } \mu_t \text{ is finite}} + \ln\ln\left(\tfrac{1}{\lambda_t}\right)\left(1 + \frac{M_t}{\mu_t}\right)\underbrace{\left\|\mathbf{w}_t\right\|}_{=\mathcal{O}(1)} + \left(1 + \frac{M_t}{\mu_t}\right)\underbrace{\left\|\tilde{\mathbf{w}}_t\right\|}_{=\mathcal{O}(1)} + \frac{M_t}{\mu_t}\left\|\mathbf{r}_{t-1}^{(\lambda)}\right\|\,,
\end{aligned}$$

and since we have by the induction assumption that $\left\|\mathbf{r}_{t-1}^{(\lambda)}\right\| = \mathcal{O}\left(\sum_{t'=1}^{t-1}\left(\prod_{n=t'}^{t-1}\frac{M_n}{\mu_n}\right)\ln\ln\left(\frac{1}{\lambda_{t'}}\right)\right)$ and also $\frac{M_n}{\mu_n} \geq 1$, we can conclude that $\left\|\mathbf{r}_t^{(\lambda)}\right\| = \mathcal{O}\left(\sum_{t'=1}^{t}\left(\prod_{n=t'}^{t}\frac{M_n}{\mu_n}\right)\ln\ln\left(\frac{1}{\lambda_{t'}}\right)\right)$. □

Remark E.1 (Applicability of our analysis for finite λ). When the singular values of the weight matrices of all tasks are between $0 < \mu \leq \mu_t \leq M_t \leq M < \infty$, the residual from our analysis becomes of the order $\left\|\mathbf{r}_t^{(\lambda)}\right\| \triangleq \left\|\mathbf{w}_t^{(\lambda)} - \ln(\frac{1}{\lambda_t})\mathbf{w}_t\right\| = \mathcal{O}\left(\sum_{t'=1}^{t} \left(M/\mu\right)^{t-t'+1} \ln\ln\left(\frac{1}{\lambda_{t'}}\right)\right)$. When this bound on the condition number, *i.e.,* M/μ, is strictly larger than 1, we get an exponential growth of the residuals. Therefore, within a few tasks under a *finite* regularization strength λ, the bound on the residuals might become even larger than the scale of the scaled SMM solutions *i.e.,* larger than $\ln(\frac{1}{\lambda_t})$. In turn, this will invalidate our analysis. Of course, in the limit of $\lambda \to 0$, our analysis still applies. As we explained in Remark 4.6, we take $\lambda \to 0$ *after* fixing the number of iterations k (or t).

The question of determining the specific value of λ that practically ensures that $\frac{\mathbf{w}_t^{(\lambda)}}{\|\mathbf{w}_t^{(\lambda)}\|} \approx \frac{\mathbf{w}_t}{\|\mathbf{w}_t\|}$, as well as understanding the true impact of $M/\mu > 1$ on the residuals, remains an intriguing and open research question.

E.1. Proofs for Regularization Strength Scheduling (Section 5.1)

Recall Proposition 5.2. There exists a construction of two jointly-separable tasks in which the iterates of the cyclic ordering do not converge to $\mathcal{W}^\star$, for any $p > 1$. Specifically, for any $p > 1$, $\|\mathbf{w}^\star\| > 1$, it holds that

$$\lim_{k\to\infty} \frac{d(\mathbf{w}_k, \mathcal{W}^\star)}{\|\mathbf{w}^\star\|} = \frac{\|\mathbf{w}^\star\|^2 (p-1)}{2 + \|\mathbf{w}^\star\|^2 (p-1)} \sqrt{1 - \frac{1}{\|\mathbf{w}^\star\|^2}} \,.$$

Proof. First, we explain our construction in the following figure.

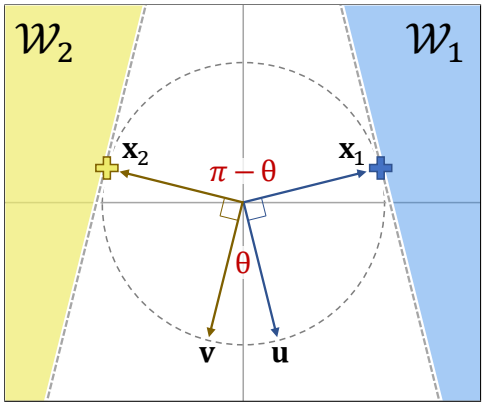


Figure 10. We consider a 2-dimensional setting of 2 tasks, one positively-labeled normalized sample per task, *i.e.,* $(\mathbf{x}_1, +1)$, $(\mathbf{x}_2, +1)$. Both $\mathbf{x}_1, \mathbf{x}_2$ are on the unit sphere (hence $R = 1$) and are symmetric w.r.t. the vertical axis.
In this setting, when $p > 1$, it can be readily seen that at each iteration we perform an orthogonal projection onto a closed affine subspace. Therefore, if we converge to $\mathcal{W}^\star$, we must converge to the minimum-norm solution $\mathbf{w}^\star$ specifically (see Halperin (1962) for instance). We thus wish to study $\lim_{k\to\infty} \|\mathbf{w}_k - \mathbf{w}^\star\|$.
To ease our notations, we again translate our space by $-\mathbf{w}^\star$ such that instead of using the "affine" halfspaces $\mathcal{W}_1, \mathcal{W}_2$, we use their "homogeneous" counterparts $\mathcal{C}_m \triangleq \mathcal{W}_m - \mathbf{w}^\star$ (see Remark D.7). Under this "change of coordinates", instead of analyzing $\|\mathbf{w}_k - \mathbf{w}^\star\|$, we can simply analyze $\|\mathbf{w}_k\|$.
Since the setting is 2-dimensional and each task has a single sample, we can parameterize the orthogonal projections as $\mathbf{P}_1 = \mathbf{u}\mathbf{u}^\top = \mathbf{I} - \mathbf{x}_1\mathbf{x}_1^\top$ and $\mathbf{P}_2 = \mathbf{v}\mathbf{v}^\top = \mathbf{I} - \mathbf{x}_2\mathbf{x}_2^\top$ (where $\|\mathbf{u}\| = \|\mathbf{v}\| = 1$). We denote the angle between the two subspaces that we project onto as $\theta \triangleq \arccos\left(\mathbf{u}^\top\mathbf{v}\right) = \pi - \arccos\left(\mathbf{x}_1^\top\mathbf{x}_2\right)$.
Notice that (before the translation) the min-norm solution is $\mathbf{w}^\star = \frac{1}{\sin(\theta/2)}\begin{bmatrix}0\\1\end{bmatrix}$.

In the Scheduled Sequential Max-Margin Scheme 3, each iterate is multiplied by $\beta \triangleq \frac{1}{p}$ before being projected onto the next subspace, equivalently to

$$\mathbf{w}_k = \mathbf{w}_{2n} = \mathbf{P}_2\Big((1-\beta)\mathbf{w}_0 + \beta\mathbf{P}_1\big((1-\beta)\mathbf{w}_0 + \beta\mathbf{w}_{2(n-1)}\big)\Big)$$

To understand this recursive expression, let $n = 1$. We get (the projections are linear) $\mathbf{w}_2 = (1-\beta)\mathbf{P}_2\mathbf{w}_0 + \beta\mathbf{P}_2\mathbf{P}_1\mathbf{w}_0$.
Then, for $n = 2$, we get $\mathbf{w}_4 = (1-\beta)\mathbf{P}_2\mathbf{w}_0 + \beta(1-\beta)\mathbf{P}_2\mathbf{P}_1\mathbf{w}_0 + \beta^2(1-\beta)\mathbf{P}_2\mathbf{P}_1\mathbf{P}_2\mathbf{w}_0 + \beta^3\left(\mathbf{P}_2\mathbf{P}_1\right)^2\mathbf{w}_0$.

Recursively, we get $\mathbf{w}_k = \mathbf{w}_{2n} = (1-\beta)\left(\sum_{k=0}^{n-1}\beta^{2k}\left(\mathbf{P}_2\mathbf{P}_1\right)^k\mathbf{P}_2\mathbf{w}_0 + \sum_{k=1}^{n-1}\beta^{2k-1}\left(\mathbf{P}_2\mathbf{P}_1\right)^k\mathbf{w}_0\right) + \beta^{2n-1}\left(\mathbf{P}_2\mathbf{P}_1\right)^2\mathbf{w}_0$.

Plugging in the parameterization of the projections, we get

$$\mathbf{w}_k = \mathbf{w}_{2n} = (1-\beta)\left(\left[\sum_{k=0}^{n-1}(\beta\cos\theta)^{2k}\right]\mathbf{v}\mathbf{v}^\top\mathbf{w}_0 + \left[\sum_{k=1}^{n-1}(\beta\cos\theta)^{2k-1}\right]\mathbf{v}\mathbf{u}^\top\mathbf{w}_0\right) + (\beta\cos\theta)^{2n-1}\mathbf{v}\mathbf{u}^\top\mathbf{w}_0\,.$$

Since $\beta = \frac{1}{p} < 1$, we have

$$\lim_{n\to\infty}\mathbf{w}_{2n} = (1-\beta)\left(\frac{\mathbf{v}^\top\mathbf{w}_0}{1-(\beta\cos\theta)^2} + \frac{\beta\mathbf{u}^\top\mathbf{w}_0\cos\theta}{1-(\beta\cos\theta)^2}\right)\mathbf{v} = (1-\beta)\left(\frac{\mathbf{v}^\top\mathbf{w}_0 + \beta\mathbf{u}^\top\mathbf{w}_0\cos\theta}{1-(\beta\cos\theta)^2}\right)\mathbf{v}\,.$$

From the definition of $\mathbf{v}, \mathbf{u}$ it holds that $\mathbf{v}^\top\mathbf{w}_0 = \mathbf{u}^\top\mathbf{w}_0 = \cos\frac{\theta}{2}\|\mathbf{w}_0\|$ and also that $\|\mathbf{w}_0\| = \frac{1}{\sin(\theta/2)}$ (after the translation).

Putting it all together, we get $\lim_{k\to\infty}\mathbf{w}_k = \lim_{n\to\infty}\mathbf{w}_{2n} = (1-\beta)\cos\frac{\theta}{2}\|\mathbf{w}_0\|\left(\frac{1+\beta\cos\theta}{1-(\beta\cos\theta)^2}\right)\mathbf{v} = \|\mathbf{w}_0\|\frac{(1-\beta)\cos\frac{\theta}{2}}{1-\beta\cos\theta}\mathbf{v}$.

Recalling that $\|\mathbf{v}\| = 1$ and returning to the original coordinate system (by reversing the aforementioned translation), we get:

$$\lim_{k\to\infty}\frac{\|\mathbf{w}_k - \mathbf{w}^\star\|}{\|\mathbf{w}_0 - \mathbf{w}^\star\|} = \frac{(1-\beta)\cos\frac{\theta}{2}}{1-\beta\cos\theta} = \frac{1-\beta}{1-\beta\left(1-\frac{2}{\|\mathbf{w}^\star\|^2}\right)}\sqrt{1-\frac{1}{\|\mathbf{w}^\star\|^2}} = \frac{p-1}{p-1+\frac{2}{\|\mathbf{w}^\star\|^2}}\sqrt{1-\frac{1}{\|\mathbf{w}^\star\|^2}}\,.$$

□

E.2. Additional Material for Weighted Regularization (Section 5.2)

Recall Proposition 5.5. Using a Fisher-information-based weighting scheme of $\mathbf{B}_t = \sum_{i=1}^{t-1} \sum_{\mathbf{x} \in S_i} \mathbf{x}\mathbf{x}^\top$, there is no forgetting in continual linear regression.

Details and proof. Consider training a linear regression model with the square loss on a sequence of tasks, obtaining iterates $\mathbf{w}_1, ..., \mathbf{w}_k$. We follow Evron et al. (2022), assuming that the tasks are jointy-realizable by a linear predictor, *i.e.,* there exists $\tilde{\mathbf{w}}$ such that for all t, $\forall (\mathbf{x}, y) \in S_t : \mathbf{x}^\top \tilde{\mathbf{w}} = y$. The forgetting after training task t is defined (as in Evron et al. (2022)) to be the average training loss of $\mathbf{w}_t$ on tasks $1, ..., t$.

Now, we identify a specific regularization scheme that guarantees no forgetting.

For linear regression with square loss, define

$$\mathbf{w}_1 \in \operatorname*{argmin}_{\mathbf{w}} \sum_{(\mathbf{x},y) \in S_1} \left(\mathbf{x}^\top \mathbf{w} - y\right)^2 ,$$

$$\mathbf{w}_t \in \operatorname*{argmin}_{\mathbf{w}} \sum_{(\mathbf{x},y) \in S_t} \left(\mathbf{x}^\top \mathbf{w} - y\right)^2 + \frac{\lambda}{2} \|\mathbf{w} - \mathbf{w}_{t-1}\|^2_{\mathbf{B}_t}, \; t \geq 2.$$

where $\mathbf{B}_t = \sum_{i=1}^{t-1} \sum_{\mathbf{x} \in S_i} \mathbf{x}\mathbf{x}^\top$. Then for any $t \in [k]$ and any $\lambda > 0$, we show that there is no forgetting.

The proof follows by induction. For $t = 1$ the statement is trivially correct. Assume by induction that the statement is correct for some task t, meaning there is no forgetting for $\mathbf{w}_t$ on tasks $1, ..., t$. We want to show that the statement is correct for task $t + 1$, *i.e.,* for $\mathbf{w}_{t+1}$ there is no forgetting on tasks $1, ..., t + 1$.

Note that for task $t + 1$ the regularization term is

$$\begin{aligned}
\frac{\lambda}{2} \left(\mathbf{w} - \mathbf{w}_t\right)^\top \mathbf{B}_{t+1} \left(\mathbf{w} - \mathbf{w}_t\right) &= \frac{\lambda}{2} \left(\mathbf{w} - \mathbf{w}_t\right)^\top \left(\sum_{i=1}^{t} \sum_{\mathbf{x} \in S_i} \mathbf{x}\mathbf{x}^\top \right) \left(\mathbf{w} - \mathbf{w}_t\right) \\
&= \frac{\lambda}{2} \sum_{i=1}^{t} \sum_{\mathbf{x} \in S_i} \left(\mathbf{w} - \mathbf{w}_t\right)^\top \mathbf{x}\mathbf{x}^\top \left(\mathbf{w} - \mathbf{w}_t\right) \\
&= \frac{\lambda}{2} \sum_{i=1}^{t} \sum_{\mathbf{x} \in S_i} \left(\mathbf{x}^\top \mathbf{w} - \mathbf{x}^\top \mathbf{w}_t\right)^2 .
\end{aligned}$$

By the realizability assumption and the inductive assumption, we have that for all $i = 1, 2, ..., t$ it holds that $\forall \left(\mathbf{x}, y\right) \in S_i : \mathbf{x}^\top \mathbf{w}_t = y$, therefore we get

$$\frac{\lambda}{2} \left(\mathbf{w} - \mathbf{w}_t\right)^\top \mathbf{B}_{t+1} \left(\mathbf{w} - \mathbf{w}_t\right) = \frac{\lambda}{2} \sum_{i=1}^{t} \sum_{(\mathbf{x},y) \in S_i} \left(\mathbf{x}^\top \mathbf{w} - y\right)^2 .$$

It follows that

$$\mathbf{w}_{t+1} \in \operatorname*{argmin}_{\mathbf{w}} \left[\sum_{(\mathbf{x},y) \in S_{t+1}} \left(\mathbf{x}^\top \mathbf{w} - y\right)^2 + \frac{\lambda}{2} \sum_{i=1}^{t} \sum_{(\mathbf{x},y) \in S_i} \left(\mathbf{x}^\top \mathbf{w} - y\right)^2 \right] .$$

Finally, by the realizability assumption, $\mathbf{w}_{t+1}$ can achieve zero loss on all tasks $1, ..., t+1$, and thus there is no forgetting. □

E.2.1. Additional experiments for Fisher-Information-based weighting schemes

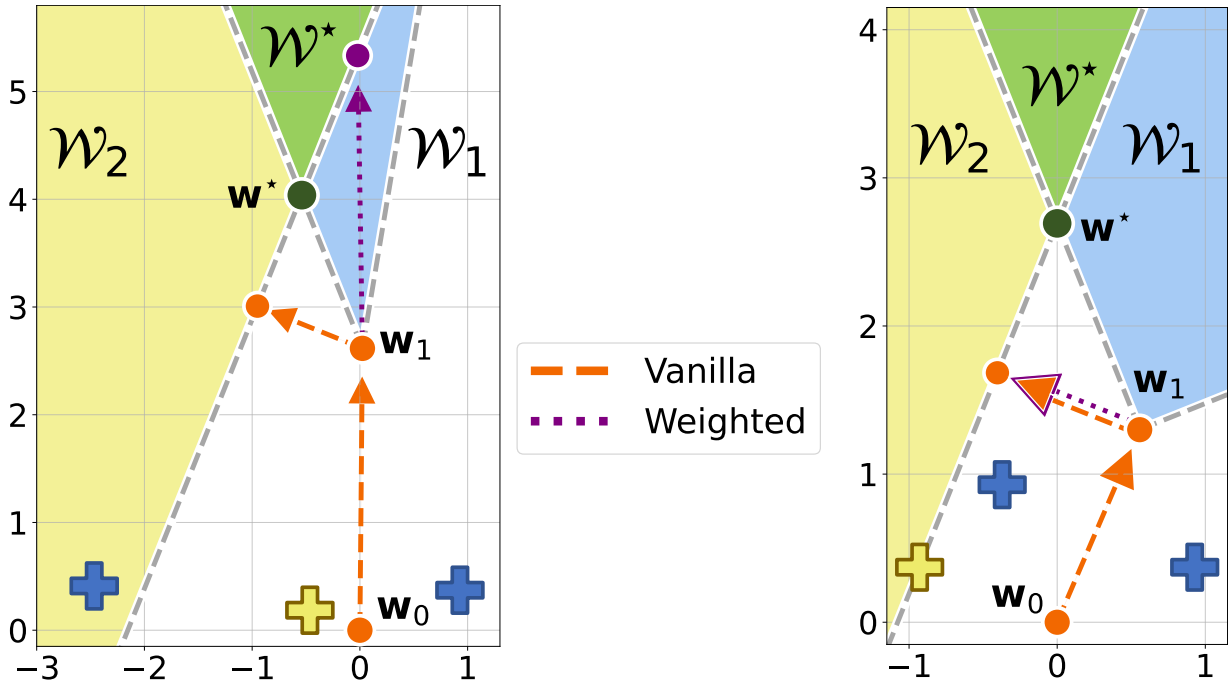
Here we demonstrate two different behaviors of Fisher-based weighting schemes.

To the right, we see a larger version of Figure 5. The two positively-labeled datapoints of the first task are $\mathbf{x}_1 = \left[-0.3714, 0.9285\right]^\top$ and $\mathbf{x}_2 = \left[0.9285, 0.3714\right]^\top$. Since these examples are orthogonal, the Fisher Information matrix is proportional to $\mathbf{x}_1\mathbf{x}_1^\top + \mathbf{x}_2\mathbf{x}_2^\top = \mathbf{I}$ (there is a multiplicative factor stemming from the normalized probability $\frac{\exp\left(y_n \tilde{\mathbf{w}}^\top \mathbf{x}_n\right)}{\exp(y_n \tilde{\mathbf{w}}^\top \mathbf{x}_n) + \exp(-y_n \tilde{\mathbf{w}}^\top \mathbf{x}_n)} = \frac{\exp(1)}{\exp(1)+\exp(-1)}$ of these two datapoints which are both support vectors). The datapoint of the second task is $\mathbf{x}_3 = \left[-0.9285, 0.3714\right]^\top$. Notice how the weighted regularization yields the same solution as the vanilla regularization, since $\mathbf{B}_2 \propto \mathbf{I}$.

To the left, we run a different experiment. The two positively-labeled datapoints of the first task are $\mathbf{x}_1 = \left[-2.4660, 0.4110\right]^\top$ and $\mathbf{x}_2 = \left[0.9285, 0.3714\right]^\top$. The datapoint of the second task is $\mathbf{x}_3 = \left[-0.4642, 0.1857\right]^\top$. In contrast to the previous example, here the Fisher Information matrix is proportional to $\begin{bmatrix} 6.9431 & -0.6687 \\ -0.6687 & 0.3068 \end{bmatrix}$, which notably assigns much less penalty to changes in the vertical axis compared to changes in the horizontal axes. As a result, the learner remains within the feasible set of the first task and projects onto the offline feasible set, effectively retaining the knowledge of the first task without any forgetting.

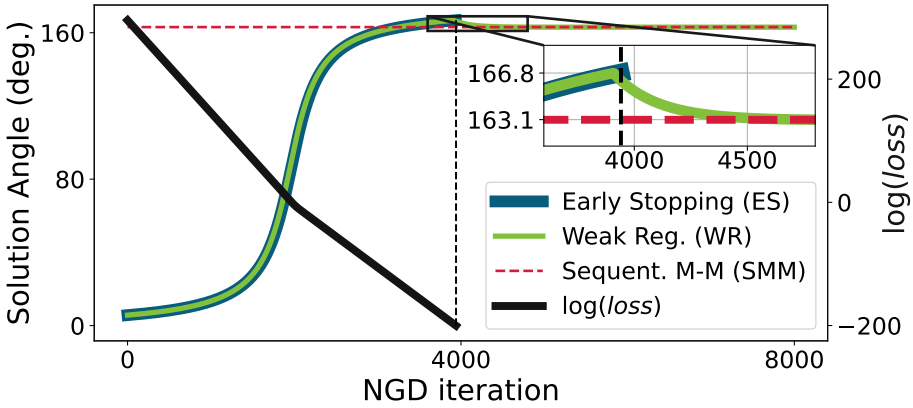
This experiment suggests that the key to avoiding forgetting in weighted regularization schemes may lie in employing ill-conditioned weighting matrices. This stands in contrast to linear *regression*, where the Fisher-information matrix always prevents forgetting (Prop. 5.5), even when the datapoints of the previous task are orthogonal and the weighting matrix is proportional to the identity matrix.

Figure 11. Solving two different task sequences while using their corresponding Fisher-Information matrices for weighting. We plot the iterates obtained by solving the weakly-regularized weighted Scheme 2 with Normalized Gradient Descent (NGD). We asserted that the obtained iterates agree with those of the weighted SMM Scheme 4.

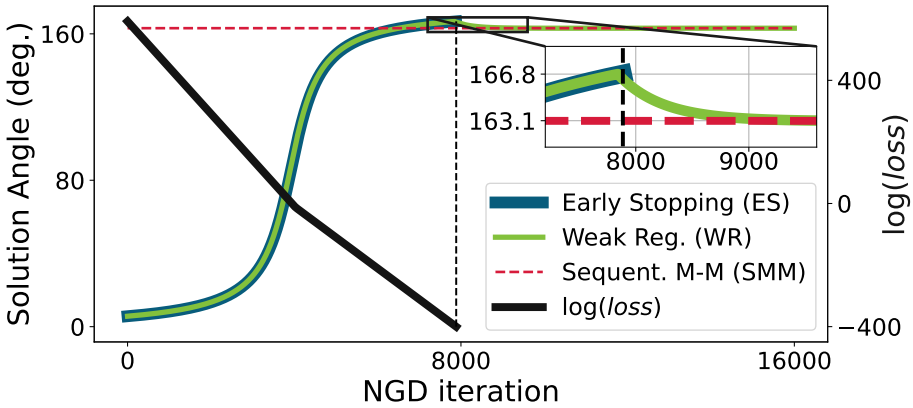


F. Additional Material for the Early Stopping Discussion (Section 7.1)

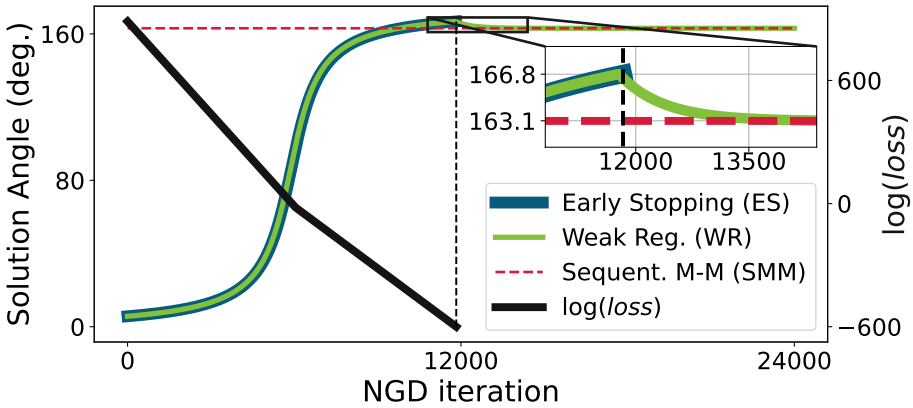
In Figure 6 we used $\lambda = \epsilon = \exp(-200)$. To make sure that this choice is small enough, we rerun the same setting as described in Figure 6 but with $\lambda = \epsilon = \exp(-400)$ and $\lambda = \epsilon = \exp(-600)$. The following figure demonstrates that the observed phenomenon remains unchanged — early stopping and weak regularization consistently lead to distinct solutions.



(a) Same as Figure 6 with $\lambda = \epsilon = \exp(-200)$



(b) $\lambda = \epsilon = \exp(-400)$



(c) $\lambda = \epsilon = \exp(-600)$

Figure 12. Repeating the experiment in Figure 6 with smaller λ and ϵ.