CONVERGENCE AND IMPLICIT BIAS OF GRADIENT DESCENT ON CONTINUAL LINEAR CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study continual learning on multiple linear classification tasks by sequentially running gradient descent (GD) for a fixed budget of iterations per task. When all tasks are jointly linearly separable and are presented in a cyclic/random order, we show the directional convergence of the trained linear classifier to the *joint (offline) max-margin* solution. This is surprising because GD training on a single task is implicitly biased towards the individual max-margin solution for the task, and the direction of the joint max-margin solution can be largely different from these individual solutions. Additionally, when tasks are given in a cyclic order, we present a non-asymptotic analysis on *cycle-averaged forgetting*, revealing that (1) alignment between tasks is indeed closely tied to catastrophic forgetting and backward knowledge transfer and (2) the amount of forgetting vanishes to zero as the cycle repeats. Lastly, we analyze the case where the tasks are no longer jointly separable and show that the model trained in a cyclic order converges to the unique minimum of the joint loss function.

1 INTRODUCTION

Continual learning (CL) aims to sequentially learn a model from a stream of tasks or datasets, to extend its knowledge continuously. The main challenge in CL is *catastrophic forgetting*, meaning that their performance on previous tasks degrades after learning new ones (McCloskey & Cohen, 1989; Goodfellow et al., 2013). It has led to a growing body of works focusing on heuristic methods of mitigating forgetting, including regularization-based methods (Kirkpatrick et al., 2017; Aljundi et al., 2018; Li & Hoiem, 2017), replay-based methods (Chaudhry et al., 2019; Lopez-Paz & Ranzato, 2017; Shin et al., 2017), and optimization-based methods (Farajtabar et al., 2020; Javed & White, 2019; Mirzadeh et al., 2020).

As CL is receiving significant attention in practice, it is also important to theoretically understand the mechanism of continual learning. A vast amount of the theoretical works on CL so far has focused on regression problems (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), whereas most of the practical application of deep learning is based on classification. Thus, theoretical analysis of continual classification methods and their learning dynamics is of significant interest and importance. Indeed, a few results study continual classification (Raghavan & Balaprakash, 2021; Kim et al., 2022; 2023; Shi & Wang, 2023), albeit focusing on theoretical perspectives that are different from ours; we review these related works in Appendix A.

This paper is mainly motivated by a recent result studying continual linear classification on a collection of jointly separable datasets (Evron et al., 2023). The authors consider continual training of a linear classifier under weak regularization, where the linear classifier is trained until convergence at every given task. By taking the limit of the regularization coefficient $\lambda \to 0$, this training procedure is shown to be equivalent (in terms of the parameter *direction* as $\lambda \to 0$) to a projection-based scheme called Sequential Max-Margin (SMM): every time we encounter a new binary classification task, we project the current model parameter vector to a convex set defined by the margin conditions of the given dataset. Then, under this framework of projection onto convex sets, the authors show linear convergence of the iterates of SMM to an *offline solution* (i.e., a classifier that solves all tasks at once) under cyclic/random ordering of the tasks. More details can be found in Appendix B.

In light of the insightful analyses by Evron et al. (2023), we now highlight some aspects of their work that motivate the setup of our interest. First of all, Evron et al. (2023) consider minimizing the regularized training loss of each task *until convergence*; however, it is far more common to spend a finite budget of iterations per task in practice (i.e., online one-pass setting, or fixed-epoch setting). Training until convergence, combined with sending the regularization coefficient $\lambda \to 0$, also raises an issue on the claimed equivalence of weakly regularized training and the projection-based scheme. As $\lambda \to 0$, the solution of the training objective diverges to infinity, which does not match the fact that the iterate of the SMM travels only for a finite distance at every stage.[1] Another noteworthy characteristic of the considered SMM scheme is that it does not always converge to the *offline max-margin solution*, i.e., the hard-margin support vector machine solution that solves *all* tasks jointly, which is known to be beneficial in terms of generalization (Vapnik, 2013). Lastly, in their concluding section, Evron et al. (2023) also suggest studying *unregularized* continual training with *early stopping* and highlight that the behavior may be different. These observations triggered our investigation into a gradient-based algorithm for continual linear classification and its convergence and algorithmic bias.

In this work, we theoretically study continual linear classification via sequentially running gradient descent (GD) on the *unregularized* logistic loss for a fixed budget of iterations at every stage.[2] When all tasks are jointly separable and revealed in cyclic order (as studied by Evron et al. (2023)), we show that sequential GD converges in the direction of the offline max-margin solution, unlike SMM. We highlight that this is an interesting result for at least two reasons:

- It reveals a clear difference between sequential GD and the projection-based SMM algorithm in terms of algorithmic bias.
- It is well-known that GD applied to an individual task has its implicit bias towards the task's own max-margin direction (Soudry et al., 2018). However, the direction of the offline max-margin solution can largely differ from the max-margin directions of individual tasks, not even lying on the subspace spanned by the individual directions (see Figure 1 and Appendix C.1).

Therefore, the convergence of sequential GD to the *offline max-margin solution* highlights that repeated continual training eventually drives the model to learn all tasks well, overcoming the biases towards individual tasks. In addition to the implicit bias result, we also characterize the convergence rate in terms of total loss and the vanishing rate of the per-cycle forgetting. Our analysis reveals a surprising but intuitive link between positive/negative task alignments and forgetting. Furthermore, we broaden the scope of our analysis to the random task ordering case and a jointly non-separable case. We summarize our main contributions below.

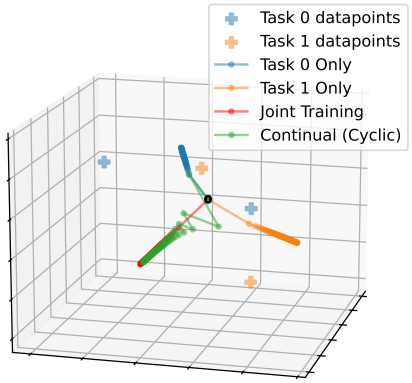


Figure 1: Trajectory of sequential GD on a two-task toy example (Appendix C.1) in which the offline max-margin direction is not on the subspace spanned by individual task max-margin solutions. Sequential GD iterates initially oscillate but quickly start to evolve along the same direction as the offline max-margin direction.

1.1 Summary of Contributions

We study continual linear classification using *sequential GD*, where the model is updated by K iterations of GD on the unregularized training loss of each given task.

- In Section 3, we study the scenario where the tasks are jointly separable and are given in a cyclic order. We prove that the joint (full) training loss asymptotically converges to zero (Theorem 3.1) and the sequential GD iterates in fact align with the *joint (offline) max-margin* solution (Theorem 3.2). We also provide non-asymptotic analysis of *cycle-averaged forgetting* and

[1] Recall that Evron et al. (2023) show their equivalence in terms of parameter *direction*.

[2] We focus on this setup instead of early stopping because it is closer to common practice in deep learning.

loss convergence and show that average forgetting per cycle J diminishes at the rate of $\mathcal{O}(\frac{\ln^4 J}{J^2})$ (Theorem 3.4), which is faster than the convergence rate of the loss $\mathcal{O}(\frac{\ln^2 J}{J})$ (Theorem 3.3). Our forgetting analysis is closely aligned with the common intuition on how task alignment/conflict impacts forgetting.

- Section 4 considers the same jointly separable setup, but the tasks given in a random order. In Theorems 4.1 and 4.2, we show that asymptotic loss convergence and directional convergence to the joint max-margin solution still happen, albeit almost surely.
- Lastly, in Section 5 we consider the case where the tasks are no longer jointly separable, which admits a unique global minimum of the joint training loss. We derive a fast non-asymptotic convergence rate of $\mathcal{O}(\frac{\ln^2 J}{J^2})$ towards the global minimum when the tasks are presented cyclically.

2 Problem Setup

In this section, we outline the problem setup considered throughout the paper.

2.1 Setup: Continual Linear Binary Classification

We consider binary classification, where each data point $\boldsymbol{x} \in \mathbb{R}^d$ has its own label $y \in \{-1, +1\}$. We assume that our learning algorithm encounters M different binary classification **tasks** in a sequential manner, and our goal is to find an **offline solution** that jointly solves all the tasks. The total dataset is denoted as $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i \in I}$, where $I := \{0, \dots, N-1\}$ is the set of indices of data. Since the dataset comprises all data pairs from M tasks, the index set I is partitioned into $I = \biguplus_{m=0}^{M-1} I_m$, where I_m is a set of indices for data points in task $m \in \{0, \dots, M-1\}$.

We consider a linear model $f(\boldsymbol{x}; \boldsymbol{w}) = \boldsymbol{x}^\top \boldsymbol{w}$, which is parameterized by a weight vector $\boldsymbol{w} \in \mathbb{R}^d$. With a loss function $\ell(u)$ that decreases to zero as $u \to \infty$, the **offline (joint) training loss** is defined as

$$\mathcal{L}(\boldsymbol{w}) := \sum_{i \in I} \ell\left(y_i f(\boldsymbol{x}_i; \boldsymbol{w})\right) = \sum_{i \in I} \ell(y_i \boldsymbol{x}_i^\top \boldsymbol{w}).$$

Likewise, loss of task $m \in \{0, \dots, M-1\}$ is defined as

$$\mathcal{L}_m(\boldsymbol{w}) := \sum_{i \in I_m} \ell(y_i \boldsymbol{x}_i^\top \boldsymbol{w}).$$

Notation. We denote the joint data matrix as $\boldsymbol{X} \in \mathbb{R}^{d \times N}$, whose columns are the d-dimensional data points $\boldsymbol{x}_i$'s. For a square matrix $\boldsymbol{A}$, we denote the maximum/minimum eigenvalue of it by $\lambda_{\max}(\boldsymbol{A})$ and $\lambda_{\min}(\boldsymbol{A})$, respectively. In particular, we write $\sigma_{\max} = \sqrt{\lambda_{\max}(\boldsymbol{X}\boldsymbol{X}^\top)}$ as the maximum singular value of $\boldsymbol{X}$. The ℓ_2 norm of a vector $\boldsymbol{v}$ is denoted as $\|\boldsymbol{v}\|$. Let $\mathbb{R}^N_{\geq 0}$ be the set of N dimensional vectors whose elements are greater or equal to zero. Also, for a couple of integers $K_1 \leq K_2$, we write $[K_1 : K_2]$ to denote a set of consecutive integers $\{K_1, K_1 + 1, \dots, K_2\}$.

2.2 Algorithm: Sequential Gradient Descent

In continual learning, we can only see data in the current stage. For each stage $t = 0, 1, \dots,$ the index set $I^{(t)}$ of data that will be used comes from one of $\{I_m\}_{m \in [0:M-1]}$. Note that the learning algorithm does *not* have the freedom to choose the next task; we assume that the task is presented to the algorithm by the "environment." During stage t, we minimize the corresponding training loss

$$\mathcal{L}^{(t)}(\boldsymbol{w}) := \sum_{i \in I^{(t)}} \ell(y_i \boldsymbol{x}_i^\top \boldsymbol{w}) \tag{1}$$

using gradient descent (GD) with a fixed learning rate η as follows:

$$\boldsymbol{w}_{k+1}^{(t)} = \boldsymbol{w}_k^{(t)} - \eta \nabla \mathcal{L}^{(t)}(\boldsymbol{w}_k^{(t)}) \quad \text{for } k \in [0 : K-1], \qquad \boldsymbol{w}_0^{(t+1)} = \boldsymbol{w}_K^{(t)}. \tag{2}$$

That is, for the task $\mathcal{L}^{(t)}$ given at stage t, we run K steps of GD updates and move on to the next task by setting the initial iterate of the next stage $\boldsymbol{w}_0^{(t+1)}$ as the last iterate of the current stage $\boldsymbol{w}_K^{(t)}$.

There are two common schemes for deciding the order of the tasks to be learned.

Cyclic task ordering. The tasks are presented in a predefined cyclic order. That is, $\mathcal{L}^{(t)} = \mathcal{L}_{t \bmod M}$.

Random task ordering. Every task is independently sampled uniformly at random. That is, for all $t \in \mathbb{N} \cup \{0\}$ and $m \in [0 : M-1]$, $\mathbb{P}(I^{(t)} = I_m) = 1/M$ holds.

Both ordering schemes have been studied theoretically and empirically (Evron et al., 2022; 2023; Cossu et al., 2022; Houyon et al., 2023). Indeed, such schemes can naturally occur in real-world scenarios. For instance, cyclic task ordering covers search engines influenced by periodic events[3] and seasonal financial data (Gultekin & Gultekin, 1983; Yang et al., 2022). Random task ordering bears a resemblance to autonomous driving in randomly recurring environments (Verwimp et al., 2023).

3 Cyclic Learning of Jointly Separable Tasks

In this section, we focus on the jointly linearly separable datasets (Evron et al., 2023). We dive deep into the case of cyclic task ordering and prove that sequential GD on separable linear classification tasks converges in direction to the offline max-margin solution of the total dataset. Additionally, through a non-asymptotic analysis on the loss convergence, we also characterize the average forgetting within cycles, and show that the forgetting vanishes to zero at a faster rate than the loss convergence.

3.1 Definitions and Assumptions

To this end, we first state some necessary assumptions and additional notation. The first assumption is that the joint dataset is linearly separable:

Assumption 3.1 (Joint Separability)**.** There exists $\boldsymbol{w} \in \mathbb{R}^d$ such that $y_i \boldsymbol{x}_i^\top \boldsymbol{w} > 0$ for $\forall i \in I$.

Under Assumption 3.1, we can state an important definition central to our analysis. We define the **joint (offline)** ℓ_2 **max-margin solution** (where we usually omit "ℓ_2" for convenience)

$$\hat{\boldsymbol{w}} := \arg\min_{\boldsymbol{w} \in \mathbb{R}^d} \|\boldsymbol{w}\|^2 \quad \text{subject to } \; y_i \boldsymbol{x}_i^\top \boldsymbol{w} \geq 1, \; \forall i \in I. \tag{3}$$

It can be shown that the optimization problem in Equation (3) has a unique solution $\hat{\boldsymbol{w}}$ (Mohri et al., 2018). Max-margin solutions are of key interest in the study of linear classification, because it is well-known that they have good generalization guarantees (Vapnik, 2013) and GD applied to a single separable binary classification problem has an implicit bias towards its ℓ_2 max-margin solution (Soudry et al., 2018). To be more specific, it is shown in Soudry et al. (2018) that the norm of GD iterates diverges to infinity, but their direction converges to $\frac{\hat{\boldsymbol{w}}}{\|\hat{\boldsymbol{w}}\|}$. In our CL setting, we consider running multiple steps of GD on one task at a time and still aim to find the joint max-margin solution that solves all tasks.

Given the definition of joint max-margin solution, we now define several key quantities. The **maximum margin** of (normalized) $\hat{\boldsymbol{w}}$ is defined as

$$\phi := \min_{i \in I} \frac{y_i \boldsymbol{x}_i^\top \hat{\boldsymbol{w}}}{\|\hat{\boldsymbol{w}}\|}. \tag{4}$$

In fact, it can be shown that $\phi = \|\hat{\boldsymbol{w}}\|$. A **support vector** is a data point $\boldsymbol{x}_i$ that attains this minimum ϕ; we define the index set of support vectors as $S := \{i \in I : y_i \boldsymbol{x}_i^\top \frac{\hat{\boldsymbol{w}}}{\|\hat{\boldsymbol{w}}\|} = \phi\}$, and define the index sets of support vectors of each task $S_m := S \cap I_m$ for $\forall m \in [0 : M-1]$. Let the support vector matrix be $\boldsymbol{X}_S \in \mathbb{R}^{d \times |S|}$, a submatrix of the data matrix $\boldsymbol{X}$ that only contains columns corresponding to support vectors. Lastly, we define the **second margin** $\theta := \min_{i \in I \setminus S} y_i \boldsymbol{x}_i^\top \hat{\boldsymbol{w}} > 1$, which will appear in our non-asymptotic analysis.

To show directional convergence to the joint max-margin solution (Theorem 3.2), we pose an additional assumption on the support vectors.

Assumption 3.2 (Non-degeneracy Condition)**.** For all $i \in S$, there exists a unique $\alpha_i > 0$ such that $\hat{\boldsymbol{w}} = \sum_{i \in S} \alpha_i \cdot y_i \boldsymbol{x}_i$.

[3] trends.google.com/trends/

Assumption 3.2 is adopted from Soudry et al. (2018). According to their analysis, this holds for almost all datasets sampled from a continuous distribution. Intuitively, for a general dataset, no more than d support vectors can be on the same hyperplane.

In the upcoming sections, we present four theorems on the convergence, implicit bias, and forgetting of sequential GD. The theorems rely on different assumptions on the loss $\ell(u)$; we collect them here. It is noteworthy that the logistic loss $\ell(u) = \ln(1 + e^{-u})$ satisfies all the assumptions listed below.

Assumption 3.3. The loss $\ell(u)$ is a positive, differentiable, β-smooth function, monotonically decreasing to zero, and $\limsup_{u\to-\infty} \ell'(u) < 0$.

Assumption 3.4 (Tight Exponential Tail)**.** The negative loss derivative $-\ell'(u)$ has a tight exponential tail. i.e., there exist positive constants μ_+, μ_-, and $\bar{u}$ such that $\forall u > \bar{u}$:

$$(1 - \exp(-\mu_- u))e^{-u} \le -\ell'(u) \le (1 + \exp(-\mu_+ u))e^{-u}$$

Assumption 3.5 (Convexity)**.** The loss $\ell(u)$ is a convex function.

3.2 Asymptotic Results: Loss Convergence & Implicit Bias to Joint Max-margin

Now, we analyze the asymptotic convergence of offline training loss and characterize the directional convergence of sequential GD (2) on jointly separable cyclic tasks. We start by understanding the asymptotic behavior of the joint task loss $\mathcal{L}(\boldsymbol{w})$.

Theorem 3.1. *Let $\{\boldsymbol{w}_k^{(t)}\}_{k\in[0:K-1],t\ge0}$ be the sequence of GD iterates (2) from any starting point $\boldsymbol{w}_0^{(0)}$, where tasks are given cyclically. Under Assumptions 3.1 and 3.3, if the learning rate satisfies $\eta < \min\left\{\frac{1}{2MK\beta\sigma_{\max}^2}, \frac{\phi^2}{4K\beta\sigma_{\max}^3(M\phi+\sigma_{\max})}\right\}$, then*

1. *Loss converges to zero:* $\lim_{t\to\infty} \mathcal{L}(\boldsymbol{w}_k^{(t)}) = 0, \forall k \in [0 : K-1]$.
2. *Every data point is eventually classified correctly:* $\lim_{t\to\infty} \boldsymbol{x}_i^\top \boldsymbol{w}_k^{(t)} = \infty, \forall k \in [0 : K-1], i \in I$.
3. *Square sum of the change of weight is finite:* $\sum_{t=0}^{\infty}\sum_{k=0}^{K-1} \|\boldsymbol{w}_{k+1}^{(t)} - \boldsymbol{w}_k^{(t)}\|^2 < \infty$.

Theorem 3.1 shows that cyclic continual learning on the jointly separable data will eventually learn all tasks, or equivalently, find an offline solution without any additional techniques such as regularization. This result matches the recent empirical findings that DNN can mitigate catastrophic forgetting when tasks are given repetitively (Lesort et al., 2023). The last part on the square sum of the change is used to prove the upcoming Theorem 3.2. We note that Theorem 3.1 does not require convexity of ℓ. The proof can be found in Appendix D.1.

Theorem 3.1 shows that the joint loss converges to zero. However, due to the joint separability (Assumption 3.1), there are multiple directions in which $\boldsymbol{w}_k^{(t)}$ could evolve to make the offline training loss decay to zero. That is, the loss convergence only guarantees finding *an* offline solution, but does not characterize *which*. Under additional assumptions of non-degeneracy and tight exponential tails, we characterize *which* direction $\boldsymbol{w}_k^{(t)}$ diverges to, and show that the model parameter in fact aligns with the joint ℓ_2 max-margin solution $\hat{\boldsymbol{w}}$ (3).

Theorem 3.2. *Let $\{\boldsymbol{w}_k^{(t)}\}_{k\in[0:K-1],t\ge0}$ be the sequence of GD iterates (2) from any starting point $\boldsymbol{w}_0^{(0)}$, where tasks are given cyclically. Under Assumptions 3.1, 3.2, 3.3, and 3.4, if the learning rate satisfies $\eta < \min\left\{\frac{1}{2MK\beta\sigma_{\max}^2}, \frac{\phi^2}{4K\beta\sigma_{\max}^3(M\phi+\sigma_{\max})}\right\}$, then $\boldsymbol{w}_k^{(t)}$ will behave as:*

$$\boldsymbol{w}_k^{(t)} = \ln\left(\tfrac{K}{M}t\right)\hat{\boldsymbol{w}} + \boldsymbol{\rho}_k^{(t)},$$

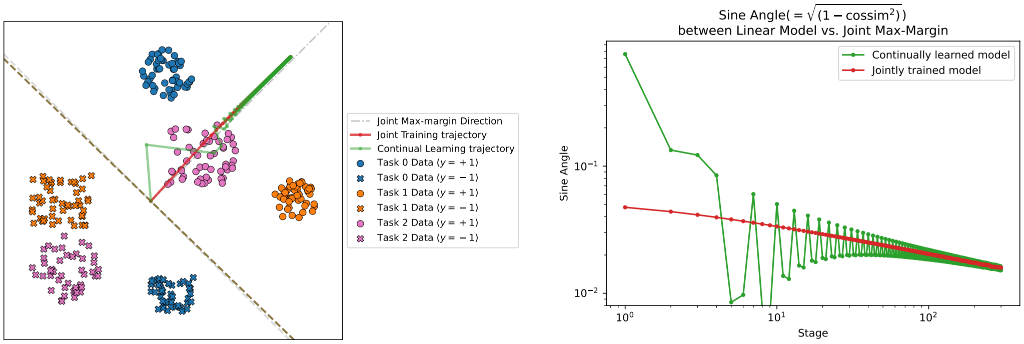
where $\|\boldsymbol{\rho}_k^{(t)}\|$ stays bounded as t grows.

The proof is in Appendix D.2. The key implication of Theorem 3.2 is that the weight vector converges in the direction of the joint max-margin solution, while diverging in magnitude in a rate $\mathcal{O}(\ln t)$:

$$\lim_{t\to\infty} \frac{\boldsymbol{w}_k^{(t)}}{\|\boldsymbol{w}_k^{(t)}\|} = \frac{\hat{\boldsymbol{w}}}{\|\hat{\boldsymbol{w}}\|}, \quad \forall k \in [0 : K-1]. \tag{5}$$

It implies that standard gradient descent without any regularization not only learns every task but also converges to the joint max-margin direction. This result suggests the potential benefits of naive training methods without common CL techniques such as regularization.

On Assumption 3.2. As noted earlier, the non-degeneracy assumption (Assumption 3.2) is borrowed from Soudry et al. (2018); the purpose of adopting this assumption is to facilitate a more complete analysis of the residual $\boldsymbol{\rho}_k^{(t)}$. In fact, in Soudry et al. (2018), the conclusion on the directional convergence (similar to (5), but for single-task GD training) continues to hold even without Assumption 3.2. In light of this, we also believe that directional convergence of sequential GD (5) will hold even without Assumption 3.2, but we did not pursue removing the assumption because it does not offer substantial additional insights.



(a) 2D visualization of data points, the training trajectory, and the decision boundaries (dashed).

(b) Sine angles, implying the implicit bias toward joint max-margin direction.

Figure 2: **Comparison between continually learned and jointly trained linear classifier.** We generate three jointly separable binary classification tasks (with 2D inputs) and run (1) sequential GD in a cyclic task ordering and (2) full-batch GD. It is well-known that the offline full-batch GD converges to the offline ℓ_2 max-margin solution (Soudry et al., 2018). We verify a similar implicit bias of sequential GD iterates (which we proved in Theorem 3.2) by observing the decrease in angle between the model weight and the joint max-margin direction (set as $(1, 1)$). We also observe similar phenomena for more general experimental setup (e.g., random task ordering): see Appendix C.2.

Beyond repetition of fixed datasets. Although we analyze continual learning in a setting where each task has a fixed dataset, the insight of our analysis extends to general setups. To show this, we conduct experiments in a setting where each task has its own (separable) data distribution and a dataset is freshly sampled at every new stage. We observe the same directional convergence behavior of sequential GD toward the true joint max-margin direction. The detailed results are in Appendix C.2.4.

Beyond linear model. We also provide experiments with shallow ReLU networks, verifying analogous insights on implicit bias and loss convergence of continually learned models: see Appendix C.4.

3.3 Non-asymptotic Results: Loss Convergence and Forgetting Bounds

In Section 3.2, we presented asymptotic results characterizing the convergence of total training loss to zero and the directional convergence of sequential GD iterates to the max-margin solutions. We now supplement these results with an additional *non-asymptotic* convergence analysis on total training loss, which we can use to obtain a non-asymptotic analysis of *cycle-averaged forgetting* as well.

As aforementioned, the main challenge in CL is mitigating catastrophic forgetting. Analyses of continual learning methods aim to show that methods decrease forgetting, theoretically or empirically. In this paper, we are interested in how strong forgetting is in our continual linear classification setup.

We start by stating a common definition of forgetting, which quantifies the amount of loss increase at the end of stage t compared to the end of K steps of GD on $\mathcal{L}^{(s)}$ executed in stage $s \leq t$.

Definition 3.6 (Forgetting)**.** The **forgetting** at stage t of the task learned in stage s $(\leq t)$ is the change of the task loss $\mathcal{L}^{(s)}$ from the moment the K GD steps were finished in stage s. That is,

$$\mathcal{F}^{(s)}(t) := \mathcal{L}^{(s)}(\boldsymbol{w}_K^{(t)}) - \mathcal{L}^{(s)}(\boldsymbol{w}_K^{(s)}).$$

Notice that forgetting is zero by definition when $t = s$. While it is usually expected that forgetting is a positive quantity, it could be also negative by definition. Such a case can happen when the tasks seen in stages between s and t are well-aligned with $\mathcal{L}^{(s)}$, so that the model improves on the task previously seen in stage s. This phenomenon is called *backward knowledge transfer*.

When CL tasks do not necessarily repeat, it is common to evaluate the average forgetting over all past stages, namely $\frac{1}{t}\sum_{s=0}^{t-1}\mathcal{F}^{(s)}(t)$. However, since we consider the case where tasks are given cyclically, it is natural to define our quantity of interest as below:

Definition 3.7 (Cycle-averaged Forgetting)**.** The **cycle-averaged forgetting** at cycle j is the average loss change of previous tasks from the stage in which it was learned. That is,

$$\mathcal{CF}(j) := \frac{1}{M}\sum_{m=0}^{M-1}\mathcal{F}^{(Mj+m)}(Mj+M-1) = \frac{1}{M}\sum_{m=0}^{M-1}\mathcal{L}_m(\boldsymbol{w}_0^{(Mj+M)}) - \mathcal{L}_m(\boldsymbol{w}_K^{(Mj+m)}).$$

By studying cycle-averaged forgetting, we would like to understand how much forgetting happens during the cyclic learning process, and how the amount of forgetting changes as we repeat the cycles.

Although the asymptotic convergence to joint max-margin solution (Theorem 3.2) suggests that the model will suffer a diminishing level of forgetting in the long run, characterizing the amount of forgetting for a given cycle count J necessitates a more careful non-asymptotic analysis of the loss convergence. For this purpose, we present an additional theorem characterizing the non-asymptotic convergence of offline training loss $\mathcal{L}$; we then build on this theorem to prove upper and lower bounds on cycle-averaged forgetting. The new convergence theorem requires the same set of assumptions as Theorem 3.1, except for an additional assumption of convex $\ell(u)$.

Theorem 3.3. *Under the same setting as Theorem 3.1 with an additional Assumption 3.5, for any* $m \in [0 : M-1]$ *and* $k \in [0 : K-1]$*, we have*

$$\mathcal{L}(\boldsymbol{w}_k^{(MJ+m)}) \le \left(|S| + \frac{\sum_{i=0}^{m-1}|S_i| + \frac{k}{K}|S_m|}{J}\right)\ell(\ln J) + \frac{\left\|\boldsymbol{w}_0^{(0)} - \hat{\boldsymbol{w}}\ln J\right\|^2}{2\eta KJ} + \frac{D_1}{J}$$
$$+ \left(|I| - |S| + \frac{\sum_{i=0}^{m-1}(|I_i| - |S_i|) + \frac{k}{K}(|I_m| - |S_m|)}{J}\right)\ell(\theta\ln J),$$

where $\theta > 1$ *is the second margin defined in Section 3.1, and*

$$D_1 := \frac{4\sigma_{\max}^2}{\phi^2}\left(\mathcal{L}(\boldsymbol{w}_0^{(0)}) + \left(1 + \frac{\eta K\sigma_{\max}^3\beta}{\phi(1-\eta MK\sigma_{\max}^2\beta)}\right)\frac{\eta K\sigma_{\max}}{\phi(1-\eta MK\sigma_{\max}^2\beta)}\left\|\nabla\mathcal{L}(\boldsymbol{w}_0^{(0)})\right\|^2\right).$$

The proof can be found in Appendix D.3. One can revisit Section 3.1 to recall the definitions of symbols such as $\sigma_{\max}$, ϕ, and β. The bound in Theorem 3.3 may be a bit difficult to parse. First of all, notice that whenever $\ell(u) \le e^{-u}$, which is true for logistic loss $\ell(u) = \ln(1+e^{-u})$, we have $\ell(\ln J) \le \frac{1}{J}$ and $\ell(\theta\ln J) \le \frac{1}{J^\theta}$. Combined with other terms, this implies an overall $\mathcal{O}(\frac{\ln^2 J}{J})$ upper bound for the offline training loss.

Next, we can notice for any fixed J, the upper bound in fact *grows* with k and m. This unusual growth of the upper bound reflects the effect of forgetting that can happen during cycles. Even though such an increase in loss does not usually occur with a small learning rate, it is not impossible. For example, when most of the tasks have individual max-margin directions different from the joint max-margin direction, this situation can occur. We demonstrate this mid-cycle increase of joint loss using a toy example in Appendix C.3.

The possible increase of loss due to forgetting becomes less of an issue as training proceeds since the terms increasing in m and k are all divided by an additional factor of J and hence decay faster than other terms. Therefore, the increase of loss bound becomes smaller for larger J, indicating smaller forgetting during cycles. Despite the possible forgetting, Theorem 3.3 indicates that if tasks are given cyclically, then the loss bound is guaranteed to decrease at the end of every cycle.

We can now use Theorem 3.3 to derive bounds on cycle-averaged forgetting we defined in Definition 3.7. We characterize how fast the cycle-averaged forgetting $\mathcal{CF}(J)$ converges to zero as the cycles replay. For this theorem, we specifically consider the logistic loss, which satisfies all loss assumptions in the paper.

Theorem 3.4. *Let $\ell(u) = \ln(1+e^{-u})$ be the logistic loss. If the learning rate satisfies $\eta < \min\left\{\frac{1}{2MK\beta\sigma_{\max}^2}, \frac{\phi^2}{4K\beta\sigma_{\max}^3(M\phi+\sigma_{\max})}\right\}$, then the cycle-averaged forgetting $\mathcal{CF}(J)$ for cycle J satisfies the following upper and lower bounds:*

$$-\eta K \cdot L(J)^2 \cdot \frac{\sum_{p\neq q} N_{p,q}}{M} \leq \mathcal{CF}(J) \leq \eta K \cdot L(J)^2 \cdot \frac{-\sum_{p\neq q} \bar{N}_{p,q}}{M},$$

where

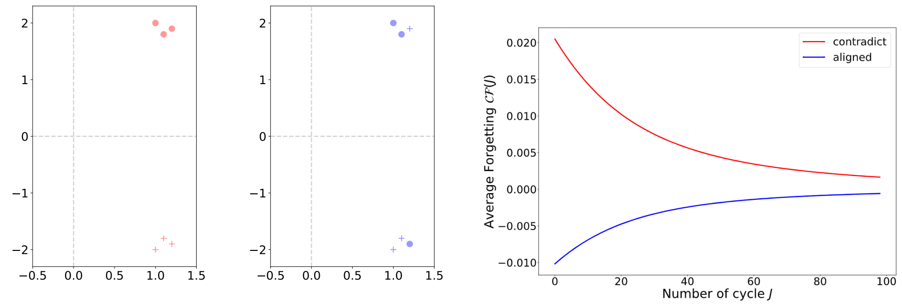
$$L(J) := \frac{1}{J}\left(\left(|S| + \frac{|I|-|S|}{J^{\theta-1}}\right)\left(1+\frac{1}{J}\right) + \frac{\|\boldsymbol{w}_0^{(0)} - \hat{\boldsymbol{w}} \ln J\|^2}{2\eta K} + D_1\right) = \mathcal{O}\left(\frac{\ln^2 J}{J}\right)$$

$$N_{p,q} := \sum_{\substack{(i,j)\in I_p\times I_q \\ \boldsymbol{x}_i^\top \boldsymbol{x}_j > 0}} \boldsymbol{x}_i^\top \boldsymbol{x}_j > 0, \quad \bar{N}_{p,q} := \sum_{\substack{(i,j)\in I_p\times I_q \\ \boldsymbol{x}_i^\top \boldsymbol{x}_j < 0}} \boldsymbol{x}_i^\top \boldsymbol{x}_j < 0.$$

The proof is in Appendix D.4. Theorem 3.4 shows a nonnegative upper bound and a nonpositive lower bound on the cycle-averaged forgetting at cycle J. Note that both upper and lower bounds decay to zero as J grows. Convergence of $\mathcal{CF}(J)$ is of rate $\mathcal{O}(\frac{\ln^4 J}{J^2})$, which is faster than the convergence rate $\mathcal{O}(\frac{\ln^2 J}{J})$ of joint loss shown in Theorem 3.3.

The bounds in Theorem 3.4 reflect how positive/negative data alignment between different tasks impact forgetting. The quantities $N_{p,q}$ and $\bar{N}_{p,q}$ capture show how similar and different (respectively) data points are, for a pair of tasks (p,q). In particular, when $\sum_{p\neq q} \bar{N}_{p,q} = 0$, it is guaranteed that average forgetting does not happen, regardless of J. Rather, training on a task will decrease the loss for all previously learned tasks, which can be thought of as an extreme form of *backward knowledge transfer*. On the other hand, when $\sum_{p\neq q} N_{p,q} = 0$, it is guaranteed that the model will suffer forgetting at every cycle; however, even in this case, Theorem 3.4 implies that repeating tasks over cycles mitigates catastrophic forgetting.

Even when the joint dataset $\mathcal{D}$ is the same, forgetting behavior can differ depending on how the data points are distributed over different tasks. This matches the former theoretical explanation of how distribution affects forgetting. For instance, Lin et al. (2023) show that a larger distance between each task's optimal solution leads to larger forgetting. For a straightforward interpretation, consider the following example of two tasks: their cycle-averaged forgetting for two different decompositions of $\mathcal{D}$ is plotted in Figure 3. We can observe that two tasks contradicting each other (i.e., large $\bar{N}_{1,2}$) results in positive forgetting, whereas two tasks aligning better (i.e., large $N_{1,2}$) exhibit negative forgetting. Nevertheless, cycle-averaged forgetting converges to zero in both cases.



(a) Contradicting case (b) Aligned case (c) Cycle-averaged Forgetting

Figure 3: We compare two continual learning scenarios with the same joint dataset $\mathcal{D} = \{(1,2),(1.1,1.8),(1.2,1.9),(1,-2),(1.1,-1.8),(1.2,-1.9)\}$, where labels are all $+1$ and hence omitted. We mark Task 1's data as '○' and Task 2's data as '+'. We used $M=2$ and $K=10$. Figure 3(a) displays a data composition that makes large $\bar{N}_{1,2}$, whereas Figure 3(b) displays a data composition that makes relatively small $\bar{N}_{1,2}$ and large $N_{1,2}$. Figure 3(c) is a plot of cycle-averaged forgetting (CF), evolving over cycles. For "contradict" scenario (red), CF is always positive and diminishing to 0. In contrast, for "aligned" scenario (blue), CF is always negative and rising to 0.

4 Random-order Learning of Jointly Separable Tasks

In this section, we consider the scenario where tasks are given in a random order, while still assuming that the tasks are jointly separable. Formally, at the end of K-th GD iteration of stage t, the next task is sampled independently and uniformly at random. Even in this case, our analysis reveals that the asymptotic results shown in Section 3.2 continue to hold *almost surely*.

We first show that the offline training loss converges to zero almost surely, which is a random-order counterpart of Theorem 3.1. The proof is in Appendix E.1.

Theorem 4.1. *Let $\{\boldsymbol{w}_k^{(t)}\}_{k\in[0:K-1],t\geq 0}$ be the sequence of GD iterates (2) from any starting point $\boldsymbol{w}_0^{(0)}$, where tasks are given randomly. Under Assumptions 3.1 and 3.3, if the learning rate satisfies $\eta < \frac{2\phi^2}{\beta\sigma_{\max}^4}$, then the following statements hold with probability 1:*

1. *Loss converges to zero:* $\lim_{t\to\infty}\mathcal{L}(\boldsymbol{w}_k^{(t)}) = 0, \forall k \in [0 : K-1]$.
2. *Every data point is classified correctly:* $\lim_{t\to\infty}\boldsymbol{x}_i^\top \boldsymbol{w}_k^{(t)} = 0, \forall k \in [0 : K-1], i \in I$.
3. *Square sum of the change of weight is finite:* $\sum_{t=0}^{\infty}\sum_{k=0}^{K-1}\|\boldsymbol{w}_{k+1}^{(t)} - \boldsymbol{w}_k^{(t)}\|^2 < \infty$.

We derive the same asymptotic loss convergence result, with a minor difference that the learning rate can be chosen independent of the number of tasks M and the iteration count K.

We now state the random-order counterpart of Theorem 3.2, which implies that the sequential GD iterates converge to joint ℓ_2 max-margin solution almost surely. The proof is in Appendix E.2.

Theorem 4.2. *Let $\{\boldsymbol{w}_k^{(t)}\}_{k\in[0:K-1],t\geq 0}$ be the sequence of GD iterates (2) from any starting point $\boldsymbol{w}_0^{(0)}$, where tasks are given randomly. Under Assumptions 3.1, 3.2, 3.3, and 3.4, if the learning rate satisfies $\eta < \frac{2\phi^2}{\beta\sigma_{\max}^4}$, then with probability 1, $\boldsymbol{w}_k^{(t)}$ will behave as:*

$$\boldsymbol{w}_k^{(t)} = \ln\left(\tfrac{K}{M}t\right)\hat{\boldsymbol{w}} + \boldsymbol{\rho}_k^{(t)},$$

where $\|\boldsymbol{\rho}_k^{(t)}\|$ stays bounded as t grows.

5 Beyond Jointly Separable Tasks

Now we turn our attention to the CL on a strictly *non-separable* set of M tasks, where the tasks are presented in a *cyclic* manner. In this section, we assume that the set of all data points spans the whole space $\mathbb{R}^d$ without loss of generality. This is a mild assumption because every gradient update happens in the span of data points. In this case, if we assume the strict non-separability on the full dataset (see Assumption 5.1), the offline training loss $\mathcal{L}(\boldsymbol{w}) = \sum_{m=0}^{M-1}\mathcal{L}_m(\boldsymbol{w})$ defined with logistic losses becomes strictly convex and coercive (i.e., $\lim_{\|\boldsymbol{w}\|\to\infty}\mathcal{L}(\boldsymbol{w}) = +\infty$); thus, it has a unique minimum $\boldsymbol{w}_\star \in \mathbb{R}^d$. We show that, under cyclic task ordering, the iterates of sequential GD converge to $\boldsymbol{w}_\star$ at a rate $\mathcal{O}(\frac{\ln^2 J}{J^2})$, which is faster than the loss convergence rate of the separable case.

The core idea of the analysis is to identify the local strong convexity of the offline training loss on a compact set on which every end-of-cycle iterates lie (Freund et al., 2018). To this end, we require a strict non-separability of the joint dataset as defined below.

Assumption 5.1 (Joint Strict Non-separability Condition (Freund et al., 2018))**.** Assume that the whole collection of data points is of full rank: $\operatorname{span}(\{\boldsymbol{x}_i : i = 0, \dots, N-1\}) = \mathbb{R}^d$. Additionally, assume that there exists $b > 0$ defined as

$$b := \min_{\boldsymbol{v}\in\mathbb{R}^d:\|\boldsymbol{v}\|=1}\sum_{i=0}^{N-1}[y_i\boldsymbol{x}_i^\top\boldsymbol{v}]^-,$$

where $[a]^- := \max\{0, -a\}$.

Note that a large b means that the joint data points are highly non-separable: for any classifier vector $\boldsymbol{v}$, there exist some data points with the incorrect prediction of the label with a large margin. We also remark that individual tasks are not necessarily strictly non-separable. Hence, our analysis covers the case where all individual tasks are separable while the full dataset is not separable.

We additionally assume some mild properties of the loss function $\ell(\cdot)$.

Assumption 5.2. The loss function $\ell : \mathbb{R} \to \mathbb{R}_+$ is a strictly convex, β-smooth function with a positive second derivative such that $\ell(u) \geq G \cdot [u]^-$ for some $G > 0$.

Note that the logistic loss $\ell(u) = \ln(1 + e^{-u})$ satisfies the assumption above with $\beta = 1/4$ and $G = 1$. From the assumptions, we have that (1) the risk of m-th task $\mathcal{L}_m(\boldsymbol{w}) = \sum_{i \in I_m} \ell(y_i \boldsymbol{x}_i^\top \boldsymbol{w})$ is convex and β_m-smooth for $\beta_m := \beta \lambda_{\max}\left(\boldsymbol{X}_m \boldsymbol{X}_m^\top\right)$ where $\boldsymbol{X}_m \in \mathbb{R}^{d \times |I_m|}$ is a data matrix of task m consisting of columns $\{\boldsymbol{x}_i : i \in I_m\}$; (2) due to the strict non-separability, the offline training loss $\mathcal{L}(\boldsymbol{w}) = \sum_{m=0}^{M-1} \mathcal{L}_m(\boldsymbol{w})$ has a unique minimum $\boldsymbol{w}_\star$. Furthermore, we can prove that the end-of-cycle iterates of the sequential GD stay bounded in a compact set $\mathcal{W}$ around $\boldsymbol{w}_\star$. Consequently, we have a local strong convexity of the offline training loss on $\mathcal{W}$. The proof is in Appendix F.1.

Lemma 5.1. *Consider learning M linear classification tasks cyclically. Suppose that Assumptions 5.1 and 5.2 hold. Let $B := \sum_{m=0}^{M-1} \beta_m$ and $V_\star := \sum_{m=0}^{M-1} \frac{1}{\beta_m} \|\nabla \mathcal{L}_m(\boldsymbol{w}_\star)\|^2$. Take a step size $\eta \leq \frac{1}{2\sqrt{2}KB}$. Then, there exists a compact set $\mathcal{W} \subset \mathbb{R}^d$ containing $\boldsymbol{w}_\star$ and every $\boldsymbol{w}_0^{(jM)}$ ($j = 0, 1, 2, \ldots$), whose radius is independent of J (the number of cycles) but depends on other parameters like b, G, B, and $V_\star$. Also, the offline training loss $\mathcal{L}$ is μ-strongly convex on $\mathcal{W}$, where*

$$\mu := \left(\min_{i \in [0:N-1], \boldsymbol{w} \in \mathcal{W}} \ell''\left(y_i \boldsymbol{x}_i^\top \boldsymbol{w}\right)\right) \cdot \lambda_{\min}\left(\boldsymbol{X}\boldsymbol{X}^\top\right) > 0. \tag{6}$$

We remark that the radius of the set $\mathcal{W}$ largely depends on the non-separability b (Assumption 5.1): loosely speaking, $\mathcal{W}$ can be arbitrarily large if b goes to zero since $\|\boldsymbol{w} - \boldsymbol{w}_\star\| = \mathcal{O}(1/b)$ for any $\boldsymbol{w} \in \mathcal{W}$. In particular, for the logistic loss ℓ, the local strong convexity coefficient μ can get small if b is small, because of (possibly) a large radius of $\mathcal{W}$. With the local strong convexity, we finally have a fast non-asymptotic convergence rate of $\tilde{\mathcal{O}}(J^{-2})$ towards the global minimum. The proof can be found in Appendix F.2.

Theorem 5.2. *Suppose we learn M tasks cyclically for $J > 1$ cycles. We adopt the notation from Lemma 5.1. If we choose a step size*

$$\eta = \min\left\{\frac{1}{2\sqrt{2}KB}, \frac{1+2\sqrt{2}}{2\sqrt{2}KJ} \ln\left(J^2 \cdot \max\left\{1, \frac{\|\boldsymbol{w}_0^{(0)} - \boldsymbol{w}_\star\|^2 \mu^3}{B^2 V_\star}\right\}\right)\right\},$$

then the final iterate of sequential GD satisfies

$$\left\|\boldsymbol{w}_0^{(MJ)} - \boldsymbol{w}_\star\right\|^2 \leq \tilde{\mathcal{O}}\left(\exp\left(-\frac{\mu J}{(1+2\sqrt{2})B}\right) \cdot \left\|\boldsymbol{w}_0^{(0)} - \boldsymbol{w}_\star\right\|^2 + \frac{B^2 V_\star \ln^2 J}{\mu^3 J^2}\right), \tag{7}$$

where we hide a poly-logarithmic factor of J in Equation (7).

Remark on the loss convergence rate. Since the $\mathcal{L}(\boldsymbol{w})$ is B-smooth, it satisfies that

$$\mathcal{L}(\boldsymbol{w}) - \mathcal{L}(\boldsymbol{w}_\star) \leq \langle \nabla \mathcal{L}(\boldsymbol{w}_\star), \boldsymbol{w} - \boldsymbol{w}_\star \rangle + \frac{B}{2}\|\boldsymbol{w} - \boldsymbol{w}_\star\|^2 = \frac{B}{2}\|\boldsymbol{w} - \boldsymbol{w}_\star\|^2. \tag{8}$$

Thus, our Theorem 5.2 naturally implies the loss convergence at the same rate (in terms of J).

Experiments on a real-world dataset. For those interested, we also provide an experiments on a real-world dataset CIFAR-10 (Krizhevsky et al., 2009), which is not guaranteed to be linearly separable: see Appendix C.5.

6 Conclusion

We considered continual linear classification by running gradient descent for a fixed number of iterations per task. When there exist solutions that can solve every task, we found that even without any regularization or CL methods, the classifier eventually converges to the joint max-margin direction. This implicit bias happens on both cyclic/random task ordering. We further presented a non-asymptotic analysis on cycle-averaged forgetting with respect to positive/negative alignments of tasks and the number of cycles. Lastly, we showed that if no linear classifier solves all tasks simultaneously, the model converges to the unique minimum of the offline training loss. As for future work, we believe the convergence on continual classification can be extended to other model structures, bridging the gap between empirical findings and theoretical understanding of the impact of task repetition. Also, our results are restricted to the "small learning rate" regime, and do not cover larger learning rates or even the "edge of stability" regime (Wu et al., 2024); relaxing this restriction is left for future work.

REFERENCES

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 139–154, 2018. 1

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

Mehdi Abbana Bennani, Thang Doan, and Masashi Sugiyama. Generalisation guarantees for continual learning with orthogonal gradient descent. *arXiv preprint arXiv:2006.11942*, 2020. 1, A

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *ICML Workshop on Multi-Task and Lifelong Reinforcement Learning*, 2019. 1

Andrea Cossu, Gabriele Graffieti, Lorenzo Pellegrini, Davide Maltoni, Davide Bacciu, Antonio Carta, and Vincenzo Lomonaco. Is class-incremental enough for continual learning? *Frontiers in Artificial Intelligence*, 5:829842, 2022. 2.2

Thang Doan, Mehdi Abbana Bennani, Bogdan Mazoure, Guillaume Rabusseau, and Pierre Alquier. A theoretical analysis of catastrophic forgetting through the ntk overlap matrix. In *International Conference on Artificial Intelligence and Statistics*, pp. 1072–1080. PMLR, 2021. 1, A

Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019. E.3, E.4

Itay Evron, Edward Moroshko, Rachel Ward, Nathan Srebro, and Daniel Soudry. How catastrophic can catastrophic forgetting be in linear regression? In *Conference on Learning Theory*, pp. 4028–4079. PMLR, 2022. 1, 2.2

Itay Evron, Edward Moroshko, Gon Buzaglo, Maroun Khriesh, Badea Marjieh, Nathan Srebro, and Daniel Soudry. Continual learning in linear classification on separable data. In *International Conference on Machine Learning*, pp. 9440–9484. PMLR, 2023. 1, 1, 1, 2.2, 3, 6, B, B, B.1, B

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3762–3773. PMLR, 2020. 1, A

Robert M Freund, Paul Grigas, and Rahul Mazumder. Condition number analysis of logistic regression, and its implications for standard first-order solution methods. *arXiv preprint arXiv:1810.08727*, 2018. 5, 5.1

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

Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013. 1

Mustafa N Gultekin and N Bulent Gultekin. Stock market seasonality: International evidence. *Journal of financial economics*, 12(4):469–481, 1983. 2.2

Joachim Houyon, Anthony Cioppa, Yasir Ghunaim, Motasem Alfarra, Anaïs Halin, Maxim Henry, Bernard Ghanem, and Marc Van Droogenbroeck. Online distillation with continual learning for cyclic domain shifts, 2023. URL https://arxiv.org/abs/2304.01239. 2.2

Khurram Javed and Martha White. Meta-learning representations for continual learning. *Advances in neural information processing systems*, 32, 2019. 1

Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018. A

Ziwei Ji and Matus Telgarsky. Characterizing the implicit bias via a primal-dual analysis. In *Algorithmic Learning Theory*, pp. 772–804. PMLR, 2021. A

Ryo Karakida and Shotaro Akaho. Learning curves for continual learning in neural networks: Self-knowledge transfer and forgetting. In *International Conference on Learning Representations*, 2022. A

Gyuhak Kim, Changnan Xiao, Tatsuya Konishi, Zixuan Ke, and Bing Liu. A theoretical study on solving continual learning. *Advances in neural information processing systems*, 35:5065–5079, 2022. 1, A

Gyuhak Kim, Changnan Xiao, Tatsuya Konishi, and Bing Liu. Learnability and algorithm for continual learning. In *International Conference on Machine Learning*, pp. 16877–16896. PMLR, 2023. 1

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114 (13):3521–3526, 2017. 1

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5, C.5

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

Timothée Lesort, Oleksiy Ostapenko, Pau Rodríguez, Diganta Misra, Md Rifat Arefin, Laurent Charlin, and Irina Rish. Challenging common assumptions about catastrophic forgetting and knowledge accumulation. In *Conference on Lifelong Learning Agents*, pp. 43–65. PMLR, 2023. 3.2

Haoran Li, Jingfeng Wu, and Vladimir Braverman. Fixed design analysis of regularization-based continual learning. In *Conference on Lifelong Learning Agents*, pp. 513–533. PMLR, 2023. 1

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 1

Sen Lin, Peizhong Ju, Yingbin Liang, and Ness Shroff. Theory on forgetting and generalization of continual learning. In *International Conference on Machine Learning*, pp. 21078–21100. PMLR, 2023. 3.3, A

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017. 1

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989. 1

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Razvan Pascanu, and Hassan Ghasemzadeh. Understanding the role of training regimes in continual learning. *Advances in Neural Information Processing Systems*, 33:7308–7320, 2020. 1

M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning, second edition*. Adaptive Computation and Machine Learning series. MIT Press, 2018. ISBN 9780262351362. URL `https://books.google.co.kr/books?id=dWB9DwAAQBAJ`. 3.1

Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3051–3059. PMLR, 2019. A, D.1, D.1.1, D.3

Binghui Peng and Andrej Risteski. Continual learning: a feature extraction formalization, an efficient algorithm, and fundamental obstructions. *Advances in Neural Information Processing Systems*, 35: 28414–28427, 2022. A

Krishnan Raghavan and Prasanna Balaprakash. Formalizing the generalization-forgetting trade-off in continual learning. *Advances in Neural Information Processing Systems*, 34:17284–17297, 2021. 1, A

Haizhou Shi and Hao Wang. A unified approach to domain incremental learning with memory: Theory and algorithm. *Advances in Neural Information Processing Systems*, 36, 2023. 1, A

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017. 1

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018. 1, 3.1, 3.1, 3.2, 2, A, D.2

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013. 1, 3.1

Eli Verwimp, Kuo Yang, Sarah Parisot, Lanqing Hong, Steven McDonagh, Eduardo Pérez-Pellitero, Matthias De Lange, and Tinne Tuytelaars. Clad: A realistic continual learning benchmark for autonomous driving. *Neural Networks*, 161:659–669, 2023. 2.2

Jingfeng Wu, Vladimir Braverman, and Jason D Lee. Implicit bias of gradient descent for logistic regression at the edge of stability. *Advances in Neural Information Processing Systems*, 36, 2024. 6, A

Yingxiang Yang, Zhihan Xiong, Tianyi Liu, Taiqing Wang, and Chong Wang. Fourier learning with cyclical data. In *International Conference on Machine Learning*, pp. 25280–25301. PMLR, 2022. 2.2

CONTENTS

A Other Related Works

Theoretical Results on Continual Learning. Several theoretical analyses have been proposed on classification. Raghavan & Balaprakash (2021) examine the generalization-forgetting trade-off by viewing it as a two-player sequential game, in which player 1 wants to maximize generalization, whereas player 2 wants to minimize forgetting. They show the existence of a balanced point where both players are satisfied with each new task and suggest a new algorithm to achieve the point. Kim et al. (2022) consider Class-Incremental Learning, where the model can see a disjoint subset of the total class at a time. They prove that good Within-task Prediction (WP) and good Task-id Prediction (TP) are necessary and sufficient for good CIL. Furthermore, they relate TP with OOD detection. Shi & Wang (2023) consider Domain-Incremental Learning, where the model can see the different domains in a class over time. They especially suggest a framework with a memory buffer that unified earlier methods.

Lin et al. (2023) distinguish empirical and population risks by drawing samples from Gaussian with true linear regression solutions. Then, they investigate the impact of overparameterization and task similarity over forgetting. Bennani et al. (2020); Doan et al. (2021); Karakida & Akaho (2022) study forgetting in NTK regime. Specifically, Bennani et al. (2020); Doan et al. (2021) analyze forgetting of orthogonal gradient descent (OGD, Farajtabar et al. (2020)), while Karakida & Akaho (2022) study continual transfer learning. Other settings such as Teacher-Student setup (Lee et al., 2021), and feature extraction (Peng & Risteski, 2022) have been considered in Task-Incremental Learing.

Implicit Bias of Gradient Descent for Linear Classification. Soudry et al. (2018) are the first to show that if data is linearly separable, gradient descent with certain loss functions converges to the max-margin direction. Nacson et al. (2019) prove the same result on the same condition but with stochastic gradient descent. Ji & Telgarsky (2018) show the same result with a slower convergence rate, resulting from the absence of degeneracy condition. They also consider cases where data is not separable, yet weight diverges to infinity. Ji & Telgarsky (2021) show a faster convergence rate under decreasing learning rate via a primal-dual analysis. While these findings require small learning rates, Wu et al. (2024) prove that gradient descent with logistic loss converges to the max-margin direction even when the learning rate is large.

B Brief Overview of Evron et al. (2023) and Comparisons

To highlight how our sequential GD algorithm differs from Evron et al. (2023), we briefly summarize the Sequential Max-Margin (SMM) framework considered in the existing paper and its theoretical results.

Evron et al. (2023) consider minimizing the regularized training loss of each task until convergence, where the loss function is chosen to be the exponential loss $\ell(u) = \exp(-u)$. Let $\left\{\boldsymbol{w}_{\lambda\text{-Re}}^{(t)}\right\}_t$ be the iterates trained by regularized continual learning with regularization coefficient λ. The algorithm can be written as follows:

$$\boldsymbol{w}_{\lambda\text{-Re}}^{(t+1)} = \arg\min_{\boldsymbol{w}\in\mathbb{R}^d} \sum_{i\in I^{(t)}} \exp\left(-y_i \boldsymbol{x}_i^\top \boldsymbol{w}\right) + \frac{\lambda}{2}\left\|\boldsymbol{w} - \boldsymbol{w}_{\lambda\text{-Re}}^{(t)}\right\|^2. \tag{9}$$

Also, let $\boldsymbol{w}_{\text{SMM}}^{(t)}$ be the weight trained by the Sequential Max-Margin algorithm. The update rule is as follows:

$$\begin{aligned} \boldsymbol{w}_{\text{SMM}}^{(t+1)} &= \arg\min_{\boldsymbol{w}\in\mathbb{R}^d} \left\|\boldsymbol{w} - \boldsymbol{w}_{\text{SMM}}^{(t)}\right\|^2 \quad \text{subject to} \quad y_i \boldsymbol{x}_i^\top \boldsymbol{w} \geq 1, \forall i \in I^{(t)} \\ &= P^{(t)}(\boldsymbol{w}_{\text{SMM}}^{(t)}). \end{aligned} \tag{10}$$

Here, the operator $P^{(t)}$ can be thought of as the orthogonal projection onto a convex set

$$\left\{\boldsymbol{w}\in\mathbb{R}^d : y_i \boldsymbol{x}_i^\top \boldsymbol{w} \geq 1, \forall i \in I^{(t)}\right\} \tag{11}$$

defined by the margin conditions on data points in $I^{(t)}$. That is, $\boldsymbol{w}_{\text{SMM}}^{(t)}$ is the same as the sequential projection onto such convex sets. Evron et al. (2023) showed the relation of $\boldsymbol{w}_{\lambda\text{-Re}}^{(t)}$ and $\boldsymbol{w}_{\text{SMM}}^{(t)}$, when the regularization coefficient $\lambda \to 0$:

Theorem B.1 (Theorem 3.1 of Evron et al. (2023))**.** *For almost all dataset, in the limit of $\lambda \to 0$, it holds that $\boldsymbol{w}_{\lambda\text{-Re}}^{(t)} \to \boldsymbol{w}_{\text{SMM}}^{(t)}$ with a residual of $O(t \log\log\left(\frac{1}{\lambda}\right))$. Therefore, at any $t = o\left(\frac{\log\left(\frac{1}{\lambda}\right)}{\log\log\left(\frac{1}{\lambda}\right)}\right)$, we get*

$$\lim_{\lambda\to 0} \frac{\boldsymbol{w}_{\lambda\text{-Re}}^{(t)}}{\left\|\boldsymbol{w}_{\lambda\text{-Re}}^{(t)}\right\|} = \frac{\boldsymbol{w}_{\text{SMM}}^{(t)}}{\left\|\boldsymbol{w}_{\text{SMM}}^{(t)}\right\|}.$$

Based on this equivalence in terms of parameter *direction*, Evron et al. (2023) expect that the behavior of $\boldsymbol{w}_{\lambda\text{-Re}}^{(t)}$ can be analyzed through the lens of $\boldsymbol{w}_{\text{SMM}}^{(t)}$ as long as λ is close to 0, since Theorem B.1 holds for all $t = o\left(\frac{\log\left(\frac{1}{\lambda}\right)}{\log\log\left(\frac{1}{\lambda}\right)}\right)$.

Given this background, we now highlight some differences between Evron et al. (2023) and our analysis. First of all, as seen in (9), Evron et al. (2023) study regularized exponential loss trained until convergence, whereas we study unregularized logistic loss trained for a fixed number of iterations. Training the weakly regularized loss until convergence, in conjunction with limit $\lambda \to 0$, sends each $\boldsymbol{w}_{\lambda\text{-Re}}^{(t)}$ to infinity. Hence, each stage requires a growing number of iterations, and the grounds for the equivalence between (9) and (10) becomes weaker, since the solutions become vastly different in terms of magnitude.

Second, thanks to the connection between weakly-regularized continual learning and SMM, Evron et al. (2023) could obtain the exact trajectory of every stage via the projection method. On the other hand, in our sequential GD setting, it is very difficult to keep track of the exact location of the iterate after one task is trained, since the iterates are updated multiple times but training stops before convergence. This makes it challenging to analyze implicit bias and forgetting via tracking the exact trajectory stage by stage. We use different proof techniques from Evron et al. (2023) to overcome this challenge. Rather than pinpointing the exact position of the iterate after each stage, we focus on the direction that sequential GD eventually converges to.

On top of that, importantly, our analysis of sequential GD reveals that training on unregularized loss using a fixed number of GD iterations results in the joint/offline max-margin solution. In contrast, although the convergence to *some* offline solutions is already shown for SMM (Evron et al., 2023), the converged offline solution can be different from the offline *max-margin* solution. In fact, in the next section (Appendix C.1), we demonstrate by a toy example that SMM can indeed converge to a point other than the joint max-margin solution.

C Experiment Details & Omitted Experimental Results

C.1 Experiment Details of Figure 1

In this section, we present a simple toy example that demonstrates interesting facts about max-margin solutions in continual linear classification:

- The joint max-margin direction of the joint dataset can be quite different from the max-margin solutions of individual tasks. Specifically, the joint solution may *not* be on the subspace spanned by the individual solutions.
- The limit of Sequantial Max-Margin (SMM) iterations can be different from the joint max-margin solution, whereas the limit direction of sequential GD does align with it.

We consider the case of $M = 2$ tasks, where the input points come from $\mathbb{R}^3$. Without loss of generality, we assume that all the labels are $+1$, and hence omit them. We let $\{(1,1,0),(1,-2,1)\}$ be the dataset of task 1, and $\{(1,0,1),(1,1,-2)\}$ be the data of task 2. One can verify that:

- Their joint max-margin direction is $(1,0,0)$.
- The max-margin direction for task 1 is $(\frac{10}{11},\frac{1}{11},\frac{3}{11})$.
- The max-margin direction for task 2 is $(\frac{10}{11},\frac{3}{11},\frac{1}{11})$.

Therefore, we can observe that the joint max-margin solution does not belong to the span of individual max-margin solutions.

We ran numerical experiments running the SMM iterations, which is done by solving the constrained minimization problems using `fmincon` in MATLAB Optimization Toolbox. The code is provided in our supplementary material. We find that SMM converges to $(\frac{12}{11},\frac{1}{11},\frac{1}{11})$; the trajectory for 10 cycles can be seen in Figure 4.

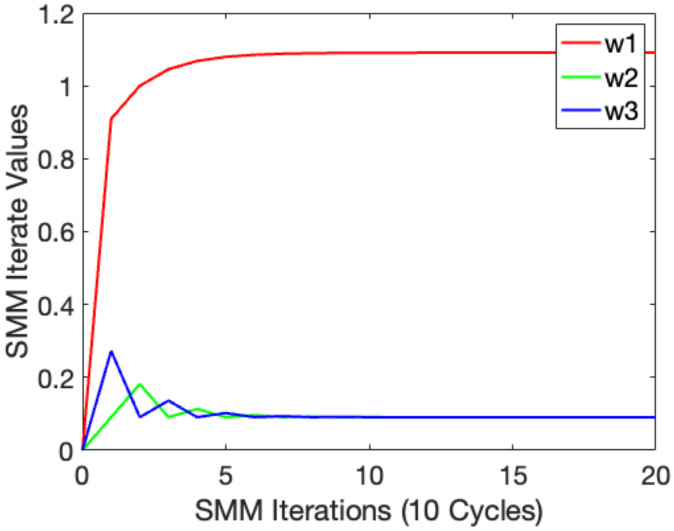


Figure 4: We run SMM iterations on the toy example by solving the projection problems using an optimization solver.

C.2 Experiment Details of Figure 2 & More Results

Here we present the experimental details of Figure 2. We also provide omitted result related to it. Then, more importantly, we extend our experimental setups beyond the cyclic task ordering and the fixed total offline dataset.

C.2.1 Experimental Detail

Data Generation. We carefully design three 2D synthetic datasets. Each dataset (of size 100) is randomly sampled from a bounded support. Below, we describe the data distribution from which we draw samples. Note that the label $y \in \{\pm 1\}$ is uniformly randomly sampled before sampling the 2D input points.

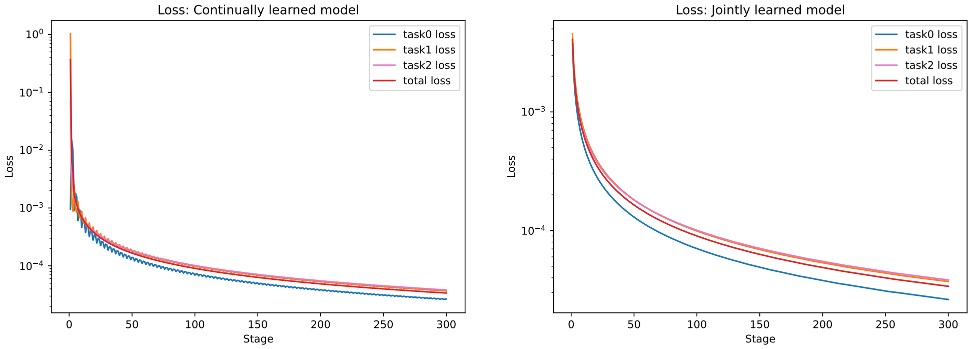
- Task 0, $\boldsymbol{x}|y = +1$: Uniform distribution on a round disk (i.e., inside of a circle) with radius 0.9 and centered at $(0.6, 4.5)$.
- Task 0, $\boldsymbol{x}|y = -1$: Uniform distribution on a rectangle $[0, 1.5] \times [-3.9, -2.7]$.
- Task 1, $\boldsymbol{x}|y = +1$: Uniform distribution on a round disk with radius 0.75 and centered at $(5.1, 0)$.
- Task 1, $\boldsymbol{x}|y = -1$: Uniform distribution on a rectangle $[-4.2, -2.1] \times [-0.9, 0.9]$.
- Task 2, $\boldsymbol{x}|y = +1$: Uniform distribution on a rectangle $[0.6, 3] \times [0.6, 2.7]$.
- Task 2, $\boldsymbol{x}|y = -1$: Uniform distribution on a disk with radius 1.2 and centered at $(-3, -2.4)$.

Among all 300 data points, we randomly choose 3 points (one for each task) and replace them by $(\boldsymbol{x} = (1.5, -2.7), y = -1)$ (for task 0), $(\boldsymbol{x} = (-2.1, 0.9), y = -1)$ (for task 1), and $(\boldsymbol{x} = (0.6, 0.6), y = +1)$ (for task 2), which are the points included in the support of the data distribution(s). These three points play the role of supporting vectors so that the joint max-margin direction becomes $\frac{\hat{\boldsymbol{w}}}{\|\hat{\boldsymbol{w}}\|} = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$, where the size of maximum margin (Equation (4)) is $\phi = 0.6\sqrt{2} > 0$ (thus, jointly separable).

Optimization. We run sequential GD for 300 stages in total. Since there are three tasks, for the cyclic ordering case, it is equivalent to $J = 100$. The step size we used is $\eta = 0.1$. Also, we allow and conduct $K = 1{,}000$ updates per stage. For the joint training case, we run full-batch GD on the union of all datasets for $MJK = 300{,}000$ steps.

C.2.2 Omitted Loss Convergence Result in Figure 2

Although we only displayed the directional convergence in the main text, we also observe the loss convergence to zero, which we proved in Theorems 3.1 and 3.3: see Figure 5. Note that we depict the loss values for a jointly trained model (with full-batch GD) every $K = 1{,}000$ gradient updates, for a fair comparison with a continually learned model (with sequential GD). It is omitted due to space limit and being relatively more obvious than directional convergence.



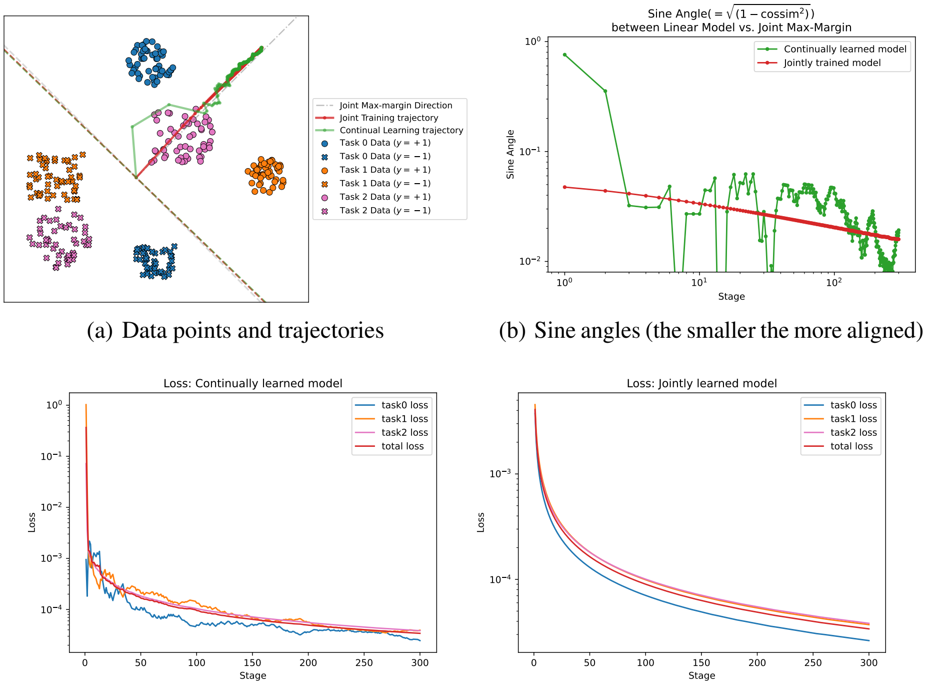
(a) Losses of continually learned model

(b) Losses of jointly trained model

Figure 5: Loss convergence results for cyclic task ordering.

C.2.3 Random Task Ordering

In Section 4, we theoretically showed that loss convergence, as well as implicit bias result, holds almost surely under the random task ordering. Indeed, we observe a similar tendency of directional convergence and loss decrease even under the random task ordering. The result is shown in Figure 6.



(a) Data points and trajectories

(b) Sine angles (the smaller the more aligned)

(c) Losses of continually learned model

(d) Losses of jointly trained model

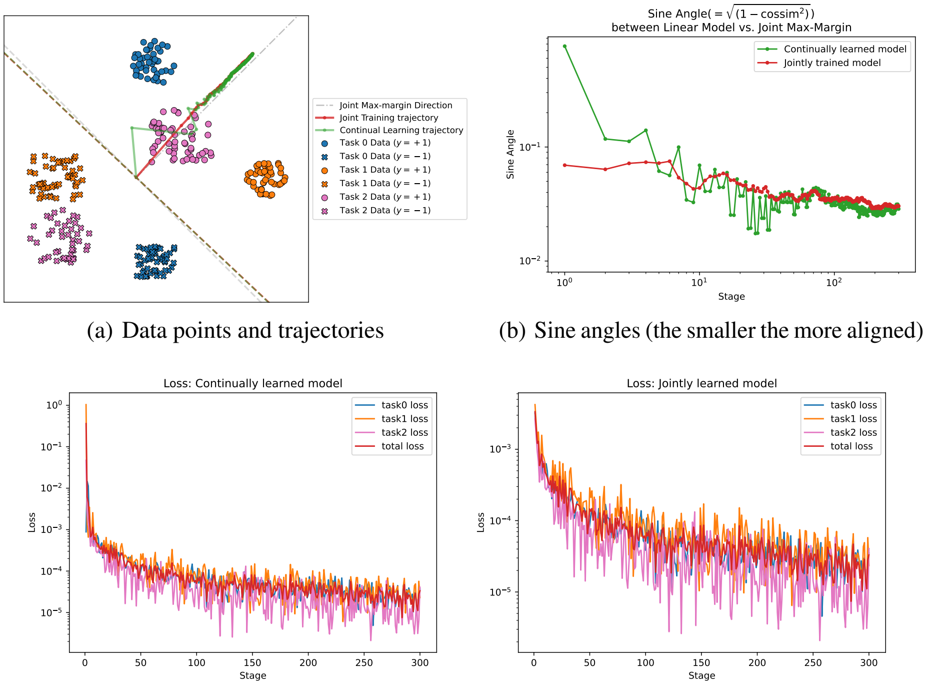
Figure 6: Experiments on 2D synthetic data under random task ordering.

C.2.4 Beyond Theoretical Setup: Towards Continual Learning on Online Data

Most theoretical analysis in this work exploits a structural assumption on the data points: there is a pre-defined set of offline dataset, which is divided into chunks and accessible one by one at each stage. Thus, exactly the same batch of data is guaranteed to be reused (surely or with high probability). Can we go beyond this repetition and apply our theoretical intuition to more general setups?

Here, we demonstrate that the results of our theoretical findings are not really limited to the task repetition setup. Instead, our insight about jointly separable continual linear classification applies to several general setups. In this section, we showcase an analogous behavior of sequential GD when the total dataset is no longer fixed throughout the continual learning process. We consider the setup where there are M different (jointly separable) data *distributions*, rather than datasets; every time we encounter a task, we have an access to a totally new samples of data points drawn from the task's distribution. For simplicity of visualization, we still stick to the bounded support cases.

An implementational difference from the previous sections is that we re-sample the data points from a predefined data distribtion at every stage. Another minor detail is that we no longer fix the three support vectors as mentioned in Appendix C.2.1: thus, at every stage, we never reuse the same data point(s) from the previous stage, almost surely. We test whether a similar trend happens even when we add the resampling process, under the same data distribution described in Appendix C.2.1. The results are shown in Figures 7 and 8 for cyclic task ordering and random task ordering cases, respectively.



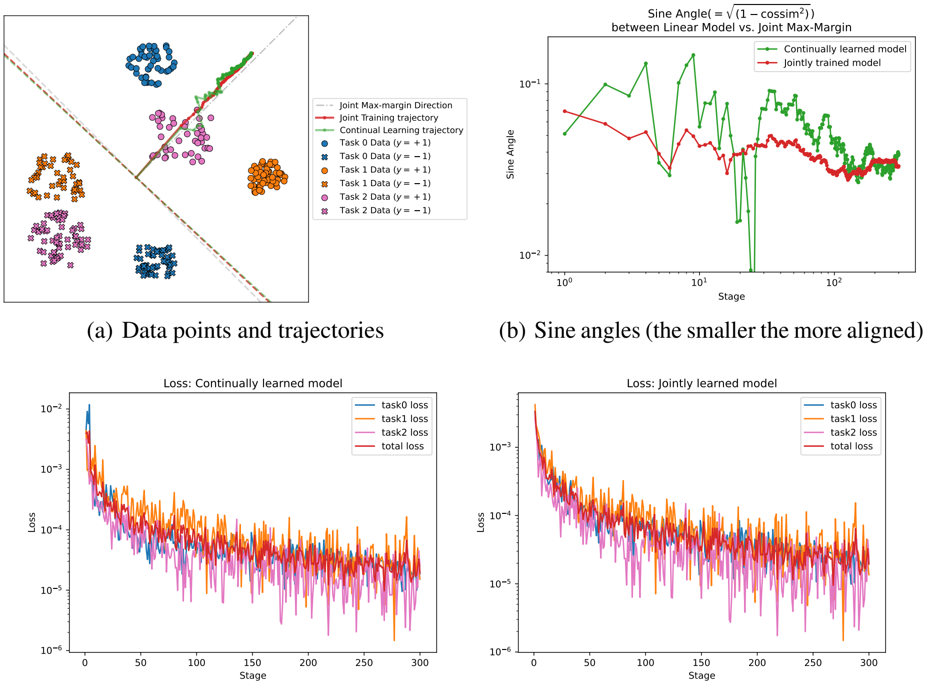
(a) Data points and trajectories

(b) Sine angles (the smaller the more aligned)

(c) Losses of continually learned model

(d) Losses of jointly trained model

Figure 7: 2D synthetic experiments: Cyclic task ordering, jointly separable online dataset (keep being drawn from a task's predefined data distribution).



(a) Data points and trajectories

(b) Sine angles (the smaller the more aligned)

(c) Losses of continually learned model

(d) Losses of jointly trained model

Figure 8: 2D synthetic experiments: Random task ordering, jointly separable online dataset (keep being drawn from a task's predefined data distribution).

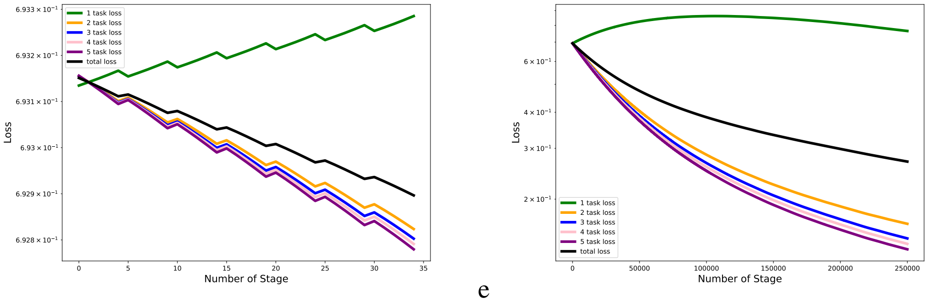
C.3 Toy Example for Increasing Loss in A Cycle

Here, we give a toy example that shows temporarily increasing joint training loss during a cycle, even with a small learning rate.

Let the datasets $\mathcal{D}_i$ $(i = 1, ..., 5)$ be as the following. Without loss of generality, we choose all labels as $+1$ without loss of generality, hence we omitted them.

$$D_1 = \{(1, -2)\},\ D_2 = \{(1, 2)\},\ D_3 = \{(1.1, 2.1)\},$$
$$D_4 = \{(1.1, 2.2)\},\ D_5 = \{(1.1, 2.3)\}.$$

In this case, the max-margin direction is $(1, 0)$, while most of the task has their individual max-margin direction around $(1, 2)$. We set $K = 10$, $\eta = 10^{-6}$ so that η satisfies the learning rate condition.



e

(a) $J = 7$

(b) $J = 50000$

Figure 9: We take average on total loss(black) for better visualization. The current task switches every 10 iterations. One cycle consists of 5 stages. Figure 9(a) shows the case where some task's loss increases within a cycle. However, it eventually decreases as Figure 9(b) shows.

When task 1 is being trained, joint training loss increases while it decreases when other tasks are being trained. This is because most of the tasks have their own max-margin direction around $(1, 2)$, dominating joint training loss.

C.4 Experiments with Neural Networks: Beyond Linear Models

We explore the possibility of extending our theoretical insight to *nonlinear* models, in particular wide two-layer ReLU networks.

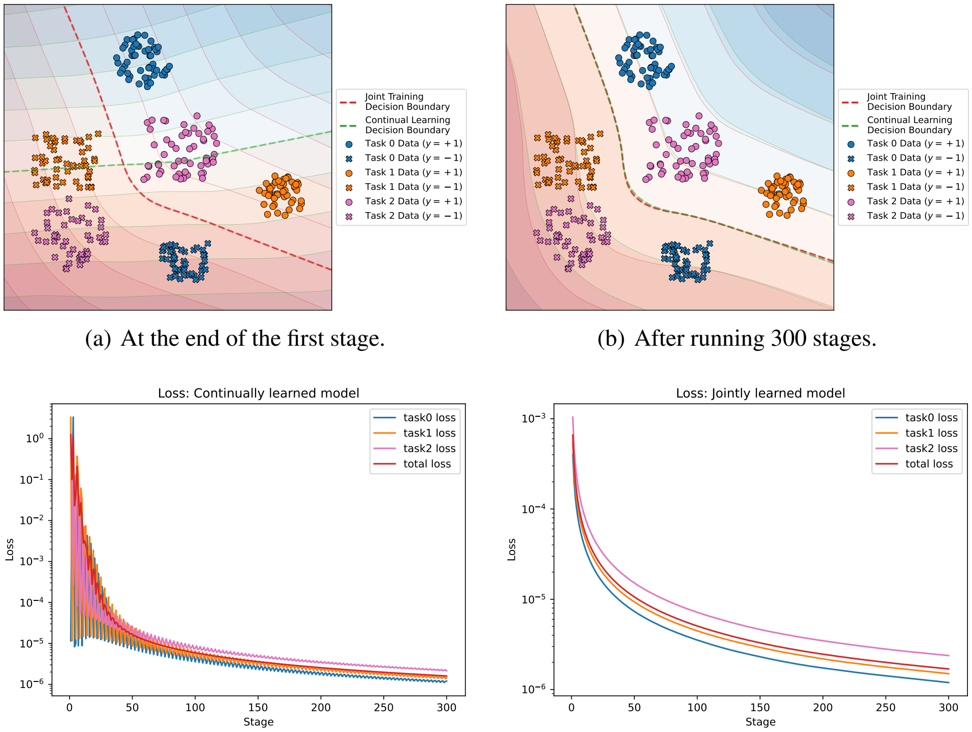
For a *linear* classifier with a single linear layer, recall that we already verified that the sequentially trained model (in cyclic/random task ordering) directionally converges to the max-margin direction. However, it is more difficult to analyze and visualize the dynamics of the multi-layer neural net's parameter values. Moreover, it might be nonsense to discuss the relationship (e.g., alignment, directional convergence) between the max-margin direction and the parameter matrices of a neural net, because the parameter matrices themselves no longer have a semantic meaning in the data space.

Instead of inspecting the parameter values, we move our attention to the *decision boundary* of the model. Observe that the decision boundary of a linear binary classifer is a hyperplane (i.e., $d - 1$ dimensional subspace) of the data space (of d-dimension), whose orthogonal complement is the span of the classifier's weight vector. Thus, the alignment between the weight vector and the max-margin direction (i.e., the implicit bias guarantee) is semantically equivalent to the alignment between the classifier's decision boundary and a hyperplane determined by the max-margin solution as a normal vector; this hyperplane can be approximated well by jointly training a single-layer linear classifier. Thus, we can still verify the similar idea of implicit bias even for a neural network by observing, not only that a continually learned model (with sequential GD, under task repitition) eventually classifies all the data points correctly, but also that the decision boundary of the continually trained model getting comparable with that of a jointly trained model (both starting from an identical initialization). Although we cannot not exactly characterize to which set of points a two-layer ReLU net's decision boundary should converge only with our theorems, it gives an effective and efficient way to confirm our findings beyond a simple linear model.

To intuitively visualize the decision boundaries, we again use the 2D synthetic datasets. Most of the experimental setting is the same as in Appendix C.2.1, except for the following three differences:

1. The classifier's architecture is a two-layer neural network $f_{\boldsymbol{\theta}} : \mathbb{R}^2 \to \mathbb{R}$ consisting of 2-dimensional input, 500 hidden ReLU neurons, and scalar output:
$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{w}_2^\top \operatorname{ReLU}(\boldsymbol{W}_1 \boldsymbol{x} + \boldsymbol{b}_1) + b_2,$$
where $\boldsymbol{\theta} = (\boldsymbol{W}_1, \boldsymbol{b}_1, \boldsymbol{w}_2, b_2) \in \mathbb{R}^{500\times 2} \times \mathbb{R}^{500} \times \mathbb{R}^{500} \times \mathbb{R}$ and $\operatorname{ReLU}(\boldsymbol{v})_i = \max\{\boldsymbol{v}_i, 0\}$.
2. To make the total dataset non-separable by a linear classifier with a positive margin but still classifiable by a neural net, we translate all datapoints with positive labels $(+1)$ by a vector $(-1.2, -1.2)$. In this case, the decision boundary should not be a straightly but bended in a curly L-shape to effecitively distinguish two classes.
3. To prevent the sequential GD from behaving similarly to a mini-batch SGD with small-scale and lazy updates, we increase K to 3,000 to guarantee that (1) the jointly trained model can correctly classify all data points within only one stage (i.e., with initial K updates), and (2) the continually learned model gets sufficiently trained on a specific task at each stage. As a result, the jointly trained model takes $MJK = 900{,}000$ iterations. ($M = 3$, $J = 100$)

As we did for a linear classifier, we classify the input data as $y = +1$ if the model output is positive and as -1 otherwise (thus, the decision boundary is a level set $\{\boldsymbol{x} \in \mathbb{R}^2 : f_{\boldsymbol{\theta}}(\boldsymbol{x}) = 0\}$). We again use the usual logistic loss $\frac{1}{N}\sum_{i=1}^{N} \ell(y_i f_{\boldsymbol{\theta}}(\boldsymbol{x}_i))$.



(a) At the end of the first stage.

(b) After running 300 stages.

(c) Losses of continually trained model.

(d) Losses of jointly trained model.

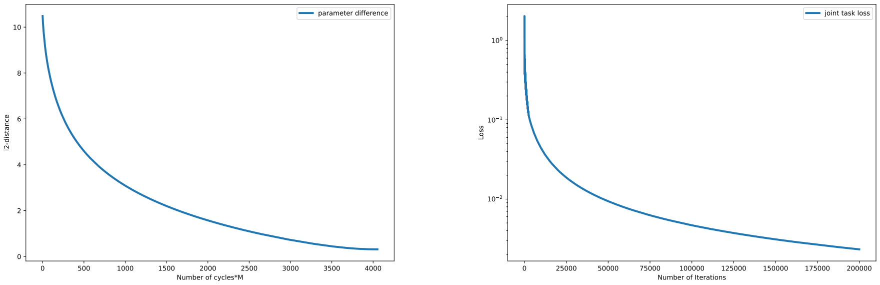
Figure 10: Two-layer ReLU network experiment under cyclic task ordering. **(Top.)** Each subfigure displays the decision boundaries (and other auxiliary level sets) of a jointly trained model (dashed red line) and a continually trained model (dashed green line). **(Bottom.)** Figure 10(c) demonstrates the large amounts of forgetting at initial few cycles and convergences of loss and (cycle-averaged) forgetting to near zero. On the other hand, Figure 10(d) shows that the training loss of the jointly trained model is already small (e.g., less than 10^{-3}) at initial stages.

The result of experiment for cyclic task order is visualized in Figure 10, exhibiting decision boundaries of a jointly trained model and a continually trained model (with sequential GD). As we expected, the updates are aggressive enough so that even a single stage (i.e., initial $K = 3{,}000$ iterations) is sufficient to perfectly classify the total dataset with the jointly trained model, and the same for the dataset of the task 0 with the continually trained model (Figure 10(a)). After some number of stages (Figure 10(b)), the both models not only correctly classify every data points, but also have an almost identical decision boundary (note that the other level sets are not necessarily the same), implying that a similar phenomenon like implicit bias is happening here. We also observe almost the same tendencies under random task ordering and even for non-repeating dataset cases (Appendix C.2.4). We omit their detailed results from the paper, but one can find them in our supplementary materials.

C.5 Experiment on a Real-world Dataset

In this section, we present a result of training linear model with CIFAR-10 (Krizhevsky et al., 2009).

We choose two classes from the CIFAR-10 dataset and design 3 tasks which have 512 data points from the two classes ('airplane', 'automobile'). Our Theorem 5.2 on linearly non-separable data like CIFAR-10 shows that sequential GD iterates should not diverge and instead converge to the global minimum $\boldsymbol{w}^*$ under the properly chosen learning rate. To estimate the distance between sequential GD iterates and the global minimum, we first train a linear model using joint task data and obtain $\boldsymbol{w}_{\mathrm{Joint}}$ as a proxy of $\boldsymbol{w}^*$; we do this because offline training is guaranteed to converge to the global minimum. Then, we train sequential GD and measure the distance between iterates and the jointly trained solution $\boldsymbol{w}_{\mathrm{Joint}}$ at the end of every stage of sequential GD.



(a) $\left\|\boldsymbol{w}_K^{(t)} - \boldsymbol{w}_{\mathrm{Joint}}\right\|$

(b) Loss of jointly trained model

Figure 11: **CIFAR-10 Experiments with linear model.** We jointly train a model for 200000 iterations to achieve the global minimum. We then train each task with cyclic ordering. We set the number of GD for each stage as 50 ($K = 50$), and run 1350 cycles ($J = 1350$). Figure 11(a) shows that sequential GD iterate converges close to $\boldsymbol{w}_{\mathrm{Joint}}$ as the training goes on. However, it does not fully converge to $\boldsymbol{w}_{\mathrm{Joint}}$, as $\boldsymbol{w}_{\mathrm{Joint}}$ is not equal to $\boldsymbol{w}^*$. Figure 11(b) reveals that the loss of the jointly trained model was decreasing after 200000 iterations.

As a result, we observe that the distance between sequential GD iterates and $\boldsymbol{w}_{\mathrm{Joint}}$ converges close to 0, even when we adopt a learning rate $\eta = 0.01$, which is not as too small as our theorem requires. Yet, we couldn't show convergence of distance to exactly 0 since the jointly trained model did not converge all the way to $\boldsymbol{w}^*$.

D Proofs for Section 3: Cyclic Task Ordering, Jointly Separable

Without loss of generality, we set $y_i = 1$ for all $i \in [N]$.

D.1 Asymptotic Loss Convergence Analysis (Proof of Theorem 3.1)

Let us restate the theorem here for the sake of readability.

Theorem 3.1. *Let $\{\boldsymbol{w}_k^{(t)}\}_{k\in[0:K-1],t\geq 0}$ be the sequence of GD iterates (2) from any starting point $\boldsymbol{w}_0^{(0)}$, where tasks are given cyclically. Under Assumptions 3.1 and 3.3, if the learning rate satisfies $\eta < \min\left\{\frac{1}{2MK\beta\sigma_{\max}^2}, \frac{\phi^2}{4K\beta\sigma_{\max}^3(M\phi+\sigma_{\max})}\right\}$, then*

1. *Loss converges to zero: $\lim_{t\to\infty}\mathcal{L}(\boldsymbol{w}_k^{(t)}) = 0, \forall k \in [0:K-1]$.*
2. *Every data point is eventually classified correctly: $\lim_{t\to\infty}\boldsymbol{x}_i^\top \boldsymbol{w}_k^{(t)} = \infty, \forall k \in [0:K-1], i \in I$.*
3. *Square sum of the change of weight is finite: $\sum_{t=0}^{\infty}\sum_{k=0}^{K-1}\|\boldsymbol{w}_{k+1}^{(t)} - \boldsymbol{w}_k^{(t)}\|^2 < \infty$.*

Here, we use the following lemma which holds in cyclic continual learning with M tasks.

Lemma D.1. *For all $t \in \mathbb{N}, m \in [0:M-1], k \in [0:K-1]$,*

$$\left\|\boldsymbol{w}_k^{(t+m)} - \boldsymbol{w}_0^{(t)} + \eta\left(K\sum_{i=0}^{m-1}\nabla\mathcal{L}^{(t+i)}(\boldsymbol{w}_0^{(t)}) + k\nabla\mathcal{L}^{(t+m)}(\boldsymbol{w}_0^{(t)})\right)\right\| \leq \frac{\eta^2(mK+k)K\sigma_{\max}^3\beta}{\phi\{1-\eta(mK+k)\sigma_{\max}^2\beta\}}\left\|\nabla\mathcal{L}(\boldsymbol{w}_0^{(t)})\right\|,$$

$$\left\|\boldsymbol{w}_k^{(t+m)} - \boldsymbol{w}_0^{(t)}\right\| \leq \frac{\eta K\sigma_{\max}}{\phi\{1-\eta(mK+k)\sigma_{\max}^2\beta\}}\left\|\nabla\mathcal{L}(\boldsymbol{w}_0^{(t)})\right\|,$$

$$\left\|\nabla\mathcal{L}(\boldsymbol{w}_k^{(t+m)}) - \nabla\mathcal{L}(\boldsymbol{w}_0^{(t)})\right\| \leq \frac{\eta K\sigma_{\max}^3\beta}{\phi\{1-\eta(mK+k)\sigma_{\max}^2\beta\}}\left\|\nabla\mathcal{L}(\boldsymbol{w}_0^{(t)})\right\|.$$

Proof. See Appendix D.1.1. □

Also, we rely on the key property of linearly separable data, which is proposed by Nacson et al. (2019).

Lemma D.2. *For any $\boldsymbol{w} \in \mathbb{R}^d$,*

$$\|\nabla\mathcal{L}(\boldsymbol{w})\| \geq \phi\sqrt{\sum_{i\in I}\left[\ell'(\boldsymbol{x}_i^\top\boldsymbol{w})\right]^2}$$

Proof. See Appendix D.1.2. □

Since $\mathcal{L}$ is a $\sigma_{\max}^2\beta$-smooth function, we get

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{w}_0^{(Mt+M)}) - \mathcal{L}(\boldsymbol{w}_0^{(Mt)}) - \frac{\sigma_{\max}^2\beta}{2}\left\|\boldsymbol{w}_0^{(Mt+M)} - \boldsymbol{w}_0^{(Mt)}\right\|^2 \\
&\leq \nabla\mathcal{L}(\boldsymbol{w}_0^{(Mt)})^\top(\boldsymbol{w}_0^{(Mt+M)} - \boldsymbol{w}_0^{(Mt)}) \\
&= \nabla\mathcal{L}(\boldsymbol{w}_0^{(Mt)})^\top(\boldsymbol{w}_0^{(Mt+M)} - \boldsymbol{w}_0^{(Mt)} - \eta K\nabla\mathcal{L}(\boldsymbol{w}_0^{(Mt)}) + \eta K\nabla\mathcal{L}(\boldsymbol{w}_0^{(Mt)})) \\
&\leq -\eta K\left\|\nabla\mathcal{L}(\boldsymbol{w}_0^{(Mt)})\right\|^2 + \left\|\nabla\mathcal{L}(\boldsymbol{w}_0^{(Mt)})\right\|\left\|\boldsymbol{w}_0^{(Mt+M)} - \boldsymbol{w}_0^{(Mt)} + \eta K\nabla\mathcal{L}(\boldsymbol{w}_0^{(Mt)})\right\|
\end{aligned}$$

By Lemma D.1,

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{w}_0^{(Mt+M)}) - \mathcal{L}(\boldsymbol{w}_0^{(Mt)}) - \frac{\sigma_{\max}^2\beta}{2}\cdot\frac{(\eta\sigma_{\max}K)^2}{\phi^2(1-\eta MK\sigma_{\max}^2\beta)^2}\left\|\nabla\mathcal{L}(\boldsymbol{w}_0^{(Mt)})\right\|^2 \\
&\leq -\eta K\left\|\nabla\mathcal{L}(\boldsymbol{w}_0^{(Mt)})\right\|^2 + \frac{\eta^2MK^2\sigma_{\max}^3\beta}{\phi(1-\eta MK\sigma_{\max}^2\beta)}\left\|\nabla\mathcal{L}(\boldsymbol{w}_0^{(Mt)})\right\|^2
\end{aligned}$$

Given that $\eta \leq \frac{1}{2MK\sigma_{\max}^2\beta}$,

$$
\begin{aligned}
&\mathcal{L}(\boldsymbol{w}_0^{(Mt+M)}) - \mathcal{L}(\boldsymbol{w}_0^{(Mt)}) && (12)\\
&\leq \eta K\{1 - \eta K \left(\frac{M\sigma_{\max}^3\beta}{\phi(1-\eta MK\sigma_{\max}^2\beta)} + \frac{\sigma_{\max}^4\beta}{2\phi^2(1-\eta MK\sigma_{\max}^2\beta)^2} \right)\} \left\| \nabla\mathcal{L}(\boldsymbol{w}_0^{(Mt)}) \right\|^2 && (13)\\
&\leq -\eta K \left(1 - \eta K \frac{2(M\phi + \sigma_{\max})\sigma_{\max}^3\beta}{\phi^2} \right) \left\| \nabla\mathcal{L}(\boldsymbol{w}_0^{(Mt)}) \right\|^2 && (14)\\
&= -\eta K \left(1 - \eta K \beta' \right) \left\| \nabla\mathcal{L}(\boldsymbol{w}_0^{(Mt)}) \right\|^2, && (15)
\end{aligned}
$$

where we set $\beta' := \frac{2(M\phi+\sigma_{\max})\sigma_{\max}^3\beta}{\phi^2}$. Given that $\eta \leq \frac{1}{2K\beta'}$, $\mathcal{L}(\boldsymbol{w}_0^{(Mt+M)}) \leq \mathcal{L}(\boldsymbol{w}_0^{(Mt)})$ holds. Also, by (15),

$$
\sum_{t=0}^{\infty} \left\| \nabla\mathcal{L}(\boldsymbol{w}_0^{(Mt)}) \right\|^2 \leq \frac{\mathcal{L}(\boldsymbol{w}_0^{(0)}) - \lim_{t\to\infty}\mathcal{L}(\boldsymbol{w}_0^{(Mt)})}{\eta K(1-\eta K\beta')} \leq \frac{\mathcal{L}(\boldsymbol{w}_0^{(0)})}{\eta K(1-\eta K\beta')} < \infty
$$

Coupled with Lemma D.1,

$$
\begin{aligned}
&\sum_{t=0}^{\infty}\sum_{m=0}^{M-1}\sum_{k=0}^{K-1} \left\| \nabla\mathcal{L}(\boldsymbol{w}_k^{(Mt+m)}) \right\|^2 \\
&\qquad \leq \sum_{t=0}^{\infty}\sum_{m=0}^{M-1}\sum_{k=0}^{K-1} \left(\left\| \nabla\mathcal{L}(\boldsymbol{w}_0^{(Mt)}) \right\| + \left\| \nabla\mathcal{L}(\boldsymbol{w}_k^{(Mt+m)}) - \nabla\mathcal{L}(\boldsymbol{w}_0^{(Mt)}) \right\| \right)^2 \\
&\qquad \leq \sum_{t=0}^{\infty}\sum_{m=0}^{M-1}\sum_{k=0}^{K-1} \left(1 + \frac{\eta K\sigma_{\max}^3\beta}{\phi\{1-\eta(mK+k)\sigma_{\max}^2\beta\}} \right)^2 \left\| \nabla\mathcal{L}(\boldsymbol{w}_0^{(Mt)}) \right\|^2 \\
&\qquad \leq \left(1 + \frac{\eta K\sigma_{\max}^3\beta}{\phi\{1-\eta MK\sigma_{\max}^2\beta\}} \right)^2 MK \sum_{t=0}^{\infty} \left\| \nabla\mathcal{L}(\boldsymbol{w}_0^{(Mt)}) \right\|^2 < \infty
\end{aligned}
$$

The boundedness of infinite sum of nonzero elements means $\lim_{t\to\infty} \left\| \nabla\mathcal{L}(\boldsymbol{w}_k^{(t)}) \right\|^2 = 0, \forall k \in [0 : K-1]$. This leads to $\lim_{t\to\infty} \ell'(x_i^\top \boldsymbol{w}_k^{(t)}) = 0, \forall i \in I, k \in [0 : K-1]$ by Lemma D.2. Since $\ell'(u) \to 0$ only when $u \to \infty$, we obtain $x_i^\top \boldsymbol{w}_k^{(t)} \to \infty, \forall i \in I, k \in [0 : K-1]$ and $\lim_{t\to\infty} \mathcal{L}(\boldsymbol{w}_k^{(t)}) = 0, \forall k \in [0 : K-1]$. Finally, we obtain that $\sum_{t=0}^{\infty}\sum_{k=0}^{K-1} \left\| \boldsymbol{w}_{k+1}^{(t)} - \boldsymbol{w}_k^{(t)} \right\|^2 < \infty$ followed by

$$
\begin{aligned}
\left\| \nabla\mathcal{L}(\boldsymbol{w}_k^{(t)}) \right\| &\geq \phi \sqrt{\sum_{i\in I} \left[\ell'(\boldsymbol{x}_i^\top \boldsymbol{w}_k^{(t)}) \right]^2} \geq \phi \sqrt{\sum_{i\in I(t)} \left[\ell'(\boldsymbol{x}_i^\top \boldsymbol{w}_k^{(t)}) \right]^2} \\
&\geq \frac{\phi}{\sigma_{\max}} \left\| \sum_{i\in I(t)} \ell'(\boldsymbol{x}_i^\top \boldsymbol{w}_k^{(t)}) x_i \right\| = \frac{\phi}{\sigma_{\max}} \eta^{-1} \left\| \boldsymbol{w}_{k+1}^{(t)} - \boldsymbol{w}_k^{(t)} \right\|,
\end{aligned}
$$

where in the first inequality, we use Lemma D.2 and in the third ineqaultiy, we use the fact $\forall \lambda_s \in \mathbb{R} : \left\| \sum_{s\in I} \lambda_s \boldsymbol{x}_s \right\|_2 \leq \sigma_{\max} \sqrt{\sum_{s\in I} \lambda_s^2}$. The last equality is true by the definition of gradient descent.

D.1.1 Proof of Lemma D.1

For all $t \in \mathbb{N}, m \in [0 : M-1], k \in [0 : K-1]$

$$
\begin{aligned}
&\left\| \boldsymbol{w}_k^{(t+m)} - \boldsymbol{w}_0^{(t)} + \eta \left(K \sum_{i=0}^{m-1} \nabla\mathcal{L}^{(t+i)}(\boldsymbol{w}_0^{(t)}) + k\nabla\mathcal{L}^{(t+m)}(\boldsymbol{w}_0^{(t)}) \right) \right\| \\
&= \left\| \eta \sum_{i=0}^{m-1}\sum_{j=0}^{K-1} \left(\nabla\mathcal{L}^{(t+i)}(\boldsymbol{w}_0^{(t)}) - \nabla\mathcal{L}^{(t+i)}(\boldsymbol{w}_j^{(t+i)}) \right) + \eta \sum_{j=0}^{k-1} \left(\nabla\mathcal{L}^{(t+m)}(\boldsymbol{w}_0^{(t)}) - \nabla\mathcal{L}^{(t+m)}(\boldsymbol{w}_j^{(t+m)}) \right) \right\|
\end{aligned}
$$

$$\begin{aligned}
&= \left\| \eta \sum_{i=0}^{m-1}\sum_{j=0}^{K-1}\sum_{s\in I^{(t+i)}} \left(\ell'(\boldsymbol{x}_s^\top \boldsymbol{w}_0^{(t)}) - \ell'(\boldsymbol{x}_s^\top \boldsymbol{w}_j^{(t+i)})\right)\boldsymbol{x}_s + \eta\sum_{j=0}^{k-1}\sum_{s\in I^{(t+m)}}\left(\ell'(\boldsymbol{x}_s^\top \boldsymbol{w}_0^{(t)}) - \ell'(\boldsymbol{x}_s^\top \boldsymbol{w}_j^{(t+m)})\right)\boldsymbol{x}_s \right\| \\
&\le \eta \sum_{i=0}^{m-1}\sum_{j=0}^{K-1}\left\|\sum_{s\in I^{(t+i)}} \left(\ell'(\boldsymbol{x}_s^\top \boldsymbol{w}_0^{(t)}) - \ell'(\boldsymbol{x}_s^\top \boldsymbol{w}_j^{(t+i)})\right)\boldsymbol{x}_s\right\| + \eta\sum_{j=0}^{k-1}\left\|\sum_{s\in I^{(t+m)}}\left(\ell'(\boldsymbol{x}_s^\top \boldsymbol{w}_0^{(t)}) - \ell'(\boldsymbol{x}_s^\top \boldsymbol{w}_j^{(t+m)})\right)\boldsymbol{x}_s\right\|
\end{aligned}$$

holds by triangle inequality. Then

$$\begin{aligned}
&\eta \sum_{i=0}^{m-1}\sum_{j=0}^{K-1}\left\|\sum_{s\in I^{(t+i)}} \left(\ell'(\boldsymbol{x}_s^\top \boldsymbol{w}_0^{(t)}) - \ell'(\boldsymbol{x}_s^\top \boldsymbol{w}_j^{(t+i)})\right)\boldsymbol{x}_s\right\| + \eta\sum_{j=0}^{k-1}\left\|\sum_{s\in I^{(t+m)}}\left(\ell'(\boldsymbol{x}_s^\top \boldsymbol{w}_0^{(t)}) - \ell'(\boldsymbol{x}_s^\top \boldsymbol{w}_j^{(t+m)})\right)\boldsymbol{x}_s\right\| \\
&\le \eta\sigma_{\max} \sum_{i=0}^{m-1}\sum_{j=0}^{K-1}\sqrt{\sum_{s\in I^{(t+i)}} \left(\ell'(\boldsymbol{x}_s^\top \boldsymbol{w}_0^{(t)}) - \ell'(\boldsymbol{x}_s^\top \boldsymbol{w}_j^{(t+i)})\right)^2} + \eta\sigma_{\max}\sum_{j=0}^{k-1}\sqrt{\sum_{s\in I^{(t+m)}}\left(\ell'(\boldsymbol{x}_s^\top \boldsymbol{w}_0^{(t)}) - \ell'(\boldsymbol{x}_s^\top \boldsymbol{w}_j^{(t+m)})\right)^2} \\
&\le \eta\sigma_{\max}\beta \sum_{i=0}^{m-1}\sum_{j=0}^{K-1}\sqrt{\sum_{s\in I^{(t+i)}} \left[\boldsymbol{x}_s^\top \left(\boldsymbol{w}_0^{(t)} - \boldsymbol{w}_j^{(t+i)}\right)\right]^2} + \eta\sigma_{\max}\beta\sum_{j=0}^{k-1}\sqrt{\sum_{s\in I^{(t+m)}}\left[\boldsymbol{x}_s^\top \left(\boldsymbol{w}_0^{(t)} - \boldsymbol{w}_j^{(t+m)}\right)\right]^2} \\
&\le \eta\sigma_{\max}^2\beta \sum_{i=0}^{m-1}\sum_{j=0}^{K-1}\left\|\boldsymbol{w}_j^{(t+i)} - \boldsymbol{w}_0^{(t)}\right\| + \eta\sigma_{\max}^2\beta\sum_{j=0}^{k-1}\left\|\boldsymbol{w}_j^{(t+m)} - \boldsymbol{w}_0^{(t)}\right\| \qquad (16)
\end{aligned}$$

The first inequality comes from the fact $\forall\lambda_s \in \mathbb{R} : \left\|\sum_{s\in I}\lambda_s\boldsymbol{x}_s\right\|_2 \le \sigma_{\max}\sqrt{\sum_{s\in I}\lambda_s^2}$. The next one comes from β-smoothness, and the last inequality holds since $\forall\boldsymbol{v} \in \mathbb{R}^d : \sum_{s\in I}(\boldsymbol{x}_s^\top\boldsymbol{v})^2 \le \sigma_{\max}^2\left\|\boldsymbol{v}\right\|^2$. Then we get

$$\begin{aligned}
&\left\|\boldsymbol{w}_k^{(t+m)} - \boldsymbol{w}_0^{(t)}\right\| \\
&\le \left\|-\eta\left(K\sum_{i=0}^{m-1}\nabla\mathcal{L}^{(t+i)}(\boldsymbol{w}_0^{(t)}) + k\nabla\mathcal{L}^{(t+m)}(\boldsymbol{w}_0^{(t)})\right)\right\| \\
&+ \left\|\boldsymbol{w}_k^{(t+m)} - \boldsymbol{w}_0^{(t)} + \eta\left(K\sum_{i=0}^{m-1}\nabla\mathcal{L}^{(t+i)}(\boldsymbol{w}_0^{(t)}) + k\nabla\mathcal{L}^{(t+m)}(\boldsymbol{w}_0^{(t)})\right)\right\| \\
&\le \eta\left\|K\sum_{i=0}^{m-1}\sum_{s\in I^{(t+i)}}\ell'(\boldsymbol{x}_s^\top\boldsymbol{w}_0^{(t)})\boldsymbol{x}_s + k\sum_{s\in I^{(t+m)}}\ell'(\boldsymbol{x}_s^\top\boldsymbol{w}_0^{(t)})\boldsymbol{x}_s\right\| \\
&+ \left\|\boldsymbol{w}_k^{(t+m)} - \boldsymbol{w}_0^{(t)} + \eta\left(K\sum_{i=0}^{m-1}\nabla\mathcal{L}^{(t+i)}(\boldsymbol{w}_0^{(t)}) + k\nabla\mathcal{L}^{(t+m)}(\boldsymbol{w}_0^{(t)})\right)\right\| \\
&\le \eta\sigma_{\max}\sqrt{\sum_{i=0}^{m-1}\sum_{s\in I^{(t+i)}}\left(K\ell'(\boldsymbol{x}_s^\top\boldsymbol{w}_0^{(t)})\right)^2 + \sum_{s\in I^{(t+m)}}\left(k\ell'(\boldsymbol{x}_s^\top\boldsymbol{w}_0^{(t)})\right)^2} \\
&+ \left\|\boldsymbol{w}_k^{(t+m)} - \boldsymbol{w}_0^{(t)} + \eta\left(K\sum_{i=0}^{m-1}\nabla\mathcal{L}^{(t+i)}(\boldsymbol{w}_0^{(t)}) + k\nabla\mathcal{L}^{(t+m)}(\boldsymbol{w}_0^{(t)})\right)\right\| \\
&\le \eta K\sigma_{\max}\sqrt{\sum_{s\in I}\left(\ell'(\boldsymbol{x}_s^\top\boldsymbol{w}_0^{(t)})\right)^2} + \left\|\boldsymbol{w}_k^{(t+m)} - \boldsymbol{w}_0^{(t)} + \eta\left(K\sum_{i=0}^{m-1}\nabla\mathcal{L}^{(t+i)}(\boldsymbol{w}_0^{(t)}) + k\nabla\mathcal{L}^{(t+m)}(\boldsymbol{w}_0^{(t)})\right)\right\|
\end{aligned}$$

Then by (16) and Lemma D.2, we obtain

$$\begin{aligned}
&\left\|\boldsymbol{w}_k^{(t+m)} - \boldsymbol{w}_0^{(t)}\right\| \\
&\le \frac{\eta K\sigma_{\max}}{\phi}\left\|\nabla\mathcal{L}(\boldsymbol{w}_0^{(t)})\right\| + \eta\sigma_{\max}^2\beta\left(\sum_{i=0}^{m-1}\sum_{j=0}^{K-1}\left\|\boldsymbol{w}_j^{(t+i)} - \boldsymbol{w}_0^{(t)}\right\| + \sum_{j=0}^{k-1}\left\|\boldsymbol{w}_j^{(t+m)} - \boldsymbol{w}_0^{(t)}\right\|\right)
\end{aligned} \qquad (17)$$

Here, we use a lemma in Nacson et al. (2019).

Lemma D.3 (Nacson et al. (2019)). *For some ϵ and θ, let $\delta_k \leq \theta + \epsilon \sum_{u=0}^{k-1} \delta_u$ holds for all k. Then*

$$\delta_k \leq \frac{\theta}{1-k\epsilon}$$

and

$$\sum_{u=0}^{k-1} \delta_u \leq \frac{k\theta}{1-k\epsilon}$$

By applying the lemma to (17), we obtain

$$\left\| \boldsymbol{w}_k^{(t+m)} - \boldsymbol{w}_0^{(t)} \right\| \leq \frac{\eta K \sigma_{\max}}{\phi\{1-\eta(mK+k)\sigma_{\max}^2\beta\}} \left\| \nabla \mathcal{L}(\boldsymbol{w}_0^{(t)}) \right\|$$

and

$$\begin{aligned}
&\left\| \boldsymbol{w}_k^{(t+m)} - \boldsymbol{w}_0^{(t)} + \eta \left(K \sum_{i=0}^{m-1} \nabla \mathcal{L}^{(t+i)}(\boldsymbol{w}_0^{(t)}) + k \nabla \mathcal{L}^{(t+m)}(\boldsymbol{w}_0^{(t)}) \right) \right\| \\
&\leq \eta \sigma_{\max}^2 \beta \left(\sum_{i=0}^{m-1} \sum_{j=0}^{K-1} \left\| \boldsymbol{w}_j^{(t+i)} - \boldsymbol{w}_0^{(t)} \right\| + \sum_{j=0}^{k-1} \left\| \boldsymbol{w}_j^{(t+m)} - \boldsymbol{w}_0^{(t)} \right\| \right) \\
&\leq \frac{\eta^2 (mK+k) K \sigma_{\max}^3 \beta}{\phi\{1-\eta(mK+k)\sigma_{\max}^2\beta\}} \left\| \nabla \mathcal{L}(\boldsymbol{w}_0^{(t)}) \right\|.
\end{aligned}$$

Finally,

$$\begin{aligned}
\left\| \nabla \mathcal{L}(\boldsymbol{w}_k^{(t+m)}) - \nabla \mathcal{L}(\boldsymbol{w}_0^{(t)}) \right\| &\leq \sigma_{\max}^2 \beta \left\| \boldsymbol{w}_k^{(t+m)} - \boldsymbol{w}_0^{(t)} \right\| \\
&\leq \frac{\eta K \sigma_{\max}^3 \beta}{\phi\{1-\eta(mK+k)\sigma_{\max}^2\beta\}} \left\| \nabla \mathcal{L}(\boldsymbol{w}_0^{(t)}) \right\|
\end{aligned}$$

D.1.2 Proof of Lemma D.2

For all $\boldsymbol{w} \in \mathbb{R}^d$,

$$\begin{aligned}
\|\nabla \mathcal{L}(\boldsymbol{w})\| &= \left\| \sum_{i \in I} \ell'(\boldsymbol{x}_i^\top \boldsymbol{w}) \boldsymbol{x}_i \right\| \\
&\geq \sqrt{\sum_{i \in I} \left[\ell'(\boldsymbol{x}_i^\top \boldsymbol{w}) \right]^2} \cdot \min_{\boldsymbol{v} \in \mathbb{R}_{\geq 0}^N : \|\boldsymbol{v}\|=1} \|X\boldsymbol{v}\|
\end{aligned}$$

Let $\hat{\boldsymbol{v}} := \arg\min_{\boldsymbol{v} \in \mathbb{R}_{\geq 0}^N : \|\boldsymbol{v}\|=1} \|X\boldsymbol{v}\|$. Then for max-margin direction $\hat{\boldsymbol{w}}$, the following holds.

$$\|X\hat{\boldsymbol{v}}\| \geq \left\| \frac{\hat{\boldsymbol{w}}}{\|\hat{\boldsymbol{w}}\|}^\top X \hat{\boldsymbol{v}} \right\| \geq \phi \|\hat{\boldsymbol{v}}\| = \phi$$

We used Cauchy-Schwarz for the first inequality, and the definition of $\hat{\boldsymbol{w}}$ for the second one.

D.2 Directional Convergence Analysis (Proof of Theorem 3.2)

In this section, we prove Theorem 3.2 and further discuss the convergence of $\boldsymbol{\rho}_k^{(t)}$ beyond boundedness.

Theorem 3.2. *Let $\{\boldsymbol{w}_k^{(t)}\}_{k \in [0:K-1], t \geq 0}$ be the sequence of GD iterates (2) from any starting point $\boldsymbol{w}_0^{(0)}$, where tasks are given cyclically. Under Assumptions 3.1, 3.2, 3.3, and 3.4, if the learning rate satisfies $\eta < \min\left\{ \frac{1}{2MK\beta\sigma_{\max}^2}, \frac{\phi^2}{4K\beta\sigma_{\max}^3(M\phi+\sigma_{\max})} \right\}$, then $\boldsymbol{w}_k^{(t)}$ will behave as:*

$$\boldsymbol{w}_k^{(t)} = \ln\left(\tfrac{K}{M}t\right) \hat{\boldsymbol{w}} + \boldsymbol{\rho}_k^{(t)},$$

where $\|\boldsymbol{\rho}_k^{(t)}\|$ stays bounded as t grows.

Note that we use Assumption 3.2, the unique existence of SVM dual variables $\boldsymbol{\alpha}$ that satisfies

$$\hat{\boldsymbol{w}} = \sum_{s\in S} \alpha_s \boldsymbol{x}_s$$

$$\forall s \in S : \alpha_s > 0, \forall s \notin S : \alpha_s = 0$$

This assumption holds for almost all data (Soudry et al., 2018).

When the tasks are given in a cyclic order, the following lemma holds. Note that the lemma does not depend on the algorithm.

Lemma D.4. *When tasks are given cyclic, there exists $\breve{\boldsymbol{w}}, m_1(t,k) \in \mathbb{R}^d$ the following holds for all $t \in \mathbb{N}$, $k \in [0 : K-1]$.*

$$K \sum_{u=1}^{t-1} \frac{1}{u} \sum_{s\in S^{(u)}} \alpha_s \boldsymbol{x}_s + \frac{k}{t} \sum_{s\in S^{(t)}} \alpha_s \boldsymbol{x}_s = \frac{K}{M} \log(\frac{t}{M}) \hat{\boldsymbol{w}} + \frac{K}{M} \breve{\boldsymbol{w}} + m_1(t,k)$$

$$m_1(t,K) := m_1(t+1, 0)$$

such that $\|m_1(t,k)\| = o(t^{-0.5+\epsilon})$, and $\|m_1(t,k+1) - m_1(t,k)\| = O(t^{-1})$ for all $k \in [0 : K-1], \epsilon > 0$, and $\breve{\boldsymbol{w}}$ only depends on the order of tasks and constant with respect to t.

Proof. See Appendix D.2.1. □

We set $m_1(t,k)$ and $\breve{\boldsymbol{w}}$ along Lemma D.4, and define $\boldsymbol{\rho}_k^{(t)}$ and $\boldsymbol{r}_k^{(t)}$ as

$$\begin{aligned}\forall k \in [0 : K-1] : \boldsymbol{w}_k^{(t)} &= \log(\frac{K}{M} t) \hat{\boldsymbol{w}} + \boldsymbol{\rho}_k^{(t)} \\ &= \log(\frac{K}{M} t) \hat{\boldsymbol{w}} + \tilde{\boldsymbol{w}} + \frac{M}{K} m_1(t,k) + \boldsymbol{r}_k^{(t)},\end{aligned}$$

$$\boldsymbol{\rho}_K^{(t)} = \boldsymbol{\rho}_0^{(t+1)}, \boldsymbol{r}_K^{(t)} = \boldsymbol{r}_0^{(t+1)},$$

where $\tilde{\boldsymbol{w}}$ is the solution of

$$\forall i \in S : \eta \exp\left(-\boldsymbol{x}_i^\top \tilde{\boldsymbol{w}}\right) = \alpha_i, \quad \bar{P}(\tilde{\boldsymbol{w}} - \boldsymbol{w}_0^{(0)}) = 0,$$

which is unique under Assumption 3.2. Then by the definition,

$$\begin{aligned}\boldsymbol{r}_k^{(t)} &= \boldsymbol{w}_k^{(t)} - \frac{M}{K}\left(\frac{K}{M} \log(\frac{K}{M} t) \hat{\boldsymbol{w}} + m_1(t,k)\right) - \tilde{\boldsymbol{w}} \\ &= \boldsymbol{w}_k^{(t)} - \frac{M}{K}\left(K \sum_{u=1}^{t-1} \frac{1}{u} \sum_{s\in S^{(u)}} \alpha_s \boldsymbol{x}_s + \frac{k}{t} \sum_{s\in S^{(t)}} \alpha_s \boldsymbol{x}_s\right) - \log K \hat{\boldsymbol{w}} - \tilde{\boldsymbol{w}} + \breve{\boldsymbol{w}}\end{aligned}$$

Under these definitions, we can get the primary lemma of $\boldsymbol{r}_k^{(t)}$.

Lemma D.5. *Under Assumption 3.1, 3.3, 3.4, and Assumption 3.2, if learning rate is $\eta < \min\{\frac{1}{2MK\beta\sigma_{\max}^2}, \frac{\phi^2}{4K\beta\sigma_{\max}^3(M\phi+\sigma_{\max})}\}$, then*

1. *$\exists \tilde{t}, C_1, C_2 > 0$ such that $\forall t > \tilde{t}$,*

$$(\boldsymbol{r}_{k+1}^{(t)} - \boldsymbol{r}_k^{(t)})^\top \boldsymbol{r}_k^{(t)} \leq C_1 t^{-\theta} + C_2 t^{-1-0.5\tilde{\mu}}, \forall k \in [0 : K-1]$$

2. *Moreover, for all $\epsilon_1 > 0$, $\exists \tilde{t}^*, C_3 > 0$ such that if $\left\|P\boldsymbol{r}_k^{(t)}\right\| \geq \epsilon_1$ and $S^{(t)} \neq \emptyset$,*

$$(\boldsymbol{r}_{k+1}^{(t)} - \boldsymbol{r}_k^{(t)})^\top \boldsymbol{r}_k^{(t)} \leq -C_3 t^{-1}, \forall t > \tilde{t}^*, k \in [0 : K-1]$$

Proof. See Appendix D.2.2. □

By the definition of $\boldsymbol{\rho}_k^{(t)} = \tilde{\boldsymbol{w}} + \frac{M}{K} m_1(t,k) + \boldsymbol{r}_k^{(t)}$, it is enough to prove $\left\|\boldsymbol{r}_k^{(t)}\right\|$ is bounded.

$$\left\|\boldsymbol{r}_{k+1}^{(t)}\right\|^2 - \left\|\boldsymbol{r}_k^{(t)}\right\|^2 = 2(\boldsymbol{r}_{k+1}^{(t)} - \boldsymbol{r}_k^{(t)})^\top \boldsymbol{r}_k^{(t)} + \left\|\boldsymbol{r}_{k+1}^{(t)} - \boldsymbol{r}_k^{(t)}\right\|^2$$

For all $k \in [0 : K-2]$, let $\boldsymbol{a}_k^{(t)} := \frac{M}{K}(m_1(t,k+1) - m_1(t,k))$. And let $\boldsymbol{a}_{K-1}^{(t)} := \log(1 + \frac{1}{t})\hat{\boldsymbol{w}} + \frac{M}{K}(m_1(t+1,0) - m_1(t,K-1))$. Since $\boldsymbol{w}_k^{(t)} = \log(\frac{K}{M}t)\hat{\boldsymbol{w}} + \tilde{\boldsymbol{w}} + \frac{M}{K}m_1(t,k) + \boldsymbol{r}_k^{(t)}$, $\left\|\boldsymbol{r}_{k+1}^{(t)} - \boldsymbol{r}_k^{(t)}\right\|^2 = \left\|\boldsymbol{w}_{k+1}^{(t)} - \boldsymbol{w}_k^{(t)} - \boldsymbol{a}_k^{(t)}\right\|^2$. Also, by Lemma D.4, $\left\|\boldsymbol{a}_k^{(t)}\right\| = O(t^{-1})$. Thus, $\exists t_1$ such that $\forall t \geq t_1, \forall k \in [0 : K-1] : \left\|\boldsymbol{a}_k^{(t)}\right\| \leq t^{-1}$.

Now we can get the following for all $T \geq t_1$.

$$\begin{aligned}
&\sum_{t=t_1}^{T}\sum_{k=0}^{K-1}\left\|\boldsymbol{r}_{k+1}^{(t)} - \boldsymbol{r}_k^{(t)}\right\|^2 = \sum_{t=t_1}^{T}\sum_{k=0}^{K-1}\left\|\boldsymbol{w}_{k+1}^{(t)} - \boldsymbol{w}_k^{(t)} - \boldsymbol{a}_k^{(t)}\right\|^2 \\
&= \sum_{t=t_1}^{T}\sum_{k=0}^{K-1}\left\|\boldsymbol{w}_{k+1}^{(t)} - \boldsymbol{w}_k^{(t)}\right\|^2 + \sum_{t=t_1}^{T}\sum_{k=0}^{K-1} 2(\boldsymbol{w}_k^{(t)} - \boldsymbol{w}_{k+1}^{(t)})^\top \boldsymbol{a}_k^{(t)} + \sum_{t=t_1}^{T}\sum_{k=0}^{K-1}\left\|\boldsymbol{a}_k^{(t)}\right\|^2 \\
&\leq \sum_{t=t_1}^{T}\sum_{k=0}^{K-1}\left\|\boldsymbol{w}_{k+1}^{(t)} - \boldsymbol{w}_k^{(t)}\right\|^2 + 2\sqrt{\sum_{t=t_1}^{T}\sum_{k=0}^{K-1}\left\|\boldsymbol{w}_k^{(t)} - \boldsymbol{w}_{k+1}^{(t)}\right\|^2 \sum_{t=t_1}^{T}\sum_{k=0}^{K-1}\left\|\boldsymbol{a}_k^{(t)}\right\|^2} + \sum_{t=t_1}^{T}\sum_{k=0}^{K-1}\left\|\boldsymbol{a}_k^{(t)}\right\|^2 \\
&\leq \sum_{t=t_1}^{T}\sum_{k=0}^{K-1}\left\|\boldsymbol{w}_{k+1}^{(t)} - \boldsymbol{w}_k^{(t)}\right\|^2 + 2\sqrt{\sum_{t=t_1}^{T}\sum_{k=0}^{K-1}\left\|\boldsymbol{w}_k^{(t)} - \boldsymbol{w}_{k+1}^{(t)}\right\|^2 \sum_{t=t_1}^{T}\sum_{k=0}^{K-1} t^{-2}} + \sum_{t=t_1}^{T}\sum_{k=0}^{K-1} t^{-2} \\
&< \infty
\end{aligned} \tag{18}$$

We use Cauchy-Schwarz inequality for the first inequality and the factor that $\sum_{t=t_1}^{T} t^{-2} < \infty$ and $\sum_{t=t_1}^{T}\sum_{k=0}^{K-1}\left\|\boldsymbol{w}_k^{(t)} - \boldsymbol{w}_{k+1}^{(t)}\right\|^2 < \infty$ by Theorem 3.1.

Combined with Lemma D.5 and the fact that $\forall c > 1 : \sum_{t=1}^{\infty} t^{-c} < \infty$, we get

$$\begin{aligned}
\left\|\boldsymbol{r}_0^{(t)}\right\|^2 - \left\|\boldsymbol{r}_0^{(t_1)}\right\|^2 &= \sum_{u=t_1}^{t-1}\sum_{k=0}^{K-1}\left(\left\|\boldsymbol{r}_{k+1}^{(u)}\right\|^2 - \left\|\boldsymbol{r}_k^{(u)}\right\|^2\right) \\
&= \sum_{u=t_1}^{t-1}\sum_{k=0}^{K-1}\left(2(\boldsymbol{r}_{k+1}^{(u)} - \boldsymbol{r}_k^{(u)})^\top \boldsymbol{r}_k^{(u)} + \left\|\boldsymbol{r}_{k+1}^{(u)} - \boldsymbol{r}_k^{(u)}\right\|^2\right) < \infty
\end{aligned}$$

Hence $\left\|\boldsymbol{r}_k^{(t)}\right\|$ is bounded.

D.2.1 Proof of Lemma D.4

$$\begin{aligned}
&K\sum_{u=1}^{t-1}\frac{1}{u}\sum_{s\in S^{(u)}}\alpha_s \boldsymbol{x}_s + \frac{k}{t}\sum_{s\in S^{(t)}}\alpha_s \boldsymbol{x}_s \\
&= K\sum_{u=1}^{\lfloor\frac{t-1}{M}\rfloor M}\frac{1}{u}\sum_{s\in S^{(u)}}\alpha_s \boldsymbol{x}_s + K\sum_{u=\lfloor\frac{t-1}{M}\rfloor M+1}^{t-1}\frac{1}{u}\sum_{s\in S^{(u)}}\alpha_s \boldsymbol{x}_s + \frac{k}{t}\sum_{s\in S^{(t)}}\alpha_s \boldsymbol{x}_s \\
&= K\sum_{u=1}^{\lfloor\frac{t-1}{M}\rfloor M}\frac{1}{u}\sum_{s\in S^{(u)}}\alpha_s \boldsymbol{x}_s + m'(t,k)
\end{aligned}$$

$$
\begin{aligned}
&= K \sum_{u=1}^{\lfloor \frac{t-1}{M} \rfloor} \left[\sum_{v=1}^{M} \frac{1}{v+M(u-1)} \left(\sum_{s\in S^{(v)}} \alpha_s \boldsymbol{x}_s \right) \right] + m'(t,k) \\
&= K \sum_{v=1}^{M} \left[\sum_{u=1}^{\lfloor \frac{t-1}{M} \rfloor} \frac{1}{v+M(u-1)} \left(\sum_{s\in S^{(v)}} \alpha_s \boldsymbol{x}_s \right) \right] + m'(t,k)
\end{aligned}
$$

Note that $m'(t,k)$ and $m'(t,k+1) - m'(t,k)$ are both $O(t^{-1})$ for all $k \in [0 : K-1]$. For every v,

$$
\begin{aligned}
&\sum_{u=1}^{\lfloor \frac{t-1}{M} \rfloor} \frac{1}{v+M(u-1)} \left(\sum_{s\in S^{(v)}} \alpha_s \boldsymbol{x}_s \right) \\
&\quad = \sum_{u=1}^{\lfloor \frac{t-1}{M} \rfloor} \left[\frac{1}{Mu} + \frac{1-\frac{v}{M}}{Mu^2+(v-M)u} \right] \left(\sum_{s\in S^{(v)}} \alpha_s \boldsymbol{x}_s \right) \\
&\quad = \left[\frac{1}{M} \left(\log \left(\lfloor \frac{t-1}{M} \rfloor \right) + \gamma + O(t^{-1}) \right) + \sum_{u=1}^{\lfloor \frac{t-1}{M} \rfloor} \frac{1-\frac{v}{M}}{Mu^2+(v-M)u} \right] \left(\sum_{s\in S^{(v)}} \alpha_s \boldsymbol{x}_s \right) \\
&\quad = \left[\frac{1}{M} \left(\log \left(\frac{t-1}{M} \right) + \gamma + O(t^{-1}) \right) + \sum_{u=1}^{\lfloor \frac{t-1}{M} \rfloor} \frac{1-\frac{v}{M}}{Mu^2+(v-M)u} \right] \left(\sum_{s\in S^{(v)}} \alpha_s \boldsymbol{x}_s \right) \\
&\quad = \left[\frac{1}{M} \left(\log \left(\frac{t}{M} \right) + \gamma + O(t^{-1}) \right) + \sum_{u=1}^{\lfloor \frac{t-1}{M} \rfloor} \frac{1-\frac{v}{M}}{Mu^2+(v-M)u} \right] \left(\sum_{s\in S^{(v)}} \alpha_s \boldsymbol{x}_s \right)
\end{aligned}
$$

where in the last three equality, we use the fact

$$
\begin{gathered}
\sum_{u=1}^{t} \frac{1}{u} = \log t + \gamma + O(t^{-1}) \\
\log(t) - \log(\lfloor t \rfloor) = O(t^{-1}) \\
\log(t) - \log(t-1) = O(t^{-1})
\end{gathered}
$$

where γ is the Euler-Mascheroni constant. Since $1 \le v \le M$, $\frac{1-\frac{v}{M}}{Mu^2+(v-M)u} \le \frac{1-\frac{v}{M}}{vu^2}$. Therefore, $\sum_u \frac{1-\frac{v}{M}}{Mu^2+(v-M)u}$ converges with a rate $O(t^{-1})$.

$$
\begin{aligned}
\sum_{u=1}^{\lfloor \frac{t-1}{M} \rfloor} \frac{1-\frac{v}{M}}{Mu^2+(v-M)u} &= \sum_{u=1}^{\infty} \frac{1-\frac{v}{M}}{Mu^2+(v-M)u} - \sum_{u=\lfloor \frac{t-1}{M} \rfloor + 1}^{\infty} \frac{1-\frac{v}{M}}{Mu^2+(v-M)u} \\
&= \sum_{u=1}^{\infty} \frac{1-\frac{v}{M}}{Mu^2+(v-M)u} + O(t^{-1})
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&K \sum_{v=1}^{M} \left[\sum_{u=1}^{\lfloor \frac{t-1}{M} \rfloor} \frac{1}{v+M(u-1)} \left(\sum_{s\in S^{(v)}} \alpha_s \boldsymbol{x}_s \right) \right] \\
&= \frac{K}{M} \left(\log \frac{t}{M} + \gamma \right) \left(\sum_{s\in S} \alpha_s \boldsymbol{x}_s \right) + K \sum_{v=1}^{M} \sum_{u=1}^{\infty} \frac{1-\frac{v}{M}}{Mu^2+(v-M)u} \left(\sum_{s\in S^{(v)}} \alpha_s \boldsymbol{x}_s \right) + m''(t) \\
&= \frac{K}{M} \left(\log \frac{t}{M} + \gamma \right) \hat{\boldsymbol{w}} + K \sum_{v=1}^{M} \sum_{u=1}^{\infty} \frac{1-\frac{v}{M}}{Mu^2+(v-M)u} \left(\sum_{s\in S^{(v)}} \alpha_s \boldsymbol{x}_s \right) + m''(t)
\end{aligned}
$$

$$= \frac{K}{M}\log(\frac{t}{M})\hat{\boldsymbol{w}} + \frac{K}{M}\breve{\boldsymbol{w}} + m''(t)$$

where $\breve{\boldsymbol{w}} := \gamma\hat{\boldsymbol{w}} + M\sum_{v=1}^{M}\sum_{u=1}^{\infty}\frac{1-\frac{v}{M}}{Mu^2+(v-M)u}\left(\sum_{s\in S^{(v)}}\alpha_s\boldsymbol{x}_s\right)$, and $m''(t) = O(t^{-1})$.

Finally, for all $k \in [0 : K-1]$ let

$$m_1(t,k) := K\sum_{u=1}^{t-1}\frac{1}{u}\sum_{s\in S^{(u)}}\alpha_s\boldsymbol{x}_s + \frac{k}{t}\sum_{s\in S^{(t)}}\alpha_s\boldsymbol{x}_s - \frac{K}{M}\log(\frac{t}{M})\hat{\boldsymbol{w}} - \frac{K}{M}\breve{\boldsymbol{w}}$$

and

$$m_1(t,K) := m_1(t+1,0)$$

Then $m_1(t,k) = m'(t,k) + m''(t) = O(t^{-1})$, and

$$\forall k \in [0 : K-1] : m_1(t,k+1) - m_1(t,k) = \frac{1}{t}\sum_{s\in S^{(t)}}\alpha_s\boldsymbol{x}_s = O(t^{-1})$$

$$m_1(t+1,0) - m_1(t,K-1) = \frac{1}{t}\sum_{s\in S^{(t)}}\alpha_s\boldsymbol{x}_s - \frac{K}{M}\log(1+t^{-1})\hat{\boldsymbol{w}} = O(t^{-1})$$

D.2.2 Proof of Lemma D.5

We use Assumption 3.4 here. That is, there exist positive constants μ_+, μ_-, and $\bar{u}$ such that $\forall u > \bar{u}$:

$$(1-\exp(-\mu_- u))e^{-u} \le -\ell'(u) \le (1+\exp(-\mu_+ u))e^{-u}$$

By definition,

$$\forall k \in [0 : K-1] : \boldsymbol{r}_k^{(t)} = \boldsymbol{w}_k^{(t)} - \frac{M}{K}\left(K\sum_{u=1}^{t-1}\frac{1}{u}\sum_{s\in S^{(u)}}\alpha_s\boldsymbol{x}_s + \frac{k}{t}\sum_{s\in S^{(t)}}\alpha_s\boldsymbol{x}_s\right) - \log K\hat{\boldsymbol{w}} - \tilde{\boldsymbol{w}} + \breve{\boldsymbol{w}}$$

$$\boldsymbol{r}_K^{(t)} = \boldsymbol{r}_0^{(t+1)}$$

Then for all $k \in [0 : K-1]$, we get

$$\begin{aligned}
\boldsymbol{r}_{k+1}^{(t)} - \boldsymbol{r}_k^{(t)} &= \boldsymbol{w}_{k+1}^{(t)} - \boldsymbol{w}_k^{(t)} - \frac{M}{Kt}\sum_{s\in S^{(t)}}\alpha_s\boldsymbol{x}_s \\
&= -\eta\sum_{s\in I^{(t)}}\ell'(\boldsymbol{x}_s^\top\boldsymbol{w}_k^{(t)})\boldsymbol{x}_s - \frac{M}{Kt}\sum_{s\in S^{(t)}}\alpha_s\boldsymbol{x}_s \\
&= -\eta\sum_{s\in I^{(t)}\setminus S^{(t)}}\ell'(\boldsymbol{x}_s^\top\boldsymbol{w}_k^{(t)})\boldsymbol{x}_s - \sum_{s\in S^{(t)}}\left[\eta\ell'(\boldsymbol{x}_s^\top\boldsymbol{w}_k^{(t)}) + \frac{M}{Kt}\alpha_s\right]\boldsymbol{x}_s
\end{aligned}$$

Hence,

$$\begin{aligned}
\left(\boldsymbol{r}_{k+1}^{(t)} - \boldsymbol{r}_k^{(t)}\right)^\top\boldsymbol{r}_k^{(t)} &= -\eta\sum_{s\in I^{(t)}\setminus S^{(t)}}\ell'(\boldsymbol{x}_s^\top\boldsymbol{w}_k^{(t)})\boldsymbol{x}_s^\top\boldsymbol{r}_k^{(t)} - \sum_{s\in S^{(t)}}\left[\eta\ell'(\boldsymbol{x}_s^\top\boldsymbol{w}_k^{(t)}) + \frac{M}{Kt}\alpha_s\right]\boldsymbol{x}_s^\top\boldsymbol{r}_k^{(t)} \\
&= -\eta\sum_{s\in I^{(t)}\setminus S^{(t)}}\ell'\left(\log(\frac{K}{M}t)\boldsymbol{x}_s^\top\hat{\boldsymbol{w}} + \frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) + \boldsymbol{x}_s^\top\tilde{\boldsymbol{w}} + \boldsymbol{x}_s^\top\boldsymbol{r}_k^{(t)}\right)\boldsymbol{x}_s^\top\boldsymbol{r}_k^{(t)} \qquad (19) \\
&\quad - \sum_{s\in S^{(t)}}\left[\eta\ell'\left(\log(\frac{K}{M}t) + \frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) + \boldsymbol{x}_s^\top\tilde{\boldsymbol{w}} + \boldsymbol{x}_s^\top\boldsymbol{r}_k^{(t)}\right) + \frac{M}{Kt}\alpha_s\right]\boldsymbol{x}_s^\top\boldsymbol{r}_k^{(t)} \qquad (20)
\end{aligned}$$

The behavior of each term can be analyzed when stage t is large. To achieve this, we first characterize five stages.

$$
\begin{aligned}
t_5 &:= \min\{t' \mid \forall t \geq t', \forall k \in [0:K-1], \forall s \in I : \boldsymbol{x}_s^\top \boldsymbol{w}_k^{(t)} \geq \bar{u}\} \\
t_6 &:= \min\{t' \mid \forall t \geq t', \forall k \in [0:K-1], \forall s \in I : \boldsymbol{x}_s^\top \boldsymbol{w}_k^{(t)} \geq 0\} \\
t_7 &:= \min\{t' \mid \forall t \geq t', \forall k \in [0:K-1], \forall s \in I : \exp\left(-\frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k)\right) \leq 2\} \\
t_8 &:= \min\{t' \mid \forall t \geq t', \forall k \in [0:K-1], \forall s \in I : \exp\left(-\frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k)\right) \geq \frac{1}{2}\} \\
t_9 &:= \min\{t' \mid \forall t \geq t', \forall k \in [0:K-1], \forall s \in I : \exp\left(-\mu_- \boldsymbol{x}_s^\top \boldsymbol{w}_k^{(t)}\right) \leq \frac{1}{2}\}
\end{aligned}
$$

Such $t_5 \sim t_9$ exist since $\forall s \in I, \forall k \in [0:K-1] : \lim_{t\to\infty} \boldsymbol{x}_s^\top \boldsymbol{w}_k^{(t)} = \infty$ by Theorem 3.1, and $\forall k \in [0:K-1] : \lim_{t\to\infty} \|m_1(t,k)\| = 0$ by Lemma D.4.

Then for all $t \geq \max\{t_5, t_6, t_7, t_8, t_9\}$, the first term (19) can be upper bounded as below:

$$
-\eta \sum_{s\in I^{(t)}\setminus S^{(t)}} \ell'(\boldsymbol{x}_s^\top \boldsymbol{w}_k^{(t)})\boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \leq -\eta \sum_{\substack{s\in I^{(t)}\setminus S^{(t)} \\ \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)}>0}} \ell'(\boldsymbol{x}_s^\top \boldsymbol{w}_k^{(t)})\boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)}
$$

$$
\leq \eta \sum_{\substack{s\in I^{(t)}\setminus S^{(t)} \\ \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)}>0}} \left(1+\exp(-\mu_+ \boldsymbol{x}_s^\top \boldsymbol{w}_k^{(t)})\right)\exp(-\boldsymbol{x}_s^\top \boldsymbol{w}_k^{(t)})\boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \qquad t \geq t_5
$$

$$
\leq \eta \sum_{\substack{s\in I^{(t)}\setminus S^{(t)} \\ \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)}>0}} 2\exp\left(-\log(\frac{K}{M}t)\boldsymbol{x}_s^\top \hat{\boldsymbol{w}} - \frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) - \boldsymbol{x}_s^\top \tilde{\boldsymbol{w}} - \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)}\right)\boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \qquad t \geq t_6
$$

$$
\leq \sum_{\substack{s\in I^{(t)}\setminus S^{(t)} \\ \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)}>0}} 2\alpha_s \exp\left(-\log(\frac{K}{M}t)\boldsymbol{x}_s^\top \hat{\boldsymbol{w}} - \frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) - \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)}\right)\boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \tag{21}
$$

$$
\leq \sum_{\substack{s\in I^{(t)}\setminus S^{(t)} \\ \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)}>0}} 2\alpha_s \exp\left(-\log(\frac{K}{M}t)\boldsymbol{x}_s^\top \hat{\boldsymbol{w}} - \frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k)\right) \tag{22}
$$

$$
\leq \sum_{\substack{s\in I^{(t)}\setminus S^{(t)} \\ \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)}>0}} 4\alpha_s \exp\left(-\log(\frac{K}{M}t)\boldsymbol{x}_s^\top \hat{\boldsymbol{w}}\right) \qquad t \geq t_7 \tag{23}
$$

$$
\leq 4N(\max_s \alpha_s)\left(\frac{Kt}{M}\right)^{-\theta} \tag{24}
$$

where in (21) we use the definition of $\tilde{\boldsymbol{w}}$, in (22) we use the fact $\forall x \geq 0 : x\exp(-x) \leq 1$, and in (24) we use $\forall s \in I^{(t)} \setminus S^{(t)} : x_s^\top \hat{\boldsymbol{w}} \geq \theta$. Now we examine the second term (20). Given $t \geq t_5$, it can be divided into two cases.

$$
-\ell'(\boldsymbol{x}_s^\top \boldsymbol{w}_k^{(t)})\boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \leq \begin{cases} \left(1+\exp(-\mu_+ \boldsymbol{x}_s^\top \boldsymbol{w}_k^{(t)})\right)\exp(-\boldsymbol{x}_s^\top \boldsymbol{w}_k^{(t)})\boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} & \text{if } \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} > 0 \\ \left(1-\exp(-\mu_- \boldsymbol{x}_s^\top \boldsymbol{w}_k^{(t)})\right)\exp(-\boldsymbol{x}_s^\top \boldsymbol{w}_k^{(t)})\boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} & \text{if } \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \leq 0 \end{cases}
$$

For each $s \in S$, define $A_{s,k}^{(t)}$ as

$$
A_{s,k}^{(t)} := \begin{cases} 1+\exp(-\mu_+ \boldsymbol{x}_s^\top \boldsymbol{w}_k^{(t)}) & \text{if } \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} > 0 \\ 1-\exp(-\mu_- \boldsymbol{x}_s^\top \boldsymbol{w}_k^{(t)}) & \text{if } \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \leq 0 \end{cases}
$$

Then, we can use

$$-\ell'(\boldsymbol{x}_s^\top \boldsymbol{w}_k^{(t)})\boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \leq A_{s,k}^{(t)} \exp(-\boldsymbol{x}_s^\top \boldsymbol{w}_k^{(t)})\boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)}$$

in any $s \in S, k \in [0 : K-1]$. Therefore the second term (20) is bounded

$$\begin{aligned}
&- \sum_{s \in S^{(t)}} \left[\eta \ell' \left(\log\left(\frac{K}{M}t\right) + \frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) + \boldsymbol{x}_s^\top \tilde{\boldsymbol{w}} + \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \right) + \frac{M}{Kt}\alpha_s \right] \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \\
&\leq \sum_{s \in S^{(t)}} \left[\eta A_{s,k}^{(t)} \exp\left(-\log\left(\frac{K}{M}t\right) - \frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) - \boldsymbol{x}_s^\top \tilde{\boldsymbol{w}} - \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \right) - \frac{M}{Kt}\alpha_s \right] \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \\
&= \sum_{s \in S^{(t)}} \left[A_{s,k}^{(t)} \frac{M\alpha_s}{Kt} \exp\left(-\frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) - \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \right) - \frac{M}{Kt}\alpha_s \right] \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \\
&= \sum_{s \in S^{(t)}} \frac{M}{Kt}\alpha_s \left[A_{s,k}^{(t)} \exp\left(-\frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) - \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \right) - 1 \right] \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)}
\end{aligned}$$

Now we analyze each $s \in S^{(t)}$ by dividing into cases. Note that $\left|\frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k)\right| = o(t^{-0.5+\epsilon})$ for all $\epsilon > 0$. Therefore if we set $\tilde{\mu} = \min\{\mu_+, \mu_-, 0.25\}$, then $\left|\frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k)\right| = o(t^{-\tilde{\mu}})$.

1. if $0 \leq \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \leq C_7 t^{-0.5\tilde{\mu}}$:

$$\begin{aligned}
&\frac{M}{Kt}\alpha_s \left[A_{s,k}^{(t)} \exp\left(-\frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) - \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \right) - 1 \right] \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \\
&\qquad \leq \left[2\exp\left(-\frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) - \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \right) - 1 \right] \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} && t \geq t_6 \\
&\qquad \leq \left[4\exp\left(-\boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \right) - 1 \right] \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} && t \geq t_7 \\
&\qquad \leq \left(\max_s \alpha_s \right) \frac{4MC_7}{K} t^{-1-0.5\tilde{\mu}}
\end{aligned}$$

   The last inequality holds by the case condition $0 \leq \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \leq C_7 t^{-0.5\tilde{\mu}}$.

2. if $-C_7 t^{-0.5\tilde{\mu}} \leq \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \leq 0$:

$$\begin{aligned}
&\frac{M}{Kt}\alpha_s \left[A_{s,k}^{(t)} \exp\left(-\frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) - \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \right) - 1 \right] \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \\
&\qquad = \frac{M}{Kt}\alpha_s \left[1 - A_{s,k}^{(t)} \exp\left(-\frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) - \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \right) \right] \left| \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \right| \\
&\qquad \leq \frac{M}{Kt}\alpha_s \left| \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \right| \leq \frac{M}{Kt}\alpha_s \cdot C_7 t^{-0.5\tilde{\mu}} \\
&\qquad \leq \left(\max_s \alpha_s \right) \frac{MC_7}{K} t^{-1-0.5\tilde{\mu}}
\end{aligned}$$

3. if $C_7 t^{-0.5\tilde{\mu}} < \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)}$:

   Here, we first examine $A_{s,k}^{(t)}$.

$$\begin{aligned}
A_{s,k}^{(t)} &= 1 + \exp(-\mu_+ \boldsymbol{x}_s^\top \boldsymbol{w}_k^{(t)}) \\
&= 1 + \exp\left(-\mu_+ \left(\log\left(\frac{K}{M}t\right) + \frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) + \boldsymbol{x}_s^\top \tilde{\boldsymbol{w}} + \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \right) \right) \\
&\leq 1 + \exp\left(-\mu_+ \left(\log\left(\frac{K}{M}t\right) + \frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) + \boldsymbol{x}_s^\top \tilde{\boldsymbol{w}} \right) \right) \\
&\leq 1 + 2^{\mu_+} \exp\left(-\mu_+ \left(\log\left(\frac{K}{M}t\right) + \boldsymbol{x}_s^\top \tilde{\boldsymbol{w}} \right) \right) && t \geq t_7 \\
&\leq 1 + C_8 t^{-\mu_+}
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\frac{M}{Kt}\alpha_s \left[A_{s,k}^{(t)} \exp\left(-\frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) - \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)}\right) - 1\right] \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \\
&\quad \le \frac{M}{Kt}\alpha_s \left[\left(1 + C_8 t^{-\mu_+}\right) \exp\left(-\frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) - \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)}\right) - 1\right] \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \\
&\quad \le \frac{M}{Kt}\alpha_s \left[\left(1 + C_8 t^{-\mu_+}\right) \exp\left(-\frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) - C_7 t^{-0.5\tilde{\mu}}\right) - 1\right] \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \qquad (25)
\end{aligned}$$

Since $t \ge t_7$, $-\frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) \le 1$. Now we use the fact $\forall x \le 1 : \exp x \le 1 + x + x^2$.

$$\exp\left(-\frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k)\right) \le 1 - \frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) + \left(\frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k)\right)^2$$

$$\exp\left(-C_7 t^{-0.5\tilde{\mu}}\right) \le 1 - C_7 t^{-0.5\tilde{\mu}} + C_7^2 t^{-\tilde{\mu}}$$

Then we get

$$\begin{aligned}
&\left(1 + C_8 t^{-\mu_+}\right) \exp\left(-\frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) - C_7 t^{-0.5\tilde{\mu}}\right) \\
&\le \left(1 - \frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) + \left(\frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k)\right)^2\right)\left(1 - C_7 t^{-0.5\tilde{\mu}}\right) + o(t^{-\mu_+}) \\
&\le 1 - \frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) + \left(\frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k)\right)^2 - C_7 t^{-0.5\tilde{\mu}} + o(t^{-\mu_+}) \\
&\le 1 - C_7 t^{-0.5\tilde{\mu}} + o(t^{-\tilde{\mu}})
\end{aligned}$$

where in the last two inequality, we use $\left|\frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k)\right| = o(t^{-\tilde{\mu}})$.
Finally, Equation (25) is bounded

$$\begin{aligned}
&\frac{M}{Kt}\alpha_s \left[\left(1 + C_8 t^{-\mu_+}\right) \exp\left(-\frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) - C_7 t^{-0.5\tilde{\mu}}\right) - 1\right] \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \\
&\le \frac{M}{Kt}\alpha_s \left[-C_7 t^{-0.5\tilde{\mu}} + o(t^{-\tilde{\mu}})\right] \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)}
\end{aligned}$$

Since $-C_7 t^{-0.5\tilde{\mu}}$ decrease to zero slower than the other term, $\exists t_+ \ge \max\{t_5, t_6, t_7, t_8, t_9\}$ such that for all $t \ge t_+$, the last term is negative.

4. if $\boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} < -C_7 t^{-0.5\tilde{\mu}}$:

Since $\boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} < 0$, it is enough to show that $A_{s,k}^{(t)} \exp\left(-\frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) - \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)}\right) > 1$ for sufficiently large t. Note that $A_{s,k}^{(t)} = 1 - \exp(-\mu_- \boldsymbol{x}_s^\top \boldsymbol{w}_k^{(t)}) > 0$ since $t \ge t_6$. If $\exp\left(-\boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)}\right) \ge 4$,

$$\begin{aligned}
&A_{s,k}^{(t)} \exp\left(-\frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) - \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)}\right) \\
&\qquad \ge 4(1 - \exp(-\mu_- \boldsymbol{x}_s^\top \boldsymbol{w}_k^{(t)})) \exp\left(-\frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k)\right) \ge 1
\end{aligned}$$

The last inequality holds by $t \ge \max\{t_8, t_9\}$. Now, if $\exp\left(-\boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)}\right) < 4$,

$$\begin{aligned}
A_{s,k}^{(t)} &= 1 - \exp\left(-\mu_-\left(\log\left(\frac{K}{M}t\right) + \frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) + \boldsymbol{x}_s^\top \tilde{\boldsymbol{w}} + \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)}\right)\right) \\
&\ge 1 - \left(\frac{4Kt}{M}\right)^{-\mu_-} \exp\left(-\mu_-\left(\frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) + \boldsymbol{x}_s^\top \tilde{\boldsymbol{w}}\right)\right)
\end{aligned}$$

$$\geq 1 - \left(\frac{8Kt}{M}\right)^{-\mu_-} \exp\left(-\mu_- \boldsymbol{x}_s^\top \tilde{\boldsymbol{w}}\right) \geq 1 - C_9 t^{-\mu_-} \qquad t \geq t_7$$

Also, by the fact $\forall x : \exp x \geq 1 + x$,

$$\exp\left(-\frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) - \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)}\right) \geq \left(1 - \frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k)\right)\left(1 - \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)}\right)$$

Combined with the former inequality,

$$\begin{aligned}
&A_{s,k}^{(t)} \exp\left(-\frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) - \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)}\right) \\
&\quad \geq \left(1 - C_9 t^{-\mu_-}\right)\left(1 - \frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k)\right)\left(1 - \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)}\right) \\
&\quad \geq \left(1 - C_9 t^{-\mu_-}\right)\left(1 + o(t^{-\tilde{\mu}})\right)\left(1 + C_7 t^{-0.5\tilde{\mu}}\right) \\
&\quad = 1 + C_7 t^{-0.5\tilde{\mu}} - o(t^{-\tilde{\mu}})
\end{aligned}$$

Since $C_7 t^{-0.5\tilde{\mu}}$ decrease to zero slower than the other term, $\exists t_- \geq \max\{t_5, t_6, t_7, t_8, t_9\}$ such that for all $t \geq t_-$, the last equation is larger than 1.

To sum up, there exist $C_1, C_2 > 0, \tilde{t} \geq \max\{t_+, t_-\}$ such that for all $t \geq \tilde{t}$,

$$(\boldsymbol{r}_{k+1}^{(t)} - \boldsymbol{r}_k^{(t)})^\top \boldsymbol{r}_k^{(t)} \leq C_1 t^{-\theta} + C_2 t^{-1-0.5\tilde{\mu}}, \forall k \in [0 : K-1]$$

Now we consider special cases to finish the lemma. For any $\epsilon_2 > 0$, the following analysis holds.

1. If $\boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \geq \epsilon_2 > 0$:
   Since $\lim_{t\to\infty} m_1(t,k) = 0$, there exist $t_1^* \geq \max\{t_+, t_-\}$ such that $\forall t \geq t_1^*, \forall s \in S, \forall k \in [0 : K-1] : \left|\frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k)\right| < 0.5\epsilon_2$. Also since $\lim_{t\to\infty} \boldsymbol{x}_s^\top \boldsymbol{w}_k^{(t)} \to \infty$, there exist $t_+^* \geq t_1^*$ such that $\forall t \geq t_+^*, \forall s \in S, \forall k \in [0 : K-1] : \exp\left(-\mu_+ \boldsymbol{x}_s^\top \boldsymbol{w}_k^{(t)}\right) \leq \exp(0.25\epsilon_2) - 1$. Therefore for $t \geq t_+^*$,

$$\begin{aligned}
&\frac{M}{Kt}\alpha_s \left[A_{s,k}^{(t)} \exp\left(-\frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) - \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)}\right) - 1\right] \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \\
&\quad \leq \frac{M}{Kt}\alpha_s \left[\left(1 + \exp(-\mu_+ \boldsymbol{x}_s^\top \boldsymbol{w}_k^{(t)})\right)\exp(-0.5\epsilon_2) - 1\right] \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \qquad t \geq t_1^* \\
&\quad \leq \frac{M}{Kt}\alpha_s \left(\exp(-0.25\epsilon_2) - 1\right) \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \qquad t \geq t_+^* \\
&\quad \leq \min_s \alpha_s \frac{M}{K}\left(\exp(-0.25\epsilon_2) - 1\right)\epsilon_2 \frac{1}{t} = -C''_+ t^{-1}
\end{aligned}$$

2. If $\boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \leq -\epsilon_2 < 0$:
   Again, since $\lim_{t\to\infty} \boldsymbol{x}_s^\top \boldsymbol{w}_k^{(t)} \to \infty$, there exist $t_-^* \geq t_1^*$ such that $\forall t \geq t_-^*, \forall s \in S, \forall k \in [0 : K-1] : 1 - \exp\left(-\mu_- \boldsymbol{x}_s^\top \boldsymbol{w}_k^{(t)}\right) \geq \exp(-0.25\epsilon_2)$. Therefore for $t \geq t_-^*$,

$$\begin{aligned}
&\frac{M}{Kt}\alpha_s \left[A_{s,k}^{(t)} \exp\left(-\frac{M}{K}\boldsymbol{x}_s^\top m_1(t,k) - \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)}\right) - 1\right] \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \\
&\quad \leq \frac{M}{Kt}\alpha_s \left[\left(1 - \exp(-\mu_- \boldsymbol{x}_s^\top \boldsymbol{w}_k^{(t)})\right)\exp(0.5\epsilon_2) - 1\right] \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \qquad t \geq t_1^* \\
&\quad \leq \frac{M}{Kt}\alpha_s \left(\exp(0.25\epsilon_2) - 1\right) \boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)} \qquad t \geq t_-^* \\
&\quad \leq -\min_s \alpha_s \frac{M}{K}\left(\exp(0.25\epsilon_2) - 1\right)\epsilon_2 \frac{1}{t} = -C''_- t^{-1}
\end{aligned}$$

In conclusion, for any $\epsilon_1 > 0$, if $\left\|P\boldsymbol{r}_k^{(t)}\right\| \geq \epsilon_1$ and $S^{(t)} \neq \emptyset$, then

$$\max_{s\in S^{(t)}} \left|\boldsymbol{x}_s^\top \boldsymbol{r}_k^{(t)}\right|^2 = \max_{s\in S^{(t)}} \left|\left(P^\top \boldsymbol{x}_s\right)^\top \boldsymbol{r}_k^{(t)}\right|^2 \geq \frac{1}{\left|S^{(t)}\right|} \sum_{s\in S^{(t)}} \left|\boldsymbol{x}_s^\top P \boldsymbol{r}_k^{(t)}\right|^2$$

$$= \frac{1}{\left|S^{(t)}\right|}\left\|X_{S^{(t)}}^{\top} P\boldsymbol{r}_k^{(t)}\right\|^2 \geq \frac{1}{\left|S^{(t)}\right|}\sigma_{\min}^2(X_{S^{(t)}})\left\|P\boldsymbol{r}_k^{(t)}\right\|^2 \geq \frac{1}{\left|S^{(t)}\right|}\sigma_{\min}^2(X_{S^{(t)}})\epsilon_1^2$$

where $X_{S^{(t)}} \in \mathbb{R}^{d\times\left|S^{(t)}\right|}$ is a matrix which has $\{x_s \mid s \in S^{(t)}\}$ as its columns. By Assumption 3.2, $\sigma_{\min}(X_{S^{(t)}})$ is non-zero. Therefore, for all $\epsilon_1 > 0$, $\exists \tilde{t}^*, C_3 > 0$ such that if $\left\|P\boldsymbol{r}_k^{(t)}\right\| \geq \epsilon_1$ and $S^{(t)} \neq \emptyset$,

$$(\boldsymbol{r}_{k+1}^{(t)} - \boldsymbol{r}_k^{(t)})^{\top}\boldsymbol{r}_k^{(t)} \leq -C_3 t^{-1}, \forall t > \tilde{t}^*, k \in [0 : K-1]$$

D.2.3 CONVERGENCE OF $\rho_k^{(t)}$

Theorem 3.2 only shows boundedness of $\boldsymbol{\rho}_k^{(t)}$. Yet, if additional mild assumption on data is given, it can be guaranteed for $\boldsymbol{\rho}_k^{(t)}$ to converge to the particular vector.

Assumption D.1. Support vectors span dataset. That is, $\operatorname{rank}\{\boldsymbol{x}_i : i \in S\} = \operatorname{rank}\{\boldsymbol{x}_i : i \in I\}$.

Proposition D.6. *Under the same setting as Theorem 3.2 with an additional Assumption D.1, the "residual" converges to $\lim_{t\to\infty}\boldsymbol{\rho}_k^{(t)} = \tilde{\boldsymbol{w}}, \forall k \in [0 : K-1]$. Here, $\tilde{\boldsymbol{w}}$ is the unique solution of the following system of equations*

$$\forall i \in S : \eta \exp\left(-\boldsymbol{x}_i^{\top}\tilde{\boldsymbol{w}}\right) = \alpha_i, \quad (I - P)(\tilde{\boldsymbol{w}} - \boldsymbol{w}_0^{(0)}) = 0,$$

where $P \in \mathbb{R}^{d\times d}$ is the orthogonal projection matrix to the space spanned by the joint support vectors indexed by S.

We set $\bar{P} = I - P$ for the convenience of proof.

Proof. By the definition of $\boldsymbol{\rho}_k^{(t)} = \tilde{\boldsymbol{w}} + \frac{M}{K}m_1(t,k) + \boldsymbol{r}_k^{(t)}$, it is enough to prove $\lim_{t\to\infty}\boldsymbol{r}_k^{(t)} = 0$.

First of all, since $\boldsymbol{w}_k^{(t)} = \log(\frac{K}{M}t)\hat{\boldsymbol{w}} + \tilde{\boldsymbol{w}} + \frac{M}{K}m_1(t,k) + \boldsymbol{r}_k^{(t)}$,

$$\begin{aligned}
\bar{P}\boldsymbol{r}_k^{(t)} &= \bar{P}\boldsymbol{w}_k^{(t)} - \log(\frac{K}{M}t)\bar{P}\hat{\boldsymbol{w}} - \bar{P}\tilde{\boldsymbol{w}} - \frac{M}{K}\bar{P}m_1(t,k) \\
&= \bar{P}\boldsymbol{w}_0^{(0)} - \log(\frac{K}{M}t)\bar{P}\hat{\boldsymbol{w}} - \bar{P}\tilde{\boldsymbol{w}} - \frac{M}{K}\bar{P}m_1(t,k) \\
&= \bar{P}\boldsymbol{w}_0^{(0)} - \bar{P}\tilde{\boldsymbol{w}} = 0
\end{aligned}$$

The first line holds under the Assumption D.1 since $\nabla\mathcal{L}(w)$ is a linear combination of the columns of X. that is, $\forall l < t : \bar{P}\nabla\mathcal{L}^{(l)}(w) = 0$. Remaining lines are true by the definition.

Second, we get to show $P\boldsymbol{r}_k^{(t)} \to 0$. By Equation (18), $\lim_{T\to\infty}\sum_{t=t_1}^{T}\sum_{k=0}^{K-1}\left\|\boldsymbol{r}_{k+1}^{(t)} - \boldsymbol{r}_k^{(t)}\right\|^2 = C_4$. That means $\forall k \in [0 : K-1] : \lim_{T\to\infty}\left\|\boldsymbol{r}_{k+1}^{(T)} - \boldsymbol{r}_k^{(T)}\right\| = 0$. Therefore, for any ϵ_0, there exists $t_2 > 0$ such that $\left\|\boldsymbol{r}_{k+1}^{(t)} - \boldsymbol{r}_k^{(t)}\right\| < \frac{\epsilon_0}{K}$ for all $t \geq t_2, k \in [0 : K-1]$. As a result,

$$\left\|P\boldsymbol{r}_0^{(t)}\right\| + \frac{k}{K}\epsilon_0 \geq \left\|P\boldsymbol{r}_k^{(t)}\right\| \geq \left\|P\boldsymbol{r}_0^{(t)}\right\| - \frac{k}{K}\epsilon_0$$

For $t \geq \max\{t_1, t_2, \tilde{t}^*\}$, if $\left\|P\boldsymbol{r}_0^{(t)}\right\| \geq \epsilon_1 + \epsilon_0$ and $S^{(t)} \neq \emptyset$, then $\forall k \in [0 : K-1] : \left\|P\boldsymbol{r}_k^{(t)}\right\| \geq \epsilon_1$. By Lemma D.5 (2),

$$\forall m \in [0 : M-1] : \sum_{u=t}^{t+m}\sum_{v=0}^{K-1}(\boldsymbol{r}_{v+1}^{(u)} - \boldsymbol{r}_v^{(u)})^{\top}\boldsymbol{r}_v^{(u)} \leq -KC_3t^{-1} + Km\left(C_1t^{-\theta} + C_2t^{-1-0.5\tilde{\mu}}\right),$$

Since t^{-1} decrease to zero slower than $t^{-\theta}$ and $t^{-1-0.5\tilde{\mu}}$, there exists $t_3 > \max\{t_1, t_2, \tilde{t}^*\}, C_4 > 0$ such that $-KC_3t^{-1} + Km\left(C_1t^{-\theta} + C_2t^{-1-0.5\tilde{\mu}}\right) \leq -C_5t^{-1}$. To sum up, for any $\epsilon_0, \epsilon_2 > 0$, there exists $t_3 > \max\{t_1, t_2, \tilde{t}^*\}$ such that if $\left\|P\boldsymbol{r}_0^{(t)}\right\| \geq \epsilon_0 + \epsilon_1$ and $S^{((t))} \neq \emptyset$, then

$$\forall m \in [0 : M-1] : \sum_{u=t}^{t+m} \sum_{v=0}^{K-1} (\boldsymbol{r}_{v+1}^{(u)} - \boldsymbol{r}_v^{(u)})^\top \boldsymbol{r}_v^{(u)} \leq -C_5 t^{-1},$$

Now, define two sets for each $k \in [0 : K-1]$

$$\mathcal{T}_k := \{t > t_3 : \left\|P\boldsymbol{r}_k^{(t)}\right\| < \epsilon_0 + \epsilon_1\}$$

$$\bar{\mathcal{T}}_k := \{t > t_3 : \left\|P\boldsymbol{r}_k^{(t)}\right\| \geq \epsilon_0 + \epsilon_1\}$$

We will finish our proof by showing $\bar{\mathcal{T}}_k$ is finite.

First, every $\mathcal{T}_k$ is not empty nor finite. If there exists some k' that $\mathcal{T}_k'$ is empty or finite, then $\exists t_{\max} \in \bar{\mathcal{T}}_k'$. Then

$$\begin{aligned}
\left\|P\boldsymbol{r}_0^{(t)}\right\|^2 - \left\|P\boldsymbol{r}_0^{(t_{\max})}\right\|^2 &= \left\|\boldsymbol{r}_0^{(t)}\right\|^2 - \left\|\boldsymbol{r}_0^{(t_{\max})}\right\|^2 \\
&= \sum_{u=t_{\max}}^{t-1} \sum_{k=0}^{K-1} \left[\left\|\boldsymbol{r}_{k+1}^{(u)}\right\|^2 - \left\|\boldsymbol{r}_k^{(u)}\right\|^2\right] \\
&= \sum_{u=t_{\max}}^{t-1} \sum_{k=0}^{K-1} \left[\left\|\boldsymbol{r}_{k+1}^{(u)} - \boldsymbol{r}_k^{(u)}\right\|^2\right] + 2 \sum_{u=t_{\max}}^{t-1} \sum_{k=0}^{K-1} (\boldsymbol{r}_{k+1}^{(u)} - \boldsymbol{r}_k^{(u)})^\top \boldsymbol{r}_k^{(u)} \\
&\leq C_4 + 2 \sum_{u=t_{\max}}^{t-1} \left(\sum_{k \neq k'} (\boldsymbol{r}_{k+1}^{(u)} - \boldsymbol{r}_k^{(u)})^\top \boldsymbol{r}_k^{(u)} + (\boldsymbol{r}_{k'+1}^{(u)} - \boldsymbol{r}_{k'}^{(u)})^\top \boldsymbol{r}_{k'}^{(u)} \right) \\
&\leq C_4 + C_6 + 2 \sum_{\substack{t_{\max} \leq u \leq t-1 \\ S^{(u)} \neq \emptyset}} (\boldsymbol{r}_{k'+1}^{(u)} - \boldsymbol{r}_{k'}^{(u)})^\top \boldsymbol{r}_{k'}^{(u)} \\
&\leq C_4 + C_6 - 2C_3 \sum_{\substack{t_{\max} \leq u \leq t-1 \\ S^{(u)} \neq \emptyset}} u^{-1}
\end{aligned}$$

The first inequality is true by Equation (18). Other inequalities hold by Lemma D.5. As t goes infinity, the upper bound goes to negative infinity. However, it contradicts to the fact that $\left\|\boldsymbol{r}_0^{(t)}\right\|$ is bounded.

Before we move on the final step, note that $\lim_{T \to \infty} \sum_{t=t_1}^{T} \sum_{k=0}^{K-1} \left\|\boldsymbol{r}_{k+1}^{(t)} - \boldsymbol{r}_k^{(t)}\right\|^2 = C_4$ implies

$$\sum_{u=t_1}^{t} \sum_{k=0}^{K-1} \left\|\boldsymbol{r}_{k+1}^{(u)} - \boldsymbol{r}_k^{(u)}\right\|^2 = C_4 - h(t)$$

where $h(t)$ is a positive function monotonic decreasing to zero.
Now, assume that there exists some k' that $\bar{\mathcal{T}}_k$ is infinite. WLOG, we set $k' = 0$. Since $\mathcal{T}_0$ is infinite, for any $t \in \bar{\mathcal{T}}_0$ there exists $t', t'' \in \mathcal{T}_0$ such that $t \in [t'+1, t''-1] \subset \bar{\mathcal{T}}_0$. We divide it into two cases: For all $t \in [t'+1, t''-1]$,

1. if $|[t'+1, t]| < M$, then $\left\|P\boldsymbol{r}_0^{(t)}\right\|^2 \leq \left\|P\boldsymbol{r}_0^{(t')}\right\|^2 + M\epsilon_0 \leq (M+1)\epsilon_0 + \epsilon_1$.

2. if $|[t'+1, t]| \geq M$, let $t^* = \min\{u \in [t'+1, t] : S^{(u)} \neq \emptyset\}$. Then

$$\left\|P\boldsymbol{r}_0^{(t)}\right\|^2 = \left\|P\boldsymbol{r}_0^{(t^*)}\right\|^2 + \sum_{u=t^*}^{t-1} \sum_{k=0}^{K-1} \left[\left\|\boldsymbol{r}_{k+1}^{(u)}\right\|^2 - \left\|\boldsymbol{r}_k^{(u)}\right\|^2\right]$$

$$
\begin{aligned}
&= \left\| P\boldsymbol{r}_0^{(t^*)} \right\|^2 + \sum_{u=t^*}^{t-1} \sum_{k=0}^{K-1} \left[\left\| \boldsymbol{r}_{k+1}^{(u)} - \boldsymbol{r}_k^{(u)} \right\|^2 + 2(\boldsymbol{r}_{k+1}^{(u)} - \boldsymbol{r}_k^{(u)})^\top \boldsymbol{r}_k^{(u)} \right] \\
&= \left\| P\boldsymbol{r}_0^{(t^*)} \right\|^2 + h(t) - h(t^*) + 2 \sum_{u=t^*}^{t-1} \sum_{k=0}^{K-1} \left[(\boldsymbol{r}_{k+1}^{(u)} - \boldsymbol{r}_k^{(u)})^\top \boldsymbol{r}_k^{(u)} \right] \\
&\le (M\epsilon_0 + \epsilon_0 + \epsilon_1)^2 + h(t) - 2C_5 \sum_{u=0}^{\lfloor \frac{t-1-t^*}{M} \rfloor} \frac{1}{Mu + t^*} \\
&\le (M\epsilon_0 + \epsilon_0 + \epsilon_1)^2 + h(t)
\end{aligned}
$$

Since $h(t)$ is monotonic decreasing function, for any $\epsilon_2 > 0$, there exists t_4 such that $\forall t \ge t_4 : h(t) < \epsilon_2$.

Therefore, $\forall t \ge \max\{t_3, t_4\} : \left\| P\boldsymbol{r}_0^{(t)} \right\|^2 \le (M\epsilon_0 + \epsilon_0 + \epsilon_1)^2 + \epsilon_2$. Since it holds for any $\epsilon_0, \epsilon_1, \epsilon_2$, it contradicts with the assumption that $\bar{\mathcal{T}}_0$ is infinite. □

D.3 Non-asymptotic Loss Convergence Analysis (Proof of Theorem 3.3)

In this section, we show non-asymptotic loss convergence, as stated below:

Theorem 3.3. *Under the same setting as Theorem 3.1 with an additional Assumption 3.5, for any $m \in [0 : M-1]$ and $k \in [0 : K-1]$, we have*

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{w}_k^{(MJ+m)}) \le{}& \left(|S| + \frac{\sum_{i=0}^{m-1} |S_i| + \frac{k}{K} |S_m|}{J} \right) \ell(\ln J) + \frac{\left\| \boldsymbol{w}_0^{(0)} - \hat{\boldsymbol{w}} \ln J \right\|^2}{2\eta K J} + \frac{D_1}{J} \\
&+ \left(|I| - |S| + \frac{\sum_{i=0}^{m-1} (|I_i| - |S_i|) + \frac{k}{K}(|I_m| - |S_m|)}{J} \right) \ell(\theta \ln J),
\end{aligned}
$$

where $\theta > 1$ is the second margin defined in Section 3.1, and

$$
D_1 := \frac{4\sigma_{\max}^2}{\phi^2} \left(\mathcal{L}(\boldsymbol{w}_0^{(0)}) + \left(1 + \frac{\eta K \sigma_{\max}^3 \beta}{\phi(1 - \eta M K \sigma_{\max}^2 \beta)} \right) \frac{\eta K \sigma_{\max}}{\phi(1 - \eta M K \sigma_{\max}^2 \beta)} \left\| \nabla \mathcal{L}(\boldsymbol{w}_0^{(0)}) \right\|^2 \right).
$$

Three major lemmas are used to prove Theorem 3.3. The first lemma is an extension of Lemma D.1. When M tasks are given cyclic, the following lemma holds.

Lemma D.7. *Let $t \in \mathbb{N}, l \in [0 : K-1]$, $m \in [0 : M-1]$ and $k \in [0 : K-1]$. If $m = 0$, then $l \ge k$. If $m = M$, then $l \le k$. For any t, l, m, k satisfying the condition,*

$$
\begin{aligned}
\left\| \boldsymbol{w}_l^{(t+m)} - \boldsymbol{w}_k^{(t)} + \eta \left((K - k + 1) \nabla \mathcal{L}^{(t)}(\boldsymbol{w}_k^{(t)}) K \sum_{i=1}^{m-1} \nabla \mathcal{L}^{(t+i)}(\boldsymbol{w}_k^{(t)}) + l \nabla \mathcal{L}^{(t+m)}(\boldsymbol{w}_k^{(t)}) \right) \right\| \\
\le \frac{\eta^2 (mK + l - k) K \sigma_{\max}^3 \beta}{\phi\{1 - \eta(mK + l - k)\sigma_{\max}^2 \beta\}} \left\| \nabla \mathcal{L}(\boldsymbol{w}_k^{(t)}) \right\|, \\
\left\| \boldsymbol{w}_l^{(t+m)} - \boldsymbol{w}_k^{(t)} \right\| \le \frac{\eta K \sigma_{\max}}{\phi\{1 - \eta(mK + l - k)\sigma_{\max}^2 \beta\}} \left\| \nabla \mathcal{L}(\boldsymbol{w}_k^{(t)}) \right\|, \\
\left\| \nabla \mathcal{L}(\boldsymbol{w}_l^{(t+m)}) - \nabla \mathcal{L}(\boldsymbol{w}_k^{(t)}) \right\| \le \frac{\eta K \sigma_{\max}^3 \beta}{\phi\{1 - \eta(mK + l - k)\sigma_{\max}^2 \beta\}} \left\| \nabla \mathcal{L}(\boldsymbol{w}_k^{(l)}) \right\|.
\end{aligned}
$$

Proof. We omitted the proof since there are only a few changes from the proof of Appendix D.1.1. □

The second and third lemmas represent two similar versions with respect to the common Gradient Descent setting, and Continual Learning setting.

Lemma D.8. *Suppose $\mathcal{L}$ is convex, and there exists $\beta \geq 0$ so that $1 - \eta\beta \geq 0$ and weights $(\boldsymbol{w}_0, \dots, \boldsymbol{w}_t)$ by $\boldsymbol{w}_{j+1} := \boldsymbol{w}_j - \eta\nabla\mathcal{L}(\boldsymbol{w}_j)$ satisfy*

$$\mathcal{L}(\boldsymbol{w}_{j+1}) \leq \mathcal{L}(\boldsymbol{w}_j) - \eta\left(1 - \eta\beta\right)\left\|\nabla\mathcal{L}(\boldsymbol{w}_j)\right\|^2$$

Then for any $\boldsymbol{z} \in \mathbb{R}^d$,

$$2\sum_{j=0}^{t-1}\eta\left(\mathcal{L}(\boldsymbol{w}_j) - \mathcal{L}(\boldsymbol{z})\right) - \sum_{j=0}^{t-1}\frac{\eta}{1-\eta\beta}\left(\mathcal{L}(\boldsymbol{w}_j) - \mathcal{L}(\boldsymbol{w}_{j+1})\right) \leq \left\|\boldsymbol{w}_0 - \boldsymbol{z}\right\|^2 - \left\|\boldsymbol{w}_t - \boldsymbol{z}\right\|^2.$$

Proof. See Appendix D.3.1. □

Lemma D.9. *Suppose $\mathcal{L}$ is convex, $\sigma_{\max}^2\beta$-smooth function and there exists $\beta' \geq 0$ so that $\eta \leq \min\{\frac{1}{2MK\sigma_{\max}^2\beta}, \frac{1}{2K\beta'}\}$ and weights $(\boldsymbol{w}_0^{(0)}, \dots, \boldsymbol{w}_{K-1}^{(0)}, \boldsymbol{w}_0^{(1)}, \dots, \boldsymbol{w}_{K-1}^{(MJ+M)})$ by $\boldsymbol{w}_{q+1}^{(p)} := \boldsymbol{w}_q^{(p)} - \eta\nabla\mathcal{L}^{(p)}(\boldsymbol{w}_q^{(p)})$, $\boldsymbol{w}_0^{(p+1)} := \boldsymbol{w}_K^{(p)}$ satisfy, for all $m \in [0 : M-1]$ and $k \in [0 : K-1]$,*

$$\mathcal{L}(\boldsymbol{w}_k^{(Mj+M+m)}) \leq \mathcal{L}(\boldsymbol{w}_k^{(Mj+m)}) - \eta K\left(1 - \eta K\beta'\right)\left\|\nabla\mathcal{L}(\boldsymbol{w}_k^{(Mj+m)})\right\|^2.$$

Then for any $\boldsymbol{z} \in \mathbb{R}^d$,

$$\begin{aligned}
&2\sum_{j=0}^{J-1}\eta K\left(\mathcal{L}(\boldsymbol{w}_k^{(Mj+M+m)}) - \mathcal{L}(\boldsymbol{z})\right)\\
&- \frac{2\eta MK\sigma_{\max}^4\beta}{\phi^2(1-\eta MK\sigma_{\max}^2\beta)^2}\sum_{j=0}^{J-1}\frac{\eta K}{1-\eta K\beta'}\left(\mathcal{L}(\boldsymbol{w}_k^{(Mj+m)}) - \mathcal{L}(\boldsymbol{w}_k^{(Mj+M+m)})\right)\\
&\leq \left\|\boldsymbol{w}_k^{(m)} - \boldsymbol{z}\right\|^2 - \left\|\boldsymbol{w}_k^{(MJ+m)} - \boldsymbol{z}\right\|^2.
\end{aligned}$$

Proof. See Appendix D.3.2. □

Note that Lemma D.9 holds only when jointly separable tasks are given cyclic, while Lemma D.8 always holds.

We follow the process of Appendix D.1 to show that it satisfies the condition in Lemma D.9. Since $\mathcal{L}$ is a $\sigma_{\max}^2\beta$-smooth function, For all $j \in [0 : J-1], m \in [0 : M-1], k \in [0 : K-1]$ we get

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{w}_k^{(Mj+M+m)}) - \mathcal{L}(\boldsymbol{w}_k^{(Mj+m)}) - \frac{\sigma_{\max}^2\beta}{2}\left\|\boldsymbol{w}_k^{(Mj+M+m)} - \boldsymbol{w}_k^{(Mj+m)}\right\|^2\\
&\leq \nabla\mathcal{L}(\boldsymbol{w}_k^{(Mj+m)})^\top(\boldsymbol{w}_k^{(Mj+M+m)} - \boldsymbol{w}_k^{(Mj+m)})\\
&= \nabla\mathcal{L}(\boldsymbol{w}_k^{(Mj+m)})^\top(\boldsymbol{w}_k^{(Mj+M+m)} - \boldsymbol{w}_k^{(Mj+m)} - \eta K\nabla\mathcal{L}(\boldsymbol{w}_k^{(Mj+m)}) + \eta K\nabla\mathcal{L}(\boldsymbol{w}_k^{(Mj+m)}))\\
&\leq -\eta K\left\|\nabla\mathcal{L}(\boldsymbol{w}_k^{(Mj+m)})\right\|^2 + \left\|\nabla\mathcal{L}(\boldsymbol{w}_k^{(Mj+m)})\right\|\left\|\boldsymbol{w}_k^{(Mj+M+m)} - \boldsymbol{w}_k^{(Mj+m)} + \eta K\nabla\mathcal{L}(\boldsymbol{w}_k^{(Mj+m)})\right\|.
\end{aligned}$$

By Lemma D.7,

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{w}_k^{(Mj+M+m)}) - \mathcal{L}(\boldsymbol{w}_k^{(Mj+m)}) - \frac{\sigma_{\max}^2\beta}{2}\cdot\frac{(\eta\sigma_{\max}K)^2}{\phi^2(1-\eta MK\sigma_{\max}^2\beta)^2}\left\|\nabla\mathcal{L}(\boldsymbol{w}_k^{(Mj+m)})\right\|^2\\
&\leq -\eta K\left\|\nabla\mathcal{L}(\boldsymbol{w}_k^{(Mj+m)})\right\|^2 + \frac{\eta^2MK^2\sigma_{\max}^3\beta}{\phi(1-\eta MK\sigma_{\max}^2\beta)}\left\|\nabla\mathcal{L}(\boldsymbol{w}_k^{(Mj+m)})\right\|^2.
\end{aligned}$$

Given that $\eta \leq \frac{1}{2MK\sigma_{\max}^2\beta}$,

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{w}_k^{(Mj+M+m)}) - \mathcal{L}(\boldsymbol{w}_k^{(Mj+m)})\\
&\leq -\eta K\{1 - \eta K\left(\frac{M\sigma_{\max}^3\beta}{\phi(1-\eta MK\sigma_{\max}^2\beta)} + \frac{\sigma_{\max}^4\beta}{2\phi^2(1-\eta MK\sigma_{\max}^2\beta)^2}\right)\}\left\|\nabla\mathcal{L}(\boldsymbol{w}_k^{(Mj+m)})\right\|^2
\end{aligned}$$

$$\begin{aligned}
&\le -\eta K\left(1-\eta K\frac{2(M\phi+\sigma_{\max})\sigma_{\max}^3\beta}{\phi^2}\right)\left\|\nabla\mathcal{L}(\boldsymbol{w}_k^{(Mj+m)})\right\|^2\\
&= -\eta K\left(1-\eta K\beta'\right)\left\|\nabla\mathcal{L}(\boldsymbol{w}_k^{(Mj+m)})\right\|^2,
\end{aligned}\tag{26}$$

where we set $\beta' := \frac{2(M\phi+\sigma_{\max})\sigma_{\max}^3\beta}{\phi^2}$.

Since Equation (26) holds for all $j\in[0:J-1], m\in[0:M-1], k\in[0:K-1]$ and $\eta<1/2K\beta'$ is given, by Lemma D.9, we get

$$\begin{aligned}
&2\sum_{j=0}^{J-1}\eta K\left(\mathcal{L}(\boldsymbol{w}_k^{(Mj+M+m)})-\mathcal{L}(\boldsymbol{z})\right)\\
&\quad-\frac{2\eta MK\sigma_{\max}^4\beta}{\phi^2(1-\eta MK\sigma_{\max}^2\beta)^2}\sum_{j=0}^{J-1}\frac{\eta K}{1-\eta K\beta'}\left(\mathcal{L}(\boldsymbol{w}_k^{(Mj+m)})-\mathcal{L}(\boldsymbol{w}_k^{(Mj+M+m)})\right)\\
&\le\left\|\boldsymbol{w}_k^{(m)}-\boldsymbol{z}\right\|^2-\left\|\boldsymbol{w}_k^{(MJ+m)}-\boldsymbol{z}\right\|^2.
\end{aligned}\tag{27}$$

Given that $\eta<\min\{\frac{1}{2MK\beta\sigma_{\max}^2},\frac{1}{2K\beta'}\}$ and $\mathcal{L}(\boldsymbol{w}_k^{(Mj+m)})$ is decreasing,

$$\begin{aligned}
&\frac{2\eta MK\sigma_{\max}^4\beta}{\phi^2(1-\eta MK\sigma_{\max}^2\beta)^2}\cdot\frac{\eta K}{1-\eta K\beta'}\left(\mathcal{L}(\boldsymbol{w}_k^{(Mj+m)})-\mathcal{L}(\boldsymbol{w}_k^{(Mj+M+m)})\right)\\
&\qquad\qquad\le\frac{8\sigma_{\max}^2}{\phi^2}\eta K\left(\mathcal{L}(\boldsymbol{w}_k^{(Mj+m)})-\mathcal{L}(\boldsymbol{w}_k^{(Mj+M+m)})\right).
\end{aligned}\tag{28}$$

Also,

$$\begin{aligned}
&\frac{8\sigma_{\max}^2}{\phi^2}\eta K\mathcal{L}(\boldsymbol{w}_k^{(MJ+m)})+2\eta KJ\mathcal{L}(\boldsymbol{w}_k^{(MJ+m)})-\frac{8\sigma_{\max}^2}{\phi^2}\eta K\mathcal{L}(\boldsymbol{w}_k^{(m)})\\
&\le\frac{8\sigma_{\max}^2}{\phi^2}\eta K\mathcal{L}(\boldsymbol{w}_k^{(MJ+m)})+2\eta K\sum_{j=1}^{J}\mathcal{L}(\boldsymbol{w}_k^{(Mj+m)})-\frac{8\sigma_{\max}^2}{\phi^2}\eta K\mathcal{L}(\boldsymbol{w}_k^{(m)})\\
&=2\eta K\sum_{j=0}^{J-1}\mathcal{L}(\boldsymbol{w}_k^{(Mj+M+m)})-\frac{8\sigma_{\max}^2}{\phi^2}\eta K\sum_{j=0}^{J-1}\left(\mathcal{L}(\boldsymbol{w}_k^{(Mj+m)})-\mathcal{L}(\boldsymbol{w}_k^{(Mj+M+m)})\right).
\end{aligned}\tag{29}$$

Combine the result (27), (28) and (29), we obtain

$$\begin{aligned}
&2\eta KJ\left(\mathcal{L}(\boldsymbol{w}_k^{(MJ+m)})-\mathcal{L}(\boldsymbol{z})\right)+\frac{8\sigma_{\max}^2}{\phi^2}\eta K\left(\mathcal{L}(\boldsymbol{w}_k^{(MJ+m)})-\mathcal{L}(\boldsymbol{w}_k^{(m)})\right)\\
&\le2\sum_{j=0}^{J-1}\eta K\left(\mathcal{L}(\boldsymbol{w}_k^{(Mj+M+m)})-\mathcal{L}(\boldsymbol{z})\right)-\frac{8\sigma_{\max}^2}{\phi^2}\eta K\sum_{j=0}^{J-1}\left(\mathcal{L}(\boldsymbol{w}_k^{(Mj+m)})-\mathcal{L}(\boldsymbol{w}_k^{(Mj+M+m)})\right)\\
&\le\left\|\boldsymbol{w}_k^{(m)}-\boldsymbol{z}\right\|^2-\left\|\boldsymbol{w}_k^{(MJ+m)}-\boldsymbol{z}\right\|^2.
\end{aligned}\tag{30}$$

Now we examine the loss change in a cycle. For any $j\in[0:M-1], l\in[0:K-1]$,

$$\mathcal{L}_j(\boldsymbol{w}_{l+1}^{(j)})\le\mathcal{L}_j(\boldsymbol{w}_l^{(j)})-\eta(1-\frac{\eta\sigma_{\max}^2\beta}{2})\left\|\nabla\mathcal{L}_j(\boldsymbol{w}_l^{(j)})\right\|^2.$$

Since $\eta<\frac{1}{2MK\beta\sigma_{\max}^2}$, $\mathcal{L}_j(\boldsymbol{w}_l^{(j)})$ decreases. Therefore for any $p\in[0:M-1], q\in[0:K-1]$,

$$\begin{aligned}
2\eta q(\mathcal{L}_p(\boldsymbol{w}_{q+1}^{(p)})-\mathcal{L}_p(\boldsymbol{z}))&\le2\sum_{l=1}^{q}\eta\left(\mathcal{L}_p(\boldsymbol{w}_l^{(p)})-\mathcal{L}_p(\boldsymbol{z})\right)\\
&=2\sum_{l=0}^{q-1}\eta\left(\mathcal{L}_p(\boldsymbol{w}_l^{(p)})-\mathcal{L}_p(\boldsymbol{z})\right)+2\sum_{l=0}^{q-1}\eta\left(\mathcal{L}_p(\boldsymbol{w}_{l+1}^{(p)})-\mathcal{L}_p(\boldsymbol{w}_l^{(p)})\right)
\end{aligned}$$

$$
\begin{aligned}
&\le 2\sum_{l=0}^{q-1}\eta\left(\mathcal{L}_p(\boldsymbol{w}_l^{(p)})-\mathcal{L}_p(\boldsymbol{z})\right)-\sum_{l=0}^{t-1}\frac{\eta}{1-\eta\beta}\left(\mathcal{L}_p(\boldsymbol{w}_l^{(p)})-\mathcal{L}_p(\boldsymbol{w}_{l+1}^{(p)})\right)\\
&\le\left\|\boldsymbol{w}_0^{(p)}-\boldsymbol{z}\right\|^2-\left\|\boldsymbol{w}_q^{(p)}-\boldsymbol{z}\right\|^2,
\end{aligned}
$$

where in the third line, we use $\frac{1}{1-\eta\beta}<2$, and in the last line, we use Lemma D.8

By summing up, we obtain

$$
\sum_{p=0}^{m-1}2\eta K\left(\mathcal{L}_p(\boldsymbol{w}_K^{(p)})-\mathcal{L}_p(\boldsymbol{z})\right)+2\eta k\left(\mathcal{L}_m(\boldsymbol{w}_k^{(m)})-\mathcal{L}_m(\boldsymbol{z})\right)\le\left\|\boldsymbol{w}_0^{(0)}-\boldsymbol{z}\right\|^2-\left\|\boldsymbol{w}_k^{(m)}-\boldsymbol{z}\right\|^2. \tag{31}
$$

At last, $\mathcal{L}(\boldsymbol{w}_k^{(m)})$ is bounded by $\mathcal{L}(\boldsymbol{w}_0^{(0)})$ as follows:

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{w}_k^{(m)})-\mathcal{L}(\boldsymbol{w}_0^{(0)})&\le\nabla\mathcal{L}(\boldsymbol{w}_0^{(0)})^\top\left(\boldsymbol{w}_k^{(m)}-\boldsymbol{w}_0^{(0)}\right)+\frac{\sigma_{\max}^2\beta}{2}\left\|\boldsymbol{w}_k^{(m)}-\boldsymbol{w}_0^{(0)}\right\|^2\\
&\le\left(\left\|\nabla\mathcal{L}(\boldsymbol{w}_0^{(0)})\right\|+\frac{\sigma_{\max}^2\beta}{2}\left\|\boldsymbol{w}_k^{(m)}-\boldsymbol{w}_0^{(0)}\right\|\right)\left\|\boldsymbol{w}_k^{(m)}-\boldsymbol{w}_0^{(0)}\right\|\\
&\le\left(1+\frac{\eta K\sigma_{\max}^3\beta}{\phi(1-\eta MK\sigma_{\max}^2\beta)}\right)\frac{\eta K\sigma_{\max}}{\phi(1-\eta MK\sigma_{\max}^2\beta)}\left\|\nabla\mathcal{L}(\boldsymbol{w}_0^{(0)})\right\|^2\\
&=D_0\left\|\nabla\mathcal{L}(\boldsymbol{w}_0^{(0)})\right\|^2,
\end{aligned}
\tag{32}
$$

where in the first inequality we use smoothness, and in the second inequality we use Cauchy-Schwarz, and in the last line we use the fact $\left\|\boldsymbol{w}_k^{(m)}-\boldsymbol{w}_0^{(0)}\right\|\le\frac{\eta K\sigma_{\max}}{\phi(1-\eta MK\sigma_{\max}^2\beta)}\left\|\nabla\mathcal{L}(\boldsymbol{w}_0^{(0)})\right\|$ held by Lemma D.7. We set $D_0:=\left(1+\frac{\eta K\sigma_{\max}^3\beta}{\phi(1-\eta MK\sigma_{\max}^2\beta)}\right)\frac{\eta K\sigma_{\max}}{\phi(1-\eta MK\sigma_{\max}^2\beta)}$.

Combine the result (30),(31),(32), we obtain

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{w}_k^{(MJ+m)})\le&\mathcal{L}(z)+\frac{1}{J}\sum_{p=0}^{m-1}\mathcal{L}_p(z)+\frac{1}{J}\cdot\frac{q}{K}\mathcal{L}_m(z)\\
&+\frac{4\sigma_{\max}^2}{\phi^2}\frac{\mathcal{L}(w_0^{(0)})+D_0\left\|\nabla\mathcal{L}(\boldsymbol{w}_0^{(0)})\right\|^2}{J}+\frac{\left\|\boldsymbol{w}_0^{(0)}-\boldsymbol{z}\right\|^2}{2\eta KJ}.
\end{aligned}
$$

Let $\boldsymbol{z}:=\hat{\boldsymbol{w}}\log J$. Using $\mathcal{L}_j(\hat{\boldsymbol{w}}\log J)\le|S_j|\,\ell(\log J)+(|I_j|-|S_j|)\,\ell(\theta\log J)$, we can finish the proof.

D.3.1 Proof of Lemma D.8

This is a well-known property about gradient descent applied to a smooth convex objective function. We contain the proof for completeness.

Suppose $\mathcal{L}$ is convex, and there exists $\beta\ge0$ so that $1-\eta\beta\ge0$ and weights $(\boldsymbol{w}_0,\dots,\boldsymbol{w}_t)$ by $\boldsymbol{w}_{j+1}:=\boldsymbol{w}_j-\eta\nabla\mathcal{L}(\boldsymbol{w}_j)$ satisfy

$$
\mathcal{L}(\boldsymbol{w}_{j+1})\le\mathcal{L}(\boldsymbol{w}_j)-\eta\left(1-\eta\beta\right)\left\|\nabla\mathcal{L}(\boldsymbol{w}_j)\right\|^2.
$$

For any j and $\boldsymbol{z}\in\mathbb{R}^d$,

$$
\begin{aligned}
\left\|\boldsymbol{w}_{j+1}-\boldsymbol{z}\right\|^2&=\left\|\boldsymbol{w}_j-\boldsymbol{z}\right\|^2+2\eta\langle\nabla\mathcal{L}(\boldsymbol{w}_j),\boldsymbol{z}-\boldsymbol{w}_j\rangle+\eta^2\left\|\nabla\mathcal{L}(\boldsymbol{w}_j)\right\|^2\\
&\le\left\|\boldsymbol{w}_j\right\|^2+2\eta(\mathcal{L}(z)-\mathcal{L}(\boldsymbol{w}_j))+\eta^2\left\|\nabla\mathcal{L}(\boldsymbol{w}_j)\right\|^2\\
&\le\left\|\boldsymbol{w}_j\right\|^2+2\eta(\mathcal{L}(z)-\mathcal{L}(\boldsymbol{w}_j))+\frac{\eta}{1-\eta\beta}\left(\mathcal{L}(\boldsymbol{w}_j)-\mathcal{L}(\boldsymbol{w}_{j+1})\right).
\end{aligned}
$$

where the first line comes from convexity and the second line comes from the condition. By adding all $j\in\{0,\cdots,t-1\}$, we get

$$
2\sum_{j=0}^{t-1}\eta\left(\mathcal{L}(\boldsymbol{w}_j)-\mathcal{L}(\boldsymbol{z})\right)-\sum_{j=0}^{t-1}\frac{\eta}{1-\eta\beta}\left(\mathcal{L}(\boldsymbol{w}_j)-\mathcal{L}(\boldsymbol{w}_{j+1})\right)\le\left\|\boldsymbol{w}_0-\boldsymbol{z}\right\|^2-\left\|\boldsymbol{w}_t-\boldsymbol{z}\right\|^2.
$$

D.3.2 Proof of Lemma D.9

Without loss of generality, we assume $m = 0, k = 0$.

Suppose $\mathcal{L}$ is convex, $\sigma_{\max}^2\beta$-smooth function and there exists $\beta' \geq 0$ so that $\eta \leq \min\{\frac{1}{2MK\sigma_{\max}^2\beta}, \frac{1}{2K\beta'}\}$ and weights $(\boldsymbol{w}_0^{(0)}, \ldots, \boldsymbol{w}_{K-1}^{(0)}, \boldsymbol{w}_0^{(1)}, \ldots, \boldsymbol{w}_{K-1}^{(Mt+M)})$ by $\boldsymbol{w}_{q+1}^{(p)} := \boldsymbol{w}_q^{(p)} - \eta \nabla \mathcal{L}^{(p)}(\boldsymbol{w}_q^{(p)})$, $\boldsymbol{w}_0^{(p+1)} := \boldsymbol{w}_K^{(p)}$ satisfy

$$\mathcal{L}(\boldsymbol{w}_0^{(Mj+M)}) \leq \mathcal{L}(\boldsymbol{w}_0^{(Mj)}) - \eta K\left(1 - \eta K \beta'\right) \left\| \nabla \mathcal{L}(\boldsymbol{w}_0^{(Mj)}) \right\|^2 .$$

For any j and $\boldsymbol{z} \in \mathbb{R}^d$,

$$\begin{aligned}
&\left\| \boldsymbol{w}_0^{(Mj+M)} - \boldsymbol{z} \right\|^2 = \left\| \boldsymbol{w}_0^{(Mj)} - \eta \sum_{p=0}^{M-1} \sum_{q=0}^{K-1} \nabla \mathcal{L}_p(\boldsymbol{w}_q^{(Mj+p)}) - z \right\|^2 \\
&= \left\| \boldsymbol{w}_0^{(Mj)} - \boldsymbol{z} \right\|^2 + 2\eta \sum_{p=0}^{M-1} \sum_{q=0}^{K-1} \langle \nabla \mathcal{L}_p(\boldsymbol{w}_q^{(Mj+p)}), \boldsymbol{z} - \boldsymbol{w}_0^{(Mj)} \rangle + \eta^2 \left\| \sum_{p=0}^{M-1} \sum_{q=0}^{K-1} \nabla \mathcal{L}_p(\boldsymbol{w}_q^{(Mj+p)}) \right\|^2 \\
&= \left\| \boldsymbol{w}_0^{(Mj)} - \boldsymbol{z} \right\|^2 + 2\eta \sum_{p=0}^{M-1} \sum_{q=0}^{K-1} \langle \nabla \mathcal{L}_p(\boldsymbol{w}_q^{(Mj+p)}), \boldsymbol{z} - \boldsymbol{w}_q^{(Mj+p)} \rangle \\
&+ 2\eta \sum_{p=0}^{M-1} \sum_{q=0}^{K-1} \langle \nabla \mathcal{L}_p(\boldsymbol{w}_q^{(Mj+p)}), \boldsymbol{w}_q^{(Mj+p)} - \boldsymbol{w}_0^{(Mj)} \rangle + \eta^2 \left\| \sum_{p=0}^{M-1} \sum_{q=0}^{K-1} \nabla \mathcal{L}_p(\boldsymbol{w}_q^{(Mj+p)}) \right\|^2 . \qquad (33)
\end{aligned}$$

By convexity,

$$\sum_{p=0}^{M-1} \sum_{q=0}^{K-1} \langle \nabla \mathcal{L}_p(\boldsymbol{w}_q^{(Mj+p)}), \boldsymbol{z} - \boldsymbol{w}_q^{(Mj+p)} \rangle \leq K\mathcal{L}(\boldsymbol{z}) - \sum_{p=0}^{M-1} \sum_{q=0}^{K-1} \mathcal{L}_p(\boldsymbol{w}_q^{(Mj+p)}).$$

Apply smoothness on (33), we get

$$\begin{aligned}
&\left\| \boldsymbol{w}_0^{(Mj+M)} - \boldsymbol{z} \right\|^2 - \left\| \boldsymbol{w}_0^{(Mj)} - \boldsymbol{z} \right\|^2 \leq 2\eta K \mathcal{L}(\boldsymbol{z}) - 2\eta \sum_{p=0}^{M-1} \sum_{q=0}^{K-1} \mathcal{L}_p(\boldsymbol{w}_q^{(Mj+p)}) \\
&+ 2\eta \sum_{p=0}^{M-1} \sum_{q=0}^{K-1} \langle \nabla \mathcal{L}_p(\boldsymbol{w}_q^{(Mj+p)}), \boldsymbol{w}_q^{(Mj+p)} - \boldsymbol{w}_0^{(Mj)} \rangle + \eta^2 \left\| \sum_{p=0}^{M-1} \sum_{q=0}^{K-1} \nabla \mathcal{L}_p(\boldsymbol{w}_q^{(Mj+p)}) \right\|^2 .
\end{aligned}$$

By $\sigma_{\max}^2\beta$-smoothness,

$$\begin{aligned}
&\sum_{p=0}^{M-1} \sum_{q=0}^{K-1} \langle \nabla \mathcal{L}_p(\boldsymbol{w}_q^{(Mj+p)}), \boldsymbol{z} - \boldsymbol{w}_q^{(Mj+p)} \rangle \\
&\qquad \geq K\mathcal{L}(\boldsymbol{z}) - \sum_{p=0}^{M-1} \sum_{q=0}^{K-1} \mathcal{L}_p(\boldsymbol{w}_q^{(Mj+p)}) - \frac{\sigma_{\max}^2\beta}{2} \sum_{p=0}^{M-1} \sum_{q=0}^{K-1} \left\| \boldsymbol{z} - \boldsymbol{w}_q^{(Mj+p)} \right\|^2 .
\end{aligned}$$

Apply smoothness on (33), and let $\boldsymbol{z} := \boldsymbol{w}_0^{(Mj+M)}$ then we get

$$\begin{aligned}
&0 \geq \left\| \boldsymbol{w}_0^{(Mj)} - \boldsymbol{w}_0^{(Mj+M)} \right\|^2 + 2\eta K \mathcal{L}(\boldsymbol{w}_0^{(Mj+M)}) \\
&- 2\eta \sum_{p=0}^{M-1} \sum_{q=0}^{K-1} \mathcal{L}_p(\boldsymbol{w}_q^{(Mj+p)}) - \eta\sigma_{\max}^2\beta \sum_{p=0}^{M-1} \sum_{q=0}^{K-1} \left\| \boldsymbol{w}_0^{(Mj+M)} - \boldsymbol{w}_q^{(Mj+p)} \right\|^2 \\
&+ 2\eta \sum_{p=0}^{M-1} \sum_{q=0}^{K-1} \langle \nabla \mathcal{L}_p(\boldsymbol{w}_q^{(Mj+p)}), \boldsymbol{w}_q^{(Mj+p)} - \boldsymbol{w}_0^{(Mj)} \rangle + \eta^2 \left\| \sum_{p=0}^{M-1} \sum_{q=0}^{K-1} \nabla \mathcal{L}_p(\boldsymbol{w}_q^{(Mj+p)}) \right\|^2 .
\end{aligned}$$

Combined with the smoothness result,

$$\begin{aligned}
&\left\|\boldsymbol{w}_0^{(Mj+M)} - \boldsymbol{z}\right\|^2 - \left\|\boldsymbol{w}_0^{(Mj)} - \boldsymbol{z}\right\|^2 \le 2\eta K \mathcal{L}(\boldsymbol{z}) - 2\eta K \mathcal{L}(\boldsymbol{w}_0^{(Mj+M)}) \\
&- \left\|\boldsymbol{w}_0^{(Mj)} - \boldsymbol{w}_0^{(Mj+M)}\right\|^2 + \eta \sigma_{\max}^2 \beta \sum_{p=0}^{M-1}\sum_{q=0}^{K-1} \left\|\boldsymbol{w}_0^{(Mj+M)} - \boldsymbol{w}_q^{(Mj+p)}\right\|^2 \\
&\le 2\eta K \mathcal{L}(\boldsymbol{z}) - 2\eta K \mathcal{L}(\boldsymbol{w}_0^{(Mj+M)}) - \left\|\boldsymbol{w}_0^{(Mj)} - \boldsymbol{w}_0^{(Mj+M)}\right\|^2 \\
&+ 2\eta M K \sigma_{\max}^2 \beta \left\|\boldsymbol{w}_0^{(Mj)} - \boldsymbol{w}_0^{(Mj+M)}\right\|^2 + 2\eta \sigma_{\max}^2 \beta \sum_{p=0}^{M-1}\sum_{q=0}^{K-1} \left\|\boldsymbol{w}_0^{(Mj)} - \boldsymbol{w}_q^{(Mj+p)}\right\|^2 \\
&\le 2\eta K \mathcal{L}(\boldsymbol{z}) - 2\eta K \mathcal{L}(\boldsymbol{w}_0^{(Mj+M)}) + 2\eta \sigma_{\max}^2 \beta \sum_{p=0}^{M-1}\sum_{q=0}^{K-1} \left\|\boldsymbol{w}_0^{(Mj)} - \boldsymbol{w}_q^{(Mj+p)}\right\|^2,
\end{aligned}$$

where in the last line we use $\eta \le \frac{1}{2MK\sigma_{\max}^2\beta}$. Finally, by Lemma D.7,

$$\begin{aligned}
&\left\|\boldsymbol{w}_0^{(Mj+M)} - \boldsymbol{z}\right\|^2 - \left\|\boldsymbol{w}_0^{(Mj)} - \boldsymbol{z}\right\|^2 \\
&\le 2\eta K \left(\mathcal{L}(\boldsymbol{z}) - \mathcal{L}(\boldsymbol{w}_0^{(Mj+M)})\right) + \frac{2\eta^3 M K^3 \sigma_{\max}^4 \beta}{\phi^2 (1 - \eta M K \sigma_{\max}^2 \beta)^2} \left\|\nabla \mathcal{L}(\boldsymbol{w}_0^{(Mj)})\right\|^2 \\
&\le 2\eta K \left(\mathcal{L}(\boldsymbol{z}) - \mathcal{L}(\boldsymbol{w}_0^{(Mj+M)})\right) - \frac{2\eta M K \sigma_{\max}^4 \beta}{\phi^2 (1 - \eta M K \sigma_{\max}^2 \beta)^2} \frac{\eta K}{1 - \eta K \beta'} \left(\mathcal{L}(\boldsymbol{w}_k^{(Mj+m)}) - \mathcal{L}(\boldsymbol{w}_k^{(Mj+M+m)})\right).
\end{aligned}$$

By adding all $j \in \{0, \cdots, J-1\}$, we can finish the proof.

D.4 Forgetting Analysis (Proof of Theorem 3.4)

We prove Theorem 3.4 here, which is restated for readability.

Theorem 3.4. *Let $\ell(u) = \ln(1 + e^{-u})$ be the logistic loss. If the learning rate satisfies $\eta < \min\left\{\frac{1}{2MK\beta\sigma_{\max}^2}, \frac{\phi^2}{4K\beta\sigma_{\max}^3(M\phi + \sigma_{\max})}\right\}$, then the cycle-averaged forgetting $\mathcal{CF}(J)$ for cycle J satisfies the following upper and lower bounds:*

$$-\eta K \cdot L(J)^2 \cdot \frac{\sum_{p\neq q} N_{p,q}}{M} \le \mathcal{CF}(J) \le \eta K \cdot L(J)^2 \cdot \frac{-\sum_{p\neq q} \bar{N}_{p,q}}{M},$$

where

$$\begin{aligned}
L(J) &:= \frac{1}{J}\left(\left(|S| + \frac{|I| - |S|}{J^{\theta-1}}\right)\left(1 + \frac{1}{J}\right) + \frac{\|\boldsymbol{w}_0^{(0)} - \hat{\boldsymbol{w}} \ln J\|^2}{2\eta K} + D_1\right) = \mathcal{O}\left(\frac{\ln^2 J}{J}\right) \\
N_{p,q} &:= \sum_{\substack{(i,j)\in I_p \times I_q \\ \boldsymbol{x}_i^\top \boldsymbol{x}_j > 0}} \boldsymbol{x}_i^\top \boldsymbol{x}_j > 0, \quad \bar{N}_{p,q} := \sum_{\substack{(i,j)\in I_p \times I_q \\ \boldsymbol{x}_i^\top \boldsymbol{x}_j < 0}} \boldsymbol{x}_i^\top \boldsymbol{x}_j < 0.
\end{aligned}$$

By Theorem 3.3, loss on cycle J is bounded as

$$\mathcal{L}(\boldsymbol{w}_k^{(MJ+m)}) \le L(J)$$

where

$$\begin{aligned}
L(J) &:= \frac{1}{J}\left(\left(|S| + \frac{|I| - |S|}{J^{\theta-1}}\right)(1 + \frac{1}{J}) + \frac{\left\|\boldsymbol{w}_0^{(0)} - \hat{\boldsymbol{w}} \log J\right\|^2}{2\eta K} + D_1\right), \\
D_1 &:= \frac{4\sigma_{\max}^2}{\phi^2}\left(\mathcal{L}(w_0^{(0)}) + D_0 \left\|\nabla \mathcal{L}(\boldsymbol{w}_0^{(0)})\right\|^2\right).
\end{aligned}$$

Therefore, the following holds:

$$\forall s \in I, \forall m \in [0:M-1], \forall k \in [0:K-1]: \quad \boldsymbol{x}_s^\top \boldsymbol{w}_k^{(MJ+m)} \geq \ell^{-1}\left(L(t)\right).$$

Now, we analyze the change of each task in one cycle. For upper bound,

$$\begin{aligned}
&\mathcal{L}_m(\boldsymbol{w}_0^{(MJ+M)}) - \mathcal{L}_m(\boldsymbol{w}_K^{(MJ+m)}) \\
&\quad \leq -\eta \sum_{p=m+1}^{M-1} \sum_{q=0}^{K-1} \nabla \mathcal{L}_m(\boldsymbol{w}_0^{(MJ+M)})^\top \nabla \mathcal{L}_p(\boldsymbol{w}_q^{(MJ+p)}) \qquad (34) \\
&\quad \leq -\eta \sum_{p=m+1}^{M-1} \sum_{q=0}^{K-1} \sum_{\substack{(i,j)\in I_m \times I_p \\ \boldsymbol{x}_i^\top \boldsymbol{x}_j < 0}} \ell'(\boldsymbol{x}_i^\top \boldsymbol{w}_0^{(MJ+M)}) \ell'(\boldsymbol{x}_j^\top \boldsymbol{w}_q^{(MJ+p)}) \boldsymbol{x}_i^\top \boldsymbol{x}_j \\
&\quad \leq -\eta \sum_{p=m+1}^{M-1} \sum_{q=0}^{K-1} \left[\ell'\left(\ell^{-1}\left(L(J)\right)\right)\right]^2 \sum_{\substack{(i,j)\in I_m \times I_p \\ \boldsymbol{x}_i^\top \boldsymbol{x}_j < 0}} \boldsymbol{x}_i^\top \boldsymbol{x}_j \qquad (35) \\
&\quad \leq -\eta K L(J)^2 \sum_{p=m+1}^{M-1} \sum_{\substack{(i,j)\in I_m \times I_p \\ \boldsymbol{x}_i^\top \boldsymbol{x}_j < 0}} \boldsymbol{x}_i^\top \boldsymbol{x}_j, \qquad (36)
\end{aligned}$$

where in (34) we use convexity, in (35) we use the condition that ℓ' is a negative function monotonically increasing to zero. (36) holds by the fact $\forall x : \ell'(x) = {\ell_{\log}}'(x) \geq -\exp(-x)$ and $\forall x : \ell^{-1}(x) = {\ell_{\log}}^{-1}(x) \geq -\log(x)$. Likewise, we can get a lower bound.

$$\begin{aligned}
&\mathcal{L}_m(\boldsymbol{w}_0^{(MJ+M)}) - \mathcal{L}_m(\boldsymbol{w}_K^{(MJ+m)}) \\
&\quad \geq -\eta \sum_{p=m+1}^{M-1} \sum_{q=0}^{K-1} \nabla \mathcal{L}_m(\boldsymbol{w}_k^{(MJ+m)})^\top \nabla \mathcal{L}_p(\boldsymbol{w}_q^{(MJ+p)}) \\
&\quad \geq -\eta \sum_{p=m+1}^{M-1} \sum_{q=0}^{K-1} \sum_{\substack{(i,j)\in I_m \times I_p \\ \boldsymbol{x}_i^\top \boldsymbol{x}_j > 0}} \ell'(\boldsymbol{x}_i^\top \boldsymbol{w}_k^{(MJ+m)}) \ell'(\boldsymbol{x}_j^\top \boldsymbol{w}_q^{(MJ+p)}) \boldsymbol{x}_i^\top \boldsymbol{x}_j \\
&\quad \geq -\eta \sum_{p=m+1}^{M-1} \sum_{q=0}^{K-1} \left[\ell'\left(\ell^{-1}\left(L(J)\right)\right)\right]^2 \sum_{\substack{(i,j)\in I_m \times I_p \\ \boldsymbol{x}_i^\top \boldsymbol{x}_j > 0}} \boldsymbol{x}_i^\top \boldsymbol{x}_j \\
&\quad \geq -\eta K L(J)^2 \sum_{p=m+1}^{M-1} \sum_{\substack{(i,j)\in I_m \times I_p \\ \boldsymbol{x}_i^\top \boldsymbol{x}_j > 0}} \boldsymbol{x}_i^\top \boldsymbol{x}_j.
\end{aligned}$$

Define

$$N_{p,q} := \sum_{\substack{(i,j)\in I_p \times I_q \\ \boldsymbol{x}_i^\top \boldsymbol{x}_j > 0}} \boldsymbol{x}_i^\top \boldsymbol{x}_j, \bar{N}_{p,q} := \sum_{\substack{(i,j)\in I_p \times I_q \\ \boldsymbol{x}_i^\top \boldsymbol{x}_j < 0}} \boldsymbol{x}_i^\top \boldsymbol{x}_j.$$

Since

$$\sum_{m=0}^{M-1} \sum_{p=m+1}^{M-1} \sum_{\substack{(i,j)\in I_m \times I_p \\ \boldsymbol{x}_i^\top \boldsymbol{x}_j > 0}} \boldsymbol{x}_i^\top \boldsymbol{x}_j = \sum_{p \neq q} N_{p,q},$$

$$\sum_{m=0}^{M-1} \sum_{p=m+1}^{M-1} \sum_{\substack{(i,j)\in I_m \times I_p \\ \boldsymbol{x}_i^\top \boldsymbol{x}_j < 0}} \boldsymbol{x}_i^\top \boldsymbol{x}_j = \sum_{p \neq q} \bar{N}_{p,q},$$

we can conclude

$$-\eta KL(J)^2 \cdot \frac{\sum_{p\neq q} N_{p,q}}{M} \leq \frac{1}{M}\sum_{m=0}^{M-1} \mathcal{F}^{(MJ+m)}(MJ+M) \leq -\eta KL(J)^2 \cdot \frac{\sum_{p\neq q} \bar{N}_{p,q}}{M}.$$

E Proofs for Section 4: Random Task Ordering, Jointly Separable

E.1 Asymptotic Loss Convergence Analysis (Proof of Theorem 4.1)

Let us restate the theorem here for the sake of readability.

Theorem 4.1. *Let $\{\boldsymbol{w}_k^{(t)}\}_{k\in[0:K-1],t\geq 0}$ be the sequence of GD iterates (2) from any starting point $\boldsymbol{w}_0^{(0)}$, where tasks are given randomly. Under Assumptions 3.1 and 3.3, if the learning rate satisfies $\eta < \frac{2\phi^2}{\beta\sigma_{\max}^4}$, then the following statements hold with probability 1:*

1. *Loss converges to zero: $\lim_{t\to\infty}\mathcal{L}(\boldsymbol{w}_k^{(t)}) = 0, \forall k \in [0:K-1]$.*
2. *Every data point is classified correctly: $\lim_{t\to\infty}\boldsymbol{x}_i^\top\boldsymbol{w}_k^{(t)} = 0, \forall k\in[0:K-1], i\in I$.*
3. *Square sum of the change of weight is finite: $\sum_{t=0}^{\infty}\sum_{k=0}^{K-1}\|\boldsymbol{w}_{k+1}^{(t)}-\boldsymbol{w}_k^{(t)}\|^2<\infty$.*

Since $\mathcal{L}$ is a $\sigma_{\max}^2\beta$-smooth function, we get

$$
\begin{aligned}
&\mathbb{E}\left[\mathcal{L}(\boldsymbol{w}_{k+1}^{(t)})\right]-\mathbb{E}\left[\mathcal{L}(\boldsymbol{w}_k^{(t)})\right]\\
&\leq \mathbb{E}\left[\nabla\mathcal{L}(\boldsymbol{w}_k^{(t)})^\top(\boldsymbol{w}_{k+1}^{(t)}-\boldsymbol{w}_k^{(t)})\right]+\frac{\sigma_{\max}^2\beta}{2}\mathbb{E}\left[\left\|\boldsymbol{w}_{k+1}^{(t)}-\boldsymbol{w}_k^{(t)}\right\|^2\right]\\
&= \mathbb{E}\left[\mathbb{E}\left[\nabla\mathcal{L}(\boldsymbol{w}_k^{(t)})^\top(\boldsymbol{w}_{k+1}^{(t)}-\boldsymbol{w}_k^{(t)})\mid\boldsymbol{w}_k^{(t)}\right]\right]+\frac{\sigma_{\max}^2\beta}{2}\mathbb{E}\left[\left\|\boldsymbol{w}_{k+1}^{(t)}-\boldsymbol{w}_k^{(t)}\right\|^2\right]\\
&= \mathbb{E}\left[\mathbb{E}\left[\nabla\mathcal{L}(\boldsymbol{w}_k^{(t)})^\top(\boldsymbol{w}_{k+1}^{(t)}-\boldsymbol{w}_k^{(t)})\mid\boldsymbol{w}_k^{(t)}\right]\right]+\frac{\sigma_{\max}^2\beta}{2}\eta^2\mathbb{E}\left[\left\|\sum_{s\in I}z_s^{(t)}\ell'(\boldsymbol{x}_s^\top\boldsymbol{w}_k^{(t)})\boldsymbol{x}_s\right\|^2\right]\\
&= -\frac{\eta}{M}\mathbb{E}\left[\left\|\nabla\mathcal{L}(\boldsymbol{w}_k^{(t)})\right\|^2\right]+\frac{\sigma_{\max}^2\beta}{2}\eta^2\mathbb{E}\left[\left\|\sum_{s\in I}z_s^{(t)}\ell'(\boldsymbol{x}_s^\top\boldsymbol{w}_k^{(t)})\boldsymbol{x}_s\right\|^2\right]\\
&\leq -\frac{\eta}{M}\mathbb{E}\left[\left\|\nabla\mathcal{L}(\boldsymbol{w}_k^{(t)})\right\|^2\right]+\frac{\sigma_{\max}^4\beta}{2}\eta^2\mathbb{E}\left[\sum_{s\in I}\left[z_s^{(t)}\ell'(\boldsymbol{x}_s^\top\boldsymbol{w}_k^{(t)})\right]^2\right]\\
&= -\frac{\eta}{M}\mathbb{E}\left[\left\|\nabla\mathcal{L}(\boldsymbol{w}_k^{(t)})\right\|^2\right]+\frac{\sigma_{\max}^4\beta}{2}\eta^2\sum_{s\in I}\mathbb{E}\left[(z_s^{(t)})^2\right]\mathbb{E}\left[\ell'(\boldsymbol{x}_s^\top\boldsymbol{w}_k^{(t)})^2\right]\\
&= -\frac{\eta}{M}\mathbb{E}\left[\left\|\nabla\mathcal{L}(\boldsymbol{w}_k^{(t)})\right\|^2\right]+\frac{\sigma_{\max}^4\beta}{2M}\eta^2\mathbb{E}\left[\sum_{s\in I}\ell'(\boldsymbol{x}_s^\top\boldsymbol{w}_k^{(t)})^2\right],
\end{aligned}
$$

where $z_s^{(t)}$ is a variable which is 1 when $\boldsymbol{x}_s$ is in the task on stage t, or 0 otherwise. The second inequality comes from the fact $\forall\lambda_s\in\mathbb{R}:\left\|\sum_{s\in I}\lambda_s\boldsymbol{x}_s\right\|_2\leq\sigma_{\max}\sqrt{\sum_{s\in I}\lambda_s^2}$.

By applying Lemma D.2, we obtain

$$
\begin{aligned}
\mathbb{E}\left[\mathcal{L}(\boldsymbol{w}_{k+1}^{(t)})\right]-\mathbb{E}\left[\mathcal{L}(\boldsymbol{w}_k^{(t)})\right] &\leq -\frac{\eta}{M}\left(1-\eta\frac{\sigma_{\max}^4\beta}{2\phi^2}\right)\mathbb{E}\left[\left\|\nabla\mathcal{L}(\boldsymbol{w}_k^{(t)})\right\|^2\right]\\
&= -\frac{\eta}{M}(1-\eta\beta'')\mathbb{E}\left[\left\|\nabla\mathcal{L}(\boldsymbol{w}_k^{(t)})\right\|^2\right],
\end{aligned}
\tag{37}
$$

where $\beta'':=\frac{\sigma_{\max}^4\beta}{2\phi^2}$. Given that $\eta\leq\frac{1}{\beta''}$,

$$
\sum_{t=0}^{\infty}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\mathcal{L}(\boldsymbol{w}_k^{(t)})\right\|^2\right]\leq\frac{\mathcal{L}(\boldsymbol{w}_0^{(0)})-\lim_{t\to\infty}\mathbb{E}\left[\mathcal{L}(\boldsymbol{w}_0^{(t)})\right]}{\frac{\eta}{M}(1-\eta\beta'')}\leq\frac{M\mathcal{L}(\boldsymbol{w}_0^{(0)})}{\eta(1-\eta\beta'')}<\infty.
$$

According to Markov inequality,

$$
\mathbb{P}\left(\sum_{t=0}^{\infty}\sum_{k=0}^{K-1}\left\|\nabla\mathcal{L}(\boldsymbol{w}_k^{(t)})\right\|^2<c\right)\geq 1-\frac{\mathbb{E}\left[\sum_{t=0}^{\infty}\sum_{k=0}^{K-1}\left\|\nabla\mathcal{L}(\boldsymbol{w}_k^{(t)})\right\|^2\right]}{c}
$$

Since $\mathbb{E}\left[\sum_{t=0}^{\infty}\sum_{k=0}^{K-1}\left\|\nabla\mathcal{L}(\boldsymbol{w}_k^{(t)})\right\|^2\right]$ is finite, if we send $c \to \infty$, we get

$$\mathbb{P}\left(\sum_{t=0}^{\infty}\sum_{k=0}^{K-1}\left\|\nabla\mathcal{L}(\boldsymbol{w}_k^{(t)})\right\|^2 < \infty\right) = 1.$$

That is, $\sum_{t=0}^{\infty}\sum_{k=0}^{K-1}\left\|\nabla\mathcal{L}(\boldsymbol{w}_k^{(t)})\right\|^2$ is bounded with probability 1. The boundedness of infinite sum of nonzero elements implies $\forall k \in [0:K-1] : \lim_{t\to 0}\left\|\nabla\mathcal{L}(\boldsymbol{w}_k^{(t)})\right\|^2 = 0$. Combined with Lemma D.2, we obtain $\lim_{t\to 0}\ell'(x_i^\top \boldsymbol{w}_k^{(t)}) = 0, \forall i \in I, k \in [0:K-1]$. Since $\ell'(u) \to 0$ only when $u \to \infty$, $x_i^\top \boldsymbol{w}_k^{(t)} \to \infty, \forall i \in I, k \in [0:K-1]$. And $\lim_{t\to\infty}\mathcal{L}(\boldsymbol{w}_k^{(t)}) = 0, \forall k \in [0:K-1]$. Finally, followed by

$$\begin{aligned}\left\|\nabla\mathcal{L}(\boldsymbol{w}_k^{(t)})\right\| &\geq \phi\sqrt{\sum_{i\in I}\left[\ell'(\boldsymbol{x}_i^\top \boldsymbol{w}_k^{(t)})\right]^2} \geq \phi\sqrt{\sum_{i\in I(t)}\left[\ell'(\boldsymbol{x}_i^\top \boldsymbol{w}_k^{(t)})\right]^2} \\ &\geq \frac{\phi}{\sigma_{\max}}\left\|\sum_{i\in I(t)}\ell'(\boldsymbol{x}_i^\top \boldsymbol{w}_k^{(t)})x_i\right\| = \frac{\phi}{\sigma_{\max}}\eta^{-1}\left\|\boldsymbol{w}_{k+1}^{(t)} - \boldsymbol{w}_k^{(t)}\right\|.\end{aligned}$$

We obtain that $\sum_{t=0}^{\infty}\sum_{k=0}^{K-1}\left\|\boldsymbol{w}_{k+1}^{(t)} - \boldsymbol{w}_k^{(t)}\right\|^2 < \infty$ with probability 1.

E.2 Directional Convergence Analysis (Proof of Theorem 4.2)

In this section, we prove Theorem 4.2 and further discuss the convergence of $\boldsymbol{\rho}_k^{(t)}$ beyond boundedness.

Theorem 4.2. *Let $\{\boldsymbol{w}_k^{(t)}\}_{k\in[0:K-1],t\geq 0}$ be the sequence of GD iterates (2) from any starting point $\boldsymbol{w}_0^{(0)}$, where tasks are given randomly. Under Assumptions 3.1, 3.2, 3.3, and 3.4, if the learning rate satisfies $\eta < \frac{2\phi^2}{\beta\sigma_{\max}^4}$, then with probability 1, $\boldsymbol{w}_k^{(t)}$ will behave as:*

$$\boldsymbol{w}_k^{(t)} = \ln\left(\tfrac{K}{M}t\right)\hat{\boldsymbol{w}} + \boldsymbol{\rho}_k^{(t)},$$

where $\|\boldsymbol{\rho}_k^{(t)}\|$ stays bounded as t grows.

We only need to prove that the two following lemmas still hold in random order.

Lemma E.1. *When tasks are given randomly, there exists $\breve{\boldsymbol{w}}, m_1(t,k) \in \mathbb{R}^d$ the following almost surely holds for all $t \in \mathbb{N}$, $k \in [0:K-1]$:*

$$K\sum_{u=1}^{t-1}\frac{1}{u}\sum_{s\in S^{(u)}}\alpha_s\boldsymbol{x}_s + \frac{k}{t}\sum_{s\in S^{(t)}}\alpha_s\boldsymbol{x}_s = \frac{K}{M}\log(\frac{t}{M})\hat{\boldsymbol{w}} + \frac{K}{M}\breve{\boldsymbol{w}} + m_1(t,k), \tag{38}$$
$$m_1(t,K) := m_1(t+1,0),$$

such that $\|m_1(t,k)\| = o(t^{-0.5+\epsilon})$, and $\|m_1(t,k+1) - m_1(t,k)\| = \mathcal{O}(t^{-1})$ for all $k \in [0:K-1], \epsilon > 0$, and $\breve{\boldsymbol{w}}$ only depends on the order of tasks and constant with respect to t.

Proof. See Appendix E.2.1. □

Using Lemma E.1, we set $m_1(t,k)$ and $\breve{\boldsymbol{w}}$ and define $\boldsymbol{\rho}_k^{(t)}$ and $\boldsymbol{r}_k^{(t)}$ as we did in cyclic order. That is,

$$\begin{aligned}\forall k \in [0:K-1] : \boldsymbol{w}_k^{(t)} &= \log(\frac{K}{M}t)\hat{\boldsymbol{w}} + \boldsymbol{\rho}_k^{(t)} \\ &= \log(\frac{K}{M}t)\hat{\boldsymbol{w}} + \tilde{\boldsymbol{w}} + \frac{M}{K}m_1(t,k) + \boldsymbol{r}_k^{(t)},\end{aligned}$$

and

$$\boldsymbol{\rho}_K^{(t)} = \boldsymbol{\rho}_0^{(t+1)}, \quad \boldsymbol{r}_K^{(t)} = \boldsymbol{r}_0^{(t+1)},$$

where $\tilde{\boldsymbol{w}}$ is the solution of

$$\forall i \in S : \eta \exp\left(-\boldsymbol{x}_i^\top \tilde{\boldsymbol{w}}\right) = \alpha_i, \quad \bar{P}(\tilde{\boldsymbol{w}} - \boldsymbol{w}_0^{(0)}) = 0.$$

which is unique under Assumption 3.2. Then by the definition,

$$\begin{aligned} \boldsymbol{r}_k^{(t)} &= \boldsymbol{w}_k^{(t)} - \frac{M}{K}\left(\frac{K}{M}\log(\frac{K}{M}t)\hat{\boldsymbol{w}} + m_1(t,k)\right) - \tilde{\boldsymbol{w}} \\ &= \boldsymbol{w}_k^{(t)} - \frac{M}{K}\left(K\sum_{u=1}^{t-1}\frac{1}{u}\sum_{s\in S^{(u)}}\alpha_s \boldsymbol{x}_s + \frac{k}{t}\sum_{s\in S^{(t)}}\alpha_s \boldsymbol{x}_s\right) - \log K\hat{\boldsymbol{w}} - \tilde{\boldsymbol{w}} + \check{\boldsymbol{w}}. \end{aligned}$$

Then we can get the second primary lemma of $\boldsymbol{r}_k^{(t)}$.

Lemma E.2. *Under Assumption 3.1, 3.2, 3.3, and 3.4, if learning rate is* $\eta < \frac{2\phi^2}{\beta\sigma_{\max}^4}$*, then*

1. $\exists \tilde{t}, C_1, C_2 > 0$ *such that* $\forall t > \tilde{t}$,
$$(\boldsymbol{r}_{k+1}^{(t)} - \boldsymbol{r}_k^{(t)})^\top \boldsymbol{r}_k^{(t)} \leq C_1 t^{-\theta} + C_2 t^{-1-0.5\tilde{\mu}}, \forall k \in [0:K-1].$$

2. *Moreover, for all* $\epsilon_1 > 0$, $\exists \tilde{t}^*, C_3 > 0$ *such that if* $\left\|P\boldsymbol{r}_k^{(t)}\right\| \geq \epsilon_1$ *and* $S^{(t)} \neq \emptyset$,
$$(\boldsymbol{r}_{k+1}^{(t)} - \boldsymbol{r}_k^{(t)})^\top \boldsymbol{r}_k^{(t)} \leq -C_3 t^{-1}, \forall t > \tilde{t}^*, k \in [0:K-1].$$

Proof. Only the learning rate is different from the cyclic case. Therefore see Appendix D.2.2. □

The remaining step is the same as the proof of Theorem 3.2. To sum up, we can set $\boldsymbol{a}_k^{(t)}$ as $\left\|\boldsymbol{r}_{k+1}^{(t)} - \boldsymbol{r}_k^{(t)}\right\|^2 = \left\|\boldsymbol{w}_{k+1}^{(t)} - \boldsymbol{w}_k^{(t)} - \boldsymbol{a}_k^{(t)}\right\|^2$. Then by Lemma E.1, $\exists t_1$ such that $\forall t \geq t_1, \forall k \in [0:K-1]: \left\|\boldsymbol{a}_k^{(t)}\right\| \leq t^{-1}$.

For all $T \geq t_1$.

$$\begin{aligned} &\sum_{t=t_1}^{T}\sum_{k=0}^{K-1}\left\|\boldsymbol{r}_{k+1}^{(t)} - \boldsymbol{r}_k^{(t)}\right\|^2 = \sum_{t=t_1}^{T}\sum_{k=0}^{K-1}\left\|\boldsymbol{w}_{k+1}^{(t)} - \boldsymbol{w}_k^{(t)} - \boldsymbol{a}_k^{(t)}\right\|^2 \\ &= \sum_{t=t_1}^{T}\sum_{k=0}^{K-1}\left\|\boldsymbol{w}_{k+1}^{(t)} - \boldsymbol{w}_k^{(t)}\right\|^2 + \sum_{t=t_1}^{T}\sum_{k=0}^{K-1}2(\boldsymbol{w}_k^{(t)} - \boldsymbol{w}_{k+1}^{(t)})^\top \boldsymbol{a}_k^{(t)} + \sum_{t=t_1}^{T}\sum_{k=0}^{K-1}\left\|\boldsymbol{a}_k^{(t)}\right\|^2 \\ &\leq \sum_{t=t_1}^{T}\sum_{k=0}^{K-1}\left\|\boldsymbol{w}_{k+1}^{(t)} - \boldsymbol{w}_k^{(t)}\right\|^2 + 2\sqrt{\sum_{t=t_1}^{T}\sum_{k=0}^{K-1}\left\|\boldsymbol{w}_k^{(t)} - \boldsymbol{w}_{k+1}^{(t)}\right\|^2\sum_{t=t_1}^{T}\sum_{k=0}^{K-1}\left\|\boldsymbol{a}_k^{(t)}\right\|^2} + \sum_{t=t_1}^{T}\sum_{k=0}^{K-1}\left\|\boldsymbol{a}_k^{(t)}\right\|^2 \\ &\leq \sum_{t=t_1}^{T}\sum_{k=0}^{K-1}\left\|\boldsymbol{w}_{k+1}^{(t)} - \boldsymbol{w}_k^{(t)}\right\|^2 + 2\sqrt{\sum_{t=t_1}^{T}\sum_{k=0}^{K-1}\left\|\boldsymbol{w}_k^{(t)} - \boldsymbol{w}_{k+1}^{(t)}\right\|^2\sum_{t=t_1}^{T}\sum_{k=0}^{K-1}t^{-2}} + \sum_{t=t_1}^{T}\sum_{k=0}^{K-1}t^{-2} \\ &< \infty. \end{aligned} \tag{39}$$

We use Cauchy-Schwarz inequality for the first inequality and the fact that $\sum_{t=t_1}^{T}t^{-2} < \infty$ and $\sum_{t=t_1}^{T}\sum_{k=0}^{K-1}\left\|\boldsymbol{w}_k^{(t)} - \boldsymbol{w}_{k+1}^{(t)}\right\|^2 < \infty$ by Theorem 4.1.

Combined with Lemma E.2 and the fact that $\forall c > 1: \sum_{t=1}^{\infty}t^{-c} < \infty$, we almost surely get

$$\left\|\boldsymbol{r}_0^{(t)}\right\|^2 - \left\|\boldsymbol{r}_0^{(t_1)}\right\|^2 = \sum_{u=t_1}^{t-1}\sum_{k=0}^{K-1}\left(\left\|\boldsymbol{r}_{k+1}^{(u)}\right\|^2 - \left\|\boldsymbol{r}_k^{(u)}\right\|^2\right)$$

$$= \sum_{u=t_1}^{t-1} \sum_{k=0}^{K-1} \left(2(\boldsymbol{r}_{k+1}^{(u)} - \boldsymbol{r}_k^{(u)})^\top \boldsymbol{r}_k^{(u)} + \left\| \boldsymbol{r}_{k+1}^{(u)} - \boldsymbol{r}_k^{(u)} \right\|^2 \right) < \infty.$$

E.2.1 Proof of Lemma E.1

Here we restate the lemma for the sake of readability.

Lemma E.1. *When tasks are given randomly, there exists $\breve{\boldsymbol{w}}, m_1(t,k) \in \mathbb{R}^d$ the following almost surely holds for all $t \in \mathbb{N}$, $k \in [0 : K-1]$:*

$$K \sum_{u=1}^{t-1} \frac{1}{u} \sum_{s \in S^{(u)}} \alpha_s \boldsymbol{x}_s + \frac{k}{t} \sum_{s \in S^{(t)}} \alpha_s \boldsymbol{x}_s = \frac{K}{M} \log(\frac{t}{M}) \hat{\boldsymbol{w}} + \frac{K}{M} \breve{\boldsymbol{w}} + m_1(t,k), \tag{38}$$
$$m_1(t,K) := m_1(t+1,0),$$

such that $\|m_1(t,k)\| = o(t^{-0.5+\epsilon})$, and $\|m_1(t,k+1) - m_1(t,k)\| = \mathcal{O}(t^{-1})$ for all $k \in [0 : K-1], \epsilon > 0$, and $\breve{\boldsymbol{w}}$ only depends on the order of tasks and constant with respect to t.

We define an (i.i.d.) random variable(s) $z_i^{(t)} := \mathbb{1}\{\boldsymbol{x}_i \in I^{(t)}\}$. Note that $\mathbb{E}[z_i^{(t)}] = \frac{1}{M}$ and $\mathrm{Var}(z_i^{(t)}) = \frac{M-1}{M^2}$ due to uniform sampling of the task index in $[0 : M-1]$. Then, we can write a sum on the right-hand side of Equation (38) as follows:

$$\begin{aligned} K \sum_{u=1}^{t-1} \frac{1}{u} \sum_{s \in S^{(u)}} \alpha_s \boldsymbol{x}_s &= K \sum_{s \in S} \left(\sum_{u=1}^{t-1} \frac{z_s^{(u)}}{u} \right) \alpha_s \boldsymbol{x}_s \\ &= K \sum_{s \in S} \left(\sum_{u=1}^{t-1} \frac{\mathbb{E}[z_s^{(u)}]}{u} + \sum_{u=1}^{t-1} \frac{z_s^{(u)} - \mathbb{E}[z_s^{(u)}]}{u} \right) \alpha_s \boldsymbol{x}_s \\ &= K \sum_{s \in S} \left(\frac{1}{M} \sum_{u=1}^{t-1} \frac{1}{u} + . \sum_{u=1}^{t-1} \frac{z_s^{(u)} - \mathbb{E}[z_s^{(u)}]}{u} \right) \alpha_s \boldsymbol{x}_s. \end{aligned}$$

Since

$$\sum_{u=1}^{t-1} \frac{1}{u} = \log t + \gamma + q(t)$$

where γ is the Euler-Mascheroni constant and $q(t) = \mathcal{O}(t^{-1})$, we have

$$K \sum_{s \in S} \left(\frac{1}{M} \sum_{u=1}^{t-1} \frac{1}{u} \right) \alpha_s \boldsymbol{x}_s = \frac{K}{M} \left(\log t + \gamma + q(t) \right) \hat{\boldsymbol{w}}.$$

Now we are going to deal with the sum

$$\sum_{u=1}^{t-1} \frac{z_s^{(u)} - \mathbb{E}[z_s^{(u)}]}{u}$$

in two aspects: (1) it converges with probability 1 as $t \to \infty$ and (2) the almost-sure vanishing rate of the "residual" (a sum from $u = t$ to ∞) is $o(t^{-0.5+\epsilon})$ for any $\epsilon > 0$. Let us look at its almost-sure convergence. To this end, we utilize the following useful proposition.

Proposition E.3 (Theorem 5.2.6 of Durrett (2019))**.** *Suppose $X_1, X_2, \ldots$ are zero-mean independent random variables. If $\sum_{n=1}^{\infty} \mathrm{Var}(X_n) < \infty$, then $\sum_{n=1}^{\infty} X_n$ converges almost surely (i.e., with probability 1).*

Observe that $X_u := \frac{z_s^{(u)} - \mathbb{E}[z_s^{(u)}]}{u}$ is a zero-mean random variables. Not only they are independent for all u, but also the sum of their variances is convergent:

$$\sum_{u=1}^{\infty} \mathrm{Var}(X_u) = \frac{M-1}{M^2} \sum_{u=1}^{\infty} \frac{1}{u^2} < \infty.$$

Thus, by Proposition E.3, the sum $\sum_{u=1}^{\infty} X_u$ converges with probability 1. Next, we want to show the vanishing rate

$$\sum_{u=t}^{\infty} X_u = o(t^{-0.5+\epsilon})$$

with probability 1, where we choose any $\epsilon > 0$. Observe that it is equivalent to show, for any $\delta > 0$,

$$\mathbb{P}\left(t^{0.5-\epsilon} \cdot \left| \sum_{u=t}^{\infty} X_u \right| > \delta \text{ for infinitely many } t \right) = 0.$$

Here we bring a renowned Borel-Cantelli Lemma.

Proposition E.4 (Borel-Cantelli lemma; Theorem 2.3.1 of Durrett (2019))**.** *Consider a sequence of events $A_1, A_2, \cdots$. If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then*

$$\mathbb{P}(\limsup_{n\to\infty} A_n) := \mathbb{P}(A_n \textit{ happens for infinitely many } n) = 0.$$

By Proposition E.4, it suffices to show

$$\forall \delta > 0, \quad \sum_{t=1}^{\infty} \mathbb{P}\left(t^{0.5-\epsilon} \cdot \left| \sum_{u=t}^{\infty} X_u \right| > \delta \right) < \infty.$$

Let us recall Hoeffding inequality here:

Proposition E.5 (Hoeffding inequality)**.** *Consider a collection of independent random variables $X_1, \cdots, X_n$ satisfying $a_i \le X_i \le b_i$ for each $i = 1, \cdots, n$ ($a_i < b_i$). Then,*

$$\mathbb{P}\left(\left| \sum_{i=1}^{n} X_i \right| \ge r \right) \le 2 \exp\left(-\frac{2r^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right).$$

Since the sum $\sum_{u=t}^{\infty} X_u$ converges almost surely, it is a well-defined random variable with probability 1, and

$$\begin{aligned}
\mathbb{P}\left(\left| \sum_{u=t}^{\infty} X_u \right| > \delta \cdot t^{-0.5+\epsilon} \right) &= \mathbb{P}\left(\left| \sum_{u=t}^{T} X_u \right| > \delta \cdot t^{-0.5+\epsilon} \text{ for all but finitely many } T \right) \\
&=: \mathbb{P}\left(\liminf_{T\to\infty} \left\{ \left| \sum_{u=t}^{T} X_u \right| > \delta \cdot t^{-0.5+\epsilon} \right\} \right) \\
&\le \liminf_{T\to\infty} \mathbb{P}\left(\left| \sum_{u=t}^{T} X_u \right| > \delta \cdot t^{-0.5+\epsilon} \right) && (40) \\
&\le \liminf_{T\to\infty} 2 \exp\left(-\frac{2\delta^2 t^{-1+2\epsilon}}{\sum_{u=t}^{T} \frac{1}{u^2}} \right) && (41) \\
&= 2 \exp\left(-\frac{2\delta^2 t^{-1+2\epsilon}}{\sum_{u=t}^{\infty} \frac{1}{u^2}} \right) && (42) \\
&\le 2 \exp\left(-\delta^2 t^{2\epsilon} \right). && (43)
\end{aligned}$$

We use the fact "$\mathbb{P}(\liminf_n A_n) \le \liminf_n \mathbb{P}(A_n)$" in Equation (40); we apply Hoeffding inequality (Proposition E.5) and the fact $-\frac{1}{Mu} \le X_u \le \frac{M-1}{Mu}$ in Equation (41); and we utilize the fact $\sum_{u=t}^{\infty} \frac{1}{u^2} \le \frac{2}{t}$ in Equation (43). Since $\exp(-\delta^2 t^{2\epsilon}) = o(t^{-2})$ for any $\epsilon > 0$ and large enough t, the sum $\sum_t \exp(-\delta^2 t^{2\epsilon})$ converges. Therefore, we have desired almost-sure convergence guarantees.

From now on, let us proceed with the proof. Using the almost-sure convergence results, let

$$\check{\boldsymbol{w}} := (\log M + \gamma)\hat{\boldsymbol{w}} + M \sum_{s\in S} \left(\sum_{u=1}^{\infty} X_u \right) \alpha_s \boldsymbol{x}_s,$$

$$m_1(t,k) := \frac{K}{M}q(t)\hat{\boldsymbol{w}} + K\sum_{s\in S}\left(\sum_{u=t}^{\infty} X_u\right)\alpha_s\boldsymbol{x}_s + \frac{k}{t}\sum_{s\in S^{(t)}}\alpha_s\boldsymbol{x}_s.$$

Then with probability 1, the statement of the lemma holds: Equation (38) holds, where $\check{\boldsymbol{w}}$ is a constant vector in terms of t, $\|m_1(t,k)\| \le o(t^{-0.5+\epsilon})$ for any $\epsilon > 0$, and

$$\|m_1(t,k+1) - m_1(t,k)\| = \mathcal{O}(t^{-1}), \quad (k = 0, ..., K-2)$$
$$\|m_1(t+1,0) - m_1(t,K-1)\| = \mathcal{O}(t^{-1}).$$

This concludes the proof of the lemma.

E.2.2 Convergence of $\rho_k^{(t)}$

We also can prove a characterization of the limit of $\boldsymbol{\rho}_k^{(t)}$, as done in Appendix D.2.3. However, when tasks are given randomly, we need an additional assumption to guarantee the convergence of $\boldsymbol{\rho}_k^{(t)}$ to the particular point.

Assumption E.1. Every task has at least one support vector. That is, $\forall m \in [0:M-1] : S_m \neq \emptyset$.

Proposition E.6. *Under the same setting of Theorem 4.2 with additional Assumptions D.1 and E.1, the "residual" converges to* $\lim_{t\to\infty}\boldsymbol{\rho}_k^{(t)} = \tilde{\boldsymbol{w}}, \forall k \in [0:K-1]$. *Here,* $\tilde{\boldsymbol{w}}$ *is the vector defined in Proposition D.6.*

Proof. First, $\bar{P}\boldsymbol{r}_k^{(t)} = \bar{P}\boldsymbol{w}_0^{(0)} - \bar{P}\tilde{\boldsymbol{w}} = 0$ holds as in cyclic case. See Appendix D.2.3.

Second, we get to show $P\boldsymbol{r}_k^{(t)} \to 0$. By Equation (39), $\lim_{T\to\infty}\sum_{t=t_1}^{T}\sum_{k=0}^{K-1}\left\|\boldsymbol{r}_{k+1}^{(t)} - \boldsymbol{r}_k^{(t)}\right\|^2 = C_4$. That means $\forall k \in [0:K-1] : \lim_{T\to\infty}\left\|\boldsymbol{r}_{k+1}^{(T)} - \boldsymbol{r}_k^{(T)}\right\| = 0$. Therefore, for any ϵ_0, there exists $t_2 > 0$ such that $\left\|\boldsymbol{r}_{k+1}^{(t)} - \boldsymbol{r}_k^{(t)}\right\| < \frac{\epsilon_0}{K}$ for all $t \ge t_2, k \in [0:K-1]$. As a result,

$$\left\|P\boldsymbol{r}_0^{(t)}\right\| + \frac{k}{K}\epsilon_0 \ge \left\|P\boldsymbol{r}_k^{(t)}\right\| \ge \left\|P\boldsymbol{r}_0^{(t)}\right\| - \frac{k}{K}\epsilon_0$$

For $t \ge \max\{t_1, t_2, \tilde{t}^*\}$, if $\left\|P\boldsymbol{r}_0^{(t)}\right\| \ge \epsilon_1 + \epsilon_0$ and $S^{(t)} \neq \emptyset$, then $\forall k \in [0:K-1] : \left\|P\boldsymbol{r}_k^{(t)}\right\| \ge \epsilon_1$. By Lemma E.2 (2),

$$\sum_{u=t-1}^{t}\sum_{v=0}^{K-1}(\boldsymbol{r}_{v+1}^{(u)} - \boldsymbol{r}_v^{(u)})^\top\boldsymbol{r}_v^{(u)} \le -KC_3t^{-1} + K\left(C_1t^{-\theta} + C_2t^{-1-0.5\tilde{\mu}}\right),$$

Since t^{-1} decrease to zero slower than $t^{-\theta}$ and $t^{-1-0.5\tilde{\mu}}$, there exists $t_3 > \max\{t_1, t_2, \tilde{t}^*\}, C_4 > 0$ such that $-KC_3t^{-1} + K\left(C_1t^{-\theta} + C_2t^{-1-0.5\tilde{\mu}}\right) \le -C_5t^{-1}$. Also $S^{(t)} \neq \emptyset$ is given by Assumption E.1. To sum up, for any $\epsilon_0, \epsilon_2 > 0$, there exists $t_3 > \max\{t_1, t_2, \tilde{t}^*\}$ such that if $\left\|P\boldsymbol{r}_0^{(t)}\right\| \ge \epsilon_0 + \epsilon_1$, then

$$\sum_{u=t-1}^{t}\sum_{v=0}^{K-1}(\boldsymbol{r}_{v+1}^{(u)} - \boldsymbol{r}_v^{(u)})^\top\boldsymbol{r}_v^{(u)} \le -C_5t^{-1},$$

Now, define two sets for each $k \in [0:K-1]$

$$\mathcal{T}_k := \{t > t_3 : \left\|P\boldsymbol{r}_k^{(t)}\right\| < \epsilon_0 + \epsilon_1\}$$
$$\bar{\mathcal{T}}_k := \{t > t_3 : \left\|P\boldsymbol{r}_k^{(t)}\right\| \ge \epsilon_0 + \epsilon_1\}$$

We will finish our proof by showing that $\bar{\mathcal{T}}_k$ is finite. Here, we use the fact that every $\mathcal{T}_k$ is infinite. The proof is the same as in the cyclic case. Since $\lim_{T\to\infty}\sum_{t=t_1}^{T}\sum_{k=0}^{K-1}\left\|\boldsymbol{r}_{k+1}^{(t)}-\boldsymbol{r}_k^{(t)}\right\|^2 = C_4$, we get

$$\sum_{u=t_1}^{t}\sum_{k=0}^{K-1}\left\|\boldsymbol{r}_{k+1}^{(u)}-\boldsymbol{r}_k^{(u)}\right\|^2 = C_4 - h(t)$$

where $h(t)$ is a positive function monotonic decreasing to zero.
Now, assume that there exists some k' that $\bar{\mathcal{T}}_k$ is infinite. WLOG, we set $k'=0$. Since $\mathcal{T}_0$ is infinite, for any $t\in\bar{\mathcal{T}}_0$ there exists $t',t''\in\mathcal{T}_0$ such that $t\in[t'+1,t''-1]\subset\bar{\mathcal{T}}_0$. We divide it into two cases: For all $t\in[t'+1,t''-1]$,

1. if $t=t'+1$, then $\left\|P\boldsymbol{r}_0^{(t)}\right\|^2 \le \left\|P\boldsymbol{r}_0^{(t')}\right\|^2+\epsilon_0\le 2\epsilon_0+\epsilon_1$.

2. if $t\ge t'+1$, then

$$\begin{aligned}\left\|P\boldsymbol{r}_0^{(t)}\right\|^2 &= \left\|P\boldsymbol{r}_0^{(t')}\right\|^2+\sum_{u=t'}^{t-1}\sum_{k=0}^{K-1}\left[\left\|\boldsymbol{r}_{k+1}^{(u)}\right\|^2-\left\|\boldsymbol{r}_k^{(u)}\right\|^2\right]\\ &= \left\|P\boldsymbol{r}_0^{(t')}\right\|^2+\sum_{u=t'}^{t-1}\sum_{k=0}^{K-1}\left[\left\|\boldsymbol{r}_{k+1}^{(u)}-\boldsymbol{r}_k^{(u)}\right\|^2+2(\boldsymbol{r}_{k+1}^{(u)}-\boldsymbol{r}_k^{(u)})^\top\boldsymbol{r}_k^{(u)}\right]\\ &= \left\|P\boldsymbol{r}_0^{(t')}\right\|^2+h(t)-h(t')+2\sum_{u=t'}^{t-1}\sum_{k=0}^{K-1}\left[(\boldsymbol{r}_{k+1}^{(u)}-\boldsymbol{r}_k^{(u)})^\top\boldsymbol{r}_k^{(u)}\right]\\ &\le(\epsilon_0+\epsilon_1)^2+h(t)-2C_5\frac{1}{t'+1}-2C_3\sum_{u=t'+2}^{t-1}\frac{1}{u}\\ &\le(\epsilon_0+\epsilon_1)^2+h(t).\end{aligned}$$

   Since $h(t)$ is monotonic decreasing function, for any $\epsilon_2>0$, there exists t_4 such that $\forall t\ge t_4$: $h(t)<\epsilon_2$.

Therefore, $\forall t\ge\max\{t_3,t_4\}$: $\left\|P\boldsymbol{r}_0^{(t)}\right\|^2\le(\epsilon_0+\epsilon_1)^2+\epsilon_2$. Since it holds for any $\epsilon_0,\epsilon_1,\epsilon_2$, it contradicts with the assumption that $\bar{\mathcal{T}}_0$ is infinite. □

F Proofs for Section 5: Cyclic Task Ordering, Jointly Non-Separable

Review on Bregman Divergence. Before we start the proofs, we briefly overview some basic properties of *Bregman divergence*.

Given a convex function $f : \mathcal{S} \to \mathbb{R}$ defined on a convex set $\mathcal{S} \subset \mathbb{R}^d$, the Bregman divergence between two points $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}$ with respect to f is defined as

$$D_f(\boldsymbol{x}, \boldsymbol{y}) := f(\boldsymbol{x}) - f(\boldsymbol{y}) - \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle .$$

Note that $D_f(\boldsymbol{x}, \boldsymbol{y}) \geq 0$ for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}$ because of the definition of convexity; when f is strictly convex, $D_f(\boldsymbol{x}, \boldsymbol{y}) = 0$ if and only if $\boldsymbol{x} = \boldsymbol{y}$. Also, if f is β-smooth, $D_f(\boldsymbol{x}, \boldsymbol{y}) \leq \frac{\beta}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2$ holds by the definition of smoothness. We often use the following useful identity that links three different points $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathcal{S}$:

$$\langle \nabla f(\boldsymbol{z}), \boldsymbol{x} - \boldsymbol{y} \rangle = [f(\boldsymbol{x}) - f(\boldsymbol{y})] - [D_f(\boldsymbol{x}, \boldsymbol{z}) - D_f(\boldsymbol{y}, \boldsymbol{z})] . \tag{44}$$

Here is another useful fact: for a convex β-smooth function f, the Bregman divergence is bound below by the squared distance between gradients.

Proposition F.1. *Let $f : \mathcal{S} \to \mathbb{R}$ be a convex, β-smooth function defined on a convex set $\mathcal{S} \subset \mathbb{R}^d$. For any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}$,*

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|^2 \leq 2\beta D_f(\boldsymbol{x}, \boldsymbol{y}).$$

Proof. Observe that $D_f(\cdot, \boldsymbol{y})$ is also a β-smooth function for any $\boldsymbol{y}$. Let $\boldsymbol{z} = \boldsymbol{x} - \frac{1}{\beta} \nabla_{\boldsymbol{x}} D_f(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x} - \frac{1}{\beta}[\nabla f(\boldsymbol{x}) - \nabla(\boldsymbol{y})]$. Then by β-smoothness and the non-negativity of $D_f(\cdot, \boldsymbol{y})$, we have

$$\begin{aligned}
0 &\leq D_f(\boldsymbol{z}, \boldsymbol{y}) \\
&\leq D_f(\boldsymbol{x}, \boldsymbol{y}) + \langle \nabla_{\boldsymbol{x}} D_f(\boldsymbol{x}, \boldsymbol{y}), \boldsymbol{z} - \boldsymbol{x} \rangle + \frac{\beta}{2} \|\boldsymbol{z} - \boldsymbol{x}\|^2 \\
&= D_f(\boldsymbol{x}, \boldsymbol{y}) - \frac{1}{\beta} \langle \nabla f(\boldsymbol{x}) - \nabla(\boldsymbol{y}), \nabla f(\boldsymbol{x}) - \nabla(\boldsymbol{y}) \rangle + \frac{1}{2\beta} \|\nabla f(\boldsymbol{x}) - \nabla(\boldsymbol{y})\|^2 \\
&= D_f(\boldsymbol{x}, \boldsymbol{y}) - \frac{1}{2\beta} \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|^2 .
\end{aligned}$$

This proves the proposition. □

Useful Inequalities. There are other two crucial inequalities for the proofs in this appendix. One is a variant of Jensen's inequality applied to a squared norm.

Proposition F.2. *For any positive numbers $\lambda_1, \cdots, \lambda_n > 0$, any vectors $\boldsymbol{u}_1, \cdots, \boldsymbol{u}_n \in \mathbb{R}^d$, and an integer $m \in [0 : n]$,*

$$\left\| \sum_{i=1}^m \boldsymbol{u}_i \right\|^2 \leq \left(\sum_{i=1}^n \lambda_i \right) \left(\sum_{i=1}^n \frac{1}{\lambda_i} \|\boldsymbol{u}_i\|^2 \right).$$

Proof. Let $\Lambda_m = \sum_{i=1}^m \lambda_i$. Then by convexity of the squared norm,

$$\begin{aligned}
\left\| \sum_{i=1}^m \boldsymbol{u}_i \right\|^2 &= \left\| \sum_{i=1}^m \frac{\lambda_i}{\Lambda_m} \left(\frac{\Lambda_m}{\lambda_i} \boldsymbol{u}_i \right) \right\|^2 \\
&\leq \sum_{i=1}^m \frac{\lambda_i}{\Lambda_m} \left\| \frac{\Lambda_m}{\lambda_i} \boldsymbol{u}_i \right\|^2 \\
&= \Lambda_m \sum_{i=1}^m \frac{1}{\lambda_i} \|\boldsymbol{u}_i\|^2 \\
&\leq \Lambda_n \sum_{i=1}^n \frac{1}{\lambda_i} \|\boldsymbol{u}_i\|^2 .
\end{aligned}$$

□

Another is about solving a recurrent inequality.

Proposition F.3. *Consider $0 < \mu \le \beta$, $V > 0$, $T > 1$, $0 < c = \Theta(1)$, $0 < m = \Theta(1)$, and $\Delta_0 \ge 0$. Suppose the following inequality holds for any positive $\alpha \le \frac{c}{\beta}$ and $t \in [0 : T-1]$:*

$$\Delta_{t+1} \le \frac{1}{1+\alpha\mu}\Delta_t + \alpha^{m+1}V.$$

If we take

$$\alpha = \min\left\{\frac{c}{\beta}, \ \frac{c+1}{\mu T}\ln\left(T^m \cdot \max\left\{1, \frac{\Delta_0\mu^{m+1}}{V}\right\}\right)\right\},$$

we have

$$\Delta_T = \mathcal{O}\left(\exp\left(-\frac{c\mu}{(c+1)\beta}T\right)\Delta_0 + \frac{V\ln^m T}{\mu^{m+1}T^m}\right).$$

Proof. Since $\alpha\mu \le \frac{c\mu}{\beta} \le c$, we have $\frac{1}{1+\alpha\mu} \le 1 - \frac{\alpha\mu}{c+1}$. By unrolling the recurrent inequality, we have

$$\begin{aligned}\Delta_T &\le \left(1-\frac{\alpha\mu}{c+1}\right)^T\Delta_0 + \alpha^{m+1}V\sum_{t=0}^{T-1}\left(1-\frac{\alpha\mu}{c+1}\right)^t \\ &\le \exp\left(-\frac{\alpha\mu}{c+1}T\right)\Delta_0 + \frac{2\alpha^m V}{\mu}.\end{aligned}$$

With the choice of α, the first exponential term is bounded as

$$\begin{aligned}\exp\left(-\frac{\alpha\mu}{c+1}T\right)\Delta_0 &\le \max\left\{\exp\left(-\frac{c\mu}{(c+1)\beta}T\right)\Delta_0, \ \frac{V}{\mu^{m+1}T^m}\right\} \\ &\le \exp\left(-\frac{c\mu}{(c+1)\beta}T\right)\Delta_0 + \frac{V}{\mu^{m+1}T^m}.\end{aligned}$$

Also, the second term is bounded as

$$\frac{2\alpha^m V}{\mu} \le \frac{2(c+1)^m V}{\mu^{m+1}T^m}\ln^m\left(T^m \cdot \max\left\{1, \frac{\Delta_0\mu^{m+1}}{V}\right\}\right).$$

Combining these two and ignoring the constant/polylogarithmic factors, we have a desired bound. □

F.1 Local Strong Convexity Analysis (Proof of Lemma 5.1)

Recall that we consider cyclic continual learning on M jointly strictly non-separable classification tasks. Let us restate the lemma here for the sake of readability.

Lemma 5.1. *Consider learning M linear classification tasks cyclically. Suppose that Assumptions 5.1 and 5.2 hold. Let $B := \sum_{m=0}^{M-1}\beta_m$ and $V_\star := \sum_{m=0}^{M-1}\frac{1}{\beta_m}\|\nabla\mathcal{L}_m(\boldsymbol{w}_\star)\|^2$. Take a step size $\eta \le \frac{1}{2\sqrt{2}KB}$. Then, there exists a compact set $\mathcal{W} \subset \mathbb{R}^d$ containing $\boldsymbol{w}_\star$ and every $\boldsymbol{w}_0^{(jM)}$ ($j = 0, 1, 2, \ldots$), whose radius is independent of J (the number of cycles) but depends on other parameters like b, G, B, and $V_\star$. Also, the offline training loss $\mathcal{L}$ is μ-strongly convex on $\mathcal{W}$, where*

$$\mu := \left(\min_{i\in[0:N-1],\boldsymbol{w}\in\mathcal{W}}\ell''\left(y_i\boldsymbol{x}_i^\top\boldsymbol{w}\right)\right)\cdot\lambda_{\min}\left(\boldsymbol{X}\boldsymbol{X}^\top\right) > 0. \tag{6}$$

To prove the boundedness of end-of-cycle iterates and the local strong convexity, we first establish a per-cycle recurrent inequality in terms of squared distance to an arbitrary comparator $\boldsymbol{u} \in \mathbb{R}^d$ and the risk values. We put $\boldsymbol{u} = \boldsymbol{w}_\star$ later.

Lemma F.4 (Backward recurrent inequality)**.** *Consider learning M linear classification tasks cyclically. Suppose that Assumption 5.2 holds. Let $B = \sum_{m=0}^{M-1}\beta_m$. If we take any step size satisfying $\eta \le \frac{1}{2\sqrt{2}KB}$, the iterates of sequential GD satisfies*

$$\left\|\boldsymbol{w}_0^{((j+1)M)} - \boldsymbol{u}\right\|^2$$

$$\leq \left\| \boldsymbol{w}_0^{(jM)} - \boldsymbol{u} \right\|^2 - 2\eta K \left[\mathcal{L}\left(\boldsymbol{w}_0^{(jM)}\right) - \mathcal{L}(\boldsymbol{u}) \right] + 2\sqrt{2}\eta^2 K^2 B \left(\sum_{m=0}^{M-1} \frac{1}{\beta_m} \left\|\nabla \mathcal{L}_m(\boldsymbol{u})\right\|^2 \right),$$

for any vector $\boldsymbol{u} \in \mathbb{R}^d$ *and for all* $j = 0, 1, \cdots$.

Proof. We defer the proof to Appendix F.1.1. We remark that this lemma holds even without assuming the non-separability. □

Observe that the following holds as a special case:

$$\left\| \boldsymbol{w}_0^{((j+1)M)} - \boldsymbol{w}_\star \right\|^2 \leq \left\| \boldsymbol{w}_0^{(jM)} - \boldsymbol{w}_\star \right\|^2 - 2\eta K \left[\mathcal{L}\left(\boldsymbol{w}_0^{(jM)}\right) - \mathcal{L}(\boldsymbol{w}_\star) \right] + 2\sqrt{2}\eta^2 K^2 B V_\star, \tag{45}$$

where $V_\star = \sum_{m=0}^{M-1} \frac{1}{\beta_m} \left\|\nabla \mathcal{L}_m(\boldsymbol{w}_\star)\right\|^2$.

The next step is to construct a compact ball $\mathcal{W}$ centered at $\boldsymbol{w}_\star$, containing every end-of-cycle iterate of sequential GD. The crucial step is to apply the non-separability coefficient $b > 0$ (Assumption 5.1).

Lemma F.5 (Boundedness of the end-of-cycle iterates)**.** *Consider learning* M *linear classification tasks cyclically. Suppose that Assumptions 5.1 and 5.2 holds. Let* $B = \sum_{m=0}^{M-1} \beta_m$ *and* $V_\star = \sum_{m=0}^{M-1} \frac{1}{\beta_m} \left\|\nabla \mathcal{L}_m(\boldsymbol{w}_\star)\right\|^2$. *If we take any step size satisfying* $\eta \leq \frac{1}{2\sqrt{2}KB}$, *the end-of-cycle iterates of sequential GD are contained in a compact set which is fixed in terms of the number of the cycle: for all* $j = 0, 1, \cdots$,

$$\begin{aligned} \boldsymbol{w}_0^{(jM)} \in \mathcal{W} &:= \left\{ \boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w} - \boldsymbol{w}_\star\|^2 \leq \left[\frac{1}{Gb} \left(\mathcal{L}(\boldsymbol{w}_\star) + \sqrt{2}\eta K B V_\star \right) + \|\boldsymbol{w}_\star\| \right]^2 + 2\sqrt{2}\eta^2 K^2 B V_\star \right\} \\ &\subseteq \left\{ \boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w} - \boldsymbol{w}_\star\|^2 \leq \left[\frac{1}{Gb} \left(\mathcal{L}(\boldsymbol{w}_\star) + \frac{V_\star}{2} \right) + \|\boldsymbol{w}_\star\| \right]^2 + \frac{V_\star}{2\sqrt{2}B} \right\}. \end{aligned}$$

Proof. The proof is done by induction based on Equation (45). We defer the proof to Appendix F.1.2. □

The last part of the proof is to compute the strong convexity coefficient of $\mathcal{L}$ on $\mathcal{W}$. Since $\mathcal{L}$ is twice differentiable, it can be directly done by computing a lower bound of the minimum Hessian eigenvalue on $\mathcal{W}$: for any $\boldsymbol{w} \in \mathcal{W}$,

$$\nabla^2 \mathcal{L}(\boldsymbol{w}) = \sum_{i=0}^{N-1} \ell''(\boldsymbol{x}_i^\top \boldsymbol{w}) \boldsymbol{x}_i \boldsymbol{x}_i^\top \succeq \left(\min_{\substack{i \in [0:N-1] \\ \boldsymbol{w} \in \mathcal{W}}} \ell''(\boldsymbol{x}_i^\top \boldsymbol{w}) \right) \boldsymbol{X}\boldsymbol{X}^\top \succeq \mu \boldsymbol{I}.$$

This concludes the proof of Lemma 5.1.

F.1.1 Proof of Lemma F.4

For the sake of readability, we restate the lemma.

Lemma F.4 (Backward recurrent inequality)**.** *Consider learning* M *linear classification tasks cyclically. Suppose that Assumption 5.2 holds. Let* $B = \sum_{m=0}^{M-1} \beta_m$. *If we take any step size satisfying* $\eta \leq \frac{1}{2\sqrt{2}KB}$, *the iterates of sequential GD satisfies*

$$\begin{aligned} &\left\| \boldsymbol{w}_0^{((j+1)M)} - \boldsymbol{u} \right\|^2 \\ &\leq \left\| \boldsymbol{w}_0^{(jM)} - \boldsymbol{u} \right\|^2 - 2\eta K \left[\mathcal{L}\left(\boldsymbol{w}_0^{(jM)}\right) - \mathcal{L}(\boldsymbol{u}) \right] + 2\sqrt{2}\eta^2 K^2 B \left(\sum_{m=0}^{M-1} \frac{1}{\beta_m} \left\|\nabla \mathcal{L}_m(\boldsymbol{u})\right\|^2 \right), \end{aligned}$$

for any vector $\boldsymbol{u} \in \mathbb{R}^d$ *and for all* $j = 0, 1, \cdots$.

We start the proof by bounding the squared distance between two iterates in the same cycle of continual learning. For $k \in [0:K]$ and $m \in [0:M-1]$,

$$\begin{aligned}
&\left\|\boldsymbol{w}_0^{(jM)} - \boldsymbol{w}_k^{(jM+m)}\right\|^2 \\
&= \eta^2 \left\| \sum_{l=0}^{m-1}\sum_{h=0}^{K-1} \nabla\mathcal{L}_l(\boldsymbol{w}_h^{(jM+l)}) + \sum_{h=0}^{k-1} \nabla\mathcal{L}_m(\boldsymbol{w}_h^{(jM+m)}) \right\|^2 \\
&\leq \eta^2 \left(\sum_{l=0}^{M-1}\sum_{h=0}^{K-1} \beta_l \right) \left(\sum_{l=0}^{M-1}\sum_{h=0}^{K-1} \frac{1}{\beta_l} \left\| \nabla\mathcal{L}_l(\boldsymbol{w}_h^{(jM+l)}) \right\|^2 \right) \\
&\leq 2\eta^2 KB \left(\sum_{l=0}^{M-1}\sum_{h=0}^{K-1} \frac{1}{\beta_l} \left[\left\| \nabla\mathcal{L}_l(\boldsymbol{w}_h^{(jM+l)}) - \nabla\mathcal{L}_l(\boldsymbol{u}) \right\|^2 + \left\| \nabla\mathcal{L}_l(\boldsymbol{u}) \right\|^2 \right] \right) \\
&\leq 4\eta^2 KB \sum_{l=0}^{M-1}\sum_{h=0}^{K-1} D_{\mathcal{L}_l}(\boldsymbol{u}, \boldsymbol{w}_h^{(jM+l)}) + 2\eta^2 K^2 B \sum_{l=0}^{M-1} \frac{1}{\beta_l} \left\| \nabla\mathcal{L}_l(\boldsymbol{u}) \right\|^2 .
\end{aligned} \tag{46}$$

We use (modified) Jensen's inequality (e.g., Proposition F.2) in the first two inequalities above; the last inequality is due to Proposition F.1.

Next, we decompose the $(j+1)$-th squared distance into j-th squared distance and more:

$$\begin{aligned}
&\left\|\boldsymbol{w}_0^{((j+1)M)} - \boldsymbol{u}\right\|^2 \\
&= \left\|\boldsymbol{w}_0^{(jM)} - \boldsymbol{u}\right\|^2 - 2\eta \sum_{m=0}^{M-1}\sum_{k=0}^{K-1} \left\langle \nabla\mathcal{L}_l(\boldsymbol{w}_k^{(jM+m)}), \boldsymbol{w}_0^{(jM)} - \boldsymbol{u} \right\rangle + \left\|\boldsymbol{w}_0^{(jM)} - \boldsymbol{w}_0^{((j+1)M)}\right\|^2 .
\end{aligned}$$

Using Equation (44) and β_m-smoothnesses,

$$\begin{aligned}
&\left\|\boldsymbol{w}_0^{((j+1)M)} - \boldsymbol{u}\right\|^2 - \left\|\boldsymbol{w}_0^{(jM)} - \boldsymbol{u}\right\|^2 \\
&= -2\eta \sum_{m=0}^{M-1}\sum_{k=0}^{K-1} \left[\mathcal{L}_m(\boldsymbol{w}_0^{(jM)}) - \mathcal{L}_m(\boldsymbol{u}) - D_{L_m}(\boldsymbol{w}_0^{(jM)}, \boldsymbol{w}_k^{(jM+m)}) + D_{L_m}(\boldsymbol{u}, \boldsymbol{w}_k^{(jM+m)}) \right] \\
&\quad + \left\|\boldsymbol{w}_0^{(jM)} - \boldsymbol{w}_0^{((j+1)M)}\right\|^2 \\
&\leq -2\eta K \left[\mathcal{L}\left(\boldsymbol{w}_0^{(jM)}\right) - \mathcal{L}(\boldsymbol{u}) \right] + \eta \sum_{m=0}^{M-1}\sum_{k=0}^{K-1} \beta_m \left\|\boldsymbol{w}_0^{(jM)} - \boldsymbol{w}_k^{(jM+m)}\right\|^2 \\
&\quad - 2\eta \sum_{m=0}^{M-1}\sum_{k=0}^{K-1} D_{L_m}(\boldsymbol{u}, \boldsymbol{w}_k^{(jM+m)}) + \left\|\boldsymbol{w}_0^{(jM)} - \boldsymbol{w}_0^{((j+1)M)}\right\|^2 \\
&\leq -2\eta K \left[\mathcal{L}\left(\boldsymbol{w}_0^{(jM)}\right) - \mathcal{L}(\boldsymbol{u}) \right] + 2\eta^2 K^2 B (1 + \eta KB) \sum_{m=0}^{M-1} \frac{1}{\beta_m} \left\| \nabla\mathcal{L}_m(\boldsymbol{u}) \right\|^2 \\
&\quad - 2\eta(1 - 2\eta KB - 2\eta^2 K^2 B^2) \sum_{m=0}^{M-1}\sum_{k=0}^{K-1} D_{L_m}(\boldsymbol{u}, \boldsymbol{w}_k^{(jM+m)}) \\
&\leq -2\eta K \left[\mathcal{L}\left(\boldsymbol{w}_0^{(jM)}\right) - \mathcal{L}(\boldsymbol{u}) \right] + 2\sqrt{2}\eta^2 K^2 B \sum_{m=0}^{M-1} \frac{1}{\beta_m} \left\| \nabla\mathcal{L}_m(\boldsymbol{u}) \right\|^2 .
\end{aligned} \tag{47}$$

In Equation (47), we use the result from Equation (46) for multiple times. The last inequality is due to our choice of step size: $\eta KB \leq \frac{1}{2\sqrt{2}} < \frac{\sqrt{3}-1}{2} < \sqrt{2} - 1$ ($\because 1 - 2q - 2q^2 \geq 0$ if $q \in \left[0, \frac{\sqrt{3}-1}{2}\right]$). This is the end of the proof.

F.1.2 Proof of Lemma F.5

For the sake of readability, we restate the lemma.

Lemma F.5 (Boundedness of the end-of-cycle iterates)**.** *Consider learning M linear classification tasks cyclically. Suppose that Assumptions 5.1 and 5.2 holds. Let $B = \sum_{m=0}^{M-1} \beta_m$ and $V_\star = \sum_{m=0}^{M-1} \frac{1}{\beta_m} \left\| \nabla \mathcal{L}_m(\boldsymbol{w}_\star) \right\|^2$. If we take any step size satisfying $\eta \leq \frac{1}{2\sqrt{2}KB}$, the end-of-cycle iterates of sequential GD are contained in a compact set which is fixed in terms of the number of the cycle: for all $j = 0, 1, \cdots$,*

$$\boldsymbol{w}_0^{(jM)} \in \mathcal{W} := \left\{ \boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w} - \boldsymbol{w}_\star\|^2 \leq \left[\frac{1}{Gb} \left(\mathcal{L}(\boldsymbol{w}_\star) + \sqrt{2}\eta K B V_\star \right) + \|\boldsymbol{w}_\star\| \right]^2 + 2\sqrt{2}\eta^2 K^2 B V_\star \right\}$$
$$\subseteq \left\{ \boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w} - \boldsymbol{w}_\star\|^2 \leq \left[\frac{1}{Gb} \left(\mathcal{L}(\boldsymbol{w}_\star) + \frac{V_\star}{2} \right) + \|\boldsymbol{w}_\star\| \right]^2 + \frac{V_\star}{2\sqrt{2}B} \right\}.$$

Also, recall the backward recurrent inequality which we write here again:

$$\left\| \boldsymbol{w}_0^{((j+1)M)} - \boldsymbol{w}_\star \right\|^2 \leq \left\| \boldsymbol{w}_0^{(jM)} - \boldsymbol{w}_\star \right\|^2 - 2\eta K \left[\mathcal{L}\left(\boldsymbol{w}_0^{(jM)} \right) - \mathcal{L}(\boldsymbol{w}_\star) \right] + 2\sqrt{2}\eta^2 K^2 B V_\star. \tag{48}$$

We choose $\boldsymbol{w}_0^{(0)}$ as we want: if $\boldsymbol{w}_0^{(0)} = \boldsymbol{0}$, since $\left\| \boldsymbol{w}_0^{(0)} - \boldsymbol{w}_\star \right\|^2 = \|\boldsymbol{w}_\star\|^2$, it is clear that $\boldsymbol{w}_0^{(0)} \in \mathcal{W}$. Now assume $\boldsymbol{w}_0^{(jM)} \in \mathcal{W}$ and proceed with induction on j: we aim to show $\boldsymbol{w}_0^{((j+1)M)} \in \mathcal{W}$. The proof is divided into two parts:

1. If the current total risk is too high, then we can show that the squared distance to $\boldsymbol{w}_\star$ will decrease.
2. The other case means that the current iterate is close enough to $\boldsymbol{w}_\star$ (due to the strict non-separability of the full dataset). Thus, the squared distance to $\boldsymbol{w}_\star$ at the next cycle will not grow that much.

Part 1: High-Risk Case. Suppose $\mathcal{L}\left(\boldsymbol{w}_0^{(jM)} \right) - \mathcal{L}(\boldsymbol{w}_\star) \geq \sqrt{2}\eta K B V_\star$. Then Equation (48) implies $\left\| \boldsymbol{w}_0^{((j+1)M)} - \boldsymbol{w}_\star \right\|^2 \leq \left\| \boldsymbol{w}_0^{(jM)} - \boldsymbol{w}_\star \right\|^2$. Thus, $\boldsymbol{w}_0^{((j+1)M)} \in \mathcal{W}$.

Part 2: Low-Risk Case. We first show that $\mathcal{L}\left(\boldsymbol{w}_0^{(jM)} \right) - \mathcal{L}(\boldsymbol{w}_\star) \leq \sqrt{2}\eta K B V_\star$ implies an upper bound on the current squared distance to $\boldsymbol{w}_\star$. Because of Assumptions 5.1 and 5.2, for any $\boldsymbol{w} \in \mathbb{R}^d$,

$$\begin{aligned} \mathcal{L}(\boldsymbol{w}) &= \sum_{i=0}^{N-1} \ell\left(\boldsymbol{x}_i^\top \boldsymbol{w} \right) \\ &\geq \sum_{i=0}^{N-1} G \left[\boldsymbol{x}_i^\top \boldsymbol{w} \right]^- \\ &= G \|\boldsymbol{w}\| \cdot \sum_{i=0}^{N-1} \left[\boldsymbol{x}_i^\top \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \right]^- \\ &\geq G \|\boldsymbol{w}\| b, \end{aligned}$$

by the definition of the non-separability $b > 0$. Thus, we have

$$\begin{aligned} \left\| \boldsymbol{w}_0^{(jM)} - \boldsymbol{w}_\star \right\| &\leq \left\| \boldsymbol{w}_0^{(jM)} \right\| + \|\boldsymbol{w}_\star\| \\ &\leq \frac{1}{Gb} \mathcal{L}\left(\boldsymbol{w}_0^{(jM)} \right) + \|\boldsymbol{w}_\star\| \\ &\leq \frac{1}{Gb} \left[\mathcal{L}(\boldsymbol{u}) + \sqrt{2}\eta K B V_\star \right] + \|\boldsymbol{w}_\star\| . \end{aligned}$$

Thus, $\boldsymbol{w}_0^{(jM)}$ lies in a strict subset of $\mathcal{W}$. Also, Equation (48) implies

$$\begin{aligned}\left\|\boldsymbol{w}_0^{((j+1)M)} - \boldsymbol{w}_\star\right\|^2 &\le \left\|\boldsymbol{w}_0^{(jM)} - \boldsymbol{w}_\star\right\|^2 + 2\sqrt{2}\eta^2 K^2 B V_\star \\ &\le \left[\frac{1}{Gb}\left[\mathcal{L}(\boldsymbol{u}) + \sqrt{2}\eta K B V_\star\right] + \|\boldsymbol{w}_\star\|\right]^2 + 2\sqrt{2}\eta^2 K^2 B V_\star.\end{aligned}$$

Thus, by the definition of $\mathcal{W}$, $\boldsymbol{w}_0^{((j+1)M)} \in \mathcal{W}$. This concludes the proof of the lemma.

F.2 Non-asymptotic Loss Convergence Analysis (Proof of Theorem 5.2)

Recall that we write $B = \sum_{m=0}^{M-1} \beta_m$ and $V_\star = \sum_{m=0}^{M-1} \frac{1}{\beta_m} \|\nabla \mathcal{L}_m(\boldsymbol{w}_\star)\|^2$. Also, in the previous sub-section, we discovered a local strong convexity (with coefficient $\mu > 0$) of the total risk function satisfied on a compact ball $\mathcal{W}$ containing $\boldsymbol{w}_\star$ and every end-of-cycle iterates of the sequential GD.

Let us restate the theorem for the sake of readability.

Theorem 5.2. *Suppose we learn M tasks cyclically for $J > 1$ cycles. We adopt the notation from Lemma 5.1. If we choose a step size*

$$\eta = \min\left\{\frac{1}{2\sqrt{2}KB}, \frac{1+2\sqrt{2}}{2\sqrt{2}KJ} \ln\left(J^2 \cdot \max\left\{1, \frac{\|\boldsymbol{w}_0^{(0)} - \boldsymbol{w}_\star\|^2 \mu^3}{B^2 V_\star}\right\}\right)\right\},$$

then the final iterate of sequential GD satisfies

$$\left\|\boldsymbol{w}_0^{(MJ)} - \boldsymbol{w}_\star\right\|^2 \le \tilde{\mathcal{O}}\left(\exp\left(-\frac{\mu J}{(1+2\sqrt{2})B}\right) \cdot \left\|\boldsymbol{w}_0^{(0)} - \boldsymbol{w}_\star\right\|^2 + \frac{B^2 V_\star \ln^2 J}{\mu^3 J^2}\right), \tag{7}$$

where we hide a poly-logarithmic factor of J in Equation (7).

The theorem states a fast $\tilde{\mathcal{O}}(J^{-2})$ rate of convergence. One could try to prove it with the backward recurrent inequality (Equation (45)), but it is difficult due to the η^2 dependency of the so-called “noise” term. We only succeeded in proving a slower $\tilde{\mathcal{O}}(J^{-1})$ rate with the backward recurrent inequality, whose proof is pretty much similar to that in this sub-section. To take a step further towards a faster rate, we should use a different way of writing the recurrent inequality, with a higher exponent for η in the “noise” term. Here is how it goes:

Lemma F.6 (Forward recurrent inequality)**.** *Consider learning M linear classification tasks cyclically. Suppose that Assumption 5.2 holds. If we take any step size satisfying $\eta \le \frac{1}{\sqrt{2}KB}$, the iterates of sequential GD satisfies*

$$\begin{aligned}&\left\|\boldsymbol{w}_0^{((j+1)M)} - \boldsymbol{u}\right\|^2 \\ &\le \left\|\boldsymbol{w}_0^{(jM)} - \boldsymbol{u}\right\|^2 - 2\eta K\left[\mathcal{L}\left(\boldsymbol{w}_0^{((j+1)M)}\right) - \mathcal{L}(\boldsymbol{u})\right] + 2\eta^3 K^3 B^2 \left(\sum_{m=0}^{M-1} \frac{1}{\beta_m} \|\nabla \mathcal{L}_m(\boldsymbol{u})\|^2\right),\end{aligned}$$

for any vector $\boldsymbol{u} \in \mathbb{R}^d$ and for all $j = 0, 1, \cdots$.

Proof. We defer the proof to Appendix F.2.1. We remark that this lemma holds even without assuming the non-separability. □

In particular, we have

$$\left\|\boldsymbol{w}_0^{((j+1)M)} - \boldsymbol{w}_\star\right\|^2 \le \left\|\boldsymbol{w}_0^{(jM)} - \boldsymbol{w}_\star\right\|^2 - 2\eta K\left[\mathcal{L}\left(\boldsymbol{w}_0^{((j+1)M)}\right) - \mathcal{L}(\boldsymbol{w}_\star)\right] + 2\eta^3 K^3 B^2 V_\star. \tag{49}$$

Applying μ-strong convexity, i.e.,

$$\mathcal{L}\left(\boldsymbol{w}_0^{((j+1)M)}\right) - \mathcal{L}(\boldsymbol{w}_\star) \ge \frac{\mu}{2}\left\|\boldsymbol{w}_0^{((j+1)M)} - \boldsymbol{w}_\star\right\|^2,$$

we eventually have a recurrent inequality purely on the squared distance to $\boldsymbol{w}_\star$:

$$\left\| \boldsymbol{w}_0^{((j+1)M)} - \boldsymbol{w}_\star \right\|^2 \le \frac{1}{1+\eta K \mu} \left\| \boldsymbol{w}_0^{(jM)} - \boldsymbol{w}_\star \right\|^2 + 2\eta^3 K^3 B^2 V_\star. \tag{50}$$

We now conclude the proof by applying Proposition F.3: plugging $\alpha \leftarrow \eta K$, $\beta \leftarrow B$, $T \leftarrow J$, $c \leftarrow \frac{1}{2\sqrt{2}}$, $V \leftarrow 2B^2 V_\star$, $m = 2$, and $\Delta_j \leftarrow \left\| \boldsymbol{w}_0^{(jM)} \right\|$ to the proposition, we have a desired result.

F.2.1 Proof of Lemma F.6

For the sake of readability, we restate the lemma, whose proof is very similar to that of Lemma F.4.

Lemma F.6 (Forward recurrent inequality)**.** *Consider learning M linear classification tasks cyclically. Suppose that Assumption 5.2 holds. If we take any step size satisfying $\eta \le \frac{1}{\sqrt{2}KB}$, the iterates of sequential GD satisfies*

$$\begin{aligned}
&\left\| \boldsymbol{w}_0^{((j+1)M)} - \boldsymbol{u} \right\|^2 \\
&\le \left\| \boldsymbol{w}_0^{(jM)} - \boldsymbol{u} \right\|^2 - 2\eta K \left[\mathcal{L}\left(\boldsymbol{w}_0^{((j+1)M)} \right) - \mathcal{L}(\boldsymbol{u}) \right] + 2\eta^3 K^3 B^2 \left(\sum_{m=0}^{M-1} \frac{1}{\beta_m} \left\| \nabla \mathcal{L}_m(\boldsymbol{u}) \right\|^2 \right),
\end{aligned}$$

for any vector $\boldsymbol{u} \in \mathbb{R}^d$ and for all $j = 0, 1, \cdots$.

We start the proof by bounding the squared distance between two iterates in the same cycle of continual learning. For $k \in [0 : K-1]$ and $m \in [0 : M-1]$,

$$\begin{aligned}
&\left\| \boldsymbol{w}_0^{((j+1)M)} - \boldsymbol{w}_k^{(jM+m)} \right\|^2 \\
&= \eta^2 \left\| \sum_{l=m+1}^{M-1} \sum_{h=0}^{K-1} \nabla \mathcal{L}_l(\boldsymbol{w}_h^{(jM+l)}) + \sum_{h=k}^{K-1} \nabla \mathcal{L}_m(\boldsymbol{w}_h^{(jM+m)}) \right\|^2 \\
&\le \eta^2 \left(\sum_{l=0}^{M-1} \sum_{h=0}^{K-1} \beta_l \right) \left(\sum_{l=0}^{M-1} \sum_{h=0}^{K-1} \frac{1}{\beta_l} \left\| \nabla \mathcal{L}_l(\boldsymbol{w}_h^{(jM+l)}) \right\|^2 \right) \\
&\le 2\eta^2 KB \left(\sum_{l=0}^{M-1} \sum_{h=0}^{K-1} \frac{1}{\beta_l} \left[\left\| \nabla \mathcal{L}_l(\boldsymbol{w}_h^{(jM+l)}) - \nabla \mathcal{L}_l(\boldsymbol{u}) \right\|^2 + \left\| \nabla \mathcal{L}_l(\boldsymbol{u}) \right\|^2 \right] \right) \\
&\le 4\eta^2 KB \sum_{l=0}^{M-1} \sum_{h=0}^{K-1} D_{\mathcal{L}_l}(\boldsymbol{u}, \boldsymbol{w}_h^{(jM+l)}) + 2\eta^2 K^2 B \sum_{l=0}^{M-1} \frac{1}{\beta_l} \left\| \nabla \mathcal{L}_l(\boldsymbol{u}) \right\|^2 \qquad (51)
\end{aligned}$$

We use (modified) Jensen's inequality (e.g., Proposition F.2) in the first two inequalities above; the last inequality is due to Proposition F.1.

Next, we decompose the j-th squared distance into $(j+1)$-th squared distance and more:

$$\left\| \boldsymbol{w}_0^{(jM)} - \boldsymbol{u} \right\|^2 \ge \left\| \boldsymbol{w}_0^{((j+1)M)} - \boldsymbol{u} \right\|^2 + 2\eta \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} \left\langle \nabla \mathcal{L}_l(\boldsymbol{w}_k^{(jM+m)}), \boldsymbol{w}_0^{((j+1)M)} - \boldsymbol{u} \right\rangle.$$

Using Equation (44) and β_m-smoothnesses,

$$\begin{aligned}
&\left\| \boldsymbol{w}_0^{((j+1)M)} - \boldsymbol{u} \right\|^2 - \left\| \boldsymbol{w}_0^{(jM)} - \boldsymbol{u} \right\|^2 \\
&\le -2\eta \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} \left[\mathcal{L}_m(\boldsymbol{w}_0^{((j+1)M)}) - \mathcal{L}_m(\boldsymbol{u}) - D_{L_m}(\boldsymbol{w}_0^{((j+1)M)}, \boldsymbol{w}_k^{(jM+m)}) + D_{L_m}(\boldsymbol{u}, \boldsymbol{w}_k^{(jM+m)}) \right] \\
&\le -2\eta K \left[\mathcal{L}\left(\boldsymbol{w}_0^{((j+1)M)} \right) - \mathcal{L}(\boldsymbol{u}) \right] + \eta \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} \beta_m \left\| \boldsymbol{w}_0^{((j+1)M)} - \boldsymbol{w}_k^{(jM+m)} \right\|^2 \\
&\quad - 2\eta \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} D_{L_m}(\boldsymbol{u}, \boldsymbol{w}_k^{(jM+m)})
\end{aligned}$$

$$\begin{aligned}
&\leq -2\eta K \left[\mathcal{L}\left(\boldsymbol{w}_0^{((j+1)M)} \right) - \mathcal{L}(\boldsymbol{u}) \right] + 2\eta^3 K^3 B^2 \sum_{m=0}^{M-1} \frac{1}{\beta_m} \left\| \nabla \mathcal{L}_m(\boldsymbol{u}) \right\|^2 \\
&\quad - 2\eta(1 - 2\eta^2 K^2 B^2) \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} D_{L_m}(\boldsymbol{u}, \boldsymbol{w}_k^{(jM+m)}) \\
&\leq -2\eta K \left[\mathcal{L}\left(\boldsymbol{w}_0^{((j+1)M)} \right) - \mathcal{L}(\boldsymbol{u}) \right] + 2\eta^3 K^3 B^2 \sum_{m=0}^{M-1} \frac{1}{\beta_m} \left\| \nabla \mathcal{L}_m(\boldsymbol{u}) \right\|^2 .
\end{aligned} \tag{52}$$

In Equation (52), we use the result from Equation (51) for multiple times. The last inequality is due to our choice of step size: $\eta K B \leq \frac{1}{\sqrt{2}}$. This is the end of the proof.