Time-Varying Optimization for Streaming Data Via Temporal Weighting

Muhammad Faraz Ul Abrar*, Nicolò Michelusi*, and Erik G. Larsson†

*School of Electrical, Computer and Energy Engineering, Arizona State University
†Department of Electrical Engineering (ISY), Linköping University

Emails: mulabrar@asu.edu, nicolo.michelusi@asu.edu, erik.g.larsson@liu.se

***Abstract*—Classical optimization theory deals with fixed, time-invariant objective functions. However, time-varying optimization has emerged as an important subject for decision-making in dynamic environments. In this work, we study the problem of learning from streaming data through a time-varying optimization lens. Unlike prior works that focus on generic formulations, we introduce a structured, *weight-based* formulation that explicitly captures the streaming-data origin of the time-varying objective, where at each time step, an agent aims to minimize a weighted average loss over all the past data samples. We focus on two specific weighting strategies: (1) uniform weights, which treat all samples equally, and (2) discounted weights, which geometrically decay the influence of older data. For both schemes, we derive tight bounds on the "tracking error" (TE), defined as the deviation between the model parameter and the time-varying optimum at a given time step, under gradient descent (GD) updates. We show that under uniform weighting, the TE vanishes asymptotically with a $\mathcal{O}(1/t)$ decay rate, whereas discounted weighting incurs a nonzero error floor controlled by the discount factor and the number of gradient updates performed at each time step. Our theoretical findings are validated through numerical simulations.**

I. Introduction

The deployment of machine learning (ML) solutions has surged across diverse domains with applications in autonomous vehicles, robotics, telecommunications, power grids, and cyber-physical systems. Conventional ML optimizes a static objective, and therefore inherently assumes a static data distribution. Yet, real-world solutions operate under dynamically evolving environments and must continuously adapt to the streaming information [1]–[3]. Examples include tracking a moving robot, localizing a mobile target, portfolio optimization, risk management in fluctuating financial markets, and adapting a controller with time-varying system dynamics. From a learning perspective, this leads to a streaming data setting in which the objective function evolves over time, resulting in a non-stationary optimization problem. The goal then becomes to track the optimum of a *time-varying* objective function $F_t(\cdot)$: [1], [4], [5]:

$$\overline{\mathbf{w}}_t^* = \arg\min_{\mathbf{w}\in\mathbb{R}^d} F_t(\mathbf{w}), \quad t \geq 0. \tag{1}$$

This research was funded in part by NSF under grant CNS-2129015. The work of E. G. Larsson was supported in part by ELLIIT, VR, and the KAW foundation.

This class of problems is typically approached using iterative optimization techniques such as gradient descent (GD) or the Newton method. However, due to computational constraints in real-time or resource-limited settings [1], [6], only a limited number of updates can typically be performed at each time step, preventing exact tracking of $\overline{\mathbf{w}}_t^*$. Consequently, a widely adopted performance metric in this context is the "tracking error" (TE) $\|\mathbf{w}_t - \overline{\mathbf{w}}_t^*\|$, defined as the distance between the current iterate $\mathbf{w}_t$ and the time-varying optimum $\overline{\mathbf{w}}_t^*$.

Early work in [5] proposed a Newton-type algorithm leveraging second-order information to achieve exponential convergence in gradient norm for time-varying objectives. Subsequent studies have shown that for μ-strongly convex and L-smooth objectives, GD can track the time-varying minimizer $\overline{\mathbf{w}}_t^*$ within an $\mathcal{O}(C)$ neighborhood, assuming a uniform bounded drift condition $\|\overline{\mathbf{w}}_{t+1}^* - \overline{\mathbf{w}}_t^*\| \leq C$ [1], [4]. A different condition was later utilized in [7], where convergence was established under the assumption $\|\nabla F_{t+1}(\mathbf{w}) - \nabla F_t(\mathbf{w})\| \leq C$ for all $\mathbf{w} \in \mathbb{R}^d$. Such "correction-only" schemes rely solely on gradient- or Newton-type updates to track the drifting minimizer [1]. By contrast, prediction–correction schemes first forecast the next optimizer $\widehat{\mathbf{w}}_{t+1}$ using information up to time t (for instance, via first-order optimality condition), and then apply gradient- or Newton-type correction steps once F_{t+1} is revealed [8]. Other approaches for prediction-based algorithms include the Kalman filter-based linear estimation and neural network-based non-linear estimation [9]. More recently, time-varying optimization has been studied in the distributed and decentralized settings (e.g., [10]–[13]). However, most of these works focus on the generic formulation (1), neglecting the inherent structure of streaming data, which may result in loose bounds on the TE.

A closely related paradigm is *continual learning* (CL), which focuses on learning an ML model on a sequence of "tasks" [3], [14]–[16]. Unlike conventional learning, CL assumes no access to future task data and only limited access to past task data samples. This constraint gives rise to the challenge of "catastrophic forgetting", where learning new tasks degrades performance on earlier ones [17]. Although CL also addresses the challenge of adapting models to evolving data, existing approaches remain largely empirical.

In this work, we bridge time-varying optimization and CL from a theoretical standpoint. Specifically, we propose

a structured formulation in which the objective at time t is defined as a weighted average of the losses on all past samples. By explicitly encoding temporal relevance in the weights, our framework captures the streaming data structure and permits tight, weight-specific bounds on TE performance. We discuss two natural weighting strategies: first, uniform weights, which assign equal weight to all past samples and model a stationary environment; second, *discounted weights*, which geometrically decay past sample contributions, prioritizing recent observations. For both schemes, we characterize the TE under GD updates. Exploiting the streaming-data structure yields sharper TE bounds than the existing generic time-varying analyses. Specifically, with uniform weights, we prove that the TE decays as $\mathcal{O}(1/t)$. Under discounted weights, we derive an explicit, nonzero asymptotic TE bound that quantifies the impact of the discount factor and the number of GD iterations per time step.

The rest of the paper is organized as follows. Section II presents the system model. Section III develops the TE analysis for the uniform and discounted weighting schemes. Section IV reports numerical results, and Section V concludes the work.

II. System Model

We consider a learning setup where a single agent (e.g., an edge or cloud server) aims to track a time-varying ML model parameter. In particular, we assume that data arrives sequentially in an online streaming fashion, where at every iteration $t\geq 1$, the agent receives a new data sample $(\mathbf{x}_t, y_t)$. Here, $\mathbf{x}_t$ denotes the feature vector and y_t represents the corresponding label. Let $f_t(\cdot)$ denote the loss associated with the data sample arrived in iteration t, such as the cross-entropy loss evaluated on the t-th sample. The goal is to obtain an ML model that minimizes the *weighted* average loss over the data accumulated thus far, formulated as:

$$\overline{\mathbf{w}}_t^* = \arg\min_{\mathbf{w}\in\mathbb{R}^d} F_t(\mathbf{w}), \text{where } F_t(\mathbf{w}) \triangleq \sum_{i=1}^{t} a_i(t) f_i(\mathbf{w}). \quad (2)$$

Here, $a_i(t)$ are the weights associated with $f_i(\cdot)$ assigned at iteration t. We assume that $0 \leq a_i(t) \leq 1$ for all i and t and $\sum_{i=1}^{t} a_i(t) = 1$ for all t, so that $F_t(\cdot)$ represents a convex combination of the individual losses associated to each data point accumulated up to time t. We note that, unlike the generic time-varying formulation (1), (2) explicitly captures the fact that the objective arises from streaming data.

We employ the gradient descent (GD) method with fixed step size to solve (2). For each time step t, the agent performs E gradient updates, which are initialized with $\mathbf{w}_{t,0} = \mathbf{w}_t$. These updates are of the form:

$$\mathbf{w}_{t,k+1} = \mathbf{w}_{t,k} - \eta \nabla F_{t+1}(\mathbf{w}_{t,k}) \quad (3)$$

$$= \mathbf{w}_{t,k} - \eta \sum_{i=1}^{t+1} a_i(t+1) \nabla f_i(\mathbf{w}_{t,k}), \quad (4)$$

for $k = 0, 1, \cdots, E-1$, where η represents the learning step size. Finally, the updated model is obtained as $\mathbf{w}_{t+1} = \mathbf{w}_{t,E}$. Note that the model update in (4) requires computing the gradients for all historical samples $\{\nabla f_i(\cdot)\}_{i\leq t+1}$. The focus of this work is to understand the fundamental limits of structured time-varying learning, and therefore, we do not assume memory constraints. Under this assumption, $\nabla F_{t+1}(\cdot)$ can be exactly computed at each time step. Analysis of memory-constrained time-varying learning is reserved for future investigation. Since the global objective $F_t(\cdot)$ keeps evolving over time, we characterize the learning performance using the "tracking error" (TE), defined as:

$$\text{TE}(t) = \|\mathbf{w}_t - \overline{\mathbf{w}}_t^*\|, \quad (5)$$

for all $t \geq 1$. Here we are interested in analyzing how the TE evolves over time, and its dependence on E. We are also interested in the asymptotic tracking error (ATE) $\triangleq \limsup_{t\to\infty} \|\mathbf{w}_t - \overline{\mathbf{w}}_t^*\|$[1] to understand the convergence behavior of model updates in (3) and the corresponding rate of convergence. These questions are addressed next.

III. Tracking Error Analysis

In this section, we characterize the tracking error associated with the GD updates in (3) to solve the time-varying learning problem in (2) for a given budget on the number of GD iterations E. In particular, we investigate the TE for two special choices of weights $\{a_i(t)\}$: (a) uniform weights, where each sample is assigned equal weights, and (b) discounted weights, in which past samples are assigned geometrically discounted weights. The first case models a stationary data environment in which all past data samples are equally important. In contrast, the latter captures a scenario in which recent observations are deemed more relevant than older ones, as is common in evolving data dynamics. The TE for both choices is analyzed under the following assumptions:

Assumption 1. *The loss function associated with each data sample is L-smooth (has Lipschitz-continuous gradients) and μ-strongly convex, that is, for all $t = 1, 2, \cdots$, $f_t(\cdot)$ satisfies*

$$\|\nabla f_t(\mathbf{x}) - \nabla f_t(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|, \quad (6)$$

$$f_t(\mathbf{y}) \geq f_t(\mathbf{x}) + \nabla f_t(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2, \quad (7)$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. Since $F_t(\cdot)$ is a convex combination of the losses up to time t, it is also L-smooth and μ-strongly convex for all $t \geq 1$.

Assumption 2. *The minimizers of the data sample losses, $\mathbf{w}_t^* = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f_t(\mathbf{w})$, are uniformly bounded, that is, $\exists\, C > 0$ such that $\|\mathbf{w}_t^*\| \leq C$, for all t.*

Assumption 1 is standard in analyzing gradient-based methods (see, e.g., [18], [19]). In contrast, Assumption 2 (bounded minimizers) is more specific. This assumption was used before in the analysis of time-varying optimization for the federated learning setting (see Assumption 5 in [13]) and is required to make the global optimizer $\overline{\mathbf{w}}_t^*$ drift slowly enough with

[1]Note that we used lim sup, since the limit may not exist.

t. In fact, as we will show in (16) and (22), it implies the often-invoked "bounded drift" assumption $\|\overline{\mathbf{w}}^*_{t+1} - \overline{\mathbf{w}}^*_t\| \leq \bar{C}$ commonly adopted in prior time-varying optimization work (e.g. [1]), but is a more primitive and transparent structural assumption on the problem.

Under Assumption 1, the TE at iteration $t+1$ can be analyzed as follows. From (3), we have

$$\begin{aligned}
&\text{TE}(t+1) = \|\mathbf{w}_{t+1} - \overline{\mathbf{w}}^*_{t+1}\| = \|\mathbf{w}_{t,E} - \overline{\mathbf{w}}^*_{t+1}\| \quad (8)\\
&= \|\mathbf{w}_{t,E-1} - \overline{\mathbf{w}}^*_{t+1} - \eta[\nabla F_{t+1}(\mathbf{w}_{t,E-1}) - \nabla F_{t+1}(\overline{\mathbf{w}}^*_{t+1})]\|\\
&\overset{(a)}{\leq} (1-\eta\mu)\|\mathbf{w}_{t,E-1} - \overline{\mathbf{w}}^*_{t+1}\|\\
&= (1-\eta\mu)\Big\|\mathbf{w}_{t,E-2} - \overline{\mathbf{w}}^*_{t+1} \quad (9)\\
&- \eta[\nabla F_{t+1}(\mathbf{w}_{t,E-2}) - \nabla F_{t+1}(\overline{\mathbf{w}}^*_{t+1})]\Big\|\\
&\overset{(b)}{\leq} (1-\eta\mu)^2\|\mathbf{w}_{t,E-2} - \overline{\mathbf{w}}^*_{t+1}\|\\
&\vdots\\
&\leq (1-\eta\mu)^E\|\mathbf{w}_t - \overline{\mathbf{w}}^*_{t+1}\|, \quad (10)
\end{aligned}$$

where $\mathbf{w}_{t,0} = \mathbf{w}_t$. Steps (a) and (b) leverage the fact that $\nabla F_{t+1}(\overline{\mathbf{w}}^*_{t+1}) = \mathbf{0}$ along with Lemma 1 (provided in the Appendix), which exploits the μ-strong convexity and L-smoothness of $F_{t+1}(\cdot)$ (Assumption 1) for a learning step size choice $\eta \in (0, \frac{2}{\mu+L}]$. The final inequality in (10) is obtained using induction over E gradient updates. Using the triangle inequality, we can further bound the error term in (10) as

$$\begin{aligned}
\|\mathbf{w}_t - \overline{\mathbf{w}}^*_{t+1}\| &\leq \|\mathbf{w}_t - \overline{\mathbf{w}}^*_t\| + \|\overline{\mathbf{w}}^*_{t+1} - \overline{\mathbf{w}}^*_t\|\\
&= \text{TE}(t) + \|\overline{\mathbf{w}}^*_{t+1} - \overline{\mathbf{w}}^*_t\|.
\end{aligned}$$

Hence, continuing from (10) and letting $\alpha \triangleq (1-\eta\mu)^E$, we obtain the following recursive expression on the TE:

$$\text{TE}(t+1) \leq \alpha\left(\text{TE}(t) + \left\|\overline{\mathbf{w}}^*_{t+1} - \overline{\mathbf{w}}^*_t\right\|\right). \quad (11)$$

Remark 1. *The expression in* (11) *reveals that the TE follows a contracting sequence (since $\alpha < 1$) except for the presence of an extra term $\alpha\|\overline{\mathbf{w}}^*_{t+1} - \overline{\mathbf{w}}^*_t\|$ capturing the scaled drift of the minimizers of the time-varying objectives.*

Using recursion, the TE in (11) can be further expressed as

$$\begin{aligned}
\text{TE}(t) &\leq \alpha^{t-1}\text{TE}(1) + \sum_{i=1}^{t-1}\alpha^{t-i}\left\|\overline{\mathbf{w}}^*_{i+1} - \overline{\mathbf{w}}^*_i\right\| \quad (12)\\
&\leq \alpha^t\|\mathbf{w}_0 - \overline{\mathbf{w}}^*_1\| + \sum_{i=1}^{t-1}\alpha^{t-i}\left\|\overline{\mathbf{w}}^*_{i+1} - \overline{\mathbf{w}}^*_i\right\|, \quad (13)
\end{aligned}$$

where we bounded $\text{TE}(1)$ via (10) to obtain the last inequality. The above bound on the TE contains two components: 1) the impact of the initialization, which diminishes geometrically with t, and 2) the error accumulated due to the drift of the minimizers of the time-varying objective functions. Next, we proceed to bound the drift of the minimizers $\left\|\overline{\mathbf{w}}^*_{i+1} - \overline{\mathbf{w}}^*_i\right\|$. Utilizing the μ-strong convexity of $F_{i+1}(\cdot)$, we obtain:

$$\begin{aligned}
\|\overline{\mathbf{w}}^*_{i+1} - \overline{\mathbf{w}}^*_i\| &\leq \frac{1}{\mu}\|\nabla F_{i+1}(\overline{\mathbf{w}}^*_{i+1}) - \nabla F_{i+1}(\overline{\mathbf{w}}^*_i)\|\\
&= \frac{1}{\mu}\|\nabla F_{i+1}(\overline{\mathbf{w}}^*_i)\|, \quad (14)
\end{aligned}$$

where the equality follows due to the optimality condition $\nabla F_{i+1}(\overline{\mathbf{w}}^*_{i+1}) = \mathbf{0}$. Since $F_t(\cdot)$ and the corresponding minimizer $\overline{\mathbf{w}}^*_t$ depends on the choice of weights $\{a_i(t)\}$, further bounding the minimizer drift in (14) necessitates specializing the analysis to specific weighting schemes, which is done in the next two subsections for uniform and discounted weights, respectively.

A. Uniform Weights

A natural strategy is to assign uniform weights to each data sample observed thus far, and hence set $a_i(t) = 1/t$ for all $i = 1, \cdots, t$, and for all t. In this case, we can express $F_{t+1}(\cdot)$ for any $t \geq 1$ as

$$\begin{aligned}
F_{t+1}(\mathbf{w}) &= \sum_{i=1}^{t+1}\frac{f_i(\mathbf{w})}{t+1} = \frac{1}{t+1}\sum_{i=1}^{t} f_i(\mathbf{w}) + \frac{1}{t+1}f_{t+1}(\mathbf{w})\\
&= \frac{t}{t+1}F_t(\mathbf{w}) + \frac{1}{t+1}f_{t+1}(\mathbf{w}). \quad (15)
\end{aligned}$$

Continuing from (14) and using (15), we can bound the minimizer drift as

$$\begin{aligned}
\|\overline{\mathbf{w}}^*_{t+1} - \overline{\mathbf{w}}^*_t\| &\leq \frac{1}{\mu}\left\|\frac{t}{t+1}\nabla F_t(\overline{\mathbf{w}}^*_t) + \frac{1}{t+1}\nabla f_{t+1}(\overline{\mathbf{w}}^*_t)\right\|\\
&= \frac{1}{\mu(t+1)}\|\nabla f_{t+1}(\overline{\mathbf{w}}^*_t)\|\\
&\overset{(a)}{\leq} \frac{L}{\mu(t+1)}\left\|\overline{\mathbf{w}}^*_t - \mathbf{w}^*_{t+1}\right\|\\
&\overset{(b)}{\leq} \frac{L}{\mu(t+1)}\|\overline{\mathbf{w}}^*_t\| + \frac{L}{\mu(t+1)}\left\|\mathbf{w}^*_{t+1}\right\|\\
&\overset{(c)}{\leq} \frac{1}{t+1}\left(1+\sqrt{\frac{L}{\mu}}\right)\frac{LC}{\mu} = \frac{C'}{t+1}, \quad (16)
\end{aligned}$$

where the equality follows due to the optimality condition $\nabla F_t(\overline{\mathbf{w}}^*_t) = \mathbf{0}$, and in the last step we defined $C' \triangleq \left(1+\sqrt{\frac{L}{\mu}}\right)\frac{LC}{\mu}$. Step (a) follows from L-smoothness of $f_{t+1}(\cdot)$ (Assumption 1), (b) uses the triangle inequality, and (c) utilizes Assumption 2 and Lemma 2 provided in the Appendix. Using the bound on the minimizer drift in (16) along with (13), the TE for uniform weights can be expressed as

$$\text{TE}(t) \leq \alpha^t\|\mathbf{w}_0 - \overline{\mathbf{w}}^*_1\| + C'\sum_{i=1}^{t-1}\frac{\alpha^{t-i}}{i+1}. \quad (17)$$

Next, we will show that for large t, the sum in (17) admits an $\mathcal{O}(1/t)$ upper bound.

Proposition 1. *Define $S(t) \triangleq \sum_{i=1}^{t-1}\frac{\alpha^{t-i}}{i+1}$, and let $A = \max\{t_0 S(t_0), \frac{2\alpha}{1-\alpha}\}$, with $t_0 = \lceil\frac{2\alpha}{1-\alpha}\rceil$. Then for all $t \geq t_0$,*

$$S(t) \;\leq\; \frac{A}{t}.$$

Proof. We can express $S(t+1)$ as

$$S(t+1) = \sum_{i=1}^{t} \frac{\alpha^{t+1-i}}{i+1} = \frac{\alpha}{t+1} + \sum_{i=1}^{t-1} \frac{\alpha^{t+1-i}}{i+1}$$
$$= \frac{\alpha}{t+1} + \alpha \sum_{i=1}^{t-1} \frac{\alpha^{t-i}}{i+1} = \frac{\alpha}{t+1} + \alpha S(t). \quad (18)$$

By the definition of A, we have $A \geq t_0 S(t_0)$; therefore it directly holds that $S(t_0) \leq \frac{A}{t_0}$. Next, as induction hypothesis we assume that $S(t) \leq \frac{A}{t}$ for some $t \geq t_0 = \lceil \frac{2\alpha}{1-\alpha} \rceil$. Then, $S(t+1)$ can be upper bounded from (18) as $S(t+1) \leq \frac{\alpha}{t+1} + \alpha \frac{A}{t}$. To prove this, it is sufficient to show that the right-hand side above is bounded by $\frac{A}{t+1}$. Reorganizing the terms, this is equivalent to showing that

$$A\big(1 - \alpha - \frac{\alpha}{t}\big) \geq \alpha.$$

Since $t \geq t_0 \geq 2\alpha/(1-\alpha)$, it follows that $1 - \alpha - \alpha/t \geq (1-\alpha)/2$, hence a sufficient condition to satisfy the previous condition is $A \geq \frac{2\alpha}{1-\alpha}$, which holds by definition of A. We have thus proved that $S(t+1) \leq A/(t+1)$. By induction, the bound holds for all $t \geq t_0$, completing the proof. □

Using Proposition 1, the TE for the uniform weights in (17) can be bounded as

$$\text{TE}(t) \leq \alpha^t \|\mathbf{w}_0 - \overline{\mathbf{w}}_1^*\| + C' \frac{A}{t}, \ \forall t \geq t_0, \quad (19)$$

where A and t_0 are given in the statement of Proposition 1.

Remark 2. *Since the first term in* (19) *decays geometrically, it can be concluded using Proposition 1 that the TE for the uniform weights decays as $\mathcal{O}\left(\frac{1}{t}\right)$ for sufficiently large t, and therefore a vanishing asymptotic TE, i.e., $\lim_{t\to\infty} TE(t) = 0$ is achieved. Moreover, we can lower bound $S(t)$ as $S(t) \geq \frac{1}{t}\sum_{i=1}^{t-1} \alpha^{t-i} = \frac{\alpha}{t}(\frac{1-\alpha^{t-1}}{1-\alpha})$, which also decays as $\mathcal{O}(\frac{1}{t})$ for t large enough. The matching upper and lower bounds confirm that $\mathcal{O}(1/t)$ cannot be improved. Remarkably, the $\mathcal{O}(1/t)$ convergence to the time-varying minimizer $\overline{\mathbf{w}}_t^*$ holds even when the sequence $\{\overline{\mathbf{w}}_t^*\}$ itself is non-convergent.*

B. Discounted Weights

Another strategy is to geometrically discount the samples observed in the past, i.e., $a_i(t) \propto \gamma^{t-i}$ for all $i \leq t$, where $0 < \gamma < 1$ is the discount factor. To ensure that $\sum_{i=1}^{t} a_i(t) = 1$, we normalize the weights yielding:

$$a_i(t) = \frac{1-\gamma}{1-\gamma^t}\gamma^{t-i}, \ \forall\, i = 1, \cdots, t. \quad (20)$$

Accordingly, for the discounted weights, we can express the global objective at iteration $t+1$ as

$$F_{t+1}(\mathbf{w}) = \sum_{i=1}^{t+1} \frac{1-\gamma}{1-\gamma^{t+1}}\gamma^{t+1-i} f_i(\mathbf{w})$$
$$= \sum_{i=1}^{t} \frac{1-\gamma}{1-\gamma^{t+1}}\gamma^{t+1-i} f_i(\mathbf{w}) + \frac{1-\gamma}{1-\gamma^{t+1}} f_{t+1}(\mathbf{w})$$
$$= \gamma \cdot \frac{1-\gamma^t}{1-\gamma^{t+1}} \left(\sum_{i=1}^{t} \frac{1-\gamma}{1-\gamma^t}\gamma^{t-i} f_i(\mathbf{w}) \right) + \frac{1-\gamma}{1-\gamma^{t+1}} f_{t+1}(\mathbf{w})$$
$$= \gamma \cdot \frac{1-\gamma^t}{1-\gamma^{t+1}} F_t(\mathbf{w}) + \frac{1-\gamma}{1-\gamma^{t+1}} f_{t+1}(\mathbf{w}). \quad (21)$$

Using (21), the minimizer drift in (14) specializes as

$$\|\overline{\mathbf{w}}_{t+1}^* - \overline{\mathbf{w}}_t^*\| \leq \frac{1}{\mu} \Big\| \frac{\gamma(1-\gamma^t)}{1-\gamma^{t+1}} \nabla F_t(\overline{\mathbf{w}}_t^*) + \frac{1-\gamma}{1-\gamma^{t+1}} \nabla f_{t+1}(\overline{\mathbf{w}}_t^*) \Big\|$$
$$= \frac{1-\gamma}{\mu(1-\gamma^{t+1})} \|\nabla f_{t+1}(\overline{\mathbf{w}}_t^*)\| \leq \frac{1-\gamma}{1-\gamma^{t+1}} C', \quad (22)$$

with C' defined as in (16), where the equality uses the optimality condition $\nabla F_t(\overline{\mathbf{w}}_t^*) = 0$, and the final inequality follows from Assumptions 1–2, Lemma 2, and the triangle inequality, in direct analogy to the steps used to obtain (16). Using the minimizer drift bound in (22), the TE in (13) specializes to

$$\text{TE}(t) \leq \alpha^t \|\mathbf{w}_0 - \overline{\mathbf{w}}_1^*\| + C'(1-\gamma) \sum_{i=1}^{t-1} \frac{\alpha^{t-i}}{1-\gamma^{i+1}}. \quad (23)$$

Proposition 2. *Define $S(t) \triangleq \sum_{i=1}^{t-1} \frac{(1-\gamma)\alpha^{t-i}}{1-\gamma^{i+1}}$ and let $A_\gamma = \max\{\frac{(1-\gamma^{t_0})S(t_0)}{1-\gamma}, \frac{2\alpha}{1-\alpha}\}$, and $t_0 = \lceil \ln(\frac{1-\alpha}{1+\alpha-2\gamma\alpha}) / \ln(\gamma) \rceil$. Then for all $t \geq t_0$ we have*

$$S(t) \leq \frac{A_\gamma(1-\gamma)}{1-\gamma^t}.$$

Furthermore,

$$\lim_{t\to\infty} S(t) = \frac{(1-\gamma)\alpha}{1-\alpha}.$$

Proof. We begin by computing $S(t+1) = \sum_{i=1}^{t} \frac{(1-\gamma)\alpha^{t+1-i}}{1-\gamma^{i+1}}$, which can also be recursively expressed using $S(t)$ as

$$S(t+1) = \frac{(1-\gamma)\alpha}{1-\gamma^{t+1}} + \sum_{i=1}^{t-1} \frac{(1-\gamma)\alpha^{t+1-i}}{1-\gamma^{i+1}}$$
$$= \frac{(1-\gamma)\alpha}{1-\gamma^{t+1}} + \alpha \sum_{i=1}^{t-1} \frac{(1-\gamma)\alpha^{t-i}}{1-\gamma^{i+1}} = \frac{(1-\gamma)\alpha}{1-\gamma^{t+1}} + \alpha S(t). \quad (24)$$

Note that by definition of A_γ, we have $A_\gamma \geq \frac{(1-\gamma^{t_0})S(t_0)}{1-\gamma}$, therefore it directly holds that $S(t_0) \leq \frac{A_\gamma(1-\gamma)}{1-\gamma^{t_0}}$. Next, we use the induction hypothesis and assume that $S(t) \leq \frac{A_\gamma(1-\gamma)}{1-\gamma^t}$ for some arbitrary $t \geq t_0 = \lceil \ln(\frac{1-\alpha}{1+\alpha-2\gamma\alpha}) / \ln(\gamma) \rceil$. Then, $S(t+1)$ can be upper bounded as $S(t+1) \leq \frac{(1-\gamma)\alpha}{1-\gamma^{t+1}} + \alpha \frac{A_\gamma(1-\gamma)}{1-\gamma^t}$. To prove the induction, it suffices to show that the right-hand side

above is bounded by $\frac{A_\gamma(1-\gamma)}{1-\gamma^{t+1}}$. Equivalently, after reorganizing the terms and simplifying:

$$A_\gamma\Big[1-\alpha-\alpha\gamma^t\frac{1-\gamma}{1-\gamma^t}\Big]\geq\alpha. \tag{25}$$

To show that this condition holds true, note that for $t\geq t_0$ and since $\gamma<1$,

$$\gamma^t\leq\gamma^{t_0}\leq\gamma^{\ln(\frac{1-\alpha}{1+\alpha-2\gamma\alpha})/\ln(\gamma)}=\frac{1-\alpha}{1+\alpha-2\gamma\alpha}.$$

Therefore, we can lower bound the left-hand side of (25) as

$$A_\gamma\Big[1-\alpha-\alpha\gamma^t\frac{1-\gamma}{1-\gamma^t}\Big]$$
$$\geq A_\gamma\Big[1-\alpha-\alpha\frac{1-\alpha}{1+\alpha-2\gamma\alpha}\frac{1-\gamma}{1-\frac{1-\alpha}{1+\alpha-2\gamma\alpha}}\Big]=A_\gamma\frac{1-\alpha}{2},$$

where the inequality is due to the bound on γ^t. Finally, by the definition of A_γ, it holds that $A_\gamma(\frac{1-\alpha}{2})\geq\alpha$, so that (25) is satisfied. We have thus proved that $S(t+1)\leq\frac{A_\gamma(1-\gamma)}{1-\gamma^{t+1}}$. By induction, the bound holds for all $t\geq t_0$.

Next, we apply Lemma 3 from the Appendix to the sequence $S(t)$ governed by (24) with $b_t\triangleq\frac{(1-\gamma)\alpha}{1-\gamma^{t+1}}$. Since $b_t\to(1-\gamma)\alpha$ as $t\to\infty$, it immediately follows from this lemma that $\lim_{t\to\infty}S(t)=\frac{(1-\gamma)\alpha}{1-\alpha}$, which completes the proof. □

Using Proposition 2, the TE for the discounted weights in (23) can be bounded as

$$\text{TE}(t)\leq\alpha^t\|\mathbf{w}_0-\overline{\mathbf{w}}_1^*\|+C'\frac{A_\gamma(1-\gamma)}{1-\gamma^t},\ \forall t\geq t_0 \tag{26}$$

with A_γ and t_0 defined in Proposition 2.

Remark 3. *Since the first term in* (23) *decays geometrically, whereas the second term is a non-vanishing term, it can be concluded using Proposition 2 that, with discounted weights, a non-vanishing ATE is achieved, i.e.,*

$$ATE_\gamma\triangleq\limsup_{t\to\infty}\|\mathbf{w}_t-\overline{\mathbf{w}}_t^*\|\leq\left(1+\sqrt{\frac{L}{\mu}}\right)\frac{LC}{\mu}\cdot\frac{(1-\gamma)\alpha}{1-\alpha}. \tag{27}$$

Remark 4. *Uniform weights treat every past sample equally, so as t grows, the influence of any one new sample on the overall objective decays like $\mathcal{O}(1/t)$. Equivalently, the minimizer drifts by $\mathcal{O}(1/t)$ each step (see* (16)*), and this drift, and hence the TE, vanishes asymptotically. By contrast, with γ-discounting, old samples are exponentially forgotten. The minimizer, therefore, continues to drift by an amount proportional to $1-\gamma=\mathcal{O}(1)$ whenever a new sample arrives (see* (22)*), so the algorithm never fully catches up, causing a non-vanishing ATE. Moreover, it is interesting to note that as $\gamma\to1$ the discounted weights converge to uniform weights, and correspondingly $ATE_\gamma\to0$.*

Proposition 3. *Let $0<\gamma<1$, $0<\eta\leq\frac{2}{\mu+L}$, and $C'>0$. For a given choice of $\epsilon>0$, one can ensure $ATE_\gamma\leq\epsilon$ by performing*

$$E\geq\frac{\ln\left(\frac{\epsilon}{C'(1-\gamma)+\epsilon}\right)}{\ln(1-\eta\mu)} \tag{28}$$

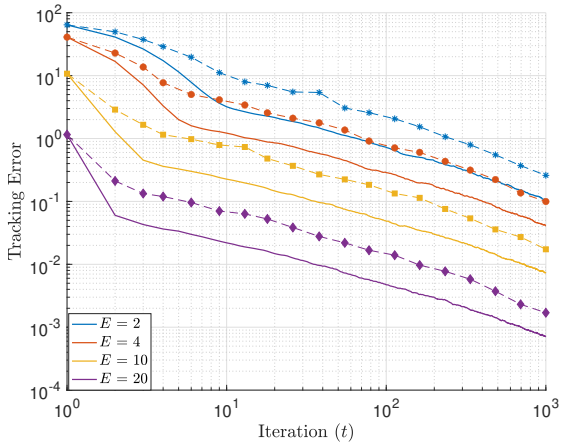

Fig. 1: Root mean squared (**solid lines**) and maximum tracking error (**dashed lines with markers**) vs. iterations for uniform weights, $\mu=0.1,\eta=2$.

gradient updates in each time step. This leads to a gradient iteration complexity of $\mathcal{O}(\ln(\frac{1}{\epsilon}))$ (when $\epsilon\ll C'(1-\gamma)$) to achieve $\epsilon-$ATE.

IV. Numerical Results

In this section, we perform experimentation to numerically verify the efficacy of our TE analysis for the considered time-varying learning problem. We investigate the TE for both the uniform and discounted weights using scalar quadratic loss functions. In particular, the loss function associated with the data arrived in iteration t is given by

$$f_t(w)=\frac{\mu}{2}(w-c_t)^2.$$

It is straightforward to see that $f_t(\cdot)$ is minimized at $w_t^*=c_t$, and the global objective $F_t(\cdot)$ is minimized at $\overline{w}_t^*=\sum_{i=1}^t a_i(t)c_i$, where $\{a_i(t)\}$ denote the weighting coefficients. We generate $\{c_t\}$ according to a bounded random walk process, that has the form:

$$c_{t+1}=\max(-C_{\max},\min(c_t+z_{t+1},C_{\max})), \tag{29}$$

for all $t=0,1,\cdots$, where we let $z_t\sim N(0,\sigma^2)$ for all $t\geq1$, and we set $c_0=0$. We set $C_{\max}=100$, $\sigma^2=100$, and $\mu=0.1$. Note that the above choice of loss functions $\{f_t\}$ ensures that the global objective $F_t(\cdot)$ satisfies both Assumptions 1 and 2 with $C=C_{\max}$.

Fig. 1 plots the root mean squared (RMS) and the maximum (worst-case) TE (over 1000 independent and identical runs) vs. iterations for the case of uniform weights, considering different choices of the number of gradient updates E. As the iteration index t increases, both the RMS and the worst-case TE curves clearly exhibit a monotonic decay consistent with the $\mathcal{O}(1/t)$ rate established in equation (19). As expected, increasing the number of gradient updates E accelerates the TE decay rate. This is because a larger E reduces the contraction factor $\alpha=(1-\eta\mu)^E$, thereby diminishing both the initialization error and

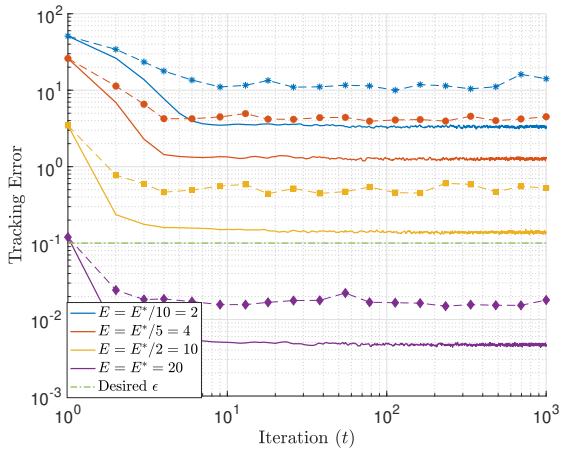

Fig. 2: Root mean squared (**solid lines**) and maximum tracking error (**dashed lines with markers**) vs. iterations for discounted weights, $\gamma = 0.7, \mu = 0.1, \eta = 2.85$.

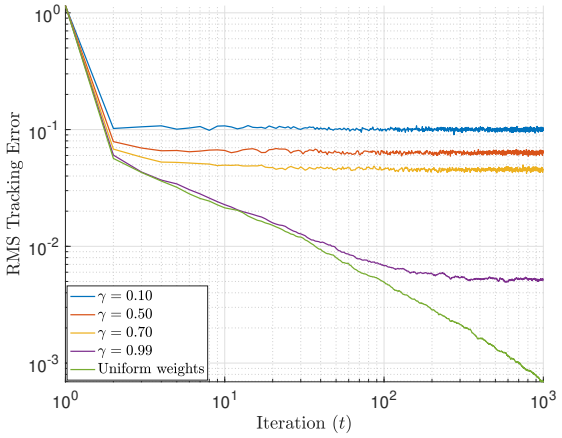

Fig. 3: Root mean squared TE vs. iterations for varying discount factor γ.

the accumulated drift-induced error in (19) more rapidly. We also note that for large t, the TE becomes dominated by the residual error from minimizer drift, which scales as $\mathcal{O}(1/t)$. Notably, doubling E from 10 to 20 reduces the empirical TE by an improvement factor of $(1-\eta\mu)^{20-10}\left[\frac{(1-\eta\mu)^{10}}{(1-\eta\mu)^{20}}\right] \approx 0.09$ (with $\eta = 2, \mu = 0.1$), quantitatively matching the theoretical prediction according to (19).

RMS and the worst-case TE for the discounted case are plotted in Fig. 2 with different choices of the number of gradient updates E. The minimum E^* derived from (28) to guarantee $\epsilon = 0.1$ ATE ensures the error floor remains below ϵ, whereas the considered choices with $E < E^*$ violate the desired condition. Consistent with Remark 3, both error metrics converge to a non-vanishing asymptotic error floor. Finally, we observe that doubling E from 10 to 20 causes a drop of $(1-\eta\mu)^{20-10}\left[\frac{(1-\eta\mu)^{10}}{(1-\eta\mu)^{20}}\right] \approx 0.03$ (with $\eta = 2.85, \mu = 0.1$) in the ATE, which matches with our theoretical bound in (26).

Fig. 3 plots the TE against iterations for the discounted weights for various choices of the discount factor γ. Although all curves approach a nonzero asymptotic error, larger γ yields a lower ATE. This behavior matches the intuition that smaller γ discounts the past samples more heavily, causing the minimizer to drift more rapidly and consequently a higher TE. Fig. 3 also verifies Remark 4 empirically, where for $\gamma = 0.99$, the TE closely matches the uniform weight case.

V. Conclusion

In this paper, we established theoretical guarantees for time-varying optimization specialized to the ML setting where data arrives in an online streaming fashion. We captured this by formulating an objective that evolves as a weighted average of past losses. We analyzed gradient-descent updates under two canonical weighting schemes, uniform weights, which yield a vanishing TE that decays as $\mathcal{O}(1/t)$, and geometrically discounted weights, leading to a non-vanishing asymptotic TE floor explicitly characterized by the discount factor and gradient iterations. Our weight-specific analysis yields tighter, interpretable expressions that clarify how the weighting scheme and the per-time-step gradient update budget shape the asymptotic error, and we confirm these predictions through numerical simulations.

Appendix

Lemma 1. *Let $g : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function that satisfies Assumption 1. Then, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, there exists a positive definite matrix $\mathbf{A}$ such that*

$$\nabla g(\mathbf{y}) - \nabla g(\mathbf{x}) = \mathbf{A}(\mathbf{y} - \mathbf{x}),$$

where the eigenvalues of $\mathbf{A}$ lie in the interval $[\mu, L]$. Moreover, for $0 < \eta \leq \frac{2}{\mu+L}$ it holds that

$$\|\mathbf{y} - \mathbf{x} - \eta\left(\nabla g(\mathbf{y}) - \nabla g(\mathbf{x})\right)\| \leq (1 - \eta\mu)\|\mathbf{y} - \mathbf{x}\|.$$

Proof. This is the mean Hessian theorem, stated and proved in Section II.E of [20]. □

Lemma 2. *Under Assumptions 1 and 2, the set of global minimizers $\{\overline{\mathbf{w}}_t^*\}_{t\geq 1}$ is uniformly bounded and satisfies:*

$$\|\overline{\mathbf{w}}_t^*\|^2 \leq \frac{L}{\mu}C^2, \tag{30}$$

Proof. Using the μ-strong convexity of $F_t(\cdot)$, we have

$$\|\overline{\mathbf{w}}_t^*\|^2 \leq \frac{2}{\mu}\left(F_t(\mathbf{0}) - F_t(\overline{\mathbf{w}}_t^*)\right) \tag{31}$$

Furthermore, by using L-smoothness of $f_i(\cdot)$, we have

$$f_i(\mathbf{0}) \leq f_i(\mathbf{w}_i^*) + \frac{L}{2}\|\mathbf{w}_i^*\|^2 \leq f_i(\mathbf{w}_i^*) + \frac{L}{2}C^2,$$

where the second inequality invokes Assumption 2. Averaging over $i = 1, \cdots, t$ with weights $a_i(t) \geq 0$ and $\sum_{i=1}^t a_i(t) = 1$, we obtain:

$$F_t(\mathbf{0}) \leq \sum_{i=1}^t a_i(t) f_i(\mathbf{w}_i^*) + \frac{L}{2}C^2 \leq \sum_{i=1}^t a_i(t) f_i(\overline{\mathbf{w}}_t^*) + \frac{L}{2}C^2,$$

where the second inequality follows since $\mathbf{w}_i^*$ minimizes $f_i(\cdot)$, which combined with (31) completes the proof. □

Lemma 3. *Let $0 < \alpha < 1$ and let $\{b_t\}_{t\geq 0}$ be a sequence of real numbers that satisfies $b_t \to b^*$ as $t \to \infty$. Define the sequence $\{x_t\}_{t\geq 0}$ by*

$$x_{t+1} = \alpha x_t + b_t, \qquad x_0 \in \mathbb{R}.$$

Then, $\{x_t\}$ is also a convergent sequence which satisfies

$$\lim_{t\to\infty} x_t = \frac{b^*}{1-\alpha}.$$

Proof. By induction on t, we can express x_t as

$$x_t = \alpha^t x_0 + \sum_{k=0}^{t-1} \alpha^{t-1-k} b_k \quad , \quad t \geq 1. \tag{32}$$

Since $b_k \to b^*$, we express b_k as $b_k = b^* + e_k$ with $e_k \triangleq b_k - b^* \to 0$. Subtracting the candidate limit $\frac{b^*}{1-\alpha}$ from both sides of (32), we have

$$\begin{aligned} x_t - \frac{b^*}{1-\alpha} &= \alpha^t x_0 + \sum_{k=0}^{t-1} \alpha^{t-1-k} e_k + b^* \sum_{k=0}^{t-1} \alpha^{t-1-k} - \frac{b^*}{1-\alpha} \\ &= \alpha^t (x_0 - \frac{b^*}{1-\alpha}) + \sum_{k=0}^{t-1} \alpha^{t-1-k} e_k \\ &\triangleq A_t + B_t. \end{aligned}$$

Next, we will show that for every $\varepsilon > 0$ there exists a τ such that $|A_t| + |B_t| < \varepsilon$ for all $t \geq \tau$, that is $x_t \to \frac{b^*}{1-\alpha}$. Let $\varepsilon > 0$ be given. Since $\alpha^t \to 0$, there exists some $\tau_1 > 0$ such that $|A_t| = \alpha^t |x_0 - \frac{b^*}{1-\alpha}| < \varepsilon/2$ for all $t \geq \tau_1$. Next, to bound $|B_t|$, we split the term B_t as $B_t = H_t + T_t$ where

$$H_t \triangleq \sum_{k=0}^{N-1} \alpha^{t-1-k} e_k, \qquad T_t \triangleq \sum_{k=N}^{t-1} \alpha^{t-1-k} e_k, \quad t \geq 1.$$

Define $\delta \triangleq \frac{(1-\alpha)\varepsilon}{4} > 0$. Since $e_k \to 0$, there exists N such that $|e_k| \leq \delta, \forall k \geq N$. With this, the tail term T_t can be bounded as

$$|T_t| \leq \delta \sum_{k=N}^{t-1} \alpha^{t-1-k} \leq \delta \sum_{j=0}^{\infty} \alpha^j = \frac{\delta}{1-\alpha} = \frac{\varepsilon}{4}.$$

To bound the head term H_t, we use the fact that, since $e_t \to 0$, it is bounded by some $\mathcal{E} > 0$ ($|e_t| \leq \mathcal{E}, \forall t$), so that

$$|H_t| \leq \sum_{k=0}^{N-1} \alpha^{t-1-k} |e_k| \leq \mathcal{E} \sum_{k=0}^{N-1} \alpha^{t-1-k} \leq \frac{\mathcal{E}}{1-\alpha} \alpha^{t-N}.$$

Next, since $\alpha^{t-N} \to 0$, as $t \to \infty$, there exists a $\tau_2 \geq N$ such that $|H_t| \leq \varepsilon/4$ for all $t \geq \tau_2$. Finally, choosing $\tau = \max\{\tau_1, \tau_2\}$, the following holds for all $t \geq \tau$,

$$|x_t - \tfrac{b^*}{1-\alpha}| \leq |A_t| + |H_t| + |T_t| < \frac{\varepsilon}{2} + \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \varepsilon.$$

Overall, we have proved that, for any $\varepsilon > 0$, there exists a τ such that $\forall t \geq \tau$, $|x_t - \frac{b^*}{1-\alpha}| < \varepsilon$, thus proving the lemma. □

REFERENCES

[1] A. Simonetto, E. Dall'Anese, S. Paternain, G. Leus, and G. B. Giannakis, "Time-varying convex optimization: Time-structured algorithms and applications," *Proceedings of the IEEE*, vol. 108, no. 11, pp. 2032–2048, 2020.

[2] A. H. Sayed, "Adaptation, learning, and optimization over networks," *Found. Trends Mach. Learn.*, vol. 7, pp. 311–801, 2014.

[3] L. Wang, X. Zhang, H. Su, and J. Zhu, "A comprehensive survey of continual learning: Theory, method and application," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5362–5383, 2024.

[4] B. Polyak, *Introduction to optimization*. Optimization Software, 1987.

[5] Y. Zhao and M. Swamy, "A novel technique for tracking time-varying minimum and its applications," in *Conference Proceedings. IEEE Canadian Conference on Electrical and Computer Engineering (Cat. No.98TH8341)*, vol. 2, 1998, pp. 910–913 vol.2.

[6] E. Dall'Anese, A. Simonetto, S. Becker, and L. Madden, "Optimization and learning with information streams: Time-varying algorithms and applications," *IEEE Signal Processing Magazine*, vol. 37, pp. 71–83, 2019.

[7] A. Y. Popkov, "Gradient methods for nonstationary unconstrained optimization problems," *Autom. Remote Control*, vol. 66, no. 6, p. 883–891, Jun. 2005.

[8] A. Simonetto, A. Mokhtari, A. Koppel, G. Leus, and A. Ribeiro, "A class of prediction-correction methods for time-varying convex optimization," *Trans. Sig. Proc.*, vol. 64, no. 17, p. 4576–4591, Sep. 2016.

[9] A. S. Charles, A. Balavoine, and C. J. Rozell, "Dynamic filtering of time-varying sparse signals via ℓ_1 minimization," *IEEE Transactions on Signal Processing*, vol. 64, no. 21, pp. 5644–5656, 2016.

[10] A. Simonetto, A. Koppel, A. Mokhtari, G. Leus, and A. Ribeiro, "Decentralized prediction-correction methods for networked time-varying convex optimization," *IEEE Transactions on Automatic Control*, vol. 62, pp. 5724–5738, 2016.

[11] C. Sun, M. Ye, and G. Hu, "Distributed time-varying quadratic optimization for multiple agents under undirected graphs," *IEEE Transactions on Automatic Control*, vol. 62, no. 7, pp. 3687–3694, 2017.

[12] S. Rahili and W. Ren, "Distributed continuous-time convex optimization with time-varying cost functions," *IEEE Transactions on Automatic Control*, vol. 62, no. 4, pp. 1590–1605, 2017.

[13] C.-H. Hu, Z. Chen, and E. G. Larsson, "Energy-efficient federated edge learning with streaming data: A Lyapunov optimization approach," *IEEE Transactions on Communications*, 2024.

[14] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 2935–2947, 2016.

[15] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Netw.*, vol. 113, no. C, p. 54–71, May 2019.

[16] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, "A continual learning survey: Defying forgetting in classification tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3366–3385, 2022.

[17] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," *Psychology of Learning and Motivation*, vol. 24, pp. 109–165, 1989.

[18] Y. Nesterov, *Lectures on Convex Optimization*, 2nd ed. Springer Publishing Company, Incorporated, 2018.

[19] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[20] E. G. Larsson and N. Michelusi, "Unified analysis of decentralized gradient descent: a contraction mapping framework," *IEEE Open Journal of Signal Processing*, pp. 1–25, 2025.