

Distributed Optimization with Streaming Data: A Temporal Weighting Perspective

Muhammad Faraz Ul Abrar, *Graduate Student Member, IEEE*, Nicolò Michelusi, *Senior Member, IEEE*,
and Erik G. Larsson, *Fellow, IEEE*

Abstract—Optimization theory is a widely used tool for intelligent decision-making. While classical optimization deals with fixed, time-invariant objective functions, many modern applications operate in dynamic environments where data arrive sequentially, and the learning objective evolves over time, often under decentralized data and communication constraints. Motivated by these trends, we study decentralized optimization from streaming data through a structured time-varying formulation in which the global objective is a temporally weighted average of losses observed across the network. We analyze multi-iteration decentralized first-order methods, including decentralized gradient descent and diffusion, where only a fixed number of communication/gradient steps can be performed before new samples arrive. For strongly convex and smooth losses, we develop guarantees for the Euclidean-norm *tracking error* through a contraction-mapping viewpoint. The resulting bounds decompose the tracking error into a fixed-point tracking component and a bias term induced by decentralization and data heterogeneity. We specialize our analysis to uniform and exponentially discounted weights, as well as their finite-memory *windowed* counterparts. The bounds explicitly characterize the roles of the temporal weighting rule, per-step iteration budget, step size, and network connectivity. Uniform weighting yields a vanishing fixed-point tracking contribution of order $\mathcal{O}(1/t)$, whereas discounted and windowed strategies generally induce non-vanishing tracking floors governed by the discount factor and effective memory, respectively. In all cases, decentralization induces an additional non-zero bias floor under a constant step size. Numerical experiments illustrate the predicted trends.

Index Terms—Decentralized optimization, streaming data, time-varying optimization, DGD, tracking error, online learning.

I. INTRODUCTION

Modern artificial intelligence (AI) systems increasingly rely on learning and optimization tools for intelligent decision-making across domains such as autonomous vehicles, robotics, telecommunications, power grids, and cyber-physical systems [2], [3]. Despite their success, many existing AI systems follow an *optimize-and-deploy* paradigm, where a model is trained for a fixed objective under an implicitly static data distribution. In practice, however, many applications are both *dynamic* and *decentralized*: data arrive sequentially over time [4]–[6], while storage, computation, and communication are distributed across multiple agents due to scale, privacy, or architectural constraints [7]–[10]. Representative examples include mobile target localization, measurement-based network optimization, streaming data processing, and adaptive control under time-varying dynamics [3], [4], [11], [12]. These settings

call for learning systems that continuously adapt to incoming information while operating cooperatively over a network.

Time-varying optimization provides a natural framework for such problems. A network of agents collaboratively tracks the minimizer of an objective that evolves over time, typically while performing only a limited number of optimization and communication steps before the objective changes again [3], [4]. Exact tracking is therefore generally infeasible, and performance is naturally measured through the *tracking error* (TE), defined at each time step as the Euclidean distance between the current iterate and the corresponding optimizer. Existing decentralized time-varying optimization methods have primarily focused on algorithmic developments for improving TE performance [12]–[20]. However, these works treat the objective evolution as generic and largely *unstructured*. In these formulations, the sequence of objectives is usually specified directly as an arbitrary time-varying process and worst-case TE guarantees are derived under generic “minimizer drift” assumptions. While broadly applicable, this viewpoint treats the temporal evolution as *exogenous* and unstructured, and may therefore be conservative when the objective changes according to a specific mechanism.

In many modern learning problems, the objective does not evolve arbitrarily. Rather, it changes because new data samples are continuously acquired across the network, see [21] and references therein. The objective evolution therefore inherits structure from the underlying data-acquisition process. This observation motivates the central question addressed in this paper: *Can the streaming-data structure be explicitly incorporated into the optimization model to obtain sharper and more interpretable tracking guarantees?*

To address this question, in this work, we study decentralized time-varying optimization from a streaming-data perspective. Each agent acquires data online, and the network objective at time t is modeled as a *temporally weighted* average of losses over the samples observed so far. By explicitly encoding the temporal relevance of past data through the weights, the resulting formulation yields *weight-specific* tracking guarantees that sharpen generic worst-case drift-based bounds. Building on this structured formulation, we analyze decentralized first-order methods under limited computation and communication per time step. In particular, we develop a unified TE analysis for a general class of temporal weighting strategies and specialize the results to two natural choices: (i) *uniform* weights, which assign equal importance to all past samples and are natural for stationary environments, and (ii) *exponentially discounted* weights, which geometrically discount old samples and emphasize recent observations. We further specialize the guarantees to their finite-memory *windowed* variants. The resulting bounds explicitly characterize the roles of the temporal

M. Faraz Ul Abrar and N. Michelusi are with the School of Electrical, Computer and Energy Engineering, Arizona State University. email: {mular, nicolo.michelusi}@asu.edu. Erik G. Larsson is with the Department of Electrical Engineering (ISY), Linköping University, 58183 Linköping, Sweden. e-mail: erik.g.larsson@liu.se. This research was funded in part by NSF under grant CNS-2129015. The work of E. G. Larsson was supported in part by ELLIT, VR, and the KAW foundation.

A preliminary version of this work is available in [1].

weighting rule, per-step iteration budget, network connectivity, step size, and data heterogeneity. For smooth and strongly convex losses, our analysis reveals a vanishing fixed-point tracking contribution of order $\mathcal{O}(1/t)$ under uniform weighting and generally non-vanishing TE floors under discounted and finite-memory windowed strategies.

A. Related Work

The field of time-varying optimization traces back to early works on non-stationary optimization [22]–[24], primarily in the single-agent setting. For smooth and strongly convex objectives, classical results establish tracking to a neighborhood whose size depends on a bounded-minimizer-drift or bounded-gradient-variation condition [3], [22], [24], [25]. Methods that solely utilize the objective available at time t to track the drifting minimizer are often termed *correction-only* schemes [3]. In contrast, *prediction-correction* schemes first predict the next optimizer using the objective at time t (e.g., via a first-order optimality condition), and then perform a *correction* update once the objective at time $t + 1$ becomes available, leading to improved asymptotic tracking performance [15], [25]–[27]. Related prediction mechanisms based on parameter estimation, including Kalman-filter and neural-network predictors, have also been studied [28], [29]. Time-varying optimization has also been investigated in decentralized multi-agent settings using primal methods, such as decentralized gradient descent (DGD) [30], and primal-dual methods, such as alternating direction method of multipliers (ADMM) [31]; see, e.g., [13]–[20], [32]. Nonetheless, these works characterize temporal variation through generic worst-case drift measures and treat the objective evolution as exogenous. In contrast, the present formulation ties the evolution of the objective explicitly to streaming data through the temporal weights.

Another closely related research direction is *online learning* [5], [33]–[35]. In online learning, a learner chooses the decision before observing the current loss, and performance is commonly measured through regret against a fixed or time-varying comparator [36]–[38]. The paradigm of *continual learning* similarly considers sequential data or “task” arrivals, with a primary focus on *catastrophic forgetting* and methods for retaining previously learned knowledge [6], [39]–[43]. While these directions are related through their sequential-data viewpoint, they primarily address information availability, regret, or task adaptation. Our focus is instead on *tracking-error* guarantees for decentralized optimization algorithms operating under communication and computation constraints.

Learning from data streams has also been studied in server-coordinated federated settings [44]–[46]. Among these, [45] is closest in spirit to our work, as it considers memory-constrained learning from streaming data and assigns relative importance weights to samples. Yet a key distinction is that [45] adopts a statistical learning viewpoint: it designs the weights to improve *generalization error* under distributional assumptions. In contrast, we provide an optimization-theoretic, *weight-specific* characterization of the TE of decentralized first-order methods, without imposing a data-generating distribution. Moreover, the setting considered in this work is fully decentralized and does not rely on a coordinating server.

B. Contributions and Organization

Our main contributions are summarized as follows:

- We extend our single-agent formulation in [47] to decentralized multi-agent networks and formulate decentralized learning from streaming data samples through temporally weighted objectives. We introduce a structured family of kernel-parameterized temporal weighting strategies, including uniform and exponentially discounted weights, together with their finite-memory windowed variants.
- Leveraging the contraction-mapping framework of [48], we develop a unified TE analysis for multi-iteration DGD and diffusion under limited communication and computation per time step.
- We develop TE guarantees that decompose the overall error into a fixed-point tracking component and a bias term capturing the effect of data heterogeneity across the network. The resulting bounds explicitly characterize the dependence on the per-step iteration budget, network connectivity, step size, and temporal weighting rule.
- We specialize the TE guarantees to four canonical weighting strategies and use numerical experiments to illustrate the predicted effects of key system parameters and temporal-weighting choices.

The remainder of the paper is organized as follows. Section II presents the system model and streaming-data objective. Section III introduces the temporal weighting strategies. Section IV describes the decentralized algorithms and performance metrics. Section V develops the TE analysis. Section VI presents numerical results. Section VII concludes the paper.

C. Notation

Scalars are denoted by italic letters (e.g., a), vectors by boldface lowercase letters (e.g., $\mathbf{w}$), and matrices by boldface uppercase letters (e.g., $\mathbf{A}$). Bold sans-serif lowercase letters are reserved for network-stacked vectors. The all-ones vector in $\mathbb{R}^N$ is denoted by $\mathbf{1}_N$. For a vector or matrix, $(\cdot)^\top$ denotes its transpose, and the Euclidean norm is denoted by $\|\cdot\|$. For a stacked vector $\mathbf{w} = [\mathbf{w}_1^\top \cdots \mathbf{w}_N^\top]^\top \in \mathbb{R}^{Nd}$, the notation $[\mathbf{w}]_n \in \mathbb{R}^d$ denotes its n -th block. The identity matrix of dimension q is denoted by $\mathbf{I}_q \in \mathbb{R}^{q \times q}$, and $\otimes$ denotes the Kronecker product. For a matrix $\mathbf{A}$, $\|\mathbf{A}\|_2$ denotes its spectral norm and $\mathbf{A}^\dagger$ its Moore–Penrose pseudoinverse. The symbols $\mathcal{R}(\mathbf{A})$ and $\mathcal{N}(\mathbf{A})$ denote the range and nullspace of $\mathbf{A}$, respectively. If $\mathbf{A}$ is symmetric, $\mathbf{A} \succeq 0$ means that $\mathbf{A}$ is positive semidefinite, and $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ denote its minimum and maximum eigenvalues. The indicator function of an event ω is denoted by $\mathbf{1}\{\omega\}$. For a scalar $x \in \mathbb{R}$, we define its positive and negative parts as $x_+ \triangleq \max\{x, 0\}$ and $x_- \triangleq \max\{-x, 0\}$. For functions $f(\epsilon)$ and $g(\epsilon)$, we write $f(\epsilon) = \mathcal{O}(g(\epsilon))$ if $f(\epsilon)$ is upper bounded by a constant multiple of $g(\epsilon)$ as $\epsilon \rightarrow 0$, and $f(\epsilon) = \Theta(g(\epsilon))$ if $f(\epsilon)$ is both upper and lower bounded by constant multiples of $g(\epsilon)$ as $\epsilon \rightarrow 0$.

II. SYSTEM MODEL AND PROBLEM FORMULATION

We consider a network of N agents collaboratively solving a time-varying optimization problem induced by streaming data, as shown in Fig. 1. At each time step $t \geq 1$, agent $n \in \{1, \dots, N\}$ observes a new data sample and incurs the

instantaneous loss $\ell_{n,t}(\mathbf{w})$ evaluated at the model parameter $\mathbf{w} \in \mathbb{R}^d$. The network seeks to *track* the minimizer of a suitably constructed weighted average of the accumulated losses (discussed next), under a limited communication/computation budget per time index. The agents communicate over an undirected, connected, time-invariant graph. In particular, the communication among agents is described by a symmetric mixing matrix $\mathbf{M} \in \mathbb{R}^{N \times N}$. Specifically, $[\mathbf{M}]_{m,n}$ is the weight used by agent m to combine information from agent n , and $[\mathbf{M}]_{m,n} = 0$ whenever agents m and n do not directly communicate. We assume that $\mathbf{M}$ satisfies

$$\mathbf{M} = \mathbf{M}^\top, \quad \mathbf{M}\mathbf{1}_N = \mathbf{1}_N. \quad (1)$$

We further assume that the eigenvalues of $\mathbf{M}$ lie in $(-1, 1]$ and are ordered as $1 = \lambda_1 > \lambda_2 \geq \dots \geq \lambda_N > -1$.¹

We next define the streaming-data objective. The network-wide instantaneous loss at time t is

$$F_t(\mathbf{w}) \triangleq \frac{1}{N} \sum_{n=1}^N \ell_{n,t}(\mathbf{w}), \quad (2)$$

capturing the average loss computed across all the agents at time t . Building on the single-agent formulation in [47], we define the time- t objective as a *temporally weighted average* of the past network-wide losses. Let $\{a_i(t)\}_{i=1}^t$ represent a nonnegative temporal weighting strategy that satisfies

$$a_i(t) \in [0, 1], \quad \forall i \leq t, \quad \sum_{i=1}^t a_i(t) = 1 \quad (3)$$

for all $t \geq 1$. Several strategies that satisfy (3) are discussed in the next section. This leads to the following time-varying optimization problem:

$$\bar{\mathbf{w}}_t^* \in \arg \min_{\mathbf{w} \in \mathbb{R}^d} \bar{F}_t(\mathbf{w}), \quad \bar{F}_t(\mathbf{w}) \triangleq \sum_{i=1}^t a_i(t) F_i(\mathbf{w}), \quad (4)$$

defined for all $t \geq 1$. We emphasize that the objective $\bar{F}_t(\cdot)$, and hence its minimizer $\bar{\mathbf{w}}_t^*$, evolves as new samples arrive across the network. Our goal is to characterize how well decentralized methods, such as DGD [30] and diffusion [11], track this moving minimizer when only a limited number of communication/computation iterations can be performed per time step. Before presenting the decentralized algorithms and the associated tracking guarantees, we first discuss the temporal weighting strategies used in (4).

III. TEMPORAL WEIGHTS FOR STREAMING LOSSES

The temporal weights $\{a_i(t)\}_{i=1}^t$ in (4) determine how past samples contribute to the current objective and hence control the stability-plasticity tradeoff in streaming environments [6]. We present two structured ways to generate such weights, and begin with a structured family of temporal weightings induced by kernel sequences, which includes several canonical choices.

¹Sufficient conditions for this spectral property, e.g., in terms of nonnegativity of $\mathbf{M}$ and non-bipartiteness of the underlying graph, are discussed in [48].

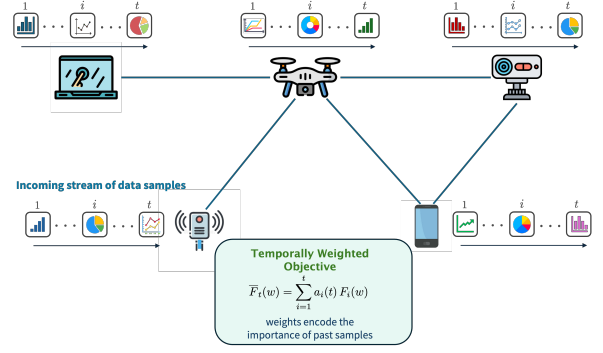


Fig. 1. Illustration of decentralized learning from streaming data. Each agent receives a local stream of samples and exchanges information with neighboring agents over a communication graph. The network tracks the minimizer of a temporally weighted objective formed from past losses.

A. Stationary kernel-induced temporal weights

Let $\{g_k\}_{k \geq 0}$ be a nonnegative kernel sequence with $g_0 > 0$, and define the associated kernel-induced temporal weights for all $t \geq 1$ as

$$a_i(t) \triangleq \frac{g_{t-i}}{S_t}, \quad S_t \triangleq \sum_{k=0}^{t-1} g_k, \quad i = 1, \dots, t. \quad (5)$$

Here, S_t normalizes the weights so the probability simplex constraint (3) holds. Note that the index $k = t - i$ can be interpreted as the *age* of sample i at time t . Since $a_i(t) \propto g_{t-i}$, the kernel $\{g_k\}$ specifies how sample importance varies with age, and (5) implements a simple *shift-and-normalize* weighting rule. Several choices follow as special cases:

- Uniform Weights: Setting $g_k = 1$ for all $k \geq 0$ yields

$$a_i(t) = \frac{1}{t}, \quad i = 1, \dots, t. \quad (6)$$

This rule is natural in stationary environments.

- Exponentially discounted weights: Setting $g_k = \gamma^k$ with a discount factor $\gamma \in (0, 1)$ results in

$$a_i(t) = \frac{(1-\gamma)\gamma^{t-i}}{1-\gamma^t}, \quad i = 1, \dots, t. \quad (7)$$

This rule emphasizes recent samples and can be naturally employed for *adaptation* in non-stationary environments; see, e.g., [4] for several applications.

- Windowed-uniform weights: A finite-memory variant of the uniform weighting strategy assigns equal weight only to the most recent m samples, corresponding to $g_k = \mathbb{1}\{0 \leq k \leq m-1\}$:

$$a_i(t) = \begin{cases} \frac{1}{m}, & t-m+1 \leq i \leq t, \\ 0, & \text{otherwise,} \end{cases} \quad \text{for } t \geq m. \quad (8)$$

- Windowed-discounted weights: Similarly, choosing $g_k = \gamma^k \mathbb{1}\{0 \leq k \leq m-1\}$ yields

$$a_i(t) = \begin{cases} \frac{(1-\gamma)\gamma^{t-i}}{1-\gamma^m}, & t-m+1 \leq i \leq t, \\ 0, & \text{otherwise.} \end{cases} \quad \text{for } t \geq m. \quad (9)$$

This provides a finite-memory approximation of exponentially discounted weighting using only the last m samples.

B. Uniform-shrinkage weighting

The stationary kernel family (5) assigns weights based on age using a fixed time-invariant kernel sequence. Another possible construction is to generate the weights recursively by shrinking all previous weights by the same factor whenever a new sample arrives. Let $\{\beta_t\}_{t \geq 1}$ satisfy $\beta_1 = 1$ and $\beta_t \in (0, 1)$ for $t \geq 2$, and define the time-varying objective function as

$$\bar{F}_{t+1}(\mathbf{w}) = (1 - \beta_{t+1}) \bar{F}_t(\mathbf{w}) + \beta_{t+1} F_{t+1}(\mathbf{w}), \quad (10)$$

for all $t \geq 1$. Equivalently, the induced weights satisfy

$$a_i(t) = \beta_i \prod_{k=i+1}^t (1 - \beta_k), \quad i = 1, \dots, t. \quad (11)$$

This family recovers uniform weighting with $\beta_{t+1} = 1/(t+1)$ and exponential discounting with $\beta_{t+1} = (1 - \gamma)/(1 - \gamma^{t+1})$.

Notably, (11) also allows *time-adaptive* weighting strategies. For example, one may choose β_t to decay from an initially constant level toward $1/t$, so that the objective first emphasizes recent samples for fast adaptation and then gradually approaches uniform weighting for stability. Such schedules depend on absolute time, not only on sample age, and are therefore not generally representable by a stationary kernel. On the other hand, since $\beta_t \in (0, 1)$ preserves strictly positive weights for all past samples, this construction does not capture the finite-window rules in (8)–(9).

IV. ALGORITHM AND PERFORMANCE METRICS

We now present the decentralized methods used to solve (4) and define the associated performance metrics that are adopted throughout the paper. We study two decentralized first-order algorithms: decentralized gradient descent (DGD) [30] and adapt-then-combine (ATC) diffusion [11].² Each agent n maintains a local parameter $\mathbf{w}_{n,t} \in \mathbb{R}^d$ at time t . Upon receiving new data samples, the objective is updated, after which the agents perform $E \geq 1$ decentralized iterations before the time advances. Each decentralized iteration consists of one mixing and one local gradient step. This models a fixed communication/computation budget per time step.

Let $\mathbf{w}_t \in \mathbb{R}^{Nd}$ denote the stacked vector of local parameters, $\mathbf{w}_t \triangleq [\mathbf{w}_{1,t}^\top \dots \mathbf{w}_{N,t}^\top]^\top$ with $\mathbf{w}_{n,t} = [\mathbf{w}_t]_n$. Next, define the block-separable instantaneous objective

$$f_t(\mathbf{w}) \triangleq \sum_{n=1}^N \ell_{n,t}(\mathbf{w}_n),$$

and its temporally weighted version

$$\bar{f}_t(\mathbf{w}) \triangleq \sum_{i=1}^t a_i(t) f_i(\mathbf{w}) = \sum_{i=1}^t a_i(t) \sum_{n=1}^N \ell_{n,i}(\mathbf{w}_n), \quad (12)$$

whose gradient has the block-stacked form, and is given by

$$\nabla \bar{f}_t(\mathbf{w}) = \sum_{i=1}^t a_i(t) \nabla f_i(\mathbf{w}) = \sum_{i=1}^t a_i(t) \begin{bmatrix} \nabla \ell_{1,i}(\mathbf{w}_1) \\ \vdots \\ \nabla \ell_{N,i}(\mathbf{w}_N) \end{bmatrix}. \quad (13)$$

²Throughout the paper, “diffusion” refers to ATC diffusion, which is closely related to combine-then-adapt diffusion; see [48] for a discussion.

The objectives $\bar{F}_t(\cdot)$ in (4) and $\bar{f}_t(\cdot)$ in (12) encode the same collection of sample losses $\{\ell_{n,i}\}$. However, $\bar{F}_t(\cdot)$ is defined over a common decision parameter $\mathbf{w} \in \mathbb{R}^d$, whereas $\bar{f}_t(\cdot)$ is defined over the *stacked collection* of local parameters $\mathbf{w} \in \mathbb{R}^{Nd}$ which need not be consensual. In particular, on the consensus subspace $\{\mathbf{1}_N \otimes \mathbf{w} : \mathbf{w} \in \mathbb{R}^d\} \subseteq \mathbb{R}^{Nd}$, we have $\bar{f}_t(\mathbf{1}_N \otimes \mathbf{w}) = \sum_{i=1}^t a_i(t) \sum_{n=1}^N \ell_{n,i}(\mathbf{w}) = N \bar{F}_t(\mathbf{w})$. Thus, on the consensus subspace, the two objectives differ only by the constant factor N and their minimizers correspond through the mapping $\mathbf{w} \mapsto \mathbf{1}_N \otimes \mathbf{w}$. Finally, we let $\bar{\mathbf{w}}_t^* \triangleq \mathbf{1}_N \otimes \bar{\mathbf{w}}_t^* \in \mathbb{R}^{Nd}$ denote the stacked version of the global minimizer in (4).

After the new samples arrive at time $t+1$, the agents perform E decentralized iterations on the updated objective. For DGD, the k -th iteration at agent n is

$$\mathbf{w}_{n,t}^{(k+1)} = \sum_{m=1}^N [\mathbf{M}]_{n,m} \mathbf{w}_{m,t}^{(k)} - \eta \sum_{i=1}^{t+1} a_i(t+1) \nabla \ell_{n,i}(\mathbf{w}_{n,t}^{(k)}), \quad (14)$$

whereas diffusion performs the adaptation and combination:

$$\tilde{\mathbf{w}}_{n,t}^{(k+1)} = \mathbf{w}_{n,t}^{(k)} - \eta \sum_{i=1}^{t+1} a_i(t+1) \nabla \ell_{n,i}(\mathbf{w}_{n,t}^{(k)}), \quad (15)$$

$$\mathbf{w}_{n,t}^{(k+1)} = \sum_{m=1}^N [\mathbf{M}]_{n,m} \tilde{\mathbf{w}}_{m,t}^{(k+1)}, \quad (16)$$

for $k = 0, \dots, E-1$. Stacking the local parameters across the network, the DGD and diffusion updates can be expressed in the unified form

$$\mathbf{w}_t^{(k+1)} = \bar{\mathbf{M}} \mathbf{w}_t^{(k)} - \eta \bar{\mathbf{Z}} \nabla \bar{f}_{t+1}(\mathbf{w}_t^{(k)}), \quad (17)$$

where $\eta > 0$ is a constant step size, and we define $\bar{\mathbf{M}} \triangleq \mathbf{M} \otimes \mathbf{I}_d$, $\bar{\mathbf{Z}} \triangleq \mathbf{Z} \otimes \mathbf{I}_d \in \mathbb{R}^{Nd \times Nd}$ and

$$\mathbf{Z} \triangleq \begin{cases} \mathbf{I}, & \text{DGD,} \\ \mathbf{M}, & \text{diffusion.} \end{cases} \quad (18)$$

For both methods, $\mathbf{w}_t^{(0)} = \mathbf{w}_t$, and the macro-update (the stacked local parameter update) after E iterations is

$$\mathbf{w}_{t+1} \triangleq \mathbf{w}_t^{(E)}. \quad (19)$$

When $E = 1$, (17)–(19) reduce to the standard single-step DGD/diffusion updates applied to the time-varying objective.

In the considered streaming setting, the objective changes whenever new data arrive; therefore, the stacked iterates $\{\mathbf{w}_t\}$ generally do not converge. We therefore measure performance through the tracking error

$$\text{TE}(t) \triangleq \|\mathbf{w}_t - \bar{\mathbf{w}}_t^*\| = \|\mathbf{w}_t - \mathbf{1}_N \otimes \bar{\mathbf{w}}_t^*\|, \quad (20)$$

and the asymptotic tracking error

$$\text{ATE} \triangleq \limsup_{t \rightarrow \infty} \|\mathbf{w}_t - \bar{\mathbf{w}}_t^*\|. \quad (21)$$

V. TRACKING ERROR ANALYSIS

In this section, we analyze the decentralized updates (17)–(19) to solve the streaming time-varying problem (4). Following [48], we adopt a contraction-mapping viewpoint, which enables a unified TE analysis for both DGD and diffusion. We first derive generic TE bounds and then specialize them to the temporal weighting strategies in Section III. Define the time-varying decentralized update operator $\phi_t: \mathbb{R}^{Nd} \rightarrow \mathbb{R}^{Nd}$ as

$$\phi_t(\mathbf{w}) \triangleq \bar{\mathbf{M}}\mathbf{w} - \eta \bar{\mathbf{Z}} \nabla \bar{f}_t(\mathbf{w}), \quad (22)$$

where $\bar{\mathbf{M}} = \mathbf{M} \otimes \mathbf{I}_d$ and $\bar{\mathbf{Z}} = \mathbf{Z} \otimes \mathbf{I}_d$. The decentralized updates in (17)–(19) with E iterations at time t can then be written as

$$\mathbf{w}_{t+1} = \underbrace{(\phi_{t+1} \circ \dots \circ \phi_{t+1})}_{E \text{ times}}(\mathbf{w}_t) \triangleq \Phi_{t+1}(\mathbf{w}_t), \quad (23)$$

where we use $\Phi_t(\cdot)$ to denote the composition of the mapping $\phi_t(\cdot)$ applied E times. For the analysis, we make the following standard assumptions; see, e.g., [48]–[50].

Assumption 1. For every agent n and time t , the loss $\ell_{n,t}(\cdot)$ is L -smooth and μ -strongly convex. Then, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$\mu \|\mathbf{x} - \mathbf{y}\| \leq \|\nabla \ell_{n,t}(\mathbf{x}) - \nabla \ell_{n,t}(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\| \quad (24)$$

We use $\kappa \triangleq L/\mu$ with $\kappa \geq 1$ to denote the condition number of the optimization problem. Since $F_t(\cdot)$ and $\bar{F}_t(\cdot)$ are convex combinations of the losses at time t and up to time t , respectively, both are also L -smooth and μ -strongly convex for all $t \geq 1$. The same properties also hold for the block-separable objectives $f_t(\cdot)$ and $\bar{f}_t(\cdot)$. That is, for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{Nd}$ and for either choice $f_t \in \{f_t, \bar{f}_t\}$,

$$\|\nabla \tilde{f}_t(\mathbf{u}) - \nabla \tilde{f}_t(\mathbf{v})\| \leq L \|\mathbf{u} - \mathbf{v}\|, \quad (25)$$

$$\tilde{f}_t(\mathbf{v}) \geq \tilde{f}_t(\mathbf{u}) + \nabla \tilde{f}_t(\mathbf{u})^\top (\mathbf{v} - \mathbf{u}) + \frac{\mu}{2} \|\mathbf{v} - \mathbf{u}\|^2. \quad (26)$$

Under Assumption 1, the analysis in Section IV.C of [48] gives the following lemma.

Lemma 1. Suppose Assumption 1 holds. Then, $\phi_t(\cdot)$ is a contraction for every t , and

$$\|\phi_t(\mathbf{u}) - \phi_t(\mathbf{v})\| \leq (1 - \eta\mu) \|\mathbf{u} - \mathbf{v}\|, \quad \forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^{Nd}. \quad (27)$$

under the step size conditions $0 < \eta \leq (1 + \lambda_N)/(L + \mu)$ for DGD and $0 < \eta \leq 2/(L + \mu)$ for diffusion. Consequently, Φ_t is a contraction with factor $\alpha \triangleq (1 - \eta\mu)^E \in (0, 1)$.

Throughout the rest of the analysis, we assume the step size conditions of Lemma 1. Hence, by Banach's fixed-point theorem [51], each $\phi_t(\cdot)$ admits a unique fixed point, given by

$$\tilde{\mathbf{w}}_t \in \mathbb{R}^{Nd} \text{ s.t. } \tilde{\mathbf{w}}_t = \phi_t(\tilde{\mathbf{w}}_t). \quad (28)$$

Since $\Phi_t(\cdot)$ is the E -fold composition of $\phi_t(\cdot)$, $\tilde{\mathbf{w}}_t = \Phi_t(\tilde{\mathbf{w}}_t)$. The sequence $\{\tilde{\mathbf{w}}_t\}$ provides a natural reference trajectory for the analysis: it represents the network state to which the decentralized method would converge if the objective $\bar{f}_t(\cdot)$ were frozen at time t . Leveraging this perspective, the tracking error in (20) can be upper-bounded as

$$\text{TE}(t) = \|\mathbf{w}_t - \bar{\mathbf{w}}_t^*\| \leq \underbrace{\|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|}_{\text{FPTE}(t)} + \underbrace{\|\tilde{\mathbf{w}}_t - \bar{\mathbf{w}}_t^*\|}_{\text{FP bias}}. \quad (29)$$

The *fixed-point tracking error* (FPTE) above captures the ability of the algorithms to track a moving fixed point under a limited per-time-step iteration budget. The second term is the fixed-point bias, capturing the distance between the fixed point and the minimizer at time t .

To bound these two terms, we next impose the following boundedness condition on the sample-wise minimizers.

Assumption 2. For each agent n and time t , let $\mathbf{w}_{n,t}^* \in \arg \min_{\mathbf{w} \in \mathbb{R}^d} \ell_{n,t}(\mathbf{w})$. There exists $C > 0$ such that $\|\mathbf{w}_{n,t}^*\| \leq C$ for all n and t .

Assumption 2 is a structural condition on the sample losses, rather than a direct bounded-drift assumption on the minimizer sequence, as commonly used in time-varying optimization; see, e.g., [3]. Similar bounded-minimizer conditions have appeared in streaming and federated learning analyses [46], [47]. Here, it ensures that the time-varying minimizer $\{\bar{\mathbf{w}}_t^*\}$ and the fixed-point sequences $\{\tilde{\mathbf{w}}_t\}$ are uniformly bounded.

Lemma 2. Under Assumptions 1 and 2, it holds that

$$\|\bar{\mathbf{w}}_t^*\| \leq C\sqrt{\kappa}. \quad (30)$$

Moreover, the fixed points $\{\tilde{\mathbf{w}}_t\}$ satisfy, for all $t \geq 1$,

$$\|\tilde{\mathbf{w}}_t\| \leq \sqrt{N} C_Z, \text{ where } C_Z \triangleq \begin{cases} C\sqrt{\kappa}, & \mathbf{Z} = \mathbf{I} \text{ (DGD)}, \\ C\kappa, & \mathbf{Z} = \mathbf{M} \text{ (diffusion)}, \\ C\sqrt{\kappa}, & \mathbf{Z} = \mathbf{M}, \mathbf{M} \succeq 0. \end{cases} \quad (31)$$

Furthermore, for all $i, t \geq 1$,

$$\|\nabla \bar{f}_t(\tilde{\mathbf{w}}_i)\|, \|\nabla f_t(\tilde{\mathbf{w}}_i)\|, \|\nabla \bar{f}_t(\bar{\mathbf{w}}_t^*)\| \leq G_Z. \quad (32)$$

where $G_Z \triangleq 2L\sqrt{N} C_Z$.

The proof of Lemma 2 can be found in Appendix. We note that the condition $\mathbf{M} \succeq 0$ is not required by the algorithms or by the TE analysis; it only yields a tighter bound in (31) for diffusion.

A. Fixed-point bias

To bound the TE, we first bound the bias term in (29). The following result is obtained by applying the fixed-point-to-minimizer bound of [48, App. A] to $\bar{f}_t(\cdot)$ and then using Lemma 2.

Proposition 1. Under the conditions above,

$$\|\tilde{\mathbf{w}}_t - \bar{\mathbf{w}}_t^*\| \leq \eta\kappa\Lambda_Z \|\nabla \bar{f}_t(\bar{\mathbf{w}}_t^*)\| \leq \eta\kappa\Lambda_Z G_Z, \quad (33)$$

where

$$\Lambda_Z \triangleq \begin{cases} \|(\mathbf{I} - \mathbf{M})^\dagger\|_2 = \frac{1}{1 - \lambda_2}, & \mathbf{Z} = \mathbf{I} \text{ (DGD)}, \\ 2\|(\mathbf{I} - \mathbf{M})^\dagger \mathbf{M}\|_2, & \mathbf{Z} = \mathbf{M} \text{ (diffusion)}. \end{cases} \quad (34)$$

For diffusion, the first inequality additionally requires $\eta \leq 1/(L\Lambda_Z)$.

The bounds in (33) show that the fixed point bias scales linearly with the step size η and depends on the condition number κ , the topology factor Λ_Z , and the gradient magnitude $\|\nabla \bar{f}_t(\bar{\mathbf{w}}_t^*)\|$, which reflects data heterogeneity across agents.

In particular, in homogeneous settings where all agents observe identical losses, they share the same minimizer, and hence $\|\nabla \bar{f}_t(\bar{\mathbf{w}}_t^*)\| = 0$, implying zero bias. Better network connectivity reduces Λ_Z ; for the complete graph $\mathbf{M} = \frac{1}{N}\mathbf{1}\mathbf{1}^\top$, one obtains $\Lambda_Z = 1$ for DGD and $\Lambda_Z = 0$ for diffusion.

B. Fixed-point tracking error and drift

We next bound the fixed-point tracking term in (29). Since $\mathbf{w}_{t+1} = \Phi_{t+1}(\mathbf{w}_t)$ and $\tilde{\mathbf{w}}_{t+1} = \Phi_{t+1}(\tilde{\mathbf{w}}_{t+1})$, the contraction of $\Phi_{t+1}(\cdot)$ from Lemma 1 yields FPTE($t+1$)

$$\begin{aligned} &= \|\mathbf{w}_{t+1} - \tilde{\mathbf{w}}_{t+1}\| \\ &= \|\Phi_{t+1}(\mathbf{w}_t) - \Phi_{t+1}(\tilde{\mathbf{w}}_{t+1})\| \\ &\leq \alpha \|\mathbf{w}_t - \tilde{\mathbf{w}}_{t+1}\| \leq \alpha \text{FPTE}(t) + \alpha \|\tilde{\mathbf{w}}_{t+1} - \tilde{\mathbf{w}}_t\|, \end{aligned} \quad (35)$$

where recall that $\alpha = (1 - \eta\mu)^E$. Unrolling (35) and using the same contraction argument yields

$$\text{FPTE}(t) \leq \alpha^t \|\mathbf{w}_0 - \tilde{\mathbf{w}}_1\| + \sum_{i=1}^{t-1} \alpha^{t-i} \|\tilde{\mathbf{w}}_{i+1} - \tilde{\mathbf{w}}_i\|. \quad (36)$$

In (36), the first term decreases geometrically to zero as $t \rightarrow \infty$, while the second term accumulates the drift of the moving fixed-point $\{\tilde{\mathbf{w}}_t\}$. Thus, controlling the FPTE reduces to controlling this drift. We next bound this drift.

Lemma 3. *Under Assumption 1, the fixed points $\{\tilde{\mathbf{w}}_t\}$ satisfy*

$$\|\tilde{\mathbf{w}}_{t+1} - \tilde{\mathbf{w}}_t\| \leq \frac{1}{\mu} \|\nabla \bar{f}_{t+1}(\tilde{\mathbf{w}}_{t+1}) - \nabla \bar{f}_t(\tilde{\mathbf{w}}_{t+1})\|. \quad (37)$$

The proof is given in Appendix. Lemma 3 relates the fixed-point drift to the temporal variation of the gradient of the time-weighted objective. In particular, if $\nabla \bar{f}_t(\cdot)$ changes slowly with t , then the corresponding fixed-point sequence $\{\tilde{\mathbf{w}}_t\}$ cannot drift rapidly.

Next, using $\nabla \bar{f}_{t+1}(\mathbf{w}) - \nabla \bar{f}_t(\mathbf{w}) = a_{t+1}(t+1)\nabla f_{t+1}(\mathbf{w}) + \sum_{i=1}^t (a_i(t+1) - a_i(t))\nabla f_i(\mathbf{w})$, the gradient bound in Lemma 2 and applying the triangle inequality, (37) implies

$$\|\tilde{\mathbf{w}}_{t+1} - \tilde{\mathbf{w}}_t\| \leq \frac{G_Z}{\mu} \left(a_{t+1}(t+1) + \sum_{i=1}^t |a_i(t+1) - a_i(t)| \right). \quad (38)$$

The following identity further simplifies (38); its proof is given in the Appendix.

Lemma 4. *Let $I_t \triangleq \sum_{i=1}^t (a_i(t+1) - a_i(t))_+$. Then,*

$$\sum_{i=1}^t |a_i(t+1) - a_i(t)| = a_{t+1}(t+1) + 2I_t. \quad (39)$$

Leveraging Lemma 4, the fixed-point drift in (38) can be rewritten as

$$\|\tilde{\mathbf{w}}_{t+1} - \tilde{\mathbf{w}}_t\| \leq \frac{2G_Z}{\mu} (a_{t+1}(t+1) + I_t). \quad (40)$$

Moreover, if the weights are non-increasing over time, i.e., $a_i(t+1) \leq a_i(t), \forall i \leq t$, then $I_t = 0$, and (40) simplifies to

$$\|\tilde{\mathbf{w}}_{t+1} - \tilde{\mathbf{w}}_t\| \leq \frac{2G_Z}{\mu} a_{t+1}(t+1). \quad (41)$$

This property is satisfied by the uniform-shrinkage family in Section III-B; it is also satisfied by the stationary-kernel family in Section III-A whenever the kernel sequence $\{g_k\}$ is non-increasing. Therefore, it holds for the four canonical weighting strategies considered in Section III, whose drift bounds are summarized next.

Proposition 2. *Under (41), the following bounds hold.*

• Uniform weights in (6):

$$\|\tilde{\mathbf{w}}_{t+1} - \tilde{\mathbf{w}}_t\| \leq \frac{2G_Z}{\mu(t+1)}. \quad (42)$$

• Exponentially discounted weights in (7):

$$\|\tilde{\mathbf{w}}_{t+1} - \tilde{\mathbf{w}}_t\| \leq \frac{2G_Z}{\mu} \frac{1-\gamma}{1-\gamma^{t+1}}. \quad (43)$$

• Windowed-uniform weights in (8): For all $t \geq m$,

$$\|\tilde{\mathbf{w}}_{t+1} - \tilde{\mathbf{w}}_t\| \leq \frac{2G_Z}{\mu m}. \quad (44)$$

• Windowed-discounted weights in (9): For all $t \geq m$,

$$\|\tilde{\mathbf{w}}_{t+1} - \tilde{\mathbf{w}}_t\| \leq \frac{2G_Z}{\mu} \frac{1-\gamma}{1-\gamma^m}. \quad (45)$$

Substituting these drift bounds into (36), and combining with the bias bound in (33), yields the TE guarantees discussed next.

C. Tracking error guarantees

We next present the TE bounds specialized to the four canonical weighting strategies. All bounds are stated under the step size conditions of Lemma 1 and, for diffusion, the additional condition in Proposition 1.

Uniform temporal weights: For the uniform temporal weighting strategy (6), utilizing (42), the following holds

$$\text{TE}(t) \leq \alpha^t \|\mathbf{w}_0 - \tilde{\mathbf{w}}_1\| + \frac{2G_Z}{\mu} \sum_{i=1}^{t-1} \frac{\alpha^{t-i}}{i+1} + \eta\kappa\Lambda_Z G_Z. \quad (46)$$

The summation term in (46) admits an $\mathcal{O}(1/t)$ upper bound, yielding the following theorem.

Theorem 1. *Define $S(t) \triangleq \sum_{i=1}^{t-1} \frac{\alpha^{t-i}}{i+1}$ with $\alpha = (1 - \eta\mu)^E \in (0, 1)$. Under uniform weights (6), the tracking error satisfies for all $t \geq t_0 \triangleq \lceil \frac{2\alpha}{1-\alpha} \rceil$*

$$\text{TE}(t) \leq \alpha^t \|\mathbf{w}_0 - \tilde{\mathbf{w}}_1\| + \frac{2G_Z}{\mu} \frac{A}{t} + \eta\kappa\Lambda_Z G_Z, \quad (47)$$

where $A \triangleq \max \left\{ t_0 S(t_0), \frac{2\alpha}{1-\alpha} \right\}$. Furthermore,

$$\limsup_{t \rightarrow \infty} \text{TE}(t) = \text{ATE} \leq \eta\kappa\Lambda_Z G_Z. \quad (48)$$

The proof is given in Appendix. Since $\alpha \in (0, 1)$, the first term in (47) decays geometrically, while the second term decays as $\mathcal{O}(1/t)$. Thus, under uniform weighting, the FPTE contribution vanishes asymptotically, and the TE is ultimately governed solely by the fixed-point bias term, which is $\mathcal{O}(\eta)$. In particular, to guarantee $\text{ATE} = \limsup_{t \rightarrow \infty} \text{TE}(t) \leq \epsilon$, it suffices to choose

$$\eta \leq \frac{\epsilon}{\kappa\Lambda_Z G_Z}. \quad (49)$$

Exponentially discounted weights: For the exponentially discounted temporal weighting strategy (7), utilizing (43) gives

$$\text{TE}(t) \leq \alpha^t \|\mathbf{w}_0 - \tilde{\mathbf{w}}_1\| + \frac{2G_Z}{\mu} \sum_{i=1}^{t-1} \alpha^{t-i} \frac{1-\gamma}{1-\gamma^{i+1}} + \eta\kappa\Lambda_Z G_Z. \quad (50)$$

The summation term in (50) admits a non-vanishing asymptotic bound, yielding the following theorem.

Theorem 2. Define $S_\gamma(t) \triangleq \sum_{i=1}^{t-1} \frac{(1-\gamma)\alpha^{t-i}}{1-\gamma^{i+1}}$ with $\alpha = (1-\eta\mu)^E \in (0, 1)$. Under discounted weights (7), the tracking error satisfies for all $t \geq t_0 \triangleq \lceil \ln(\frac{1-\alpha}{1-\alpha-2\gamma\alpha}) / \ln(\gamma) \rceil$

$$\text{TE}(t) \leq \alpha^t \|\mathbf{w}_0 - \tilde{\mathbf{w}}_1\| + \frac{2G_Z}{\mu} A_\gamma \frac{1-\gamma}{1-\gamma^t} + \eta\kappa\Lambda_Z G_Z, \quad (51)$$

where $A_\gamma \triangleq \max\left\{\frac{(1-\gamma^{t_0})S_\gamma(t_0)}{1-\gamma}, \frac{2\alpha}{1-\alpha}\right\}$. Furthermore,

$$\text{ATE} \leq \frac{2G_Z}{\mu} \frac{(1-\gamma)\alpha}{1-\alpha} + \eta\kappa\Lambda_Z G_Z. \quad (52)$$

The proof is given in Appendix. Unlike uniform weighting, exponentially discounted weighting induces a *non-vanishing* FPTE contribution in addition to the bias floor. This is because old samples are exponentially forgotten, and, hence, the fixed point continues to drift by an amount proportional to $1-\gamma$ (cf. (43)). As $\gamma \rightarrow 1$, this additional floor vanishes and the discounted rule approaches uniform weighting.

For a target $\epsilon > 0$, suppose we choose $0 < \eta \leq \epsilon/(2\kappa\Lambda_Z G_Z)$ so that the bias term $\eta\kappa\Lambda_Z G_Z$ is at most $\epsilon/2$. Then, from (52), it is sufficient to choose

$$E \geq \left\lceil \frac{\ln\left(\frac{\epsilon}{\epsilon + \frac{4G_Z}{\mu}(1-\gamma)}\right)}{\ln(1-\eta\mu)} \right\rceil, \quad (53)$$

to guarantee $\text{ATE} \leq \epsilon$.

Windowed temporal weights: We next consider the finite-memory windowed variants. Unlike the full-memory weightings, windowed strategies retain only a fixed number of past samples, thereby reducing memory and computational requirements. This finite-memory constraint, however, generally induces a non-vanishing FPTE. We first state the result for windowed-discounted weights; the windowed-uniform case is then recovered as the limiting case $\gamma \rightarrow 1$. Using the drift bound in (45), together with the geometric sum $\sum_{i=1}^{t-1} \alpha^{t-i} = \alpha(1-\alpha^{t-1})/(1-\alpha)$, we obtain the following bound.

Theorem 3. Under windowed-discounted weights (9), for all $t \geq m$, the tracking error satisfies

$$\text{TE}(t) \leq \alpha^t \|\mathbf{w}_0 - \tilde{\mathbf{w}}_1\| + \frac{2G_Z}{\mu} \frac{1-\gamma}{1-\gamma^m} \frac{\alpha(1-\alpha^{t-1})}{1-\alpha} + \eta\kappa\Lambda_Z G_Z. \quad (54)$$

Consequently,

$$\text{ATE} \leq \frac{2G_Z}{\mu} \frac{1-\gamma}{1-\gamma^m} \frac{\alpha}{1-\alpha} + \eta\kappa\Lambda_Z G_Z. \quad (55)$$

For windowed-discounted weights, the FPTE contribution in (55) is controlled by the effective-memory factor $(1-\gamma)/(1-\gamma^m)$. As $m \rightarrow \infty$, this factor approaches $1-\gamma$, matching

the bound in (52) for full-memory exponentially discounted weighting.

For a target $\epsilon > 0$, suppose $\eta \leq \epsilon/(2\kappa\Lambda_Z G_Z)$. Then, from (55), it is sufficient to choose

$$E \geq \left\lceil \frac{\ln\left(\frac{\epsilon}{\epsilon + \frac{4G_Z}{\mu} \cdot \frac{1-\gamma}{1-\gamma^m}}\right)}{\ln(1-\eta\mu)} \right\rceil, \quad (56)$$

to guarantee $\text{ATE} \leq \epsilon$.

Since $\lim_{\gamma \rightarrow 1} (1-\gamma)/(1-\gamma^m) = 1/m$, the windowed-uniform guarantees follow directly as a special case.

Corollary 1. Under windowed-uniform weights (8), for all $t \geq m$,

$$\text{TE}(t) \leq \alpha^t \|\mathbf{w}_0 - \tilde{\mathbf{w}}_1\| + \frac{2G_Z}{\mu m} \frac{\alpha(1-\alpha^{t-1})}{1-\alpha} + \eta\kappa\Lambda_Z G_Z. \quad (57)$$

Consequently,

$$\text{ATE} \leq \frac{2G_Z}{\mu m} \frac{\alpha}{1-\alpha} + \eta\kappa\Lambda_Z G_Z. \quad (58)$$

Moreover, for a target $\epsilon > 0$, if $0 < \eta \leq \epsilon/(2\kappa\Lambda_Z G_Z)$, it is sufficient to choose

$$E \geq \left\lceil \frac{\ln\left(\frac{\epsilon}{\epsilon + \frac{4G_Z}{\mu m}}\right)}{\ln(1-\eta\mu)} \right\rceil \quad (59)$$

to guarantee $\text{ATE} \leq \epsilon$.

For fixed η and E , the finite-memory FPTE contribution in (58) has a $1/m$ prefactor, whereas the overall ATE also includes the fixed-point bias term. Hence, increasing the window length reduces the additional finite-memory contribution. In the limit $m \rightarrow \infty$, this contribution vanishes, and the bound reduces to the ATE bound obtained under full-memory uniform weighting in (48).

The iteration complexity needed to achieve $\text{ATE} \leq \epsilon$ is summarized next.

Remark 1. Under uniform weighting, the FPTE contribution vanishes for any $E \geq 1$, and the ATE in (48) is controlled solely by the bias term. Hence, the condition in (49) shows that the ATE can be made arbitrarily small by reducing the step size η . In contrast, for the discounted and windowed strategies, the ATE contains a non-vanishing FPTE floor in addition to the bias term. The sufficient conditions in (53), (56), and (59) reveal a tradeoff between the step size η and the number of decentralized iterations per time-step E . For fixed η , the sufficient threshold on E for achieving an FPTE floor of order ϵ scales as $\Theta(\ln(1/\epsilon))$. However, achieving an arbitrarily small total ATE also requires reducing the bias term, which motivates choosing $\eta = \Theta(\epsilon)$. In this case, $-\ln(1-\eta\mu) = \Theta(\epsilon)$, and the sufficient per-step iteration budgets in (53)–(56) scale as $\Theta(\ln(1/\epsilon)/\epsilon)$.

VI. NUMERICAL RESULTS

We now present numerical experiments illustrating the tracking behavior predicted by the analysis. While the derived bounds are not expected to be tight in general, the experiments

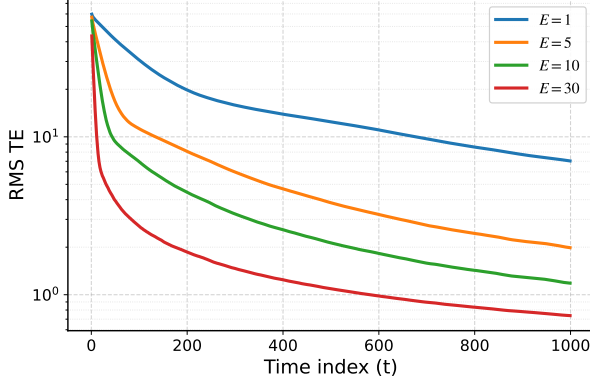


Fig. 2. Diffusion with uniform temporal weights: RMS TE vs. time index for number of iterations per time-step E .

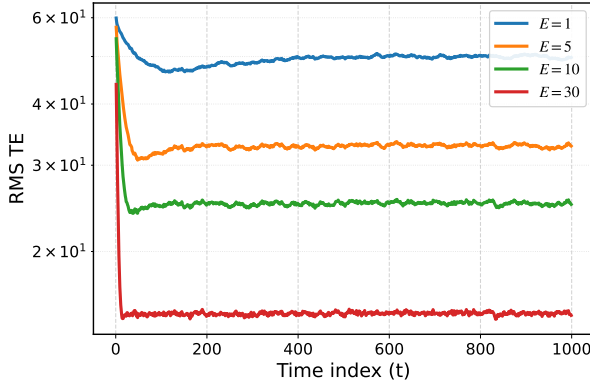


Fig. 3. Diffusion with exponentially discounted temporal weights, $\gamma = 0.7$: RMS TE vs. time index for number of iterations per time-step E .

are intended to show that they are meaningful and capture the qualitative behavior of the considered decentralized system. At time t , agent $n \in \{1, \dots, N\}$ incurs the quadratic loss

$$\ell_{n,t}(\mathbf{w}) = \frac{1}{2}(\mathbf{w} - \mathbf{c}_{n,t})^\top \mathbf{A}_{n,t}(\mathbf{w} - \mathbf{c}_{n,t}), \quad (60)$$

where $\mathbf{c}_{n,t} \in \mathbb{R}^d$ models the streaming data process and $\mathbf{A}_{n,t} \in \mathbb{R}^{d \times d}$ is positive definite. We adopt quadratic losses because they provide a controlled and interpretable setting and admit a closed-form expression for the time-varying minimizer. This choice is appropriate because our objective is not to benchmark a particular learning architecture or dataset, but to study the qualitative implications of the theoretical analysis. We generate $\mathbf{A}_{n,t} = \text{diag}(\lambda_{n,t}^1, \dots, \lambda_{n,t}^d)$, where $\lambda_{n,t}^j \sim \text{Unif}[\mu, L]$ independently across n, t , and j . Hence, $\ell_{n,t}$ in (60) is μ -strongly convex and L -smooth. For this quadratic model, the temporally weighted global objective $\bar{F}_t(\mathbf{w}) = \sum_{i=1}^t a_i(t) F_i(\mathbf{w})$ admits the closed-form minimizer $\bar{\mathbf{w}}_t^* = (\sum_{i=1}^t a_i(t) \sum_{n=1}^N \mathbf{A}_{n,i})^{-1} (\sum_{i=1}^t a_i(t) \sum_{n=1}^N \mathbf{A}_{n,i} \mathbf{c}_{n,i})$, which allows exact evaluation of the TE.

The temporal evolution is generated through a bounded Gaussian random walk. For each coordinate $j = 1, \dots, d$, $[\mathbf{c}_{n,t+1}]_j = \max\{-C_{\max}, \min([\mathbf{c}_{n,t}]_j + [\mathbf{z}_{n,t+1}]_j, C_{\max})\}$, where $\mathbf{z}_{n,t} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ are independent and identically distributed across agents and time. We initialize $\mathbf{c}_{n,0}$ uniformly over $[-C_{\max}, C_{\max}]^d$ and set $\mathbf{w}_{n,0} = \mathbf{0}$ for all n . In all

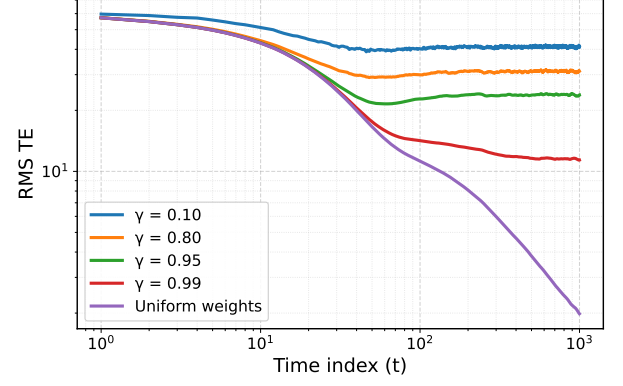


Fig. 4. Diffusion with discounted weights: RMS TE vs. time index for different discount factors γ , with $E = 5$. The uniform-weight curve is included as a benchmark.

experiments, we use $N = 50$ agents, dimension $d = 100$, step size $\eta = 0.1$, $C_{\max} = 10$, $\sigma^2 = 1$, $\mu = 0.01$, and $L = 0.1$. All the experiments are run for $T = 1000$ time-steps. Agents are placed uniformly at random in a disk of radius one, and an undirected random geometric graph is formed by connecting agents within distance $r_{\text{th}} = 0.9$, increased if needed to ensure connectivity. Given the connected graph, we construct the Metropolis-Hastings mixing matrix

$$[\mathbf{M}]_{ij} = \begin{cases} \frac{1}{1 + \max\{\deg(i), \deg(j)\}}, & (i, j) \in \mathcal{E}, i \neq j, \\ 1 - \sum_{j \neq i} [\mathbf{M}]_{ij}, & i = j, \\ 0, & \text{otherwise,} \end{cases} \quad (61)$$

where $\deg(i)$ is the degree of agent i . This construction yields a symmetric doubly stochastic mixing matrix satisfying the assumptions in Section II. All curves report the root-mean-squared (RMS) TE $\sqrt{\frac{1}{R} \sum_{r=1}^R \text{TE}_r(t)^2}$, averaged over $R = 100$ Monte Carlo runs, where $\text{TE}_r(t)$ is computed according to (20). The communication graph is generated once and kept fixed across all Monte Carlo runs; randomness across runs comes from the streaming data process and loss parameters.

Figures 2 and 3 compare different numbers of decentralized diffusion iterations per time-step E for uniform and exponentially discounted temporal weights, respectively. Under uniform weighting, the influence of each new sample decreases with time, leading to a decaying fixed-point drift and a vanishing FPTE. Increasing E accelerates the transient decay through the smaller contraction factor $\alpha = (1 - \eta\mu)^E$. The TE therefore decreases steadily before approaching the bias-dominated regime predicted by Theorem 1. By contrast, under discounted weighting with $\gamma = 0.7$, the TE settles to a substantially larger steady-state level, consistent with Theorem 2. This behavior arises because old samples are exponentially forgotten, so the fixed-point drift does not vanish, and the TE contains a non-zero FPTE contribution in addition to the data-heterogeneity-induced bias. Larger E lowers this floor by improving tracking of the moving fixed point.

Figure 4 studies the effect of the discount factor γ for diffusion with fixed $E = 5$. Smaller γ places more emphasis

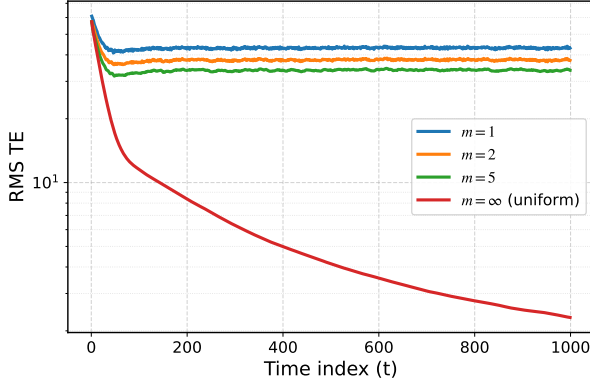


Fig. 5. DGD with windowed-uniform temporal weights: RMS TE vs. time index for different window lengths m , with $E = 5$. The case $m = \infty$ corresponds to standard uniform weighting.

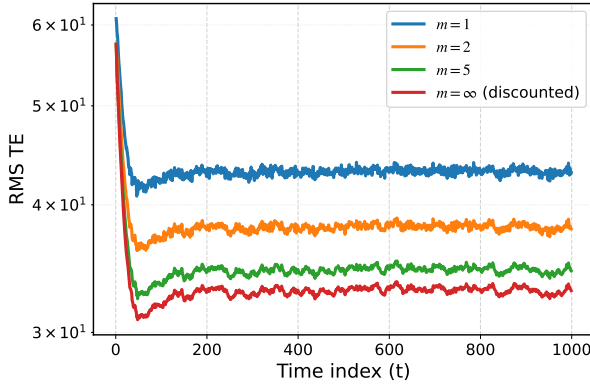


Fig. 6. DGD with windowed-discounted temporal weights, $\gamma = 0.7$: RMS TE vs. time index for different window lengths m , with $E = 5$. The case $m = \infty$ corresponds to standard discounted weighting.

on recent samples, making the effective objective more adaptive but also more variable; this leads to a larger asymptotic TE, consistent with Theorem 2. As γ increases, however, the *effective memory* grows, and the TE floor decreases. In the limit $\gamma \rightarrow 1$, discounted weighting approaches uniform weighting, explaining why the curve for $\gamma = 0.99$ closely follows the uniform-weight curve. Moreover, the continued decay of the uniform-weight curve is consistent with Theorem 1: after the geometrically decaying initialization transient, the FPTE contribution decreases as $\mathcal{O}(1/t)$, until the overall TE becomes limited by the constant-step-size bias floor.

Figures 5 and 6 illustrate finite-memory windowing with DGD. For windowed-uniform weights, increasing the window length m lowers the TE floor, in agreement with the $1/m$ dependence in Corollary 1. In contrast to the fixed-window curves, the full-memory uniform curve, corresponding to $m = \infty$, continues to decrease. This behavior is consistent with Theorem 1, which gives an $\mathcal{O}(1/t)$ FPTE contribution under full-memory uniform weighting. For windowed-discounted weights, increasing m improves performance only up to the effective memory induced by the discount factor. Notably, due to exponential forgetting, even moderate window lengths, e.g., $m > 5$, nearly match the full-memory discounted curve, consistent with the factor $(1 - \gamma)/(1 - \gamma^m)$ in (55).

VII. CONCLUSION

We investigated decentralized optimization with streaming data through a temporal-weighting formulation. For both DGD and diffusion, we developed a unified analysis that yields explicit tracking guarantees by decomposing the tracking error into a fixed-point tracking component and a network-induced bias term. We also introduced a structured class of kernel-based temporal weights, encompassing uniform and exponentially discounted rules together with memory-efficient windowed variants. Specializing the bounds, we showed that uniform weighting yields a vanishing fixed-point tracking contribution of order $\mathcal{O}(1/t)$, leaving only the decentralization-induced bias floor, whereas discounted and windowed schemes generally exhibit non-vanishing error floors governed by the discount factor and the window length, respectively. The resulting bounds explicitly show how temporal weighting, network topology, and the per-time-step iteration budget jointly determine the attainable asymptotic tracking performance. Numerical experiments corroborate the qualitative trends predicted by the theory.

APPENDIX

Lemma 5 (Mean Hessian Theorem [48]). *Let $g : \mathbb{R}^{Nd} \rightarrow \mathbb{R}$ be differentiable (but not necessarily twice differentiable), μ -strongly convex, and L -smooth. Then, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{Nd}$, there exists a symmetric matrix $\mathbf{A}_{\mathbf{x}, \mathbf{y}} \in \mathbb{R}^{Nd \times Nd}$ (dependent on $\mathbf{x}$ and $\mathbf{y}$) such that*

$$\mu \mathbf{I} \preceq \mathbf{A}_{\mathbf{x}, \mathbf{y}} \preceq L \mathbf{I}, \quad (62)$$

$$\nabla g(\mathbf{y}) - \nabla g(\mathbf{x}) = \mathbf{A}_{\mathbf{x}, \mathbf{y}}(\mathbf{y} - \mathbf{x}). \quad (63)$$

The proof of Lemma 5 can be found in [48].

Lemma 6. *Let $0 < \alpha < 1$ and let $\{b_t\}_{t \geq 0} \subset \mathbb{R}$ be a sequence that satisfies $b_t \rightarrow b^*$ as $t \rightarrow \infty$. Define the sequence $\{x_t\}_{t \geq 0}$ by*

$$x_{t+1} = \alpha x_t + b_t, \quad x_0 \in \mathbb{R}. \quad (64)$$

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

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

Proof. Since $b_t \rightarrow b^*$, we express b_t as $b_t = b^* + e_t$ with $e_t \triangleq b_t - b^* \rightarrow 0$. Subtracting the candidate limit $\frac{b^*}{1 - \alpha}$ from both sides of (64), letting $\delta_t = x_t - \frac{b^*}{1 - \alpha}$ gives $\delta_{t+1} = \alpha \delta_t + e_t$. Using induction on t , we further obtain

$$\delta_t = \alpha^t \delta_0 + \sum_{k=0}^{t-1} \alpha^{t-1-k} e_k. \quad (65)$$

Next, we will show that for every $\varepsilon > 0$ there exists a τ such that $|\delta_t| < \varepsilon$, $\forall t \geq \tau$, which implies $\delta_t \rightarrow 0$, hence $x_t \rightarrow \frac{b^*}{1 - \alpha}$. Let $\varepsilon > 0$ be given. Since $e_t \rightarrow 0$, there exists N such that $|e_t| \leq (1 - \alpha)\varepsilon/2$ $\forall t \geq N$, and, since every convergent sequence is bounded, there exists $\mathcal{E} > 0$ such that $|e_t| \leq \mathcal{E}$, $\forall t$. Under such N , for $t \geq N$, we rewrite δ_t as

$$\delta_t = \alpha^t \delta_0 + \underbrace{\sum_{k=0}^{N-1} \alpha^{t-1-k} e_k}_{H_t} + \underbrace{\sum_{k=N}^{t-1} \alpha^{t-1-k} e_k}_{T_t}.$$

Since $|e_k| \leq (1 - \alpha)\varepsilon/2, \forall k \geq N$, we bound $|T_t|$ as

$$|T_t| \leq (1 - \alpha) \frac{\varepsilon}{2} \sum_{k=N}^{t-1} \alpha^{t-1-k} \leq (1 - \alpha) \frac{\varepsilon}{2} \sum_{j=0}^{\infty} \alpha^j = \frac{\varepsilon}{2}.$$

Since $|e_t| \leq \varepsilon, \forall t$, it follows that

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

Therefore, for all $t \geq N$, it holds that

$$|\delta_t| \leq \alpha^t |\delta_0| + \frac{\varepsilon}{1 - \alpha} \alpha^{t-N} + \frac{\varepsilon}{2}. \quad (66)$$

Since $\alpha^t \rightarrow 0$, there exists some $\tau \geq N$ such that $\alpha^t |\delta_0| < \varepsilon/4$ and $\frac{\varepsilon}{1 - \alpha} \alpha^{t-N} < \varepsilon/4$, for all $t \geq \tau$, hence $|\delta_t| < \varepsilon, \forall t \geq \tau$. Hence $\delta_t \rightarrow 0$, which proves the lemma. $\square$

Proof of Lemma 3. Fix any $t \geq 1$. From the definition of $\phi_t(\cdot)$ in (22) and the fixed-point relation (28), we have

$$(\mathbf{I} - \bar{\mathbf{M}})\tilde{\mathbf{w}}_t + \eta \bar{\mathbf{Z}} \nabla \bar{f}_t(\tilde{\mathbf{w}}_t) = \mathbf{0}, \quad (67)$$

where $\bar{\mathbf{Z}} = \mathbf{I}_{Nd}$ for DGD and $\bar{\mathbf{Z}} = \bar{\mathbf{M}} = \mathbf{M} \otimes \mathbf{I}_d$ for diffusion. For notational convenience, define $\mathbf{d}_t \triangleq \tilde{\mathbf{w}}_{t+1} - \tilde{\mathbf{w}}_t$.

• *DGD ($\bar{\mathbf{Z}} = \mathbf{I}$):* Writing (67) at times t and $t + 1$ and subtracting, we obtain:

$$(\mathbf{I} - \bar{\mathbf{M}})\mathbf{d}_t + \eta(\nabla \bar{f}_{t+1}(\tilde{\mathbf{w}}_{t+1}) - \nabla \bar{f}_t(\tilde{\mathbf{w}}_t)) = \mathbf{0}. \quad (68)$$

Applying Lemma 5 to $\bar{f}_t(\cdot)$ at the points $\tilde{\mathbf{w}}_t$ and $\tilde{\mathbf{w}}_{t+1}$, there exists a symmetric matrix $\mathbf{A}$ with $\mu \mathbf{I} \preceq \mathbf{A} \preceq L \mathbf{I}$ such that

$$\nabla \bar{f}_t(\tilde{\mathbf{w}}_t) - \nabla \bar{f}_t(\tilde{\mathbf{w}}_{t+1}) = \mathbf{A}(\tilde{\mathbf{w}}_t - \tilde{\mathbf{w}}_{t+1}). \quad (69)$$

Substituting (69) into (68) gives

$$[(\mathbf{I} - \bar{\mathbf{M}}) + \eta \mathbf{A}]\mathbf{d}_t = -\eta(\nabla \bar{f}_{t+1}(\tilde{\mathbf{w}}_{t+1}) - \nabla \bar{f}_t(\tilde{\mathbf{w}}_{t+1})).$$

Since the eigenvalues of $\bar{\mathbf{M}} = \mathbf{M} \otimes \mathbf{I}_d$ are those of $\mathbf{M}$, each repeated d times, and $\lambda_{\max}(\mathbf{M}) = 1$, it follows that $\mathbf{I} - \bar{\mathbf{M}} \succeq 0$. Moreover, since $\mathbf{A} \succeq \mu \mathbf{I}$, we have $\lambda_{\min}((\mathbf{I} - \bar{\mathbf{M}}) + \eta \mathbf{A}) \geq \eta \mu$, and therefore it holds that

$$\|\mathbf{d}_t\| \leq \frac{\eta}{\eta \mu} \|\nabla \bar{f}_{t+1}(\tilde{\mathbf{w}}_{t+1}) - \nabla \bar{f}_t(\tilde{\mathbf{w}}_{t+1})\|$$

which yields (37) for any $\eta > 0$.

• *Diffusion ($\bar{\mathbf{Z}} = \bar{\mathbf{M}}$):* Specializing (67) yields

$$(\mathbf{I} - \bar{\mathbf{M}})\tilde{\mathbf{w}}_t + \eta \bar{\mathbf{Z}} \nabla \bar{f}_t(\tilde{\mathbf{w}}_t) = \mathbf{0}.$$

Subtracting the relations at t and $t + 1$ and rearranging yields

$$\mathbf{d}_t = \bar{\mathbf{M}}\mathbf{d}_t - \eta \bar{\mathbf{M}}(\nabla \bar{f}_{t+1}(\tilde{\mathbf{w}}_{t+1}) - \nabla \bar{f}_t(\tilde{\mathbf{w}}_t)). \quad (70)$$

Next, applying Lemma 5 and utilizing (69), (70) becomes

$$\mathbf{d}_t = \bar{\mathbf{M}}((\mathbf{I} - \eta \mathbf{A})\mathbf{d}_t - \eta(\nabla \bar{f}_{t+1}(\tilde{\mathbf{w}}_{t+1}) - \nabla \bar{f}_t(\tilde{\mathbf{w}}_{t+1}))).$$

Taking norms and using $\|\bar{\mathbf{M}}\|_2 = \|\mathbf{M} \otimes \mathbf{I}_d\|_2 = \|\mathbf{M}\|_2 \|\mathbf{I}_d\|_2 = 1$ (since $\mathbf{M}$ is symmetric and doubly stochastic) gives

$$\|\mathbf{d}_t\| \leq \|\mathbf{I} - \eta \mathbf{A}\|_2 \|\mathbf{d}_t\| + \eta \|\nabla \bar{f}_{t+1}(\tilde{\mathbf{w}}_{t+1}) - \nabla \bar{f}_t(\tilde{\mathbf{w}}_{t+1})\|.$$

Next, it can be shown that under $0 < \eta \leq 2/(L + \mu)$, we have $\|\mathbf{I} - \eta \mathbf{A}\|_2 = \max\{|1 - \eta \mu|, |1 - \eta L|\} \leq 1 - \eta \mu$, so

$$\eta \mu \|\mathbf{d}_t\| \leq \eta \|\nabla \bar{f}_{t+1}(\tilde{\mathbf{w}}_{t+1}) - \nabla \bar{f}_t(\tilde{\mathbf{w}}_{t+1})\|,$$

which again yields (37). $\square$

Proof of Lemma 4. For notational convenience, define $\Delta_i(t) \triangleq a_i(t + 1) - a_i(t)$, for $i = 1, \dots, t$. Then, we have

$$\begin{aligned} \sum_{i=1}^t \Delta_i(t) &= \sum_{i=1}^t a_i(t + 1) - \sum_{i=1}^t a_i(t) \\ &= (1 - a_{t+1}(t + 1)) - 1 = -a_{t+1}(t + 1). \end{aligned} \quad (71)$$

Next, we can express $\Delta_i(t) = (\Delta_i(t))_+ - (\Delta_i(t))_-$. Summing over $i = 1, \dots, t$ yields

$$\sum_{i=1}^t (\Delta_i(t))_+ - \sum_{i=1}^t (\Delta_i(t))_- = -a_{t+1}(t + 1),$$

where the equality follows from (71). Hence,

$$\sum_{i=1}^t (\Delta_i(t))_- = \sum_{i=1}^t (\Delta_i(t))_+ + a_{t+1}(t + 1) = I_t + a_{t+1}(t + 1),$$

since $I_t = \sum_{i=1}^t (\Delta_i(t))_+$. Leveraging $|\Delta_i(t)| = (\Delta_i(t))_+ + (\Delta_i(t))_-$, we have

$$\begin{aligned} \sum_{i=1}^t |\Delta_i(t)| &= \sum_{i=1}^t ((\Delta_i(t))_+ + (\Delta_i(t))_-) \\ &= I_t + (I_t + a_{t+1}(t + 1)) = a_{t+1}(t + 1) + 2I_t, \end{aligned}$$

which proves (39). $\square$

Proof of Lemma 2. We prove the three claims in (30)–(32) as follows:

• *Bound on the global minimizer $\bar{\mathbf{w}}_t^*$:* Since $\bar{F}_t(\cdot)$ is μ -strongly convex (Assumption 1) with minimizer $\bar{\mathbf{w}}_t^*$, we have, for any $\mathbf{w} \in \mathbb{R}^d$,

$$\bar{F}_t(\mathbf{w}) \geq \bar{F}_t(\bar{\mathbf{w}}_t^*) + \frac{\mu}{2} \|\mathbf{w} - \bar{\mathbf{w}}_t^*\|^2.$$

Specializing the above inequality at $\mathbf{w} = \mathbf{0}$, we obtain:

$$\|\bar{\mathbf{w}}_t^*\|^2 \leq \frac{2}{\mu} (\bar{F}_t(\mathbf{0}) - \bar{F}_t(\bar{\mathbf{w}}_t^*)). \quad (72)$$

Using the definition of $\bar{F}_t(\mathbf{w}) = \sum_{i=1}^t a_i(t) F_i(\mathbf{w})$ and $F_i(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^N \ell_{n,i}(\mathbf{w})$, we have

$$\begin{aligned} \bar{F}_t(\mathbf{0}) - \bar{F}_t(\bar{\mathbf{w}}_t^*) &= \sum_{i=1}^t a_i(t) \frac{1}{N} \sum_{n=1}^N (\ell_{n,i}(\mathbf{0}) - \ell_{n,i}(\bar{\mathbf{w}}_t^*)) \\ &\leq \sum_{i=1}^t a_i(t) \frac{1}{N} \sum_{n=1}^N (\ell_{n,i}(\mathbf{0}) - \ell_{n,i}(\mathbf{w}_{n,i}^*)), \end{aligned}$$

where the inequality follows since $\mathbf{w}_{n,i}^* = \arg \min_{\mathbf{w} \in \mathbb{R}^d} \ell_{n,i}(\mathbf{w})$. Furthermore, utilizing L -smoothness (Assumption 1) and optimality of $\mathbf{w}_{n,i}^*$, it follows that

$$\ell_{n,i}(\mathbf{0}) - \ell_{n,i}(\mathbf{w}_{n,i}^*) \leq \frac{L}{2} \|\mathbf{0} - \mathbf{w}_{n,i}^*\|^2 \leq \frac{L}{2} C^2,$$

where the second inequality invokes Assumption 2. Next, using $\sum_{i=1}^t a_i(t) = 1$ gives

$$\bar{F}_t(\mathbf{0}) - \bar{F}_t(\bar{\mathbf{w}}_t^*) \leq \frac{L}{2} C^2. \quad (73)$$

Finally, substituting (73) into (72) yields $\|\bar{\mathbf{w}}_t^*\| \leq C \sqrt{L/\mu} = C \sqrt{\kappa}$, proving (30).

• *Bound on the fixed point $\tilde{\mathbf{w}}_t$:*

– (DGD: $\mathbf{Z} = \mathbf{I}$). From the definition of $\phi_t(\cdot)$ in (22) and the fixed-point relation (28), for DGD it holds that

$$(\mathbf{I} - \overline{\mathbf{M}})\tilde{\mathbf{w}}_t + \eta \nabla \bar{f}_t(\tilde{\mathbf{w}}_t) = \mathbf{0}. \quad (74)$$

Utilizing the μ -strong convexity of $\bar{f}_t(\cdot)$ (Assumption 1), (26) can be specialized as

$$\bar{f}_t(\mathbf{0}) \geq \bar{f}_t(\tilde{\mathbf{w}}_t) - \nabla \bar{f}_t(\tilde{\mathbf{w}}_t)^\top \tilde{\mathbf{w}}_t + \frac{\mu}{2} \|\tilde{\mathbf{w}}_t\|^2$$

Next, using (74) to substitute $\nabla \bar{f}_t(\tilde{\mathbf{w}}_t) = -(1/\eta)(\mathbf{I} - \overline{\mathbf{M}})\tilde{\mathbf{w}}_t$ further gives

$$\bar{f}_t(\mathbf{0}) \geq \bar{f}_t(\tilde{\mathbf{w}}_t) + \frac{1}{\eta} \tilde{\mathbf{w}}_t^\top (\mathbf{I} - \overline{\mathbf{M}})\tilde{\mathbf{w}}_t + \frac{\mu}{2} \|\tilde{\mathbf{w}}_t\|^2. \quad (75)$$

Since $\mathbf{I} - \overline{\mathbf{M}} \succeq 0$, the middle term is nonnegative. Hence,

$$\|\tilde{\mathbf{w}}_t\|^2 \leq \frac{2}{\mu} (\bar{f}_t(\mathbf{0}) - \bar{f}_t(\tilde{\mathbf{w}}_t)). \quad (76)$$

Next, using the definition of $\bar{f}_t(\cdot)$ and $\tilde{\mathbf{w}}_t = [\tilde{\mathbf{w}}_t]_1^\top \cdots [\tilde{\mathbf{w}}_t]_N^\top$, we have

$$\begin{aligned} \bar{f}_t(\mathbf{0}) - \bar{f}_t(\tilde{\mathbf{w}}_t) &= \sum_{i=1}^t a_i(t) \sum_{n=1}^N (\ell_{n,i}(\mathbf{0}) - \ell_{n,i}([\tilde{\mathbf{w}}_t]_n)) \\ &\stackrel{(a)}{\leq} \sum_{i=1}^t a_i(t) \sum_{n=1}^N (\ell_{n,i}(\mathbf{0}) - \ell_{n,i}(\mathbf{w}_{n,i}^*)) \\ &\stackrel{(b)}{\leq} \sum_{i=1}^t a_i(t) \sum_{n=1}^N \frac{L}{2} \|\mathbf{w}_{n,i}^*\|^2 \stackrel{(c)}{\leq} \frac{NL}{2} C_Z^2, \end{aligned} \quad (77)$$

where (a) follows since $\mathbf{w}_{n,i}^*$ minimizes $\ell_{n,i}(\cdot)$, (b) follows from L -smoothness (Assumption 1), and (c) invokes Assumption 2 along with the fact that $\sum_{i=1}^t a_i(t) = 1$. Substituting (77) into (76) gives $\|\tilde{\mathbf{w}}_t\| \leq C \sqrt{\frac{NL}{\mu}} = \sqrt{N} C_Z$, which is (31) with $C_Z = C\sqrt{\kappa}$ and holds for all $t \geq 1$.

– (Diffusion: $\mathbf{Z} = \mathbf{M}$). From (22) and (28), the fixed point satisfies $\tilde{\mathbf{w}}_t = \phi_t(\tilde{\mathbf{w}}_t)$ with $\phi_t(\mathbf{w}) = \overline{\mathbf{M}}\mathbf{w} - \eta \overline{\mathbf{M}} \nabla \bar{f}_t(\mathbf{w})$ for diffusion. Hence,

$$\begin{aligned} \|\tilde{\mathbf{w}}_t\| &= \|\phi_t(\tilde{\mathbf{w}}_t)\| \\ &\stackrel{(a)}{\leq} \|\phi_t(\tilde{\mathbf{w}}_t) - \phi_t(\mathbf{0})\| + \|\phi_t(\mathbf{0})\| \stackrel{(b)}{\leq} \rho \|\tilde{\mathbf{w}}_t\| + \|\phi_t(\mathbf{0})\|, \end{aligned}$$

where (a) follows from adding and subtracting $\phi_t(\mathbf{0})$ and using the triangle inequality, and (b) invokes Lemma 1, whereby $\phi_t(\cdot)$ is a contraction with factor $\rho \in (0, 1)$. Therefore,

$$\|\tilde{\mathbf{w}}_t\| \leq \frac{1}{1-\rho} \|\phi_t(\mathbf{0})\|. \quad (78)$$

Next, $\phi_t(\mathbf{0}) = -\eta \overline{\mathbf{M}} \nabla \bar{f}_t(\mathbf{0})$, and since $\|\overline{\mathbf{M}}\|_2 = 1$ (for symmetric doubly stochastic $\overline{\mathbf{M}}$), we have

$$\|\phi_t(\mathbf{0})\| \leq \eta \|\nabla \bar{f}_t(\mathbf{0})\|.$$

Moreover, from (13), the n -th block of $\nabla \bar{f}_t(\mathbf{0})$ equals $\sum_{i=1}^t a_i(t) \nabla \ell_{n,i}(\mathbf{0})$. By L -smoothness (Assumption 1) and the optimality condition $\nabla \ell_{n,i}(\mathbf{w}_{n,i}^*) = \mathbf{0}$,

$$\|\nabla \ell_{n,i}(\mathbf{0})\| = \|\nabla \ell_{n,i}(\mathbf{0}) - \nabla \ell_{n,i}(\mathbf{w}_{n,i}^*)\| \leq L \|\mathbf{0} - \mathbf{w}_{n,i}^*\| \leq LC,$$

where the last inequality utilizes Assumption 2. Therefore,

$$\|\nabla \bar{f}_t(\mathbf{0})\|^2 = \sum_{n=1}^N \left\| \sum_{i=1}^t a_i(t) \nabla \ell_{n,i}(\mathbf{0}) \right\|^2 \leq \sum_{n=1}^N (LC)^2 = NL^2 C^2,$$

where the inequality follows from Jensen's inequality and $\sum_{i=1}^t a_i(t) = 1$, yielding $\|\nabla \bar{f}_t(\mathbf{0})\| \leq \sqrt{N} LC$. Substituting into (78) gives $\|\tilde{\mathbf{w}}_t\| \leq \frac{\eta}{1-\rho} \sqrt{N} LC$. Finally, under the step size range $0 < \eta \leq 2/(L + \mu)$, we have $\rho = 1 - \eta\mu$ (cf. Lemma 1), and therefore, for all $t \geq 1$, $\|\tilde{\mathbf{w}}_t\| \leq C\sqrt{N} \kappa$, which is (31) for $\mathbf{Z} = \mathbf{M}$.

– (Sharper diffusion bound when $\mathbf{M} \succeq 0$). If, in addition, $\mathbf{M} \succeq 0$, the diffusion fixed-point satisfies

$$(\mathbf{I} - \overline{\mathbf{M}})\tilde{\mathbf{w}}_t + \eta \overline{\mathbf{M}} \nabla \bar{f}_t(\tilde{\mathbf{w}}_t) = \mathbf{0}. \quad (79)$$

Define the orthogonal projectors $P_{\mathcal{R}(\overline{\mathbf{M}})} \triangleq \overline{\mathbf{M}} \overline{\mathbf{M}}^\dagger$ and $P_{\mathcal{N}(\overline{\mathbf{M}})} \triangleq \mathbf{I} - \overline{\mathbf{M}} \overline{\mathbf{M}}^\dagger$, where $\overline{\mathbf{M}}^\dagger$ is the Moore–Penrose pseudoinverse of $\overline{\mathbf{M}}$. Pre-multiplying (79) by $P_{\mathcal{N}(\overline{\mathbf{M}})}$ yields

$$P_{\mathcal{N}(\overline{\mathbf{M}})} \tilde{\mathbf{w}}_t = \mathbf{0},$$

where we leverage the facts that $P_{\mathcal{N}(\overline{\mathbf{M}})} \overline{\mathbf{M}} = (\mathbf{I} - \overline{\mathbf{M}} \overline{\mathbf{M}}^\dagger) \overline{\mathbf{M}} = \overline{\mathbf{M}} - \overline{\mathbf{M}} \overline{\mathbf{M}}^\dagger \overline{\mathbf{M}} = \mathbf{0}$ and $P_{\mathcal{N}(\overline{\mathbf{M}})} (\mathbf{I} - \overline{\mathbf{M}}) = P_{\mathcal{N}(\overline{\mathbf{M}})} - P_{\mathcal{N}(\overline{\mathbf{M}})} \overline{\mathbf{M}} = P_{\mathcal{N}(\overline{\mathbf{M}})}$. Therefore, $\tilde{\mathbf{w}}_t \in \mathcal{R}(\overline{\mathbf{M}})$. Hence, there exists $\mathbf{y}_t \in \mathbb{R}^{Nd}$ such that

$$\tilde{\mathbf{w}}_t = \overline{\mathbf{M}} \mathbf{y}_t. \quad (80)$$

Next, pre-multiply (79) by $\mathbf{y}_t^\top$ to obtain

$$\mathbf{y}_t^\top (\mathbf{I} - \overline{\mathbf{M}}) \tilde{\mathbf{w}}_t + \eta \mathbf{y}_t^\top \overline{\mathbf{M}} \nabla \bar{f}_t(\tilde{\mathbf{w}}_t) = 0.$$

Since $\overline{\mathbf{M}}$ is symmetric, $\mathbf{y}_t^\top \overline{\mathbf{M}} = (\overline{\mathbf{M}} \mathbf{y}_t)^\top = \tilde{\mathbf{w}}_t^\top$, we further obtain

$$\nabla \bar{f}_t(\tilde{\mathbf{w}}_t)^\top \tilde{\mathbf{w}}_t = -\frac{1}{\eta} \mathbf{y}_t^\top (\mathbf{I} - \overline{\mathbf{M}}) \tilde{\mathbf{w}}_t = -\frac{1}{\eta} \mathbf{y}_t^\top (\mathbf{I} - \overline{\mathbf{M}}) \overline{\mathbf{M}} \mathbf{y}_t, \quad (81)$$

where the last equality uses (80). Since $\mathbf{M} \succeq 0$, $\overline{\mathbf{M}}$ is also positive semidefinite and its eigenvalues are those of $\mathbf{M}$. Hence, since $\lambda(\overline{\mathbf{M}}) \subset [0, 1]$, the matrix $(\mathbf{I} - \overline{\mathbf{M}}) \overline{\mathbf{M}}$ is positive semidefinite (its eigenvalues are $\lambda(1 - \lambda) \geq 0$), and therefore,

$$-\nabla \bar{f}_t(\tilde{\mathbf{w}}_t)^\top \tilde{\mathbf{w}}_t = \frac{1}{\eta} \mathbf{y}_t^\top (\mathbf{I} - \overline{\mathbf{M}}) \overline{\mathbf{M}} \mathbf{y}_t \geq 0. \quad (82)$$

Next, utilizing the μ -strong convexity of $\bar{f}_t(\cdot)$ (Assumption 1), (26) can be specialized as

$$\bar{f}_t(\mathbf{0}) \geq \bar{f}_t(\tilde{\mathbf{w}}_t) - \nabla \bar{f}_t(\tilde{\mathbf{w}}_t)^\top \tilde{\mathbf{w}}_t + \frac{\mu}{2} \|\tilde{\mathbf{w}}_t\|^2 \geq \bar{f}_t(\tilde{\mathbf{w}}_t) + \frac{\mu}{2} \|\tilde{\mathbf{w}}_t\|^2,$$

where the last inequality uses (82). Rearranging yields

$$\|\tilde{\mathbf{w}}_t\|^2 \leq \frac{2}{\mu} (\bar{f}_t(\mathbf{0}) - \bar{f}_t(\tilde{\mathbf{w}}_t)),$$

which is identical to (76). Thus, following the same steps as in (76)–(77) yields the same bound as that obtained for DGD: $\|\tilde{\mathbf{w}}_t\| \leq C \sqrt{\frac{NL}{\mu}} = \sqrt{N} C_Z$, which is (31) with $C_Z = C\sqrt{\kappa}$.

• *Gradient Bounds $\|\nabla \bar{f}_t(\tilde{\mathbf{w}}_i)\|$ and $\|\nabla f_t(\tilde{\mathbf{w}}_i)\|$:* We begin by recalling the μ -strong convexity of $\bar{f}_t(\cdot)$ under Assumption 1. Let $\tilde{\mathbf{m}}_t^* = \arg \min_{\mathbf{w} \in \mathbb{R}^{Nd}} \bar{f}_t(\mathbf{w})$ denote the minimizer

of $\bar{f}_t(\cdot)$. Next, since $\bar{f}_t(\mathbf{w}) = \sum_{n=1}^N \sum_{i=1}^t a_i(t) \ell_{n,i}([\mathbf{w}]_n)$ is block-separable, each d -dimensional block $[\bar{\mathbf{m}}^*]_n$ minimizes the function $\sum_{i=1}^t a_i(t) \ell_{n,i}(\mathbf{w})$ with $\mathbf{w} \in \mathbb{R}^d$. As a result, by using the same argument used to bound $\bar{\mathbf{w}}_t^*$ in (72)–(73), it is easy to verify that $\|[\bar{\mathbf{m}}^*]_n\| \leq C\sqrt{\kappa}$, and hence

$$\|\bar{\mathbf{m}}_t^*\| \leq \sqrt{N} C_Z. \quad (83)$$

Next, L -smoothness (Assumption 1) of $\bar{f}_t(\cdot)$ and the optimality condition $\nabla \bar{f}_t(\bar{\mathbf{m}}_t^*) = \mathbf{0}$ yields, for any $i, t \geq 1$,

$$\begin{aligned} \|\nabla \bar{f}_t(\tilde{\mathbf{w}}_i)\| &= \|\nabla \bar{f}_t(\tilde{\mathbf{w}}_i) - \nabla \bar{f}_t(\bar{\mathbf{m}}_t^*)\| \\ &\leq L\|\tilde{\mathbf{w}}_i - \bar{\mathbf{m}}_t^*\| \stackrel{(a)}{\leq} L(\|\tilde{\mathbf{w}}_i\| + \|\bar{\mathbf{m}}_t^*\|) \stackrel{(b)}{\leq} 2L\sqrt{N} C_Z, \end{aligned}$$

where (a) invokes the triangle inequality and (b) uses (83) and $\|\tilde{\mathbf{w}}_t\| \leq \sqrt{N} C_Z$. This proves the first inequality in (32) with $G_Z = 2L\sqrt{N} C_Z$.

Let $\mathbf{m}_t^* = \arg \min_{\mathbf{w}} f_t(\mathbf{w})$. Since $f_t(\mathbf{w}) = \sum_{n=1}^N \ell_{n,t}([\mathbf{w}]_n)$ is block-wise separable, it follows that $\mathbf{m}_t^* = [(\mathbf{w}_{1,t}^* \cdots \mathbf{w}_{N,t}^*)^\top]^\top$, hence $\|\mathbf{m}_t^*\| \leq \sqrt{N} C \leq \sqrt{N} C_Z$. Using L -smoothness (Assumption 1) of $f_t(\cdot)$ and $\nabla f_t(\mathbf{m}_t^*) = \mathbf{0}$ yields

$$\|\nabla f_t(\tilde{\mathbf{w}}_i)\| \leq L\|\tilde{\mathbf{w}}_i - \mathbf{m}_t^*\| \leq L(\|\tilde{\mathbf{w}}_i\| + \|\mathbf{m}_t^*\|) \leq 2L\sqrt{N} C_Z,$$

for any $i, t \geq 1$, which proves the second inequality in (32). Finally, since $\bar{\mathbf{w}}_t^* = \mathbf{1}_N \otimes \bar{\mathbf{w}}_t^*$, (30) implies $\|\bar{\mathbf{w}}_t^*\| = \|\mathbf{1}_N \otimes \bar{\mathbf{w}}_t^*\| = \sqrt{N} \|\bar{\mathbf{w}}_t^*\| \leq C\sqrt{N\kappa} \leq \sqrt{N} C_Z$. Using again the L -smoothness of $\bar{f}_t(\cdot)$ and the optimality condition $\nabla \bar{f}_t(\bar{\mathbf{m}}_t^*) = \mathbf{0}$, for all $t \geq 1$, we have

$$\begin{aligned} \|\nabla \bar{f}_t(\bar{\mathbf{w}}_t^*)\| &= \|\nabla \bar{f}_t(\bar{\mathbf{w}}_t^*) - \nabla \bar{f}_t(\bar{\mathbf{m}}_t^*)\| \\ &\leq L\|\bar{\mathbf{w}}_t^* - \bar{\mathbf{m}}_t^*\| \leq L(\|\bar{\mathbf{w}}_t^*\| + \|\bar{\mathbf{m}}_t^*\|) \leq 2L\sqrt{N} C_Z. \end{aligned}$$

Utilizing the definition of G_Z yields the desired bound. $\square$

Proof of Theorem 1. We first show that for all $t \geq t_0$ it holds that

$$S(t) \leq \frac{A}{t}. \quad (84)$$

Note that we can express $S(t+1)$ as

$$S(t+1) = \sum_{i=1}^t \frac{\alpha^{t+1-i}}{i+1} = \frac{\alpha}{t+1} + \alpha S(t). \quad (85)$$

By the definition of A , we have $A \geq t_0 S(t_0)$, hence it directly holds that $S(t_0) \leq \frac{A}{t_0}$. Next, as induction hypothesis assume that $S(t) \leq \frac{A}{t}$ for some $t \geq t_0 = \lceil \frac{2\alpha}{1-\alpha} \rceil$. Then, from (85), $S(t+1)$ can be upper bounded as $S(t+1) \leq \frac{\alpha}{t+1} + \alpha \frac{A}{t}$. To complete the argument, it is sufficient to show that the right-hand side above is at most $\frac{A}{t+1}$, which is equivalent to showing that

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

Since $t \geq t_0 \geq 2\alpha/(1-\alpha)$, it follows that $1 - \alpha - \alpha/t \geq (1-\alpha)/2$. Thus, the desired inequality holds whenever $A \geq 2\alpha/(1-\alpha)$, which is true by definition of A . Therefore, it follows by induction that (84) holds for all $t \geq t_0$. Leveraging this bound in (46) yields the desired TE bound (47). Finally, taking $\limsup_{t \rightarrow \infty}$ in (47) gives the stated ATE bound, completing the proof. $\square$

Proof of Theorem 2. We first show that, for all $t \geq t_0$,

$$S_\gamma(t) \leq A_\gamma \frac{1-\gamma}{1-\gamma^t}. \quad (86)$$

Note that $S_\gamma(t+1) = \sum_{i=1}^t \frac{(1-\gamma)\alpha^{t+1-i}}{1-\gamma^{t+1}}$ can be written recursively as

$$S_\gamma(t+1) = \frac{(1-\gamma)\alpha}{1-\gamma^{t+1}} + \alpha S_\gamma(t). \quad (87)$$

By definition of A_γ , we have $A_\gamma \geq \frac{(1-\gamma^t)S_\gamma(t_0)}{1-\gamma}$, thus it holds that $S_\gamma(t_0) \leq \frac{A_\gamma(1-\gamma)}{1-\gamma^{t_0}}$. Next, we assume as induction hypothesis that $S_\gamma(t) \leq \frac{A_\gamma(1-\gamma)}{1-\gamma^t}$ for some $t \geq t_0 = \lceil \ln(\frac{1-\alpha}{1+\alpha-2\gamma\alpha}) / \ln(\gamma) \rceil$. Then, from (87), $S_\gamma(t+1)$ can be upper bounded as $S_\gamma(t+1) \leq \frac{(1-\gamma)\alpha}{1-\gamma^{t+1}} + \alpha \frac{A_\gamma(1-\gamma)}{1-\gamma^t}$. To complete the induction, it suffices to show that the right-hand side above is at most $\frac{A_\gamma(1-\gamma)}{1-\gamma^{t+1}}$. After rearranging and simplifying, this condition is equivalent to

$$A_\gamma \left[1 - \alpha - \alpha\gamma^t \frac{1-\gamma}{1-\gamma^t} \right] \geq \alpha. \quad (88)$$

Since $t \geq t_0$ and $\gamma \in (0, 1)$, the definition of t_0 implies

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

Therefore, we can lower bound the left-hand side of (88) as

$$\begin{aligned} &A_\gamma \left[1 - \alpha - \alpha\gamma^t \frac{1-\gamma}{1-\gamma^t} \right] \\ &\geq A_\gamma \left[1 - \alpha - \alpha \frac{1-\alpha}{1+\alpha-2\gamma\alpha} \frac{1-\gamma}{1-\frac{1-\alpha}{1+\alpha-2\gamma\alpha}} \right] = A_\gamma \frac{1-\alpha}{2}, \end{aligned}$$

where the inequality is due to the bound on γ^t . Finally, by the definition of A_γ , we have $A_\gamma(\frac{1-\alpha}{2}) \geq \alpha$, and thus (88) holds. We have thus proved that $S_\gamma(t+1) \leq \frac{A_\gamma(1-\gamma)}{1-\gamma^{t+1}}$. Hence, by induction, (86) holds for all $t \geq t_0$. Substituting (86) into (50) yields (51).

Next, from (87), the sequence $\{S_\gamma(t)\}$ satisfies

$$S_\gamma(t+1) = \alpha S_\gamma(t) + b_t,$$

with $b_t \triangleq \frac{(1-\gamma)\alpha}{1-\gamma^{t+1}}$. Since $b_t \rightarrow (1-\gamma)\alpha$ and $\alpha \in (0, 1)$, Lemma 6 immediately yields

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

Applying $\limsup_{t \rightarrow \infty}$ to both sides of (50), and using the limit of $S_\gamma(t)$ gives (52), which completes the proof. $\square$

REFERENCES

- [1] M. F. U. Abrar, N. Michelusi, and E. G. Larsson, "Decentralized time-varying optimization for streaming data via temporal weighting," 2026, to appear at the 2026 IEEE Asilomar Conference on Signals, Systems, and Computers. [Online]. Available: <https://arxiv.org/abs/2605.06971>
- [2] T. Yang, X. Yi, J. Wu, Y. Yuan, D. Wu, Z. Meng, Y. Hong, H. Wang, Z. Lin, and K. H. Johansson, "A survey of distributed optimization," *Annual Reviews in Control*, vol. 47, pp. 278–305, 2019.
- [3] 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.
- [4] 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.

- [5] S. C. Hoi, D. Sahoo, J. Lu, and P. Zhao, "Online learning: A comprehensive survey," *Neurocomput.*, vol. 459, no. C, p. 249–289, Oct. 2021.
- [6] L. Wang, X. Zhang, H. Su, and J. Zhu, "A comprehensive survey of continual learning: Theory, method and application," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5362–5383, 2024.
- [7] L. Yuan, Z. Wang, L. Sun, P. S. Yu, and C. G. Brinton, "Decentralized federated learning: A survey and perspective," *IEEE Internet of Things Journal*, vol. 11, no. 21, pp. 34617–34638, 2024.
- [8] M. Chen, D. Gündüz, K. Huang, W. Saad, M. Bennis, A. V. Feljan, and H. V. Poor, "Distributed learning in wireless networks: Recent progress and future challenges," *IEEE Journal on Selected Areas in Comms.*, vol. 39, no. 12, pp. 3579–3605, 2021.
- [9] N. Michelusi, "Non-coherent over-the-air decentralized gradient descent," *IEEE Trans. on Signal Processing*, vol. 72, pp. 4618–4634, 2024.
- [10] M. F. Ul Abrar and N. Michelusi, "Biased federated learning under wireless heterogeneity," *IEEE Trans. on Wireless Comms.*, vol. 25, pp. 16449–16462, 2026.
- [11] A. H. Sayed, "Adaptation, learning, and optimization over networks," *Found. Trends Mach. Learn.*, vol. 7, no. 4–5, p. 311–801, Jul. 2014.
- [12] L. Bai, C. Sun, Z. Feng, and G. Hu, "Distributed continuous-time resource allocation with time-varying resources under quadratic cost functions," in *2018 IEEE Conference on Decision and Control (CDC)*, 2018, pp. 823–828.
- [13] Q. Ling and A. Ribeiro, "Decentralized dynamic optimization through the alternating direction method of multipliers," *IEEE Trans. on Signal Processing*, vol. 62, no. 5, pp. 1185–1197, 2014.
- [14] C. Xi and U. A. Khan, "Distributed dynamic optimization over directed graphs," in *2016 IEEE 55th Conference on Decision and Control (CDC)*, 2016, pp. 245–250.
- [15] A. Simonetto, A. Koppel, A. Mokhtari, G. Leus, and A. Ribeiro, "Decentralized prediction-correction methods for networked time-varying convex optimization," *IEEE Trans. on Automatic Control*, vol. 62, no. 11, pp. 5724–5738, 2017.
- [16] C. Sun, M. Ye, and G. Hu, "Distributed time-varying quadratic optimization for multiple agents under undirected graphs," *IEEE Trans. on Automatic Control*, vol. 62, no. 7, pp. 3687–3694, 2017.
- [17] S. Rahili and W. Ren, "Distributed continuous-time convex optimization with time-varying cost functions," *IEEE Trans. on Automatic Control*, vol. 62, no. 4, pp. 1590–1605, 2017.
- [18] B. Wang, S. Sun, and W. Ren, "Distributed continuous-time algorithms for optimal resource allocation with time-varying quadratic cost functions," *IEEE Trans. on Control of Network Systems*, vol. 7, no. 4, pp. 1974–1984, 2020.
- [19] B. Huang, Y. Zou, Z. Meng, and W. Ren, "Distributed time-varying convex optimization for a class of nonlinear multiagent systems," *IEEE Trans. on Automatic Control*, vol. 65, no. 2, pp. 801–808, 2020.
- [20] S. Sun, J. Xu, and W. Ren, "Distributed continuous-time algorithms for time-varying constrained convex optimization," *IEEE Trans. on Automatic Control*, vol. 68, no. 7, pp. 3931–3946, 2023.
- [21] Y. Li, H. Wang, W. Xu, T. Xiao, H. Liu, M. Tu, Y. Wang, X. Yang, R. Zhang, S. Yu, S. Guo, and R. Li, "Unleashing the power of continual learning on non-centralized devices: A survey," *IEEE Comms. Surveys & Tutorials*, vol. 28, pp. 1059–1098, 2026.
- [22] B. Polyak, *Introduction to optimization*. Optimization Software, 1987.
- [23] 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*, vol. 2, 1998, pp. 910–913 vol.2.
- [24] A. Y. Popkov, "Gradient methods for nonstationary unconstrained optimization problems," *Autom. Remote Control*, vol. 66, no. 6, p. 883–891, Jun. 2005.
- [25] 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.
- [26] M. Fazlyab, S. Paternain, V. M. Preciado, and A. Ribeiro, "Prediction-correction interior-point method for time-varying convex optimization," *IEEE Trans. on Automatic Control*, vol. 63, no. 7, pp. 1973–1986, 2018.
- [27] A. Simonetto and E. Dall'Anese, "Prediction-correction algorithms for time-varying constrained optimization," *IEEE Trans. on Signal Processing*, vol. 65, no. 20, pp. 5481–5494, 2017.
- [28] A. S. Charles, A. Balavoine, and C. J. Rozell, "Dynamic filtering of time-varying sparse signals via ℓ_1 minimization," *IEEE Trans. on Signal Processing*, vol. 64, no. 21, pp. 5644–5656, 2016.
- [29] A. Simonetto and P. Massioni, "Nonlinear optimization filters for stochastic time-varying convex optimization," *International Journal of Robust and Nonlinear Control*, vol. 34, no. 12, pp. 8065–8089, 2024.
- [30] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.
- [31] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, "On the linear convergence of the admm in decentralized consensus optimization," *Trans. Sig. Proc.*, vol. 62, no. 7, p. 1750–1761, Apr. 2014.
- [32] K. Yuan, W. Xu, and Q. Ling, "Can primal methods outperform primal-dual methods in decentralized dynamic optimization?" *IEEE Trans. on Signal Processing*, vol. 68, pp. 4466–4480, 2020.
- [33] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in *Proceedings of the Twentieth International Conference on Machine Learning*, ser. ICML'03. AAAI Press, 2003, p. 928–935.
- [34] S. Shalev-Shwartz, "Online learning and online convex optimization," *Found. Trends Mach. Learn.*, vol. 4, pp. 107–194, 2012.
- [35] E. Hazan, "Introduction to online convex optimization," *Found. Trends Optim.*, vol. 2, pp. 157–325, 2016.
- [36] A. Jadbabaie, A. Rakhlin, S. Shahrampour, and K. Sridharan, "Online Optimization: Competing with Dynamic Comparators," in *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, ser. Proceedings of Machine Learning Research, vol. 38, 2015, pp. 398–406.
- [37] O. Besbes, Y. Gur, and A. Zeevi, "Non-stationary stochastic optimization," *Oper. Res.*, vol. 63, no. 5, p. 1227–1244, Oct. 2015.
- [38] S. Shahrampour and A. Jadbabaie, "Distributed online optimization in dynamic environments using mirror descent," *IEEE Trans. on Automatic Control*, vol. 63, no. 3, pp. 714–725, 2018.
- [39] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 2935–2947, 2016.
- [40] 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.
- [41] 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 Trans. on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3366–3385, 2022.
- [42] 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.
- [43] B. Wu, Z. Ding, and J. Huang, "A review of continual learning in edge ai," *IEEE Trans. on Network Science and Engineering*, vol. 13, pp. 6571–6588, 2026.
- [44] A. Mitra, H. Hassani, and G. J. Pappas, "Online federated learning," in *2021 60th IEEE Conference on Decision and Control (CDC)*, 2021, pp. 4083–4090.
- [45] O. Marfoq, G. Neglia, L. Kameni, and R. Vidal, "Federated learning for data streams," in *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, F. Ruiz, J. Dy, and J.-W. van de Meent, Eds., vol. 206. PMLR, 25–27 Apr 2023, pp. 8889–8924.
- [46] C.-H. Hu, Z. Chen, and E. G. Larsson, "Energy-efficient federated edge learning with streaming data: A Lyapunov optimization approach," *IEEE Trans. on Comms.*, vol. 73, no. 2, pp. 1142–1156, 2025.
- [47] M. F. Ul Abrar, N. Michelusi, and E. G. Larsson, "Time-varying optimization for streaming data via temporal weighting," in *2025 59th Asilomar Conference on Signals, Systems, and Computers*, 2025, pp. 1343–1349.
- [48] 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.
- [49] Y. Nesterov, *Lectures on Convex Optimization*, 2nd ed. Springer Publishing Company, Incorporated, 2018.
- [50] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge university press, 2004.
- [51] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed. New York, NY: McGraw-Hill, 1976, vol. 3.