

IMPROVING ONLINE-TO-NONCONVEX CONVERSION FOR SMOOTH OPTIMIZATION VIA DOUBLE OPTIMISM

Francisco Patitucci^{1*} Ruichen Jiang^{1*} Aryan Mokhtari^{1,2}

¹The University of Texas at Austin ²Google Research

{fpatitucci, rjiang}@utexas.edu, mokhtari@austin.utexas.edu

ABSTRACT

A recent breakthrough in nonconvex optimization is the online-to-nonconvex conversion framework of Cutkosky et al. (2023), which reformulates the task of finding an ε -first-order stationary point as an online learning problem. When both the gradient and the Hessian are Lipschitz continuous, instantiating this framework with two different online learners achieves a complexity of $\mathcal{O}(\varepsilon^{-1.75} \log(1/\varepsilon))$ in the deterministic case and a complexity of $\mathcal{O}(\varepsilon^{-3.5})$ in the stochastic case. However, this approach suffers from several limitations: (i) the deterministic method relies on a complex double-loop scheme that solves a fixed-point equation to construct hint vectors for an optimistic online learner, introducing an extra logarithmic factor; (ii) the stochastic method assumes a bounded second-order moment of the stochastic gradient, which is stronger than standard variance bounds; and (iii) different online learning algorithms are used in the two settings.

In this paper, we address these issues by introducing an online optimistic gradient method based on a novel *doubly optimistic hint function*. Specifically, we use the gradient at an extrapolated point as the hint, motivated by two optimistic assumptions: that the difference between the hint and the target gradient remains near constant, and that consecutive update directions change slowly due to smoothness. Our method eliminates the need for a double loop and removes the logarithmic factor. Furthermore, by simply replacing full gradients with stochastic gradients and under the standard assumption that their variance is bounded by σ^2 , we obtain a unified algorithm with complexity $\mathcal{O}(\varepsilon^{-1.75} + \sigma^2 \varepsilon^{-3.5})$, smoothly interpolating between the best-known deterministic rate and the optimal stochastic rate.

1 INTRODUCTION

In this paper, we consider an unconstrained minimization problem

$$\min_{\mathbf{x} \in \mathbb{R}^d} F(\mathbf{x}), \quad (1)$$

where $F : \mathbb{R}^d \rightarrow \mathbb{R}$ is a smooth but possibly non-convex function bounded from below. With access to a deterministic gradient oracle, it is well known that gradient descent (GD) can find an ε -stationary point within $\mathcal{O}(\varepsilon^{-2})$ iterations when the gradient of F is Lipschitz continuous, which is worst-case optimal among all first-order methods (Carmon et al., 2020). In the stochastic setting, where the oracle returns an unbiased noisy gradient with variance bounded by σ^2 , stochastic gradient descent (SGD) achieves a complexity of $\mathcal{O}(\varepsilon^{-2} + \sigma^2 \varepsilon^{-4})$ (Ghadimi & Lan, 2013), which is likewise worst-case optimal (Arjevani et al., 2023).

Interestingly, when both the gradient and the Hessian of F are Lipschitz continuous, it becomes possible to outperform standard GD or SGD. In the deterministic setting, Carmon et al. (2017) proposed an accelerated gradient-based method that exploits directions of negative curvature, achieving a complexity of $\mathcal{O}(\varepsilon^{-1.75} \log(1/\varepsilon))$. Concurrently, Agarwal et al. (2017) introduced an approximate variant of the cubic regularization method (Nesterov & Polyak, 2006), which uses only first-order gradients and access to Hessian-vector products, and achieves a complexity of $\mathcal{O}(\varepsilon^{-1.75} \log(d/\varepsilon))$.

*Equal contribution

More recently, Li & Lin (2022; 2023) eliminated the logarithmic factor from these bounds by incorporating restarting techniques into accelerated gradient and heavy-ball methods, achieving the state-of-the-art complexity of $\mathcal{O}(\varepsilon^{-1.75})$. However, the methods proposed in (Carmon et al., 2017; Agarwal et al., 2017) are complex and rely on multiple subroutines to exploit negative curvature, while the restarting techniques in (Li & Lin, 2022; 2023) involve lengthy derivations, and their extension to the stochastic setting remains unclear. In the stochastic setting, similar improvements over SGD have been achieved using different algorithmic techniques (Allen-Zhu, 2018a;b; Tripuraneni et al., 2018; Fang et al., 2019). To our knowledge, the best known result is due to Cutkosky & Mehta (2020), who proposed normalized SGD with momentum and extrapolation, achieving $\mathcal{O}(\varepsilon^{-2} + \sigma^2 \varepsilon^{-3.5})$. While this method attains the optimal rate of $\mathcal{O}(\sigma^2 \varepsilon^{-3.5})$ for the stochastic setting (Arjevani et al., 2020), it fails to achieve the accelerated rate of $\mathcal{O}(\varepsilon^{-1.75})$ for the deterministic setting. Notably, to the best of our knowledge, no single unified algorithm simultaneously attains the best-known guarantees in both settings.

To address this gap, a promising step toward unification is the *online-to-nonconvex* (O2NC) conversion proposed in (Cutkosky et al., 2023), which offers a general and powerful framework for nonconvex optimization across various settings. The core idea is to reduce the problem of finding a stationary point of a nonconvex function to an online learning problem over the update directions. This reduction enables converting an online learning algorithm into a nonconvex optimization method, which greatly simplifies convergence analysis. In the smooth setting considered in this paper, by instantiating two different online learning algorithms, Cutkosky et al. (2023) establish a complexity of $\mathcal{O}(\varepsilon^{-1.75} \log(1/\varepsilon))$ in the deterministic case and $\mathcal{O}(G^2 \varepsilon^{-3.5})$ in the stochastic case, where G^2 is an upper bound on the second-order moment of the stochastic gradient. Despite its conceptual elegance, the O2NC framework has several limitations. First, in the deterministic setting, the algorithm relies on a fixed-point iteration as a subroutine, introducing an extra logarithmic factor in the convergence rate and adding complexity to the algorithmic design. Second, in the stochastic setting, it requires a stronger assumption that the second-order moment of the stochastic gradient is uniformly bounded, which can be restrictive in practice. Third, the framework requires different online learning algorithms for the stochastic and deterministic settings, limiting its modularity and complicating potential extensions.

Our contribution. Building on the O2NC framework, we propose a simple yet effective algorithm that replaces the fixed-point iteration subroutine in (Cutkosky et al., 2023) with a single stochastic gradient call at an extrapolated point, achieving a complexity of $\mathcal{O}(\varepsilon^{-1.75} + \sigma^2 \varepsilon^{-3.5})$. This eliminates the additional logarithmic factor and smoothly interpolates between the best-known deterministic and stochastic rates. Our core idea is a novel *doubly optimistic* online gradient method, which aims to predict the gradient using a carefully crafted hint function. Inspired by standard optimistic methods (Chiang et al., 2012; Rakhlin & Sridharan, 2013; Joulani et al., 2020), our first optimistic assumption is that the discrepancy between the gradient and the hint evolves slowly over time. Crucially, we introduce a second optimism that the update direction itself also evolves slowly, which we exploit using the function’s smoothness to guide the design of the hint. To the best of our knowledge, this is the first algorithm to seamlessly attain the best-known guarantees in both deterministic and stochastic settings, resolving a key limitation of the original O2NC framework.

1.1 ADDITIONAL RELATED WORK

Deterministic setting Beyond the works reviewed in the introduction, Marumo & Takeda (2024a;b) built on the restarting technique from (Li & Lin, 2022; 2023) to develop parameter-free algorithms using line search, thereby removing the need for prior knowledge of problem-specific constants. Complementing these upper bounds, Carmon et al. (2021) established a lower bound of $\Omega(\varepsilon^{-\frac{12}{7}})$ for functions with Lipschitz gradient and Hessian. This leaves a gap of $\mathcal{O}(\varepsilon^{-\frac{1}{28}})$ between the best-known upper and lower bounds in this setting. Several works also aimed to find second-order stationary points (Agarwal et al., 2017; Carmon et al., 2018; Jin et al., 2018; Allen-Zhu & Li, 2018; Xu et al., 2017; Royer & Wright, 2018; Royer et al., 2020) using gradient or Hessian-vector product oracles, but their complexity remains no better than $\mathcal{O}(\varepsilon^{-1.75})$.

Stochastic setting Several works have shown improved guarantees by either strengthening the oracle or imposing additional restrictions on the noise. Specifically, with a stronger oracle that has access to Hessian-vector products, Arjevani et al. (2020) proposed a variance-reduction algorithm

that combines stochastic gradients with Hessian-vector products, attaining the optimal complexity of $\mathcal{O}(\sigma^2 \varepsilon^{-3})$. Moreover, another line of work relies on the mean-squared smoothness assumption on stochastic gradients, i.e., $\mathbb{E}[\|\nabla f(\mathbf{x}; \xi) - \nabla f(\mathbf{y}; \xi)\|^2] \leq \bar{L}^2 \|\mathbf{x} - \mathbf{y}\|^2$ for all \mathbf{x} and \mathbf{y} with some constant \bar{L} ; under this condition, variance-reduction-based methods such as SPIDER (Fang et al., 2018) and SNVRG (Zhou et al., 2020) achieved a complexity of $\mathcal{O}(\sigma \varepsilon^{-3})$. Under a slightly stricter assumption that $\nabla f(\mathbf{x}; \xi)$ is Lipschitz continuous with probability one, Cutkosky & Orabona (2019) established the complexity of $\mathcal{O}(\varepsilon^{-2} + \sigma \varepsilon^{-3})$. It is also worth mentioning that the above methods require a multi-point stochastic oracle, which computes the stochastic gradients at two different points with the same random seed ξ . Finally, the lower bounds in (Arjevani et al., 2023) established that these complexity results are worst-case optimal.

Online-to-nonconvex conversion Apart from the smooth nonconvex optimization considered in this paper, the original O2NC conversion by Cutkosky et al. (2023) also addressed the complexity of nonsmooth nonconvex optimization first considered in Zhang et al. (2020), yielding an optimal stochastic algorithm. The framework has been extended and refined in several follow-up works (Zhang & Cutkosky, 2024; Ahn et al., 2024a), and it has been shown that various practical algorithms, such as SGD with momentum and Adam, can be interpreted within the O2NC framework (Zhang & Cutkosky, 2024; Ahn et al., 2024b; Ahn & Cutkosky, 2024; Ahn et al., 2024a).

Comparison with Jiang et al. (2025) A recent work by Jiang et al. (2025) also built upon the online-to-non-convex conversion framework introduced by Cutkosky et al. (2023), leveraging optimistic online gradient descent to tackle the resulting online learning problem. However, their focus is on quasi-Newton methods in the deterministic setting, whereas our work focuses on first-order methods in both stochastic and deterministic settings. While that approach shares a similar high-level structure and relies on common core lemmas as ours, they differ fundamentally in their hint constructions. Specifically, in our paper, our hint is based on the stochastic gradient at an extrapolated point, whereas Jiang et al. (2025) employed a second-order approximation, where the approximate Hessian matrix is updated via an online learning algorithm in the space of matrices. Due to the different hint constructions, our analysis exploits negative terms in the regret of optimistic online gradient descent to obtain tighter regret bounds—terms that are neglected in (Jiang et al., 2025).

2 PRELIMINARIES AND BACKGROUND

Throughout the paper, we assume that F satisfies two key assumptions described below. Unless otherwise specified, we use $\|\cdot\|$ to denote the ℓ_2 -norm for vectors and the operator norm of matrices.

Assumption 2.1. We have $\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\| \leq L_1 \|\mathbf{x} - \mathbf{y}\|$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

Assumption 2.2. We have $\|\nabla^2 F(\mathbf{x}) - \nabla^2 F(\mathbf{y})\| \leq L_2 \|\mathbf{x} - \mathbf{y}\|$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

In the stochastic setting, we assume access to an unbiased gradient oracle with bounded variance.

Assumption 2.3. We assume that we have access to a stochastic gradient oracle satisfying

$$\mathbb{E}_{\xi \sim \mathcal{D}} [\nabla f(\mathbf{x}; \xi)] = \nabla F(\mathbf{x}) \quad \text{and} \quad \mathbb{E}_{\xi \sim \mathcal{D}} \|\nabla f(\mathbf{x}; \xi) - \nabla F(\mathbf{x})\|^2 \leq \sigma^2, \quad \forall \mathbf{x} \in \mathbb{R}^d, \quad (2)$$

where \mathcal{D} is a data distribution.

2.1 ONLINE-TO-NONCONVEX CONVERSION

As discussed, the online-to-nonconvex conversion (O2NC) framework proposed in Cutkosky et al. (2023) provides a unified approach that reformulates the task of finding a stationary point of a nonconvex function F as an online convex optimization (OCO) problem. By instantiating this framework with different online learning algorithms to solve the resulting OCO problem, one can derive various algorithms for nonconvex optimization. In this section, we first recap the core ideas behind the O2NC framework.

Consider the general update rule $\mathbf{x}_n = \mathbf{x}_{n-1} + \Delta_n$, where we assume that the update direction Δ_n has a bounded norm, i.e., $\|\Delta_n\| \leq D$ for a constant $D > 0$. The starting point of Cutkosky et al. (2023) is to apply the fundamental theorem of calculus to characterize the function decrease between two consecutive points as

$$F(\mathbf{x}_n) - F(\mathbf{x}_{n-1}) = \langle \nabla_n, \mathbf{x}_n - \mathbf{x}_{n-1} \rangle = \langle \nabla_n, \Delta_n \rangle, \quad (3)$$

where in the above expression $\nabla_n = \int_0^1 \nabla F(\mathbf{x}_{n-1} + s(\mathbf{x}_n - \mathbf{x}_{n-1})) ds$ is the average gradient along the line segment between \mathbf{x}_{n-1} and \mathbf{x}_n . Given the equality in (3) and upper bound on the norm of Δ_n , the ideal choice of Δ_n to maximize the function value decrease is to set it as $\Delta_n = -D \frac{\nabla_n}{\|\nabla_n\|}$, which leads to a function decrease of $-D\|\nabla_n\|$. However, the implementation of this update poses several challenges, which we outline below along with potential solutions.

First, computing ∇_n involves evaluating an integral over a segment, which is computationally expensive. One remedy is to approximate the average gradient over the segment by evaluating the gradient at a randomly selected point along the line segment connecting \mathbf{x}_{n-1} and \mathbf{x}_n ; in expectation, this yields the same value as ∇_n . An alternative, deterministic approach is to use the gradient at the midpoint $\frac{1}{2}(\mathbf{x}_{n-1} + \mathbf{x}_n)$ as an approximation to ∇_n . Under the assumption of Lipschitz continuity of the Hessian, it can be shown that the resulting error is $\mathcal{O}(LD^2)$. In this work, we adopt the second approach and define the midpoint as $\mathbf{w}_n = \frac{1}{2}(\mathbf{x}_{n-1} + \mathbf{x}_n)$.

However, two key challenges remain. First, in the stochastic setting, we do not have access to the gradient $\nabla F(\mathbf{w}_n)$, but only to a stochastic gradient oracle (see Assumption 2.3). To resolve this, we use an unbiased stochastic estimate of the gradient at the midpoint, denoted by $\mathbf{g}_n = \nabla f(\mathbf{w}_n; \xi_n)$, where $\xi_n \sim \mathcal{D}$. In the rest of the paper, we assume access to a stochastic gradient oracle with variance σ^2 . By setting $\sigma^2 = 0$, our analysis and algorithm naturally recover the deterministic setting.

The second and perhaps most critical challenge, which arises in both deterministic and stochastic settings, is that computing the (stochastic) gradient at the midpoint \mathbf{w}_n requires knowledge of the next iterate \mathbf{x}_n , which itself depends on the update direction Δ_n . Since \mathbf{x}_n is not available when choosing Δ_n , this creates a circular dependency. To overcome this, we cast the selection of Δ_n as an *online learning problem*, where Δ_n is the action taken at round n , and the loss is defined as $\langle \mathbf{g}_n, \Delta_n \rangle$. This loss reflects how well the chosen direction aligns with the negative of the (stochastic) gradient evaluated at the midpoint, thus guiding the updates even in the absence of \mathbf{x}_n at decision time. Concretely, the online-to-nonconvex conversion framework proceeds as follows

- Update $\mathbf{x}_n = \mathbf{x}_{n-1} + \Delta_n$ and $\mathbf{w}_n = \mathbf{x}_{n-1} + \frac{1}{2}\Delta_n$;
- Construct a stochastic estimate $\mathbf{g}_n = \nabla f(\mathbf{w}_n; \xi_n)$ with $\xi_n \sim \mathcal{D}$;
- Feed the loss $\langle \mathbf{g}_n, \Delta_n \rangle$ to an online learning algorithm to obtain Δ_{n+1} with $\|\Delta_{n+1}\| \leq D$.

A key result from Cutkosky et al. (2023) establishes that convergence to stationarity in Problem (1) can be related to a specific notion of regret in the online learning problem above. Specifically, suppose the process runs for $M = KT$ iterations, divided into K episodes of length T each. We then define the *K-shifting regret* as

$$\text{Reg}_T(\mathbf{u}^1, \dots, \mathbf{u}^K) = \sum_{k=1}^K \sum_{n=(k-1)T+1}^{kT} \langle \mathbf{g}_n, \Delta_n - \mathbf{u}^k \rangle, \quad (4)$$

where $\{\mathbf{u}^k\}_{k=1}^K$ is a sequence of comparator vectors to be specified. In the next proposition, we bound the gradient norm at the average iterates in terms of $\text{Reg}_T(\mathbf{u}^1, \dots, \mathbf{u}^K)$. Due to space constraints, we defer the detailed discussions and the proof to the Appendix.

Proposition 2.1. *Suppose that Assumptions 2.2 and 2.3 hold. For $k = 1, \dots, K$, define $\bar{\mathbf{w}}^k = \frac{1}{T} \sum_{n=(k-1)T+1}^{kT} \mathbf{w}_n$ and $\mathbf{u}^k = -D \frac{\sum_{n=(k-1)T+1}^{kT} \mathbf{g}_n}{\|\sum_{n=(k-1)T+1}^{kT} \mathbf{g}_n\|}$. Recall the definition of $\text{Reg}_T(\mathbf{u}^1, \dots, \mathbf{u}^K)$ in (4). Then we have*

$$\mathbb{E} \left[\frac{1}{K} \sum_{k=1}^K \|\nabla F(\bar{\mathbf{w}}^k)\| \right] \leq \frac{F(\mathbf{x}_0) - F^*}{DKT} + \frac{\mathbb{E} [\text{Reg}_T(\mathbf{u}^1, \dots, \mathbf{u}^K)]}{DKT} + \frac{L_2}{48} D^2 + \frac{L_2}{2} T^2 D^2 + \frac{\sigma}{\sqrt{T}}.$$

In Proposition 2.1, the returned points $\{\bar{\mathbf{w}}^k\}_{k=1}^K$ represent the average iterates within each episode, and the comparators $\{\mathbf{u}^k\}_{k=1}^K$ are defined based on the sum of stochastic gradients in each episode. Moreover, the proposition shows that the average gradient norm at $\{\bar{\mathbf{w}}^k\}_{k=1}^K$ can be bounded in terms of the *K-shifting regret* with respect to these comparators. Hence, to obtain a gradient complexity bound for finding a stationary point of the function F , it suffices to use an online learning algorithm

to minimize the regret. This approach yields an explicit update for Δ_n , and consequently for the iterates. Specifically, the online learning problem we need to solve is presented in the box below.

In Section 3, we introduce our *online doubly optimistic gradient method* along with a new hint function, designed to efficiently solve the above online learning problem in both deterministic and stochastic settings. Before that, we briefly review the online learning algorithms proposed in Cutkosky et al. (2023), originally developed for these settings, and highlight their limitations. Notably, the approaches differ between the deterministic and stochastic cases to achieve the best possible regret bounds and corresponding gradient complexity.

Online Learning Problem

For $n = 1, \dots, KT$:

- The learner chooses $\Delta_n \in \mathbb{R}^d$ such that $\|\Delta_n\| \leq D$;
- Sample $\xi_n \sim \mathcal{D}$ and compute $\mathbf{g}_n = \nabla f(\frac{1}{2}(\mathbf{x}_n + \mathbf{x}_{n-1}); \xi_n)$, where $\mathbf{x}_n = \mathbf{x}_{n-1} + \Delta_n$;
- The learner observes the loss $\ell_n(\Delta_n) = \langle \mathbf{g}_n, \Delta_n \rangle$.

Goal: Minimize the regret $\text{Reg}_T(\mathbf{u}^1, \dots, \mathbf{u}^K) = \sum_{k=1}^K \sum_{n=(k-1)T+1}^{kT} \langle \mathbf{g}_n, \Delta_n - \mathbf{u}^k \rangle$.

Deterministic setting. In the deterministic setting, Cutkosky et al. (2023) employs an optimistic gradient method, which relies on a hint vector \mathbf{h}_n to approximate the true gradient \mathbf{g}_n as closely as possible. The regret of these methods is governed by the approximation error between \mathbf{h}_n and \mathbf{g}_n , as shown later in Lemma 3.1. Hence, a natural choice for \mathbf{h}_n is to leverage the smoothness of F and set $\mathbf{h}_n = \mathbf{g}_{n-1}$. However, this only leads to a complexity of $\mathcal{O}(\varepsilon^{-\frac{1}{6}})$ that is suboptimal compared to the state-of-the-art, as discussed in Section 3.2. To achieve a better rate, Cutkosky et al. (2023) build on a variant of optimistic gradient descent from Chen et al. (2021) and propose a more sophisticated hint construction. Specifically, starting from $\hat{\Delta} = 0$, they follow the update rule $\Delta_n = \Pi_{\{\|\Delta\| \leq D\}}(\hat{\Delta}_n - \eta \mathbf{h}_n)$, $\hat{\Delta}_{n+1} = \Pi_{\{\|\Delta\| \leq D\}}(\hat{\Delta}_n - \eta \mathbf{g}_n)$ where $\Pi_{\{\|\Delta\| \leq D\}}(\cdot)$ denotes projection onto the Euclidean ball of radius D . As shown by the update rule, the algorithm maintains two sequences: $\hat{\Delta}_n$ and Δ_n . The auxiliary sequence $\hat{\Delta}_n$ is designed to closely approximate Δ_n and plays a key role in constructing the hint vector \mathbf{h}_n . Specifically, \mathbf{h}_n is computed through an inner loop that solves the following fixed-point equation for \mathbf{h} , which depends on $\hat{\Delta}_n$, $\mathbf{h} = \nabla F(\mathbf{x}_{n-1} + \frac{1}{2} \Pi_{\{\|\Delta\| \leq D\}}[\hat{\Delta}_n - \eta \mathbf{h}])$. While successfully achieving the complexity of $\mathcal{O}(\varepsilon^{-1.75} \log(\frac{1}{\varepsilon}))$, this hint construction introduces two drawbacks: the inner loop adds an extra logarithmic factor to the final complexity bound, and the fixed-point formulation complicates its extension to the stochastic setting.

Stochastic setting. In the stochastic setting, Cutkosky et al. (2023) proposes using standard projected online gradient descent (OGD) to solve the online learning problems. The update rule follows $\Delta_{n+1} = \Pi_{\{\|\Delta\| \leq D\}}(\Delta_n - \eta \mathbf{g}_n)$. Under the assumption that the stochastic gradient has a bounded second moment, i.e., $\mathbb{E}[\|\nabla f(\mathbf{x}; \xi)\|^2] \leq G^2$ for all \mathbf{x} , they prove that the regret of OGD can be bounded by $\text{Reg}_T(\mathbf{u}^1, \dots, \mathbf{u}^K) = \mathcal{O}(KGD\sqrt{T})$. Under Assumption 2.2 and with proper tuning of the hyperparameters, this leads to an improved complexity of $\mathcal{O}(\varepsilon^{-3.5})$. However, note that this rate does not improve in the deterministic setting, where $\sigma = 0$. Moreover, the bounded second moment condition required to guarantee regret bounds using OGD can be a restrictive assumption to impose, which is not necessary in other existing methods Cutkosky & Mehta (2020).

To address the limitations identified above in both the stochastic and deterministic settings, we develop our ‘‘Online Doubly Optimistic Gradient’’ method. In the following section, we explain how the algorithm leverages two layers of optimism to overcome these challenges and provide a comprehensive analysis of its complexity.

3 PROPOSED ALGORITHM

This section introduces our proposed algorithm and the key ideas behind its design. Building on the O2NC framework Cutkosky et al. (2023), we reformulate the problem of finding a stationary point of

F as an online learning task. To solve the resulting OCO problem, we present a modified variant of the Online Optimistic Gradient method, distinct from the one discussed in the previous section. In addition, as our main algorithmic contribution, we propose a novel hint function derived from a two-level optimism scheme, detailed below. Hence, we refer to our approach as the ‘‘Online Doubly Optimistic Gradient’’ method.

3.1 ONLINE OPTIMISTIC GRADIENT METHOD

As mentioned earlier, we also aim to solve the online learning problem described in Section 2.1. Unlike the optimistic template used in Cutkosky et al. (2023), which requires two projections per iteration, we adopt a simpler variant of the optimistic gradient method that requires only a single projection. Specifically, for $n > 1$, we update the direction Δ_{n+1} using the following rule

$$\Delta_{n+1} = \Pi_{\|\Delta\| \leq D} (\Delta_n - \eta \mathbf{h}_{n+1} - \eta(\mathbf{g}_n - \mathbf{h}_n)), \quad \forall n > 1, \quad (5)$$

and for $n = 1$, we initialize with $\Delta_1 = \arg \min_{\|\Delta\| \leq D} \langle \mathbf{h}_1, \Delta \rangle$.

To motivate the update rule introduced above, observe that a natural approach would be to base the update of Δ_{n+1} on the current gradient \mathbf{g}_{n+1} , but this is not feasible because \mathbf{g}_{n+1} is revealed only after Δ_{n+1} has already been selected. To address this challenge, we leverage online optimistic learning techniques from Rakhlin & Sridharan (2013); Joulani et al. (2020), introducing the first layer of optimism in our *doubly optimistic algorithm*. These methods rely on a hint vector \mathbf{h}_n that approximates \mathbf{g}_n as accurately as possible. The intuition of this approach assumes that the deviation between the true gradient and the hint remains relatively stable across iterations, i.e., $\mathbf{g}_{n+1} - \mathbf{h}_{n+1} \approx \mathbf{g}_n - \mathbf{h}_n$. Accordingly, one can estimate $\mathbf{g}_{n+1} \approx \mathbf{h}_{n+1} + \mathbf{g}_n - \mathbf{h}_n$, and the update combines an optimistic prediction using \mathbf{h}_{n+1} with a correction based on the observed error $\mathbf{g}_n - \mathbf{h}_n$.

Next, we present the K -shifting regret bound for the optimistic gradient method described in (5). This result holds for any hint function and, accordingly, depends on the approximation error between \mathbf{g}_n and \mathbf{h}_n , as shown in the following lemma.

Lemma 3.1. *Consider executing the update described in (5) with a constant step size η . Then the K -shifting regret can be bounded by*

$$\text{Reg}_T(\mathbf{u}^1, \dots, \mathbf{u}^K) \leq \frac{4KD^2}{\eta} + \frac{5\eta}{2} \sum_{n=1}^{KT} \|\mathbf{g}_n - \mathbf{h}_n\|^2 - \frac{1}{4\eta} \sum_{n=2}^{KT} \|\Delta_n - \Delta_{n-1}\|^2$$

Lemma 3.1 highlights the crucial role of the hint vector \mathbf{h}_n : the more accurately it approximates the true gradient \mathbf{g}_n , the more favorable the convergence behavior of the algorithm becomes. Precise hints tighten the regret bound, while inaccurate approximations can substantially degrade performance. Lemma 3.1 also establishes the foundation for the second layer of optimism in our algorithm. In particular, it features a negative term, $\|\Delta_n - \Delta_{n-1}\|^2$, which was previously neglected in Cutkosky et al. (2023). Under the smoothness assumption, we can reasonably expect that $\Delta_n \approx \Delta_{n-1}$. Our second layer of optimism capitalizes on this assumption and incorporates the previously neglected term, establishing a connection between this term and $\|\mathbf{g}_n - \mathbf{h}_n\|^2$ under the Lipschitz continuity of the gradient. This connection allows us to use telescoping techniques, thereby improving upon previous convergence rates within this framework. We now address the critical question of how to construct the hint sequence \mathbf{h}_n to minimize the regret bound established in Lemma 3.1.

3.2 HINT CONSTRUCTION

For ease of exposition, we first assume access to a deterministic gradient oracle; we will return to the stochastic setting at the end of this section. As explained in Section 3.1, our primary objective in hint construction is to estimate $\mathbf{g}_n = \nabla F(\mathbf{w}_n)$ with maximum accuracy using only information available up to \mathbf{x}_{n-1} , as illustrated in Figure 1. Leveraging the function’s smoothness properties, an intuitive first approximation is to reuse the previous gradient estimate, assuming $\mathbf{g}_n \approx \mathbf{g}_{n-1}$ —effectively positing that the gradient at \mathbf{w}_n approximates the gradient at \mathbf{w}_{n-1} . This leads to the following update rule, which corresponds to the standard optimistic online gradient descent

$$\Delta_{n+1} = \Pi_{\|\Delta\| \leq D} (\Delta_n - \eta(2\mathbf{g}_n - \mathbf{g}_{n-1})).$$

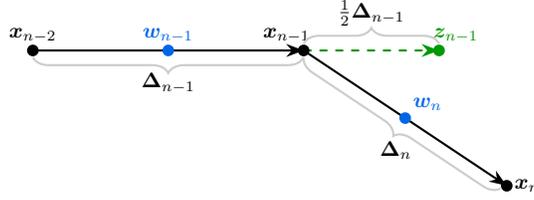


Figure 1: Illustration for the extrapolated point in our algorithm.

However, this straightforward approximation fails to exploit the negative term $\|\Delta_n - \Delta_{n-1}\|^2$ in Lemma 3.1. With this approach, $\|\mathbf{g}_n - \mathbf{h}_n\| \leq L\|\mathbf{w}_n - \mathbf{w}_{n-1}\|$ becomes proportional to $\|\Delta_n + \Delta_{n-1}\|$, preventing the application of telescoping arguments. Consequently, this method yields an overall gradient complexity of $\mathcal{O}(\varepsilon^{-11/6})$ in the deterministic setting (Cutkosky et al., 2023), which is worse than the best-known bound of $\mathcal{O}(\varepsilon^{-7/4})$.

Cutkosky et al. (2023) improved this rate under the O2NC framework by constructing the hint \mathbf{h}_n through a fixed-point iteration scheme. Their approach develops an approximation $\hat{\Delta}_n$ of Δ_n via multiple iteration steps, estimating $\mathbf{g}_n \approx \nabla F(\mathbf{x}_{n-1} + \frac{1}{2}\hat{\Delta}_n)$. While this method achieves an improved convergence rate of $\mathcal{O}(\varepsilon^{-7/4} \log(1/\varepsilon))$ in the deterministic setting, it introduces an additional logarithmic factor. Moreover, this construction is complex and presents challenges when extended to stochastic environments. These limitations motivate our search for an alternative that maintains theoretical guarantees while offering greater simplicity and broader applicability.

Our method not only removes the logarithmic factor in the deterministic case but also matches the known lower bound in the stochastic case, achieving a complexity of $\mathcal{O}(\varepsilon^{-1.75} + \sigma^2\varepsilon^{-3.5})$. We leverage our second layer of optimism to improve upon previous results. Referring to Figure 1, our goal is to estimate the gradient at \mathbf{w}_n using only information available up to \mathbf{x}_{n-1} . Since $\mathbf{w}_n = \mathbf{x}_{n-1} + \frac{1}{2}\Delta_n$ and we can reasonably approximate $\Delta_n \approx \Delta_{n-1}$ in the smooth setting, we propose to estimate the gradient at \mathbf{w}_n using the gradient at an extrapolated point $\mathbf{z}_{n-1} = \mathbf{x}_{n-1} + \frac{1}{2}\Delta_{n-1}$. This leads to the following elegant construction for the hint

$$\mathbf{h}_n = \nabla F(\mathbf{z}_{n-1}) = \nabla F(\mathbf{x}_{n-1} + \frac{1}{2}\Delta_{n-1}). \quad (6)$$

We use the convention $\Delta_0 = 0$, which implies that $\mathbf{h}_1 = \nabla F(\mathbf{x}_0)$. This approximation has a clear mathematical justification from the regret analysis. Recalling from Lemma 3.1 the negative term $\|\Delta_n - \Delta_{n-1}\|^2$ that was previously neglected in Cutkosky et al. (2023), we can now see that under this hint construction and using Assumption 2.1, we have $\|\mathbf{g}_n - \mathbf{h}_n\| \propto \|\mathbf{w}_n - \mathbf{z}_{n-1}\|$, and by construction $\|\mathbf{w}_n - \mathbf{z}_{n-1}\| = \frac{1}{2}\|\Delta_n - \Delta_{n-1}\|$. With appropriate stepsize selection, this allows us to leverage telescoping between these two terms and obtain tighter convergence rates.

The novelty of our hint construction lies in its simplicity: it requires only one gradient call compared to the multi-step procedure in Cutkosky et al. (2023), and it significantly simplifies the analysis. In particular, this makes it especially well-suited for the stochastic setting, where the only modification needed is replacing the deterministic gradient with its stochastic counterpart, i.e., $\mathbf{h}_{n+1} = \nabla f(\mathbf{z}_n; \xi_n)$, where ξ_n is the random sample drawn at iteration n . In this setting, we continue to leverage the telescoping argument, with only an additional term accounting for the stochastic gradient variance. The regret incurred by our hint is formalized in the following lemma.

Lemma 3.2. *Consider executing the update described in (5) with the hint construction defined in (6) (or its stochastic counterpart in the stochastic setting), using step size $\eta \leq \frac{1}{\sqrt{3}L_1}$. Then we can bound one episode of the K -shifting regret, as*

$$\mathbb{E} \left[\sum_{n=(k-1)T+1}^{kT} \langle \mathbf{g}_n, \Delta_n - \mathbf{u}^k \rangle \right] \leq \frac{4D^2}{\eta} + \frac{9L_1^2\eta D^2}{2} + 6\eta T\sigma^2.$$

The stochastic variant of our proposed method is summarized in Algorithm 1. The deterministic version can be obtained by simply replacing the stochastic gradients with their deterministic coun-

Algorithm 1 Online Doubly Optimistic Gradient Algorithm**Require:** Initial point \mathbf{x}_0 , $K, T \in \mathbb{N}$, radius D **Initialize:** $\Delta_1 = -D \frac{\nabla f(\mathbf{x}_0; \xi_0)}{\|\nabla f(\mathbf{x}_0; \xi_0)\|}$, $\mathbf{h}_1 = \nabla f(\mathbf{x}_0; \xi_0)$

- 1: **for** $n = 1$ to KT **do**
- 2: Set $\mathbf{x}_n = \mathbf{x}_{n-1} + \Delta_n$, $\mathbf{w}_n = \mathbf{x}_{n-1} + \frac{1}{2}\Delta_n$, $\mathbf{z}_n = \mathbf{x}_n + \frac{1}{2}\Delta_n$
- 3: Set $\Delta_{n+1} = \Pi_{\|\Delta\| \leq D} (\Delta_n - \eta_n \nabla f(\mathbf{z}_n; \xi_n) - \eta_n (\nabla f(\mathbf{w}_n; \xi_n) - \nabla f(\mathbf{z}_{n-1}; \xi_{n-1})))$
- 4: **end for**
- 5: Set $\mathbf{w}_t^k = \mathbf{w}_{(k-1)T+t}$ for $k = 1, \dots, K$ and $t = 1, \dots, T$
- 6: Set $\bar{\mathbf{w}}^k = \frac{1}{T} \sum_{t=1}^T \mathbf{w}_t^k$ for $k = 1, \dots, K$
- 7: **return** $\hat{\mathbf{w}} = \begin{cases} \text{Deterministic Oracle: } \arg \min_{\bar{\mathbf{w}} \in \{\bar{\mathbf{w}}^1, \dots, \bar{\mathbf{w}}^K\}} \|\nabla F(\bar{\mathbf{w}})\|, \\ \text{Stochastic Oracle: Uniformly sample } \{\bar{\mathbf{w}}^1, \dots, \bar{\mathbf{w}}^K\}. \end{cases}$

terparts. In the following section, we present a comprehensive complexity analysis to establish the convergence rate of the algorithm.

3.3 COMPLEXITY ANALYSIS

Thus far, we have presented our proposed ‘‘Online Doubly Optimistic Gradient’’ method in Algorithm 1, and discussed the construction of the hint vector \mathbf{h}_n . Next, leveraging Lemma 3.2 and Proposition 2.1, we establish the convergence rate of our proposed method.

Theorem 3.3. *If Assumptions 2.1–2.3 hold and we set $D = \Theta(\min\{(ML_2^{\frac{1}{5}}\sigma^{\frac{4}{5}})^{-\frac{5}{7}}, (ML_1^{\frac{2}{3}}L_2^{\frac{1}{3}})^{-\frac{3}{7}}\})$, $T = \Theta(\min\{\max\{[(\frac{\sigma}{L_2 D^2})^{\frac{2}{5}}], [(\frac{L_1}{L_2 D})^{\frac{1}{3}}]\}, \frac{M}{2}\})$, $K = \lfloor \frac{M}{T} \rfloor$, and $\eta = \Theta(\frac{1}{\sqrt{L_1^2 + \sigma^2 T/D^2}})$ in Algorithm 1, then*

$$\mathbb{E} \left[\frac{1}{K} \sum_{k=1}^K \|\nabla F(\bar{\mathbf{w}}^k)\| \right] = \mathcal{O} \left((F(\mathbf{x}_0) - F^*)^{\frac{2}{7}} L_2^{\frac{1}{7}} \frac{\sigma^{\frac{4}{7}}}{M^{\frac{2}{7}}} + L_1^{\frac{2}{7}} L_2^{\frac{1}{7}} \frac{(F(\mathbf{x}_0) - F^*)^{\frac{4}{7}}}{M^{\frac{4}{7}}} \right). \quad (7)$$

Proof Sketch. Based on Proposition 2.1, achieving the final convergence rate requires bounding the K -shifting regret. As a corollary of Lemma 3.2, we get $\mathbb{E} [\text{Reg}_T(\mathbf{u}^1, \dots, \mathbf{u}^K)] \leq \frac{4KD^2}{\eta} + \frac{9L_1^2\eta KD^2}{2} + 6\eta KT\sigma^2$. Combining this regret bound with Proposition 2.1, we obtain

$$\mathbb{E} \left[\frac{1}{K} \sum_{k=1}^K \|\nabla F(\bar{\mathbf{w}}^k)\| \right] \leq \frac{F(\mathbf{x}_0) - F^*}{DKT} + \frac{4D}{T\eta} + \frac{9L_1^2\eta D}{2T} + \frac{6\eta\sigma^2}{D} + \frac{L_2}{48} D^2 + \frac{L_2}{2} T^2 D^2 + \frac{\sigma}{\sqrt{T}}.$$

Finally, by selecting the parameters η , D , and T according to the theorem’s prescription, we arrive at the stated convergence rate. The definitions of D and T involve a max and min, respectively, to interpolate between the deterministic and stochastic regimes. This allows the final bound to adapt to the variance σ of the stochastic oracle, interpolating between both settings. \square

We show that after M iterations, the gradient norm at the best iterate satisfies $\min_{1 \leq k \leq K} \mathbb{E} [\|\nabla F(\bar{\mathbf{w}}^k)\|] = \mathcal{O} \left(\frac{\sigma^{4/7}}{M^{2/7}} + \frac{1}{M^{4/7}} \right)$, which yields a total gradient complexity of $\mathcal{O}(\sigma^2 \varepsilon^{-3.5} + \varepsilon^{-1.75})$. This matches the lower bound $\Omega(\sigma^2 \varepsilon^{-3.5})$ in the stochastic regime Arjevani et al. (2023) and recovers the best known rate $\mathcal{O}(\varepsilon^{-1.75})$ in the deterministic setting, smoothly interpolating between these two regimes.

Remark 3.1. While our paper focuses on the smooth setting, Algorithm 1 also recovers the optimal convergence rate in the nonsmooth setting with properly chosen parameters. In this case, we assume that $\mathbb{E} [\|\nabla f(\mathbf{x}; \xi)\|^2] \leq G^2$ for all $\mathbf{x} \in \mathbb{R}^d$ and follow the stationary notation of Cutkosky et al. (2023). Under this assumption, Algorithm 1 achieves a complexity of $\mathcal{O}(\varepsilon^{-3}\delta^{-1})$ for finding (δ, ε) -stationary points. This guarantee is obtained by adapting our analysis to the nonsmooth regime. Specifically, we replace Proposition 2.1 with Theorem 8 of Cutkosky et al. (2023), which serves as its nonsmooth analogue, and in the regret bound of Lemma 3.1 we discard the negative term and bound $\|\mathbf{g}_n - \mathbf{h}_n\|^2$ directly using the bounded second-moment assumption. The remaining steps follow the same structure as the analysis in Cutkosky et al. (2023).

4 ADAPTIVE STEP SIZE SCHEME

In the previous section, we analyzed Algorithm 1 with a constant step size η and established its convergence rate. However, as shown in Theorem 3.3, the choice of η depends on the global gradient Lipschitz constant and the episode length T , which can be overly conservative and leads to slow convergence. To address this, it is often desirable to select the step size adaptively to exploit the local Lipschitz continuity of the objective function. In this section, we extend Algorithm 1 to incorporate an adaptive step size scheme that automatically adjusts the step size based on past gradient information.

Inspired by the adaptive step size analyses developed in Kavis et al. (2019); Antonakopoulos et al. (2022); Levy et al. (2018), we define the learning rate as

$$\eta_n = \frac{\gamma D}{\sqrt{\alpha + \sum_{i=1}^{n \bmod T} \|\mathbf{g}_i^k - \mathbf{h}_i^k\|^2}}, \quad (8)$$

where $\gamma > 0$ controls the initial step size and $\alpha > 0$ is a small constant added for numerical stability. Here, $\mathbf{g}_i^k = \mathbf{g}_{(k-1)T+i}$ and $\mathbf{h}_i^k = \mathbf{h}_{(k-1)T+i}$, with k indicating the episode and i the iteration index within that episode. The sum $\sum_{i=1}^{n \bmod T} \|\mathbf{g}_i^k - \mathbf{h}_i^k\|^2$ is reset at the beginning of each episode, which facilitates the theoretical analysis of regret in the Online Learning Problem 1.

The key challenge in analyzing the adaptive step size scheme in (8) is that the step size η_n does not admit a direct upper bound, making it difficult to leverage the negative term in the regret analysis. Specifically, similar to Lemma 3.1, we can bound the regret for the first episode as (the subsequent episodes follow by similar arguments) $\sum_{n=1}^T \langle \mathbf{g}_n, \Delta_n - \mathbf{u}^1 \rangle \leq \frac{3D^2}{\eta_T} + \sum_{n=1}^T \left(\frac{3\eta_n}{2} \|\mathbf{g}_n - \mathbf{h}_n\|^2 - \frac{1}{4\eta_n} \|\Delta_n - \Delta_{n-1}\|^2 \right)$. In the constant step size setting, as shown in the proof of Lemma 3.2, the key to applying the telescoping argument is the ability to choose $\eta \leq \frac{1}{\sqrt{3}L_1}$, ensuring that the negative term $-\frac{1}{4\eta} \|\Delta_n - \Delta_{n-1}\|^2$ balances the deterministic approximation error $\frac{3\eta}{2} \|\nabla F(\mathbf{w}_n) - \nabla F(\mathbf{z}_{n-1})\|^2$. However, in the adaptive setting, we have no guarantee that η_n is bounded by $\frac{1}{\sqrt{3}L_1}$, making the telescoping argument not directly applicable.

To overcome this difficulty, we adopt a thresholding strategy inspired by prior work in Kavis et al. (2019); Antonakopoulos et al. (2022); Levy et al. (2018). At a high level, we partition the iterations within each episode into two categories. For iterations where η_n is below a certain threshold, we show that the negative term dominates the positive one. Conversely, when η_n is large, the definition of (8) implies that the denominator $\sum_{i=1}^{n \bmod T} \|\mathbf{g}_i^k - \mathbf{h}_i^k\|^2$ is small, which allows us to control the cumulative contribution of the positive terms. We refer the reader to the appendix for further details.

Under the adaptive step size η_n , we establish the following lemma, which bounds the regret incurred during a single episode of the K -shifting regret. Without loss of generality, we present the result for the first episode; the analysis extends straightforwardly to any subsequent episode.

Lemma 4.1. *Consider executing the update described in (5), using \mathbf{h}_n as defined in (6) (or its stochastic counterpart in the stochastic setting), and the adaptive step size defined in (8). Then we can bound the first episode of the K -shifting regret as*

$$\sum_{n=1}^T \mathbb{E} [\langle \mathbf{g}_n, \Delta_n - \mathbf{u}^1 \rangle] \leq 8 \left(\frac{3}{\gamma} + \gamma \right) D \sqrt{T} \sigma + 16 \left(\frac{3}{\gamma} + \gamma \right)^{\frac{3}{2}} \hat{L}_1 \gamma^{\frac{1}{2}} D^2,$$

where $\hat{L}_1 = \max_{1 \leq n \leq T} \frac{\|\mathbf{g}_n - \mathbf{h}_n\|}{\|\mathbf{w}_n - \mathbf{z}_{n-1}\|}$.

Lemma 4.1 plays the same role as Lemma 3.2 in the constant step size setting. Notably, instead of relying on the global gradient Lipschitz constant, the bound depends on \hat{L}_1 , which can be interpreted as a local Lipschitz constant and may be significantly smaller in practice. We are now ready to present the final convergence guarantee for this adaptive variant of the ‘‘Online Doubly Optimistic Gradient’’ method.

Theorem 4.2. *Suppose Assumptions 2.3, 2.1, and 2.2 hold. If we run Algorithm 1 with \mathbf{h}_n as described in (6) (or its stochastic counterpart in the stochastic setting), the values for D , T , and K are selected as the ones in Theorem 3.3, and η_n as described in (8), replacing L_1 with \hat{L}_1^* , where*

$\hat{L}_1^* = \max_{1 \leq n \leq M} \frac{\|\mathbf{g}_n - \mathbf{h}_n\|}{\|\mathbf{w}_n - \mathbf{z}_{n-1}\|}$, then the following holds

$$\mathbb{E} \left[\frac{1}{K} \sum_{k=1}^K \|\nabla F(\bar{\mathbf{w}}^k)\| \right] = \mathcal{O} \left((F(\mathbf{x}_0) - F^*)^{\frac{2}{7}} L_2^{\frac{1}{7}} \frac{\sigma^{\frac{4}{7}}}{M^{\frac{2}{7}}} + (\hat{L}_1^*)^{\frac{2}{7}} L_2^{\frac{1}{7}} \frac{(F(\mathbf{x}_0) - F^*)^{\frac{4}{7}}}{M^{\frac{4}{7}}} \right). \quad (9)$$

Theorem 4.2 shows that our adaptive step size scheme maintains the same complexity of $\mathcal{O}(\varepsilon^{-1.75} + \sigma^2 \varepsilon^{-3.5})$ without requiring manual tuning of the learning rate, while offering improved dependence on the gradient Lipschitz constant compared to the constant step size scheme. Moreover, Cutkosky & Mehta (2020) posed an open problem of designing an algorithm that achieves this complexity without knowledge of the variance σ^2 . Our result represents a first step towards this goal by introducing an adaptive step size. A promising future direction is to further adapt the choices of D and T , ultimately aiming to make the algorithm fully parameter-free.

5 CONCLUSION

In this paper, we build on the online-to-nonconvex conversion framework of Cutkosky et al. (2023) and propose a simple algorithm for solving smooth nonconvex optimization problems, which only requires two (stochastic) gradient queries per iteration. Our key contribution is a doubly optimistic hint function for the online optimistic gradient method, thus eliminating the fixed point iteration subroutine from (Cutkosky et al., 2023). Assuming that both the gradient and the Hessian of the objective are Lipschitz continuous and that the stochastic gradient has bounded variance σ^2 , our algorithm achieves a complexity of $\mathcal{O}(\varepsilon^{-1.75} + \sigma^2 \varepsilon^{-3.5})$. This result smoothly interpolates between the best-known deterministic rate and the worst-case optimal stochastic rate. In addition, we develop an adaptive step size strategy for our algorithm, marking a first step toward a potential future direction: the design of fully adaptive, parameter-free algorithms that match this performance without relying on prior knowledge of problem-specific parameters.

ACKNOWLEDGMENTS

This work was supported in part by the NSF CAREER Award CCF-2338846, the NSF AI Institute for Foundations of Machine Learning (IFML), and the NSF AI Institute for Future Edge Networks and Distributed Intelligence (AI-EDGE).

REFERENCES

- Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding approximate local minima faster than gradient descent. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1195–1199, 2017.
- Kwangjun Ahn and Ashok Cutkosky. Adam with model exponential moving average is effective for nonconvex optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- Kwangjun Ahn, Gagik Magakyan, and Ashok Cutkosky. General framework for online-to-nonconvex conversion: Schedule-free sgd is also effective for nonconvex optimization. *arXiv preprint arXiv:2411.07061*, 2024a.
- Kwangjun Ahn, Zhiyu Zhang, Yunbum Kook, and Yan Dai. Understanding adam optimizer via online learning of updates: Adam is ftrl in disguise. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 619–640, 2024b.
- Zeyuan Allen-Zhu. How to make the gradients small stochastically: Even faster convex and nonconvex sgd. *Advances in Neural Information Processing Systems*, 31, 2018a.
- Zeyuan Allen-Zhu. Natasha 2: Faster non-convex optimization than sgd. *Advances in neural information processing systems*, 31, 2018b.
- Zeyuan Allen-Zhu and Yuanzhi Li. Neon2: Finding local minima via first-order oracles. *Advances in Neural Information Processing Systems*, 31, 2018.

- Kimon Antonakopoulos, Ali Kavis, and Volkan Cevher. Extra-newton: A first approach to noise-adaptive accelerated second-order methods. *Advances in Neural Information Processing Systems*, 35:29859–29872, 2022.
- Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Ayush Sekhari, and Karthik Sridharan. Second-order information in non-convex stochastic optimization: Power and limitations. In *Conference on Learning Theory*, pp. 242–299. PMLR, 2020.
- Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1): 165–214, 2023.
- Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. “convex until proven guilty”: Dimension-free acceleration of gradient descent on non-convex functions. In *International conference on machine learning*, pp. 654–663. PMLR, 2017.
- Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018.
- Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, 184(1):71–120, 2020.
- Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points ii: first-order methods. *Mathematical Programming*, 185(1):315–355, 2021.
- Liyu Chen, Haipeng Luo, and Chen-Yu Wei. Impossible tuning made possible: A new expert algorithm and its applications. In *Conference on Learning Theory*, pp. 1216–1259. PMLR, 2021.
- Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In *Conference on Learning Theory*, pp. 6–1. JMLR Workshop and Conference Proceedings, 2012.
- Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In *International conference on machine learning*, pp. 2260–2268. PMLR, 2020.
- Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex SGD. *Advances in neural information processing systems*, 32, 2019.
- Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. In *International Conference on Machine Learning*, pp. 6643–6670. PMLR, 2023.
- Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in neural information processing systems*, 31, 2018.
- Cong Fang, Zhouchen Lin, and Tong Zhang. Sharp analysis for nonconvex SGD escaping from saddle points. In *Conference on Learning Theory*, pp. 1192–1234. PMLR, 2019.
- Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.
- Ruichen Jiang, Aryan Mokhtari, and Francisco Patitucci. Improved complexity for smooth nonconvex optimization: a two-level online learning approach with quasi-newton methods. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pp. 2225–2236, 2025.
- Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In *Conference On Learning Theory*, pp. 1042–1085. PMLR, 2018.
- Pooria Joulani, András György, and Csaba Szepesvári. A modular analysis of adaptive (non-) convex optimization: Optimism, composite objectives, variance reduction, and variational bounds. *Theoretical Computer Science*, 808:108–138, 2020.

- Ali Kavis, Kfir Y Levy, Francis Bach, and Volkan Cevher. Unixgrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization. *Advances in neural information processing systems*, 32, 2019.
- Kfir Y Levy, Alp Yurtsever, and Volkan Cevher. Online adaptive methods, universality and acceleration. *Advances in neural information processing systems*, 31, 2018.
- Huan Li and Zhouchen Lin. Restarted nonconvex accelerated gradient descent: No more polylogarithmic factor in the $O(\epsilon^{-7/4})$ complexity. In *International Conference on Machine Learning*, pp. 12901–12916. PMLR, 2022.
- Huan Li and Zhouchen Lin. Restarted nonconvex accelerated gradient descent: No more polylogarithmic factor in the in the $O(\epsilon^{-7/4})$ complexity. *Journal of Machine Learning Research*, 24(157): 1–37, 2023.
- Naoki Marumo and Akiko Takeda. Parameter-free accelerated gradient descent for nonconvex minimization. *SIAM Journal on Optimization*, 34(2):2093–2120, 2024a.
- Naoki Marumo and Akiko Takeda. Universal heavy-ball method for nonconvex optimization under hölder continuous Hessians. *Mathematical Programming*, pp. 1–29, 2024b.
- Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical programming*, 108(1):177–205, 2006.
- Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory*, pp. 993–1019. PMLR, 2013.
- Clément W Royer and Stephen J Wright. Complexity analysis of second-order line-search algorithms for smooth nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1448–1477, 2018.
- Clément W Royer, Michael O’Neill, and Stephen J Wright. A Newton-CG algorithm with complexity guarantees for smooth unconstrained optimization. *Mathematical Programming*, 180:451–488, 2020.
- Nilesh Tripuraneni, Mitchell Stern, Chi Jin, Jeffrey Regier, and Michael I Jordan. Stochastic cubic regularization for fast nonconvex optimization. *Advances in neural information processing systems*, 31, 2018.
- Yi Xu, Rong Jin, and Tianbao Yang. Neon+: Accelerated gradient methods for extracting negative curvature for non-convex optimization. *arXiv preprint arXiv:1712.01033*, 2017.
- Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *International Conference on Machine Learning*, pp. 11173–11182. PMLR, 2020.
- Qinzi Zhang and Ashok Cutkosky. Random scaling and momentum for non-smooth non-convex optimization. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 58780–58799, 2024.
- Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic nested variance reduction for nonconvex optimization. *Journal of machine learning research*, 21(103):1–63, 2020.

APPENDIX

A MORE DETAILS ON ONLINE-TO-NONCONVEX CONVERSION AND RELATED PROOFS

In this section, we elaborate on the details of the online-to-nonconvex conversion under the assumption that we have access to a deterministic oracle, with the goal of building intuition. In the next subsection, these results will be extended to the case where only a stochastic oracle is available.

As introduced in Section 2.1, to eliminate the randomness from the analysis, we define the gradient estimate $\mathbf{g}_n = \nabla F(\mathbf{w}_n)$ where $\mathbf{w}_n = \frac{1}{2}(\mathbf{x}_{n-1} + \mathbf{x}_n)$. This modification introduces an error in approximating $F(\mathbf{x}_n) - F(\mathbf{x}_{n-1})$ by $\mathbf{g}_n^\top \Delta_n$. However, as established in the next lemma, under Assumption 2.2 and for sufficiently small D , this error becomes negligible. The proof is available in (Jiang et al., 2025, Appendix A.1).

Lemma A.1. *Consider $\mathbf{x}_n = \mathbf{x}_{n-1} + \Delta_n$ where $\|\Delta_n\| \leq D$. Further, define $\mathbf{g}_n = \nabla F(\mathbf{w}_n)$ where $\mathbf{w}_n = \frac{1}{2}(\mathbf{x}_{n-1} + \mathbf{x}_n)$. If Assumption 2.2 holds, then $F(\mathbf{x}_{n-1}) - F(\mathbf{x}_n) \geq -\mathbf{g}_n^\top \Delta_n - \frac{L_2 D^3}{48}$.*

Lemma A.1 provides a formal connection between \mathbf{g}_n , Δ_n , and the function value decrease. While the choice $-\mathbf{g}_n$ yields the steepest local descent, as discussed in Section 2.1, the key challenge is that \mathbf{g}_n cannot be computed prior to selecting Δ_n , since it depends on \mathbf{x}_n , which itself requires Δ_n . This observation motivates an alternative interpretation: the process of choosing Δ_n can be cast as an instance of *online learning*, where each Δ_n serves as a decision variable, and the loss incurred at step n is given by the linear form $\mathbf{g}_n^\top \Delta_n$.

To rigorously relate the online learning perspective of choosing Δ_n to the objective of identifying a stationary point, we appeal to Lemma A.1. This result establishes that for any fixed comparator \mathbf{u} , after T iterations satisfies the following bound:

$$F(\mathbf{x}_T) - F(\mathbf{x}_0) \leq \sum_{n=1}^T \mathbf{g}_n^\top (\Delta_n - \mathbf{u}) + \sum_{n=1}^T \mathbf{g}_n^\top \mathbf{u} + \frac{TL_2 D^3}{48}.$$

Setting the arbitrary vector as $\mathbf{u} = -D(\sum_{n=1}^T \mathbf{g}_n) / \|\sum_{n=1}^T \mathbf{g}_n\|$, it can be shown:

$$\left\| \frac{1}{T} \sum_{n=1}^T \mathbf{g}_n \right\| \leq \frac{F(\mathbf{x}_0) - F(\mathbf{x}_T)}{DT} + \frac{1}{DT} \sum_{n=1}^T \mathbf{g}_n^\top (\Delta_n - \mathbf{u}) + \frac{L_2 D^2}{48}. \quad (10)$$

While the connection between the average gradient and the cumulative regret over T iterations has been established, our primary objective remains the identification of an ε -first-order stationary point. To bridge this gap, we now present a formal relationship linking the average of the gradients to the gradient evaluated at the mean iterate. The following lemma serves a similar purpose to (Cutkosky et al., 2023, Proposition 15).

Lemma A.2. *If $\mathbf{g}_n = \nabla F(\mathbf{w}_n)$ and Assumption 2.2 holds, then $\|\nabla F(\bar{\mathbf{w}})\| \leq \|\frac{1}{t} \sum_{n=1}^t \mathbf{g}_n\| + \frac{L_2}{2} t^2 D^2$, where $\bar{\mathbf{w}} = \frac{1}{t} \sum_{n=1}^t \mathbf{w}_n$.*

The proof of this lemma is provided in (Jiang et al., 2025, Appendix A.2). By combining this result with the expression in (10), we obtain a connection between the norm of the gradient at the averaged iterate and the regret bound appearing on the right-hand side. It is important to note that the error incurred when approximating the average of the gradients over t iterations by the gradient evaluated at the average iterate grows quadratically with the window length t . To control this approximation error, we reset the averaging process every T iterations. At the same time, we must account for the fact that certain terms in the upper bound scale inversely with T , requiring a careful choice of T to achieve the tightest possible convergence guarantees.

A.1 PROOF OF PROPOSITION 2.1

Note that when we have access to a stochastic oracle, the gradient becomes $\mathbf{g}_n = \nabla f(\mathbf{w}_n, \xi_n)$. Consequently, Lemma A.1 takes the following form:

$$F(\mathbf{x}_{n-1}) - F(\mathbf{x}_n) \geq -\nabla F(\mathbf{w}_n)^\top \Delta_n - \frac{L_2 D^3}{48}.$$

This expression can be further related to \mathbf{g}_n as follows:

$$F(\mathbf{x}_{n-1}) - F(\mathbf{x}_n) \geq -\langle \nabla F(\mathbf{w}_n) - \mathbf{g}_n, \Delta_n \rangle - \mathbf{g}_n^\top \Delta_n - \frac{L_2 D^3}{48}.$$

Taking expectations and using the fact that \mathbf{g}_n is an unbiased estimator of $\nabla F(\mathbf{w}_n)$, we get:

$$\mathbb{E}[F(\mathbf{x}_{n-1})] - \mathbb{E}[F(\mathbf{x}_n)] \geq -\mathbb{E}[\mathbf{g}_n^\top \Delta_n] - \frac{L_2 D^3}{48}.$$

By telescoping over T iterations and adding and subtracting $\sum_{n=1}^T \mathbf{g}_n^\top \mathbf{u}$:

$$\mathbb{E}[F(\mathbf{x}_T)] - F(\mathbf{x}_0) \leq \mathbb{E} \left[\sum_{n=1}^T \langle \mathbf{g}_n, \Delta_n - \mathbf{u} \rangle \right] + \mathbb{E} \left[\sum_{n=1}^T \mathbf{g}_n^\top \mathbf{u} \right] + \frac{TL_2 D^3}{48}.$$

Choosing the comparator vector as $\mathbf{u} = -D(\sum_{n=1}^T \mathbf{g}_n) / \|\sum_{n=1}^T \mathbf{g}_n\|$, we obtain the following bound

$$\mathbb{E} \left[\left\| \frac{1}{T} \sum_{n=1}^T \mathbf{g}_n \right\| \right] \leq \frac{F(\mathbf{x}_0) - \mathbb{E}[F(\mathbf{x}_T)]}{DT} + \frac{1}{DT} \mathbb{E} \left[\sum_{n=1}^T \mathbf{g}_n^\top (\Delta_n - \mathbf{u}) \right] + \frac{L_2 D^2}{48}. \quad (11)$$

Now consider the k -th episode ($k = 1, 2, \dots, K$) from $n = (k-1)T + 1$ to $n = kT$. By applying the inequality in (11), we obtain

$$\begin{aligned} \mathbb{E} \left[\left\| \frac{1}{T} \sum_{n=(k-1)T+1}^{kT} \mathbf{g}_n \right\| \right] &\leq \frac{\mathbb{E}[F(\mathbf{x}_{(k-1)T})] - \mathbb{E}[F(\mathbf{x}_{kT})]}{DT} + \frac{1}{DT} \mathbb{E} \left[\sum_{n=(k-1)T+1}^{kT} \mathbf{g}_n^\top (\Delta_n - \mathbf{u}^k) \right] \\ &\quad + \frac{L_2 D^2}{48}. \end{aligned} \quad (12)$$

Moreover, recall that $\bar{\mathbf{w}}^k = \frac{1}{T} \sum_{n=(k-1)T+1}^{kT} \mathbf{w}_n$ and it follows from Lemma A.2 that $\|\nabla F(\bar{\mathbf{w}}^k)\| \leq \left\| \frac{1}{T} \sum_{n=(k-1)T+1}^{kT} \nabla F(\mathbf{w}_n) \right\| + \frac{L_2}{2} T^2 D^2$. Recall that Lemma A.2 was derived under the deterministic setting where $\mathbf{g}_n = \nabla F(\mathbf{w}_n)$, but now $\mathbf{g}_n = \nabla f(\mathbf{w}_n, \xi)$. However, we can state:

$$\begin{aligned} \|\nabla F(\bar{\mathbf{w}}^k)\| &\leq \left\| \frac{1}{T} \sum_{n=(k-1)T+1}^{kT} \nabla F(\mathbf{w}_n) \right\| + \frac{L_2}{2} T^2 D^2 \\ &= \left\| \frac{1}{T} \sum_{n=(k-1)T+1}^{kT} (\nabla F(\mathbf{w}_n) - \mathbf{g}_n + \mathbf{g}_n) \right\| + \frac{L_2}{2} T^2 D^2 \\ &\leq \left\| \frac{1}{T} \sum_{n=(k-1)T+1}^{kT} (\nabla F(\mathbf{w}_n) - \mathbf{g}_n) \right\| + \left\| \frac{1}{T} \sum_{n=(k-1)T+1}^{kT} \mathbf{g}_n \right\| + \frac{L_2}{2} T^2 D^2. \end{aligned}$$

The final inequality follows from the triangle inequality. In expectation, we can state

$$\mathbb{E} [\|\nabla F(\bar{\mathbf{w}}^k)\|] \leq \frac{\sigma}{\sqrt{T}} + \mathbb{E} \left[\left\| \frac{1}{T} \sum_{n=(k-1)T+1}^{kT} \mathbf{g}_n \right\| \right] + \frac{L_2}{2} T^2 D^2. \quad (13)$$

This bound follows from applying Assumption 2.3, which allows us to establish

$$\mathbb{E} \left[\left\| \frac{1}{T} \sum_{n=(k-1)T+1}^{kT} \nabla F(\mathbf{w}_n) - \mathbf{g}_n \right\| \right] \leq \frac{\sigma}{\sqrt{T}}.$$

Using (12) together with (13) leads to

$$\begin{aligned} \mathbb{E} [\|\nabla F(\bar{\mathbf{w}}^k)\|] &\leq \frac{\mathbb{E}[F(\mathbf{x}_{(k-1)T})] - \mathbb{E}[F(\mathbf{x}_{kT})]}{DT} + \frac{1}{DT} \mathbb{E} \left[\sum_{n=(k-1)T+1}^{kT} \mathbf{g}_n^\top (\Delta_n - \mathbf{u}^k) \right] \\ &\quad + \frac{L_2 D^2}{48} + \frac{L_2}{2} T^2 D^2 + \frac{\sigma}{\sqrt{T}}. \end{aligned}$$

Summing over $k = 1$ to K and dividing by K , we obtain

$$\begin{aligned} \frac{1}{K} \sum_{k=1}^K \mathbb{E} [\|\nabla F(\bar{\mathbf{w}}^k)\|] &\leq \frac{F(\mathbf{x}_0) - \mathbb{E}[F(\mathbf{x}_M)]}{KDT} + \frac{1}{KDT} \mathbb{E} \left[\sum_{k=1}^K \sum_{n=(k-1)T+1}^{kT} \mathbf{g}_n^\top (\Delta_n - \mathbf{u}^k) \right] \\ &\quad + \frac{L_2 D^2}{48} + \frac{L_2}{2} T^2 D^2 + \frac{\sigma}{\sqrt{T}}. \end{aligned}$$

Finally, noting that $\mathbb{E}[F(\mathbf{x}_M)] \geq F^*$ completes the proof.

B PROOFS FOR ONLINE DOUBLY OPTIMISTIC GRADIENT WITH CONSTANT STEP SIZE

In this section, we provide the detailed proofs corresponding to Section 3, where the *Online Doubly Optimistic Gradient* method with a constant step size was introduced, along with its complexity analysis. While some of the proofs closely follow arguments from Jiang et al. (2025), we include them here for completeness.

B.1 PROOF OF LEMMA 3.1

To prepare for the proof of Lemma 3.1, we first prove an auxiliary result.

Lemma B.1. *Consider the update rule in (5). Then for any \mathbf{u} such that $\|\mathbf{u}\| \leq D$, it holds that*

$$\begin{aligned} \langle \mathbf{g}_{n+1}, \Delta_{n+1} - \mathbf{u} \rangle &\leq \frac{\|\Delta_n - \mathbf{u}\|^2}{2\eta} - \frac{\|\Delta_{n+1} - \mathbf{u}\|^2}{2\eta} + \langle \mathbf{g}_{n+1} - \mathbf{h}_{n+1}, \Delta_{n+1} - \mathbf{u} \rangle \\ &\quad - \langle \mathbf{g}_n - \mathbf{h}_n, \Delta_n - \mathbf{u} \rangle + \eta \|\mathbf{g}_n - \mathbf{h}_n\|^2 - \frac{1}{4\eta} \|\Delta_{n+1} - \Delta_n\|^2. \end{aligned}$$

Proof. By using the optimality condition, for any $\mathbf{u} \in \{\Delta \in \mathbb{R}^d : \|\Delta\| \leq D\}$, we have

$$\langle \Delta_{n+1} - \Delta_n + \eta \mathbf{h}_{n+1} + \eta(\mathbf{g}_n - \mathbf{h}_n), \mathbf{u} - \Delta_{n+1} \rangle \geq 0.$$

Rearranging terms, we obtain

$$\begin{aligned} \langle \mathbf{g}_{n+1}, \Delta_{n+1} - \mathbf{u} \rangle &\leq \langle \mathbf{g}_{n+1} - \mathbf{h}_{n+1}, \Delta_{n+1} - \mathbf{u} \rangle - \langle \mathbf{g}_n - \mathbf{h}_n, \Delta_{n+1} - \mathbf{u} \rangle \\ &\quad + \frac{1}{2\eta} \|\Delta_n - \mathbf{u}\|^2 - \frac{1}{2\eta} \|\Delta_{n+1} - \Delta_n\|^2 - \frac{1}{2\eta} \|\Delta_{n+1} - \mathbf{u}\|^2, \end{aligned}$$

where we have used the three-point equality $\langle \Delta_{n+1} - \Delta_n, \mathbf{u} - \Delta_{n+1} \rangle = \frac{1}{2} \|\Delta_n - \mathbf{u}\|^2 - \frac{1}{2} \|\Delta_{n+1} - \Delta_n\|^2 - \frac{1}{2} \|\Delta_{n+1} - \mathbf{u}\|^2$. Moreover, we have:

$$-\langle \mathbf{g}_n - \mathbf{h}_n, \Delta_{n+1} - \mathbf{u} \rangle = -\langle \mathbf{g}_n - \mathbf{h}_n, \Delta_n - \mathbf{u} \rangle + \langle \mathbf{g}_n - \mathbf{h}_n, \Delta_n - \Delta_{n+1} \rangle.$$

We can further bound the second term as follows:

$$\langle \mathbf{g}_n - \mathbf{h}_n, \Delta_n - \Delta_{n+1} \rangle \leq \|\mathbf{g}_n - \mathbf{h}_n\| \|\Delta_n - \Delta_{n+1}\| \leq \eta \|\mathbf{g}_n - \mathbf{h}_n\|^2 + \frac{1}{4\eta} \|\Delta_{n+1} - \Delta_n\|^2,$$

where the final inequality follows from the weighted Young's inequality. Putting all of the above inequalities together yields the desired bound. \square

Using Lemma B.1, we can bound the regret for each episode in the following lemma.

Lemma B.2. *Consider the optimistic update in (5). Then for $k = 1$ and any \mathbf{u}^1 such that $\|\mathbf{u}^1\| \leq D$, we have*

$$\sum_{n=1}^T \langle \mathbf{g}_n, \Delta_n - \mathbf{u}^1 \rangle \leq \frac{2D^2}{\eta} + \sum_{n=1}^T \eta \|\mathbf{g}_n - \mathbf{h}_n\|^2 - \sum_{n=2}^T \frac{1}{4\eta} \|\Delta_n - \Delta_{n-1}\|^2. \quad (14)$$

For $k \geq 2$ and any \mathbf{u}^k such that $\|\mathbf{u}^k\| \leq D$, we have

$$\begin{aligned} \sum_{n=(k-1)T+1}^{kT} \langle \mathbf{g}_n, \Delta_n - \mathbf{u}^k \rangle &\leq \frac{4D^2}{\eta} + \frac{\eta}{2} \|\mathbf{g}_{(k-1)T} - \mathbf{h}_{(k-1)T}\|^2 + \sum_{n=(k-1)T}^{kT} \eta \|\mathbf{g}_n - \mathbf{h}_n\|^2 \\ &\quad - \sum_{n=(k-1)T+1}^{kT} \frac{1}{4\eta} \|\Delta_n - \Delta_{n-1}\|^2. \end{aligned} \quad (15)$$

Proof. To prove (14), we set $\mathbf{u} = \mathbf{u}^1$ and sum the inequality in Lemma B.1 from $n = 1$ to $n = T - 1$:

$$\begin{aligned} \sum_{n=2}^T \langle \mathbf{g}_n, \Delta_n - \mathbf{u}^1 \rangle &\leq \frac{1}{2\eta} \|\Delta_1 - \mathbf{u}^1\|^2 - \frac{1}{2\eta} \|\Delta_T - \mathbf{u}^1\|^2 + \langle \mathbf{g}_T - \mathbf{h}_T, \Delta_T - \mathbf{u}^1 \rangle \\ &\quad - \langle \mathbf{g}_1 - \mathbf{h}_1, \Delta_1 - \mathbf{u}^1 \rangle + \sum_{n=1}^{T-1} \left(\eta \|\mathbf{g}_n - \mathbf{h}_n\|^2 - \frac{1}{4\eta} \|\Delta_{n+1} - \Delta_n\|^2 \right). \end{aligned}$$

Moreover, recall that $\Delta_1 = \arg \min_{\|\Delta\| \leq D} \langle \mathbf{h}_1, \Delta \rangle$. Thus, this implies that $\langle \mathbf{h}_1, \Delta_1 - \mathbf{u}^1 \rangle \leq 0$.

$$\langle \mathbf{g}_T - \mathbf{h}_T, \Delta_T - \mathbf{u}^1 \rangle \leq \|\mathbf{g}_T - \mathbf{h}_T\| \|\Delta_T - \mathbf{u}^1\| \leq \frac{\eta}{2} \|\mathbf{g}_T - \mathbf{h}_T\|^2 + \frac{1}{2\eta} \|\Delta_T - \mathbf{u}^1\|^2.$$

Combining all the inequalities, we obtain that

$$\sum_{n=1}^T \langle \mathbf{g}_n, \Delta_n - \mathbf{u}^1 \rangle \leq \frac{1}{2\eta} \|\Delta_1 - \mathbf{u}^1\|^2 + \frac{\eta}{2} \|\mathbf{g}_T - \mathbf{h}_T\|^2 + \sum_{n=1}^{T-1} \left(\eta \|\mathbf{g}_n - \mathbf{h}_n\|^2 - \frac{1}{4\eta} \|\Delta_{n+1} - \Delta_n\|^2 \right).$$

Using $\frac{\eta}{2} \|\mathbf{g}_T - \mathbf{h}_T\|^2 \leq \eta \|\mathbf{g}_T - \mathbf{h}_T\|^2$ and $\sum_{n=1}^{T-1} \frac{1}{4\eta} \|\Delta_{n+1} - \Delta_n\|^2 = \sum_{n=2}^T \frac{1}{4\eta} \|\Delta_n - \Delta_{n-1}\|^2$ lead to:

$$\sum_{n=1}^T \langle \mathbf{g}_n, \Delta_n - \mathbf{u}^1 \rangle \leq \frac{1}{2\eta} \|\Delta_1 - \mathbf{u}^1\|^2 + \sum_{n=1}^T \eta \|\mathbf{g}_n - \mathbf{h}_n\|^2 - \sum_{n=2}^T \frac{1}{4\eta} \|\Delta_n - \Delta_{n-1}\|^2.$$

Finally, since we have $\|\Delta_1\| = D$ and $\|\mathbf{u}^1\| \leq D$, this leads to (14).

Now we move to (15). To simplify the notation, we let $\Delta_t^k := \Delta_{(k-1)T+t}$, $\mathbf{g}_t^k := \mathbf{g}_{(k-1)T+t}$ and $\mathbf{h}_t^k = \mathbf{h}_{(k-1)T+t}$. Then by setting $\mathbf{u} = \mathbf{u}^k$ and summing the inequality in Lemma B.1 from $n = (k-1)T$ to $n = kT - 1$, we obtain:

$$\begin{aligned} \sum_{t=1}^T \langle \mathbf{g}_t^k, \Delta_t^k - \mathbf{u}^k \rangle &\leq \frac{\|\Delta_0^k - \mathbf{u}^k\|^2}{2\eta} - \frac{\|\Delta_T^k - \mathbf{u}^k\|^2}{2\eta} + \langle \mathbf{g}_T^k - \mathbf{h}_T^k, \Delta_T^k - \mathbf{u}^k \rangle \\ &\quad - \langle \mathbf{g}_T^{k-1} - \mathbf{h}_T^{k-1}, \Delta_T^{k-1} - \mathbf{u}^k \rangle + \sum_{n=(k-1)T}^{kT-1} \left(\eta \|\mathbf{g}_n - \mathbf{h}_n\|^2 - \frac{1}{4\eta} \|\Delta_{n+1} - \Delta_n\|^2 \right). \end{aligned}$$

Following a similar argument as in the proof of (14), we have the following inequalities:

- $\langle \mathbf{g}_T^k - \mathbf{h}_T^k, \Delta_T^k - \mathbf{u}^k \rangle \leq \frac{\eta}{2} \|\mathbf{g}_T^k - \mathbf{h}_T^k\|^2 + \frac{1}{2\eta} \|\Delta_T^k - \mathbf{u}^k\|^2.$
- $\langle \mathbf{g}_T^{k-1} - \mathbf{h}_T^{k-1}, \Delta_T^{k-1} - \mathbf{u}^k \rangle \leq \frac{\eta}{2} \|\mathbf{g}_T^{k-1} - \mathbf{h}_T^{k-1}\|^2 + \frac{1}{2\eta} \|\Delta_T^{k-1} - \mathbf{u}^k\|^2.$
- $\sum_{n=(k-1)T}^{kT-1} \frac{1}{4\eta} \|\Delta_{n+1} - \Delta_n\|^2 = \sum_{n=(k-1)T+1}^{kT} \frac{1}{4\eta} \|\Delta_n - \Delta_{n-1}\|^2.$

Thus, combining all the inequalities above, we obtain:

$$\begin{aligned}
\sum_{t=1}^T \langle \mathbf{g}_t^k, \Delta_t^k - \mathbf{u}^k \rangle &\leq \frac{1}{2\eta} \|\Delta_0^k - \mathbf{u}^k\|^2 + \frac{\eta}{2} \|\mathbf{g}_T^k - \mathbf{h}_T^k\|^2 + \frac{\eta}{2} \|\mathbf{g}_T^{k-1} - \mathbf{h}_T^{k-1}\|^2 + \frac{1}{2\eta} \|\Delta_T^{k-1} - \mathbf{u}^k\|^2 \\
&\quad + \sum_{n=(k-1)T}^{kT-1} \eta \|\mathbf{g}_n - \mathbf{h}_n\|^2 - \sum_{n=(k-1)T+1}^{kT} \frac{1}{4\eta} \|\Delta_n - \Delta_{n-1}\|^2 \\
&\leq \frac{\|\Delta_0^k - \mathbf{u}^k\|^2}{2\eta} + \frac{\eta \|\mathbf{g}_T^{k-1} - \mathbf{h}_T^{k-1}\|^2}{2} + \frac{\|\Delta_T^{k-1} - \mathbf{u}^k\|^2}{2\eta} + \sum_{n=(k-1)T}^{kT} \eta \|\mathbf{g}_n - \mathbf{h}_n\|^2 \\
&\quad - \sum_{n=(k-1)T+1}^{kT} \frac{1}{4\eta} \|\Delta_n - \Delta_{n-1}\|^2.
\end{aligned}$$

Finally, since we have $\|\Delta_0^k\| \leq D$, $\|\mathbf{u}\| \leq D$, and $\|\Delta_T^{k-1}\| \leq D$, we get $\frac{\|\Delta_0^k - \mathbf{u}^k\|^2}{2\eta} + \frac{\|\Delta_T^{k-1} - \mathbf{u}^k\|^2}{2\eta} \leq \frac{4D^2}{\eta}$. This completes the proof of (15). \square

Now we are ready to prove Lemma 3.1.

Proof of Lemma 3.1. Summing the inequality in (14) and the inequality in (15) for $k = 2, 3, \dots, K$ in Lemma B.2, we have:

$$\begin{aligned}
\text{Reg}_T(\mathbf{u}^1, \dots, \mathbf{u}^K) &= \sum_{k=1}^K \sum_{n=1}^T \langle \mathbf{g}_n^k, \Delta_n^k - \mathbf{u}^k \rangle \\
&\leq \frac{4KD^2}{\eta} + \sum_{k=1}^K \frac{\eta}{2} \|\mathbf{g}_T^{k-1} - \mathbf{h}_T^{k-1}\|^2 + \sum_{n=1}^{KT} \eta \|\mathbf{g}_n - \mathbf{h}_n\|^2 - \sum_{n=2}^{KT} \frac{1}{4\eta} \|\Delta_n - \Delta_{n-1}\|^2 \\
&\quad + \sum_{k=1}^K \eta \|\mathbf{g}_T^{k-1} - \mathbf{h}_T^{k-1}\|^2 \\
&\leq \frac{4KD^2}{\eta} + \sum_{k=1}^K \frac{3\eta}{2} \|\mathbf{g}_T^{k-1} - \mathbf{h}_T^{k-1}\|^2 + \sum_{n=1}^{KT} \eta \|\mathbf{g}_n - \mathbf{h}_n\|^2 - \sum_{n=2}^{KT} \frac{1}{4\eta} \|\Delta_n - \Delta_{n-1}\|^2 \\
&\leq \frac{4KD^2}{\eta} + \frac{5\eta}{2} \sum_{n=1}^{KT} \|\mathbf{g}_n - \mathbf{h}_n\|^2 - \sum_{n=2}^{KT} \frac{1}{4\eta} \|\Delta_n - \Delta_{n-1}\|^2.
\end{aligned}$$

Where in the last inequality we use the fact that $\sum_{k=1}^K \frac{3\eta}{2} \|\mathbf{g}_T^{k-1} - \mathbf{h}_T^{k-1}\|^2 \leq \sum_{n=1}^{kT} \frac{3\eta}{2} \|\mathbf{g}_n - \mathbf{h}_n\|^2$. This completes the proof. \square

B.2 PROOF OF LEMMA 3.2

To prepare for the proof of Lemma 3.2, we first prove an auxiliary result.

Lemma B.3. *Let \mathbf{h}_n be defined as in (6). Then the following bound holds*

$$\mathbb{E} [\|\mathbf{g}_n - \mathbf{h}_n\|^2] \leq 6\sigma^2 + \frac{3L_1^2}{4} \mathbb{E} \|\Delta_n - \Delta_{n-1}\|^2.$$

Proof. To bound $\mathbb{E} \|\mathbf{g}_n - \mathbf{h}_n\|^2$, we begin by recalling the definitions of the respective vectors: $\mathbf{g}_n = \nabla f(\mathbf{w}_n; \xi_n)$ and $\mathbf{h}_n = \nabla f(\mathbf{z}_{n-1}; \xi_{n-1})$. Thus,

$$\begin{aligned} \mathbb{E} [\|\mathbf{g}_n - \mathbf{h}_n\|^2] &= \mathbb{E} \|\nabla f(\mathbf{w}_n; \xi_n) - \nabla f(\mathbf{z}_{n-1}; \xi_{n-1})\|^2 \\ &= \mathbb{E} \|\nabla f(\mathbf{w}_n; \xi_n) - \nabla f(\mathbf{z}_{n-1}; \xi_{n-1}) \\ &\quad + \nabla F(\mathbf{z}_{n-1}) - \nabla F(\mathbf{z}_{n-1}) \\ &\quad + \nabla F(\mathbf{w}_n) - \nabla F(\mathbf{w}_n)\|^2, \end{aligned}$$

where the last equality comes from adding and subtracting its deterministic counterparts. Applying Young's inequality

$$\begin{aligned} \mathbb{E} [\|\mathbf{g}_n - \mathbf{h}_n\|^2] &\leq 3 \mathbb{E} [\|\nabla f(\mathbf{z}_{n-1}; \xi_{n-1}) - \nabla F(\mathbf{z}_{n-1})\|^2] \\ &\quad + 3 \mathbb{E} [\|\nabla f(\mathbf{w}_n; \xi_n) - \nabla F(\mathbf{w}_n)\|^2] \\ &\quad + 3 \mathbb{E} [\|\nabla F(\mathbf{w}_n) - \nabla F(\mathbf{z}_{n-1})\|^2] \\ &\leq 6\sigma^2 + \frac{3L_1^2}{4} \mathbb{E} \|\Delta_n - \Delta_{n-1}\|^2. \end{aligned}$$

The last inequality follows from Assumption 2.3 that allows us to bound the variance of the first two terms, and the last term comes from leveraging Assumption 2.1, $\|\nabla F(\mathbf{x}_{n-1} + \frac{1}{2}\Delta_n) - \nabla F(\mathbf{x}_{n-1} + \frac{1}{2}\Delta_{n-1})\|^2 \leq L_1^2 \|\frac{1}{2}\Delta_n - \frac{1}{2}\Delta_{n-1}\|^2$. \square

To prove Lemma 3.2, we first analyze the case $k \geq 2$ and subsequently show that the bound also holds for $k = 1$. Our analysis focuses on the stochastic setting. However, by Lemma B.2, inequality (15) is valid in both the deterministic and stochastic cases. Therefore, we have:

$$\begin{aligned} \sum_{n=(k-1)T+1}^{kT} \langle \mathbf{g}_n, \Delta_n - \mathbf{u}^k \rangle &\leq \frac{4D^2}{\eta} + \frac{\eta}{2} \|\mathbf{g}_{(k-1)T} - \mathbf{h}_{(k-1)T}\|^2 + \sum_{n=(k-1)T}^{kT} \eta \|\mathbf{g}_n - \mathbf{h}_n\|^2 \\ &\quad - \sum_{n=(k-1)T+1}^{kT} \frac{1}{4\eta} \|\Delta_n - \Delta_{n-1}\|^2. \end{aligned}$$

Taking expectations on both sides yields

$$\begin{aligned} \sum_{n=(k-1)T+1}^{kT} \mathbb{E} \langle \mathbf{g}_n, \Delta_n - \mathbf{u}^k \rangle &\leq \frac{4D^2}{\eta} + \frac{\eta}{2} \mathbb{E} \|\mathbf{g}_{(k-1)T} - \mathbf{h}_{(k-1)T}\|^2 + \sum_{n=(k-1)T}^{kT} \eta \mathbb{E} \|\mathbf{g}_n - \mathbf{h}_n\|^2 \\ &\quad - \sum_{n=(k-1)T+1}^{kT} \frac{1}{4\eta} \mathbb{E} \|\Delta_n - \Delta_{n-1}\|^2. \end{aligned} \tag{16}$$

Leveraging Lemma B.3

$$\begin{aligned} \sum_{n=(k-1)T+1}^{kT} \mathbb{E} \langle \mathbf{g}_n, \Delta_n - \mathbf{u}^k \rangle &\leq \frac{4D^2}{\eta} + \frac{\eta}{2} \left(6\sigma^2 + \frac{3L_1^2}{4} \mathbb{E} \|\Delta_{(k-1)T} - \Delta_{(k-1)T-1}\|^2 \right) \\ &\quad + \sum_{n=(k-1)T}^{kT} \eta \left(6\sigma^2 + \frac{3L_1^2}{4} \mathbb{E} \|\Delta_n - \Delta_{n-1}\|^2 \right) \\ &\quad - \sum_{n=(k-1)T+1}^{kT} \frac{1}{4\eta} \mathbb{E} \|\Delta_n - \Delta_{n-1}\|^2. \end{aligned}$$

Rearranging the terms leads to

$$\begin{aligned} \sum_{n=(k-1)T+1}^{kT} \mathbb{E} \langle \mathbf{g}_n, \Delta_n - \mathbf{u}^k \rangle &\leq \frac{4D^2}{\eta} + 3\sigma^2\eta + \frac{9L_1^2\eta}{8} \mathbb{E} \|\Delta_{(k-1)T} - \Delta_{(k-1)T-1}\|^2 \\ &+ \sum_{n=(k-1)T}^{kT} 6\sigma^2\eta + \sum_{n=(k-1)T+1}^{kT} \left(\frac{3L_1^2\eta}{4} - \frac{1}{4\eta} \right) \mathbb{E} \|\Delta_n - \Delta_{n-1}\|^2. \end{aligned}$$

Note that as long as $\eta \leq \frac{1}{\sqrt{3}L_1}$, it holds that $\sum_{n=(k-1)T+1}^{kT} \left(\frac{3L_1^2\eta}{4} - \frac{1}{4\eta} \right) \mathbb{E} \|\Delta_n - \Delta_{n-1}\|^2 \leq 0$. Finally, since $\|\Delta_n\| \leq D$ for all n , we have:

$$\begin{aligned} \sum_{n=(k-1)T+1}^{kT} \mathbb{E} \langle \mathbf{g}_n, \Delta_n - \mathbf{u}^k \rangle &\leq \frac{4D^2}{\eta} + \frac{9L_1^2D^2\eta}{2} + \left(T + \frac{3}{2} \right) 6\sigma^2\eta \\ &\leq \frac{4D^2}{\eta} + \frac{9L_1^2D^2\eta}{2} + 18\sigma^2T\eta, \end{aligned} \quad (17)$$

where the last inequality follows from the fact that $T \geq 1$, completing the proof.

For completeness, we analyze the case $k = 1$. As will be shown, the upper bound (17) also holds in this case. From Lemma B.2, inequality (14) gives

$$\sum_{n=1}^T \langle \mathbf{g}_n, \Delta_n - \mathbf{u}^1 \rangle \leq \frac{2D^2}{\eta} + \sum_{n=1}^T \eta \|\mathbf{g}_n - \mathbf{h}_n\|^2 - \sum_{n=2}^T \frac{1}{4\eta} \|\Delta_n - \Delta_{n-1}\|^2.$$

Taking expectations on both sides yields

$$\sum_{n=1}^T \mathbb{E} \langle \mathbf{g}_n, \Delta_n - \mathbf{u}^1 \rangle \leq \frac{2D^2}{\eta} + \sum_{n=1}^T \eta \mathbb{E} \|\mathbf{g}_n - \mathbf{h}_n\|^2 - \sum_{n=2}^T \frac{1}{4\eta} \mathbb{E} \|\Delta_n - \Delta_{n-1}\|^2. \quad (18)$$

Leveraging Lemma B.3

$$\sum_{n=1}^T \mathbb{E} \langle \mathbf{g}_n, \Delta_n - \mathbf{u}^1 \rangle \leq \frac{2D^2}{\eta} + \sum_{n=1}^T \eta \left(6\sigma^2 + \frac{3L_1^2}{4} \mathbb{E} \|\Delta_n - \Delta_{n-1}\|^2 \right) - \sum_{n=2}^T \frac{1}{4\eta} \mathbb{E} \|\Delta_n - \Delta_{n-1}\|^2.$$

Rearranging terms (with the convention $\Delta_0 = 0$) leads to

$$\sum_{n=1}^T \mathbb{E} \langle \mathbf{g}_n, \Delta_n - \mathbf{u}^1 \rangle \leq \frac{2D^2}{\eta} + \sum_{n=1}^T 6\sigma^2\eta + \frac{3L_1^2\eta}{4} \mathbb{E} \|\Delta_1\|^2 + \sum_{n=2}^T \left(\frac{3L_1^2\eta}{4} - \frac{1}{4\eta} \right) \mathbb{E} \|\Delta_n - \Delta_{n-1}\|^2.$$

Note that as long as $\eta \leq \frac{1}{\sqrt{3}L_1}$, it holds that $\sum_{n=(k-1)T+2}^{kT} \left(\frac{3L_1^2\eta}{4} - \frac{1}{4\eta} \right) \mathbb{E} \|\Delta_n - \Delta_{n-1}\|^2 \leq 0$. Finally, since $\|\Delta_n\| \leq D$ for all n , we have:

$$\begin{aligned} \sum_{n=(k-1)T+1}^{kT} \mathbb{E} \langle \mathbf{g}_n, \Delta_n - \mathbf{u}^k \rangle &\leq \frac{2D^2}{\eta} + \frac{3L_1^2D^2\eta}{4} + 6T\sigma^2\eta \\ &\leq \frac{4D^2}{\eta} + \frac{9L_1^2D^2\eta}{2} + 18\sigma^2T\eta, \end{aligned}$$

where the last inequality simply enlarges constants. This confirms that (17) also holds when $k = 1$.

B.3 PROOF OF THEOREM 3.3

According to Proposition 2.1, attaining the final convergence rate necessitates bounding the K -shifting regret. We obtain the following result as a direct consequence of Lemma 3.2

$$\mathbb{E} [\text{Reg}_T(\mathbf{u}^1, \dots, \mathbf{u}^K)] \leq \frac{4KD^2}{\eta} + \frac{9L_1^2\eta KD^2}{2} + 18\eta KT\sigma^2.$$

Integrating the aforementioned regret bound with Proposition 2.1, we arrive at

$$\mathbb{E} \left[\frac{1}{K} \sum_{k=1}^K \|\nabla F(\bar{\mathbf{w}}^k)\| \right] \leq \frac{F(\mathbf{x}_0) - F^*}{DKT} + \frac{4D}{T\eta} + \frac{9L_1^2\eta D}{2T} + \frac{18\eta}{D}\sigma^2 + \frac{L_2}{48}D^2 + \frac{L_2}{2}T^2D^2 + \frac{\sigma}{\sqrt{T}}.$$

Rearranging terms

$$\mathbb{E} \left[\frac{1}{K} \sum_{k=1}^K \|\nabla F(\bar{\mathbf{w}}^k)\| \right] \leq \frac{F(\mathbf{x}_0) - F^*}{DKT} + \frac{4D}{T\eta} + \left(\frac{9L_1^2D}{2T} + \frac{18}{D}\sigma^2 \right) \eta + L_2T^2D^2 + \frac{\sigma}{\sqrt{T}}. \quad (19)$$

From the upper bound in (19), we must select the hyperparameters η, D, K, T , subject to the constraint $KT \leq M$, where M denotes the total iteration budget and $K, T \in \mathbb{N}$. To obtain the tightest possible bound, we will first choose η , then determine T and K , and finally set D .

To proceed, we balance the two η -dependent terms to determine the optimal η , and scale it appropriately so that the condition $\eta \leq \frac{1}{\sqrt{3}L_1}$, required by Lemma 3.2, is satisfied. This yields the choice $\eta = \frac{1}{\sqrt{3L_1^2 + \frac{12T}{D^2}\sigma^2}}$. Substituting this expression into the bound gives

$$\mathbb{E} \left[\frac{1}{K} \sum_{k=1}^K \|\nabla F(\bar{\mathbf{w}}^k)\| \right] \leq \frac{F(\mathbf{x}_0) - F^*}{DKT} + \frac{11D}{2T} \sqrt{3L_1^2 + \frac{12T}{D^2}\sigma^2} + L_2T^2D^2 + \frac{\sigma}{\sqrt{T}}.$$

Applying the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$, and simplifying yields

$$\mathbb{E} \left[\frac{1}{K} \sum_{k=1}^K \|\nabla F(\bar{\mathbf{w}}^k)\| \right] \leq \frac{F(\mathbf{x}_0) - F^*}{DKT} + \frac{10L_1D}{T} + \frac{20\sigma}{\sqrt{T}} + L_2T^2D^2. \quad (20)$$

Note that the bound involves four terms dependent on T . If we set $T = \min \left(\max \left(\left\lceil \left(\frac{20\sigma}{L_2D^2} \right)^{\frac{2}{5}} \right\rceil, \left\lceil \left(\frac{10L_1}{L_2D} \right)^{\frac{1}{3}} \right\rceil \right), \frac{M}{2} \right)$ and $K = \lfloor \frac{M}{T} \rfloor$, we can bound the first three terms of the inequality (20) as follows:

- $KT \geq (\frac{M}{T} - 1)T \geq M - T \geq M/2$ which implies $\frac{F(\mathbf{x}_0) - F^*}{DKT} \leq 2\frac{F(\mathbf{x}_0) - F^*}{DM}$;
- $\frac{10L_1D}{T} \leq \max \left(10^{\frac{2}{3}}L_1^{\frac{2}{3}}L_2^{\frac{1}{3}}D^{\frac{4}{3}}, \frac{20L_1D}{M} \right)$;
- $\frac{20\sigma}{\sqrt{T}} \leq \max \left(20^{\frac{4}{5}}L_2^{\frac{1}{5}}D^{\frac{2}{5}}\sigma^{\frac{4}{5}}, \frac{20\sqrt{2}\sigma}{\sqrt{M}} \right)$;

The max terms arise from the fact that the second and third terms in (20) are inversely proportional to T . Since T is clipped by $M/2$, these terms become inversely proportional to $\min \left(D^{-\frac{1}{3}}, \frac{M}{2} \right)$ and $\sqrt{\min \left(\left(\frac{\sigma}{D^2} \right)^{\frac{2}{5}}, \frac{M}{2} \right)}$, respectively.

We upper bound $L_2T^2D^2$ as follows:

$$\begin{aligned} L_2T^2D^2 &\leq L_2 \max \left(\left\lceil \left(\frac{20\sigma}{L_2D^2} \right)^{\frac{2}{5}} \right\rceil, \left\lceil \left(\frac{10L_1}{L_2D} \right)^{\frac{1}{3}} \right\rceil \right)^2 D^2 \\ &\leq L_2 \left\lceil \left(\frac{20\sigma}{L_2D^2} \right)^{\frac{2}{5}} \right\rceil^2 D^2 + L_2 \left\lceil \left(\frac{10L_1}{L_2D} \right)^{\frac{1}{3}} \right\rceil^2 D^2. \end{aligned}$$

Using the fact that $\lceil x \rceil^2 \leq (x+1)^2 \leq 2x^2 + 2$ for any $x \geq 0$, we obtain

$$\begin{aligned}
L_2 T^2 D^2 &\leq L_2 \left(2 \left(\frac{20\sigma}{L_2 D^2} \right)^{\frac{4}{5}} + 2 \right) D^2 + L_2 \left(2 \left(\frac{10L_1}{L_2 D} \right)^{\frac{2}{3}} + 2 \right) D^2 \\
&= 2L_2 \left(\frac{20\sigma}{L_2 D^2} \right)^{\frac{4}{5}} D^2 + 2L_2 \left(\frac{10L_1}{L_2 D} \right)^{\frac{2}{3}} D^2 + 4L_2 D^2 \\
&\leq 22L_2^{\frac{1}{5}} \sigma^{\frac{4}{5}} D^{\frac{2}{5}} + 10L_1^{\frac{2}{3}} L_2^{\frac{1}{3}} D^{\frac{4}{3}} + 4L_2 D^2
\end{aligned}$$

Substituting the above inequalities into the bound in (20) yields

$$\begin{aligned}
\mathbb{E} \left[\frac{1}{K} \sum_{k=1}^K \|\nabla F(\bar{\mathbf{w}}^k)\| \right] &\leq 2 \frac{F(\mathbf{x}_0) - F^*}{DM} + \max \left(10^{\frac{2}{3}} L_1^{\frac{2}{3}} L_2^{\frac{1}{3}} D^{\frac{4}{3}}, \frac{20L_1 D}{M} \right) \\
&\quad + \max \left(20^{\frac{4}{5}} L_2^{\frac{1}{5}} D^{\frac{2}{5}} \sigma^{\frac{4}{5}}, \frac{20\sqrt{2}\sigma}{\sqrt{M}} \right) + 22L_2^{\frac{1}{5}} \sigma^{\frac{4}{5}} D^{\frac{2}{5}} + 10L_1^{\frac{2}{3}} L_2^{\frac{1}{3}} D^{\frac{4}{3}} + 4L_2 D^2.
\end{aligned}$$

Using $\max(a, b) \leq a + b$, and after rounding constants, we get:

$$\mathbb{E} \left[\frac{1}{K} \sum_{k=1}^K \|\nabla F(\bar{\mathbf{w}}^k)\| \right] \leq 2 \frac{F(\mathbf{x}_0) - F^*}{DM} + \frac{20L_1 D}{M} + \frac{20\sqrt{2}\sigma}{\sqrt{M}} + 33L_2^{\frac{1}{5}} \sigma^{\frac{4}{5}} D^{\frac{2}{5}} + 15L_1^{\frac{2}{3}} L_2^{\frac{1}{3}} D^{\frac{4}{3}} + 4L_2 D^2.$$

If we set $D = \min \left\{ \left(\frac{2(F(\mathbf{x}_0) - F^*)}{33L_2^{\frac{1}{5}} \sigma^{\frac{4}{5}} M} \right)^{5/7}, \left(\frac{2(F(\mathbf{x}_0) - F^*)}{15ML_1^{\frac{2}{3}} L_2^{\frac{1}{3}}} \right)^{3/7} \right\}$. Given this selection of D , the resulting upper bounds are as follows:

- $33L_2^{\frac{1}{5}} D^{\frac{2}{5}} \sigma^{\frac{4}{5}} \leq \frac{33^{\frac{5}{7}} 2^{\frac{2}{7}} (F(\mathbf{x}_0) - F^*)^{\frac{2}{7}}}{M^{\frac{2}{7}}} L_2^{\frac{1}{7}} \sigma^{\frac{4}{7}} \leq \frac{15(F(\mathbf{x}_0) - F^*)^{\frac{2}{7}}}{M^{\frac{2}{7}}} L_2^{\frac{1}{7}} \sigma^{\frac{4}{7}};$
- $15L_1^{\frac{2}{3}} L_2^{\frac{1}{3}} D^{\frac{4}{3}} \leq \frac{15^{\frac{3}{7}} 2^{\frac{4}{7}} (F(\mathbf{x}_0) - F^*)^{\frac{4}{7}}}{M^{\frac{4}{7}}} \leq \frac{5(F(\mathbf{x}_0) - F^*)^{\frac{4}{7}}}{M^{\frac{4}{7}}} L_1^{\frac{2}{7}} L_2^{\frac{1}{7}};$
- $2 \frac{F(\mathbf{x}_0) - F^*}{DM} \leq 2 \frac{7.5^{\frac{3}{7}} (F(\mathbf{x}_0) - F^*)^{\frac{4}{7}}}{M^{\frac{4}{7}}} + 2 \frac{16.5^{\frac{5}{7}} (F(\mathbf{x}_0) - F^*)^{\frac{2}{7}}}{M^{\frac{2}{7}}} L_2^{\frac{1}{7}} \sigma^{\frac{4}{7}} \leq \frac{5(F(\mathbf{x}_0) - F^*)^{\frac{4}{7}}}{M^{\frac{4}{7}}} L_1^{\frac{2}{7}} L_2^{\frac{1}{7}} + \frac{15(F(\mathbf{x}_0) - F^*)^{\frac{2}{7}}}{M^{\frac{2}{7}}} L_2^{\frac{1}{7}} \sigma^{\frac{4}{7}};$
- $\frac{20L_1 D}{M} \leq \frac{20L_1^{\frac{5}{7}} (F(\mathbf{x}_0) - F^*)^{\frac{3}{7}}}{7.5^{\frac{3}{7}} L_2^{\frac{1}{7}} M^{\frac{10}{7}}} \leq 9 \frac{L_1^{\frac{5}{7}} (F(\mathbf{x}_0) - F^*)^{\frac{3}{7}}}{L_2^{\frac{1}{7}} M^{\frac{10}{7}}};$
- $4L_2 D^2 \leq \frac{4L_2^{\frac{5}{7}} (F(\mathbf{x}_0) - F^*)^{\frac{6}{7}}}{7.5^{\frac{6}{7}} M^{\frac{6}{7}} L_1^{\frac{4}{7}}} \leq \frac{L_2^{\frac{5}{7}} (F(\mathbf{x}_0) - F^*)^{\frac{6}{7}}}{L_1^{\frac{4}{7}} M^{\frac{6}{7}}}.$

Substituting these inequalities into the bound yields

$$\begin{aligned}
\mathbb{E} \left[\frac{1}{K} \sum_{k=1}^K \|\nabla F(\bar{\mathbf{w}}^k)\| \right] &\leq \frac{10(F(\mathbf{x}_0) - F^*)^{\frac{4}{7}}}{M^{\frac{4}{7}}} L_1^{\frac{2}{7}} L_2^{\frac{1}{7}} + \frac{30(F(\mathbf{x}_0) - F^*)^{\frac{2}{7}}}{M^{\frac{2}{7}}} L_2^{\frac{1}{7}} \sigma^{\frac{4}{7}} \\
&\quad + 9 \frac{L_1^{\frac{5}{7}} (F(\mathbf{x}_0) - F^*)^{\frac{3}{7}}}{L_2^{\frac{1}{7}} M^{\frac{10}{7}}} + \frac{L_2^{\frac{5}{7}} (F(\mathbf{x}_0) - F^*)^{\frac{6}{7}}}{L_1^{\frac{4}{7}} M^{\frac{6}{7}}} + \frac{20\sqrt{2}\sigma}{\sqrt{M}}
\end{aligned}$$

This final bound confirms the convergence rate stated in the main theorem, as the last three terms are of higher order and thus do not influence the rate.

C PROOFS FOR ONLINE DOUBLY OPTIMISTIC GRADIENT METHOD WITH ADAPTIVE STEP SIZE

In this section, we provide detailed proofs corresponding to Section 4, where the *Online Doubly Optimistic Gradient* method with an adaptive step size was introduced, along with its complexity analysis. The section is organized into several subsections, presenting useful inequalities and lemmas that build up to the proofs of the two main results from Section 4, namely Lemma 4.1 and Theorem 4.2.

C.1 HELPFUL INEQUALITIES

Lemma C.1. *Let $a_i \geq 0 \forall i$, and that there exists at least one index $j \in [1, T]$ with $a_j > 0$. Then the following inequality holds*

$$\sqrt{\sum_{i=1}^T a_i} \leq \sum_{i=1}^T \frac{a_i}{\sqrt{\sum_{j=1}^i a_j}} \leq 2\sqrt{\sum_{i=1}^T a_i}.$$

Lemma C.1 appears in Kavis et al. (2019); Antonakopoulos et al. (2022); Levy et al. (2018); for the sake of completeness, we additionally provide its proof here.

Proof. We prove the lower and upper bounds separately.

Lower bound: We observe that

$$\sum_{i=1}^T a_i = \sum_{i=1}^T a_i \left(\frac{\sqrt{\sum_{j=1}^i a_j}}{\sqrt{\sum_{j=1}^i a_j}} \right) \leq \sum_{i=1}^T a_i \left(\frac{\sqrt{\sum_{j=1}^T a_j}}{\sqrt{\sum_{j=1}^i a_j}} \right),$$

where the second inequality follows from the fact that $a_i \geq 0 \forall i$. Dividing both sides by $\sqrt{\sum_{j=1}^T a_j}$ gives the desired bound.

Upper bound: Define $S_j = \sum_{i=1}^j a_i$, with the convention that $S_0 = 0$. For each i , it holds that

$$\frac{a_i}{2\sqrt{S_i}} \leq \frac{a_i}{\sqrt{S_i} + \sqrt{S_{i-1}}} = \frac{S_i - S_{i-1}}{\sqrt{S_i} + \sqrt{S_{i-1}}} = \sqrt{S_i} - \sqrt{S_{i-1}},$$

where the first inequality uses $S_{i-1} \leq S_i$, the first equality uses $a_i = S_i - S_{i-1}$, and the last equality follows from the difference of squares. Summing both sides of the inequality from $i = 1$ to $i = T$, yields:

$$\sum_{i=1}^T \frac{a_i}{2\sqrt{S_i}} \leq \sum_{i=1}^T \sqrt{S_i} - \sqrt{S_{i-1}} = \sqrt{S_T},$$

where the first equality follows from telescoping the series, and note that by definition $S_0 = 0$. Multiplying both sides by 2 gives the desired inequality. \square

Lemma C.2. *Let $a, b \in \mathbb{R}$, and define $M = \min\{(a+b)^2, a^2\}$. Then the following inequality holds:*

$$(a+b)^2 \leq 2M + 2b^2.$$

Proof. We have both $(a+b)^2 \leq 2a^2 + 2b^2$ and $(a+b)^2 \leq 2(a+b)^2 + 2b^2$. Combining these two gives us the desired inequality. \square

C.2 PROOF OF LEMMA 4.1

Prior to proving Lemma 4.1, we state the following proposition, which is an immediate consequence of Lemma B.1.

Proposition C.3. *Consider the update in (5), with an adaptive step size η_n . Then for any \mathbf{u} such that $\|\mathbf{u}\| \leq D$, it holds that*

$$\begin{aligned} \langle \mathbf{g}_{n+1}, \Delta_{n+1} - \mathbf{u} \rangle &\leq \frac{1}{2\eta_n} \|\Delta_n - \mathbf{u}\|^2 - \frac{1}{2\eta_n} \|\Delta_{n+1} - \mathbf{u}\|^2 + \langle \mathbf{g}_{n+1} - \mathbf{h}_{n+1}, \Delta_{n+1} - \mathbf{u} \rangle \\ &\quad - \langle \mathbf{g}_n - \mathbf{h}_n, \Delta_n - \mathbf{u} \rangle + \eta_n \|\mathbf{g}_n - \mathbf{h}_n\|^2 - \frac{1}{4\eta_n} \|\Delta_{n+1} - \Delta_n\|^2. \end{aligned}$$

Remark C.1. To simplify the notation in the remainder of the proof, we define $\bar{\mathbf{g}}_n := \nabla F(\mathbf{w}_n)$ and $\bar{\mathbf{h}}_n := \nabla F(\mathbf{z}_{n-1})$.

We are now prepared to present the proof of Lemma 4.1.

Proof of Lemma 4.1. The first step in proving Lemma 4.1 is to bound the regret within the episode. Analogously to Lemma B.2, we can formulate the following lemma for the first episode using an adaptive step size.

Lemma C.4. *Consider the optimistic update in (5) with an adaptive step size η_n . Then for $k = 1$ and any \mathbf{u}^1 such that $\|\mathbf{u}^1\| \leq D$, we have*

$$\sum_{n=1}^T \langle \mathbf{g}_n, \Delta_n - \mathbf{u}^1 \rangle \leq \frac{3D^2}{\eta_T} + \sum_{n=1}^T \left(\eta_n \|\mathbf{g}_n - \mathbf{h}_n\|^2 - \frac{1}{4\eta_n} \|\Delta_n - \Delta_{n-1}\|^2 \right).$$

The proof of this lemma is provided in Appendix C.3. Furthermore, the following lemma allows us to upper bound the sum $\sum_{n=1}^T \eta_n \|\mathbf{g}_n - \mathbf{h}_n\|^2$ in terms of $\sum_{n=1}^T \|\mathbf{g}_n - \mathbf{h}_n\|^2$, effectively decoupling the product of random variables. The proof can be found in Appendix C.3

Lemma C.5. *Using the definition of η_n from (8), we can derive the following bound*

$$\frac{3D}{\eta_T} + \sum_{n=1}^T \eta_n \|\mathbf{g}_n - \mathbf{h}_n\|^2 \leq 2\sqrt{2} \left(\frac{3D}{2\gamma} + \gamma D \right) \sqrt{\alpha + \sum_{n=1}^T \|\mathbf{g}_n - \mathbf{h}_n\|^2}$$

In addition, define $\mu_n = \min(\|\mathbf{g}_n - \mathbf{h}_n\|^2, \|\bar{\mathbf{g}}_n - \bar{\mathbf{h}}_n\|^2)$. Applying Lemma C.2, we obtain the inequality $\|\mathbf{g}_n - \mathbf{h}_n\|^2 \leq 2\mu_n + 4\|\mathbf{g}_n - \bar{\mathbf{g}}_n\|^2 + 4\|\mathbf{h}_n - \bar{\mathbf{h}}_n\|^2$. This allows us to upper bound the right-hand side of the expression as follows

$$\sqrt{\alpha + \sum_{n=1}^T \|\mathbf{g}_n - \mathbf{h}_n\|^2} \leq 2\sqrt{\sum_{n=1}^T (\|\mathbf{g}_n - \bar{\mathbf{g}}_n\|^2 + \|\mathbf{h}_n - \bar{\mathbf{h}}_n\|^2)} + \sqrt{2} \sqrt{\alpha + \sum_{n=1}^T \mu_n}.$$

This effectively decouples the noise in $\|\mathbf{g}_n - \mathbf{h}_n\|$ into separate, more manageable components. We now focus on bounding the expectation of the first term in the preceding inequality. By applying Jensen's inequality and invoking Assumption 2.3, we obtain: $\mathbb{E}\left[2\sqrt{\sum_{n=1}^T (\|\mathbf{g}_n - \bar{\mathbf{g}}_n\|^2 + \|\mathbf{h}_n - \bar{\mathbf{h}}_n\|^2)}\right] \leq 2\sqrt{\sum_{n=1}^T \mathbb{E}[\|\mathbf{g}_n - \bar{\mathbf{g}}_n\|^2 + \|\mathbf{h}_n - \bar{\mathbf{h}}_n\|^2]} \leq 2\sqrt{2}\sigma\sqrt{T}$. Therefore, it only remains to upper bound

$$4 \left(\frac{3D}{2\gamma} + \gamma D \right) \sqrt{\alpha + \sum_{n=1}^T \mu_n - \sum_{n=1}^T \frac{1}{4\eta_n} \|\Delta_n - \Delta_{n-1}\|^2}. \quad (21)$$

Note that by the definition of η_n in (8), we have $\eta_n \leq \frac{\gamma D}{\sqrt{\alpha + \sum_{i=1}^n \mu_i}}$. Also, by the definition $\hat{L}_1^* = \max_{1 \leq n \leq M} \frac{\|\mathbf{g}_n - \mathbf{h}_n\|}{\|\mathbf{w}_n - \mathbf{z}_{n-1}\|}$, we have $\frac{1}{4} \|\Delta_n - \Delta_{n-1}\|^2 \geq \frac{1}{(\hat{L}_1^*)^2} \|\bar{\mathbf{g}}_n - \bar{\mathbf{h}}_n\|^2 \geq \frac{1}{(\hat{L}_1^*)^2} \mu_n$. Hence, we obtain that $\sum_{n=1}^T \frac{1}{4\eta_n} \|\Delta_n - \Delta_{n-1}\|^2 \geq \sum_{n=1}^T \frac{\sqrt{\alpha + \sum_{i=1}^n \mu_i}}{\gamma D (\hat{L}_1^*)^2} \mu_n$. Hence, the term in (21) can be upper bounded by $4 \left(\frac{3D}{2\gamma} + \gamma D \right) \sqrt{\alpha + \sum_{n=1}^T \mu_n - \sum_{n=1}^T \frac{\sqrt{\alpha + \sum_{i=1}^n \mu_i}}{\gamma D (\hat{L}_1^*)^2} \mu_n}$, which in turn is bounded in the following lemma. Its proof can be found in Appendix C.5

Lemma C.6. *Given the previously defined quantities μ_n and \hat{L}_1^* , we have the following*

$$4 \left(\frac{3D}{2\gamma} + \gamma D \right) \sqrt{\alpha + \sum_{n=1}^T \mu_n - \sum_{n=1}^T \frac{\sqrt{\alpha + \sum_{i=1}^n \mu_i}}{\gamma D (\hat{L}_1^*)^2} \mu_n} \leq 16 \left(\frac{3}{2\gamma} + \gamma \right)^{\frac{3}{2}} \hat{L}_1^* \gamma^{\frac{1}{2}} D^2,$$

Combining all the pieces together yields

$$\sum_{n=1}^T \mathbb{E}[\langle \mathbf{g}_n, \Delta_n - \mathbf{u}^1 \rangle] \leq 8 \left(\frac{3}{\gamma} + \gamma \right) D \sqrt{T} \sigma + 16 \left(\frac{3}{2\gamma} + \gamma \right)^{\frac{3}{2}} \gamma^{\frac{1}{2}} D^2,$$

and this completes the proof. \square

C.3 PROOF OF LEMMA C.4

Setting $\mathbf{u} = \mathbf{u}^1$, we sum the inequality from Proposition C.3 over $n = 1$ to $n = T - 1$

$$\begin{aligned} \sum_{n=2}^T \langle \mathbf{g}_n, \Delta_n - \mathbf{u}^1 \rangle &\leq \langle \mathbf{g}_T - \mathbf{h}_T, \Delta_T - \mathbf{u}^1 \rangle - \langle \mathbf{g}_1 - \mathbf{h}_1, \Delta_1 - \mathbf{u}^1 \rangle + \sum_{n=1}^{T-1} \eta_n \|\mathbf{g}_n - \mathbf{h}_n\|^2 \\ &\quad + \sum_{n=1}^{T-1} \left(\frac{1}{2\eta_{n+1}} - \frac{1}{2\eta_n} \right) \|\Delta_{n+1} - \mathbf{u}^1\|^2 + \frac{1}{2\eta_1} \|\Delta_1 - \mathbf{u}^1\|^2 - \frac{1}{2\eta_T} \|\Delta_T - \mathbf{u}^1\|^2 \\ &\quad - \sum_{n=1}^{T-1} \frac{1}{4\eta_n} \|\Delta_{n+1} - \Delta_n\|^2. \end{aligned}$$

First, note that by the definitions of \mathbf{h}_1 and Δ_1 , we have $\Delta_1 = \arg \min_{\|\Delta\| \leq D} \langle \mathbf{h}_1, \Delta \rangle$. This implies that $\langle \mathbf{h}_1, \Delta_1 - \mathbf{u}^1 \rangle \leq 0$. Therefore

$$\begin{aligned} \sum_{n=1}^T \langle \mathbf{g}_n, \Delta_n - \mathbf{u}^1 \rangle &\leq \langle \mathbf{g}_T - \mathbf{h}_T, \Delta_T - \mathbf{u}^1 \rangle + \sum_{n=1}^{T-1} \eta_n \|\mathbf{g}_n - \mathbf{h}_n\|^2 + \sum_{n=1}^{T-1} \left(\frac{1}{2\eta_{n+1}} - \frac{1}{2\eta_n} \right) \|\Delta_{n+1} - \mathbf{u}^1\|^2 \\ &\quad + \frac{1}{2\eta_1} \|\Delta_1 - \mathbf{u}^1\|^2 - \frac{1}{2\eta_T} \|\Delta_T - \mathbf{u}^1\|^2 - \sum_{n=1}^{T-1} \frac{1}{4\eta_n} \|\Delta_{n+1} - \Delta_n\|^2. \end{aligned}$$

Additionally, applying the Cauchy-Schwarz and Young's inequalities, we can bound

$$\begin{aligned} \langle \mathbf{g}_T - \mathbf{h}_T, \Delta_T - \mathbf{u}^1 \rangle &\leq \|\mathbf{g}_T - \mathbf{h}_T\| \|\Delta_T - \mathbf{u}^1\| \\ &\leq \frac{\eta_T}{2} \|\mathbf{g}_T - \mathbf{h}_T\|^2 + \frac{1}{2\eta_T} \|\Delta_T - \mathbf{u}^1\|^2. \end{aligned}$$

Combining all the inequalities, we obtain

$$\begin{aligned} \sum_{n=1}^T \langle \mathbf{g}_n, \Delta_n - \mathbf{u}^1 \rangle &\leq \frac{\eta_T}{2} \|\mathbf{g}_T - \mathbf{h}_T\|^2 + \sum_{n=1}^{T-1} \left(\eta_n \|\mathbf{g}_n - \mathbf{h}_n\|^2 - \frac{1}{4\eta_n} \|\Delta_{n+1} - \Delta_n\|^2 \right) \\ &\quad + \sum_{n=1}^{T-1} \left(\frac{1}{2\eta_{n+1}} - \frac{1}{2\eta_n} \right) \|\Delta_{n+1} - \mathbf{u}^1\|^2 + \frac{1}{2\eta_1} \|\Delta_1 - \mathbf{u}^1\|^2. \end{aligned}$$

As η_n is a decreasing sequence, by definition (8), we have $\left(\frac{1}{2\eta_{n+1}} - \frac{1}{2\eta_n} \right) \geq 0$. Additionally, $\|\Delta_n\| \leq D$ and $\|\mathbf{u}^1\| \leq D$. Therefore, we have $\sum_{n=1}^{T-1} \left(\frac{1}{2\eta_{n+1}} - \frac{1}{2\eta_n} \right) \|\Delta_{n+1} - \mathbf{u}^1\|^2 + \frac{1}{2\eta_1} \|\Delta_1 - \mathbf{u}^1\|^2 \leq 4D^2 \sum_{n=1}^{T-1} \left(\frac{1}{2\eta_{n+1}} - \frac{1}{2\eta_n} \right) + \frac{4D^2}{2\eta_1} = \frac{2D^2}{\eta_T}$. Hence

$$\begin{aligned} \sum_{n=1}^T \langle \mathbf{g}_n, \Delta_n - \mathbf{u}^1 \rangle &\leq \frac{\eta_T}{2} \|\mathbf{g}_T - \mathbf{h}_T\|^2 + \sum_{n=1}^{T-1} \left(\eta_n \|\mathbf{g}_n - \mathbf{h}_n\|^2 - \frac{1}{4\eta_n} \|\Delta_{n+1} - \Delta_n\|^2 \right) + \frac{2D^2}{\eta_T} \\ &= \frac{2D^2}{\eta_T} + \frac{\eta_T}{2} \|\mathbf{g}_T - \mathbf{h}_T\|^2 + \sum_{n=2}^{T-1} \left(\eta_n \|\mathbf{g}_n - \mathbf{h}_n\|^2 - \frac{1}{4\eta_{n-1}} \|\Delta_n - \Delta_{n-1}\|^2 \right) \\ &\quad - \frac{1}{4\eta_{T-1}} \|\Delta_T - \Delta_{T-1}\|^2 + \eta_1 \|\mathbf{g}_1 - \mathbf{h}_1\|^2 \end{aligned}$$

Note that $\frac{\eta_T}{2} \|\mathbf{g}_T - \mathbf{h}_T\|^2 \leq \eta_T \|\mathbf{g}_T - \mathbf{h}_T\|^2$. Therefore, we can state the following upper bound

$$\begin{aligned}
\sum_{n=1}^T \langle \mathbf{g}_n, \Delta_n - \mathbf{u}^1 \rangle &\leq \frac{2D^2}{\eta_T} + \sum_{n=2}^T \left(\eta_n \|\mathbf{g}_n - \mathbf{h}_n\|^2 - \frac{1}{4\eta_{n-1}} \|\Delta_n - \Delta_{n-1}\|^2 \right) + \eta_1 \|\mathbf{g}_1 - \mathbf{h}_1\|^2 \\
&= \frac{2D^2}{\eta_T} + \sum_{n=2}^T \left(\eta_n \|\mathbf{g}_n - \mathbf{h}_n\|^2 - \frac{1}{4\eta_n} \|\Delta_n - \Delta_{n-1}\|^2 \right) + \eta_1 \|\mathbf{g}_1 - \mathbf{h}_1\|^2 \\
&\quad + \frac{1}{4} \sum_{n=2}^T \left(\frac{1}{\eta_n} - \frac{1}{\eta_{n-1}} \right) \|\Delta_n - \Delta_{n-1}\|^2 \\
&\leq \frac{2D^2}{\eta_T} + \sum_{n=2}^T \left(\eta_n \|\mathbf{g}_n - \mathbf{h}_n\|^2 - \frac{1}{4\eta_n} \|\Delta_n - \Delta_{n-1}\|^2 \right) + \eta_1 \|\mathbf{g}_1 - \mathbf{h}_1\|^2 \\
&\quad + \frac{D^2}{\eta_T} - \frac{D^2}{\eta_1}
\end{aligned}$$

Here, the first equality follows by adding and subtracting the term $\sum_{n=2}^T \frac{1}{4\eta_n} \|\Delta_n - \Delta_{n-1}\|^2$, while the final inequality uses the facts that $\left(\frac{1}{\eta_n} - \frac{1}{\eta_{n-1}}\right) \geq 0$ and $\|\Delta_n - \Delta_{n-1}\|^2 \leq 4D^2$. Moreover, we observe that $-\frac{D^2}{\eta_1} \leq -\frac{1}{4\eta_1} \|\Delta_1 - \Delta_0\|^2$. Therefore

$$\sum_{n=1}^T \langle \mathbf{g}_n, \Delta_n - \mathbf{u}^1 \rangle \leq \frac{3D^2}{\eta_T} + \sum_{n=1}^T \left(\eta_n \|\mathbf{g}_n - \mathbf{h}_n\|^2 - \frac{1}{4\eta_n} \|\Delta_n - \Delta_{n-1}\|^2 \right).$$

This completes the proof.

C.4 PROOF OF LEMMA C.5

Using η_n as defined in (8), and noting that we are analyzing the first episode (so $n \bmod T$ simplifies to n), we have

$$\frac{3D}{\eta_T} + \sum_{n=1}^T \eta_n \|\mathbf{g}_n - \mathbf{h}_n\|^2 \leq \frac{3D}{\gamma} \sqrt{\alpha + \sum_{i=1}^T \|\mathbf{g}_i - \mathbf{h}_i\|^2} + \sum_{n=1}^T \frac{\gamma D \|\mathbf{g}_n - \mathbf{h}_n\|^2}{\sqrt{\alpha + \sum_{i=1}^n \|\mathbf{g}_i - \mathbf{h}_i\|^2}}.$$

Applying the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$, we obtain

$$\frac{3D}{\eta_T} + \sum_{n=1}^T \eta_n \|\mathbf{g}_n - \mathbf{h}_n\|^2 \leq \frac{3D}{\gamma} \sqrt{\alpha} + \frac{3D}{\gamma} \sqrt{\sum_{i=1}^T \|\mathbf{g}_i - \mathbf{h}_i\|^2} + \sum_{n=1}^T \frac{\gamma D \|\mathbf{g}_n - \mathbf{h}_n\|^2}{\sqrt{\sum_{i=1}^n \|\mathbf{g}_i - \mathbf{h}_i\|^2}}.$$

Applying the upper bound corresponding to the lower bound from Lemma C.1, we get

$$\frac{3D}{\eta_T} + \sum_{n=1}^T \eta_n \|\mathbf{g}_n - \mathbf{h}_n\|^2 \leq \frac{3D}{\gamma} \sqrt{\alpha} + \frac{3D}{\gamma} \sqrt{\sum_{i=1}^T \|\mathbf{g}_i - \mathbf{h}_i\|^2} + 2\gamma D \sqrt{\sum_{i=1}^T \|\mathbf{g}_i - \mathbf{h}_i\|^2}.$$

Utilizing the fact that $(\sqrt{a} + \sqrt{b})^2 \leq 2(a+b)$ leads to

$$\frac{3D}{\eta_T} + \sum_{n=1}^T \eta_n \|\mathbf{g}_n - \mathbf{h}_n\|^2 \leq 2\sqrt{2} \left(\frac{3D}{2\gamma} + \gamma D \right) \sqrt{\alpha + \sum_{n=1}^T \|\mathbf{g}_n - \mathbf{h}_n\|^2}.$$

C.5 PROOF OF LEMMA C.6

Proof. Applying Lemma C.1, we obtain

$$\begin{aligned} & 4 \left(\frac{3D}{2\gamma} + \gamma D \right) \sqrt{\alpha + \sum_{n=1}^T \mu_n} - \sum_{n=1}^T \frac{\sqrt{\alpha + \sum_{i=1}^n \mu_i}}{\gamma D (\hat{L}_1^*)^2} \mu_n \\ & \leq 4 \left(\frac{3D}{2\gamma} + \gamma D \right) \sqrt{\alpha} + 4 \left(\frac{3D}{2\gamma} + \gamma D \right) \frac{\sum_{n=1}^T \mu_n}{\sqrt{\alpha + \sum_{i=1}^n \mu_i}} - \sum_{n=1}^T \frac{\sqrt{\alpha + \sum_{i=1}^n \mu_i}}{(\hat{L}_1^*)^2 \gamma D} \mu_n. \end{aligned}$$

Reorganizing the terms, we further have

$$\begin{aligned} & 4 \left(\frac{3D}{2\gamma} + \gamma D \right) \sqrt{\alpha + \sum_{n=1}^T \mu_n} - \sum_{n=1}^T \frac{\sqrt{\alpha + \sum_{i=1}^n \mu_i}}{\gamma D (\hat{L}_1^*)^2} \mu_n \\ & \leq 4 \left(\frac{3D}{2\gamma} + \gamma D \right) \sqrt{\alpha} + \sum_{n=1}^T \left(4 \left(\frac{3D}{2\gamma} + \gamma D \right) - \frac{\alpha + \sum_{i=1}^n \mu_i}{(\hat{L}_1^*)^2 \gamma D} \right) \frac{\mu_n}{\sqrt{\alpha + \sum_{i=1}^n \mu_i}}. \quad (22) \end{aligned}$$

Observe that the quantity $\left(4 \left(\frac{3D}{2\gamma} + \gamma D \right) - \frac{\alpha + \sum_{i=1}^n \mu_i}{4(\hat{L}_1^*)^2 \gamma D} \right)$ defines a decreasing sequence with respect to n . Define T^* as the smallest integer such that for all $n > T^*$, the following holds

$$4 \left(\frac{3D}{2\gamma} + \gamma D \right) - \frac{\alpha + \sum_{i=1}^n \mu_i}{(\hat{L}_1^*)^2 \gamma D} \leq 0$$

Remark C.2. For sufficiently small α , we can ensure that the following two inequalities hold, where T^* is an integer between 1 and T :

Then, the following two scenarios hold, depending on the value of n

$$\begin{aligned} & 4 \left(\frac{3D}{2\gamma} + \gamma D \right) (\hat{L}_1^*)^2 \gamma D \leq \alpha + \sum_{i=1}^n \mu_i \quad \forall n > T^* \\ & 4 \left(\frac{3D}{2\gamma} + \gamma D \right) (\hat{L}_1^*)^2 \gamma D \geq \alpha + \sum_{i=1}^n \mu_i \quad \forall n \leq T^* \end{aligned} \quad (23)$$

Returning to equation (22), with this definition of T^* , we have

$$\begin{aligned} & 4 \left(\frac{3D}{2\gamma} + \gamma D \right) \sqrt{\alpha + \sum_{n=1}^T \mu_n} - \sum_{n=1}^T \frac{\sqrt{\alpha + \sum_{i=1}^n \mu_i}}{\gamma D (\hat{L}_1^*)^2} \mu_n \leq 4 \left(\frac{3D}{2\gamma} + \gamma D \right) \sqrt{\alpha} \\ & \quad + \sum_{n=1}^{T^*} \left(4 \left(\frac{3D}{2\gamma} + \gamma D \right) - \frac{\alpha + \sum_{i=1}^n \mu_i}{(\hat{L}_1^*)^2 \gamma D} \right) \frac{\mu_n}{\sqrt{\alpha + \sum_{i=1}^n \mu_i}}. \end{aligned}$$

By discarding the negative terms, we obtain

$$\begin{aligned} & 4 \left(\frac{3D}{2\gamma} + \gamma D \right) \sqrt{\alpha + \sum_{n=1}^T \mu_n} - \sum_{n=1}^T \frac{\sqrt{\alpha + \sum_{i=1}^n \mu_i}}{\gamma D (\hat{L}_1^*)^2} \mu_n \leq 4 \left(\frac{3D}{2\gamma} + \gamma D \right) \sqrt{\alpha} \\ & \quad + \sum_{n=1}^{T^*} 4 \left(\frac{3D}{2\gamma} + \gamma D \right) \frac{\mu_n}{\sqrt{\alpha + \sum_{i=1}^n \mu_i}}. \end{aligned}$$

Observing that $\frac{\mu_n}{\sqrt{\alpha + \sum_{i=1}^n \mu_i}} \leq \frac{\mu_n}{\sqrt{\sum_{i=1}^n \mu_i}}$, and applying Lemma C.1 to this sequence, we obtain

$$4 \left(\frac{3D}{2\gamma} + \gamma D \right) \sqrt{\alpha + \sum_{n=1}^T \mu_n} - \sum_{n=1}^T \frac{\sqrt{\alpha + \sum_{i=1}^n \mu_i}}{\gamma D (\hat{L}_1^*)^2} \mu_n \leq 4 \left(\frac{3D}{2\gamma} + \gamma D \right) \sqrt{\alpha} \\ + 8 \left(\frac{3D}{2\gamma} + \gamma D \right) \sqrt{\sum_{n=1}^{T^*} \mu_n}.$$

Combining terms

$$4 \left(\frac{3D}{2\gamma} + \gamma D \right) \sqrt{\alpha + \sum_{n=1}^T \mu_n} - \sum_{n=1}^T \frac{\sqrt{\alpha + \sum_{i=1}^n \mu_i}}{\gamma D (\hat{L}_1^*)^2} \mu_n \leq 8 \left(\frac{3D}{2\gamma} + \gamma D \right) \sqrt{\alpha + \sum_{n=1}^{T^*} \mu_n}.$$

Invoking the upper bound for T^* from (23),

$$4 \left(\frac{3D}{2\gamma} + \gamma D \right) \sqrt{\alpha + \sum_{n=1}^T \mu_n} - \sum_{n=1}^T \frac{\sqrt{\alpha + \sum_{i=1}^n \mu_i}}{\gamma D (\hat{L}_1^*)^2} \mu_n \leq 8 \left(\frac{3D}{2\gamma} + \gamma D \right) \sqrt{4 \left(\frac{3D}{2\gamma} + \gamma D \right) (\hat{L}_1^*)^2 \gamma D} \\ = 16 \left(\frac{3}{2\gamma} + \gamma \right)^{\frac{3}{2}} \hat{L}_1^* \gamma^{\frac{1}{2}} D^2.$$

This concludes the proof. \square

C.6 PROOF OF THEOREM 4.2

Note that the algorithm restarts after every T iterations. Therefore, the bounds proven for the first episode apply to each subsequent episode, and as a corollary of Lemma 4.1 we obtain $\mathbb{E} [\text{Reg}_T(\mathbf{u}^1, \dots, \mathbf{u}^K)] \leq 8 \left(\frac{3}{2\gamma} + \gamma \right) DK \sqrt{T} \sigma + 16 \left(\frac{3}{2\gamma} + \gamma \right)^{\frac{3}{2}} \hat{L}_1^* \gamma^{\frac{1}{2}} K D^2$. Together with Proposition 2.1, we can show that:

$$\frac{1}{K} \sum_{k=1}^K \|\nabla F(\bar{\mathbf{w}}^k)\| \leq \frac{F(\mathbf{x}_0) - F^*}{DM} + 8C_1 \frac{\sigma}{\sqrt{T}} + 16C_1^{\frac{3}{2}} \hat{L}_1^* \gamma^{\frac{1}{2}} \frac{D}{T} + \frac{L_2}{48} D^2 + \frac{L_2}{2} T^2 D^2 + \frac{\sigma}{\sqrt{T}},$$

where $C_1 = \left(\frac{3}{2\gamma} + \gamma \right)$, also define $C_2 = \left(\frac{12}{\gamma} + 8\gamma + 1 \right)$.

Equivalently

$$\frac{1}{K} \sum_{k=1}^K \|\nabla F(\bar{\mathbf{w}}^k)\| \leq \frac{F(\mathbf{x}_0) - F^*}{DKT} + C_2 \frac{\sigma}{\sqrt{T}} + 16C_1^{\frac{3}{2}} \hat{L}_1^* \gamma^{\frac{1}{2}} \frac{D}{T} + L_2 T^2 D^2. \quad (24)$$

Based on the upper bound in (24), the hyperparameters D, K, T must be chosen under the constraint $KT \leq M$, where M denotese the total iteration budget and $K, T \in \mathbb{N}$. To obtain the sharpest bound, we will first determine T and K , and then specify D .

If we set $T = \min \left(\max \left(\left\lceil \left(\frac{C_2 \sigma}{L_2 D^2} \right)^{\frac{2}{5}} \right\rceil, \left\lceil \left(16C_1^{\frac{3}{2}} \frac{\hat{L}_1^* \gamma^{\frac{1}{2}}}{L_2 D} \right)^{\frac{1}{3}} \right\rceil \right), \frac{M}{2} \right)$ and $K = \lfloor \frac{M}{T} \rfloor$, we can bound the first three terms of the inequality (24) as follows:

- $KT \geq \left(\frac{M}{T} - 1 \right) T \geq M - T \geq M/2$ which implies $\frac{F(\mathbf{x}_0) - F^*}{DKT} \leq 2 \frac{F(\mathbf{x}_0) - F^*}{DM}$;
- $16C_1^{\frac{3}{2}} \hat{L}_1^* \gamma^{\frac{1}{2}} \frac{D}{T} \leq \max \left(16^{\frac{2}{3}} L_2^{\frac{1}{3}} (\hat{L}_1^*)^{\frac{2}{3}} \gamma^{\frac{1}{3}} C_1 D^{\frac{4}{3}}, 32C_1^{\frac{3}{2}} \hat{L}_1^* \gamma^{\frac{1}{2}} \frac{D}{M} \right) \leq \max \left(7L_2^{\frac{1}{3}} (\hat{L}_1^*)^{\frac{2}{3}} \gamma^{\frac{1}{3}} C_1 D^{\frac{4}{3}}, 32C_1^{\frac{3}{2}} \hat{L}_1^* \gamma^{\frac{1}{2}} \frac{D}{M} \right)$;
- $C_2 \frac{\sigma}{\sqrt{T}} \leq \max \left(C_2^{\frac{4}{5}} L_2^{\frac{1}{5}} D^{\frac{2}{5}} \sigma^{\frac{4}{5}}, C_2 \sqrt{2} \frac{\sigma}{\sqrt{M}} \right)$.

We upper bound $L_2 T^2 D^2$ as follows:

$$\begin{aligned} L_2 T^2 D^2 &\leq L_2 \max \left(\left[\left(\frac{C_2 \sigma}{L_2 D^2} \right)^{\frac{2}{5}} \right], \left[\left(16 C_1^{\frac{3}{2}} \frac{\hat{L}_1^* \gamma^{\frac{1}{2}}}{L_2 D} \right)^{\frac{1}{3}} \right] \right)^2 D^2 \\ &\leq L_2 \left[\left(\frac{C_2 \sigma}{L_2 D^2} \right)^{\frac{2}{5}} \right]^2 D^2 + L_2 \left[\left(16 C_1^{\frac{3}{2}} \frac{\hat{L}_1^* \gamma^{\frac{1}{2}}}{L_2 D} \right)^{\frac{1}{3}} \right]^2 D^2. \end{aligned}$$

Using the fact that $\lceil x \rceil^2 \leq (x+1)^2 \leq 2x^2 + 2$ for any $x \geq 0$, we obtain

$$\begin{aligned} L_2 T^2 D^2 &\leq L_2 \left(2 \left(\frac{C_2 \sigma}{L_2 D^2} \right)^{\frac{4}{5}} + 2 \right) D^2 + L_2 \left(2 \left(16 C_1^{\frac{3}{2}} \frac{\hat{L}_1^* \gamma^{\frac{1}{2}}}{L_2 D} \right)^{\frac{2}{3}} + 2 \right) D^2 \\ &= 2 L_2 \left(\frac{C_2 \sigma}{L_2 D^2} \right)^{\frac{4}{5}} D^2 + 2 L_2 \left(16 C_1^{\frac{3}{2}} \frac{\hat{L}_1^* \gamma^{\frac{1}{2}}}{L_2 D} \right)^{\frac{2}{3}} D^2 + 4 L_2 D^2 \\ &\leq 2 C_2^{\frac{4}{5}} L_2^{\frac{1}{5}} \sigma^{\frac{4}{5}} D^{\frac{2}{5}} + 13 C_1 (\hat{L}_1^*)^{\frac{2}{3}} L_2^{\frac{1}{3}} D^{\frac{4}{3}} + 4 L_2 D^2. \end{aligned}$$

Substituting these values into (24) yields

$$\begin{aligned} \frac{1}{K} \sum_{k=1}^K \|\nabla F(\bar{\mathbf{w}}^k)\| &\leq 2 \frac{F(\mathbf{x}_0) - F^*}{DM} + \max \left(7 L_2^{\frac{1}{3}} (\hat{L}_1^*)^{\frac{2}{3}} \gamma^{\frac{1}{3}} C_1 D^{\frac{4}{3}}, 32 C_1^{\frac{3}{2}} \hat{L}_1^* \gamma^{\frac{1}{2}} \frac{D}{M} \right) \\ &\quad + \max \left(C_2^{\frac{4}{5}} L_2^{\frac{1}{5}} D^{\frac{2}{5}} \sigma^{\frac{4}{5}}, C_2 \sqrt{2} \frac{\sigma}{\sqrt{M}} \right) + 2 C_2^{\frac{4}{5}} L_2^{\frac{1}{5}} \sigma^{\frac{4}{5}} D^{\frac{2}{5}} + 13 C_1 (\hat{L}_1^*)^{\frac{2}{3}} L_2^{\frac{1}{3}} D^{\frac{4}{3}} \\ &\quad + 4 L_2 D^2. \end{aligned}$$

Using $\max(a, b) \leq a + b$, and after rounding constants, we get:

$$\begin{aligned} \frac{1}{K} \sum_{k=1}^K \|\nabla F(\bar{\mathbf{w}}^k)\| &\leq 2 \frac{F(\mathbf{x}_0) - F^*}{DM} + 32 C_1^{\frac{3}{2}} \hat{L}_1^* \gamma^{\frac{1}{2}} \frac{D}{M} + C_2 \sqrt{2} \frac{\sigma}{\sqrt{M}} \\ &\quad + 3 C_2^{\frac{4}{5}} L_2^{\frac{1}{5}} \sigma^{\frac{4}{5}} D^{\frac{2}{5}} + 20 C_1 (\hat{L}_1^*)^{\frac{2}{3}} L_2^{\frac{1}{3}} D^{\frac{4}{3}} + 4 L_2 D^2. \end{aligned}$$

Setting $D = \min \left\{ \left(2 \frac{F(\mathbf{x}_0) - F^*}{3 M C_2^{\frac{4}{5}} L_2^{\frac{1}{5}} \sigma^{\frac{4}{5}}} \right)^{\frac{5}{7}}, \left(\frac{F(\mathbf{x}_0) - F^*}{10 M (\hat{L}_1^*)^{\frac{2}{3}} L_2^{\frac{1}{3}} \gamma^{\frac{1}{3}} C_1} \right)^{\frac{3}{7}} \right\}$ leads to the following upper bounds:

- $3 C_2^{\frac{4}{5}} L_2^{\frac{1}{5}} D^{\frac{2}{5}} \sigma^{\frac{4}{5}} \leq 3^{\frac{5}{7}} 2^{\frac{2}{7}} \frac{(F(\mathbf{x}_0) - F^*)^{\frac{2}{7}}}{M^{\frac{2}{7}}} C_2^{\frac{4}{7}} L_2^{\frac{1}{7}} \sigma^{\frac{4}{7}} \leq 3 \frac{(F(\mathbf{x}_0) - F^*)^{\frac{2}{7}}}{M^{\frac{2}{7}}} C_2^{\frac{4}{7}} L_2^{\frac{1}{7}} \sigma^{\frac{4}{7}};$
- $20 L_2^{\frac{1}{3}} (\hat{L}_1^*)^{\frac{2}{3}} \gamma^{\frac{1}{3}} C_1 D^{\frac{4}{3}} \leq 20 \gamma^{\frac{1}{7}} C_1^{\frac{3}{7}} (\hat{L}_1^*)^{\frac{2}{7}} L_2^{\frac{1}{7}} \frac{(F(\mathbf{x}_0) - F^*)^{\frac{4}{7}}}{10^{\frac{4}{7}} M^{\frac{4}{7}}} \leq 6 \gamma^{\frac{1}{7}} C_1^{\frac{3}{7}} (\hat{L}_1^*)^{\frac{2}{7}} L_2^{\frac{1}{7}} \frac{(F(\mathbf{x}_0) - F^*)^{\frac{4}{7}}}{M^{\frac{4}{7}}};$
- $2 \frac{F(\mathbf{x}_0) - F^*}{DM} \leq 2 \cdot 10^{\frac{3}{7}} \gamma^{\frac{1}{7}} C_1^{\frac{3}{7}} (\hat{L}_1^*)^{\frac{2}{7}} L_2^{\frac{1}{7}} \frac{(F(\mathbf{x}_0) - F^*)^{\frac{4}{7}}}{M^{\frac{4}{7}}} + 2 \cdot 1.5^{\frac{5}{7}} \frac{(F(\mathbf{x}_0) - F^*)^{\frac{2}{7}}}{M^{\frac{2}{7}}} C_2^{\frac{4}{7}} L_2^{\frac{1}{7}} \sigma^{\frac{4}{7}} \leq 6 \gamma^{\frac{1}{7}} C_1^{\frac{3}{7}} (\hat{L}_1^*)^{\frac{2}{7}} L_2^{\frac{1}{7}} \frac{(F(\mathbf{x}_0) - F^*)^{\frac{4}{7}}}{M^{\frac{4}{7}}} + 3 \frac{(F(\mathbf{x}_0) - F^*)^{\frac{2}{7}}}{M^{\frac{2}{7}}} C_2^{\frac{4}{7}} L_2^{\frac{1}{7}} \sigma^{\frac{4}{7}};$
- $32 C_1^{\frac{3}{2}} \hat{L}_1^* \gamma^{\frac{1}{2}} \frac{D}{M} \leq \frac{32}{10^{\frac{3}{7}}} \frac{C_1^{\frac{15}{14}} (\hat{L}_1^*)^{\frac{5}{7}} \gamma^{\frac{5}{14}} (F(\mathbf{x}_0) - F^*)^{\frac{3}{7}}}{M^{\frac{10}{7}} L_2^{\frac{1}{7}}} \leq 12 \frac{C_1^{\frac{15}{14}} (\hat{L}_1^*)^{\frac{5}{7}} \gamma^{\frac{5}{14}} (F(\mathbf{x}_0) - F^*)^{\frac{3}{7}}}{M^{\frac{10}{7}} L_2^{\frac{1}{7}}};$

$$\bullet 4L_2D^2 \leq \frac{4}{10^{\frac{6}{7}}} \frac{L_2^{\frac{5}{7}} (F(\mathbf{x}_0) - F^*)^{\frac{6}{7}}}{M^{\frac{6}{7}} (\hat{L}_1^*)^{\frac{4}{7}} \gamma^{\frac{2}{7}} C_1^{\frac{6}{7}}} \leq \frac{L_2^{\frac{5}{7}} (F(\mathbf{x}_0) - F^*)^{\frac{6}{7}}}{M^{\frac{6}{7}} (\hat{L}_1^*)^{\frac{4}{7}} \gamma^{\frac{2}{7}} C_1^{\frac{6}{7}}}.$$

Finally, substituting this value into the bound yields

$$\begin{aligned} \frac{1}{K} \sum_{k=1}^K \|\nabla F(\bar{\mathbf{w}}^k)\| &\leq 12\gamma^{\frac{1}{7}} C_1^{\frac{3}{7}} (\hat{L}_1^*)^{\frac{2}{7}} L_2^{\frac{1}{7}} \frac{(F(\mathbf{x}_0) - F^*)^{\frac{4}{7}}}{M^{\frac{4}{7}}} + 6 \frac{(F(\mathbf{x}_0) - F^*)^{\frac{2}{7}}}{M^{\frac{2}{7}}} C_2^{\frac{4}{7}} L_2^{\frac{1}{7}} \sigma^{\frac{4}{7}} \\ &\quad + 12 \frac{C_1^{\frac{15}{14}} (\hat{L}_1^*)^{\frac{5}{7}} \gamma^{\frac{5}{14}} (F(\mathbf{x}_0) - F^*)^{\frac{3}{7}}}{M^{\frac{10}{7}} L_2^{\frac{1}{2}}} + \frac{L_2^{\frac{5}{7}} (F(\mathbf{x}_0) - F^*)^{\frac{6}{7}}}{M^{\frac{6}{7}} (\hat{L}_1^*)^{\frac{4}{7}} \gamma^{\frac{2}{7}} C_1^{\frac{6}{7}}} + C_2 \sqrt{2} \frac{\sigma}{\sqrt{M}}. \end{aligned}$$

This final bound confirms the convergence rate stated in the main theorem, as the last three terms are of higher order and thus do not influence the rate.

D USE OF LARGE LANGUAGE MODELS

Large Language models were used lightly to refine wording and correct grammatical errors in parts of the paper.