Random Scaling and Momentum for Non-smooth Non-convex Optimization

Qinzi Zhang [1] **Ashok Cutkosky** [1]

Abstract

Training neural networks requires optimizing a loss function that may be highly irregular, and in particular neither convex nor smooth. Popular training algorithms are based on stochastic gradient descent with momentum (SGDM), for which classical analysis applies only if the loss is either convex or smooth. We show that a very small modification to SGDM closes this gap: simply scale the update at each time point by an exponentially distributed random scalar. The resulting algorithm achieves optimal convergence guarantees. Intriguingly, this result is not derived by a specific analysis of SGDM: instead, it falls naturally out of a more general framework for converting online convex optimization algorithms to non-convex optimization algorithms.

1. Introduction

Non-convex optimization algorithms are one of the fundamental tools in modern machine learning, as training neural network models requires optimizing a non-convex loss function. This paper provides a new theoretical framework for building such algorithms. The simplest application of this framework almost exactly recapitulates the standard algorithm used in practice: stochastic gradient descent with momentum (SGDM).

The goal of any optimization algorithm used to train a neural network is to minimize a potentially non-convex objective function. Formally, given $F : \mathbb{R}^d \to \mathbb{R}$, the problem is to solve

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} F(\boldsymbol{x}) = \mathbb{E}_z[f(\boldsymbol{x}, z)],$$

where f is a stochastic estimator of F. In practice, $\boldsymbol{x}$ denotes the parameters of a neural network model, and z denotes the data point. Following the majority of the literature, we focus on first-order stochastic optimization algorithms, which can only access to the stochastic gradient $\nabla f(\boldsymbol{x}, z)$ as an estimate of the unknown true gradient $\nabla F(\boldsymbol{x})$. We measure the "cost" of an algorithm by counting the number of stochastic gradient evaluations it requires to achieve some desired convergence guarantee. We will frequently refer to this count as the number of "iterations" employed by the algorithm.

When the objective function is non-convex, finding a global minimum can be intractable. To navigate this complexity, many prior works have imposed various smoothness assumptions on the objective. These include, but not limited to, first-order smoothness (Ghadimi & Lan, 2013; Carmon et al., 2017; Arjevani et al., 2022; Carmon et al., 2019), second-order smoothness (Tripuraneni et al., 2018; Carmon et al., 2018; Fang et al., 2019; Arjevani et al., 2020), and mean-square smoothness (Allen-Zhu, 2018; Fang et al., 2018; Cutkosky & Orabona, 2019; Arjevani et al., 2022). Instead of finding the global minimum, the smoothness conditions allow us to find an ϵ-stationary point $\boldsymbol{x}$ of F such that $\|\nabla F(\boldsymbol{x})\| \le \epsilon$.

The optimal rates for smooth non-convex optimization are now well-understood. When the objective is smooth, stochastic gradient descent (SGD) requires $O(\epsilon^{-4})$ iterations to find ϵ-stationary point, matching the optimal rate (Arjevani et al., 2019). When F is second-order smooth, a variant of SGD augmented with occasional random perturbations achieves the optimal rate $O(\epsilon^{-7/2})$ (Fang et al., 2019; Arjevani et al., 2020). Moreover, when F is mean-square smooth, variance-reduction algorithms, such as SPIDER (Fang et al., 2018) and SNVRG (Zhou et al., 2018), achieve the optimal rate $O(\epsilon^{-3})$ (Arjevani et al., 2019). All of these algorithms can be viewed as variants of SGD.

In addition to these theoretical optimality results, SGD and its variants are also incredibly effective in practice across a wide variety of deep learning tasks. Among these variants, the family of momentum algorithms have become particularly popular (Sutskever et al., 2013; Kingma & Ba, 2014; You et al., 2017; 2019; Cutkosky & Orabona, 2019; Cutkosky & Mehta, 2020; Ziyin et al., 2021). Under smoothness conditions, the momentum algorithms also achieve the same optimal rates.

However, modern deep learning models frequently incorpo-

[1] Department of Electrical and Computer Engineering, Boston University, Boston, USA. Correspondence to: Qinzi Zhang <qinziz@bu.edu>, Ashok Cutkosky <ashok@cutkosky.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

rate a range of non-smooth architectures, including elements like ReLU, max pooling, and quantization layers. These components result in a non-smooth optimization objective, violating the fundamental assumption of a vast majority of prior works. Non-smooth optimization is fundamentally more difficult than its smooth counterpart, as in the worst-case Kornowski & Shamir (2022b) show that it is actually impossible to find a neighborhood around ϵ-stationary. This underscores the need for a tractable convergence criterion in non-smooth non-convex optimization.

One line of research in non-smooth non-convex optimization studies weakly-convex objectives (Davis & Drusvyatskiy, 2019; Mai & Johansson, 2020), with a focus on finding ϵ-stationary points of the Moreau envelope of the objectives. It has been demonstrated that various algorithms, including the proximal subgradient method and SGDM, can achieve the optimal rate of $O(\epsilon^{-4})$ for finding an ϵ-stationary point of the Moreau envelope. However, it is important to note that the assumption of weak convexity is crucial for the convergence notion involving the Moreau envelope. Our interest lies in solving non-smooth non-convex optimization without relying on the weak convexity assumption.

To this end, Zhang et al. (2020) proposed employing Goldstein stationary points (Goldstein, 1977) as a convergence target in non-smooth non-convex (and non-weakly-convex) optimization. This approach has been widely accepted by recent works studying non-smooth optimization (Kornowski & Shamir, 2022a; Lin et al., 2022; Kornowski & Shamir, 2023; Cutkosky et al., 2023). Formally, $\boldsymbol{x}$ is a (δ, ϵ)-Goldstein stationary point if there exists a random vector $\boldsymbol{y}$ such that $\mathbb{E}[\boldsymbol{y}] = \boldsymbol{x}$, $\|\boldsymbol{y} - \boldsymbol{x}\| \le \delta$ almost surely, and $\|\mathbb{E}[\nabla F(\boldsymbol{y})]\| \le \epsilon$.[1] The best-possible rate for finding a (δ, ϵ)-Goldstein stationary point is $O(\delta^{-1}\epsilon^{-3})$ iterations. This rate was only recently achieved by Cutkosky et al. (2023), who developed an "online-to-non-convex conversion" (O2NC) technique that converts online convex optimization (OCO) algorithms to non-smooth non-convex stochastic optimization algorithms. Building on this background, we will relax the definition of stationarity and extend this O2NC technique to eventually develop a convergence analysis of SGDM in the non-smooth and non-convex setting.

1.1. Our Contribution

In this paper, we introduce a new notion of stationarity for non-smooth non-convex objectives. Our notion is a natural relaxation of the Goldstein stationary point, but will allow for more flexible algorithm design. Intuitively, the problem with the Goldstein stationary point is that to verify that a point $\boldsymbol{x}$ is a stationary point, one must evaluate the gradient many times inside a ball of some small radius δ about $\boldsymbol{x}$. This means that algorithms finding such points usually make fairly conservative updates to sufficiently explore this ball: in essence, they work by verifying each iterate is *not* close to a stationary point before moving on to the next iterate. Algorithms used in practice do not typically behave this way, and our relaxed definition will not require us to employ such behavior.

Using our new criterion, we propose a general framework, "Exponentiated O2NC", that converts OCO algorithms to non-smooth optimization algorithms. This framework is an extension of the O2NC technique of Cutkosky et al. (2023) that distinguishes itself through two significant improvements.

Firstly, the original O2NC method requires the OCO algorithm to constrain all of its iterates to a small ball of radius roughly $\delta\epsilon^2$. This approach is designed to ensure that the parameters within any period of ϵ^{-2} iterations remain inside a ball of radius δ. The algorithm then uses these ϵ^{-2} gradient evaluations inside a ball of radius δ to check if the current iterate is a stationary point (i.e., if the average gradient has norm less than ϵ). Our new criterion, however, obviates the need for such explicit constraints, intuitively allowing our algorithms to make larger updates when far from a stationary point.

Secondly, O2NC does not evaluate gradients at the actual iterates. Instead, gradients are evaluated at an intermediate variable $\boldsymbol{w}_n$ lying between the two iterates $\boldsymbol{x}_n$ and $\boldsymbol{x}_{n+1}$. This conflicts with essentially all practical algorithms, and moreover imposes an extra memory burden. In contrast, our algorithm evaluates gradients exactly at each iterate, which simplifies implementation and improves space complexity.

Armed with this improved framework, we proposed an unconstrained variant of online gradient descent, which is derived from the family of online mirror descent with composite loss. When applied within this algorithm, our framework produces an algorithm that is exactly equal to stochastic gradient descent with momentum (SGDM), subject to an additional random scaling on the update. Notably, it also achieves the optimal rate under our new criterion.

To summarize, this paper has the following contributions:

- We introduce a relaxed convergence criterion for non-smooth optimization that recovers all useful properties of Goldstein stationary point.
- We propose a modified online-to-non-convex conversion framework that does not require intermediate states.
- We apply our new conversion to the most standard

[1] To be consistent with our proposed definition, we choose to present the definition of (δ, ϵ)-Goldstein stationary point involving a random vector $\boldsymbol{y}$. This presentation is equivalent to the original definition proposed by (Zhang et al., 2020).

OCO algorithm: "online gradient descent". The resulting method achieves optimal convergence guarantees as is almost exactly the same as the standard SGDM algorithm. The only difference is that the updates of SGDM are now scaled by an exponential random variable. This is especially remarkable because the machinery that we employ does not particularly resemble SGDM until it is finally all put together.

2. Preliminaries

Notations Bold font $\boldsymbol{x}$ denotes a vector in $\mathbb{R}^d$ and $\|\boldsymbol{x}\|$ denotes its Euclidean norm. We define $B_d(\boldsymbol{x}, r) = \{\boldsymbol{y} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{y}\| \le r\}$ and sometimes drop the subscript d when the context is clear. We use $[n]$ as an abbreviation for $\{1, 2, \dots, n\}$. We adopt the standard big-O notation, and $f \lesssim g$ denotes $f = O(g)$. $\mathcal{P}(S)$ denotes the set of all distributions over a measurable set S.

Stochastic Optimization Given a function $F : \mathbb{R}^d \to \mathbb{R}$, F is G-Lipschitz if $|F(\boldsymbol{x}) - F(\boldsymbol{y})| \le G\|\boldsymbol{x} - \boldsymbol{y}\|, \forall \boldsymbol{x}, \boldsymbol{y}$. Equivalently, when F is differentiable, F is G-Lipschitz if $\|\nabla F(\boldsymbol{x})\| \le G, \forall \boldsymbol{x}$. F is H-smooth if F is differentiable and ∇F is H-Lipschitz; F is ρ-second-order-smooth if F is twice differentiable and $\nabla^2 F$ is ρ-Lipschitz.

Assumption 2.1. We assume that our objective function $F : \mathbb{R}^d \to \mathbb{R}$ is differentiable and G Lipschitz, and given an initial point $\boldsymbol{x}_0$, $F(\boldsymbol{x}_0) - \inf F(\boldsymbol{x}) \le F^*$ for some known F^*. We also assume the stochastic gradient satisfies $\mathbb{E}[\nabla f(\boldsymbol{x}, z) \,|\, \boldsymbol{x}] = \nabla F(\boldsymbol{x}), \mathbb{E}\,\|\nabla F(\boldsymbol{x}) - \nabla f(\boldsymbol{x}, z)\|^2 \le \sigma^2$ for all $\boldsymbol{x}, z$. Finally, we assume that F is *well-behaved* in the sense of (Cutkosky et al., 2023): for any points $\boldsymbol{x}$ and $\boldsymbol{y}$, $F(\boldsymbol{x}) - F(\boldsymbol{y}) = \int_0^1 \langle \nabla F(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})), \boldsymbol{y} - x\rangle \, dt$.

Online Learning An online convex optimization (OCO) algorithm is an iterative algorithm that uses the following procedure: in each iteration n, the algorithm plays an action Δ_n and then receives a convex loss function ℓ_n The goal is to minimize the regret w.r.t. some comparator $\boldsymbol{u}$, defined as

$$\mathrm{Regret}_n(\boldsymbol{u}) := \textstyle\sum_{t=1}^n \ell_t(\Delta_t) - \ell_t(\boldsymbol{u}).$$

The most basic OCO algorithm is online gradient descent: $\Delta_{n+1} = \Delta_n - \eta \nabla \ell_n(\Delta_n)$, which guarantees $\mathrm{Regret}_N(\boldsymbol{u}) = O(\sqrt{N})$ for appropriate η. Notably, in OCO we make *no assumptions* about the dynamics of ℓ_n. They need not be stochastic, and could even be adversarially generated. We will be making use of algorithms that obtain *anytime* regret bounds. That is, for all n and any sequence of $\boldsymbol{u}_1, \boldsymbol{u}_2, \dots$, it is possible to bound $\mathrm{Regret}_n(\boldsymbol{u}_n)$ by some appropriate quantities (that may be function of n). This is no great burden: almost all online convex optimization algorithms have anytime regret bounds. For readers interested in more details, please refer to (Cesa-Bianchi & Lugosi, 2006; Hazan, 2019; Orabona, 2019).

2.1. Non-smooth Optimization

Suppose F is differentiable. $\boldsymbol{x}$ is an ϵ-stationary point of F if $\|\nabla F(\boldsymbol{x})\| \le \epsilon$. This definition is the standard criterion for smooth non-convex optimization. For non-smooth non-convex optimization, the standard criterion is the following: $\boldsymbol{x}$ is an (δ, ϵ)-Goldstein stationary point of F if there exists $S \subset \mathbb{R}^d$ and $P \in \mathcal{P}(S)$ such that $\boldsymbol{y} \sim P$ satisfies $\mathbb{E}[\boldsymbol{y}] = \boldsymbol{x}$, $\|\boldsymbol{y} - \boldsymbol{x}\| \le \delta$ almost surely, and $\|\,\mathbb{E}[\nabla F(\boldsymbol{y})]\| \le \epsilon$.[2] Next, we formally define (c, ϵ)-stationary point, our proposed new criterion for non-smooth optimization.

Definition 2.2. Suppose $F : \mathbb{R}^d \to \mathbb{R}$ is differentiable, $\boldsymbol{x}$ is a (c, ϵ)-stationary point of F if $\|\nabla F(\boldsymbol{x})\|_c \le \epsilon$, where

$$\|\nabla F(\boldsymbol{x})\|_c = \inf_{\substack{S \subset \mathbb{R}^d \\ \boldsymbol{y} \sim P \in \mathcal{P}(S) \\ \mathbb{E}[\boldsymbol{y}] = \boldsymbol{x}}} \|\,\mathbb{E}[\nabla F(\boldsymbol{y})]\| + c \cdot \mathbb{E}\|\boldsymbol{y} - \boldsymbol{x}\|^2.$$

In other words, if $\boldsymbol{x}$ is a (c, ϵ)-stationary point, then there exists $S \subset \mathbb{R}^d$, $P \in \mathcal{P}(S)$ such that $\boldsymbol{y} \sim P$ satisfies $\mathbb{E}[\boldsymbol{y}] = \boldsymbol{x}$, $\mathbb{E}\,\|\boldsymbol{y} - \boldsymbol{x}\|^2 \le \epsilon/c$, and $\|\,\mathbb{E}[\nabla F(\boldsymbol{y})]\| \le \epsilon$. To see how this definition is related to the previous (ϵ, δ)-Goldstein stationary point definition, consider the case when $c = \epsilon/\delta^2$. Then this new definition of (c, ϵ)-stationary point is almost identical to (δ, ϵ)-Goldstein stationary point, except that it relaxes the constraint from $\|\boldsymbol{y} - \boldsymbol{x}\| \le \delta$ to $\mathbb{E}\,\|\boldsymbol{y} - \boldsymbol{x}\|^2 \le \delta^2$.

To further motivate this definition, we demonstrate that (c, ϵ)-stationary point retains desirable properties of Goldstein stationary points. Firstly, the following result shows that, similar to Goldstein stationary points, (c, ϵ)-stationary points can also be reduced to first-order stationary points with proper choices of c when the objective is smooth or second-order smooth.

Lemma 2.3. *Suppose F is H-smooth. If $\|\nabla F(\boldsymbol{x})\|_c \le \epsilon$ where $c = H^2\epsilon^{-1}$, then $\|\nabla F(\boldsymbol{x})\| \le 2\epsilon$.*
Suppose F is ρ-second-order-smooth. If $\|\nabla F(\boldsymbol{x})\|_c \le \epsilon$ where $c = \rho/2$, then $\|\nabla F(\boldsymbol{x})\| \le 2\epsilon$.

As an immediate implication, suppose an algorithm achieves $O(c^{1/2}\epsilon^{-7/2})$ rate for finding a (c, ϵ)-stationary point. Then Lemma 2.3 implies that, with $c = O(\epsilon^{-1})$, the algorithm automatically achieves the optimal rate of $O(\epsilon^{-4})$ for smooth objectives (Arjevani et al., 2019). Similarly, with $c = O(1)$, it achieves the optimal rate of $O(\epsilon^{-7/2})$ for second-order smooth objectives (Arjevani et al., 2020).

Secondly, we show in the following lemma that (c, ϵ)-stationary points can also be reduced to Goldstein stationary points when the objective is Lipschitz.

[2] The original definition of (δ, ϵ) Goldstein stationary point proposed by (Zhang et al., 2020) does not require the condition $\mathbb{E}[\boldsymbol{y}] = \boldsymbol{x}$. However, as shown in (Cutkosky et al., 2023), this condition allows us to reduce a Goldstein stationary point to an ϵ-stationary point when the loss is second-order smooth. Hence we also keep this condition.

Algorithm 1 O2NC (Cutkosky et al., 2023)

```
Input: OCO algorithm $\mathcal{A}$, initial state $\boldsymbol{x}_0$, parameters $N, K, T \in \mathbb{N}$ such that $N = KT$.
for $n \leftarrow 1, 2, \dots, N$ do
    Receive $\Delta_n$ from $\mathcal{A}$.
    Update $\boldsymbol{x}_n \leftarrow \boldsymbol{x}_{n-1} + \Delta_n$ and $\boldsymbol{w}_n \leftarrow \boldsymbol{x}_{n-1} + s_n\Delta_n$, where $s_n \sim \mathrm{Unif}([0,1])$ i.i.d.
    Compute $\boldsymbol{g}_n \leftarrow \nabla f(\boldsymbol{x}_n, z_n)$.
    Send loss $\ell_n(\Delta) = \langle \boldsymbol{g}_n, \Delta \rangle$ to $\mathcal{A}$.
    // For output only (update every $T$ iteration):
    If $n = kT$, compute $\overline{\boldsymbol{w}}_k = \frac{1}{T}\sum_{t=0}^{T-1} \boldsymbol{w}_{n-t}$.
end for
Output $\overline{\boldsymbol{w}} \sim \mathrm{Unif}(\{\overline{\boldsymbol{w}}_k : k \in [K]\})$.
```

Lemma 2.4. *Suppose F is G-Lipschitz. For any $c, \epsilon, \delta > 0$, a (c, ϵ)-stationary point is also a (δ, ϵ')-Goldstein stationary point where $\epsilon' = (1 + \frac{2G}{c\delta^2})\epsilon$.*

2.2. Online-to-non-convex Conversion

Since our algorithm is an extension of the online-to-non-convex conversion (O2NC) technique proposed by (Cutkosky et al., 2023), we briefly review the original O2NC algorithm. The pseudocode is outlined in Algorithm 1, with minor adjustments in notations for consistency with our presentation.

At its essence, O2NC shifts the challenge of optimizing a non-convex and non-smooth objective into minimizing regret. The intuition is as follows. By adding a uniform perturbation $s_n \in [0,1]$, $\langle \nabla f(\boldsymbol{x}_{n-1} + s_n\Delta_n, z_n), \Delta_n \rangle = \langle \boldsymbol{g}_n, \Delta_n \rangle$ is an unbiased estimator of $F(\boldsymbol{x}_n) - F(\boldsymbol{x}_{n-1})$, effectively capturing the "training progress". Consequently, by minimizing the regret, which is equivalent to minimizing $\sum_{n=1}^N \langle \boldsymbol{g}_n, \Delta_n \rangle$, the algorithm automatically identifies the most effective update step Δ_n.

2.3. Paper Organization

In Section 3, we present the general online-to-non-convex framework, Exponentiated O2NC. We first explain the intuitions behind the algorithm design, and then we provide the convergence analysis in Section 3.1. In Section 4, we provide an explicit instantiation of our framework, and see that the resulting algorithm is essentially the standard SGDM. In Section 5, we present a lower bound for finding (c, ϵ)-stationary point. In Section 6, we present empirical evaluations.

3. Exponentiated Online-to-non-convex

In this section, we present our improved online-to-non-convex framework, Exponentiated O2NC, and explain the key techniques we employed to improve the algorithm. The pseudocode is presented in Algorithm 2.

Algorithm 2 Exponentiated O2NC

```
Input: OCO algorithm $\mathcal{A}$, initial state $\boldsymbol{x}_0$, parameters $N \in \mathbb{N}, \beta \in (0,1)$, regularizers $\mathcal{R}_n(\Delta)$.
for $n \leftarrow 1, 2, \dots, N$ do
    Receive $\Delta_n$ from $\mathcal{A}$.
    Update $\boldsymbol{x}_n \leftarrow \boldsymbol{x}_{n-1} + s_n\Delta_n$, where $s_n \sim \mathrm{Exp}(1)$ i.i.d.
    Compute $\boldsymbol{g}_n \leftarrow \nabla f(\boldsymbol{x}_n, z_n)$.
    Send loss $\ell_n(\Delta) = \langle \beta^{-n}\boldsymbol{g}_n, \Delta \rangle + \mathcal{R}_n(\Delta)$ to $\mathcal{A}$.
    // For output only (does *not* affect training):
    Update $\overline{\boldsymbol{x}}_n = \frac{\beta - \beta^n}{1-\beta^n}\overline{\boldsymbol{x}}_{n-1} + \frac{1-\beta}{1-\beta^n}\boldsymbol{x}_n$.
    Equivalently, $\overline{\boldsymbol{x}}_n = \sum_{t=1}^n \beta^{n-t}\boldsymbol{x}_t \cdot \frac{1-\beta}{1-\beta^n}$.
end for
Output $\overline{\boldsymbol{x}} \sim \mathrm{Unif}(\{\overline{\boldsymbol{x}}_n : n \in [N]\})$.
```

Random Scaling One notable feature of Algorithm 2 is that the update Δ_n is scaled by an exponential random variable s_n. Formally, we have the following result:

Lemma 3.1. *Let $s \sim \mathrm{Exp}(\lambda)$ for some $\lambda > 0$, then*

$$\mathbb{E}_s[F(\boldsymbol{x} + s\Delta) - F(\boldsymbol{x})] = \mathbb{E}_s[\langle \nabla F(\boldsymbol{x} + s\Delta), \Delta \rangle]/\lambda.$$

In Algorihtm 2, we set $s_n \sim \mathrm{Exp}(1)$ and then define $\boldsymbol{x}_n = \boldsymbol{x}_{n-1} + s_n\Delta_n$. Thus, Lemma 3.1 implies that

$$\begin{aligned}\mathbb{E}[F(\boldsymbol{x}_n) - F(\boldsymbol{x}_{n-1})] &= \mathbb{E}\langle \nabla F(\boldsymbol{x}_n), \Delta_n \rangle \\ &= \mathbb{E}\langle \nabla F(\boldsymbol{x}_n), \boldsymbol{x}_n - \boldsymbol{x}_{n-1} \rangle.\end{aligned}$$

In other words, we can estimate the "training progress" $F(\boldsymbol{x}_n) - F(\boldsymbol{x}_{n-1})$ by directly computing the stochastic gradient at iterate $\boldsymbol{x}_n$. By exploiting favorable properties of the exponential distribution, we dispense with the need for the "auxiliary point" $\boldsymbol{w}_n$ employed by O2NC.

We'd like to highlight the significance of this result. The vast majority of smooth non-convex optimization analysis depends on the assumption that $F(\boldsymbol{x})$ is locally linear, namely $F(\boldsymbol{x}_n) - F(\boldsymbol{x}_{n-1}) \approx \langle \nabla F(\boldsymbol{x}_n), \boldsymbol{x}_n - \boldsymbol{x}_{n-1} \rangle$. Under various smoothness assumptions, the error in this approximation can be controlled via bounds on the remainder in a Taylor series. For example, when F is smooth, then $F(\boldsymbol{x}_n) - F(\boldsymbol{x}_{n-1}) = \langle F(\boldsymbol{x}_n), \boldsymbol{x}_n - \boldsymbol{x}_{n-1} \rangle + O(\|\boldsymbol{x}_n - \boldsymbol{x}_{n-1}\|^2)$. However, since smoothness is necessary for bounding Taylor approximation error, such analysis technique is inapplicable in non-smooth optimization. In contrast, by scaling an exponential random variable to the update, we directly establish a linear equation that $\mathbb{E}[F(\boldsymbol{x}_n) - F(\boldsymbol{x}_{n-1})] = \mathbb{E}\langle \nabla F(\boldsymbol{x}_n), \boldsymbol{x}_n - \boldsymbol{x}_{n-1} \rangle$, effectively eliminating any additional error that Taylor approximation might incur.

A randomized approach such as ours is also recommended in the recent findings by Jordan et al. (2023), who demonstrated that randomization is *necessary* for achieving a dimension-free rate in non-smooth optimization. In particular, any deterministic algorithm suffers an additional dimension dependence of $\Omega(d)$.

Although employing exponential random scaling might seem unconventional, we justify this scaling by noting that $s_n \sim \mathrm{Exp}(1)$ satisfies $\mathbb{E}[s_n] = 1$ and $\mathbb{P}\{s_n \geq t\} = \exp(-t)$ (in particular, $\mathbb{P}\{s_n \leq 5\} \geq 0.99$). In other words, with high probability, the scaling factor behaves like a constant scaling on the update. To corroborate the efficacy of random scaling, we have conducted a series of empirical tests, the details of which are discussed in Section 6.

Exponentiated and Regularized Losses The most significant feature of Exponentiated O2NC (and from which it derives its name) is the loss function: $\ell_n(\Delta) = \langle \beta^{-n}\boldsymbol{g}_n, \Delta\rangle + \mathcal{R}_n(\Delta)$. This loss consists of two parts: intuitively, the exponentially upweighted linear loss $\langle \beta^{-n}\boldsymbol{g}_n, \Delta\rangle$ measures the "training progress" $F(\boldsymbol{x}_n) - F(\boldsymbol{x}_{n-1})$ (as discussed in Lemma 3.1), and $\mathcal{R}_n(\Delta)$ serves as an stabilizer that prevents the iterates from irregular behaviors. We will elaborate the role of each component later. To illustrate how Exponential O2NC works, let $\boldsymbol{u}_n$ be the optimal choice of Δ_n in hindsight. Then by minimizing the regret $\mathrm{Regret}_n(\boldsymbol{u}_n)$ w.r.t. $\boldsymbol{u}_n$, Algorithm 2 automatically chooses the best possible update Δ_n that is closest to $\boldsymbol{u}_n$.

Exponentially Weighted Gradients For now, set aside the regularizer $\mathcal{R}_n$ and focus on the linear term $\langle \beta^{-n}\boldsymbol{g}_n, \Delta\rangle$. To provide an intuition why we upweight the gradient by an exponential factor β^{-n}, we provide a brief overview for the convergence analysis of our algorithm. For simplicity of illustration, we assume $\boldsymbol{g}_n = \nabla F(\boldsymbol{x}_n)$ and $\mathcal{R}_n = 0$.

Let $S_n = \{\boldsymbol{x}_t\}_{t=1}^n$ and let $\boldsymbol{y}_n$ be distributed over S_n such that $\mathbb{P}\{\boldsymbol{y}_n = \boldsymbol{x}_t\} = p_{n,t} := \beta^{n-t} \cdot \frac{1-\beta}{1-\beta^n}$. Our strategy will be to show that this set S_n and random variable $\boldsymbol{y}_n$ satisfy the conditions to make $\overline{\boldsymbol{x}}_n$ a (c, ϵ) stationary point where $\overline{\boldsymbol{x}}_n$ is defined in Algorithm 2. To start, note that this distribution satisfies $\overline{\boldsymbol{x}}_n = \mathbb{E}[\boldsymbol{y}_n]$. Next, since there is always non-zero probability that $\boldsymbol{y}_n = \boldsymbol{x}_1$, it's not possible to obtain a deterministic bound of $\|\boldsymbol{y}_n - \overline{\boldsymbol{x}}_n\| \leq \delta$ for some small δ (as would be required if we were trying to establish (δ, ϵ) Goldstein stationarity). However, since $\boldsymbol{y}_n$ is exponentially more likely to be a later iterate (close to $\boldsymbol{x}_n$) than an early iterate (far from $\boldsymbol{x}_n$), intuitively $\mathbb{E}\|\boldsymbol{y}_n - \overline{\boldsymbol{x}}_n\|^2$ should not be too big. Formalizing this intuition forms a substantial part of our analysis.

In the convergence analysis, we will show $\overline{\boldsymbol{x}}$ is a (c, ϵ)-stationary point by bounding $\|\nabla F(\overline{\boldsymbol{x}}_n)\|_c$ (defined in Definition 2.2) for all n. The condition $\mathbb{E}[\boldsymbol{y}_n] = \overline{\boldsymbol{x}}_n$ is already satisfied by construction of $\boldsymbol{y}_n$, and it remains to bound the expected gradient $\|\mathbb{E}[\nabla F(\boldsymbol{y}_n)]\|$ and the variance $\mathbb{E}\|\boldsymbol{y}_n - \overline{\boldsymbol{x}}_n\|^2$. While the regularizer $\mathcal{R}_n$ is imposed to control the variance, the exponentiated gradients is employed to bound the expected gradient. In particular, this is achieved by reducing the difficult task of minimizing the expected gradient of a non-smooth non-convex objective to a relatively easier (and very heavily studied) one: minimizing the regret w.r.t. exponentiated losses $\ell_t(\Delta) = \langle \beta^{-t}\boldsymbol{g}_t, \Delta\rangle$. To elaborate further, let's consider a simplified illustration as follows.

Recall that $p_{n,t} = \beta^{n-t} \cdot \frac{1-\beta}{1-\beta^n}$. By construction of $\boldsymbol{y}_n$,

$$\mathbb{E}[\nabla F(\boldsymbol{y}_n)] = \textstyle\sum_{t=1}^n p_{n,t} \nabla F(\boldsymbol{x}_t).$$

Next, for each $n \in [N]$, we define

$$\boldsymbol{u}_n = -D \frac{\sum_{t=1}^n p_{n,t} \nabla F(\boldsymbol{x}_t)}{\|\sum_{t=1}^n p_{n,t} \nabla F(\boldsymbol{x}_t)\|} \tag{1}$$

for some D to be specified later. As a remark, $\boldsymbol{u}_n$ minimizes $\langle \mathbb{E}[\nabla F(\boldsymbol{y}_n)], \Delta\rangle$ over all possible Δ such that $\|\Delta\| = D$, therefore representing the optimal update in iterate n that leads to the fastest convergence.

With $\boldsymbol{u}_n$ defined in (1), it follows that

$$\begin{aligned}\frac{1}{D}\sum_{t=1}^n p_{n,t}\langle \nabla F(\boldsymbol{x}_t), -\boldsymbol{u}_n\rangle &= \left\|\sum_{t=1}^n p_{n,t} \nabla F(\boldsymbol{x}_t)\right\| \\ &= \|\mathbb{E}[\nabla F(\boldsymbol{y}_n)]\|.\end{aligned}$$

Recall that we assume $\boldsymbol{g}_t = \nabla F(\boldsymbol{x}_t)$ for simplicity. Moreover, in later convergence analysis, we will carefully show that $\sum_{n=1}^N \sum_{t=1}^n p_{n,t}\langle \nabla F(\boldsymbol{x}_t), -\Delta_t\rangle \lesssim 1 - \beta$ (see Appendix C). Consequently,

$$\begin{aligned}&\frac{1}{N}\sum_{n=1}^N \|\mathbb{E}\nabla F(\boldsymbol{y}_n)\| \\ &= \frac{1}{DN}\sum_{n=1}^N\sum_{t=1}^n p_{n,t}\langle \nabla F(\boldsymbol{x}_t), \Delta_t - \boldsymbol{u}_n\rangle \\ &\quad - \frac{1}{DN}\sum_{n=1}^N\sum_{t=1}^n p_{n,t}\langle \nabla F(\boldsymbol{x}_t), \Delta_t\rangle \\ &\lesssim \frac{1-\beta}{DN}\left(1 + \sum_{n=1}^N \beta^n \mathrm{Regret}_n(\boldsymbol{u}_n)\right).\end{aligned}$$

Here $\mathrm{Regret}_n(\boldsymbol{u}_n) = \sum_{t=1}^n \langle \beta^{-t}\boldsymbol{g}_t, \Delta_t - \boldsymbol{u}_n\rangle$ denotes the regret w.r.t. the exponentiated losses $\ell_t(\Delta) = \langle \beta^{-t}\boldsymbol{g}_t, \Delta\rangle$ for $t = 1, \ldots, n$ (assuming $\mathcal{R}_n = 0$) and comparator $\boldsymbol{u}_n$ defined in (1). Notably, the expected gradient is now bounded by the weighted average of the sequence of static regrets, $\mathrm{Regret}_n(\boldsymbol{u}_n)$. Consequently, a good OCO algorithm that effectively minimizes the regret is closely aligned with our goal of minimizing the expected gradient.

Variance Regularization As aforementioned, we impose the regularizer $\mathcal{R}_n(\Delta) = \frac{\mu_n}{2}\|\Delta\|^2$ to control the variance $\mathbb{E}\,\|\boldsymbol{y}_n - \overline{\boldsymbol{x}}_n\|^2$. Formally, the following result establishes a reduction from bounding variance to bounding the norm of Δ_t, thus motivating the choice of the regularizer.

Lemma 3.2. *For any $\beta \in (0,1)$,*

$$\mathbb{E}_s \sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{y}_n} \|\boldsymbol{y}_n - \overline{\boldsymbol{x}}_n\|^2 \leq \sum_{n=1}^{N} \frac{12}{(1-\beta)^2}\|\Delta_n\|^2.$$

This suggests that bounding $\|\Delta_n\|^2$ is sufficient to bound the variance of $\boldsymbol{y}_n$. Therefore, we impose the regularizer $\mathcal{R}_n(\Delta) = \frac{\mu_n}{2}\|\Delta\|^2$, for some constant μ_n to be determined later, to ensure that $\|\Delta_n\|^2$ remains small, effectively controlling the variance of $\boldsymbol{y}_n$.

Furthermore, we'd like to highlight that Lemma 3.2 provides a strictly better bound on the variance of $\boldsymbol{y}_n$ compared to the possible maximum deviation $\max \|\boldsymbol{y}_n - \overline{\boldsymbol{x}}_n\|$. For illustration, assume Δ_t's are orthonormal, then $\max \|\boldsymbol{y}_N - \overline{\boldsymbol{x}}_N\| \approx \|\boldsymbol{x}_1 - \boldsymbol{x}_N\| = O(N)$. On the other hand, Lemma 3.2 implies that for $n \sim \mathrm{Unif}([N])$, $\mathbb{E}_n[\mathrm{Var}(\boldsymbol{y}_n)] = O(\frac{1}{(1-\beta)^2})$. In particular, we will show that $1-\beta = N^{-1/2}$ when the objective is smooth. Consequently, $\mathbb{E}\,\|\boldsymbol{y}_n - \overline{\boldsymbol{x}}_n\| = O(\sqrt{N})$, which is strictly tighter than the deterministic bound of $\max \|\boldsymbol{y}_N - \overline{\boldsymbol{x}}_N\| = O(N)$.

This further motivates why we choose this specific distribution for $\boldsymbol{y}_n$: the algorithm does not need to be conservative all the time and can occasionally make relatively large step, breaking the deterministic constraint that $\|\boldsymbol{y}_n - \overline{\boldsymbol{x}}_n\| \leq \delta$, while still satisfying $\mathrm{Var}(\boldsymbol{y}_n) \leq \delta^2$.

3.1. Convergence Analysis

Now we present the main convergence theorem of Algorithm 2. This is a very general theorem, and we will prove the convergence bound of a more specific algorithm (Theorem 4.2) based on this result. A more formally stated version of this theorem and its proof can be found in Appendix C.

Theorem 3.3. *Follow Assumption 2.1. Let $\mathrm{Regret}_n(\boldsymbol{u}_n)$ denote the regret w.r.t. $\ell_t(\Delta) = \langle \beta^{-t}\boldsymbol{g}_t, \Delta\rangle + \mathcal{R}_t(\Delta)$ for $t = 1, \dots, n$ and comparator $\boldsymbol{u}_n$ defined in (1). Define $\mathcal{R}_t(\Delta) = \frac{\mu_t}{2}\|\Delta\|^2$, $\mu_t = \frac{24cD}{\alpha^2}\beta^{-t}$, and $\alpha = 1-\beta$, then*

$$\mathbb{E}\,\|\nabla F(\overline{\boldsymbol{x}})\|_c \lesssim \frac{F^*}{DN} + \frac{G+\sigma}{\alpha N} + \sigma\sqrt{\alpha} + \frac{cD^2}{\alpha^2} + \mathbb{E}\,\frac{\beta^{N+1}\mathrm{Regret}_N(\boldsymbol{u}_N) + \alpha\sum_{n=1}^{N}\beta^n \mathrm{Regret}_n(\boldsymbol{u}_n)}{DN}.$$

Here the second line denotes the weighted average of the sequence of static regrets, $\mathrm{Regret}_n(\boldsymbol{w}_n)$, w.r.t. the exponentiated and regularized loss $\ell_t(\Delta) = \langle \beta^{-t}\boldsymbol{g}_t, \Delta\rangle$ and comparator $\boldsymbol{u}_n$ defined in (1), as we discussed earlier. To see an immediate implication of Theorem 3.3, assume the average regret is no larger than the terms in the first line. Then by proper tuning $D = \frac{1}{\sqrt{\alpha N}}$ and $\alpha = \max\{\frac{1}{N^{2/3}}, \frac{c^{2/7}}{N^{4/7}}\}$, we have $\mathbb{E}\,\|\nabla F(\overline{\boldsymbol{w}})\|_c \lesssim \frac{1}{N^{1/3}} + \frac{c^{1/7}}{N^{2/7}}$.

4. Recovering SGDM: Exponentiated O2NC with OMD

In the previous sections, we have shown that Exponentiated O2NC can convert any OCO algorithm into a non-convex optimization algorithm in such a way that small regret bounds transform into convergence guarantees. So, the natural next step is to instantiate Exponentiated O2NC with some particular OCO algorithm. In this section we carry out this task and discover that the resulting method not only achieves optimal convergence guarantees, but is also *nearly identical* to the standard SGDM optimization algorithm!

The OCO algorithm we will use to instantiate Exponentiated O2NC is a simple variant of "online mirror descent" (OMD) (Beck & Teboulle, 2003), which a standard OCO algorithm. However, typical OMD analysis involves clipping the outputs Δ_n to lie in some pre-specified constraint set. We instead employ a minor modification to the standard algorithm to obviate the need for such clipping.

Inspired by (Duchi et al., 2010), we choose our OCO algorithm from the family of Online Mirror Descent (OMD) with composite loss. Given a sequence of gradients $\tilde{\boldsymbol{g}}_t := \beta^{-t}\boldsymbol{g}_t$ and convex functions $\psi_t(\Delta), \mathcal{R}_t(\Delta), \phi_t(\Delta)$, OMD with composite loss defines Δ_{t+1} as:

$$\arg\min_{\Delta} \langle \tilde{\boldsymbol{g}}_t, \Delta\rangle + D_{\psi_t}(\Delta, \Delta_t) + \underbrace{\mathcal{R}_{t+1}(\Delta) + \phi_t(\Delta)}_{\text{composite loss}}.$$

Here D_{ψ_t} denotes the Bregman divergence of ψ_t, and $\mathcal{R}_{t+1}(\Delta) + \phi_t(\Delta)$ denotes the composite loss. The composite loss consists of two components. Firstly, $\mathcal{R}_{t+1}(\Delta) = \frac{\mu_{t+1}}{2}\|\Delta\|^2$ controls the variance of $\boldsymbol{y}_n$, as discussed in Section 3. Secondly, OMD is known to struggle under unconstrained domain setting (Orabona & Pál, 2016), but this can be fixed with proper regularization, as done in Fang et al. (2021) (implicitly), and Jacobsen & Cutkosky (2022) (explicitly). Following a similar approach, we set $\phi_t(\Delta) = (\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t})\|\Delta\|^2$ to prevent the norm of $\|\Delta\|$ from being too large.

With $\psi_t(\Delta) = \frac{1}{2\eta_t}\|\Delta\|^2$ where $0 < \eta_{t+1} \leq \eta_t$, Theorem 4.1 provides a regret bound for this specific OCO algorithm.

Theorem 4.1. *Let $\Delta_1 = 0$ and $\Delta_{t+1} = \arg\min_\Delta \langle \tilde{\boldsymbol{g}}_t, \Delta\rangle +$*

$\frac{1}{2\eta_t}\|\Delta - \Delta_t\|^2 + \frac{\mu_{t+1}}{2}\|\Delta\|^2 + (\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t})\|\Delta\|^2$. *Then*

$$\begin{aligned}&\sum_{t=1}^{n}\langle \tilde{\boldsymbol{g}}_t, \Delta_t - \boldsymbol{u}\rangle + \mathcal{R}_t(\Delta_t) - \mathcal{R}_t(\boldsymbol{u}) \\ &\leq \left(\frac{2}{\eta_{n+1}} + \frac{\mu_{n+1}}{2}\right)\|\boldsymbol{u}\|^2 + \frac{1}{2}\sum_{t=1}^{n}\eta_t\|\tilde{\boldsymbol{g}}_t\|^2.\end{aligned}$$

Note that the implicit OMD update described in Theorem 4.1 can be explicitly represented as follows:

$$\Delta_{t+1} = \frac{\Delta_t - \eta_t\tilde{\boldsymbol{g}}_t}{1 + \eta_t\mu_{t+1} + \eta_t(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t})}. \tag{2}$$

When $\mathcal{R}_t = 0$ (implying $\mu_t = 0$), the update formula in (2) simplifies to an approximation of *online gradient descent* (Zinkevich, 2003), albeit with an additional scaling.

4.1. Reduction of Exponentiated O2NC

Upon substituting $\tilde{\boldsymbol{g}}_t = \beta^{-t}\boldsymbol{g}_t$ where $\boldsymbol{g}_t = \nabla f(\boldsymbol{x}_t, z_t)$, Theorem 4.1 provides a regret bound for $\mathrm{Regret}_n(\boldsymbol{u}_n)$ in the convergence bound in Theorem 3.3. Consequently, we can bound $\mathbb{E}\,\|\nabla F(\overline{\boldsymbol{x}})\|_c$ for Exponentiated O2NCwith the unconstrained variant of OMD as the OCO subroutine (with update formula described in (2)). Formally, we have the following result:

Theorem 4.2. *Follow Assumption 2.1 and consider any $c > 0$. Let $\Delta_1 = 0$ and update Δ_t by*

$$\Delta_{t+1} = \frac{\Delta_t - \eta_t\beta^{-t}\boldsymbol{g}_t}{1 + \eta_t\mu_{t+1} + \eta_t(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t})}.$$

Let $\mu_t = \beta^{-t}\mu$, $\eta_t = \beta^t\eta$, $\beta = 1 - \alpha$, $\mu = \frac{24F^*c}{(G+\sigma)\alpha^{5/2}N}$, $\eta = \frac{2F^*}{(G+\sigma)^2N}$, $\alpha = \max\{N^{-2/3}, \frac{(F^*)^{4/7}c^{2/7}}{(G+\sigma)^{6/7}N^{4/7}}\}$. Then for N large enough such that $\alpha \leq \frac{1}{2}$,

$$\mathbb{E}\,\|\nabla F(\overline{\boldsymbol{x}})\|_c \lesssim \frac{G+\sigma}{N^{1/3}} + \frac{(F^*)^{2/7}(G+\sigma)^{4/7}c^{1/7}}{N^{2/7}}.$$

As an immediate implication, upon solving $\mathbb{E}\,\|\nabla F(\overline{\boldsymbol{x}})\|_c \leq \epsilon$ for N, we conclude that Algorithm 2 instantiated with unconstrained OGD finds (c, ϵ)-stationary point within $N = O(\max\{(G+\sigma)^3\epsilon^{-3}, F^*(G+\sigma)^2c^{1/2}\epsilon^{-7/2}\})$ iterations. Moreover, in Section 5 we will show that this rate is optimal.

Furthermore, as discussed in Section 2, with $c = O(\epsilon^{-1})$, this algorithm achieves the optimal rate of $O(\epsilon^{-4})$ when F is smooth; with $c = O(1)$, this algorithm also achieves the optimal rate of $O(\epsilon^{-7/2})$ when F is second-order smooth. Remarkably, these optimal rates automatically follows from the reduction from (c, ϵ)-stationary point to ϵ-stationary point (see Lemma 2.3), and neither the algorithm nor the analysis is modified to achieve these rates.

4.2. Unraveling the update to discover SGDM

Furthermore, upon substituting the definition of η_t, μ_t (and neglecting constants G, σ, F^*), the update in Theorem 4.2 can be rewritten as

$$\Delta_{t+1} = \frac{\Delta_t - \eta\boldsymbol{g}_t}{1 + \frac{1}{\beta}(\eta\mu + \alpha)}$$

Let $\Delta_t = -\frac{\beta\eta}{\eta\mu+\alpha}\boldsymbol{m}_t$, then we can rewrite the update of Exponentiated O2NC with OGD as follows:

$$\begin{aligned}\boldsymbol{m}_{t+1} &= \frac{\beta}{1+\eta\mu}\boldsymbol{m}_t + \frac{\alpha+\eta\mu}{1+\eta\mu}\boldsymbol{g}_t, \\ \boldsymbol{x}_{t+1} &= \boldsymbol{x}_t - s_{n+1}\cdot\frac{\beta\eta}{\eta\mu+\alpha}\boldsymbol{m}_{t+1}.\end{aligned} \tag{3}$$

Remarkably, this update formula recovers the standard SGDM update, with the slight modification of an additional exponential random variable s_{n+1}: let $\tilde{\beta} = \frac{\beta}{1+\eta\mu}$, which denotes the effective momentum constant, and let $\tilde{\eta} = \frac{\beta\eta}{\eta\mu+\alpha}$ be the effective learning rate, then (4) becomes

$$\begin{aligned}\boldsymbol{m}_{t+1} &= \tilde{\beta}\boldsymbol{m}_t + (1-\tilde{\beta})\boldsymbol{g}_t, \\ \boldsymbol{x}_{t+1} &= \boldsymbol{x}_t - s_{t+1}\cdot\tilde{\eta}\boldsymbol{m}_{t+1}.\end{aligned} \tag{4}$$

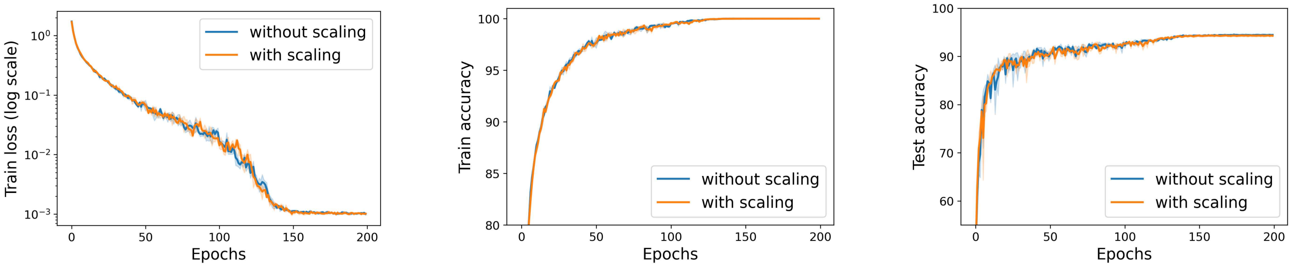
Smooth case As discussed earlier, when F is smooth, we set $c = O(\epsilon^{-1})$ to recover the optimal rate $N = O(\epsilon^{-4})$. This implies $c = O(N^{1/4})$. Consequently, we can check the parameters defined in Theorem 4.2 have order $\alpha = O(N^{-1/2})$, $\eta = O(N^{-1})$, and $\mu = O(N^{1/2})$ (note that $\eta\mu \approx \alpha$). Therefore, the effective momentum constant is roughly $\tilde{\beta} \approx 1 - \frac{1}{\sqrt{N}}$, and the effective learning rate is roughly $\tilde{\eta} \approx \frac{1}{\sqrt{N}}$. Interestingly, these values align with prior works (Cutkosky & Mehta, 2020).

Second-order smooth case When F is second-order smooth and we set $c = O(1)$, we can check that $\alpha = O(N^{-4/7})$, $\eta = O(N^{-1})$, and $\mu = O(N^{3/7})$ (again $\eta\mu \approx \alpha$). Consequently, the effective momentum should be set to $\tilde{\beta} \approx 1 - \frac{1}{N^{4/7}}$ and the effective learning rate should be $\tilde{\eta} \approx \frac{1}{N^{3/7}}$. It is interesting to note that in both smooth and second-order smooth cases, $(1-\tilde{\beta})\tilde{\eta} \approx \frac{1}{N}$.

5. Lower Bounds for finding (c, ϵ)-stationary points

In this section we leverage Lemma 2.3 to build a lower bound for finding (c, ϵ)-stationary points. Inuitively, Lemma 2.3 suggests that $O(c^{1/2}\epsilon^{-7/2})$ is the optimal rate for finding (c, ϵ)-stationary point. We can indeed prove its optimality using the lower bound construction by Arjevani et al. (2019) and Cutkosky et al. (2023).

Specifically, Arjevani et al. (2019) proved the following result: For any constants H, F^*, σ, ϵ, there exists objective



(a) Train Loss

(b) Train Accuracy

(c) Test Accuracy

Figure 1: Experiments on CIFAR-10 with ResNet-18 Network. The curves represent the average performance of each optimizer in three trials, and the shaded regions denote the standard deviation.

F and stochastic gradient estimator ∇f such that (i) F is H-smooth, $F(\boldsymbol{x}_0) - \inf F(\boldsymbol{x}) \leq F^*$, and $\mathbb{E}\,\|\nabla F(\boldsymbol{x}) - \nabla f(\boldsymbol{x}, z)\|^2 \leq \sigma^2$; and (ii) any randomized algorithm using ∇f requires $O(F^*\sigma^2 H\epsilon^{-4})$ iterations to find an ϵ-stationary point of F. As a caveat, such construction does not ensure that F is Lipschitz. Fortunately, Cutkosky et al. (2023) extended the lower bound construction so that the same lower holds and F is in addition $\sqrt{F^* H}$-Lipschitz.

Consequently, for any F^*, G, c, ϵ, define $H = \sqrt{c\epsilon}$ and $\sigma = G$ and assume $\sqrt{F^* H} \leq G$. Then by the lower bound construction, there exists F and $\mathcal{O}$ such that F is H-smooth, G-Lipschitz, $F(\boldsymbol{x}_0) - \inf F(\boldsymbol{x}) \leq F^*$, and $\mathbb{E}\,\|\nabla F(\boldsymbol{x}) - \nabla f(\boldsymbol{x}, z)\|^2 \leq G^2$. Lipschitzness and variance bound together imply $\mathbb{E}\,\|\nabla f(\boldsymbol{x}, z)\|^2 \leq 2G^2$. Moreover, finding an ϵ-stationary of F requires $\Omega(F^*\sigma^2 H\epsilon^{-4}) = \Omega(F^* G^2 c^{1/2}\epsilon^{-7/2})$ iterations (since $\sigma = G$, $H = \sqrt{c\epsilon}$).

Finally, note that $H = \sqrt{c\epsilon}$ satisfies $c = H^2\epsilon^{-1}$. Therefore by Lemma 2.3, a (c, ϵ)-stationary point of F is also an ϵ-stationary of F, implying that finding (c, ϵ)-stationary requires at least $\Omega(F^* G^2 c^{1/2}\epsilon^{-7/2})$ iterations as well. Putting these together, we have the following result:

Corollary 5.1. *For any F^*, c, ϵ and $G \geq \sqrt{F^*}(c\epsilon)^{1/4}$, there exists objective F and stochastic gradient ∇f such that (i) F is G-Lipschitz, $F(\boldsymbol{x}_0) - \inf F(\boldsymbol{x}) \leq F^*$, and $\mathbb{E}\,\|\nabla f(\boldsymbol{x}, z)\|^2 \leq 2G^2$; and (ii) any randomized algorithm using ∇f requires $\Omega(F^* G^2 c^{1/2}\epsilon^{-7/2})$ iterations to find (c, ϵ)-stationary point of F.*

6. Experiments

In the preceding sections, we theoretically demonstrated that scaling the learning rate by an exponential random variable s_n allows SGDM to satisfy convergence guarantees for non-smooth non-convex optimization. To validate this finding empirically, we implemented the SGDM algorithm with random scaling and assessed its performance against the standard SGDM optimizer without random scaling. Our evaluation involved the ResNet-18 model (He et al., 2016) on the CIFAR-10 image classification benchmark (Krizhevsky & Hinton, 2009). For the hyperparameters, we configured the learning rate at 0.01, the momentum constant at 0.9, and the weight decay at 5×10^{-4}. These settings are optimized for training the ResNet model on the CIFAR-10 dataset using SGDM. We use the same hyperparameters for our modified SGDM with random scaling.

For each optimizer, we ran the experiment three times under the same setting to minimize variability. We recorded the train loss, train accuracy, test loss, and test accuracy (refer to Figure 1). We also recorded the performance of the best iterate, e.g., the lowest train/test loss and the highest train/test accuracy, in each trial (see Table 1).

Table 1: Performance of the best iterate in each trial.

Random Scaling	No	Yes
Train loss ($\times 10^{-4}$)	9.82 ± 0.21	9.55 ± 0.37
Train accuracy (%)	100.0 ± 0.0	100.0 ± 0.0
Test loss ($\times 10^{-2}$)	21.6 ± 0.1	22.0 ± 0.4
Test accuracy (%)	94.6 ± 0.1	94.4 ± 0.2

These results show that the performance of SGDM with random scaling aligns closely with that of standard SGDM.

7. Conclusion

We introduced (c, ϵ)-stationary point, a relaxed definition of Goldstein stationary point, as a new notion of convergence criterion in non-smooth non-convex stochastic optimization. Furthermore, we proposed Exponentiated O2NC, a modified online-to-non-convex framework, by setting exponential random variable as scaling factor and adopting exponentiated and regularized loss. When applied with unconstrained online gradient descent, this framework produces an algorithm that recovers standard SGDM with random scaling and finds (c, ϵ)-stationary point within $O(c^{1/2}\epsilon^{-7/2})$ iterations. Notably, the algorithm automatically achieves the optimal rate of $O(\epsilon^{-4})$ for smooth objectives and $O(\epsilon^{-7/2})$ for second-order smooth objectives.

One interesting open problem is designing an adaptive algorithm with our Exponentiated O2NC framework. Since our framework, when applied with the simplest OCO algorithm online gradient descent, yields SGDM, a natural question emerges: what if we replace online gradient descent with an adaptive online learning algorithm, such as AdaGrad? Ideally, applied with AdaGrad as the OCO subroutine and with proper tuning, Exponentiated O2NC could recover Adam's update mechanism. However, the convergence analysis for this scenario is complex and demands a nuanced approach, especially considering the intricacies associated with the adaptive learning rate. In this vein, concurrent work by Ahn et al. (2024) applies a similar concept of online-to-non-convex conversion and connects the Adam algorithm to a principled online learning family known as Follow-The-Regularized-Leader (FTRL).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here

References

Ahn, K., Zhang, Z., Kook, Y., and Dai, Y. Understanding adam optimizer via online learning of updates: Adam is ftrl in disguise, 2024.

Allen-Zhu, Z. Natasha 2: Faster non-convex optimization than sgd. In *Advances in neural information processing systems*, pp. 2675–2686, 2018.

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*, 2019.

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Sekhari, A., and Sridharan, K. Second-order information in non-convex stochastic optimization: Power and limitations. In *Conference on Learning Theory*, pp. 242–299, 2020.

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, pp. 1–50, 2022.

Beck, A. and Teboulle, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. "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.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. Lower bounds for finding stationary points i. *Mathematical Programming*, pp. 1–50, 2019.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Chen, G. and Teboulle, M. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.

Cutkosky, A. and Mehta, H. Momentum improves normalized sgd. In *International Conference on Machine Learning*, 2020.

Cutkosky, A. and Orabona, F. Momentum-based variance reduction in non-convex sgd. In *Advances in Neural Information Processing Systems*, pp. 15210–15219, 2019.

Cutkosky, A., Mehta, H., and Orabona, F. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. In *International Conference on Machine Learning (ICML)*, 2023.

Davis, D. and Drusvyatskiy, D. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019. doi: 10.1137/18M1178244.

Duchi, J. C., Shalev-Shwartz, S., Singer, Y., and Tewari, A. Composite objective mirror descent. In *COLT*, volume 10, pp. 14–26. Citeseer, 2010.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pp. 689–699, 2018.

Fang, C., Lin, Z., and Zhang, T. Sharp analysis for nonconvex sgd escaping from saddle points. In *Conference on Learning Theory*, pp. 1192–1234, 2019.

Fang, H., Harvey, N. J. A., Portella, V. S., and Friedlander, M. P. Online mirror descent and dual averaging: keeping pace in the dynamic case, 2021.

Ghadimi, S. and Lan, G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Goldstein, A. A. Optimization of lipschitz continuous functions. *Math. Program.*, 13(1):14–22, dec 1977. ISSN 0025-5610. doi: 10.1007/BF01584320. URL https://doi.org/10.1007/BF01584320.

Hazan, E. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*, 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jacobsen, A. and Cutkosky, A. Parameter-free mirror descent. In Loh, P.-L. and Raginsky, M. (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 4160–4211. PMLR, 2022. URL https://proceedings.mlr.press/v178/jacobsen22a.html.

Jordan, M. I., Kornowski, G., Lin, T., Shamir, O., and Zampetakis, M. Deterministic nonsmooth nonconvex optimization, 2023.

Kingma, D. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kornowski, G. and Shamir, O. On the complexity of finding small subgradients in nonsmooth optimization. *arXiv preprint arXiv:2209.10346*, 2022a.

Kornowski, G. and Shamir, O. Oracle complexity in nonsmooth nonconvex optimization. *Journal of Machine Learning Research*, 23(314):1–44, 2022b.

Kornowski, G. and Shamir, O. An algorithm with optimal dimension-dependence for zero-order nonsmooth nonconvex stochastic optimization, 2023.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images, 2009.

Lin, T., Zheng, Z., and Jordan, M. I. Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization, 2022.

Mai, V. and Johansson, M. Convergence of a stochastic gradient method with momentum for non-smooth non-convex optimization. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6630–6639. PMLR, 13–18 Jul 2020.

Orabona, F. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Orabona, F. and Pál, D. Scale-free online learning. *arXiv preprint arXiv:1601.01974*, 2016.

Sutskever, I., Martens, J., Dahl, G., and Hinton, G. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pp. 1139–1147, 2013.

Tripuraneni, N., Stern, M., Jin, C., Regier, J., and Jordan, M. I. Stochastic cubic regularization for fast nonconvex optimization. In *Advances in neural information processing systems*, pp. 2899–2908, 2018.

You, Y., Gitman, I., and Ginsburg, B. Scaling sgd batch size to 32k for imagenet training. *arXiv preprint arXiv:1708.03888*, 6:12, 2017.

You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Keutzer, K., and Hsieh, C.-J. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019.

Zhang, J., Lin, H., Jegelka, S., Jadbabaie, A., and Sra, S. Complexity of finding stationary points of nonsmooth nonconvex functions. 2020.

Zhou, D., Xu, P., and Gu, Q. Stochastic nested variance reduction for nonconvex optimization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 3925–3936, 2018.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 928–936, 2003.

Ziyin, L., Wang, Z. T., and Ueda, M. Laprop: Separating momentum and adaptivity in adam, 2021.

A. Proofs in Section 2

A.1. Proof of Lemma 2.3

Lemma 2.3. *Suppose F is H-smooth. If $\|\nabla F(\boldsymbol{x})\|_c \le \epsilon$ where $c = H^2\epsilon^{-1}$, then $\|\nabla F(\boldsymbol{x})\| \le 2\epsilon$.*
Suppose F is ρ-second-order-smooth. If $\|\nabla F(\boldsymbol{x})\|_c \le \epsilon$ where $c = \rho/2$, then $\|\nabla F(\boldsymbol{x})\| \le 2\epsilon$.

Proof. Suppose $\|\nabla F(\boldsymbol{x})\|_c \le \epsilon$, then there exists $P \in \mathcal{P}(S), y \sim P$ such that $\mathbb{E}[\boldsymbol{y}] = \boldsymbol{x}$, $\|\mathbb{E}\nabla F(\boldsymbol{y})\| \le \epsilon$ and $c\,\mathbb{E}\|\boldsymbol{y}-\boldsymbol{x}\|^2 \le \epsilon$.

Assume F is H-smooth. By Jensen's inequality, $\mathbb{E}\|\boldsymbol{y}-\boldsymbol{x}\| \le \sqrt{\epsilon/c} = \epsilon/H$ with $c = H^2\epsilon^{-1}$. Consequently,

$$\begin{aligned}
\|\nabla F(\boldsymbol{x})\| &\le \|\mathbb{E}\nabla F(\boldsymbol{y})\| + \|\mathbb{E}[\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})]\| \\
&\le \|\mathbb{E}\nabla F(\boldsymbol{y})\| + \mathbb{E}\|\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})\| && \text{(Jensen's inequality)} \\
&\le \|\mathbb{E}\nabla F(\boldsymbol{y})\| + H\,\mathbb{E}\|\boldsymbol{x}-\boldsymbol{y}\| && \text{(smoothness)} \\
&\le \epsilon + H\cdot\epsilon/H = 2\epsilon.
\end{aligned}$$

Next, assume F is ρ-second-order smooth. By Taylor approximation, there exists some $\boldsymbol{z}$ such that $\nabla F(\boldsymbol{x}) = \nabla F(\boldsymbol{y}) + \nabla^2 F(\boldsymbol{x})(\boldsymbol{x}-\boldsymbol{y}) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{y})^T\nabla^3 F(\boldsymbol{z})(\boldsymbol{x}-\boldsymbol{y})$. Note that $\mathbb{E}[\nabla^2 F(\boldsymbol{x})(\boldsymbol{x}-\boldsymbol{y})] = \nabla^2 F(\boldsymbol{x})\,\mathbb{E}[\boldsymbol{x}-\boldsymbol{y}] = 0$. Consequently,

$$\begin{aligned}
\|\nabla F(\boldsymbol{x})\| &\le \|\mathbb{E}\nabla F(\boldsymbol{y})\| + \|\mathbb{E}[\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})]\| \\
&\le \|\mathbb{E}\nabla F(\boldsymbol{y})\| + \mathbb{E}\|\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{y})^T\nabla^3 F(\boldsymbol{z})(\boldsymbol{x}-\boldsymbol{y})\| && \text{(Jensen's inequality)} \\
&\le \|\mathbb{E}\nabla F(\boldsymbol{y})\| + \tfrac{\rho}{2}\,\mathbb{E}\|\boldsymbol{x}-\boldsymbol{y}\|^2 && \text{(second-order-smooth)} \\
&\le \epsilon + \tfrac{\rho}{2}\cdot\epsilon/c = 2\epsilon. && (c = \rho/2)
\end{aligned}$$

Together these prove the reduction from a (c,ϵ)-stationary point to an ϵ-stationary point. □

A.2. Proof of Lemma 2.4

Lemma 2.4. *Suppose F is G-Lipschitz. For any $c, \epsilon, \delta > 0$, a (c,ϵ)-stationary point is also a (δ,ϵ')-Goldstein stationary point where $\epsilon' = (1 + \frac{2G}{c\delta^2})\epsilon$.*

Proof. By definition of (c,ϵ)-stationary, there exists some distribution of $\boldsymbol{y}$ such that $\mathbb{E}[\boldsymbol{y}] = \boldsymbol{x}$, $\sigma^2 := \mathbb{E}\|\boldsymbol{y}-\boldsymbol{x}\|^2 \le \epsilon/c$, and $\|\mathbb{E}\nabla F(\boldsymbol{y})\| \le \epsilon$. By Chebyshev's inequality,

$$\begin{aligned}
\mathbb{P}\{\|\boldsymbol{y}-\boldsymbol{x}\| \ge \delta\} &= \mathbb{P}\left\{\|\boldsymbol{y}-\mathbb{E}[\boldsymbol{y}]\| \ge \frac{\delta}{\sigma}\cdot\sigma\right\} \\
&\le \mathbb{P}\left\{\|\boldsymbol{y}-\mathbb{E}[\boldsymbol{y}]\| \ge \frac{\delta}{\sqrt{\epsilon/c}}\cdot\sigma\right\} \le \frac{\epsilon}{c\delta^2}.
\end{aligned}$$

Next, we can construct a clipped random vector $\hat{\boldsymbol{y}}$ of $\boldsymbol{y}$ such that $\hat{\boldsymbol{y}} = \boldsymbol{y}$ if $\|\boldsymbol{y}-\boldsymbol{x}\| < \delta$, $\|\hat{\boldsymbol{y}}-\boldsymbol{x}\| \le \delta$ almost surely, and $\mathbb{E}[\hat{\boldsymbol{y}}] = \boldsymbol{x}$. In particular, note that $\mathbb{P}\{\hat{\boldsymbol{y}} \ne \boldsymbol{y}\} \le \mathbb{P}\{\|\boldsymbol{y}-\boldsymbol{x}\| \ge \delta\} \le \frac{\epsilon}{c\delta^2}$. Since F is G-Lipschitz,

$$\begin{aligned}
\|\mathbb{E}[\nabla F(\hat{\boldsymbol{y}}) - \nabla F(\boldsymbol{y})]\| &= \mathbb{P}\{\hat{\boldsymbol{y}} \ne \boldsymbol{y}\}\|\mathbb{E}[\nabla F(\hat{\boldsymbol{y}}) - \nabla F(\boldsymbol{y})|\hat{\boldsymbol{y}} \ne \boldsymbol{y}]\| \\
&\le 2G\cdot\mathbb{P}\{\hat{\boldsymbol{y}} \ne \boldsymbol{y}\} \le 2G\cdot\frac{\epsilon}{c\delta^2}.
\end{aligned}$$

Consequently $\|\mathbb{E}[\nabla F(\hat{\boldsymbol{y}})]\| \le \|\mathbb{E}[\nabla F(\boldsymbol{y})]\| + \|\mathbb{E}[\nabla F(\hat{\boldsymbol{y}}) - \nabla F(\boldsymbol{y})]\| \le \epsilon + \frac{2G\epsilon}{c\delta^2}$. This proves that x is also a $(\delta, \epsilon + \frac{2G\epsilon}{c\delta^2})$-Goldstein stationary point. □

B. Proofs in Section 3

B.1. Proof of Lemma 3.2

The proof consists of two composite lemmas. Recall the following notations: $S_n = \{\boldsymbol{x}_t\}_{t\in[n]}$, $\boldsymbol{y}_n \sim P_n$ where $P_n(\boldsymbol{x}_t) = \beta^{n-t} \cdot \frac{1-\beta}{1-\beta^n}$, and $\overline{\boldsymbol{x}}_n = \sum_{t=1}^n \beta^{n-t}\boldsymbol{x}_t \cdot \frac{1-\beta}{1-\beta^n}$. Also note two useful change of summation identities:

$$\sum_{n=1}^{N}\sum_{t=1}^{n} = \sum_{1\le t\le n\le N} = \sum_{t=1}^{N}\sum_{n=t}^{N}, \qquad\qquad \sum_{i=1}^{n}\sum_{i'=1}^{i-1}\sum_{t=i'+1}^{i} = \sum_{1\le i'<t\le i\le n} = \sum_{t=1}^{n}\sum_{i=t}^{n}\sum_{i'=1}^{t-1}.$$

Proposition B.1. $\mathbb{E}_{\boldsymbol{y}_n,s}\,\|\boldsymbol{y}_n - \overline{\boldsymbol{x}}_n\|^2 \le \sum_{t=1}^n \lambda_{n,t}\|\Delta_t\|^2$, *where*

$$\lambda_{n,t} = 4\sum_{i=t}^{n}\sum_{i'=1}^{t-1} p_{n,i}p_{n,i'}(i-i'), \quad p_{n,i} = P_n(\boldsymbol{x}_i) = \beta^{n-i}\cdot\frac{1-\beta}{1-\beta^n}. \tag{5}$$

Proof. By distribution of $\boldsymbol{y}_n$, we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{y}_n}\|\boldsymbol{y}_n - \overline{\boldsymbol{x}}_n\|^2 &= \sum_{i=1}^{n} p_{n,i}\|\boldsymbol{x}_i - \overline{\boldsymbol{x}}_n\|^2 \\ &= \sum_{i=1}^{n} p_{n,i}\left\|\sum_{i'=1}^{n} p_{n,i'}(\boldsymbol{x}_i - \boldsymbol{x}_{i'})\right\|^2 \\ &\le \sum_{i=1}^{n}\sum_{i'=1}^{n} p_{n,i}p_{n,i'}\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|^2 = 2\sum_{i=1}^{n}\sum_{i'=1}^{i-1} p_{n,i}p_{n,i'}\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|^2.\end{aligned}$$

The inequality uses convexity of $\|\cdot\|^2$. Next, upon unrolling the recursive update $\boldsymbol{x}_t = \boldsymbol{x}_{t-1} + \boldsymbol{s}_t\Delta_t$,

$$\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|^2 = \left\|\sum_{t=i'+1}^{i} s_t\Delta_t\right\|^2 \le (i-i')\sum_{t=i'+1}^{i} s_t^2\|\Delta_t\|^2.$$

Note that s_t and Δ_t are independent and $s_t \sim \mathrm{Exp}(1)$, so $\mathbb{E}_s[s_t^2\|\Delta_t\|^2] = \mathbb{E}_s[s_t^2]\|\Delta_t\|^2 = 2\|\Delta_t\|^p$. Consequently, upon substituting this back and applying change of summation, we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{y}_n,s}\|\boldsymbol{y}_n - \overline{\boldsymbol{x}}_n\|^2 &\le 4\sum_{i=1}^{n}\sum_{i'=1}^{i-1}\sum_{t=i'+1}^{i} p_{n,i}p_{n,i'}(i-i')\|\Delta_t\|^2 \\ &= \sum_{t=1}^{n}\left(4\sum_{i=t}^{n}\sum_{i'=1}^{t-1} p_{n,i}p_{n,i'}(i-i')\right)\|\Delta_t\|^2.\end{aligned}$$

We then conclude the proof by substituting the definition of $\lambda_{n,t}$. □

Proposition B.2. *Define* $\lambda_{n,t}$ *as in* (5)*, then* $\sum_{n=t}^N \lambda_{n,t} \le \frac{12}{(1-\beta)^2}$.

Proof. In the first part of the proof, we find a good upper bound of $\lambda_{n,t}$. We can rearrange the definition of $\lambda_{n,t}$ as follows.

$$\begin{aligned}\lambda_{n,t} &= 4\left(\frac{1-\beta}{1-\beta^n}\right)^2\sum_{i=t}^{n}\sum_{i'=1}^{t-1}\beta^{n-i}\beta^{n-i'}(i-i') && (\text{let } j = i-i') \\ &= 4\left(\frac{1-\beta}{1-\beta^n}\right)^2\sum_{i=t}^{n}\sum_{j=i-t+1}^{i-1}\beta^{n-i}\beta^{n-i+j}\cdot j && (\text{let } k = n-i) \\ &= 4\left(\frac{1-\beta}{1-\beta^n}\right)^2\sum_{k=0}^{n-t}\beta^{2k}\sum_{j=n-k-t+1}^{n-k-1} j\beta^j. && (6)\end{aligned}$$

The second line uses change of variable that $j = i - i'$, and the third line uses $k = n - i$. Next,

$$\begin{aligned}\sum_{j=n-k-t+1}^{n-k-1} j\beta^j = \beta \sum_{j=n-k-t+1}^{n-k-1} \frac{d}{d\beta}\beta^j &= \beta \cdot \frac{d}{d\beta}\left(\sum_{j=n-k-t+1}^{n-k-1} \beta^j\right) \\ &= \beta \cdot \frac{d}{d\beta}\left(\frac{\beta^{n-k-t+1} - \beta^{n-k}}{1-\beta}\right) \\ &= \frac{\beta^{a-k+1} - \beta^{b-k+1}}{(1-\beta)^2} + \frac{(a-k)\beta^{a-k} - (b-k)\beta^{b-k}}{1-\beta},\end{aligned}$$

where $a = n - t + 1, b = n$. Upon substituting this back into (6), we have

$$\begin{aligned}\lambda_{n,t} &= 4\left(\frac{1-\beta}{1-\beta^n}\right)^2 \sum_{k=0}^{n-t} \beta^{2k}\left(\frac{\beta^{a-k+1} - \beta^{b-k+1}}{(1-\beta)^2} + \frac{a\beta^{a-k} - b\beta^{b-k}}{1-\beta} - k\frac{\beta^{a-k} - \beta^{b-k}}{1-\beta}\right) \\ &= 4\left(\frac{1-\beta}{1-\beta^n}\right)^2 \sum_{k=0}^{n-t} \left(\frac{\beta^{a+1} - \beta^{b+1}}{(1-\beta)^2} + \frac{a\beta^a - b\beta^b}{1-\beta}\right)\beta^k - \frac{\beta^a - \beta^b}{1-\beta} \cdot k\beta^k. \qquad (7)\end{aligned}$$

For the first term, $\sum_{k=0}^{n-t} \beta^k = \frac{1-\beta^{n-t+1}}{1-\beta} = \frac{1-\beta^a}{1-\beta}$. For the second term,

$$\sum_{k=0}^{n-t} k\beta^k = \beta \cdot \frac{d}{d\beta}\left(\sum_{k=0}^{n-t} \beta^k\right) = \beta \cdot \frac{d}{d\beta}\left(\frac{1-\beta^a}{1-\beta}\right) = \frac{\beta - \beta^{a+1}}{(1-\beta)^2} - \frac{a\beta^a}{1-\beta}.$$

Upon substituting this back into (7) and simplifying the expression, we have

$$\begin{aligned}\lambda_{n,t} &= 4\left(\frac{1-\beta}{1-\beta^n}\right)^2 \cdot \left[\left(\frac{\beta^{a+1} - \beta^{b+1}}{(1-\beta)^2} + \frac{a\beta^a - b\beta^b}{1-\beta}\right) \cdot \frac{1-\beta^a}{1-\beta} - \frac{\beta^a - \beta^b}{1-\beta} \cdot \left(\frac{\beta - \beta^{a+1}}{(1-\beta)^2} - \frac{a\beta^a}{1-\beta}\right)\right] \\ &= 4\frac{(a\beta^a - b\beta^b)(1-\beta^a) + a\beta^a(\beta^a - \beta^b)}{(1-\beta^n)^2} = \ldots = 4\frac{a\beta^a(1-\beta^b) - b\beta^b(1-\beta^a)}{(1-\beta^n)^2}.\end{aligned}$$

Upon substituting $a = n - t + 1$ and $b = n$, we conclude the first half of the proof with

$$\lambda_{n,t} \le 4\frac{a\beta^a(1-\beta^b)}{(1-\beta^n)^2} \le 4 \cdot \frac{(n-t+1)\beta^{n-t+1}}{1-\beta^n}.$$

In the second part, we use this inequality to bound $\sum_{n=t}^N \lambda_{n,t}$. Define $K = \lceil \frac{1}{1-\beta} \rceil$, then

$$\sum_{n=t}^N \lambda_{n,t} = \mathbb{1}_{\{t \le K-1\}} \cdot \sum_{n=t}^{K-1} \lambda_{n,t} + \sum_{n=\max\{t,K\}}^N \lambda_{n,t}. \qquad (8)$$

For the first summation in (8), for all $t \le n \le K - 1$, we have

$$\lambda_{n,t} \le 4 \cdot \frac{(n-t+1)\beta^{n-t+1}}{1-\beta^n} \overset{(i)}{\le} 4 \cdot \frac{(n-t+1)\beta^{n-t+1}}{1-\beta^{n-t+1}} \overset{(ii)}{\le} 4 \cdot \frac{1 \cdot \beta^1}{1-\beta^1} \le \frac{4}{1-\beta}.$$

(i) holds because $\frac{1}{1-\beta^n}$ is decreasing w.r.t. n. (ii) holds because $f(x) = \frac{x\beta^x}{1-\beta^x}$ is decreasing for $x \ge 0$ and $\beta \in (0,1)$, so $f(n-t+1) \le f(1)$ since $n - t + 1 \ge 1$. Recall that $K - 1 \le \frac{1}{1-\beta}$, then the first summation in (8) can be bounded by

$$\mathbb{1}_{\{t \le K-1\}} \cdot \sum_{n=t}^{K-1} \lambda_{n,t} \le \sum_{n=1}^{K-1} \frac{4}{1-\beta} \le \frac{4}{(1-\beta)^2}. \qquad (9)$$

For the second summation in (8), for all $n \ge K \ge \frac{1}{1-\beta}$,

$$\frac{1}{1-\beta^n} \overset{(i)}{\le} \frac{1}{1-\beta^{\frac{1}{1-\beta}}} \overset{(ii)}{\le} \lim_{x \to 1} \frac{1}{1-x^{\frac{1}{1-x}}} = \frac{e}{e-1} \le 2.$$

(i) holds because $\frac{1}{1-\beta^n}$ is decreasing. (ii) holds because $f(x) = \frac{1}{1-x^{\frac{1}{1-x}}}$ is increasing for $x \geq 0$, so $f(\beta) \leq \lim_{x\to 1} f(x)$ for all $\beta \in (0,1)$. Consequently, the second summation in (8) can be bounded by

$$\sum_{n=\max\{t,K\}}^{N} \lambda_{n,t} \leq \sum_{n=\max\{t,K\}}^{N} 4 \cdot \frac{(n-t+1)\beta^{n-t+1}}{1-\beta^n} \leq 8 \sum_{n=t}^{N} (n-t+1)\beta^{n-t+1} = 8 \sum_{n=1}^{N-t} n\beta^n \tag{10}$$

By change of summation,

$$\sum_{n=1}^{N} n\beta^n = \sum_{n=1}^{N}\sum_{i=1}^{n} \beta^n = \sum_{i=1}^{N}\sum_{n=i}^{N} \beta^n \leq \sum_{i=1}^{N} \frac{\beta^i}{1-\beta} \leq \frac{1}{(1-\beta)^2}.$$

We then conclude the proof by substituting (9), (10) into (8). □

Lemma 3.2. *For any $\beta \in (0,1)$,*

$$\mathbb{E}_s \sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{y}_n} \|\boldsymbol{y}_n - \overline{\boldsymbol{x}}_n\|^2 \leq \sum_{n=1}^{N} \frac{12}{(1-\beta)^2} \|\Delta_n\|^2.$$

Proof. By Proposition B.1 and Proposition B.2, we have

$$\mathbb{E}_s \sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{y}_n} \|\boldsymbol{y}_n - \overline{\boldsymbol{x}}_n\|^2 \overset{(i)}{\leq} \sum_{n=1}^{N}\sum_{t=1}^{n} \lambda_{n,t} \|\Delta_t\|^2 \overset{(ii)}{=} \sum_{t=1}^{N} \left(\sum_{n=t}^{N} \lambda_{n,t} \right) \|\Delta_t\|^2 \overset{(iii)}{\leq} \sum_{t=1}^{N} \frac{12}{(1-\beta)^2} \|\Delta_t\|^2.$$

Here (i) is from Proposition B.1, (ii) is from change of summation, and (iii) is from Proposition B.2. □

B.2. Proof of Lemma 3.1

Lemma 3.1. *Let $s \sim \mathrm{Exp}(\lambda)$ for some $\lambda > 0$, then*

$$\mathbb{E}_s[F(\boldsymbol{x} + s\Delta) - F(\boldsymbol{x})] = \mathbb{E}_s[\langle \nabla F(\boldsymbol{x} + s\Delta), \Delta \rangle]/\lambda.$$

Proof. Denote $p(s) = \lambda \exp(-\lambda s)$ as the pdf of s. Upon expanding the expectation, we can rewrite the LHS as

$$\begin{aligned}
\mathbb{E}_s[F(\boldsymbol{x} + s\Delta) - F(\boldsymbol{x})] &= \int_0^\infty [F(\boldsymbol{x} + s\Delta) - F(\boldsymbol{x})] p(s)\, ds \\
&\overset{(i)}{=} \int_0^\infty \left(\int_0^s \langle \nabla F(\boldsymbol{x} + t\Delta), \Delta \rangle \, dt \right) p(s)\, ds \\
&= \int_0^\infty \int_0^\infty \langle \nabla F(\boldsymbol{x} + t\Delta), \Delta \rangle \mathbb{1}\{t \leq s\} p(s)\, dt ds \\
&= \int_0^\infty \left(\int_t^\infty p(s)\, ds \right) \langle \nabla F(\boldsymbol{x} + t\Delta), \Delta \rangle \, dt \\
&\overset{(ii)}{=} \int_0^\infty \frac{p(t)}{\lambda} \langle \nabla F(\boldsymbol{x} + t\Delta), \Delta \rangle \, dt \\
&= \frac{1}{\lambda} \mathbb{E}_s[\langle \nabla F(\boldsymbol{x} + s\Delta), \Delta \rangle].
\end{aligned}$$

Here the (i) applies fundamental theorem of calculus on $g(s) = F(\boldsymbol{x} + s\Delta) - F(\boldsymbol{x})$ with $g'(s) = \langle \nabla F(\boldsymbol{x} + s\Delta), \Delta \rangle$ and (ii) uses the following identity for exponential distribution: $\int_t^\infty p(s) ds = \exp(-\lambda t) = p(t)/\lambda$. □

C. Proof of Theorem 3.3

We restate the formal version of Theorem 3.3 as follows. Recall that $S_n = \{\boldsymbol{x}_t\}_{t \in [n]}$, $\boldsymbol{y}_n \sim P_n$ where $P_n(\boldsymbol{x}_t) = \beta^{n-t} \cdot \frac{1-\beta}{1-\beta^n}$, and $\overline{\boldsymbol{x}}_n = \sum_{t=1}^{n} \beta^{n-t} \boldsymbol{x}_t \cdot \frac{1-\beta}{1-\beta^n}$.

Theorem C.1. *Suppose F is G-Lipschitz, $F(\boldsymbol{x}_0) - \inf F(\boldsymbol{x}) \le F^*$, and the stochastic gradients satisfy $\mathbb{E}[\nabla f(\boldsymbol{x}, z) \mid \boldsymbol{x}] = \nabla F(\boldsymbol{x})$ and $\mathbb{E}\|\nabla F(\boldsymbol{x}) - \nabla f(\boldsymbol{x}, z)\|^2 \le \sigma^2$ for all $\boldsymbol{x}, z$. Define the comparator $\boldsymbol{u}_n$ and the regret $\operatorname{Regret}_n(\boldsymbol{u})$ of the regularized losses ℓ_t as follows:*

$$\boldsymbol{u}_n = -D \cdot \frac{\sum_{t=1}^n \beta^{n-t}\nabla F(\boldsymbol{x}_t)}{\|\sum_{t=1}^n \beta^{n-t}\nabla F(\boldsymbol{x}_t)\|}, \qquad \operatorname{Regret}_n(\boldsymbol{u}) = \sum_{t=1}^n \langle \beta^{-t}\boldsymbol{g}_t, \Delta_t - \boldsymbol{u}\rangle + \mathcal{R}_t(\Delta_t) - \mathcal{R}_t(\boldsymbol{u}).$$

Also define the regularizor as $\mathcal{R}_t(\boldsymbol{w}) = \frac{\mu_t}{2}\|\boldsymbol{w}\|^2$ where $\mu_t = \mu\beta^{-t}$, $\mu = \frac{24cD}{\alpha^2}$ and $\alpha = 1-\beta$. Then

$$\mathbb{E}\|\nabla F(\overline{\boldsymbol{x}})\|_c \le \frac{F^*}{DN} + \frac{2G+\sigma}{\alpha N} + \sigma\sqrt{\alpha} + \frac{12cD^2}{\alpha^2} + \frac{1}{DN}\left(\beta^{N+1}\,\mathbb{E}\operatorname{Regret}_N(\boldsymbol{u}_N) + \alpha\sum_{n=1}^N \beta^n\,\mathbb{E}\operatorname{Regret}_n(\boldsymbol{u}_n).\right).$$

Proof. We start with the change of summation. Note that

$$\begin{aligned}\sum_{n=1}^N\sum_{t=1}^n \beta^{n-t}(1-\beta)(F(\boldsymbol{x}_t) - F(\boldsymbol{x}_{t-1})) &= \sum_{t=1}^N\left(\sum_{n=t}^N \beta^{n-t}\right)(1-\beta)(F(\boldsymbol{x}_t) - F(\boldsymbol{x}_{t-1}))\\ &= \sum_{t=1}^N (1-\beta^{N-t+1})(F(\boldsymbol{x}_t) - F(\boldsymbol{x}_{t-1}))\\ &= F(\boldsymbol{x}_N) - F(\boldsymbol{x}_0) - \sum_{t=1}^N \beta^{N-t+1}(F(\boldsymbol{x}_t) - F(\boldsymbol{x}_{t-1})).\end{aligned}$$

Upon rearranging and applying the assumption that $F(\boldsymbol{x}_0) - F(\boldsymbol{x}_N) \le F(\boldsymbol{x}_0) - \inf F(\boldsymbol{x}) \le F^*$, we have

$$-F^* \le \mathbb{E}\sum_{n=1}^N\sum_{t=1}^n \beta^{n-t}(1-\beta)(F(\boldsymbol{x}_t) - F(\boldsymbol{x}_{t-1})) + \mathbb{E}\sum_{t=1}^N \beta^{N-t+1}(F(\boldsymbol{x}_t) - F(\boldsymbol{x}_{t-1})). \tag{11}$$

First, we bound the first summation in (11). Denote $\mathcal{F}_t$ as the σ-algebra of $\boldsymbol{x}_t$. Note that $\Delta_t \in \mathcal{F}_t$ and $z_t \notin \mathcal{F}_t$, so by the assumption that $\mathbb{E}[\nabla f(\boldsymbol{x}, z) \mid \boldsymbol{x}] = \nabla F(\boldsymbol{x})$,

$$\mathbb{E}[\boldsymbol{g}_t \mid \mathcal{F}_t] = \mathbb{E}[\nabla f(\boldsymbol{x}_t, z_t) \mid \mathcal{F}_t] = \nabla F(\boldsymbol{x}_t) \implies \mathbb{E}\langle \nabla F(\boldsymbol{x}_t), \Delta_t\rangle = \mathbb{E}\langle \boldsymbol{g}_t, \Delta_t\rangle.$$

By Lemma 3.1, $\mathbb{E}[F(\boldsymbol{x}_t) - F(\boldsymbol{x}_{t-1})] = \mathbb{E}\langle \nabla F(\boldsymbol{x}_t), \Delta_t\rangle$. Upon adding and subtracting, we have

$$\begin{aligned}\mathbb{E}[F(\boldsymbol{x}_t) - F(\boldsymbol{x}_{t-1})] &= \mathbb{E}\langle \nabla F(\boldsymbol{x}_t) - \boldsymbol{g}_t + \boldsymbol{g}_t, \Delta_t - \boldsymbol{u}_n + \boldsymbol{u}_n\rangle\\ &= \mathbb{E}\left[\langle \nabla F(\boldsymbol{x}_t), \boldsymbol{u}_n\rangle\rangle + \langle \nabla F(\boldsymbol{x}_t) - \boldsymbol{g}_t, -\boldsymbol{u}_n\rangle + \langle \boldsymbol{g}_t, \Delta_t - \boldsymbol{u}_n\rangle\right].\end{aligned}$$

Consequently, the first summation in (11) can be written as

$$\mathbb{E}\sum_{n=1}^N\sum_{t=1}^n \beta^{n-t}(1-\beta)\left(\langle \nabla F(\boldsymbol{x}_t), \boldsymbol{u}_n\rangle + \langle \nabla F(\boldsymbol{x}_t) - \boldsymbol{g}_t, -\boldsymbol{u}_n\rangle + \langle \boldsymbol{g}_t, \Delta_t - \boldsymbol{u}_n\rangle\right). \tag{12}$$

For the first term, upon substituting the definition of $\boldsymbol{u}_n$, we have

$$\begin{aligned}\sum_{t=1}^n \beta^{n-t}(1-\beta)\langle \nabla F(\boldsymbol{x}_t), \boldsymbol{u}_n\rangle &= (1-\beta)\left\langle \sum_{t=1}^n \beta^{n-t}\nabla F(\boldsymbol{x}_t), -D\frac{\sum_{t=1}^n \beta^{n-t}\nabla F(\boldsymbol{x}_t)}{\|\sum_{t=1}^n \beta^{n-t}\nabla F(\boldsymbol{x}_t)\|}\right\rangle\\ &= (1-\beta^n)\cdot -D\left\|\frac{\sum_{t=1}^n \beta^{n-t}\nabla F(\boldsymbol{x}_t)}{\sum_{t=1}^n \beta^{n-t}}\right\|\\ &= -D(1-\beta^n)\|\mathbb{E}_{\boldsymbol{y}_n}\nabla F(\boldsymbol{y}_n)\|\end{aligned}$$

Since $\|\nabla F(\boldsymbol{x}_t)\| \le G$ for all t, $\|\mathbb{E}_{\boldsymbol{y}_n}\nabla F(\boldsymbol{y}_n)\| \le G$ as well. Therefore, we have

$$\le -D\|\mathbb{E}_{\boldsymbol{y}_n}\nabla F(\boldsymbol{y}_n)\| + DG\beta^n.$$

Since $\beta < 1$, $\sum_{n=1}^N \beta^n \le \frac{1}{1-\beta}$. Therefore, upon summing over n, the first term in (12) becomes

$$\mathbb{E}\sum_{n=1}^N\sum_{t=1}^n \beta^{n-t}(1-\beta)\langle \nabla F(\boldsymbol{x}_t), \boldsymbol{u}_n\rangle \le \left(-D\sum_{n=1}^N \mathbb{E}\,\|\mathbb{E}_{\boldsymbol{y}_n}\nabla F(\boldsymbol{y}_n)\|\right) + \frac{DG}{1-\beta}. \tag{13}$$

For the second term, by Cauchy-Schwarz inequality,

$$\mathbb{E}\sum_{t=1}^n \beta^{n-t}\langle \nabla F(\boldsymbol{x}_t) - \boldsymbol{g}_t, -\boldsymbol{u}_n\rangle \le \sqrt{\mathbb{E}\left\|\sum_{t=1}^n \beta^{n-t}(\nabla F(\boldsymbol{x}_t) - \boldsymbol{g}_t)\right\|^2 \mathbb{E}\,\|\boldsymbol{u}_n\|^2}.$$

Since $\mathbb{E}[\nabla F(\boldsymbol{x}_t) - \boldsymbol{g}_t \,|\, \mathcal{F}_t] = 0$, by martingale identity and the assumption that $\mathbb{E}\,\|\nabla F(\boldsymbol{x}) - \nabla f(\boldsymbol{x}, z)\|^2 \le \sigma^2$,

$$\mathbb{E}\left\|\sum_{t=1}^n \beta^{n-t}(\nabla F(\boldsymbol{x}_t) - \boldsymbol{g}_t)\right\|^2 = \sum_{t=1}^n \mathbb{E}\,\|\beta^{n-t}(\nabla F(\boldsymbol{x}_t) - \boldsymbol{g}_t)\|^2 \le \sum_{t=1}^n \sigma^2\beta^{2(n-t)} \le \frac{\sigma^2}{1-\beta^2}.$$

Upon substituting $\|\boldsymbol{u}_n\| = D$ and $\frac{1}{1-\beta^2} \le \frac{1}{1-\beta}$, the second term in (12) becomes

$$\mathbb{E}\sum_{n=1}^N\sum_{t=1}^n \beta^{n-t}(1-\beta)\langle \nabla F(\boldsymbol{x}_t) - \boldsymbol{g}_t, -\boldsymbol{u}_n\rangle \le \sum_{n=1}^N (1-\beta)\cdot\frac{\sigma D}{\sqrt{1-\beta^2}} \le \sigma D N\sqrt{1-\beta}. \tag{14}$$

For the third term, upon adding and subtracting $\mathcal{R}_t$ and substituting the definition of $\mathrm{Regret}_n(\boldsymbol{u})$, we have

$$\begin{aligned}
&\mathbb{E}\sum_{n=1}^N\sum_{t=1}^n \beta^{n-t}(1-\beta)\langle \boldsymbol{g}_t, \Delta_t - \boldsymbol{u}_n\rangle \\
&= \mathbb{E}\sum_{n=1}^N\sum_{t=1}^n (1-\beta)\beta^n\left(\langle \beta^{-t}\boldsymbol{g}_t, \Delta_t - \boldsymbol{u}_n\rangle + \mathcal{R}_t(\Delta_t) - \mathcal{R}_t(\boldsymbol{u}_n) - \mathcal{R}_t(\Delta_t) + \mathcal{R}_t(\boldsymbol{u}_n)\right) \\
&= \mathbb{E}\sum_{n=1}^N (1-\beta)\beta^n \mathrm{Regret}_n(\boldsymbol{u}_n) + \mathbb{E}\sum_{n=1}^N\sum_{t=1}^n (1-\beta)\beta^n(-\mathcal{R}_t(\Delta_t) + \mathcal{R}_t(\boldsymbol{u}_n)).
\end{aligned} \tag{15}$$

Upon substituting (13), (14) and (15) into (12), the first summation in (11) becomes

$$\begin{aligned}
&\sum_{n=1}^N\sum_{t=1}^n \beta^{n-t}(1-\beta)(F(\boldsymbol{x}_t) - F(\boldsymbol{x}_{t-1})) \\
&\le \left(-D\sum_{n=1}^N \mathbb{E}\,\|\mathbb{E}_{\boldsymbol{y}_n}\nabla F(\boldsymbol{y}_n)\|\right) + \frac{DG}{1-\beta} + \sigma D N\sqrt{1-\beta} \\
&\quad + \mathbb{E}\sum_{n=1}^N (1-\beta)\beta^n \mathrm{Regret}_n(\boldsymbol{u}_n) + \mathbb{E}\sum_{n=1}^N\sum_{t=1}^n (1-\beta)\beta^n(-\mathcal{R}_t(\Delta_t) + \mathcal{R}_t(\boldsymbol{u}_n)).
\end{aligned} \tag{16}$$

Next, we consider the second summation in (11). Since $\mathbb{E}\,\|\boldsymbol{g}_t\| \le \mathbb{E}\,\|\nabla F(\boldsymbol{x}_t)\| + \mathbb{E}\,\|\nabla F(\boldsymbol{x}_t) - \boldsymbol{g}_t\| \le G + \sigma$ and $\mathbb{E}\langle \nabla F(\boldsymbol{x}_t), \Delta_t\rangle = \mathbb{E}\langle \boldsymbol{g}_t, \Delta_t\rangle$, we have

$$\begin{aligned}
\mathbb{E}[F(\boldsymbol{x}_t) - F(\boldsymbol{x}_{t-1})] = \mathbb{E}\langle \nabla F(\boldsymbol{x}_t), \Delta_t\rangle &= \mathbb{E}\langle \boldsymbol{g}_t, \Delta_t - \boldsymbol{u}_N\rangle + \mathbb{E}\langle \boldsymbol{g}_t, \boldsymbol{u}_N\rangle \\
&\le \mathbb{E}\langle \boldsymbol{g}_t, \Delta_t - \boldsymbol{u}_N\rangle + D(G+\sigma).
\end{aligned}$$

Following the same argument in (15) by adding and subtracting $\mathcal{R}_t$, the second summation becomes

$$\begin{aligned}
\mathbb{E}\sum_{t=1}^N \beta^{N-t+1}(F(\boldsymbol{x}_t) - F(\boldsymbol{x}_{t-1})) &= \mathbb{E}\sum_{t=1}^N \beta^{N+1}\langle \beta^{-t}\boldsymbol{g}_t, \Delta_t - \boldsymbol{u}_N\rangle + \beta^{N-t+1}D(G+\sigma) \\
&\le \beta^{N+1}\,\mathbb{E}\,\mathrm{Regret}_N(\boldsymbol{u}_N) + \frac{D(G+\sigma)}{1-\beta} + \mathbb{E}\sum_{t=1}^N \beta^{N+1}(-\mathcal{R}_t(\Delta_t) + \mathcal{R}_t(\boldsymbol{u}_N)).
\end{aligned} \tag{17}$$

Combining (16) and (17) into (11) gives

$$
\begin{aligned}
-F^* \leq & \left(-D\sum_{n=1}^{N}\mathbb{E}\,\|\mathbb{E}_{\boldsymbol{y}_n}\nabla F(\boldsymbol{y}_n)\|\right) + \frac{DG}{1-\beta} + \sigma DN\sqrt{1-\beta} \\
& + \mathbb{E}\sum_{n=1}^{N}(1-\beta)\beta^n \mathrm{Regret}_n(\boldsymbol{u}_n) + \mathbb{E}\sum_{n=1}^{N}\sum_{t=1}^{n}(1-\beta)\beta^n(-\mathcal{R}_t(\Delta_t)+\mathcal{R}_t(\boldsymbol{u}_N)) \\
& + \beta^{N+1}\,\mathbb{E}\,\mathrm{Regret}_N(\boldsymbol{u}_N) + \frac{D(G+\sigma)}{1-\beta} + \mathbb{E}\sum_{t=1}^{N}\beta^{N+1}(-\mathcal{R}_t(\Delta_t)+\mathcal{R}_t(\boldsymbol{u}_N)).
\end{aligned} \tag{18}
$$

As the final step, we simplify the terms involving $\mathcal{R}_t$. Recall that $\mathcal{R}_t(\boldsymbol{w}) = \frac{\mu_t}{2}\|\boldsymbol{w}\|^2$, so $\mathcal{R}_t(\boldsymbol{u}_n) = \frac{\mu_t}{2}D^2$ is independent of n. Hence, by change of summation,

$$
\begin{aligned}
& \mathbb{E}\sum_{n=1}^{N}\sum_{t=1}^{n}(1-\beta)\beta^n(-\mathcal{R}_t(\Delta_t)+\mathcal{R}_t(\boldsymbol{u}_n)) + \mathbb{E}\sum_{t=1}^{N}\beta^{N+1}(-\mathcal{R}_t(\Delta_t)+\mathcal{R}_t(\boldsymbol{u}_N)) \\
&= \mathbb{E}\sum_{t=1}^{N}\underbrace{\left(\sum_{n=t}^{N}\beta^n\right)(1-\beta)}_{=\beta^t-\beta^{N+1}}\left(-\frac{\mu_t}{2}\|\Delta_t\|^2 + \frac{\mu_t}{2}D^2\right) + \mathbb{E}\sum_{t=1}^{N}\beta^{N+1}\left(-\frac{\mu_t}{2}\|\Delta_t\|^2+\frac{\mu_t}{2}D^2\right) \\
&= \mathbb{E}\sum_{t=1}^{N}\beta^t\left(-\frac{\mu_t}{2}\|\Delta_t\|^2+\frac{\mu_t}{2}D^2\right)
\end{aligned}
$$

Recall Lemma 3.2 that $\mathbb{E}\sum_{n=1}^{N}\mathbb{E}_{\boldsymbol{y}_n}\|\boldsymbol{y}_n - \overline{\boldsymbol{x}}_n\|^2 \leq \mathbb{E}\sum_{t=1}^{N}\frac{12}{(1-\beta)^2}\|\Delta_t\|^2$. Upon substituting $\mu_t = \frac{24cD^2}{(1-\beta)^2}\beta^{-t}$, we have

$$
\begin{aligned}
&= \mathbb{E}\sum_{t=1}^{N}\left(-\frac{12cD}{(1-\beta)^2}\|\Delta_t\|^2 + \frac{12cD^3}{(1-\beta)^2}\right) \\
&\leq \left(-cD\,\mathbb{E}\sum_{n=1}^{N}\mathop{\mathbb{E}}_{\boldsymbol{y}_n}\|\boldsymbol{y}_n-\overline{\boldsymbol{x}}_n\|^2\right) + \frac{12cD^3N}{(1-\beta)^2}.
\end{aligned}
$$

Substituting this back into (18) with $\alpha = 1-\beta$, we have

$$
\begin{aligned}
-F^* \leq & -D\,\mathbb{E}\left[\sum_{n=1}^{N}\|\mathbb{E}_{\boldsymbol{y}_n}\nabla F(\boldsymbol{y}_n)\| + c\cdot\mathbb{E}_{\boldsymbol{y}_n}\|\boldsymbol{y}_n-\overline{\boldsymbol{x}}_n\|^2\right] + \frac{DG}{\alpha} + \sigma DN\sqrt{\alpha} + \frac{D(G+\sigma)}{\alpha} + \frac{12cD^3N}{\alpha^2} \\
& + \beta^{N+1}\,\mathbb{E}\,\mathrm{Regret}_N(\boldsymbol{u}_N) + \alpha\sum_{n=1}^{N}\beta^n\,\mathbb{E}\,\mathrm{Regret}_n(\boldsymbol{u}_n).
\end{aligned}
$$

By definition of $\|\nabla F(\cdot)\|_c$ defined in Definition 2.2, $\|\nabla F(\overline{\boldsymbol{x}}_n)\|_c \leq \|\mathbb{E}_{\boldsymbol{y}_n}\nabla F(\boldsymbol{y}_n)\| + c\cdot\mathbb{E}_{\boldsymbol{y}_n}\|\boldsymbol{y}_n-\overline{\boldsymbol{x}}_n\|^2$. Moreover, since $\overline{\boldsymbol{x}}$ is uniform over $\overline{\boldsymbol{x}}_n$, $\mathbb{E}\,\|\nabla F(\overline{\boldsymbol{x}})\|_{2,c} = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\,\|\nabla F(\overline{\boldsymbol{x}}_n)\|_{2,c}$ We then conclude the proof by rearranging the equation and dividing both sides by DN. □

D. Proofs in Section 4

D.1. Proof of Theorem 4.1

Only in this subsection, to be more consistent with the notations in online learning literature, we use $\boldsymbol{w}$ for weights instead of Δ as we used in the main text.

To prove the regret bound, we first provide a one-step inequality of OMD with composite loss. Given a convex and continuously differentiable function ψ, recall the Bregman divergence of ψ is defined as

$$
D_\psi(\boldsymbol{x},\boldsymbol{y}) = \psi(\boldsymbol{x}) - \psi(\boldsymbol{y}) - \langle\nabla\psi(\boldsymbol{y}), \boldsymbol{y}-\boldsymbol{x}\rangle.
$$

Note that $\nabla_{\boldsymbol{x}} D_\psi(\boldsymbol{x}, \boldsymbol{y}) = \nabla\psi(\boldsymbol{x}) - \nabla\psi(\boldsymbol{y})$. Moreover, as proved in (Chen & Teboulle, 1993), D_ψ satisfies the following three-point identity:

$$D_\psi(\boldsymbol{z}, \boldsymbol{x}) + D_\psi(\boldsymbol{x}, \boldsymbol{y}) - D_\psi(\boldsymbol{z}, \boldsymbol{y}) = \langle \nabla\psi(\boldsymbol{y}) - \nabla\psi(\boldsymbol{x}), \boldsymbol{z} - \boldsymbol{x} \rangle.$$

Lemma D.1. *Let ψ, ϕ be convex, and define $\boldsymbol{w}_{t+1} = \arg\min_{\boldsymbol{w}} \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w} \rangle + D_\psi(\boldsymbol{w}, \boldsymbol{w}_t) + \phi(\boldsymbol{w})$. Then for any $\boldsymbol{u}$,*

$$\langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{u} \rangle \le \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{w}_{t+1} \rangle + D_\psi(\boldsymbol{u}, \boldsymbol{w}_t) - D_\psi(\boldsymbol{u}, \boldsymbol{w}_{t+1}) - D_\psi(\boldsymbol{w}_{t+1}, \boldsymbol{w}_t) + \phi(\boldsymbol{u}) - \phi(\boldsymbol{w}_{t+1}).$$

Proof. Let $f(\boldsymbol{w}) = \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w} \rangle + D_\psi(\boldsymbol{w}, \boldsymbol{w}_t) + \phi(\boldsymbol{w})$. Since ψ, ϕ are convex, so is f. Therefore, $\boldsymbol{w}_{t+1} = \arg\min_{\boldsymbol{w}} f(\boldsymbol{w})$ implies that for all $\boldsymbol{u}$,

$$\begin{aligned}
0 &\le \langle \nabla f(\boldsymbol{w}_{t+1}), \boldsymbol{u} - \boldsymbol{w}_{t+1} \rangle \\
&= \langle \tilde{\boldsymbol{g}}_t + \nabla\psi(\boldsymbol{w}_{t+1}) - \nabla\psi(\boldsymbol{w}_t) + \nabla\phi(\boldsymbol{w}_{t+1}), \boldsymbol{u} - \boldsymbol{w}_{t+1} \rangle \\
&= \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{u} - \boldsymbol{w}_t \rangle + \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{w}_{t+1} \rangle + \langle \nabla\psi(\boldsymbol{w}_{t+1}) - \nabla\psi(\boldsymbol{w}_t), \boldsymbol{u} - \boldsymbol{w}_{t+1} \rangle + \langle \nabla\phi(\boldsymbol{w}_{t+1}), \boldsymbol{u} - \boldsymbol{w}_{t+1} \rangle.
\end{aligned}$$

Since ϕ is convex, $\langle \phi(\boldsymbol{w}_{t+1}), \boldsymbol{u} - \boldsymbol{w}_{t+1} \rangle \le \phi(\boldsymbol{u}) - \phi(\boldsymbol{w}_{t+1})$. Moreover, by the three-point identity with $\boldsymbol{z} = \boldsymbol{u}, \boldsymbol{x} = \boldsymbol{w}_{t+1}, \boldsymbol{y} = \boldsymbol{w}_t$, we have

$$\langle \nabla\psi(\boldsymbol{w}_t) - \nabla\psi(\boldsymbol{w}_{t+1}), \boldsymbol{u} - \boldsymbol{w}_{t+1} \rangle = D_\psi(\boldsymbol{u}, \boldsymbol{w}_{t+1}) + D_\psi(\boldsymbol{w}_{t+1}, \boldsymbol{w}) - D_\psi(\boldsymbol{u}, \boldsymbol{w}_t).$$

Substituting these back and rearranging the inequality then conclude the proof. □

We restate the formal version of Theorem 4.1 as follows.

Theorem D.2. *Given a sequence of $\{\tilde{\boldsymbol{g}}_t\}_{t=1}^\infty$, a sequence of $\{\eta_t\}_{t=1}^\infty$ such that $0 < \eta_{t+1} \le \eta_t$, and a sequence of $\{\mu_t\}_{t=1}^\infty$ such that $\mu_t \ge 0$, let $\mathcal{R}_t(\boldsymbol{w}) = \frac{\mu_t}{2}\|\boldsymbol{w}\|^2$, $\phi_t(\boldsymbol{w}) = (\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t})\|\boldsymbol{w}\|^2$, $\boldsymbol{w}_1 = 0$ and $\boldsymbol{w}_t$ updated by*

$$\boldsymbol{w}_{t+1} = \arg\min_{\boldsymbol{w}} \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w} \rangle + \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}_t\|^2 + \phi_t(\boldsymbol{w}) + \mathcal{R}_{t+1}(\boldsymbol{w}).$$

Then for any $n \in \mathbb{N}$,

$$\sum_{t=1}^n \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{u} \rangle + \mathcal{R}_t(\boldsymbol{w}_t) - \mathcal{R}_t(\boldsymbol{u}) \le \left(\frac{2}{\eta_{n+1}} + \frac{\mu_{n+1}}{2} \right) \|\boldsymbol{u}\|^2 + \frac{1}{2}\sum_{t=1}^n \eta_t \|\tilde{\boldsymbol{g}}_t\|^2.$$

Proof. Denote $\psi_t(\boldsymbol{w}) = \frac{1}{2\eta_t}\|\boldsymbol{w}\|^2$. Since $\psi_t, \phi_t, \mathcal{R}_t$ are all convex and $D_{\psi_t}(\boldsymbol{w}, \boldsymbol{w}_t) = \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}_t\|^2$, Lemma D.1 holds, which gives

$$\begin{aligned}
\langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{u} \rangle \le{}& \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{w}_{t+1} \rangle + D_{\psi_t}(\boldsymbol{u}, \boldsymbol{w}_t) - D_{\psi_t}(\boldsymbol{u}, \boldsymbol{w}_{t+1}) - D_{\psi_t}(\boldsymbol{w}_{t+1}, \boldsymbol{w}_t) \\
&+ \phi_t(\boldsymbol{u}) - \phi_t(\boldsymbol{w}_{t+1}) + \mathcal{R}_{t+1}(\boldsymbol{u}) - \mathcal{R}_{t+1}(\boldsymbol{w}_{t+1}).
\end{aligned}$$

Equivalently,

$$\begin{aligned}
\langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{u} \rangle + \mathcal{R}_t(\boldsymbol{w}_t) - \mathcal{R}_t(\boldsymbol{u}) \le{}& \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{w}_{t+1} \rangle + D_{\psi_t}(\boldsymbol{u}, \boldsymbol{w}_t) - D_{\psi_t}(\boldsymbol{u}, \boldsymbol{w}_{t+1}) - D_{\psi_t}(\boldsymbol{w}_{t+1}, \boldsymbol{w}_t) \\
&+ \phi_t(\boldsymbol{u}) - \phi_t(\boldsymbol{w}_{t+1}) + \mathcal{R}_t(\boldsymbol{w}_t) - \mathcal{R}_{t+1}(\boldsymbol{w}_{t+1}) + \mathcal{R}_{t+1}(\boldsymbol{u}) - \mathcal{R}_t(\boldsymbol{u}). \qquad (19)
\end{aligned}$$

By Young's inequality,

$$\langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{w}_{t+1} \rangle - D_{\psi_t}(\boldsymbol{w}_{t+1}, \boldsymbol{w}_t) \le \frac{\eta_t}{2}\|\tilde{\boldsymbol{g}}_t\|^2 + \frac{1}{2\eta_t}\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|^2 - \frac{1}{2\eta_t}\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|^2 = \frac{\eta_t}{2}\|\tilde{\boldsymbol{g}}_t\|^2.$$

Next, note that

$$D_{\psi_t}(\boldsymbol{u}, \boldsymbol{w}_t) - D_{\psi_t}(\boldsymbol{u}, \boldsymbol{w}_{t+1}) = D_{\psi_t}(\boldsymbol{u}, \boldsymbol{w}_t) - D_{\psi_{t+1}}(\boldsymbol{u}, \boldsymbol{w}_{t+1}) + D_{\psi_{t+1}}(\boldsymbol{u}, \boldsymbol{w}_{t+1}) - D_{\psi_t}(\boldsymbol{u}, \boldsymbol{w}_{t+1}).$$

Since $\|\boldsymbol{u} - \boldsymbol{w}_{t+1}\|^2 \le 2\|\boldsymbol{u}\|^2 + 2\|\boldsymbol{w}_{t+1}\|^2$ and $\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \ge 0$,

$$\begin{aligned}
&D_{\psi_{t+1}}(\boldsymbol{u}, \boldsymbol{w}_{t+1}) - D_{\psi_t}(\boldsymbol{u}, \boldsymbol{w}_{t+1}) + \phi_t(\boldsymbol{u}) - \phi_t(\boldsymbol{w}_{t+1}) \\
&= \left(\frac{1}{2\eta_{t+1}} - \frac{1}{2\eta_t}\right) \|\boldsymbol{u} - \boldsymbol{w}_{t+1}\|^2 + \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) (\|\boldsymbol{u}\|^2 - \|\boldsymbol{w}_{t+1}\|^2) \le \left(\frac{2}{\eta_{t+1}} - \frac{2}{\eta_t}\right) \|\boldsymbol{u}\|^2.
\end{aligned}$$

Upon substituting back into (19), we have

$$\begin{aligned}
\langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{u} \rangle + \mathcal{R}_t(\boldsymbol{w}_t) - \mathcal{R}_t(\boldsymbol{u}) \le \frac{\eta_t}{2}\|\tilde{\boldsymbol{g}}_t\|^2 + D_{\psi_t}(\boldsymbol{u}, \boldsymbol{w}_t) - D_{\psi_{t+1}}(\boldsymbol{u}, \boldsymbol{w}_{t+1}) + \left(\frac{2}{\eta_{t+1}} - \frac{2}{\eta_t}\right) \|\boldsymbol{u}\|^2 \\
+ \mathcal{R}_t(\boldsymbol{w}_t) - \mathcal{R}_{t+1}(\boldsymbol{w}_{t+1}) + \mathcal{R}_{t+1}(\boldsymbol{u}) - \mathcal{R}_t(\boldsymbol{u}).
\end{aligned}$$

Upon telescoping this one-step inequality, we have

$$\begin{aligned}
&\sum_{t=1}^n \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{u} \rangle + \mathcal{R}_t(\boldsymbol{w}_t) - \mathcal{R}_t(\boldsymbol{u}) \\
&\le \left(\sum_{t=1}^n \frac{\eta_t}{2}\|\tilde{\boldsymbol{g}}_t\|^2\right) + D_{\psi_1}(\boldsymbol{u}, \boldsymbol{w}_1) - D_{\psi_{n+1}}(\boldsymbol{u}, \boldsymbol{w}_{n+1}) + \left(\frac{2}{\eta_{n+1}} - \frac{2}{\eta_1}\right) \|\boldsymbol{u}\|^2 \\
&\quad + \mathcal{R}_1(\boldsymbol{w}_1) - \mathcal{R}_{n+1}(\boldsymbol{w}_{n+1}) + \mathcal{R}_{n+1}(\boldsymbol{u}) - \mathcal{R}_1(\boldsymbol{u}).
\end{aligned}$$

We then conclude the proof by using $\boldsymbol{w}_1 = 0$, $D_{\psi_t}(\boldsymbol{u}, \boldsymbol{w}) = \frac{1}{2\eta_t}\|\boldsymbol{u} - \boldsymbol{w}\|^2$ and $\mathcal{R}_n(\boldsymbol{w}) = \frac{\mu_t}{2}\|\boldsymbol{w}\|^2$ to simplify

$$\begin{aligned}
&D_{\psi_1}(\boldsymbol{u}, \boldsymbol{w}_1) - D_{\psi_{n+1}}(\boldsymbol{u}, \boldsymbol{w}_{n+1}) + \left(\frac{2}{\eta_{n+1}} - \frac{2}{\eta_1}\right) \|\boldsymbol{u}\|^2 \\
&\le \frac{1}{2\eta_1}\|\boldsymbol{u}\|^2 + \left(\frac{2}{\eta_{n+1}} - \frac{2}{\eta_1}\right) \|\boldsymbol{u}\|^2 \le \frac{2}{\eta_{n+1}}\|\boldsymbol{u}\|^2
\end{aligned}$$

and $\mathcal{R}_1(\boldsymbol{w}_1) - \mathcal{R}_{n+1}(\boldsymbol{w}_{n+1}) + \mathcal{R}_{n+1}(\boldsymbol{u}) - \mathcal{R}_1(\boldsymbol{u}) \le \mathcal{R}_{n+1}(\boldsymbol{u}) + \mathcal{R}_1(\boldsymbol{w}_1) = \frac{\mu_{n+1}}{2}\|\boldsymbol{u}\|^2$. □

D.2. Proof of Theorem 4.2

Theorem 4.2. *Follow Assumption 2.1 and consider any $c > 0$. Let $\Delta_1 = 0$ and update Δ_t by*

$$\Delta_{t+1} = \frac{\Delta_t - \eta_t \beta^{-t} \boldsymbol{g}_t}{1 + \eta_t \mu_{t+1} + \eta_t (\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t})}.$$

Let $\mu_t = \beta^{-t}\mu$, $\eta_t = \beta^t \eta$, $\beta = 1 - \alpha$, $\mu = \frac{24F^* c}{(G+\sigma)\alpha^{5/2}N}$, $\eta = \frac{2F^*}{(G+\sigma)^2 N}$, $\alpha = \max\{N^{-2/3}, \frac{(F^*)^{4/7}c^{2/7}}{(G+\sigma)^{6/7}N^{4/7}}\}$. Then for N large enough such that $\alpha \le \frac{1}{2}$,

$$\mathbb{E}\,\|\nabla F(\overline{\boldsymbol{x}})\|_c \lesssim \frac{G + \sigma}{N^{1/3}} + \frac{(F^*)^{2/7}(G+\sigma)^{4/7}c^{1/7}}{N^{2/7}}.$$

Proof. First, define $D = \frac{F^*}{(G+\sigma)\sqrt{\alpha}N}$, $\mu = \frac{24cD}{\alpha^2}$ and $\eta = \frac{2D\sqrt{\alpha}}{G+\sigma}$. Note that these definitions are equivalent to $\mu = \frac{24F^* c}{(G+D)\alpha^{5/2}N}$ and $\eta = \frac{2F^*}{(G+\sigma)^2 N}$ as defined in the theorem.

Next, note that both Theorem C.1 and Theorem D.2 hold since the explicit update of Δ_{t+1} is equivalent to

$$\Delta_{t+1} = \arg\min_{\Delta} \langle \beta^{-t}\boldsymbol{g}_t, \Delta \rangle + \frac{1}{2\eta_t}\|\Delta - \Delta_t\|^2 + \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) \|\Delta\|^2 + \frac{\mu_{t+1}}{2}\|\Delta\|^2.$$

Also recall that $\text{Regret}_n(\boldsymbol{u}_n) = \sum_{t=1}^n \langle \beta^{-t}\boldsymbol{g}_t, \Delta_t - \boldsymbol{u}_n \rangle + \mathcal{R}_t(\Delta_t) - \mathcal{R}_t(\boldsymbol{u}_n)$. Therefore, upon substituting $\tilde{\boldsymbol{g}}_t = \beta^{-t}\boldsymbol{g}_t$, $\eta_t = \beta^t\eta$, $\mu_t = \beta^{-t}\mu$ and $\|\boldsymbol{u}_n\| = D$ into Theorem D.2, we have

$$\begin{aligned}
\mathbb{E}\,\text{Regret}_n(\boldsymbol{u}_n) &\le \left(\frac{2}{\eta_{n+1}} + \frac{\mu_{n+1}}{2}\right) \mathbb{E}\,\|\boldsymbol{u}\|^2 + \frac{1}{2}\sum_{t=1}^n \eta_t\,\mathbb{E}\,\|\tilde{\boldsymbol{g}}_t\|^2 \\
&= \left(\frac{2}{\eta} + \frac{\mu}{2}\right) D^2 \beta^{-(n+1)} + \frac{\eta}{2}\sum_{t=1}^n \beta^{-t}\,\mathbb{E}\,\|\boldsymbol{g}_t\|^2.
\end{aligned}$$

By Assumption 2.1, $\mathbb{E}\,\|\boldsymbol{g}_t\|^2 = \mathbb{E}\,\|\nabla F(\boldsymbol{x}_t)\|^2 + \mathbb{E}\,\|\nabla F(\boldsymbol{x}_t) - \boldsymbol{g}_t\|^2 \le G^2 + \sigma^2$. Moreover, $\sum_{t=1}^n \beta^{-t} \le \frac{\beta^{-n}}{1-\beta}$. Therefore,

$$\beta^{n+1}\,\mathbb{E}\,\mathrm{Regret}_n(\boldsymbol{u}_n) \le \left(\frac{2}{\eta} + \frac{\mu}{2}\right) D^2 + \frac{\eta(G^2+\sigma^2)}{2\alpha}$$

Upon substituting $\eta = \frac{2D\sqrt{\alpha}}{G+\sigma}$ (note that $\frac{G^2+\sigma^2}{G+\sigma} \le G+\sigma$) and $\mu = \frac{24cD}{\alpha^2}$, we have

$$\le \frac{2D(G+\sigma)}{\sqrt{\alpha}} + \frac{12cD^3}{\alpha^2}.$$

Consequently, with $\alpha \le \frac{1}{2}$ (so that $\beta^{-1} \le 2$), we have

$$\begin{aligned}&\frac{1}{DN}\left(\beta^{N+1}\,\mathbb{E}\,\mathrm{Regret}_N(\boldsymbol{u}_N) + \alpha\sum_{n=1}^N \beta^n\,\mathbb{E}\,\mathrm{Regret}_n(\boldsymbol{u}_n)\right)\\ &\le \frac{1+2\alpha N}{DN}\left(\frac{2D(G+\sigma)}{\sqrt{\alpha}} + \frac{12cD^3}{\alpha^2}\right) \lesssim \frac{G+\sigma}{N} + \frac{cD^2}{\alpha^2 N} + (G+\sigma)\sqrt{\alpha} + \frac{cD^2}{\alpha}.\end{aligned}$$

Upon substituting this into the convergence guarantee in Theorem C.1, we have

$$\begin{aligned}\mathbb{E}\,\|\nabla F(\overline{\boldsymbol{x}})\|_c &\le \frac{F^*}{DN} + \frac{2G+\sigma}{\alpha N} + \sigma\sqrt{\alpha} + \frac{12cD^2}{\alpha^2} + \frac{1}{DN}\left(\beta^{N+1}\,\mathbb{E}\,\mathrm{Regret}_N(\boldsymbol{u}_N) + \alpha\sum_{n=1}^N \beta^n\,\mathbb{E}\,\mathrm{Regret}_n(\boldsymbol{u}_n)\right)\\ &\lesssim \frac{F^*}{DN} + \frac{G+\sigma}{\alpha N} + (G+\sigma)\sqrt{\alpha} + \frac{cD^2}{\alpha^2}\end{aligned}$$

With $D = \frac{F^*}{(G+\sigma)\sqrt{\alpha}N}$ and $\alpha = \max\{N^{-2/3}, \frac{(F^*)^{4/7}c^{2/7}}{(G+\sigma)^{6/7}N^{4/7}}\}$, we have

$$\lesssim \frac{G+\sigma}{\alpha N} + (G+\sigma)\sqrt{\alpha} + \frac{(F^*)^2 c}{(G+\sigma)^2\alpha^3 N^2} \lesssim \frac{G+\sigma}{N^{1/3}} + \frac{(F^*)^{2/7}(G+\sigma)^{4/7}c^{1/7}}{N^{2/7}}. \qquad \square$$