

---

# Gradient-Variation Online Adaptivity for Accelerated Optimization with Hölder Smoothness

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

Smoothness is known to be crucial for acceleration in offline optimization, and for problem-dependent regret that scale with gradient variations in online learning. Interestingly, these two problems are actually closely connected — accelerated optimization can be understood from the lens of gradient-variation online learning. In this paper, we systematically investigate online learning with *Hölder smooth* functions, a class encompassing both smooth and non-smooth (Lipschitz) functions, and further explore its implications for offline optimization. First, we propose an online algorithm with optimal gradient-variation regret for convex functions, which implies an optimal accelerated method for stochastic convex optimization under Hölder smoothness. Then, we extend the results in three aspects: (i) strongly convex functions, (ii) non-stationary online learning where the comparator sequence changes over time; (iii) universal online learning where the curvature of functions is unknown. In all three cases, we have achieved the first gradient-variation regret that can well interpolate the results between the smooth and Lipschitz regimes, and recover the optimal results in each case. Notably, our proposed algorithms do not require prior knowledge of the Hölder smoothness parameter, greatly improving the adaptivity of existing methods that depend on this parameter, even when designed specifically for smooth functions. Finally, we demonstrate several implications for offline optimization through carefully tailored online-to-batch conversions.

## 1 Introduction

Online Convex Optimization (OCO) [Hazan, 2022] is a powerful and versatile framework for learning with data streams, typically modeled as an iterative game between a player and the environment. At iteration  $t \in [T]$ , the player chooses a decision  $\mathbf{x}_t$  from a bounded convex feasible domain  $\mathcal{X} \subseteq \mathbb{R}^d$ . Simultaneously, the environment reveals a convex function  $f_t : \mathcal{X} \rightarrow \mathbb{R}$ , and the player incurs a loss  $f_t(\mathbf{x}_t)$ . The player receives certain information to update  $\mathbf{x}_{t+1}$ , aiming to optimize the *regret*

$$\text{REG}_T \triangleq \sum_{t=1}^T f_t(\mathbf{x}_t) - \min_{\mathbf{x} \in \mathcal{X}} \sum_{t=1}^T f_t(\mathbf{x}). \quad (1)$$

When the online functions are *Lipschitz*, it is known that minimax optimal regret bounds are  $\mathcal{O}(\sqrt{T})$  and  $\mathcal{O}(\frac{1}{\lambda} \log T)$  for convex and  $\lambda$ -strongly convex functions respectively [Zinkevich, 2003; Hazan et al., 2007]. When the online functions are *smooth*, we can further obtain *problem-dependent* regret guarantees, which enjoy better bounds in easy problem instances while maintaining the same minimax optimality in the worst case [de Rooij et al., 2014; Foster et al., 2015]. Among many problem-dependent quantities, a particular one called *gradient variations* draws much attention [Chiang et al.,

Table 1: Comparisons with existing regret bounds. We achieve the gradient-variation regret under four setups, each recovering the best-known results with either smoothness or non-smoothness (Lipschitzness). For universal regret, we use  $\min\{\cdot, \cdot\}$  to denote the minimum of the bounds for convex and strongly convex functions.

Setups	Our Results with ( $L_\nu, \nu$ )-Hölder Smoothness	Recovering Existing Results with	
		Smoothness ( $\nu = 1$ )	Non-smoothness ( $\nu = 0$ )
Convex	$\mathcal{O}(\sqrt{V_T} + L_\nu T^{\frac{1-\nu}{2}})$ [Theorem 1]	$\mathcal{O}(\sqrt{V_T})$ [Chiang et al., 2012]	$\mathcal{O}(\sqrt{T})$ [Zinkevich, 2003]
$\lambda$ -Strongly Convex	$\mathcal{O}(\frac{1}{\lambda} \log V_T + \frac{1}{\lambda} L_\nu^2 (\log T)^{1-\nu})$ [Theorem 3]	$\mathcal{O}(\frac{1}{\lambda} \log V_T)$ [Chiang et al., 2012]	$\mathcal{O}(\frac{1}{\lambda} \log T)$ [Hazan et al., 2007]
Dynamic Regret	$\mathcal{O}(\sqrt{V_T(1+P_T)} + L_\nu(1+P_T)^{\frac{1+\nu}{2}} T^{\frac{1-\nu}{2}})$ [Theorem 5]	$\mathcal{O}(\sqrt{V_T(1+P_T)} + P_T)$ [Zhao et al., 2024]	$\mathcal{O}(\sqrt{T(1+P_T)})$ [Zhang et al., 2018]
Universal Regret	$\min\left\{\mathcal{O}(\sqrt{V_T} + L_\nu T^{\frac{1-\nu}{2}}), \mathcal{O}(\frac{1}{\lambda} \log V_T + \frac{1}{\lambda} L_\nu^2 (\log T)^{\frac{1-\nu}{2}})\right\}$ [Theorem 6]	$\min\{\mathcal{O}(\sqrt{V_T}), \mathcal{O}(\frac{1}{\lambda} \log V_T)\}$ [Yan et al., 2024]	$\min\{\mathcal{O}(\sqrt{T}), \mathcal{O}(\frac{1}{\lambda} \log T)\}$ [Wang et al., 2019]

2012; Yang et al., 2014], which is defined to capture how the problem evolves over time,

$$V_T \triangleq \sum_{t=2}^T \sup_{\mathbf{x} \in \mathcal{X}} \|\nabla f_t(\mathbf{x}) - \nabla f_{t-1}(\mathbf{x})\|^2. \quad (2)$$

It is established that optimal gradient-variation regret for convex and  $\lambda$ -strongly convex functions are  $\mathcal{O}(\sqrt{V_T})$  and  $\mathcal{O}(\frac{1}{\lambda} \log V_T)$  [Chiang et al., 2012], respectively. There has been significant subsequent development in more complex environments [Zhao et al., 2020, 2024; Sachs et al., 2023; Yan et al., 2023; Xie et al., 2024]. Gradient-variation online learning has garnered significant interest, not only because its development substantially enriches the field of OCO through the analysis of trajectory dynamics, but also due to its fundamental connections to a wide range of optimization problems. Gradient-variation adaptivity has been shown to be crucial for fast convergence in minimax games [Syrkanis et al., 2015; Zhang et al., 2022], and recent studies reveal its key role in bridging adversarial and stochastic convex optimization [Sachs et al., 2022; Chen et al., 2024]. In this paper, we further demonstrate that gradient-variation online learning is closely connected to offline smooth accelerated optimization, building on the works of Cutkosky [2019] and Kavis et al. [2019].

Consider the offline optimization problem  $\min_{\mathbf{x} \in \mathcal{X}} \ell(\mathbf{x})$ , where  $\ell : \mathbb{R}^d \rightarrow \mathbb{R}$  is a convex objective and  $\mathcal{X}$  is a convex feasible domain. A standard approach to obtain  $\ell(\bar{\mathbf{x}}_T) - \min_{\mathbf{x} \in \mathcal{X}} \ell(\mathbf{x}) \leq \varepsilon_T$  is to apply the online-to-batch conversion [Cesa-Bianchi et al., 2004] with an appropriate online algorithm. For instance, by using online gradient descent [Zinkevich, 2003], one can directly obtain a convergence rate of  $\varepsilon_T = \mathcal{O}(1/\sqrt{T})$  for optimizing Lipschitz functions. However, for smooth functions, the connection between online learning and offline optimization is less straightforward. Recent breakthroughs [Cutkosky, 2019; Kavis et al., 2019] demonstrate that a *stabilized* online-to-batch conversion, combined with a gradient-variation online algorithm, can obtain the optimal *accelerated* convergence rate of  $\varepsilon_T = \mathcal{O}(1/T^2)$  for convex and smooth functions [Nesterov, 2018].

Motivated by the significant differences in the optimal rates and algorithms for Lipschitz and smooth regimes, previous works have explored intermediate function classes, named *Hölder smoothness* [Nesterov, 2015]. Here we consider in Euclidean space: a function  $\ell : \mathbb{R}^d \rightarrow \mathbb{R}$  is  $(L_\nu, \nu)$ -Hölder smooth with respect to the  $\ell_2$ -norm with  $L_\nu > 0$  and  $\nu \in [0, 1]$ , if

$$\|\nabla \ell(\mathbf{x}) - \nabla \ell(\mathbf{y})\|_2 \leq L_\nu \|\mathbf{x} - \mathbf{y}\|_2^\nu, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \quad (3)$$

It implies the  $L$ -smoothness with  $L_\nu = L$  when  $\nu = 1$  and  $G$ -Lipchitzness with  $L_\nu = 2G$  when  $\nu = 0$ . An optimization algorithm is called *universal* [Nesterov, 2015] if it can automatically achieve the best possible convergence guarantee without requiring Hölder smoothness parameters  $(L_\nu, \nu)$ . A relaxed definition of universality only requires achieving optimal rates for the two extreme cases of  $L$ -smooth ( $\nu = 1$ ) and  $G$ -Lipschitz ( $\nu = 0$ ) functions, without prior knowledge of  $L$  and  $G$ .

In this paper, we conduct a systematic study of gradient-variation online learning under  $(L_\nu, \nu)$ -Hölder smoothness. We demonstrate that, optimistic online gradient descent [Chiang et al., 2012] equipped with specific time-varying step sizes can obtain regret bounds of  $\mathcal{O}(\sqrt{V_T} + L_\nu T^{(1-\nu)/2})$  for convex functions, and  $\mathcal{O}(\frac{1}{\lambda} \log V_T + \frac{1}{\lambda} L_\nu^2 (\log T)^{1-\nu})$  for  $\lambda$ -strongly convex functions. When online functions are smooth, the above results recover the best known guarantees of  $\mathcal{O}(\sqrt{V_T})$  and  $\mathcal{O}(\frac{1}{\lambda} \log V_T)$  for convex and  $\lambda$ -strongly convex functions [Yang et al., 2014]; and when online

functions are Lipschitz, the results imply the minimax optimal  $\mathcal{O}(\sqrt{T})$  and  $\mathcal{O}(\frac{1}{\lambda} \log T)$  [Abernethy et al., 2008], respectively. The results are summarized in the first three rows in Table 1. Importantly, our proposed online algorithms do *not* require prior knowledge of  $L_\nu$  and  $\nu$ , thereby paving the way for developing universal and accelerated methods in offline optimization.

Furthermore, combining our proposed online algorithms with carefully designed online-to-batch conversions can yield favorable implications for offline accelerated optimization. Specifically,

- (1) For stochastic convex optimization, under the standard assumption that the variance of the stochastic gradient is bounded by an unknown constant  $\sigma$ , we obtain the *universal* algorithm with an *optimal* convergence rate of  $\mathcal{O}(L_\nu D^{1+\nu} T^{-(1+3\nu)/2} + \sigma D/\sqrt{T})$  for  $(L_\nu, \nu)$ -Hölder smooth functions, where  $D$  is the domain diameter. This matches the best known result of Rodomanov et al. [2024]. This interpolated result automatically recovers the optimal rates of  $\mathcal{O}(LD^2/T^2 + \sigma D/\sqrt{T})$  for  $L$ -smooth functions, and  $\mathcal{O}((G + \sigma)D/\sqrt{T})$  for  $G$ -Lipschitz functions.
- (2) For deterministic optimization with  $\lambda$ -strongly convex functions, we present the *first* universal method with an accelerated convergence. Our method simultaneously achieves the near-optimal accelerated rate of  $\mathcal{O}(\exp(-T/(\sqrt{\kappa} \cdot \log \kappa)))$  for  $L$ -smooth functions and the  $\mathcal{O}((\log T)/T)$  rate for Lipschitz functions, where  $\kappa = L/\lambda$  is the condition number. This improves upon the previously best-known universal result by Levy [2017], who achieved a non-accelerated  $\mathcal{O}(\exp(-T/\kappa) \cdot T/\kappa)$  rate for smooth case and the same  $\mathcal{O}((\log T)/T)$  rate for Lipschitz case. This is done via a novel detection method that estimates the smoothness parameter on the fly.

We finally extend our online learning results to more complex regret measures, including *dynamic regret* that competes with changing comparators [Zhang et al., 2018], and *universal regret* that does not rely on curvature information of online functions [van Erven and Koolen, 2016].<sup>1</sup> We establish the first results for these two challenging performance measures under Hölder smoothness.

- (3) For dynamic regret minimization, we prove an  $\mathcal{O}(\sqrt{V_T(1 + P_T)} + L_\nu(1 + P_T)^{(1+\nu)/2} T^{(1-\nu)/2})$  dynamic regret, where  $P_T$  models environmental non-stationarity. Our guarantee recovers the best-known results of smooth and Lipschitz functions [Zhang et al., 2018; Zhao et al., 2024].
- (4) For universal regret minimization, we achieve the first guarantee — an  $\mathcal{O}(\sqrt{V_T} + L_\nu T^{(1-\nu)/2})$  regret for convex functions and an  $\mathcal{O}(\frac{1}{\lambda} \log V_T + \frac{1}{\lambda} L_\nu^2 (\log T)^{\frac{1-\nu}{1+\nu}})$  regret for  $\lambda$ -strongly convex functions. These can recover the optimal results for Lipschitz and smooth functions [van Erven and Koolen, 2016; Wang et al., 2019; Yan et al., 2024], respectively.

The results are summarized in the last two rows in Table 1. Importantly, all the above results are achieved without requiring Hölder smoothness parameters. In contrast, the previous best-known gradient-variation methods for convex and smooth functions still depend on the smoothness parameter  $L$  [Chiang et al., 2012; Zhao et al., 2024]. Our results enhance the adaptivity of gradient-variation online learning, which we believe has broader implications for offline optimization.

**Organization.** The rest is structured as follows. Section 2 introduces the problem setup. Section 3 explores gradient-variation online learning and its implications to optimization with Hölder smoothness. Section 4 extends to strongly convex optimization. Section 5 studies two more challenging online learning scenarios. Section 6 concludes the paper. All proofs are provided in the appendices.

## 2 Problem Setup and Preliminary

In this section, we first formulate the problem setup and list assumptions in Section 2.1 and then introduce two more complex regret measures (dynamic regret and universal regret) in Section 2.2.

### 2.1 Problem Setup and Assumptions

**Notations.** We denote by  $\|\cdot\|$  the  $\ell_2$ -norm in default.  $\tilde{\mathcal{O}}(\cdot)$  omits poly-logarithmic factors on leading terms. We use  $[N]$  to represent set  $\{1, 2, \dots, N\}$  and use  $\sum_t$  as abbreviation for  $\sum_{t \in [T]}$ . We define  $\sum_{i=a}^b c_i \triangleq 0$  if  $a > b$ . Let  $\mathcal{D}_\psi(\mathbf{x}, \mathbf{y}) \triangleq \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla \psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$  denote the Bregman divergence associated with the convex regularizer  $\psi : \mathcal{X} \rightarrow \mathbb{R}$ . We use  $a \lesssim b$  to represent  $a = \mathcal{O}(b)$ .

<sup>1</sup>In offline optimization, “universality” refers to adaptivity to smoothness, allowing algorithms to perform optimally without prior knowledge of smoothness parameters, whereas in online learning, “universality” denotes adaptivity to convexity, enabling algorithms to achieve optimal regret bounds across different function classes.

**Offline Optimization Setup.** The optimization problem is formulated as  $\min_{\mathbf{x} \in \mathcal{X}} \ell(\mathbf{x})$ . We assume the learner has access to the oracle of the gradient information  $\nabla \ell(\cdot)$ , denoted by  $\mathbf{g}(\cdot)$ , where there are two settings: (i) the deterministic setting that  $\mathbf{g}(\cdot)$  exactly equals to  $\nabla \ell(\cdot)$ ; and (ii) the stochastic setting that  $\mathbf{g}(\cdot)$  is the unbiased estimate of  $\nabla \ell(\cdot)$ , such that  $\mathbb{E}[\mathbf{g}(\mathbf{x}) \mid \mathbf{x}] = \nabla \ell(\mathbf{x})$ , and at the same time we import the classic assumption of bounded variance, that is,  $\mathbb{E}[\|\mathbf{g}(\mathbf{x}) - \nabla \ell(\mathbf{x})\|^2 \mid \mathbf{x}] \leq \sigma^2$  holds for all  $\mathbf{x} \in \mathcal{X}$ . In this paper, we consider the convergence of the sub-optimality gap, i.e.,  $\ell(\mathbf{x}_T^\dagger) - \min_{\mathbf{x} \in \mathcal{X}} \ell(\mathbf{x})$ , where  $\mathbf{x}_T^\dagger$  is some statistic of the optimizing sequence  $\{\mathbf{x}_t\}_{t=1}^T$ .

The following bounded domain assumption will be used throughout this paper. We focus on this setting since it is common in the literature [Hazan, 2022], and studying gradient variation under Hölder smoothness is already challenging and non-trivial in the bounded domain.

**Assumption 1** (Bounded Domain). The feasible domain  $\mathcal{X} \subseteq \mathbb{R}^d$  is non-empty and closed with the diameter bounded by  $D$ , that is,  $\|\mathbf{x} - \mathbf{y}\|_2 \leq D$  for all  $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ .

To better illustrate the universality of our methods and to distinguish the degree of dependence on the smoothness information, we propose the following definitions of *weak/strong universality*.

**Definition 1** (Weak/Strong Universality). A method is *weakly universal* if it automatically adapts to standard  $L$ -smoothness or  $G$ -Lipschitz, without requiring  $L$  and  $G$ . A method is *strongly universal* if it automatically adapts to  $(L_\nu, \nu)$ -Hölder smoothness without requiring  $L_\nu$  and  $\nu$ .

## 2.2 More Challenging Regret Measures

In this paper, we also investigate the two more complex regret measures under Hölder smoothness.

**Dynamic Regret.** In non-stationary online learning scenario where data are evolving and the optimal decisions are drifting over time, the *dynamic regret* [Zhang et al., 2018] is proposed that competes with time-varying comparators  $\mathbf{u}_1, \dots, \mathbf{u}_T \in \mathcal{X}$ , defined as:

$$\text{D-REG}_T(\mathbf{u}_1, \dots, \mathbf{u}_T) \triangleq \sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{u}_t). \quad (4)$$

A dynamic regret upper bound usually scales with certain non-stationarity measure, such as the path length of comparators, defined as  $P_T \triangleq \sum_{t=2}^T \|\mathbf{u}_t - \mathbf{u}_{t-1}\|$  [Zhang et al., 2018; Zhao et al., 2024].

**Universal Regret.** Universal online learning [van Erven and Koolen, 2016] studies the problem of lacking curvature information and aiming to design an online algorithm, denoted by  $\mathcal{A}$ , that behaves in the following way: (i) when all the online functions are convex, the regret bound of  $\mathcal{A}$  is the same as that of methods specifically designed for convex functions; and (ii) when all the online functions are strongly convex,  $\mathcal{A}$  achieves the same regret as strongly convex regret minimization algorithms. Below we formally define the *universal regret*, which measures the decisions  $\{\mathbf{x}_t^{\mathcal{A}}\}_{t=1}^T$  generated by an online algorithm  $\mathcal{A}$ , where all online functions belong to some *unknown* function class  $\mathcal{F}$ :

$$\text{U-REG}_T^{\mathcal{A}}(\mathcal{F}) \triangleq \sum_{t=1}^T f_t(\mathbf{x}_t^{\mathcal{A}}) - \min_{\mathbf{x} \in \mathcal{X}} \sum_{t=1}^T f_t(\mathbf{x}), \quad \{f_t\}_{t=1}^T \subseteq \mathcal{F}, \quad \mathcal{F} \in \{\mathcal{F}_c, \mathcal{F}_{sc}^\lambda\}, \quad (5)$$

where  $\mathcal{F}_c$  represents convex function class, and  $\mathcal{F}_{sc}^\lambda$  represents  $\lambda$ -strongly convex function class.

## 3 Convex Optimization with Hölder Smoothness

In this section, we achieve the gradient-variation regret bound with Hölder smoothness in Section 3.1, then apply our method to obtain the best known results for stochastic optimization in Section 3.2.

### 3.1 Gradient-Variation Online Convex Optimization with Hölder Smoothness

We leverage the optimistic online gradient descent (optimistic OGD) [Chiang et al., 2012] as our algorithmic framework, which enjoys gradient-variation regret. Before receiving  $\mathbf{x}_t$  and performing the classical online gradient descent update step using  $\nabla f_t(\mathbf{x}_t)$  [Zinkevich, 2003], optimistic OGD performs an additional update step using the prediction for the upcoming gradient, denoted by

155  $M_t \in \mathbb{R}^d$ , which is often set as the last observed gradient  $\nabla f_{t-1}(\mathbf{x}_{t-1})$ . Specifically, optimistic  
 156 OGD maintains two decision sequences  $\{\mathbf{x}_t\}_{t=1}^T, \{\widehat{\mathbf{x}}_t\}_{t=1}^T$ , and updates by

$$\mathbf{x}_t = \Pi_{\mathcal{X}}[\widehat{\mathbf{x}}_t - \eta_t M_t], \quad \widehat{\mathbf{x}}_{t+1} = \Pi_{\mathcal{X}}[\widehat{\mathbf{x}}_t - \eta_t \nabla f_t(\mathbf{x}_t)], \quad (6)$$

157 where  $\eta_t > 0$  is a time-varying step size, and we denote by  $\Pi_{\mathcal{X}}[\mathbf{y}] \triangleq \arg \min_{\mathbf{x} \in \mathcal{X}} \|\mathbf{x} - \mathbf{y}\|_2$  the  
 158 Euclidean projection operator for any  $\mathbf{y} \in \mathcal{X}$ . Setting  $M_t = \nabla f_{t-1}(\mathbf{x}_{t-1})$ , optimistic OGD can  
 159 obtain gradient-variation regret under standard smoothness via classical analysis [Chiang et al., 2012]:

$$\text{REG}_T \lesssim \frac{1}{\eta_T} + \sum_{t=1}^T \eta_t \|\nabla f_t(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_{t-1})\|^2 - \sum_{t=2}^T \frac{1}{\eta_{t-1}} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2. \quad (7)$$

160 On the right-hand side, the second term  $\sum_t \eta_t \|\nabla f_t(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_{t-1})\|^2$  is an adaptivity term  
 161 measuring the deviation between the two gradients, and the last one is a negative stability term.  
 162 The adaptivity term can be bounded by  $\|\nabla f_t(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_t)\|^2 + \|\nabla f_{t-1}(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_{t-1})\|^2$ ,  
 163 where the first part can be converted to the desired gradient variation Eq. (2) and the second part is  
 164 bounded by  $L^2 \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2$  under standard  $L$ -smoothness assumption, and thus can be canceled  
 165 out by the negative term in Eq. (7). When the smoothness parameter  $L$  is given, this cancellation  
 166 is straightforward by ensuring  $\eta_t \lesssim 1/L$ , which is the reason why most existing gradient-variation  
 167 techniques require the prior knowledge of the smoothness parameter  $L$ .

168 However, when it comes to gradient-variation online learning with Hölder smoothness, we *cannot*  
 169 directly apply the definition in Eq. (3) as we did with standard smoothness, because it would yield  
 170  $\|\nabla f_{t-1}(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_{t-1})\|^2 \leq L_\nu^2 \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^{2\nu}$ , which mismatches with the negative term. To  
 171 this end, we present a key lemma regarding Hölder smoothness as a kind of “inexact” smoothness [De-  
 172 volder et al., 2014; Nesterov, 2015], which has a similar form to standard smoothness except for an  
 173 additional corruption term. The proof is in Appendix A.2.

174 **Lemma 1.** Suppose function  $f$  is  $(L_\nu, \nu)$ -Hölder smooth, then for any  $\delta > 0$ , denoting by  $L =$   
 175  $\delta^{\frac{\nu-1}{1+\nu}} L_\nu^{\frac{2}{1+\nu}}$ , it holds that for all  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ :

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \leq L^2 \|\mathbf{x} - \mathbf{y}\|^2 + 4L\delta. \quad (8)$$

176 When smoothness holds, i.e.,  $\nu = 1$ , our lemma recovers the standard smoothness assumption when  
 177  $\delta$  approaches 0. When functions are  $G$ -Lipschitz, i.e.,  $\nu = 0$  and  $L_\nu = 2G$ , by treating the right-hand  
 178 side as a function for  $\delta$  and calculating the minimum, the lemma results in  $\|\nabla f_t(\mathbf{x}) - \nabla f_t(\mathbf{y})\|^2 \lesssim G^2$ ,  
 179 providing an upper bound that depends only on  $G$ .

180 Although we can obtain  $L^2 \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2$  from the adaptivity term in Eq. (7) that matches the negative  
 181 term in Lemma 1, we still cannot directly perform cancellation by explicitly setting the clipping  
 182  $\eta_t \lesssim 1/L$ . This is because this  $L$  depends on  $\delta$  in Lemma 1, which exists only in analysis and is  
 183 algorithmically *unavailable*. To this end, inspired by Kavis et al. [2019], we adopt the following  
 184 adaptive step sizes which allows us to perform *virtual* clipping in the analysis:

$$\eta_t = \frac{D}{2\sqrt{A_{t-1}}}, \quad A_t \triangleq \|\nabla f_1(\mathbf{x}_1)\|^2 + \sum_{s=2}^t \|\nabla f_s(\mathbf{x}_s) - M_s\|^2. \quad (9)$$

185 The rationale behind this configuration is that, since  $\eta_t$  in Eq. (9) is non-increasing, it will eventually  
 186 become smaller than  $1/L$  after certain rounds (denoted as  $t_0$ ). This implies that, for  $t > t_0$ , the  
 187 analysis allows the aforementioned cancellation to occur. For the summation before  $t_0$  of the  
 188 adaptivity term in Eq. (7), we can directly bound it by  $\sqrt{A_{t_0-1}}$  that is inversely proportional to  $\eta_{t_0}$ ,  
 189 thereby further bounded by a constant because  $\eta_{t_0}$  is larger than  $1/L$ .

190 Finally, we provide the following regret guarantee with the proof in Appendix A.3.

191 **Theorem 1.** Under Assumption 1, and assuming online functions are convex and  $(L_\nu, \nu)$ -Hölder  
 192 smooth, optimistic OGD in Eq. (6) with  $M_1 = \mathbf{0}$ ,  $M_t = \nabla f_{t-1}(\mathbf{x}_{t-1})$  for all  $t \geq 2$ , and the step  
 193 sizes specified in Eq. (9) for all  $t \in [T]$ , ensures the following regret bound:

$$\text{REG}_T \leq \mathcal{O}\left(D\sqrt{V_T} + LD^{1+\nu}T^{\frac{1-\nu}{2}} + D\|\nabla f_1(\mathbf{x}_1)\|\right).$$

194 Theorem 1 implies optimal guarantees for both smooth and Lipschitz functions even in terms of  
 195 the dependence on the domain diameter  $D$ : (i) when the online functions are  $L$ -smooth, i.e.,  $(L, 1)$ -  
 196 Hölder smooth, our result recovers the optimal bound of  $\mathcal{O}(D\sqrt{V_T} + LD^2)$  [Chiang et al., 2012];  
 197 and (ii) when the online functions are  $G$ -Lipschitz, i.e.,  $(2G, 0)$ -Hölder smooth, our result also  
 198 recovers the worst-case minimax optimal guarantee  $\mathcal{O}(GD\sqrt{T})$  [Zinkevich, 2003].



**Remark 1.** We emphasize that, unlike previous works that require the parameter  $L$  of smooth functions [Chiang et al., 2012; Orabona, 2019; Yan et al., 2023; Zhao et al., 2024], our algorithm is *strongly universal* (as defined in Definition 1), as it does *not* require the Hölder smoothness parameters, while can still achieve optimal guarantees for both smooth and Lipschitz functions.  $\triangleleft$

### 3.2 Implication to Offline Convex Optimization

In this part, we demonstrate the effectiveness of the gradient-variation adaptivity presented in Section 3.1 by leveraging it to achieve acceleration for offline optimization under Hölder smoothness via a powerful online-to-batch conversion [Cutkosky, 2019]. For completeness, we first recall the conversion framework, which is useful to directly translate the regret bounds to convergence rates.

Specifically, consider the optimization problem of  $\min_{\mathbf{x} \in \mathcal{X}} \ell(\mathbf{x})$  with a gradient oracle  $\mathbf{g}(\cdot)$  provided. We require an online learning algorithm  $\mathcal{A}$  and a sequence of positive weights  $\{\alpha_t\}_{t=1}^T$ . The conversion goes as follows. At each iteration  $t$ , it computes a weighted average of past decisions  $\bar{\mathbf{x}}_t = \frac{1}{\alpha_{1:t}} \sum_{s=1}^t \alpha_s \mathbf{x}_s$  with  $\alpha_{1:t} \triangleq \sum_{s=1}^t \alpha_s$ , queries the gradient  $\mathbf{g}(\bar{\mathbf{x}}_t)$ , and then construct the online function  $f_t(\mathbf{x}) \triangleq \alpha_t \langle \mathbf{g}(\bar{\mathbf{x}}_t), \mathbf{x} \rangle$ . This function  $f_t(\cdot)$  is passed to the online algorithm  $\mathcal{A}$  to obtain  $\mathbf{x}_{t+1}$  for the next iteration. After  $T$  iterations, the online-to-batch conversion outputs the final decision  $\bar{\mathbf{x}}_T = \frac{1}{\alpha_{1:T}} \sum_{t=1}^T \alpha_t \mathbf{x}_t$ . The conversion ensures the following convergence guarantee:

$$\mathbb{E}[\ell(\bar{\mathbf{x}}_T)] - \ell(\mathbf{x}_*) \leq \frac{1}{\alpha_{1:T}} \mathbb{E} \left[ \sum_{t=1}^T \alpha_t \langle \mathbf{g}(\bar{\mathbf{x}}_t), \mathbf{x}_t - \mathbf{x}_* \rangle \right], \quad (10)$$

where  $\mathbf{x}_* \in \mathcal{X}$  is an arbitrary point in the feasible set. The right-hand-side of Eq. (10) is the expected regret of the online algorithm divided by the total weight  $\alpha_{1:T}$ . To this end, with the gradient-variation adaptivity of optimistic OGD, we present the accelerated result for optimization problem under Hölder smoothness in the following. The proof can be found in Appendix A.4.

**Theorem 2.** Under Assumption 1, for the optimization problem of  $\min_{\mathbf{x} \in \mathcal{X}} \ell(\mathbf{x})$  with stochastic setting, assume the objective  $\ell$  is convex and  $(L_\nu, \nu)$ -Hölder smooth. By employing the online-to-batch conversion algorithm with  $f_t(\mathbf{x}) \triangleq \alpha_t \langle \mathbf{g}(\bar{\mathbf{x}}_t), \mathbf{x} \rangle$  and  $\alpha_t = t$  for all  $t \in [T]$ , instantiating the online algorithm  $\mathcal{A}$  as optimistic OGD in Eq. (6) with  $M_1 = \mathbf{0}$ ,  $M_t = \alpha_t \mathbf{g}(\bar{\mathbf{x}}_t)$  and  $\tilde{\mathbf{x}}_t = \frac{1}{\alpha_{1:t}} (\sum_{s=1}^{t-1} \alpha_s \mathbf{x}_s + \alpha_t \mathbf{x}_{t-1})$ , setting the step size  $\eta_t = D/(2\sqrt{A_{t-1}})$ , we have the convergence rate for any  $\mathbf{x}_* \in \mathcal{X}$ ,

$$\mathbb{E}[\ell(\bar{\mathbf{x}}_T)] - \ell(\mathbf{x}_*) \leq \mathcal{O} \left( \frac{L_\nu D^{1+\nu}}{T^{\frac{1+3\nu}{2}}} + \frac{\sigma D}{\sqrt{T}} + \frac{D \|\nabla \ell(\mathbf{x}_1)\|}{T^2} \right).$$

For  $L$ -smooth and  $G$ -Lipschitz functions, our results recover the rates  $\mathcal{O}(LD^2/T^2 + \sigma D/\sqrt{T})$  and  $\mathcal{O}((G + \sigma)D/\sqrt{T})$ , thus aligning with those of UniXGrad [Kavis et al., 2019], the first optimal and (weakly) universal method. Moreover, our method is actually *strongly universal* due to its adaptivity to more general Hölder smoothness, matching the best-known results of Rodomanov et al. [2024].

**Remark 2.** Besides the constrained setting, recent studies investigated Hölder smoothness in the more challenging unconstrained optimization. Orabona [2023] leveraged a strongly universal method and achieved a non-accelerated rate of  $\mathcal{O}(L_\nu \|\mathbf{x}_*\|^{1+\nu}/T^{(1+\nu)/2})$ , while Li and Lan [2023] obtained an accelerated  $\mathcal{O}(L_\nu \|\mathbf{x}_*\|^{1+\nu}/T^{(1+3\nu)/2})$  with the pre-specified accuracy and appropriate initialization. We leave extending our method to the unconstrained optimization as an interesting future direction.  $\triangleleft$

## 4 Strongly Convex Optimization with Hölder Smoothness

This section considers strongly convex optimization with Hölder smoothness. Section 4.1 provides the first gradient-variation regret in this case, and Section 4.2 leverages it to obtain a universal method.

### 4.1 Gradient-Variation Online Strongly Convex Optimization with Hölder Smoothness

In this part we study the case where the online functions are strongly convex and Hölder smooth. A function  $f : \mathcal{X} \rightarrow \mathbb{R}$  is  $\lambda$ -strongly convex with respect to  $\ell_2$ -norm if it satisfies  $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\lambda}{2} \|\mathbf{x} - \mathbf{y}\|_2^2$  for all  $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ . In the following theorem we obtain the online gradient-variation adaptivity with Hölder smoothness, with the proof in Appendix B.1.

---

**Algorithm 1** Universal Accelerated Strongly Convex Optimization

---

**Input:** Strong convexity curvature  $\lambda$ , total iteration number  $T$ .

1: **Initialization:**  $\alpha_1 = 1, \bar{L}_2 = \lambda, \bar{L} \triangleq \lambda/(2 \exp((\ln T)/T) - 1)^2, \bar{\mathbf{x}}_1 = \mathbf{x}_1 \in \mathcal{X}, M_1 = \mathbf{0}$ .

2: **for**  $t = 1$  **to**  $T$  **do**

3:    $\alpha_{t+1} = \alpha_{1:t} \cdot \sqrt{\lambda/(4\bar{L}_{t+1})}$  with  $\bar{L}_{t+1}$  specified in (11)

4:    $\mathbf{g}_t = \alpha_t \nabla \ell(\bar{\mathbf{x}}_t) + \lambda \alpha_t (\mathbf{x}_t - \bar{\mathbf{x}}_t)$  with  $\bar{\mathbf{x}}_t = \frac{1}{\alpha_{1:t}} (\sum_{s=1}^{t-1} \alpha_s \mathbf{x}_s + \alpha_t \mathbf{x}_t)$

5:    $M_{t+1} = \alpha_{t+1} \nabla \ell(\bar{\mathbf{x}}_t) + \lambda \alpha_{t+1} (\mathbf{x}_t - \tilde{\mathbf{x}}_{t+1})$  with  $\tilde{\mathbf{x}}_{t+1} = \frac{1}{\alpha_{1:t+1}} (\sum_{s=1}^t \alpha_s \mathbf{x}_s + \alpha_{t+1} \mathbf{x}_t)$

6:    $\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}[\mathbf{x}_t - \eta_t (\mathbf{g}_t - M_{t+1})]$  with  $\eta_t = \frac{1}{\lambda \alpha_{1:t}}$

7: **end for**

**Output:**  $\arg \min_{\bar{\mathbf{x}}_t: t \in [T]} \ell(\bar{\mathbf{x}}_t)$

---

242 **Theorem 3.** Under Assumption 1, assuming online functions are  $\lambda$ -strongly convex and  $(L_\nu, \nu)$ -  
243 Hölder smooth, optimistic OGD in Eq. (6) with  $M_1 = \mathbf{0}$ ,  $M_t = \nabla f_{t-1}(\mathbf{x}_{t-1})$  for all  $t \geq 2$ , and step  
244 size  $\eta_t = \frac{3}{\lambda t}$  for all  $t \in [T]$ , ensures the following regret bound:

$$\text{REG}_T \leq \mathcal{O} \left( \frac{\hat{G}_{\max}^2}{\lambda} \log \left( 1 + \frac{V_T}{\hat{G}_{\max}^2} \right) + \frac{L_\nu^2 D^{2\nu}}{\lambda} (\log T)^{1-\nu} + \frac{\|\nabla f_1(\mathbf{x}_1)\|^2}{\lambda} \right),$$

245 where  $\hat{G}_{\max}^2 \triangleq \max_{t \in [T-1]} \sup_{\mathbf{x} \in \mathcal{X}} \|\nabla f_t(\mathbf{x}) - \nabla f_{t+1}(\mathbf{x})\|^2$ .

246 This theorem recovers best-known results under both smoothness and Lipschitzness. In the  $L$ -smooth  
247 case, we recover the best known regret of  $\mathcal{O}(\frac{\hat{G}_{\max}^2}{\lambda} \log(1 + V_T/\hat{G}_{\max}^2) + \frac{1}{\lambda} L^2 D^2)$  [Chen et al., 2024];  
248 and in the  $G$ -Lipschitz case, we also recovers the optimal regret of  $\mathcal{O}(\frac{G^2}{\lambda} \log T)$  [Hazan et al., 2007;  
249 Abernethy et al., 2008].

## 250 4.2 Implication to Offline Strongly Convex Optimization

251 In this part, we develop a weakly universal algorithm for deterministic strongly convex optimization.  
252 This is achieved by leveraging the gradient-variation adaptivity with a carefully specified online-to-  
253 batch conversion, and a smoothness-detection scheme with a novel analysis.

254 We first introduce the intuition of our method. Denoting by  $L_\ell$  the smoothness parameter of the  
255 objective  $\ell(\cdot)$ , which is *unknown*. Recall that in the online-to-batch conversion, the gradient of the  
256 online function  $\nabla f_t(\cdot)$  includes the term  $\alpha_t \nabla \ell(\bar{\mathbf{x}}_t)$ . Consequently, by designing an appropriate  
257 optimism, we obtain an empirical gradient variations related to  $\|\alpha_t \nabla \ell(\bar{\mathbf{x}}_t) - \alpha_t \nabla \ell(\bar{\mathbf{x}}_{t-1})\|^2$ , which  
258 is at most  $2\alpha_t^2 L_\ell \mathcal{D}_\ell(\bar{\mathbf{x}}_{t-1}, \bar{\mathbf{x}}_t)$  by smoothness [Nesterov, 2018, Theorem 2.1.5]. Therefore, the proper  
259 setting for  $\alpha_t$  relies on the smoothness parameter  $L_\ell$ , which is unknown. This motivates us to directly  
260 use the *empirical smoothness*  $L_t \triangleq \|\nabla \ell(\bar{\mathbf{x}}_t) - \nabla \ell(\bar{\mathbf{x}}_{t-1})\|^2 / (2\mathcal{D}_\ell(\bar{\mathbf{x}}_{t-1}, \bar{\mathbf{x}}_t))$  for  $L_\ell$ . However, we  
261 cannot directly use  $L_t$  to set  $\alpha_t$ , because  $\alpha_t$  is required just before  $L_t$  is observed in the algorithmic  
262 updates. To this end, we instead propose an estimate for  $L_t$ , denoted by  $\hat{L}_t$ , to set the  $\alpha_t$ . After  $L_t$   
263 is calculated, we determine whether to update our estimate in the next iteration based on its relationship  
264 with  $L_t$ , which can be formulated as:

$$\forall t \geq 2, \quad \hat{L}_{t+1} \triangleq \begin{cases} \min\{\bar{L}, 4\hat{L}_t\}, & \text{if } \hat{L}_t < L_t \triangleq \frac{\|\nabla \ell(\bar{\mathbf{x}}_t) - \nabla \ell(\bar{\mathbf{x}}_{t-1})\|^2}{2\mathcal{D}_\ell(\bar{\mathbf{x}}_{t-1}, \bar{\mathbf{x}}_t)}, \\ \hat{L}_t, & \text{otherwise.} \end{cases} \quad (11)$$

265 The intuition behind the detection scheme is — when our estimate  $\hat{L}_t < L_t \leq L_\ell$ , it means there is  
266 still a gap between our estimated  $\hat{L}_t$  and the underlying  $L_\ell$ , so we multiply it by a constant factor. And  
267 when  $L_t \geq L_t$ , our estimate may already be close to  $L_\ell$  so we keep using it. Therefore, our estimate  
268  $\hat{L}_t$  automatically adapts to  $\mathcal{O}(L_\ell)$  on the fly within about  $\mathcal{O}(\log L_\ell)$  iterations of increasing due to  
269 the geometry increasing rate. Moreover, in the case of the non-smoothness where the arbitrarily large  
270  $L_t$  would ruin our estimate, we design a threshold  $\bar{L}$  in the meanwhile, to safeguard a non-smooth  
271 convergence rate of  $\tilde{\mathcal{O}}(1/T)$ . Algorithm 1 summarizes the updates and we provide the convergence  
272 guarantee in Theorem 4 with the proof in Appendix B.3.

273 **Theorem 4.** For the optimization problem of  $\min_{\mathbf{x} \in \mathcal{X}} \ell(\mathbf{x})$  with deterministic setting, assume the  
274 objective  $\ell(\cdot)$  is  $\lambda$ -strongly convex. Then when  $\ell(\cdot)$  is  $G$ -Lipchitz, Algorithm 1 ensures

$$\min_{t \in [T]} \ell(\bar{\mathbf{x}}_t) - \ell(\mathbf{x}_*) \leq \mathcal{O} \left( \frac{G^2 \log T}{\lambda T} \right).$$

275 When  $\ell(\cdot)$  is  $L_\ell$ -smooth, denoting by  $\kappa \triangleq L_\ell/\lambda$ , [Algorithm 1](#) also ensures:

$$\min_{t \in [T]} \ell(\bar{\mathbf{x}}_t) - \ell(\mathbf{x}_*) \leq \mathcal{O} \left( \frac{\|\nabla \ell(\mathbf{x}_1)\|^2 + \hat{G}_{\max}^2}{\lambda} \exp \left( \frac{-T}{(1 + 2\sqrt{\kappa}) \lceil \log_2 \sqrt{\kappa} \rceil} \right) \right),$$

276 where  $\hat{G}_{\max}^2 \triangleq \max_{t \in [T-1]} \|\nabla \ell(\bar{\mathbf{x}}_t) - \nabla \ell(\bar{\mathbf{x}}_{t+1})\|^2$  is at most  $L_\ell^2 D^2$  under [Assumption 1](#), or it can  
277 be upper bounded by  $4G^2$  if  $\ell(\cdot)$  is  $G$ -Lipchitz.

278 [Theorem 4](#) is *weakly universal*, i.e., it maintains the respective near-optimal convergence rates in  
279 both smooth and Lipschitz cases, without knowing parameters  $L_\ell$  or  $G$ . Moreover, if provided  
280 with  $L_\ell$ , then by setting  $\alpha_t \triangleq \alpha_{1:t-1} \cdot \sqrt{\lambda/(4L_\ell)}$ , we can achieve a better convergence rate of  
281  $\mathcal{O}(\exp(-T/(1 + 2\sqrt{\kappa})))$  proved in [Appendix B.4](#). Interestingly, it matches the result of [Wei and](#)  
282 [Chen \[2025\]](#), where their “over relaxation” coincides with our *one-step* variant of optimistic OGD.

283 **Remark 3.** To the best of our knowledge, [Levy \[2017\]](#) is the previously best-known universal  
284 for strongly convex optimization, in which an adaptive normalized gradient descent is employed  
285 with online-to-batch conversion weights inversely proportional to the square of the gradient norm.  
286 For deterministic strongly convex optimization over a bounded domain, [Levy \[2017\]](#) achieved an  
287  $\mathcal{O}((\log T)/T)$  convergence rate for the Lipschitz function, and an  $\mathcal{O}(\exp(-T/\kappa) \cdot T/\kappa)$  rate for a  
288 smooth and Lipschitz objective. Our work improves upon their result by achieving the *first* accelerated  
289 rate of  $\mathcal{O}(\exp(-T/(\sqrt{\kappa} \cdot \log \kappa)))$  for the smooth function in universal strongly convex optimization.  
290 However, our method relies on a smoothness detection scheme based on the observed gradients, which  
291 only works in the deterministic setting. Extending it to the stochastic setting remains challenging.  $\triangleleft$

292 **Remark 4.** Designing a *strongly universal* method for strongly convex optimization, i.e., adapting  
293 to Hölder smoothness, remains an open problem. Notably, given the prior knowledge of the Hölder  
294 smoothness parameter, [Devolder et al. \[2013\]](#) have established a sample-complexity-based rate that  
295 recovers the (near-)optimal rate for smooth and non-smooth cases, which may serve as a starting  
296 point. Furthermore, the aforementioned works [[Levy, 2017](#); [Devolder et al., 2013](#)] and ours assume a  
297 bounded domain. While [Lan et al. \[2023\]](#) have proposed a universal method for the unbounded setting,  
298 it still relies on a pre-specified initialization related to the actual level of smoothness. Developing a  
299 fully universal method in the unconstrained setting remains a significant challenge.  $\triangleleft$

## 300 5 Complex Regret Measures in Modern Online Learning

301 In [Section 5.1](#), we study the non-stationary online learning with Hölder smoothness. In [Section 5.2](#),  
302 we investigate the universal online learning with Hölder smoothness.

### 303 5.1 Dynamic Regret with Hölder Smoothness

304 To optimize the dynamic regret in [Eq. \(4\)](#) with Hölder smoothness, we follow the best known method  
305 of [Zhao et al. \[2024\]](#) but perform more refined design and analysis to make it *strongly universal*.  
306 Specifically, we retain the original meta-base two-layer structure, consisting of (i) a group of  $N$   
307 base-learners maintained simultaneously, each runs the optimistic OGD algorithm with *time-varying*  
308 step sizes; and (ii) a meta-algorithm to adaptively combine the outputs of base-learners. The key  
309 improvement to achieve strong universality lies in the adaptive step sizes for both meta-algorithm and  
310 base-learners, with a *virtual clipping* in the analysis following the same spirit as [Eq. \(9\)](#).

311 The algorithm updates as follows. At iteration  $t$ , the decision  $\mathbf{x}_t$  is generated by weighted combining  
312 all base-learners’ local decisions  $\{\mathbf{x}_{t,i}\}_{i=1}^N$  with the meta-algorithm’s weight vector  $\mathbf{p}_t \in \Delta_N$ , i.e.,  
313  $\mathbf{x}_t \triangleq \sum_{i=1}^N p_{t,i} \mathbf{x}_{t,i}$ . After submitting  $\mathbf{x}_t$ , the  $i$ -th base-learner updates its local decision by:

$$\hat{\mathbf{x}}_{t+1,i} = \Pi_{\mathcal{X}} [\hat{\mathbf{x}}_{t,i} - \eta_{t,i} \nabla f_t(\mathbf{x}_t)], \quad \mathbf{x}_{t+1,i} = \Pi_{\mathcal{X}} [\hat{\mathbf{x}}_{t+1,i} - \eta_{t+1,i} \nabla f_t(\mathbf{x}_t)], \quad (12)$$

314 where  $\eta_{t,i}$  is the *time-varying* step size. The meta-algorithm, starting with  $\mathbf{p}_1$  where  $p_{1,i} = 1/N$  for  
315 all  $i \in [N]$ , calculates the weight vector  $\mathbf{p}_{t+1} \in \Delta_N$  by optimistic Hedge [[Syrkanis et al., 2015](#)]  
316 with a time-varying learning rate  $\varepsilon_t$ , that is,

$$p_{t+1,i} \propto \exp \left( -\varepsilon_t \left( \sum_{s=1}^t \ell_{s,i} + m_{t+1,i} \right) \right), \quad \begin{cases} \ell_{t,i} \triangleq \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_{t,i} \rangle + \lambda_{t,i} c_{t,i}, \\ m_{t+1,i} \triangleq \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_{t,i} \rangle + \lambda_{t+1,i} c_{t+1,i}, \end{cases} \quad (13)$$



with the feedback loss  $\ell_t \in \mathbb{R}^N$  and optimism  $\mathbf{m}_{t+1} \in \mathbb{R}^N$ , where  $c_{t,i} \triangleq \|\mathbf{x}_{t,i} - \mathbf{x}_{t-1,i}\|^2$  is the correction term following Zhao et al. [2024], with coefficient  $\lambda_{t,i} = (8\eta_{t,i})^{-1}$  for all  $t \geq 2$  and  $c_{1,i} \triangleq 0$ . Consequently, we provide the following gradient-variation dynamic regret bound under Hölder smoothness of our algorithm, and defer the proof and complete algorithm to Appendix C.1.

**Theorem 5.** Under Assumption 1, assuming online functions are convex and  $(L_\nu, \nu)$ -Hölder smooth, algorithm aforementioned consisting of (i)  $N = \lceil \log_2(1+T) \rceil$  base-learners where the  $i$ -th updates by Eq. (12), with the step size  $\eta_{t,i} = 2^{i-1}D/\sqrt{A_{t-1}}$  and  $A_t$  specified in Eq. (9) for all  $t \in [T]$  and  $i \in [N]$ ; and (ii) meta-algorithm that updates by Eq. (13) with the learning rate  $\varepsilon_t = \sqrt{(3 + \ln N)/(D^2 A_{t-1})}$ , ensures the following dynamic regret (as defined in Eq. (4)) bound:

$$\text{D-REG}_T(\mathbf{u}_1, \dots, \mathbf{u}_T) \leq \mathcal{O}\left(\sqrt{(D^2 + DP_T)}\left(\sqrt{V_T} + \|\nabla f_1(\mathbf{x}_1)\|\right) + L_\nu(D^2 + DP_T)^{\frac{1+\nu}{2}} T^{\frac{1-\nu}{2}}\right).$$

Theorem 5 enjoys optimal results for both smooth and Lipschitz continuous online functions. Specifically, when the online functions are smooth, i.e.,  $\nu = 1$ , we recover the state-of-the-art result of  $\mathcal{O}(\sqrt{V_T(1+P_T)} + P_T)$  [Zhao et al., 2020, 2024]; and when the online functions are Lipschitz, i.e.,  $\nu = 0$ , we obtain the optimal result of  $\mathcal{O}(\sqrt{T(1+P_T)})$  [Zhang et al., 2018].

**Remark 5** (Improved Adaptivity). Our method builds primarily on Zhao et al. [2024], but requires fewer parameters, making it more adaptive. Specifically, their method deploys base-learners with fixed-step-size optimistic OGD, where the choice of step sizes requires the smoothness parameter  $L$ . In contrast, our method can achieve the gradient-variation dynamic regret under more general Hölder smooth functions, yet avoiding the need for Hölder smoothness parameters. This is achieved by adaptive time-varying step sizes with more carefully designed configurations for base-learners.  $\triangleleft$

## 5.2 Universal Regret with Hölder Smoothness

We now consider universal online learning, where the homogeneous online functions are either convex or  $\lambda$ -strongly convex, but the curvature information is *unknown* to algorithm. We find that the state-of-the-art work of Yan et al. [2024] is already capable of handling with Hölder smoothness with a more refined analysis, which has not been realized before. This is summarized in the following theorem, with the complete algorithm and proof in Appendix C.2.

**Theorem 6.** Under Assumption 1, assuming online functions are  $(L_\nu, \nu)$ -Hölder smooth, and convex or  $\lambda$ -strongly convex with  $\lambda \in [1/T, 1]$  but without the knowledge of the type. The algorithm proposed in Yan et al. [2024, Algorithm 1] specified for convex and strongly convex functions, denoted by  $\mathcal{A}$ , ensures the following universal regret (as defined in Eq. (5)) bound:

$$\text{U-REG}_T^{\mathcal{A}}(\mathcal{F}_c) \leq \mathcal{O}\left(D\sqrt{V_T} + L_\nu D^{1+\nu} T^{\frac{1-\nu}{2}} + D\|\nabla f_1(\mathbf{x}_1)\|\right).$$

and

$$\text{U-REG}_T^{\mathcal{A}}(\mathcal{F}_{sc}^\lambda) \leq \mathcal{O}\left(\frac{\hat{G}_{\max}^2}{\lambda} \ln\left(1 + \frac{V_T}{\hat{G}_{\max}^2}\right) + \frac{L_\nu^2 D^{2\nu}}{\lambda} (\log T)^{\frac{1-\nu}{1+\nu}} + \frac{\max_{t \in [T]} \sup_{\mathbf{x} \in \mathcal{X}} \|\nabla f_t(\mathbf{x})\|^2}{\lambda}\right),$$

where  $\hat{G}_{\max}^2 \triangleq \max_{t \in [T-1]} \sup_{\mathbf{x} \in \mathcal{X}} \|\nabla f_t(\mathbf{x}) - \nabla f_{t+1}(\mathbf{x})\|^2$ .

This theorem recovers not only best-known results under both smoothness and Lipschitz continuity, but also an interpolated regret bound with Hölder smoothness, without any information of online functions' curvature and Hölder smoothness parameters. To achieve this result, we import another property of Hölder smoothness [Rodomanov et al., 2024, Theorem A.2] with a more careful derivation.

## 6 Conclusion

In this work, we investigate gradient-variation online adaptivity with Hölder smoothness and its implications to offline optimization. For online learning with Hölder smoothness, we propose the first gradient-variation regret bounds for convex and strongly convex online functions, and further extend to non-stationary online learning and universal online learning for more robust results. For offline optimization, our convergence rates match the existing optimal result for convex functions, and significantly improve the non-accelerated rate for strongly convex functions.

An interesting future direction is to study gradient-variation online learning with Hölder smoothness in unconstrained setting, and further apply it to offline optimization. Another important direction is to better develop offline optimization algorithms by utilizing more online adaptivity results.

## References

- J. D. Abernethy, P. L. Bartlett, A. Rakhlin, and A. Tewari. Optimal strategies and minimax lower bounds for online convex games. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, pages 415–424, 2008.
- N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transaction on Information Theory*, 50(9):2050–2057, 2004.
- G. Chen and M. Teboulle. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.
- S. Chen, Y.-J. Zhang, W.-W. Tu, P. Zhao, and L. Zhang. Optimistic online mirror descent for bridging stochastic and adversarial online convex optimization. *Journal of Machine Learning Research*, 25(178):1 – 62, 2024.
- C.-K. Chiang, T. Yang, C.-J. Lee, M. Mahdavi, C.-J. Lu, R. Jin, and S. Zhu. Online optimization with gradual variations. In *Proceedings of the 25th Conference on Learning Theory (COLT)*, pages 6.1–6.20, 2012.
- A. Cutkosky. Anytime online-to-batch, optimism and acceleration. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 1446–1454, 2019.
- S. de Rooij, T. van Erven, P. D. Grünwald, and W. M. Koolen. Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15(1):1281–1316, 2014.
- O. Devolder, F. Glineur, and Y. Nesterov. First-order methods with inexact oracle: the strongly convex case. *CORE Discussion Papers*, 2013016:47, 2013.
- O. Devolder, F. Glineur, and Y. Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146:37–75, 2014.
- D. J. Foster, A. Rakhlin, and K. Sridharan. Adaptive online learning. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pages 3375–3383, 2015.
- E. Hazan. *Introduction to Online Convex Optimization*. MIT Press, 2nd edition, 2022.
- E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.
- P. Joulani, A. Györfy, and C. Szepesvári. A modular analysis of adaptive (non-)convex optimization: Optimism, composite objectives, variance reduction, and variational bounds. *Theoretical Computer Science*, 808:108–138, 2020.
- A. Kavis, K. Y. Levy, F. R. Bach, and V. Cevher. UniXGrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 6257–6266, 2019.
- G. Lan, Y. Ouyang, and Z. Zhang. Optimal and parameter-free gradient minimization methods for convex and nonconvex optimization. *arXiv preprint arXiv:2310.12139*, 2023.
- K. Levy. Online to offline conversions, universality and adaptive minibatch sizes. *Advances in Neural Information Processing Systems*, 30:1613–1622, 2017.
- T. Li and G. Lan. A simple uniformly optimal method without line search for convex optimization. *ArXiv preprint*, arxiv:2310.10082, 2023.
- H. Luo and R. E. Schapire. Achieving all with no parameters: AdaNormalHedge. In *Proceedings of the 28th Annual Conference Computational Learning Theory (COLT)*, pages 1286–1304, 2015.
- H. B. McMahan and M. J. Streeter. Adaptive bound optimization for online convex optimization. In *Proceedings of the 23rd Conference on Learning Theory (COLT)*, pages 244–256, 2010.

- 405 Y. Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Program-*  
406 *ming*, 152(1):381–404, 2015.
- 407 Y. Nesterov. *Lectures on Convex Optimization*, volume 137. Springer, 2018.
- 408 F. Orabona. A Modern Introduction to Online Learning. *arXiv preprint arXiv:1912.13213*, 2019.
- 409 F. Orabona. Normalized gradients for all. *arXiv preprint arXiv:2308.05621*, 2023.
- 410 A. Rodomanov, X. Jiang, and S. U. Stich. Universality of adagrad stepsizes for stochastic optimization:  
411 Inexact oracle, acceleration and variance reduction. *Advances in Neural Information Processing*  
412 *Systems*, 37:26770–26813, 2024.
- 413 S. Sachs, H. Hadiji, T. van Erven, and C. A. Guzmán. Between stochastic and adversarial online  
414 convex optimization: Improved regret bounds via smoothness. In *Advances in Neural Information*  
415 *Processing Systems 35 (NeurIPS)*, pages 691–702, 2022.
- 416 S. Sachs, H. Hadiji, T. van Erven, and C. Guzman. Accelerated rates between stochastic and  
417 adversarial online convex optimization. *ArXiv preprint*, arXiv:2303.03272, 2023.
- 418 V. Syrgkanis, A. Agarwal, H. Luo, and R. E. Schapire. Fast convergence of regularized learning  
419 in games. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pages 2989–2997,  
420 2015.
- 421 T. van Erven and W. M. Koolen. Metagrad: Multiple learning rates in online learning. In *Advances*  
422 *in Neural Information Processing Systems 29 (NIPS)*, pages 3666–3674, 2016.
- 423 G. Wang, S. Lu, and L. Zhang. Adaptivity and optimality: A universal algorithm for online convex  
424 optimization. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*  
425 *(UAI)*, pages 659–668, 2019.
- 426 C.-Y. Wei, Y.-T. Hong, and C.-J. Lu. Tracking the best expert in non-stationary stochastic envi-  
427 ronments. In *Advances in Neural Information Processing Systems 29 (NIPS)*, pages 3972–3980,  
428 2016.
- 429 J. Wei and L. Chen. Accelerated over-relaxation heavy-ball method: Achieving global accelerated  
430 convergence with broad generalization. In *The Thirteenth International Conference on Learning*  
431 *Representations*, 2025. URL <https://openreview.net/forum?id=SWEqzy7IQB>.
- 432 Y.-F. Xie, P. Zhao, and Z.-H. Zhou. Gradient-variation online learning under generalized smoothness.  
433 In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, pages 37865–37899, 2024.
- 434 Y.-H. Yan, P. Zhao, and Z.-H. Zhou. Universal online learning with gradient variations: A multi-layer  
435 online ensemble approach. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*,  
436 pages 37682–37715, 2023.
- 437 Y.-H. Yan, P. Zhao, and Z.-H. Zhou. A simple and optimal approach for universal online learning  
438 with gradient variations. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*,  
439 pages 11132–11163, 2024.
- 440 T. Yang, M. Mahdavi, R. Jin, and S. Zhu. Regret bounded by gradual variation for online convex  
441 optimization. *Machine Learning*, 95(2):183–223, 2014.
- 442 L. Zhang, S. Lu, and Z.-H. Zhou. Adaptive online learning in dynamic environments. In *Advances in*  
443 *Neural Information Processing Systems 31 (NeurIPS)*, pages 1330–1340, 2018.
- 444 M. Zhang, P. Zhao, H. Luo, and Z.-H. Zhou. No-regret learning in time-varying zero-sum games. In  
445 *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pages 26772–  
446 26808, 2022.
- 447 P. Zhao, Y.-J. Zhang, L. Zhang, and Z.-H. Zhou. Dynamic regret of convex and smooth functions. In  
448 *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 12510–12520, 2020.

- 449 P. Zhao, Y.-J. Zhang, L. Zhang, and Z.-H. Zhou. Adaptivity and non-stationarity: Problem-dependent  
450 dynamic regret for online convex optimization. *Journal of Machine Learning Research*, 25(98):1 –  
451 52, 2024.
- 452 M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In  
453 *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pages 928–936,  
454 2003.

## A Omitted Details for Section 3

In this section, we first provide some useful lemmas for Hölder smoothness, then give the proofs of theorems in Section 3.

### A.1 Useful Lemmas for Hölder Smoothness

This part provides several useful lemmas for Hölder smoothness.

**Lemma 2** (Lemma 1 of [Nesterov \[2015\]](#)). *Let convex function  $f : \mathcal{X} \rightarrow \mathbb{R}$  over the convex set  $\mathcal{X}$  be  $(L_\nu, \nu)$ -Hölder smooth.<sup>2</sup> Then for any  $\delta > 0$ , denoting by  $L = \delta^{\frac{\nu-1}{1+\nu}} L_\nu^{\frac{2}{1+\nu}}$ , for all  $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ :*

$$f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2 + \delta. \quad (14)$$

**Lemma 3** (Theorem 1 of [Devolder et al. \[2014\]](#)). *If convex function  $f : \mathcal{X} \rightarrow \mathbb{R}$  over the convex set  $\mathcal{X}$  satisfies that, there exists positive constants  $L$  and  $\delta$  such that, for all  $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ :*

$$f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2 + \delta, \quad (15)$$

then for all  $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ :

$$\frac{1}{2L} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \leq f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \delta. \quad (16)$$

**Lemma 4** (Theorem A.2. of [Rodomanov et al. \[2024\]](#)). *If convex function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  over  $\mathbb{R}^d$  satisfies that, there exists positive constants  $L$  and  $\delta$  such that, for all  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ :*

$$f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2 + \delta, \quad (17)$$

then for all  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ :

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \leq 2LD_f(\mathbf{x}, \mathbf{y}) + 2L\delta. \quad (18)$$

### A.2 Proof of Lemma 1

*Proof.* Since  $f$  is  $(L_\nu, \nu)$ -Hölder smooth, by combining [Lemma 2](#) and [Lemma 3](#), for any  $\delta > 0$ , denoting by  $L = \delta^{\frac{\nu-1}{1+\nu}} L_\nu^{\frac{2}{1+\nu}}$ , for all  $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ :

$$\frac{1}{2L} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \stackrel{(16)}{\leq} f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \delta \stackrel{(14)}{\leq} \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2 + 2\delta. \quad (19)$$

Multiplying both sides of the inequality by  $2L$  completes the proof.  $\square$

### A.3 Proof of Theorem 1

*Proof.* Applying [Lemma 12](#) with comparators  $\mathbf{u}_t = \mathbf{x}_\star = \arg \min_{\mathbf{x} \in \mathcal{X}} \sum_{t=1}^T f_t(\mathbf{x})$  for all  $t \in [T]$ ,

$$\begin{aligned} \text{REG}_T &= \sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{x}_\star) \leq \sum_{t=1}^T \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_\star \rangle \\ &\leq \sum_{t=1}^T \eta_{t+1} \|\nabla f_t(\mathbf{x}_t) - M_t\|^2 + \frac{D^2}{\eta_{T+1}} - \sum_{t=2}^T \frac{1}{8\eta_{t+1}} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 \\ &\leq 3D\sqrt{A_T} - \sum_{t=2}^T \frac{1}{8\eta_{t+1}} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2, \end{aligned} \quad (20)$$

where  $A_t \triangleq \|\nabla f_1(\mathbf{x}_1)\|^2 + \sum_{s=2}^t \|\nabla f_s(\mathbf{x}_s) - M_s\|^2$ , and we apply [Lemma 14](#) in the last line.

<sup>2</sup>Though  $\mathcal{X}$  is supposed to be closed in [Nesterov \[2015\]](#), this lemma holds for  $\mathcal{X} = \mathbb{R}^d$  with the same proof.



475 If  $\sqrt{A_T} \leq 2LD$ , we finish the proof trivially, so in the following, we assume  $\sqrt{A_T} > 2LD$ .  
 476 Define  $t_0$  that, if  $\sqrt{A_1} > 2LD$ , let  $t_0 = 1$ , otherwise let  $t_0 = \min\{t : t \in [T-1], \sqrt{A_{t+1}} > 2LD\}$ .  
 477 Then we have  $\sqrt{A_{t_0}} \leq \|\nabla f_1(\mathbf{x}_1)\| + 2LD$ , while for all  $t_0 + 1 \leq t \leq T$  it holds that  $\sqrt{A_t} > 2LD$ .  
 478 Because all online functions are  $(L_\nu, \nu)$ -Hölder smooth and applying [Lemma 1](#), we show the  
 479 following decomposition for  $\alpha\sqrt{A_T}$  with constant  $\alpha > 0$ . For any  $\delta > 0$  that only exists in analysis,  
 480 denoting by  $L = \delta^{\frac{\nu-1}{1+\nu}} L_\nu^{\frac{2}{1+\nu}}$ :

$$\begin{aligned} \alpha\sqrt{A_T} &\leq \alpha\sqrt{A_{t_0}} + \alpha\sqrt{\sum_{t=t_0+1}^T \|\nabla f_t(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_t) + \nabla f_{t-1}(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_{t-1})\|_*^2} \\ &\leq \alpha\sqrt{A_{t_0}} + \alpha\sqrt{2V_T} + \alpha\sqrt{2L^2 \sum_{t=t_0+1}^T \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + 8L \sum_{t=t_0+1}^T \delta} \\ &\leq \alpha\sqrt{A_{t_0}} + \alpha\sqrt{2V_T} + \alpha^2 L + \frac{L}{2} \sum_{t=t_0+1}^T \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + \alpha\sqrt{8L\delta T}. \end{aligned}$$

481 With this decomposition, we prove the regret bound in the following with  $\alpha = 3D$ :

$$\begin{aligned} \text{REG}_T &\leq 3D\sqrt{A_{t_0}} + 3D\sqrt{2V_T} + 9LD^2 + \sum_{t=t_0+1}^T \left( \frac{L}{2} - \frac{1}{8\eta_{t+1}} \right) \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + 3D\sqrt{8L\delta T} \\ &\leq 3D\sqrt{2V_T} + 15LD^2 + 3D\|\nabla f_1(\mathbf{x}_1)\| + 3D\sqrt{8L\delta T}. \end{aligned}$$

482 Then by choosing  $\delta = L_\nu D^{1+\nu} T^{-\frac{1+\nu}{2}}$  (that only exists in analysis), we obtain

$$\text{REG}_T \leq \mathcal{O} \left( D\sqrt{V_T} + L_\nu D^{1+\nu} T^{\frac{1-\nu}{2}} + D\|\nabla f_1(\mathbf{x}_1)\| \right),$$

483 which completes the proof.  $\square$

#### 484 A.4 Proof of Theorem 2

485 *Proof.* With optimistic OGD as the online algorithm, by defining  $f_t(\mathbf{x}) \triangleq \langle \alpha_t \mathbf{g}(\bar{\mathbf{x}}_t), \mathbf{x} \rangle$ , we have:

$$\sum_{t=1}^T \alpha_t \langle \mathbf{g}(\bar{\mathbf{x}}_t), \mathbf{x}_t - \mathbf{x}_* \rangle = \sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{x}_*) \stackrel{(20)}{\leq} 3D\sqrt{A_T} - \sum_{t=2}^T \frac{1}{8\eta_{t+1}} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2.$$

486 Now we trivially assume  $\sqrt{A_T} > 4LD$ , and define  $t_0 \in [T-1]$  that, if  $\sqrt{A_1} > 4LD$ , let  $t_0 = 1$ ,  
 487 otherwise let  $t_0 = \min\{t : t \in [T-1], \sqrt{A_{t+1}} > 4LD\}$ . Then we have  $\sqrt{A_{t_0}} \leq \|\nabla f_1(\mathbf{x}_1)\| + 4LD$ ,  
 488 while for all  $t_0 + 1 \leq t \leq T$  it holds that  $\sqrt{A_t} > 4LD$ . Continuing with our previous inequality:

$$\begin{aligned} &\sum_{t=1}^T \alpha_t \langle \mathbf{g}(\bar{\mathbf{x}}_t), \mathbf{x}_t - \mathbf{x}_* \rangle \\ &\leq 3D\sqrt{A_{t_0}} + 3D\sqrt{\sum_{t=t_0+1}^T \alpha_t^2 \|\mathbf{g}(\bar{\mathbf{x}}_t) - \mathbf{g}(\tilde{\mathbf{x}}_t)\|^2} - \sum_{t=2}^T \frac{1}{8\eta_{t+1}} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 \\ &\leq 3D\sqrt{A_{t_0}} + 3D\sqrt{\sum_{t=t_0+1}^T 3\alpha_t^2 \|\nabla \ell(\bar{\mathbf{x}}_t) - \nabla \ell(\tilde{\mathbf{x}}_t)\|^2} - \sum_{t=2}^T \frac{1}{8\eta_{t+1}} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 \\ &\quad + 3D\sqrt{\sum_{t=t_0+1}^T 3\alpha_t^2 \|\nabla \ell(\bar{\mathbf{x}}_t) - \mathbf{g}(\bar{\mathbf{x}}_t)\|^2} + 3D\sqrt{\sum_{t=t_0+1}^T 3\alpha_t^2 \|\nabla \ell(\tilde{\mathbf{x}}_t) - \mathbf{g}(\tilde{\mathbf{x}}_t)\|^2}, \end{aligned}$$

where we use  $\|\mathbf{a} + \mathbf{b} + \mathbf{c}\|^2 \leq 3\|\mathbf{a}\|^2 + 3\|\mathbf{b}\|^2 + 3\|\mathbf{c}\|^2$  for any  $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^d$ . Now by taking expectation and using Jensen's inequality, i.e.,  $\mathbb{E}_x[\sqrt{x}] \leq \sqrt{\mathbb{E}_x[x]}$ , we have

$$\begin{aligned} & \mathbb{E} \left[ \sum_{t=1}^T \alpha_t \langle \mathbf{g}(\bar{\mathbf{x}}_t), \mathbf{x}_t - \mathbf{x}_\star \rangle \right] \\ & \leq \mathbb{E} \left[ 3D \sqrt{\sum_{t=t_0+1}^T 3\alpha_t^2 \|\nabla \ell(\bar{\mathbf{x}}_t) - \nabla \ell(\tilde{\mathbf{x}}_t)\|^2} - \sum_{t=2}^T \frac{1}{8\eta_{t+1}} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 \right] \\ & \quad + 3D \|\nabla \ell(\mathbf{x}_1)\| + 12LD^2 + 12\sqrt{2}\sigma DT^{\frac{3}{2}}, \end{aligned}$$

where we apply  $\mathbb{E}[\|\mathbf{g}(\mathbf{x}) - \nabla \ell(\mathbf{x})\|^2 \mid \mathbf{x}] \leq \sigma^2$ . By Lemma 1 and the definitions of  $\bar{\mathbf{x}}_t, \tilde{\mathbf{x}}_t$ ,

$$\begin{aligned} \alpha_t^2 \|\nabla \ell(\bar{\mathbf{x}}_t) - \nabla \ell(\tilde{\mathbf{x}}_t)\|^2 & \stackrel{(8)}{\leq} \alpha_t^2 L^2 \|\bar{\mathbf{x}}_t - \tilde{\mathbf{x}}_t\|^2 + 4\alpha_t^2 L\delta = \frac{\alpha_t^4 L^2}{\alpha_{1:t}^2} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + 4\alpha_t^2 L\delta \\ & \leq 4L^2 \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + 4t^2 L\delta. \end{aligned}$$

Then we have

$$\begin{aligned} & 3D \sqrt{\sum_{t=t_0+1}^T 3\alpha_t^2 \|\nabla \ell(\bar{\mathbf{x}}_t) - \nabla \ell(\tilde{\mathbf{x}}_t)\|^2} - \sum_{t=2}^T \frac{1}{8\eta_{t+1}} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 \\ & \leq 6D \sqrt{3L^2 \sum_{t=t_0+1}^T \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2} - \sum_{t=2}^T \frac{1}{8\eta_{t+1}} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + 12\sqrt{2}D\sqrt{L\delta}T^{\frac{3}{2}} \\ & \leq 27LD^2 + \sum_{t=t_0+1}^T \left( L - \frac{1}{8\eta_{t+1}} \right) \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + 12\sqrt{2}D\sqrt{L\delta}T^{\frac{3}{2}} \\ & \leq 27LD^2 + 12\sqrt{2}D\sqrt{L\delta}T^{\frac{3}{2}}. \end{aligned}$$

Therefore, by combining the above inequalities we obtain

$$\begin{aligned} \mathbb{E}[\ell(\bar{\mathbf{x}}_T)] - \ell(\mathbf{x}_\star) & \leq \frac{1}{\alpha_{1:T}} \mathbb{E} \left[ \sum_{t=1}^T \alpha_t \langle \mathbf{g}(\bar{\mathbf{x}}_t), \mathbf{x}_t - \mathbf{x}_\star \rangle \right] \\ & \leq \frac{6D \|\nabla \ell(\mathbf{x}_1)\| + 78LD^2}{T^2} + \frac{24\sqrt{2}D\sqrt{L\delta} + 24\sqrt{2}\sigma D}{\sqrt{T}}. \end{aligned}$$

Then by setting  $\delta = L_\nu D^{1+\nu} T^{-\frac{(3+3\nu)}{2}}$ , we achieve the convergence rate of

$$\mathbb{E}[\ell(\bar{\mathbf{x}}_T)] - \ell(\mathbf{x}_\star) \leq \mathcal{O} \left( \frac{L_\nu D^{1+\nu}}{T^{\frac{1+3\nu}{2}}} + \frac{\sigma D}{\sqrt{T}} + \frac{D \|\nabla \ell(\mathbf{x}_1)\|}{T^2} \right),$$

which completes the proof.  $\square$

## B Omitted Details for Section 4

In this section, we give the proofs of theorems in Section 4.

### B.1 Proof of Theorem 3

*Proof.* We apply Lemma 11 with  $\mathbf{u}_t = \mathbf{x}_\star = \arg \min_{\mathbf{x} \in \mathcal{X}} \sum_{t=1}^T f_t(\mathbf{x})$ :

$$\begin{aligned} \text{REG}_T & = \sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{x}_\star) \leq \sum_{t=1}^T \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_\star \rangle - \frac{\lambda}{2} \sum_{t=1}^T \|\mathbf{x}_\star - \mathbf{x}_t\|^2 \\ & \stackrel{(41)}{\leq} \underbrace{\sum_{t=1}^T \frac{1}{2\eta_t} (\|\mathbf{x}_\star - \hat{\mathbf{x}}_t\|^2 - \|\mathbf{x}_\star - \hat{\mathbf{x}}_{t+1}\|^2)}_{\text{TERM-A}} - \frac{\lambda}{2} \sum_{t=1}^T \|\mathbf{x}_\star - \mathbf{x}_t\|^2 \end{aligned}$$

$$+ \underbrace{\sum_{t=1}^T \eta_t \|\nabla f_t(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_{t-1})\|^2}_{\text{TERM-B}} - \underbrace{\sum_{t=1}^T \frac{1}{2\eta_t} (\|\mathbf{x}_t - \hat{\mathbf{x}}_t\|^2 + \|\mathbf{x}_t - \hat{\mathbf{x}}_{t+1}\|^2)}_{\text{TERM-C}}.$$

500 We first investigate TERM-A. Since  $\eta_t = \frac{3}{\lambda t}$ ,

$$\begin{aligned} \text{TERM-A} &\leq \frac{\|\mathbf{x}_\star - \hat{\mathbf{x}}_1\|^2}{2\eta_1} + \sum_{t=2}^T \left( \frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) \|\mathbf{x}_\star - \hat{\mathbf{x}}_t\|^2 - \frac{\lambda}{2} \sum_{t=1}^T \|\mathbf{x}_\star - \mathbf{x}_t\|^2 \\ &\leq \frac{\lambda}{6} \sum_{t=1}^{T-1} (\|\mathbf{x}_\star - \hat{\mathbf{x}}_{t+1}\|^2 - 2\|\mathbf{x}_\star - \mathbf{x}_t\|^2) \leq \frac{\lambda}{3} \sum_{t=1}^{T-1} \|\mathbf{x}_t - \hat{\mathbf{x}}_{t+1}\|^2 \\ &\leq \frac{\lambda}{3} \sum_{t=1}^{T-1} \eta_t^2 \|\nabla f_t(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_{t-1})\|^2 \leq \text{TERM-B}, \end{aligned}$$

501 where in the second line we use  $\hat{\mathbf{x}}_1 = \mathbf{x}_1$ . And in the last line above we apply [Lemma 9](#) [[Chiang et al., 2012](#)]. TERM-C can be bounded as:

$$\text{TERM-C} \geq \sum_{t=2}^T \left( \frac{1}{2\eta_t} \|\mathbf{x}_t - \hat{\mathbf{x}}_t\|^2 + \frac{1}{2\eta_{t-1}} \|\mathbf{x}_{t-1} - \hat{\mathbf{x}}_t\|^2 \right) \geq \sum_{t=2}^T \frac{1}{4\eta_{t-1}} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2.$$

503 Then by combining TERM-A, TERM-B and TERM-C together and applying [Lemma 1](#) with arbitrary

504  $\delta > 0$  that only exists in analysis, and denoting by  $L = \delta^{\frac{\nu-1}{1+\nu}} L_\nu^{\frac{2}{1+\nu}}$ , we obtain:

$$\begin{aligned} \text{REG}_T &\leq \sum_{t=1}^T \frac{6}{\lambda t} \|\nabla f_t(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_{t-1})\|^2 - \sum_{t=2}^T \frac{\lambda(t-1)}{12} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 \quad (21) \\ &\leq \sum_{t=2}^T \frac{12}{\lambda t} (\|\nabla f_t(\mathbf{x}_t) - \nabla f_t(\mathbf{x}_{t-1})\|^2 + \|\nabla f_t(\mathbf{x}_{t-1}) - \nabla f_{t-1}(\mathbf{x}_{t-1})\|^2) \\ &\quad - \frac{\lambda}{12} \sum_{t=2}^T \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + \frac{6\|\nabla f_1(\mathbf{x}_1)\|^2}{\lambda} \\ &\leq \sum_{t=2}^T \frac{12}{\lambda t} \sup_{\mathbf{x} \in \mathcal{X}} \|\nabla f_t(\mathbf{x}) - \nabla f_{t-1}(\mathbf{x})\|^2 + \sum_{t=1}^{T-1} \left( \frac{12L^2}{\lambda t} - \frac{\lambda}{12} \right) \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \\ &\quad + \frac{48L\delta}{\lambda} \ln T + \frac{6\|\nabla f_1(\mathbf{x}_1)\|^2}{\lambda}. \end{aligned}$$

505 The first two terms can be well controlled by two technical lemmas ([Lemma 15](#), [Lemma 16](#)), hence:

$$\begin{aligned} \text{REG}_T &\leq \frac{12\hat{G}_{\max}^2}{\lambda} \ln \left( 1 + \frac{V_T}{\hat{G}_{\max}^2} \right) + \frac{24\hat{G}_{\max}^2}{\lambda} + \frac{24L^2D^2}{\lambda} \ln \left( 1 + \frac{12L}{\lambda} \right) \\ &\quad + \frac{48L\delta}{\lambda} \ln T + \frac{6\|\nabla f_1(\mathbf{x}_1)\|^2}{\lambda}, \end{aligned}$$

506 where we define  $\hat{G}_{\max}^2 \triangleq \max_{t \in [T-1]} \sup_{\mathbf{x} \in \mathcal{X}} \|\nabla f_t(\mathbf{x}) - \nabla f_{t+1}(\mathbf{x})\|^2$ . By solving the trade-off:

507  $L\delta \ln T = L^2D^2$  with  $L = \delta^{\frac{\nu-1}{1+\nu}} L_\nu^{\frac{2}{1+\nu}}$ , we obtain  $\delta = L_\nu D^{1+\nu} (\ln T)^{-\frac{1+\nu}{2}}$ , and conclude that:

$$\begin{aligned} \text{REG}_T &\leq \frac{12\hat{G}_{\max}^2}{\lambda} \ln \left( 1 + \frac{V_T}{\hat{G}_{\max}^2} \right) + \frac{24\hat{G}_{\max}^2}{\lambda} + \frac{6\|\nabla f_1(\mathbf{x}_1)\|^2}{\lambda} \\ &\quad + \frac{24L_\nu^2D^{2\nu}}{\lambda} (\ln T)^{1-\nu} \left( 2 + \ln \left( 1 + \frac{12L_\nu}{\lambda D^{1-\nu}} (\ln T)^{\frac{1-\nu}{2}} \right) \right) \\ &= \mathcal{O} \left( \frac{\hat{G}_{\max}^2}{\lambda} \log \left( 1 + \frac{V_T}{\hat{G}_{\max}^2} \right) + \frac{L_\nu^2D^{2\nu}}{\lambda} (\log T)^{1-\nu} + \frac{\|\nabla f_1(\mathbf{x}_1)\|^2}{\lambda} \right), \end{aligned}$$

508 where  $\ln(1 + \frac{12L_\nu}{\lambda D^{1-\nu}} (\ln T)^{\frac{1-\nu}{2}}) \leq \ln(1 + \frac{12L_\nu}{\lambda D^{1-\nu}}) + \frac{1-\nu}{2} \ln \ln T = \mathcal{O}(1)$ , because it only consists  
509 of the logarithm of the constant  $\frac{L_\nu}{\lambda D^{1-\nu}}$ , and we treat double logarithmic factors in  $T$  as a constant,  
510 following previous studies [[Luo and Schapire, 2015](#); [Zhao et al., 2024](#)].  $\square$

## 511 B.2 Online-to-batch Conversion for Strongly Convex Functions

512 **Lemma 5** (Online-to-batch Conversion for Strongly Convex Functions). *Let the objective  $\ell(\cdot) : \mathcal{X} \rightarrow$*   
 513  *$\mathbb{R}$  be  $\lambda$ -strongly convex. By employing the online-to-batch conversion algorithm with online function*  
 514  *$f_t(\mathbf{x}) \triangleq \alpha_t \langle \nabla \ell(\bar{\mathbf{x}}_t), \mathbf{x} \rangle + \frac{\lambda \alpha_t}{2} \|\mathbf{x} - \bar{\mathbf{x}}_t\|^2$ , we have, for any  $\mathbf{x}_\star \in \mathcal{X}$ :*

$$\ell(\bar{\mathbf{x}}_T) - \ell(\mathbf{x}_\star) \leq \frac{1}{\alpha_{1:T}} \sum_{t=1}^T (f_t(\mathbf{x}_t) - f_t(\mathbf{x}_\star) - \alpha_{1:t-1} \mathcal{D}_\ell(\bar{\mathbf{x}}_{t-1}, \bar{\mathbf{x}}_t)). \quad (22)$$

515 *Proof.* This lemma is the variant of the stabilized online-to-batch conversion [Cutkosky, 2019] for  
 516 strongly convex functions. We start from the equality:

$$\begin{aligned} \ell(\bar{\mathbf{x}}_T) - \ell(\mathbf{x}_\star) &= \frac{\alpha_1 \ell(\bar{\mathbf{x}}_1)}{\alpha_{1:T}} + \sum_{t=2}^T \frac{\alpha_{1:t} \ell(\bar{\mathbf{x}}_t) - \alpha_{1:t-1} \ell(\bar{\mathbf{x}}_{t-1})}{\alpha_{1:T}} - \ell(\mathbf{x}_\star) \\ &= \frac{1}{\alpha_{1:T}} \sum_{t=1}^T \alpha_t (\ell(\bar{\mathbf{x}}_t) - \ell(\mathbf{x}_\star)) + \frac{1}{\alpha_{1:T}} \sum_{t=2}^T \alpha_{1:t-1} (\ell(\bar{\mathbf{x}}_t) - \ell(\bar{\mathbf{x}}_{t-1})) \\ &\leq \frac{1}{\alpha_{1:T}} \sum_{t=1}^T \alpha_t \left( \langle \nabla \ell(\bar{\mathbf{x}}_t), \bar{\mathbf{x}}_t - \mathbf{x}_\star \rangle - \frac{\lambda}{2} \|\bar{\mathbf{x}}_t - \mathbf{x}_\star\|^2 \right) \\ &\quad + \frac{1}{\alpha_{1:T}} \sum_{t=2}^T \alpha_{1:t-1} (\langle \nabla \ell(\bar{\mathbf{x}}_t), \bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t-1} \rangle - \mathcal{D}_\ell(\bar{\mathbf{x}}_{t-1}, \bar{\mathbf{x}}_t)) \\ &= \frac{1}{\alpha_{1:T}} \sum_{t=1}^T \alpha_t \left( \langle \nabla \ell(\bar{\mathbf{x}}_t), \bar{\mathbf{x}}_t - \mathbf{x}_\star \rangle - \frac{\lambda}{2} \|\bar{\mathbf{x}}_t - \mathbf{x}_\star\|^2 \right) \\ &\quad + \frac{1}{\alpha_{1:T}} \sum_{t=2}^T \alpha_t \langle \nabla \ell(\bar{\mathbf{x}}_t), \bar{\mathbf{x}}_t - \mathbf{x}_t \rangle - \frac{1}{\alpha_{1:T}} \sum_{t=2}^T \alpha_{1:t-1} \mathcal{D}_\ell(\bar{\mathbf{x}}_{t-1}, \bar{\mathbf{x}}_t) \\ &\leq \frac{1}{\alpha_{1:T}} \sum_{t=1}^T \alpha_t \left( \langle \nabla \ell(\bar{\mathbf{x}}_t), \mathbf{x}_t - \mathbf{x}_\star \rangle + \frac{\lambda}{2} \|\mathbf{x}_t - \bar{\mathbf{x}}_t\|^2 - \frac{\lambda}{2} \|\bar{\mathbf{x}}_t - \mathbf{x}_\star\|^2 \right) - \frac{1}{\alpha_{1:T}} \sum_{t=2}^T \alpha_{1:t-1} \mathcal{D}_\ell(\bar{\mathbf{x}}_{t-1}, \bar{\mathbf{x}}_t) \\ &= \frac{1}{\alpha_{1:T}} \sum_{t=1}^T (f_t(\mathbf{x}_t) - f_t(\mathbf{x}_\star) - \alpha_{1:t-1} \mathcal{D}_\ell(\bar{\mathbf{x}}_{t-1}, \bar{\mathbf{x}}_t)), \end{aligned}$$

517 where in the inequality we use the definition of  $\lambda$ -strong convexity and Bregman divergence, after  
 518 which we use the property of  $\alpha_{1:t-1}(\bar{\mathbf{x}}_{t-1} - \bar{\mathbf{x}}_t) = \alpha_t(\bar{\mathbf{x}}_t - \mathbf{x}_t)$  in Theorem 1 of Cutkosky [2019].  
 519 The second inequality is by directly adding the positive term  $\frac{\lambda \alpha_t}{2 \alpha_{1:T}} \|\mathbf{x}_t - \bar{\mathbf{x}}_t\|^2$ .  $\square$

## 520 B.3 Proof of Theorem 4

521 In this subsection, we provide the proof of Theorem 4. To better illustrate the ideas, we divide it  
 522 into four parts. Appendix B.3.1 provides an equivalent definition of weights  $\alpha_t$ , i.e., in the terms  
 523 of  $\alpha_t \triangleq \beta_t \alpha_{1:t-1}$ , and an upper bound for any-time convergence rate in Lemma 6. Appendix B.3.2  
 524 provides some crucial properties for Lemma 6. Then in Appendix B.3.3 and Appendix B.3.4 we  
 525 apply Lemma 6 to prove the smooth and non-smooth cases in Theorem 4, respectively.

### 526 B.3.1 Part 1: Basic Lemma

527 In the following, we prove a general upper bound for the convergence error of any iteration, which  
 528 demonstrates that convergence error comes from the quantities within the bad event

529 **Lemma 6.** *For the settings in Theorem 4, the weights  $\alpha_t$  have an equivalent definition:  $\alpha_t \triangleq \beta_t \alpha_{1:t-1}$*   
 530 *for all  $t \geq 2$ , where  $\beta_1 \triangleq 1$ ,  $\beta_2 \triangleq \frac{1}{2}$ ,  $\beta \triangleq \exp(\frac{1}{T} \ln T) - 1 \in [\frac{1}{T} \ln T, \frac{2}{T} \ln T]$ , and for  $t \geq 2$ :*

$$\beta_{t+1} \triangleq \begin{cases} \beta_t, & \beta_t^2 \leq \frac{\lambda}{4L_t}, \\ \max \left\{ \bar{\beta}, \frac{\beta_t}{2} \right\}, & \beta_t^2 > \frac{\lambda}{4L_t}. \end{cases} \quad (23)$$

For all  $\tau \in [T]$ , define  $\mathcal{S}_\tau \triangleq \{t : 2 \leq t \leq \tau, \beta_t^2 > \frac{\lambda}{4L_t}\} \cup \{1\}$  (which includes the iterations when the estimate  $\hat{L}_t$  is found unsuitable and needed to be updated), then algorithm in Theorem 4 ensures:

$$\ell(\bar{\mathbf{x}}_\tau) - \ell(\mathbf{x}_\star) \leq \sum_{t \in \mathcal{S}_\tau} \frac{2\beta_t^2 \hat{G}_{\max}^2}{\lambda \prod_{s=t}^\tau (1 + \beta_s)}, \quad (24)$$

where  $\hat{G}_{\max}^2 \triangleq \max\{\|\nabla \ell(\mathbf{x}_1)\|^2, \max_{t \in \mathcal{S}_\tau} \|\nabla \ell(\bar{\mathbf{x}}_t) - \nabla \ell(\bar{\mathbf{x}}_{t-1})\|^2\}$ .

*Proof.* The equivalent definition of  $\alpha_t$  regrading  $\beta_t$  is proved by substituting the definition of  $\alpha_t$ .

With  $f_t(\mathbf{x}) \triangleq \alpha_t \langle \nabla \ell(\bar{\mathbf{x}}_t), \mathbf{x} \rangle + \frac{\lambda \alpha_t}{2} \|\mathbf{x} - \bar{\mathbf{x}}_t\|^2$ , by Lemma 5 we have:

$$\ell(\bar{\mathbf{x}}_T) - \ell(\mathbf{x}_\star) \leq \frac{1}{\alpha_{1:T}} \sum_{t=1}^T (f_t(\mathbf{x}_t) - f_t(\mathbf{x}_\star) - \alpha_{1:t-1} \mathcal{D}_\ell(\bar{\mathbf{x}}_{t-1}, \bar{\mathbf{x}}_t)).$$

By Lemma 13, with the definitions  $\eta_t = \frac{1}{\lambda \alpha_{1:t}}$ ,  $M_t = \alpha_t \nabla \ell(\bar{\mathbf{x}}_{t-1}) + \lambda \alpha_t (\mathbf{x}_{t-1} - \tilde{\mathbf{x}}_t)$ ,  $\tilde{\mathbf{x}}_t = \frac{1}{\alpha_{1:t}} (\sum_{s=1}^{t-1} \alpha_s \mathbf{x}_s + \alpha_t \mathbf{x}_{t-1})$ , we arrive at

$$\begin{aligned} & \sum_{t=1}^T (f_t(\mathbf{x}_t) - f_t(\mathbf{x}_\star)) - \sum_{t=1}^T \alpha_{1:t-1} \mathcal{D}_\ell(\bar{\mathbf{x}}_{t-1}, \bar{\mathbf{x}}_t) \\ & \leq \sum_{t=1}^T \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_\star \rangle - \frac{\lambda}{2} \sum_{t=1}^T \alpha_t \|\mathbf{x}_t - \mathbf{x}_\star\|^2 - \sum_{t=1}^T \alpha_{1:t-1} \mathcal{D}_\ell(\bar{\mathbf{x}}_{t-1}, \bar{\mathbf{x}}_t) \\ & \leq \sum_{t=1}^T \eta_t \|\nabla f_t(\mathbf{x}_t) - M_t\|^2 - \sum_{t=1}^T \frac{1}{4\eta_t} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \\ & \quad - \sum_{t=1}^T \alpha_{1:t-1} \mathcal{D}_\ell(\bar{\mathbf{x}}_{t-1}, \bar{\mathbf{x}}_t) + \sum_{t=2}^T \left( \frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} - \frac{\lambda \alpha_t}{2} \right) \|\mathbf{x}_t - \mathbf{x}_\star\|^2 \\ & = \sum_{t=2}^T \frac{\alpha_t^2}{\lambda \alpha_{1:t}} \|\nabla \ell(\bar{\mathbf{x}}_t) - \nabla \ell(\bar{\mathbf{x}}_{t-1}) + \lambda (\mathbf{x}_t - \mathbf{x}_{t-1} - \bar{\mathbf{x}}_t + \tilde{\mathbf{x}}_t)\|^2 - \sum_{t=2}^T \frac{1}{4\eta_{t-1}} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 \\ & \quad - \sum_{t=1}^T \alpha_{1:t-1} \mathcal{D}_\ell(\bar{\mathbf{x}}_{t-1}, \bar{\mathbf{x}}_t) + \frac{\alpha_1}{\lambda} \|\nabla \ell(\mathbf{x}_1)\|^2 \quad (\text{by setting } \eta_t = \frac{1}{\lambda \alpha_{1:t}}) \\ & \leq \sum_{t=2}^T \frac{2\alpha_t^2}{\lambda \alpha_{1:t}} \|\nabla \ell(\bar{\mathbf{x}}_t) - \nabla \ell(\bar{\mathbf{x}}_{t-1})\|^2 + \sum_{t=2}^T \frac{2\alpha_t^2 \lambda}{\alpha_{1:t}} \left\| \left( 1 - \frac{\alpha_t}{\alpha_{1:t}} \right) (\mathbf{x}_t - \mathbf{x}_{t-1}) \right\|^2 \\ & \quad - \sum_{t=2}^T \frac{\lambda \alpha_{1:t-1}}{4} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 - \sum_{t=1}^T \alpha_{1:t-1} \mathcal{D}_\ell(\bar{\mathbf{x}}_{t-1}, \bar{\mathbf{x}}_t) + \frac{\alpha_1}{\lambda} \|\nabla \ell(\mathbf{x}_1)\|^2 \\ & \leq \underbrace{\sum_{t \in \mathcal{S}_T} \frac{2\alpha_t^2 \hat{G}_{\max}^2}{\lambda \alpha_{1:t}}}_{\text{TERM-A}} + \underbrace{\sum_{t \notin \mathcal{S}_T} \left( \frac{4\alpha_t^2 L_t}{\lambda \alpha_{1:t}} - \alpha_{1:t-1} \right) \mathcal{D}_\ell(\bar{\mathbf{x}}_{t-1}, \bar{\mathbf{x}}_t)}_{\text{TERM-B}} \\ & \quad + \underbrace{\sum_{t=2}^T \left( \frac{2\alpha_t^2 \alpha_{1:t-1}^2 \lambda}{\alpha_{1:t}^3} - \frac{\lambda \alpha_{1:t-1}}{4} \right) \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2}_{\text{TERM-C}}, \end{aligned}$$

where we use definition  $L_t \triangleq \frac{\|\nabla \ell(\bar{\mathbf{x}}_t) - \nabla \ell(\bar{\mathbf{x}}_{t-1})\|^2}{2\mathcal{D}_\ell(\bar{\mathbf{x}}_{t-1}, \bar{\mathbf{x}}_t)}$ . In the above, TERM-A includes those terms that cannot be sufficiently canceled out like TERM-B, whose indices correspond exactly to the iterations  $\mathcal{S}_T$  where the estimate  $\hat{L}_t$  needs updating.

For the TERM-B where  $t \notin \mathcal{S}_T$  and  $\beta_t^2 \leq \frac{\lambda}{4L_t}$ , by  $\alpha_t = \alpha_1 \prod_{s=2}^t (1 + \beta_s)$  for all  $t \geq 2$ , we have

$$\frac{\alpha_t^2}{\alpha_{1:t-1} \alpha_{1:t}} = \frac{\beta_t^2}{1 + \beta_t} \leq \beta_t^2 \leq \frac{\lambda}{4L_t}, \implies \frac{4\alpha_t^2 L_t}{\lambda \alpha_{1:t}} \leq \alpha_{1:t-1}, \implies \text{TERM-B} \leq 0.$$



542 For the TERM-C, since  $0 < \beta_t \leq 1$ , we arrive at

$$\frac{\beta_t^2}{(1 + \beta_t)^3} \leq \frac{1}{8} \implies \frac{2\alpha_t^2 \alpha_{1:t-1}^2 \lambda}{\alpha_{1:t}^3} \leq \frac{\lambda \alpha_{1:t-1}}{4} \implies \text{TERM-C} \leq 0.$$

543 Now notice that the above proof holds for all  $T \in \mathbb{N}_+$ , therefore for all  $\tau \in [T]$  we have

$$\ell(\bar{\mathbf{x}}_\tau) - \ell(\mathbf{x}_*) \leq \frac{1}{\alpha_{1:\tau}} \sum_{t \in \mathcal{S}_\tau} \frac{2\alpha_t^2 \hat{G}_{\max}^2}{\lambda \alpha_{1:t}} \leq \sum_{t \in \mathcal{S}_\tau} \frac{2\beta_t^2 \hat{G}_{\max}^2}{\lambda \prod_{s=t}^\tau (1 + \beta_s)}.$$

544 The proof is finished.  $\square$

### 545 B.3.2 Part 2: Some Crucial Properties

546 Before continuing the proof of [Theorem 4](#), let's investigate the bound in [Lemma 6](#). For any  $\tau \in [T]$ ,  
 547 denoting by  $m_\tau \triangleq |\mathcal{S}_\tau| - 1$  and write  $\mathcal{S}_\tau = \{t_0, t_1, t_2, \dots, t_{m_\tau}\}$  the index set where  $1 = t_0 < t_1 <$   
 548  $t_2 < \dots < t_{m_\tau} \leq \tau$ . Now we investigate what the sequence  $\{\beta_t\}_{t=1}^\tau$  looks like. First of all for all  
 549  $t \in \mathcal{S}_T$ , we have  $\beta_t \geq \bar{\beta}$  and  $\beta_t > \sqrt{\lambda/(4L_\ell)} \geq \sqrt{\lambda/(4L_\ell)} = 1/(2\sqrt{\kappa})$ , denoting by  $\kappa = L_\ell/\lambda$ .  
 550 Now we consider the following two cases:

551 **Case 1:**  $\sqrt{\lambda/(4L_\ell)} \geq \bar{\beta}$ . For all  $t \in \mathcal{S}_T$ ,  $\beta_t > \sqrt{\lambda/(4L_\ell)} \geq \bar{\beta}$ ; then for all  $i \in [m_T]$  and  
 552  $t \in (t_{i-1}, t_i]$ ,  $\beta_t = 2^{-i}$ , which implies  $m_T \leq \lceil \log_2 \sqrt{\kappa} \rceil - 1$ .

553 **Case 2:**  $\sqrt{\lambda/(4L_\ell)} < \bar{\beta}$ . Denoting  $\bar{\tau}$  the minimum index in  $[T]$  that  $\beta_{\bar{\tau}} = \bar{\beta}$  if exists, otherwise let  
 554  $\bar{\tau} = T + 1$ . Then for all  $t_i < \bar{\tau}$  and  $t \in (t_{i-1}, t_i]$ ,  $\beta_t = 2^{-i}$ ; and for all  $t \geq \bar{\tau}$ ,  $\beta_t = \bar{\beta}$ .

### 555 B.3.3 Part 3: $L_\ell$ -smooth Case

556 In this part we prove the exponential rate with  $L_\ell$ -smoothness in [Theorem 4](#).

557 *Proof.* We consider the two cases separately.

558 **Case 1:**  $\sqrt{\lambda/(4L_\ell)} \geq \bar{\beta}$ . As discussed before, we know for all  $i \in [m_T]$  and  $t \in (t_{i-1}, t_i]$ ,  
 559  $\beta_t = 2^{-i}$ , and  $m_T \leq \lceil \log_2 \sqrt{\kappa} \rceil - 1$ . Now by [Lemma 6](#),

$$\begin{aligned} \ell(\bar{\mathbf{x}}_\tau) - \ell(\mathbf{x}_*) &\stackrel{(24)}{\leq} \sum_{t \in \mathcal{S}_\tau} \frac{2\beta_t^2 \hat{G}_{\max}^2}{\lambda \prod_{s=t}^\tau (1 + \beta_s)} = \sum_{i=0}^{m_\tau} \frac{2^{1-2i} \hat{G}_{\max}^2}{\lambda \prod_{s=t_i}^\tau (1 + \beta_s)} \\ &\leq \sum_{i=0}^{m_\tau} \frac{2^{1-2i} \hat{G}_{\max}^2}{\lambda (1 + 1/(2\sqrt{\kappa}))^{\tau-t_i+1}} \leq \frac{4\hat{G}_{\max}^2}{\lambda (1 + 1/(2\sqrt{\kappa}))^{\tau-t_{m_\tau}+1}}, \end{aligned}$$

560 where the second inequality we use  $\beta_t > 1/(2\sqrt{\kappa})$ , and the third inequality we use  $t_i \leq t_{m_\tau}$  for all  
 561  $i \leq m_\tau$  and  $\sum_{i=0}^\infty 2^{-2i} < 2$ . Now notice that since  $1 = t_0 < t_1 < t_2 < \dots < t_{m_T} \leq T$  and  $m_T \leq$   
 562  $\lceil \log_2 \sqrt{\kappa} \rceil - 1$ , then either there exists  $j \in [m_T]$  that  $t_j - t_{j-1} \geq \frac{T}{m_T+1}$  or  $T - t_{m_T} + 1 \geq \frac{T}{m_T+1}$ ,  
 563 then  $\max_{\tau \in [T]} (\tau - t_{m_\tau} + 1) \geq \frac{T}{m_T+1} \geq \frac{T}{\lceil \log_2 \sqrt{\kappa} \rceil}$ ,

$$\min_{\tau \in [T]} \ell(\bar{\mathbf{x}}_\tau) - \ell(\mathbf{x}_*) \leq \frac{4\hat{G}_{\max}^2}{\lambda} \left(1 + \frac{1}{2\sqrt{\kappa}}\right)^{\frac{-T}{\lceil \log_2 \sqrt{\kappa} \rceil}} \leq \frac{4\hat{G}_{\max}^2}{\lambda} \exp\left(\frac{-T}{(1 + 2\sqrt{\kappa}) \lceil \log_2 \sqrt{\kappa} \rceil}\right),$$

564 where we use  $(1 + x^{-1})^{-T} = (1 - 1/(1 + x))^T \leq \exp(-T/(1 + x))$  for all  $x > 0$ .

565 **Case 2:**  $\sqrt{\lambda/(4L_\ell)} < \bar{\beta}$ . Once  $T > 10$ , we have  $\bar{\beta} = \exp(\frac{1}{T} \ln T) - 1 < \frac{1}{4}$ . Combining with  
 566  $\sqrt{\lambda/(4L_\ell)} < \bar{\beta}$  yields  $\sqrt{L_\ell/\lambda} = \sqrt{\kappa} > 2$ , then we have, the exponential rate:

$$\left(1 + \frac{1}{2\sqrt{\kappa}}\right)^{\frac{-T}{\lceil \log_2 \sqrt{\kappa} \rceil}} \geq (1 + \bar{\beta})^{\frac{-T}{\lceil \log_2 \sqrt{\kappa} \rceil}} = T^{\frac{-T}{T \lceil \log_2 \sqrt{\kappa} \rceil}} = T^{\frac{-1}{\lceil \log_2 \sqrt{\kappa} \rceil}} \geq \frac{1}{\sqrt{T}} \geq \Omega\left(\frac{\log T}{T}\right).$$

567 Therefore, in this case the exponential rate is dominated by the rate of the non-smooth case.  $\square$

### 568 B.3.4 Part 4: $G$ -Lipschitz Case

569 In this part we prove the rate with  $G$ -Lipschitzness in [Theorem 4](#).

570 *Proof.* By [Lemma 6](#), we have

$$\ell(\bar{\mathbf{x}}_\tau) - \ell(\mathbf{x}_*) \leq \sum_{t \in \mathcal{S}_\tau} \frac{2\beta_t^2 \hat{G}_{\max}^2}{\lambda \prod_{s=t}^\tau (1 + \beta_s)}.$$

571 As discussed in [Appendix B.3.2](#), if  $\beta_T = \bar{\beta}$ , denoting by  $\bar{\tau}$  the minimum index in  $[T]$  that  $\beta_{\bar{\tau}} = \bar{\beta}$ ,  
 572 otherwise let  $\bar{\tau} = T + 1$ . Now we have, for all  $t_i < \bar{\tau}$  and  $t \in (t_{i-1}, t_i]$ ,  $\beta_t = 2^{-i}$ ; and for all  $t \geq \bar{\tau}$ ,  
 573  $\beta_t = \bar{\beta}$ . Then the above summation terms, when  $\tau \geq \bar{\tau}$ , can be bounded as:

$$\begin{aligned} \ell(\bar{\mathbf{x}}_\tau) - \ell(\mathbf{x}_*) &\leq \sum_{t \in \mathcal{S}_\tau, t < \bar{\tau}} \frac{2\beta_t^2 \hat{G}_{\max}^2}{\lambda \prod_{s=t}^\tau (1 + \beta_s)} + \sum_{t \in \mathcal{S}_\tau, \bar{\tau} \leq t \leq \tau} \frac{2\beta_t^2 \hat{G}_{\max}^2}{\lambda \prod_{s=t}^\tau (1 + \beta_s)} \\ &\leq \sum_{t \in \mathcal{S}_\tau, t < \bar{\tau}} \frac{2\beta_t^2 \hat{G}_{\max}^2}{\lambda \prod_{s=t}^\tau (1 + \beta_s)} + \sum_{\bar{\tau} \leq t \leq \tau} \frac{2\bar{\beta}^2 \hat{G}_{\max}^2}{\lambda \prod_{s=t}^\tau (1 + \bar{\beta})} \leq \sum_{t \in \mathcal{S}_\tau, t < \bar{\tau}} \frac{2\beta_t^2 \hat{G}_{\max}^2}{\lambda \prod_{s=t}^\tau (1 + \beta_s)} + \frac{2\hat{G}_{\max}^2 \bar{\beta}}{\lambda}, \end{aligned}$$

574 where the last inequality uses the summation of a geometric series. Recall that  $\bar{\beta} \leq \frac{2 \ln T}{T} = \mathcal{O}(\frac{\log T}{T})$ .

575 In the next we will prove that, either there exists an iteration  $\tau < \bar{\tau}$  with the sub-optimality gap  
 576 bounded by the desired  $\mathcal{O}(\frac{\log T}{T})$ , where we don't need to consider those terms with  $t \geq \bar{\tau}$ ; or the  
 577 last-iterate convergence rate is bounded by  $\mathcal{O}(\frac{\log T}{T})$ , where we have already proven those terms with  
 578  $t \geq \bar{\tau}$ . Now we consider  $\tau < \bar{\tau}$ , which implies for each  $t_i \in \mathcal{S}_\tau$ ,  $\beta_{t_i} = 2^{-i}$ , then  $|\mathcal{S}_\tau| \leq \mathcal{O}(\log T)$ .

579 In the following, we prove by contradiction. Assume the proposition  $\mathcal{P}$ : neither the best-iterate  
 580 convergence rate of  $\mathcal{O}(\frac{\log T}{T})$  holds for some  $\tau < \bar{\tau}$ , nor the last-iterate convergence rate is  $\mathcal{O}(\frac{\log T}{T})$ .  
 581 Then  $\mathcal{P}$  implies the proposition  $\mathcal{Q}$ : for  $\mathcal{S}_\tau = \{t_0, t_1, \dots, t_{m_\tau}\}$ , there exists index  $0 \leq j \leq m_\tau$   
 582 satisfying the following two conditions simultaneously:

583 **Condition 1.** There should *not* exist  $\tau' < t_j$  that: for all  $t \in \mathcal{S}_{\tau'}$ ,  $\beta_t^2 / \prod_{s=t}^{\tau'} (1 + \beta_s) < 2/T$  always  
 584 holds. Otherwise the sub-optimality gap at iteration  $\tau'$  is already bounded by  $\mathcal{O}(\frac{\log T}{T})$ ;

585 **Condition 2.** At the last iteration  $T$ , we should have, for this index  $j$ :

$$\frac{\beta_{t_j}^2}{(1 + \bar{\beta})^{T-t_j}} = \frac{2^{-2j}}{T^{\frac{T-t_j}{T}}} > \frac{2}{T}, \implies t_j > \frac{(1 + 2j)T}{\log_2 T}. \quad (25)$$

586 Otherwise we obtain the last-iterate convergence rate of  $\mathcal{O}(\frac{\log T}{T})$  by the following inequality:

$$\sum_{t \in \mathcal{S}_T, t < \bar{\tau}} \frac{\beta_t^2}{\prod_{s=t}^T (1 + \beta_s)} \leq \sum_{t \in \mathcal{S}_T, t < \bar{\tau}} \frac{\beta_t^2}{(1 + \bar{\beta})^{T-t}} \leq \frac{2m_{\bar{\tau}}}{T} \leq \mathcal{O}\left(\frac{\log T}{T}\right).$$

587 In the following, we will clarify that the sequence  $\{t_0, t_1, \dots, t_{m_\tau}\}$  cannot possibly satisfy the  
 588 proposition  $\mathcal{Q}$ . This will conflict with the proposition  $\mathcal{P}$ , thereby finishing our proof.

589 The idea of proving proposition  $\mathcal{Q}$  cannot be satisfied is by doing the “best” construction of  
 590  $\{t_0, t_1, \dots, t_{m_\tau}\}$ , and show that even the best one still fails. To this end, now we try to construct  
 591  $\{t_0, t_1, \dots, t_{m_\tau}\}$  in order to investigate whether the proposition  $\mathcal{Q}$  is achievable.

592 To first get an intuition of the “best” construction, let's fix some  $j$  and assume it satisfies the above two  
 593 conditions, which motivates us the second condition in (25) requires  $t_j$  should be as large as possible  
 594 while do not conflict with the first condition. This gives us an construction strategy based on *greedy*:  
 595 during our construction by order, we only import the new time stamp  $t_j$  when the first condition  
 596 is about to fail, otherwise we can increase  $t_j$  which will not worsen our construction. Then for all  
 597  $\tau' \in [t_j, t_{j+1})$ , there is *only* index  $j$  ensuring the first condition, i.e.,  $\beta_{t_j}^2 / \prod_{s=t_j}^{\tau'} (1 + \beta_s) \geq 2/T$ .  
 598 Based on this strategy, we perform the construction.

599 **Case of  $j = 0$ .** We have  $t_0 = 1$  by definition, and only need to investigate the second condition  
 600 in (25), that is,  $1 \geq T/\log_2 T$ , which is not true.

601 **Case of  $j = 1$ .** By the strategy,  $t_1$  should be as large as possible while do not conflict with the first  
 602 condition, which gives the only constraint by applying the first condition at  $\tau = t_1 - 1$ :

$$\frac{\beta_{t_0}^2}{\prod_{s=t_0}^{t_1-1}(1+\beta_s)} = \frac{1}{(1+2^{-1})^{t_1-t_0}} \geq \frac{2}{T}, \implies t_1 - t_0 \leq \log_{1.5} \frac{T}{2},$$

603 by which, the second condition in (25), that is,  $t_1 > 3T/\log_2 T$ , is not true.

604 **Case of  $j \geq 1$ .** Now let's consider by induction. For index  $j \geq 1$ , assume that as best as we could  
 605 by greedy, which means only index  $j$  ensures that the first condition is satisfied, by which substituting  
 606  $\tau = t_{j+1} - 1$  into the first condition, we have

$$\frac{\beta_{t_j}^2}{\prod_{s=t_j}^{t_{j+1}-1}(1+\beta_s)} = \frac{2^{-2j}}{(1+2^{-j-1})^{t_{j+1}-t_j}} \geq \frac{2}{T}, \implies t_{j+1} - t_j \leq \frac{\ln \frac{T}{2^{1+2j}}}{\ln(1+2^{-j-1})}. \quad (26)$$

607 Moreover, by induction assume that the second condition in (25) fails for  $j$ :

$$t_j \leq \frac{(1+2j)T}{\log_2 T}. \quad (27)$$

608 Then we show that index  $j+1$  also cannot satisfy the second condition by combining (26) and (27):

$$t_{j+1} \leq \frac{(1+2j)T}{\log_2 T} + \frac{\ln \frac{T}{2^{1+2j}}}{\ln(1+2^{-j-1})} \leq \frac{(1+2(j+1))T}{\log_2 T},$$

609 where the proof of the second inequality will be given later.

610

611 Therefore, by induction, we prove that the proposition  $\mathcal{Q}$  is impossible. Hence we prove the existence  
 612 of the best-iterate convergence rate of  $\mathcal{O}(\frac{\log T}{T})$ .

613 Finally, we prove the following inequality always holds for all  $j \geq 1$ :

$$\frac{(1+2j)T}{\log_2 T} + \frac{\ln \frac{T}{2^{1+2j}}}{\ln(1+2^{-j-1})} \leq \frac{(1+2(j+1))T}{\log_2 T}, \iff \frac{\ln \frac{T}{2^{1+2j}}}{\ln(1+2^{-j-1})} \leq \frac{2T}{\log_2 T}.$$

614 Defining  $x = 2^{-j}$ , then we are going to prove:

$$\frac{\ln(Tx^2/2)}{\ln(1+x/2)} \leq \frac{2T}{\log_2 T}, \iff h(x) \triangleq (\log_2 T) \ln(Tx^2/2) - 2T \ln(1+x/2) \leq 0.$$

615 Now because  $\max_{x \in (0,1]} h(x) < 0$ , finally we finish the proof.  $\square$

#### 616 B.4 Algorithm 1 with a Given Smoothness Parameter

617 We provide the following corollary when Algorithm 1 is given the smoothness parameter  $L_\ell$ .

618 **Corollary 1.** For the optimization problem of  $\min_{\mathbf{x} \in \mathcal{X}} \ell(\mathbf{x})$  with deterministic setting, assume the  
 619 objective  $\ell(\cdot)$  is  $\lambda$ -strongly convex and  $L_\ell$ -smooth, denoting by  $\kappa \triangleq L_\ell/\lambda$ . Algorithm 1 by setting  
 620  $\alpha_t = \alpha_{1:t-1} \cdot \sqrt{\lambda/(4L_\ell)}$  ensures:

$$\ell(\bar{\mathbf{x}}_T) - \ell(\mathbf{x}_*) \leq \mathcal{O} \left( \frac{\|\nabla \ell(\mathbf{x}_1)\|^2}{\lambda} \exp \left( \frac{-T}{1+2\sqrt{\kappa}} \right) \right).$$

621 *Proof.* By Lemma 6, since  $\beta_t \equiv \sqrt{\lambda/(4L_\ell)} \leq \sqrt{\lambda/(4L_t)}$  for all  $t \geq 2$ , we have  $S_T = \{1\}$ , then

$$\ell(\bar{\mathbf{x}}_T) - \ell(\mathbf{x}_*) \stackrel{(24)}{\leq} \frac{2\|\nabla \ell(\mathbf{x}_1)\|^2}{\lambda(1+\sqrt{\lambda/(4L_\ell)})^{T-1}} \leq \mathcal{O} \left( \frac{\|\nabla \ell(\mathbf{x}_1)\|^2}{\lambda} \exp \left( \frac{-T}{1+2\sqrt{\kappa}} \right) \right),$$

622 where we use  $(1+x^{-1})^{-T} = (1-1/(1+x))^T \leq \exp(-T/(1+x))$  for all  $x > 0$ .  $\square$

### 623 C Omitted Details for Section 5

624 In this section, we give the proofs of theorems in Section 5.

---

**Algorithm 2** Dynamic Regret Minimization under Hölder Smoothness

---

**Input:** Domain diameter  $D$ , total iterations  $T$

- 1: **Initialization:**  $N = \lceil \log_2(1 + T) \rceil$ , starting points  $\{\mathbf{x}_{1,i}\}_{i=1}^N, \{\widehat{\mathbf{x}}_{1,i}\}_{i=1}^N, p_{1,i} = \frac{1}{N}$
  - 2: **for**  $t = 1$  **to**  $T$  **do**
  - 3:   Submit  $\mathbf{x}_t = \sum_{i=1}^N p_{t,i} \mathbf{x}_{t,i}$ , receive  $\nabla f_t(\mathbf{x}_t)$
  - 4:   The  $i$ -th base-learner updates to  $\mathbf{x}_{t+1,i}$  by Eq. (12), for all  $i \in [N]$  ▷ Base update
  - 5:   Meta-algorithm updates to  $\mathbf{p}_{t+1}$  by Eq. (13) ▷ Meta update
  - 6: **end for**
- 

### 625 C.1 Proof of Theorem 5

626 To begin with, we summarize Algorithm 2 for minimizing dynamic regret under Hölder smoothness  
627 as specified in Theorem 5. Before proving Theorem 5, we provide the theoretical guarantee for the  
628 meta-algorithm, i.e., Optimistic Hedge [Syrkanis et al., 2015].

629 **Lemma 7** (Theorem 7.35 of Orabona [2019]). *The regret of Optimistic Hedge specified at Eq. (13)*  
630 *with a time-varying learning rate  $\varepsilon_t > 0$  to any expert  $i \in [N]$  ensures:*

$$\sum_{t=1}^T \langle \mathbf{p}_t, \ell_t \rangle - \sum_{t=1}^T \ell_{t,i} \leq \frac{\ln N}{\varepsilon_T} + \sum_{t=1}^T \langle \ell_t - \mathbf{m}_t, \mathbf{p}_t - \mathbf{p}_{t+1} \rangle - \sum_{t=1}^T \frac{1}{2\varepsilon_{t-1}} \|\mathbf{p}_t - \mathbf{p}_{t+1}\|_1^2.$$

631 In the following, we provide the proof for Theorem 5.

632 *Proof of Theorem 5.* We decompose the dynamic regret into the meta-regret and base-regret as

$$\begin{aligned} \text{D-REG}_T(\mathbf{u}_1, \dots, \mathbf{u}_T) &= \sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{u}_t) = \sum_{t=1}^T \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u}_t \rangle - \sum_{t=1}^T \mathcal{D}_{f_t}(\mathbf{u}_t, \mathbf{x}_t) \\ &= \underbrace{\sum_{t=1}^T \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t,i_*} \rangle}_{\text{META-REG}} + \underbrace{\sum_{t=1}^T \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_{t,i_*} - \mathbf{u}_t \rangle}_{\text{BASE-REG}} - \sum_{t=1}^T \mathcal{D}_{f_t}(\mathbf{u}_t, \mathbf{x}_t). \end{aligned} \quad (28)$$

633 Notably, the above meta-base regret decomposition holds for any base-learner's index  $i_* \in [N]$ . We  
634 first investigate the meta-regret:

$$\begin{aligned} \text{META-REG} &= \sum_{t=1}^T \sum_{i=1}^N p_{t,i} \cdot \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_{t,i} \rangle - \sum_{t=1}^T \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_{t,i_*} \rangle \\ &= \sum_{t=1}^T \langle \mathbf{p}_t, \ell_t \rangle - \sum_{t=1}^T \ell_{t,i_*} - \sum_{t=1}^T \sum_{i=1}^N \lambda_{t,i} p_{t,i} c_{t,i} + \sum_{t=1}^T \lambda_{t,i_*} c_{t,i_*}, \end{aligned}$$

635 where we denote by  $c_{t,i} = \|\mathbf{x}_{t,i} - \mathbf{x}_{t-1,i}\|^2$ . By applying Lemma 7 with  $\varepsilon_t = \sqrt{\frac{3+\ln N}{D^2 A_t}}$ , we obtain:

$$\begin{aligned} &\sum_{t=1}^T \langle \mathbf{p}_t, \ell_t \rangle - \sum_{t=1}^T \ell_{t,i_*} \\ &\leq \max_{\mathbf{p} \in \Delta_N} \psi_{T+1}(\mathbf{p}) + \sum_{t=1}^T \langle \ell_t - \mathbf{m}_t, \mathbf{p}_t - \mathbf{p}_{t+1} \rangle - \sum_{t=1}^T \frac{1}{2\varepsilon_{t-1}} \|\mathbf{p}_t - \mathbf{p}_{t+1}\|_1^2 \\ &\leq \frac{3 + \ln N}{\varepsilon_T} + \sum_{t=1}^T \frac{\varepsilon_t}{2} \|\ell_t - \mathbf{m}_t\|_\infty^2 + \sum_{t=1}^T \frac{1}{2\varepsilon_t} \|\mathbf{p}_t - \mathbf{p}_{t+1}\|_1^2 - \sum_{t=1}^T \frac{1}{2\varepsilon_t} \|\mathbf{p}_t - \mathbf{p}_{t+1}\|_1^2 \\ &\leq \frac{3 + \ln N}{\varepsilon_T} + \frac{D^2}{2} \sum_{t=1}^T \varepsilon_t \|\nabla f_t(\mathbf{x}_t) - \mathbf{M}_t\|_2^2 \leq 2D\sqrt{(3 + \ln N)A_T}. \end{aligned}$$

Notice that  $\ln N = \mathcal{O}(\log \log T)$  has double logarithmic factors in  $T$ , which is treated as a constant following previous studies [Luo and Schapire, 2015; Zhao et al., 2024]. Now the meta-regret:

$$\text{META-REG} \leq 2D\sqrt{(3 + \ln N)A_T} - \sum_{t=1}^T \sum_{i=1}^N \lambda_{t,i} p_{t,i} c_{t,i} + \sum_{t=1}^T \lambda_{t,i_*} c_{t,i_*}. \quad (29)$$

For base-regret, denoting by  $d_i = 2^{i-1}D$ , then we set  $\eta_{t,i} = d_i/\sqrt{A_{t-1}}$ . By applying Lemma 12 and choosing  $i_* \in [N]$  such that  $d_{i_*} \leq \sqrt{D^2 + DP_T} \leq 2d_{i_*}$ , we have:

$$\begin{aligned} \text{BASE-REG} &\leq \sum_{t=1}^T \eta_{t+1,i_*} \|\nabla f_t(\mathbf{x}_t) - M_t\|^2 + \frac{D^2 + DP_T}{\eta_{T+1,i_*}} - \sum_{t=2}^T \frac{1}{8\eta_{t+1,i_*}} \|\mathbf{x}_{t,i_*} - \mathbf{x}_{t-1,i_*}\|^2 \\ &\leq 4\sqrt{(D^2 + DP_T)A_T} - \sum_{t=2}^T \frac{1}{8\eta_{t,i_*}} c_{t,i_*}, \end{aligned} \quad (30)$$

where we apply Lemma 14 in the second inequality. Now defining  $\lambda_{t,i} = \frac{1}{8\eta_{t,i}}$  and  $\alpha = \sqrt{56D^2 + 8(\ln N)D^2 + 32DP_T}$ , substituting (29) and (30) into (28), and dropping negative terms,

$$\text{D-REG}_T(\mathbf{u}_1, \dots, \mathbf{u}_T) \leq \alpha\sqrt{A_T} - \sum_{t=1}^T \mathcal{D}_{f_t}(\mathbf{u}_t, \mathbf{x}_t). \quad (31)$$

We decompose  $A_T$  as:

$$\begin{aligned} A_T &= \|\nabla f_1(\mathbf{x}_1)\|^2 + \sum_{t=2}^T \left( \|\nabla f_t(\mathbf{x}_t) - \nabla f_t(\mathbf{u}_t) + \nabla f_t(\mathbf{u}_t) - \nabla f_t(\mathbf{u}_{t-1}) \right. \\ &\quad \left. + \nabla f_t(\mathbf{u}_{t-1}) - \nabla f_{t-1}(\mathbf{u}_{t-1}) + \nabla f_{t-1}(\mathbf{u}_{t-1}) - \nabla f_{t-1}(\mathbf{x}_{t-1})\|^2 \right) \\ &\leq \|\nabla f_1(\mathbf{x}_1)\|^2 + 16L \sum_{t=1}^T \mathcal{D}_{f_t}(\mathbf{u}_t, \mathbf{x}_t) + 4DL^2P_T + 4V_T + 32L\delta T, \end{aligned} \quad (32)$$

where we apply  $\|\sum_{i=1}^n \mathbf{a}_i\|^2 \leq n \sum_{i=1}^n \|\mathbf{a}_i\|^2$ , Lemma 1, and Lemma 4. Substituting (32) into (31):

$$\begin{aligned} \text{D-REG}_T(\mathbf{u}_1, \dots, \mathbf{u}_T) &\stackrel{(31)}{\leq} \alpha\sqrt{A_T} - \sum_{t=1}^T \mathcal{D}_{f_t}(\mathbf{u}_t, \mathbf{x}_t) \\ &\stackrel{(32)}{\leq} \alpha\sqrt{\|\nabla f_1(\mathbf{x}_1)\|^2 + 4DL^2P_T + 4V_T + 32L\delta T} + 2\sqrt{4\alpha^2L \sum_{t=1}^T \mathcal{D}_{f_t}(\mathbf{u}_t, \mathbf{x}_t) - \sum_{t=1}^T \mathcal{D}_{f_t}(\mathbf{u}_t, \mathbf{x}_t)} \\ &\leq \alpha\sqrt{\|\nabla f_1(\mathbf{x}_1)\|^2 + 4DL^2P_T + 4V_T + 32L\delta T} + 4\alpha^2L \\ &\leq \mathcal{O}\left(\sqrt{(D^2 + DP_T)V_T} + \|\nabla f_1(\mathbf{x}_1)\|\sqrt{D^2 + DP_T} + L(D^2 + DP_T) + \sqrt{(D^2 + DP_T)L\delta T}\right), \end{aligned}$$

where we use AM-GM, i.e.,  $2\sqrt{ab} - b \leq a$  for  $a, b \geq 0$ , to cancel the Bregman divergence in the third inequality. Now with  $L = \delta^{\frac{\nu-1}{1+\nu}} L_\nu^{\frac{2}{1+\nu}}$ , and choosing  $\delta = L_\nu(D^2 + DP_T)^{\frac{1+\nu}{2}} T^{-\frac{1+\nu}{2}}$ , we achieve

$$\begin{aligned} \text{D-REG}_T(\mathbf{u}_1, \dots, \mathbf{u}_T) &\leq \mathcal{O}\left(\sqrt{(D^2 + DP_T)V_T} + L_\nu(D^2 + DP_T)^{\frac{1+\nu}{2}} T^{\frac{1-\nu}{2}} + \|\nabla f_1(\mathbf{x}_1)\|\sqrt{D^2 + DP_T}\right), \end{aligned}$$

which completes the proof.  $\square$

## C.2 Proof of Theorem 6

To begin with, we give the details of the algorithm [Yan et al., 2024] specified for convex and strongly convex functions. This is a two-layer structure, where there are  $N$  base-learners, each implemented with optimistic OGD running on carefully designed surrogate functions to generate local decisions



---

**Algorithm 3** Universal Regret Minimization under Hölder Smoothness

---

**Input:** Domain diameter  $D$ , total iterations  $T$

- 1: **Initialization:**  $N = \lceil \log_2 T \rceil + 2$ , starting points  $\{\mathbf{x}_{1,i}\}_{i=1}^N, \{\widehat{\mathbf{x}}_{1,i}\}_{i=1}^N, p_{1,i} = \frac{1}{N}$
  - 2: **for**  $t = 1$  **to**  $T$  **do**
  - 3:   Submit  $\mathbf{x}_t = \sum_{i=1}^N p_{t,i} \mathbf{x}_{t,i}$ , receive  $\nabla f_t(\mathbf{x}_t)$
  - 4:   Construct surrogate functions  $h_{t,i}(\cdot)$  defined in Eq. (34) and Eq. (35)
  - 5:   The  $i$ -th base-learner updates to  $\mathbf{x}_{t+1,i}$  by Eq. (33), for all  $i \in [N]$  ▷ Base update
  - 6:   Meta-algorithm updates to  $p_{t+1}$  by Eq. (36) ▷ Meta update
  - 7: **end for**
- 

651  $\{\mathbf{x}_{t,i}\}_{i=1}^N$ . The meta-algorithm runs OPTIMISTIC-ADAPT-ML-PROD [Wei et al., 2016] to calculate  
 652 the weight  $p_t \in \Delta_N$  for combining the local decisions. The procedure is summarized in Algorithm 3.

653 Specifically, the  $i$ -th base-learner starts from  $\widehat{\mathbf{x}}_{1,i} = \mathbf{x}_{1,i}$ , and updates by

$$\widehat{\mathbf{x}}_{t+1,i} = \Pi_{\mathcal{X}}[\widehat{\mathbf{x}}_{t,i} - \eta_{t,i} \nabla h_{t,i}(\mathbf{x}_t)], \quad \mathbf{x}_{t+1,i} = \Pi_{\mathcal{X}}[\widehat{\mathbf{x}}_{t+1,i} - \eta_{t+1,i} \nabla h_{t,i}(\mathbf{x}_t)], \quad (33)$$

654 where the surrogate functions  $h_{t,i}(\mathbf{x})$  and the step sizes  $\eta_{t,i}$  are tailored for specific class of online  
 655 functions which is either convex or strongly convex whose curvature lies within the range  $[1/T, 1]$ .<sup>3</sup>

656 **Base-learners for Strongly Convex Functions.** For  $\lambda$ -strongly convex functions handled by the first  
 657  $N - 1$  base-learners, we discretize the curvature  $\lambda$  into the candidate pool  $\{\lambda_i = 2^{1-i} : i \in [N - 1]\}$ ,  
 658 and construct the surrogate functions  $h_{t,i}(\mathbf{x})$  and step sizes  $\eta_{t,i}$  for  $i \in [N - 1]$  as:

$$h_{t,i}(\mathbf{x}) \triangleq \langle \nabla f_t(\mathbf{x}_t), \mathbf{x} \rangle + \frac{\lambda_i}{4} \|\mathbf{x} - \mathbf{x}_t\|^2, \quad \eta_{t,i} = \frac{6}{\lambda_i t}. \quad (34)$$

659 **Base-learner for Convex Functions.** For convex functions handled by the  $N$ -th base-learner, we  
 660 construct the surrogate functions  $h_{t,N}(\mathbf{x})$  and step sizes  $\eta_{t,N}$  as:

$$h_{t,N}(\mathbf{x}) \triangleq \langle \nabla f_t(\mathbf{x}_t), \mathbf{x} \rangle, \quad \eta_{t,N} = \frac{D}{2\sqrt{A_{t-1}}}, \quad (35)$$

661 where  $A_t \triangleq \|\nabla f_1(\mathbf{x}_1)\|^2 + \sum_{s=2}^t \|\nabla f_s(\mathbf{x}_s) - \nabla f_{s-1}(\mathbf{x}_{s-1})\|^2$ . Therefore, there are  $N = \mathcal{O}(\log T)$   
 662 base-learners in total:  $\lceil \log_2 T \rceil + 1$  for strongly convex functions and 1 for convex functions.

663 **Meta-algorithm.** Meta learner uses OPTIMISTIC-ADAPT-ML-PROD to update the weight  $p_{t+1} \in$   
 664  $\Delta_N$  by the following update rule:

$$p_{t+1,i} \propto \varepsilon_{t,i} \exp(\varepsilon_{t,i} m_{t+1,i}) W_{t,i}, \quad (36)$$

$$W_{t,i} = (W_{t-1,i} \exp(\varepsilon_{t-1,i} r_{t,i} - \varepsilon_{t-1,i}^2 (r_{t,i} - m_{t,i})))^{\frac{\varepsilon_{t,i}}{\varepsilon_{t-1,i}}},$$

665 where we denote by  $\ell_{t,i} \triangleq \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_{t,i} \rangle$  the loss of the  $i$ -th dimension, and the meta-algorithm  
 666 inputs: (i) the instantaneous regret  $r_{t,i} = \langle \ell_t, \mathbf{p}_t \rangle - \ell_{t,i}$ ; (ii) a time-varying learning rate  $\varepsilon_{t,i} =$   
 667  $\min\{1/8, \sqrt{\ln N / \sum_{s=1}^t (r_{s,i} - m_{s,i})^2}\}$ ; (iii) optimism  $m_{t,i} = 0$  for strongly convex base-learners,  
 668 i.e.,  $i \in [N - 1]$ , and  $m_{t,N} = \langle \nabla f_{t-1}(\mathbf{x}_{t-1}), \mathbf{x}_t - \mathbf{x}_{t,N} \rangle$  for the  $N$ -th convex base-learner. Then we  
 669 give the theoretical guarantee of this meta-algorithm, i.e., OPTIMISTIC-ADAPT-ML-PROD.

670 **Lemma 8** (Theorem 3.4 of Wei et al. [2016]). *The OPTIMISTIC-ADAPT-ML-PROD algorithm that*  
 671 *updates by Eq. (36), ensures the regret with respect to any  $i \in [N]$  satisfies:*

$$\sum_{t=1}^T \langle \ell_t, \mathbf{p}_t - \mathbf{e}_i \rangle \leq C_0 \sqrt{1 + \sum_{t=1}^T (r_{t,i} - m_{t,i})^2} + C_1, \quad (37)$$

672 where  $r_{t,i} = \langle \ell_t, \mathbf{p}_t \rangle - \ell_{t,i}$ ,  $\mathbf{e}_i$  denotes the  $i$ -th standard basis vector,  $C_0 = \sqrt{\ln N} + \ln(1 + \frac{N}{e}(1 +$   
 673  $\ln(T + 1))) / \sqrt{\ln N}$ , and  $C_1 = \frac{1}{4}(\ln N + \ln(1 + \frac{N}{e}(1 + \ln(T + 1)))) + 2\sqrt{\ln N} + 16 \ln N$ .

---

<sup>3</sup> This assumption is natural because if the curvature  $\lambda < 1/T$ , the optimal regret bound, i.e.,  $\mathcal{O}(\frac{1}{\lambda} \log T)$ , leading to a vacuous  $\Omega(T)$  regret. Besides, functions with strong convexity curvature larger than 1 are also 1-strongly convex, so clipping the curvature to 1 will only worsen the regret by a constant factor.

674 In Lemma 8,  $C_0$  and  $C_1$  are in order of  $\mathcal{O}(\log \log T)$ , thus can be treated as ignorable constants,  
 675 following previous studies [Luo and Schapire, 2015; Zhao et al., 2024]. Now with the algorithm  
 676 summarized in Algorithm 3, we provide the proof for Theorem 6.

677 *Proof of Theorem 6.* Following the analysis of Yan et al. [2024], we further consider the Hölder  
 678 smoothness. To begin with, we give the following decomposition of the empirical gradient variation:

$$\begin{aligned}
 & \sum_{t=2}^T \|\nabla f_t(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_{t-1})\|^2 \\
 &= \sum_{t=2}^T \|\nabla f_t(\mathbf{x}_t) - \nabla f_t(\mathbf{x}_*) + \nabla f_t(\mathbf{x}_*) - \nabla f_{t-1}(\mathbf{x}_*) + \nabla f_{t-1}(\mathbf{x}_*) - \nabla f_{t-1}(\mathbf{x}_{t-1})\|^2 \\
 &\leq 3V_T + 12L \sum_{t=1}^T \mathcal{D}_{f_t}(\mathbf{x}_*, \mathbf{x}_t) + 12L\delta T,
 \end{aligned} \tag{38}$$

679 where we apply Lemma 4 in the inequality. In the following, we first bound for convex functions.

680 **For Convex Functions.** We decompose the regret as

$$\begin{aligned}
 \text{REG}_T &= \sum_{t=1}^T \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_* \rangle - \sum_{t=1}^T \mathcal{D}_{f_t}(\mathbf{x}_*, \mathbf{x}_t) \\
 &= \underbrace{\sum_{t=1}^T \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t,i_*} \rangle}_{\text{META-REG}} + \underbrace{\sum_{t=1}^T \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_{t,i_*} - \mathbf{x}_* \rangle}_{\text{BASE-REG}} - \sum_{t=1}^T \mathcal{D}_{f_t}(\mathbf{x}_*, \mathbf{x}_t),
 \end{aligned} \tag{39}$$

681 where for base-regret, we can define surrogate functions  $h_t(\mathbf{x}) = \langle \nabla f_t(\mathbf{x}_t), \mathbf{x} \rangle$ .

682 And for the meta regret, since we define  $\ell_{t,i} \triangleq \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_{t,i} \rangle$  the loss of the  $i$ -th dimension, and

683  $\mathbf{x}_t = \sum_{i=1}^N p_{t,i} \mathbf{x}_{t,i}$ , denoting by  $\mathbf{e}_i$  the  $i$ -th standard basis vector, we can bound it as:

$$\begin{aligned}
 \text{META-REG} &= \sum_{t=1}^T \langle \ell_t, \mathbf{p}_t - \mathbf{e}_{i_*} \rangle \stackrel{(37)}{\leq} C_0 \sqrt{1 + \sum_{t=1}^T (r_{t,i_*} - m_{t,i_*})^2} + C_1 \\
 &= C_0 \sqrt{1 + \sum_{t=1}^T \langle \nabla f_t(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_{t-1}), \mathbf{x}_t - \mathbf{x}_{t,i_*} \rangle^2} + C_1 \\
 &\leq C_0 \sqrt{1 + D^2 \sum_{t=1}^T \|\nabla f_t(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_{t-1})\|^2} + C_1 \\
 &\stackrel{(38)}{\leq} C_0 \sqrt{1 + D^2 \|\nabla f_1(\mathbf{x}_1)\|^2 + 3D^2 V_T + 12LD^2 \sum_{t=1}^T \mathcal{D}_{f_t}(\mathbf{x}_*, \mathbf{x}_t) + 12LD^2 \delta T} + C_1 \\
 &\leq \mathcal{O} \left( D \sqrt{\|\nabla f_1(\mathbf{x}_1)\|^2 + V_T + L\delta T} \right) + 3C_0 C_2 L D^2 + \frac{C_0}{C_2} \sum_{t=1}^T \mathcal{D}_{f_t}(\mathbf{x}_*, \mathbf{x}_t).
 \end{aligned}$$

684 In the first line, we apply Lemma 8. In the second line we use the definition of  $r_{t,i_*}$  and  $m_{t,i_*}$ , i.e.,  
 685  $r_{t,i_*} = \langle \ell_t, \mathbf{p}_t \rangle - \ell_{t,i_*}$  and  $m_{t,i_*} = \langle \nabla f_{t-1}(\mathbf{x}_{t-1}), \mathbf{x}_t - \mathbf{x}_{t,i_*} \rangle$  for the  $i_*$ -th convex base-learner.  
 686 The third line uses Hölder's inequality and Assumption 1. The fourth line applies the decomposition  
 687 in (38). The last line uses the inequality  $2\sqrt{ab} \leq C_2 a + b/C_2$ , with  $C_2 > 0$  to be determined later.

688 For the base regret, since we run Optimistic OGD over the convex surrogate functions  $h_{t,i_*}(\mathbf{x})$  with  
 689 the same algorithm configs as Theorem 1, by (20), we obtain:

$$\text{BASE-REG} \stackrel{(20)}{\leq} 3D \sqrt{\|\nabla h_{1,i_*}(\mathbf{x}_{1,i_*})\|^2 + \sum_{t=2}^T \|\nabla h_{t,i_*}(\mathbf{x}_{t,i_*}) - \nabla h_{t-1,i_*}(\mathbf{x}_{t-1,i_*})\|^2}$$

$$\begin{aligned}
&= 3D \sqrt{\|\nabla f_1(\mathbf{x}_1)\|^2 + \sum_{t=2}^T \|\nabla f_t(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_{t-1})\|^2} \\
&\stackrel{(38)}{\leq} 3D \sqrt{\|\nabla f_1(\mathbf{x}_1)\|^2 + 3V_T + 12L \sum_{t=1}^T \mathcal{D}_{f_t}(\mathbf{x}_*, \mathbf{x}_t) + 12L\delta T} \\
&\leq \mathcal{O}\left(D\sqrt{\|\nabla f_1(\mathbf{x}_1)\|^2 + V_T + L\delta T}\right) + 9C_3LD^2 + \frac{3}{C_3} \sum_{t=1}^T \mathcal{D}_{f_t}(\mathbf{x}_*, \mathbf{x}_t).
\end{aligned}$$

690 In the first line, we use the regret analysis of Optimistic OGD for convex functions in (20). The  
691 second line substitutes the definition  $h_{t,i_*}(\mathbf{x}) \triangleq \langle \nabla f_t(\mathbf{x}_t), \mathbf{x} \rangle$  for convex base-learner. The third  
692 line applies the decomposition in (38). The last line uses the inequality  $2\sqrt{ab} \leq C_3a + b/C_3$ , with  
693  $C_3 > 0$  to be determined later.

694 Now substituting the meta regret and base regret into (39), we have

$$\begin{aligned}
\text{REG}_T &\leq \mathcal{O}\left(D\sqrt{\|\nabla f_1(\mathbf{x}_1)\|^2 + V_T + L\delta T}\right) + (3C_0C_2 + 9C_3)LD^2 \\
&\quad + \left(\frac{C_0}{C_2} + \frac{3}{C_3} - 1\right) \sum_{t=1}^T \mathcal{D}_{f_t}(\mathbf{x}_*, \mathbf{x}_t).
\end{aligned}$$

695 Then by setting  $C_2 = 2C_0$ ,  $C_3 = 6$ , and  $\delta = L_\nu D^{1+\nu} T^{-\frac{1+\nu}{2}}$  that only exists in analysis, we have

$$\text{REG}_T \leq \mathcal{O}\left(D\sqrt{V_T} + L_\nu D^{1+\nu} T^{\frac{1-\nu}{2}} + D\|\nabla f_1(\mathbf{x}_1)\|\right).$$

696 **For Strongly-Convex Functions.** Following Yan et al. [2024], for  $i_* \in [N-1]$  such that  $\lambda_{i_*} \leq \lambda \leq$   
697  $2\lambda_{i_*}$ , we decompose the regret as

$$\begin{aligned}
\text{REG}_T &= \sum_{t=1}^T \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_* \rangle - \frac{1}{2} \sum_{t=1}^T \mathcal{D}_{f_t}(\mathbf{x}_*, \mathbf{x}_t) - \frac{1}{2} \sum_{t=1}^T \mathcal{D}_{f_t}(\mathbf{x}_*, \mathbf{x}_t) \\
&\leq \sum_{t=1}^T \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t,i_*} + \mathbf{x}_{t,i_*} - \mathbf{x}_* \rangle - \frac{\lambda}{4} \sum_{t=1}^T \|\mathbf{x}_t - \mathbf{x}_*\|^2 - \frac{1}{2} \sum_{t=1}^T \mathcal{D}_{f_t}(\mathbf{x}_*, \mathbf{x}_t) \\
&\leq \underbrace{\sum_{t=1}^T \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t,i_*} \rangle - \frac{\lambda_{i_*}}{4} \sum_{t=1}^T \|\mathbf{x}_t - \mathbf{x}_{t,i_*}\|^2}_{\text{META-REG}} \\
&\quad + \underbrace{\sum_{t=1}^T (h_{t,i_*}(\mathbf{x}_{t,i_*}) - h_{t,i_*}(\mathbf{x}_*)) - \frac{1}{2} \sum_{t=1}^T \mathcal{D}_{f_t}(\mathbf{x}_*, \mathbf{x}_t)}_{\text{BASE-REG}}. \tag{40}
\end{aligned}$$

698 In the second line we use  $\mathcal{D}_{f_t}(\mathbf{x}, \mathbf{y}) \geq \frac{\lambda}{2} \|\mathbf{x} - \mathbf{y}\|^2$  for  $\lambda$ -strongly convex function  $f_t$ . The third line  
699 we use  $\lambda \geq \lambda_{i_*}$  and define the surrogate  $h_{t,i}(\mathbf{x}) \triangleq \langle \nabla f_t(\mathbf{x}_t), \mathbf{x} \rangle + \frac{\lambda_i}{4} \|\mathbf{x} - \mathbf{x}_t\|^2$  for strongly convex  
700 base-learners.

701 The meta regret can be bounded as:

$$\begin{aligned}
\text{META-REG} &\leq C_0 \sqrt{1 + \sum_{t=1}^T (r_{t,i_*} - m_{t,i_*})^2 + C_1 - \frac{\lambda_{i_*}}{4} \sum_{t=1}^T \|\mathbf{x}_t - \mathbf{x}_{t,i_*}\|^2} \\
&= C_0 \sqrt{1 + \sum_{t=1}^T \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t,i_*} \rangle^2 + C_1 - \frac{\lambda_{i_*}}{4} \sum_{t=1}^T \|\mathbf{x}_t - \mathbf{x}_{t,i_*}\|^2} \\
&\leq C_0 \sqrt{1 + \tilde{G}_{\max}^2 \sum_{t=1}^T \|\mathbf{x}_t - \mathbf{x}_{t,i_*}\|^2 + C_1 - \frac{\lambda_{i_*}}{4} \sum_{t=1}^T \|\mathbf{x}_t - \mathbf{x}_{t,i_*}\|^2}
\end{aligned}$$

$$\leq C_0 + C_1 + \frac{2C_0^2 \tilde{G}_{\max}^2}{\lambda_{i_*}} - \frac{\lambda_{i_*}}{8} \sum_{t=1}^T \|\mathbf{x}_t - \mathbf{x}_{t,i_*}\|^2.$$

702 In the first line we again use the regret analysis of OPTIMISTIC-ADAPT-ML-PROD [Wei et al.,  
703 2016, Theorem 3.4] with two constants  $C_0$  and  $C_1$ . The second line we substitute the definition of  
704  $r_{t,i_*} = \langle \ell_t, \mathbf{p}_t \rangle - \ell_{t,i_*}$  and  $m_{t,i_*} = 0$  for the  $i_*$ -th strongly convex base-learner. In the third line we  
705 define  $\tilde{G}_{\max}^2 \triangleq \max_{t \in [T]} \|\nabla f_t(\mathbf{x}_t)\|^2$ . The last line uses AM-GM, i.e.,  $2\sqrt{ab} - b \leq a$  for  $a, b \geq 0$ .

706 For the base regret, since we run Optimistic OGD over the  $\frac{\lambda_{i_*}}{2}$ -strongly convex surrogate functions  
707  $h_{t,i_*}(\mathbf{x})$  with the same algorithm configs as Theorem 3, by (21), we obtain:

$$\begin{aligned} \text{BASE-REG} &\stackrel{(21)}{\leq} \sum_{t=1}^T \frac{12}{\lambda_{i_*} t} \|\nabla h_{t,i_*}(\mathbf{x}_{t,i_*}) - \nabla h_{t-1,i_*}(\mathbf{x}_{t-1,i_*})\|^2 \\ &= \sum_{t=1}^T \frac{12}{\lambda_{i_*} t} \left\| \nabla f_t(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_{t-1}) + \frac{\lambda_{i_*}}{2} (\mathbf{x}_{t,i_*} - \mathbf{x}_t) - \frac{\lambda_{i_*}}{2} (\mathbf{x}_{t-1,i_*} - \mathbf{x}_{t-1}) \right\|^2 \\ &\leq \sum_{t=1}^T \frac{36}{\lambda_{i_*} t} \|\nabla f_t(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_{t-1})\|^2 + \sum_{t=1}^T \frac{18\lambda_{i_*}}{t} \|\mathbf{x}_{t,i_*} - \mathbf{x}_t\|^2 \\ &\leq \sum_{t=2}^T \frac{108}{\lambda_{i_*} t} \|\nabla f_t(\mathbf{x}_*) - \nabla f_{t-1}(\mathbf{x}_*)\|^2 + \frac{432L\delta(1 + \ln T)}{\lambda_{i_*}} + \sum_{t=1}^T \frac{432L}{\lambda_{i_*} t} \mathcal{D}_{f_t}(\mathbf{x}_*, \mathbf{x}_t) \\ &\quad + \sum_{t=1}^T \frac{18\lambda_{i_*}}{t} \|\mathbf{x}_{t,i_*} - \mathbf{x}_t\|^2 + \frac{36}{\lambda_{i_*}} \|\nabla f_1(\mathbf{x}_1)\|^2. \end{aligned}$$

708 In the first line we apply the regret analysis of Optimistic OGD for convex functions in (21). The  
709 second line substitutes the definition of  $h_{t,i_*}(\mathbf{x})$  for the  $i_*$ -th strongly convex base-learner, i.e.,  
710  $h_{t,i_*}(\mathbf{x}) \triangleq \langle \nabla f_t(\mathbf{x}_t), \mathbf{x} \rangle + \frac{\lambda_{i_*}}{4} \|\mathbf{x} - \mathbf{x}_t\|^2$ . The fourth line apply the same decomposition as in (38),  
711 and  $\sum_{t=1}^T \frac{1}{t} \leq 1 + \ln T$ .

712 Now combining the meta regret and base regret in (40), and applying two technical lemmas to perform  
713 cancellation (Lemma 15 and Lemma 16), we have:

$$\begin{aligned} \text{REG}_T &\leq \sum_{t=2}^T \frac{108}{\lambda_{i_*} t} \sup_{\mathbf{x} \in \mathcal{X}} \|\nabla f_t(\mathbf{x}) - \nabla f_{t-1}(\mathbf{x})\|^2 + \frac{432L\delta(1 + \ln T)}{\lambda_{i_*}} + \sum_{t=1}^T \left( \frac{432L}{\lambda_{i_*} t} - \frac{1}{2} \right) \mathcal{D}_{f_t}(\mathbf{x}_*, \mathbf{x}_t) \\ &\quad + \sum_{t=1}^T \left( \frac{18\lambda_{i_*}}{t} - \frac{\lambda_{i_*}}{8} \right) \|\mathbf{x}_{t,i_*} - \mathbf{x}_t\|^2 + \frac{36}{\lambda_{i_*}} \|\nabla f_1(\mathbf{x}_1)\|^2 + C_0 + C_1 + \frac{2C_0^2 \tilde{G}_{\max}^2}{\lambda_{i_*}} \\ &\leq \mathcal{O} \left( \frac{\tilde{G}_{\max}^2}{\lambda_{i_*}} \ln \left( 1 + \frac{V_T}{\tilde{G}_{\max}^2} \right) + \frac{L\delta \ln T}{\lambda_{i_*}} + \frac{LL_\nu D^{1+\nu}}{\lambda_{i_*}} \ln \left( 1 + \frac{L}{\lambda_{i_*}} \right) + \lambda_{i_*} D^2 + \frac{\tilde{G}_{\max}^2}{\lambda_{i_*}} \right). \end{aligned}$$

714 In the second inequality, we define  $\hat{G}_{\max}^2 \triangleq \max_{t \in [T-1]} \sup_{\mathbf{x} \in \mathcal{X}} \|\nabla f_t(\mathbf{x}) - \nabla f_{t+1}(\mathbf{x})\|^2$ , and use  
715 the property of  $(L_\nu, \nu)$ -Hölder smooth function  $f_t$  that  $\mathcal{D}_{f_t}(\mathbf{x}, \mathbf{y}) \leq L_\nu D^{1+\nu}$  [Nesterov, 2015] when  
716 applying Lemma 16. By solving the trade-off:  $L\delta \ln T = LL_\nu D^{1+\nu}$ , we arrive at:

$$\text{REG}_T \leq \mathcal{O} \left( \frac{\hat{G}_{\max}^2}{\lambda} \ln \left( 1 + \frac{V_T}{\hat{G}_{\max}^2} \right) + \frac{L_\nu^2 D^{2\nu}}{\lambda} (\log T)^{\frac{1-\nu}{1+\nu}} + \frac{\max_{t \in [T]} \|\nabla f_t(\mathbf{x}_t)\|^2}{\lambda} \right),$$

717 where we treat  $\ln(1 + L_\nu(\ln T)^{(1-\nu)/(1+\nu)})/(\lambda D^{1-\nu}) = \mathcal{O}(1)$ , because it only consists of the loga-  
718 rithm of the constant  $\frac{L_\nu}{\lambda D^{1-\nu}}$ , and we treat double logarithmic factors in  $T$  as a constant, following  
719 previous studies [Luo and Schapire, 2015; Zhao et al., 2024]. The proof is finished.  $\square$

## 720 D Supporting Lemmas

721 In this section, we provide some fundamental lemmas for this paper.

## 722 D.1 Lemmas for Optimistic OGD Algorithms

723 In the following, we provide useful lemmas for optimistic OGD and its one-step variant.

724 **Lemma 9** (Proposition 7 of [Chiang et al. \[2012\]](#)). *Consider the following two updates: (i)  $\mathbf{x} =$*   
 725  *$\arg \min_{\mathbf{x} \in \mathcal{X}} \{\langle \mathbf{g}, \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c})\}$ , and (ii)  $\mathbf{x}' = \arg \min_{\mathbf{x} \in \mathcal{X}} \{\langle \mathbf{g}', \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c})\}$ , where the regu-*  
 726 *larizer  $\psi : \mathcal{X} \rightarrow \mathbb{R}$  is  $\lambda$ -strongly convex function with respect to  $\|\cdot\|$ , we have  $\lambda \|\mathbf{x} - \mathbf{x}'\| \leq \|\mathbf{g} - \mathbf{g}'\|_*$ .*

727 **Lemma 10** (Bregman proximal inequality, Lemma 3.2 of [Chen and Teboulle \[1993\]](#)). *Consider*  
 728 *the following update:  $\mathbf{x} = \arg \min_{\mathbf{x} \in \mathcal{X}} \{\langle \mathbf{g}, \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c})\}$ , where the regularizer  $\psi : \mathcal{X} \rightarrow \mathbb{R}$  is*  
 729 *convex function, then for all  $\mathbf{u} \in \mathcal{X}$ , we have  $\langle \mathbf{g}, \mathbf{x} - \mathbf{u} \rangle \leq \mathcal{D}_\psi(\mathbf{u}, \mathbf{c}) - \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}) - \mathcal{D}_\psi(\mathbf{x}, \mathbf{c})$ .*

730 **Lemma 11** (Theorem 1 of [Zhao et al. \[2024\]](#)). *Under Assumption 1, Optimistic OGD specialized*  
 731 *at Eq. (6), that starts at  $\hat{\mathbf{x}}_1 \in \mathcal{X}$  and updates by*

$$\mathbf{x}_t = \Pi_{\mathcal{X}} [\hat{\mathbf{x}}_t - \eta_t M_t], \quad \hat{\mathbf{x}}_{t+1} = \Pi_{\mathcal{X}} [\hat{\mathbf{x}}_t - \eta_t \nabla f_t(\mathbf{x}_t)],$$

732 ensures that

$$\begin{aligned} \sum_{t=1}^T \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u}_t \rangle &\leq \underbrace{\sum_{t=1}^T \langle \nabla f_t(\mathbf{x}_t) - M_t, \mathbf{x}_t - \hat{\mathbf{x}}_{t+1} \rangle}_{\text{TERM-A}} + \underbrace{\sum_{t=1}^T \frac{1}{2\eta_t} (\|\mathbf{u}_t - \hat{\mathbf{x}}_t\|^2 - \|\mathbf{u}_t - \hat{\mathbf{x}}_{t+1}\|^2)}_{\text{TERM-B}} \\ &\quad - \underbrace{\sum_{t=1}^T \frac{1}{2\eta_t} (\|\mathbf{x}_t - \hat{\mathbf{x}}_{t+1}\|^2 + \|\mathbf{x}_t - \hat{\mathbf{x}}_t\|^2)}_{\text{TERM-C}}, \end{aligned} \quad (41)$$

733 where  $\mathbf{u}_1, \dots, \mathbf{u}_T \in \mathcal{X}$  are arbitrary comparators.

734 **Lemma 12.** *Under Assumption 1, Optimistic OGD specialized at Eq. (6) with non-increasing step*  
 735 *sizes  $\eta_t$ , ensures that*

$$\sum_{t=1}^T \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u}_t \rangle \leq \sum_{t=1}^T \eta_{t+1} \|\nabla f_t(\mathbf{x}_t) - M_t\|^2 + \frac{D^2 + DP_T}{\eta_{T+1}} - \sum_{t=2}^T \frac{1}{8\eta_{t+1}} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2, \quad (42)$$

736 where  $P_T \triangleq \sum_{t=2}^T \|\mathbf{u}_t - \mathbf{u}_{t-1}\|$  is the path length.

737 **Lemma 13** (One-step Variant of Optimistic OGD, [\[Joulani et al., 2020\]](#)). *Under Assumption 1, the*  
 738 *one-step variant of optimistic OGD that starts at  $\mathbf{x}_1 \in \mathcal{X}$  and updates by*

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}} [\mathbf{x}_t - \eta_t (\nabla f_t(\mathbf{x}_t) - M_t + M_{t+1})], \quad (43)$$

739 ensures that, for all  $\mathbf{u} \in \mathcal{X}$ :

$$\begin{aligned} \sum_{t=1}^T \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u} \rangle &\leq \sum_{t=1}^T \left( \langle \nabla f_t(\mathbf{x}_t) - M_t, \mathbf{x}_t - \mathbf{x}_{t+1} \rangle - \frac{1}{2\eta_t} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \right) \\ &\quad + \sum_{t=2}^T \left( \frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) \|\mathbf{x}_t - \mathbf{u}\|^2 + \frac{1}{2\eta_1} \|\mathbf{x}_1 - \mathbf{u}\|^2. \end{aligned}$$

740 *Proof of Lemma 12.* By [Lemma 11 \[Zhao et al., 2024\]](#), we consider each term:

$$\begin{aligned} \text{TERM-A} &\leq \sum_{t=1}^T \eta_{t+1} \|\nabla f_t(\mathbf{x}_t) - M_t\|_*^2 + \sum_{t=1}^T \left( \frac{1}{4\eta_{t+1}} - \frac{1}{4\eta_t} \right) \|\mathbf{x}_t - \hat{\mathbf{x}}_{t+1}\|^2 + \sum_{t=1}^T \frac{1}{4\eta_t} \|\mathbf{x}_t - \hat{\mathbf{x}}_{t+1}\|^2 \\ &\leq \sum_{t=1}^T \eta_{t+1} \|\nabla f_t(\mathbf{x}_t) - M_t\|_*^2 + \frac{D^2}{4\eta_{T+1}} + \sum_{t=1}^T \frac{1}{4\eta_t} \|\mathbf{x}_t - \hat{\mathbf{x}}_{t+1}\|^2, \\ \text{TERM-B} &\leq \frac{D^2}{2\eta_1} + \sum_{t=2}^T \left( \frac{1}{2\eta_t} \|\mathbf{u}_t - \hat{\mathbf{x}}_t\|^2 - \frac{1}{2\eta_t} \|\mathbf{u}_{t-1} - \hat{\mathbf{x}}_t\|^2 + \frac{1}{2\eta_t} \|\mathbf{u}_{t-1} - \hat{\mathbf{x}}_t\|^2 - \frac{1}{2\eta_{t-1}} \|\mathbf{u}_{t-1} - \hat{\mathbf{x}}_t\|^2 \right) \\ &\leq \frac{D^2}{2\eta_1} + \sum_{t=2}^T \frac{1}{2\eta_t} (\|\mathbf{u}_t - \hat{\mathbf{x}}_t\|^2 - \|\mathbf{u}_{t-1} - \hat{\mathbf{x}}_t\|^2) + \sum_{t=2}^T \left( \frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) D^2 \end{aligned}$$

$$\begin{aligned}
&\leq \frac{D^2}{2\eta_T} + \sum_{t=2}^T \frac{1}{2\eta_t} \|\mathbf{u}_t - \mathbf{u}_{t-1}\| \cdot \|\mathbf{u}_t - \hat{\mathbf{x}}_t + \mathbf{u}_{t-1} - \hat{\mathbf{x}}_t\| \\
&\leq \frac{D^2}{2\eta_{T+1}} + \frac{DP_T}{\eta_{T+1}}, \\
\text{TERM-C} &\geq \sum_{t=1}^T \frac{1}{4\eta_t} \|\mathbf{x}_t - \hat{\mathbf{x}}_{t+1}\|^2 + \sum_{t=2}^T \frac{1}{4\eta_{t-1}} (\|\mathbf{x}_{t-1} - \hat{\mathbf{x}}_t\|^2 + \|\mathbf{x}_t - \hat{\mathbf{x}}_t\|^2) \\
&\geq \sum_{t=1}^T \frac{1}{4\eta_t} \|\mathbf{x}_t - \hat{\mathbf{x}}_{t+1}\|^2 + \sum_{t=2}^T \left( \frac{1}{8\eta_{t-1}} - \frac{1}{8\eta_t} \right) \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + \sum_{t=2}^T \frac{1}{8\eta_t} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 \\
&\geq \sum_{t=1}^T \frac{1}{4\eta_t} \|\mathbf{x}_t - \hat{\mathbf{x}}_{t+1}\|^2 - \frac{D^2}{8\eta_T} + \sum_{t=2}^T \left( \frac{1}{8\eta_t} - \frac{1}{8\eta_{t+1}} \right) \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + \sum_{t=2}^T \frac{1}{8\eta_{t+1}} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 \\
&\geq \sum_{t=1}^T \frac{1}{4\eta_t} \|\mathbf{x}_t - \hat{\mathbf{x}}_{t+1}\|^2 - \frac{D^2}{4\eta_{T+1}} + \sum_{t=2}^T \frac{1}{8\eta_{t+1}} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2,
\end{aligned}$$

741 where we apply [Assumption 1](#) and the condition that  $\eta_t$  is non-increasing. Combining TERM-A,  
742 TERM-B and TERM-C finishes the proof.  $\square$

743 *Proof of [Lemma 13](#).* By [Lemma 10](#) with  $\psi(\mathbf{x}) = \frac{1}{2\eta} \|\mathbf{x}\|^2$ , the update [Eq. \(43\)](#) implies for all  $\mathbf{u} \in \mathcal{X}$ :

$$\langle \nabla f_t(\mathbf{x}_t) - M_t + M_{t+1}, \mathbf{x}_{t+1} - \mathbf{u} \rangle \leq \frac{1}{2\eta_t} (\|\mathbf{u} - \mathbf{x}_t\|^2 - \|\mathbf{u} - \mathbf{x}_{t+1}\|^2 - \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2).$$

744 Then by rearranging and taking summation from  $t = 1$  to  $T$ , we arrive at:

$$\begin{aligned}
&\sum_{t=1}^T \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u} \rangle \\
&\leq \sum_{t=1}^T \langle \nabla f_t(\mathbf{x}_t) - M_t, \mathbf{x}_t - \mathbf{x}_{t+1} \rangle + \langle M_1, \mathbf{x}_1 \rangle - \langle M_{T+1}, \mathbf{x}_{T+1} \rangle + \sum_{t=1}^T \langle M_{t+1} - M_t, \mathbf{u} \rangle \\
&\quad + \sum_{t=2}^T \left( \frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) \|\mathbf{x}_t - \mathbf{u}\|^2 + \frac{1}{2\eta_1} \|\mathbf{x}_1 - \mathbf{u}\|^2 - \sum_{t=1}^T \frac{1}{2\eta_t} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \\
&\leq \sum_{t=1}^T \left( \langle \nabla f_t(\mathbf{x}_t) - M_t, \mathbf{x}_t - \mathbf{x}_{t+1} \rangle - \frac{1}{2\eta_t} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \right) \\
&\quad + \sum_{t=2}^T \left( \frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) \|\mathbf{x}_t - \mathbf{u}\|^2 + \frac{1}{2\eta_1} \|\mathbf{x}_1 - \mathbf{u}\|^2,
\end{aligned}$$

745 where we define  $M_1 \triangleq \mathbf{0}$  and  $M_{T+1} \triangleq \mathbf{0}$ .  $\square$

## 746 D.2 Useful Lemmas

747 This part provides some useful lemmas for mathematical analysis.

748 **Lemma 14** ([McMahan and Streeter \[2010\]](#)). Suppose non-negative sequence  $\{a_t\}_{t=1}^T$  and constant  
749  $\delta > 0$ , then we have

$$\sum_{t=1}^T \frac{a_t}{\sqrt{\delta + \sum_{s=1}^t a_s}} \leq 2\sqrt{\delta + \sum_{t=1}^T a_t}. \quad (44)$$

750 **Lemma 15.** Suppose non-negative sequence  $\{a_t\}_{t=1}^T$ . Define  $a_{\max} = \max_{t \in [T]} a_t$  and assume  
751  $a_{\max} > 0$ , then we have

$$\sum_{t=1}^T \frac{a_t}{t} \leq a_{\max} \ln \left( 1 + \frac{1}{a_{\max}} \sum_{t=1}^T a_t \right) + 2a_{\max}.$$



752 **Lemma 16.** Suppose  $A > 0$  and non-negative sequence  $\{b_t\}_{t=1}^T$  and denote by  $b_{\max} =$   
 753  $\max_{t \in [T]} b_t > 0$ . Then it holds that

$$\sum_{t=1}^T \left( \frac{A}{t} - 1 \right) b_t \leq b_{\max} \cdot A \ln(1 + A).$$

754 *Proof of Lemma 15.* Define  $\tau = \lceil \frac{1}{a_{\max}} \sum_{t=1}^T a_t \rceil \in [T]$ . We have

$$\sum_{t=1}^{\tau} \frac{a_t}{t} \leq a_{\max} \sum_{t=1}^{\tau} \frac{1}{t} \leq a_{\max} \left( 1 + \int_{x=1}^{\tau} \frac{1}{x} dx \right) \leq a_{\max} \ln \left( 1 + \frac{1}{a_{\max}} \sum_{t=1}^T a_t \right) + a_{\max}.$$

755 If  $\tau < T$ , we also have

$$\sum_{t=\tau+1}^T \frac{a_t}{t} \leq \frac{1}{\tau} \sum_{t=\tau+1}^T a_t \leq \frac{a_{\max}}{\sum_{t=1}^T a_t} \sum_{t=1}^T a_t = a_{\max}.$$

756 □

757 *Proof of Lemma 16.* Define  $\tau = \min\{T, \lfloor A \rfloor\}$  and trivially assume  $\tau \geq 1$ , then we have

$$\frac{1}{b_{\max}} \sum_{t=1}^T \left( \frac{A}{t} - 1 \right) b_t \leq \sum_{t=1}^{\tau} \left( \frac{A}{t} - 1 \right) \leq A \left( 1 + \int_{s=1}^{\tau} \frac{1}{s} ds \right) - \tau = A + A \ln \tau - \tau,$$

758 whose maximum is  $A \ln A \leq A \ln(1 + A)$ . □

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: We have claimed the paper's contribution in both the abstract and the introduction part.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: This paper clearly outlines the assumptions that may be required for the theorems and regards some of them as limitations (such as bounded domain assumption or deterministic setting), and discusses the future areas for improvement in the corresponding sections.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The problem setups and assumptions are provided in [Section 2](#). We provide theoretical guarantees in [Section 3](#), [Section 4](#) and [Section 5](#), and all the corresponding proofs can be found in [Appendix A](#), [Appendix B](#) and [Appendix C](#).

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: This paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
- (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: This paper does not include experiments requiring code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: This paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: This paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification: This paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: The research conducted in the paper conforms with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: This paper is a purely theoretical work, and we do not find specific societal impacts that should be highlighted here.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This paper does not include experiments (data or models), and thus poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: This paper does not use existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

### 13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: This paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.



- 1069
- 1070
- 1071
- 1072
- 1073
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
  - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

1074 **16. Declaration of LLM usage**

1075 Question: Does the paper describe the usage of LLMs if it is an important, original, or  
1076 non-standard component of the core methods in this research? Note that if the LLM is used  
1077 only for writing, editing, or formatting purposes and does not impact the core methodology,  
1078 scientific rigorousness, or originality of the research, declaration is not required.

1079 Answer: [NA]

1080 Justification: The core method development in this research does not involve LLMs as any  
1081 important, original, or non-standard components.

1082 Guidelines:

- 1083
- 1084
- 1085
- 1086
- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
  - Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.