
Regret Bounds without Lipschitz Continuity: Online Learning with Relative-Lipschitz Losses

Yihan Zhou*

University of British Columbia

Victor S. Portella*

University of British Columbia

Mark Schmidt

University of British Columbia
CCAI Affiliate Chair (Amii)

Nicholas J. A. Harvey

University of British Columbia

Abstract

In online convex optimization (OCO), Lipschitz continuity of the functions is commonly assumed in order to obtain sublinear regret. Moreover, many algorithms have only logarithmic regret when these functions are also strongly convex. Recently, researchers from convex optimization proposed the notions of “relative Lipschitz continuity” and “relative strong convexity”. Both of the notions are generalizations of their classical counterparts. It has been shown that subgradient methods in the relative setting have performance analogous to their performance in the classical setting.

In this work, we consider OCO for relative Lipschitz and relative strongly convex functions. We extend the known regret bounds for classical OCO algorithms to the relative setting. Specifically, we show regret bounds for the follow the regularized leader algorithms and a variant of online mirror descent. Due to the generality of these methods, these results yield regret bounds for a wide variety of OCO algorithms. Furthermore, we further extend the results to algorithms with extra regularization such as regularized dual averaging.

1 Introduction

In online convex optimization (OCO), at each of many rounds a player has to pick a point from a convex set while an adversary chooses a convex function that penalizes the player’s choice. More precisely, in each round $t \in \mathbb{N}$, the player picks a point x_t from a fixed convex set $\mathcal{X} \subseteq \mathbb{R}^n$ and an adversary picks a convex function f_t depending on x_t . At the end of the round, the player suffers a loss of $f_t(x_t)$. Besides modeling a wide range of online learning problems [Shalev-Shwartz, 2011], algorithms for OCO are often used in batch optimization problems due to their low computational cost per iteration. For example, the widely used stochastic gradient descent (SGD) algorithm can be viewed as a special case of online gradient descent [Hazan, 2016, Chapter 3] and AdaGrad [Duchi et al., 2011] is a foundational adaptive gradient descent method originally proposed in the OCO setting. The performance measure usually used for OCO algorithms is the *regret*. It is the difference between the cost incurred to the player and a comparison point $z \in \mathcal{X} \subseteq \mathbb{R}^n$ (usually with minimum cumulative loss), that is to say,

$$\text{Regret}_T(z) := \sum_{t=1}^T f_t(x_t) - \sum_{t=1}^T f_t(z).$$

*Equal contributions.

Classical results show that if the cost functions are Lipschitz continuous, then there are algorithms which suffer at most $O(\sqrt{T})$ regret in T rounds [Zinkevich, 2003]. Additionally, if the cost functions are strongly convex, there are algorithms that suffer at most $O(\log T)$ regret in T rounds [Hazan et al., 2007b]). However, not all loss functions that appear in applications, such as in inverse Poisson problems [Antonakopoulos et al., 2020] and support vector machines training [Lu, 2019], satisfy these conditions on the entire feasible set.

Recently, there has been a line of work investigating the performance of optimization methods beyond conventional assumptions [Bauschke et al., 2017, Lu et al., 2018, Lu, 2019]. Intriguingly, much of this line of work proposes relaxed assumptions under which classical algorithms enjoy convergence rates similar to the ones from the classical setting.

In particular, Lu [2019] proposed the notion of relative Lipschitz-continuity and showed how mirror descent (with properly chosen regularizer/mirror map) converges at a rate of $O(1/\sqrt{T})$ in T iterations for non-smooth relative Lipschitz-continuous functions. Furthermore, they show a $O(1/T)$ convergence rate when the function is also relatively strongly-convex (a notion proposed by Lu et al. [2018]). Although the former result can be translated to a $O(\sqrt{T})$ regret bound for *online mirror descent* (OMD), the latter does not directly yield regret bounds in the online setting. Moreover, Orabona and Pál [2018] showed that OMD is not suitable when we do not know a priori the number of iterations since it may suffer linear regret in this case. Finally, at present it is not known how foundational OCO algorithms such as *follow the regularized leader* (FTRL) [Shalev-Shwartz, 2011, Hazan, 2016] and *regularized dual averaging* [Xiao, 2010] (RDA) perform in the relative setting.

Our results. We analyze the performance of two general OCO algorithms: FTRL and dual-stabilized OMD (DS-OMD, see [Fang et al., 2020]). We give $O(\sqrt{T})$ regret bounds in T rounds for relative Lipschitz loss functions. Moreover, this is the first paper to show $O(\log T)$ regret if the loss functions are also relative strongly-convex.¹ In addition, we are able to extend these bounds for problems with composite loss functions, such as adding the ℓ_1 -norm to induce sparsity. The generality of these algorithms lead to regret bounds for a wide variety of OCO algorithms (see Shalev-Shwartz [2011], Hazan [2016] for some reductions). We demonstrate this flexibility by deriving convergence rates for *dual averaging* Nesterov [2009] and *regularized dual averaging* [Xiao, 2010].

1.1 Related Work

Analyses of gradient descent methods in the differentiable convex setting usually require the objective function f to be Lipschitz smooth, that is, the gradient of the objective function f is Lipschitz continuous. Bauschke et al. [2017] proposed a generalized Lipschitz smoothness condition, called *relative Lipschitz smoothness*, using Bregman divergences of a fixed reference function. They proposed a proximal mirror descent method² called NoLips with a $O(1/T)$ convergence rate for such functions. Van Nguyen [2017] independently developed similar ideas for analyzing the convergence of a Bregman proximal gradient method applied to convex composite functions in Banach spaces. Bolte et al. [2018] extended the framework of Bauschke et al. [2017] to the non-convex setting. Building upon this work, Lu et al. [2018] slightly relaxed the definition of relative smoothness and gave simpler analyses for mirror descent and dual averaging. Hanzely and Richtárik [2018] propose and analyse coordinate and stochastic gradient descent methods for relatively smooth functions. These ideas were later applied to non-convex problems by Mukkamala and Ochs [2019]. More recently, Gao et al. [2020] analysed the coordinate descent method with composite Lipschitz smooth objectives. Unlike those prior works, in this paper we focus on the online case with non-differentiable loss functions.

For non-differentiable convex optimization, Lipschitz continuity of the objective function is usually needed to obtain a $O(1/\sqrt{T})$ convergence guarantee for classical methods. Lu [2019] showed that this condition can be relaxed to what they called relative Lipschitz continuity of the objective function. Under this latter assumption, they gave $O(1/\sqrt{T})$ convergence rates for deterministic and

¹This can be seen as analogous to the known logarithmic regret bounds when the loss functions are strongly convex [Hazan et al., 2007b].

²They propose an algorithm in the general case with composite functions, but when we set $f := 0$ in their algorithm it boils down to classical mirror descent. In this case the novelty comes from the convergence analysis at a $O(1/T)$ rate without the use of classical Lipschitz smoothness.

stochastic mirror descent. In a similar vein, Grimmer [2019] showed how projected subgradient descent enjoys a $O(1/\sqrt{T})$ convergence rate without Lipschitz continuity given that one has some control on the norm of the subgradients. None of these works considered online algorithms. Although the results from Lu [2019] for mirror descent can be adapted to the online setting, it is not clear how other foundational OCO algorithms such as FTRL or RDA perform in this setting.

Antonakopoulos et al. [2020] generalized the Lipschitz continuity condition from the perspective of Riemannian geometry. They proposed the notion of Riemann-Lipschitz continuity (RLC) and analyzed how OCO algorithms perform in this setting. They showed $O(\sqrt{T})$ regret bounds for both FTRL and OMD with RLC cost functions in both the online and stochastic settings. In Appendix A we discuss in detail the relationship between RLC and relative Lipschitzness and how some of our regret bounds compare to those due to Antonakopoulos et al. [2020]. In related work, Maddison et al. [2018] relaxed the Lipschitz smoothness condition by proposing a new family of optimization methods motivated from physics, to be more specific, the conformal Hamiltonian dynamics.

Moreover, in the presence of both Lipschitz continuity and strong convexity we can obtain $O(1/T)$ convergence rates in classical convex optimization [Bubeck, 2015, Section 3.4.1] and $O(\log T)$ regret in the online case [Hazan et al., 2007b]. By replacing the squared norm in the usual strong convexity inequality by a Bregman divergence of a fixed reference function yields the notion of *relative strong convexity*. This idea dates back to the work of Hazan et al. [2007a]. In recent work, Lu et al. [2018] showed algorithms with $O(1/T)$ convergence rates in the offline setting when the objective function is both relative Lipschitz continuous and relative strongly convex. Still, this latter work does not obtain regret bounds for the online case. Hazan et al. [2007a] analyze the online case and show logarithmic regret bounds for online mirror descent when the cost functions are strongly convex relative to the mirror map. However, they assume (classical) strong convexity of the mirror map, which ultimately implies that the cost function need also be strongly convex.³ To the best of our knowledge, this is the first work studying conditions beyond strong convexity (and exp-concavity [Hazan et al., 2007b]) to obtain logarithmic regret bounds.

2 Formal Definitions

Throughout this paper, \mathbb{R}^n denotes a n -dimensional real vector space endowed with an inner-product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$. We take $\mathcal{X} \subseteq \mathbb{R}^n$ to be a fixed convex set. The **dual norm** of $\|\cdot\|$ is defined by $\|x\|_* := \sup_{y \in \mathbb{R}^n: \|y\| \leq 1} \langle x, y \rangle$ for each $x \in \mathbb{R}^n$. Moreover, for any convex function $f: \mathcal{X} \rightarrow \mathbb{R}$ and any $x \in \mathbb{R}^n$, a vector $g \in \mathbb{R}^n$ is a **subgradient** of f at x if G satisfies the *subgradient inequality*

$$f(z) \geq f(x) + \langle g, x - z \rangle, \quad \forall z \in \mathbb{R}^n. \quad (2.1)$$

We denote by $\partial f(x)$ the set of all subgradients of f at x , called the **subdifferential** of f at x . The **normal cone** of \mathcal{X} at a point $x \in \mathcal{X}$ is the set $N_{\mathcal{X}}(x) := \{a \in \mathbb{R}^n : \langle a, z - x \rangle \leq 0 \text{ for all } z \in \mathcal{X}\}$.

Let $R: \mathcal{D} \rightarrow \mathbb{R}$ be a convex function such that it is differentiable in $\mathcal{D}^\circ := \text{int } \mathcal{D}$ and such that we have $\mathcal{X} \subseteq \mathcal{D}^\circ$. The **Bregman divergence** (with respect to R) is given by

$$D_R(x, y) := R(x) - R(y) - \langle \nabla R(y), x - y \rangle, \quad \forall x \in \mathcal{D}, y \in \mathcal{D}^\circ.$$

An interesting and useful identity regarding Bregman divergences, sometimes called *three-point identity* [Bubeck, 2015], is

$$D_R(x, y) + D_R(z, x) - D_R(z, y) = \langle \nabla R(x) - \nabla R(y), x - z \rangle, \quad \forall z \in \mathcal{D}, \forall x, y \in \mathcal{D}^\circ. \quad (2.2)$$

Although the Bregman divergence with respect to R is not a metric, we can still interpret D_R as a way of measuring distances through the lens of R . An instructive example is the Bregman divergence associated with the squared ℓ_2 -norm $R := \frac{1}{2} \|\cdot\|_2^2$. In this case, we have $D_R(x, y) = \frac{1}{2} \|x - y\|_2^2$ for all $x, y \in \mathbb{R}^n$, that is, the divergence boils down to the squared ℓ_2 -distance. In light of this, a possible way to generalize Lipschitz continuity and strong convexity is to replace the norm in the classical definitions by the square root of the Bregman divergence [Lu et al., 2018].

³More precisely, the regret bound in [Hazan et al., 2007a, Theorem 1] requires the cost functions $(g_t)_{t \in \mathbb{N}}$ to be strongly convex relative to the mirror map f . In turn, the result also requires f to be strongly convex (in the classical sense) with respect to a fixed norm $\|\cdot\|$. This implies that the cost functions $(g_t)_{t \in \mathbb{N}}$ are strongly convex w.r.t. $\|\cdot\|$ as well.

First, recall that a function $f: \mathcal{X} \rightarrow \mathbb{R}$ is **L -Lipschitz continuous** with respect to $\|\cdot\|$ on $\mathcal{X}' \subseteq \mathcal{X}$ if

$$|f(x) - f(y)| \leq L\|x - y\|, \quad \forall x, y \in \mathcal{X}'.$$

Additionally, if f is convex, then the above definition implies⁴ that $\|g\|_* \leq L$ for all $x \in \mathcal{X}$ and all $g \in \partial f(x)$. Recall as well that a convex function $f: \mathcal{X} \rightarrow \mathbb{R}$ is **M -strongly convex** with respect to $\|\cdot\|$ on $\mathcal{X}' \subseteq \mathcal{X}$ for some $M > 0$ if

$$f(y) \geq f(x) + \langle g, y - x \rangle + \frac{M}{2}\|y - x\|^2, \quad \forall x, y \in \mathcal{X}', \forall g \in \partial f(x).$$

Let us now state generalizations of the above definitions due to Lu et al. [2018] and Lu [2019].

Definition 2.1 (Relative Lipschitz continuity). A convex function $f: \mathcal{X} \rightarrow \mathbb{R}$ is **L -Lipschitz continuous** relative to R if

$$\langle g, x - y \rangle \leq L\sqrt{2D_R(y, x)}, \quad \forall x, y \in \mathcal{X}, \forall g \in \partial f(x).$$

In particular, if $f: \mathcal{X} \rightarrow \mathbb{R}$ is L -Lipschitz continuous relative to R , then

$$f(x) - f(y) \stackrel{(2.1)}{\leq} \langle g, x - y \rangle \leq L\sqrt{2D_R(y, x)}, \quad \forall x, y \in \mathcal{X}, \forall g \in \partial f(x). \quad (2.3)$$

The original definition of Lu [2019] requires $\|g\|_* \|x - y\| \leq L\sqrt{2D_R(x, y)}$ for all $x, y \in \mathcal{X}$ and $g \in \partial f(x)$. Since $\langle a, b \rangle \leq \|a\|_* \|b\|$ for any $a, b \in \mathbb{R}^n$, the above definition is slightly more general and does not depend on the choice of a norm.

Definition 2.2 (Relative strong convexity [Lu et al., 2018]). A convex function $f: \mathcal{X} \rightarrow \mathbb{R}$ is **M -strongly convex** relative to R if

$$f(y) \geq f(x) + \langle g, y - x \rangle + MD_R(y, x), \quad \forall y, x \in \mathcal{X}, \forall g \in \partial f(x). \quad (2.4)$$

A notable special case of relative Lipschitz-continuity or relative strong convexity is when we pick $R := \frac{1}{2}\|\cdot\|_2^2$ and the classical definitions with respect to the ℓ_2 -norm are recovered.

Example (A function that is relative Lipschitz but not Lipschitz). Consider the function f given by $f(x) := x^2$ for each $x \in \mathbb{R}$. Since the derivative of f is unbounded on \mathbb{R} , it is not Lipschitz continuous on the entire line. Define the function R by $R(x) := 2x^4$ for all $x \in \mathbb{R}$. Then,

$$D_R(y, x) = 2y^4 - 2x^4 - 8x^3(y - x) = \frac{1}{2}(x^2 - y^2)^2 + x^2(x - y)^2 \geq x^2(x - y)^2, \quad \forall x, y \in \mathbb{R}.$$

Thus, $(f'(x)(x - y))^2 = 4x^2(x - y)^2 \leq 2 \cdot 2D_R(y, x)$ for any $x, y \in \mathbb{R}^n$. That is, f is $\sqrt{2}$ -Lipschitz continuous relative to R .

Lu [2019] discusses more substantial examples in detail, such as training of support vector machines, and finding a point in the intersection of several ellipsoids. Furthermore, he also gives a systematic way of picking a reference function for any objective functions whose subgradients at x have ℓ_2 -norm bounded by a polynomial in $\|x\|_2$. This useful construction allows many optimization problems to benefit from algorithms that are designed for the relative setting.

2.1 Conventions and Assumptions used Throughout the Paper

We collect here some additional notation and assumptions used throughout the paper.⁵ First, $\mathcal{X} \subseteq \mathbb{R}^n$ denotes a closed convex set and $\{f_t\}_{t \geq 1}$ denotes a sequence of convex functions such that $f_t: \mathcal{X} \rightarrow \mathbb{R}$ is subdifferentiable⁶ on \mathcal{X} for each $t \geq 1$. We denote by $\{\eta_t\}_{t \geq 0}$ a sequence of scalars such that $\eta_t \geq \eta_{t+1} > 0$ for each $t \geq 0$. Moreover, $\mathcal{D} \subseteq \mathbb{R}^n$ denotes a convex set with non-empty interior $\mathcal{D}^\circ := \text{int}(\mathcal{D})$ such that $\mathcal{X} \subseteq \mathcal{D}^\circ$. This latter set will be the domain of the regularizer for FTRL and of the mirror map for OMD. Namely, in Section 3 we denote by $R: \mathcal{D} \rightarrow \mathbb{R}$ the *regularizer* of FTRL, a convex function which is differentiable on \mathcal{D}° . In Section 5 we denote by $\Phi: \mathcal{D} \rightarrow \mathbb{R}$ the *mirror map* of online mirror descent (whose precise definition we postpone to Section 5).

⁴On the boundary of \mathcal{X} this implication is not as strong: we can only guarantee the existence of one subgradient with small norm. For our purposes this will not be of fundamental importance. For a more precise statement see [Ben-Tal and Nemirovski, 2001, §5.3]

⁵The only exception is Lemma 3.1, which does not need convexity or differentiability of any of the functions.

⁶This is not too restrictive since convex functions are subdifferentiable on the relative interior of their domains [Rockafellar, 1997, Theorem 23.4].

3 Follow the Regularized Leader

The *follow the regularized leader* (FTRL) algorithm is a classical method for OCO. At each round, FTRL picks a point that minimizes the cost incurred by the previously seen functions plus a regularizer convex function (an *FTRL regularizer*). Intuitively, the latter helps the choices of the algorithm not to change too widely from one round to the next. In Algorithm 1 we formally outline the FTRL algorithm. It is well known [Hazan, 2016] that, in a game with T rounds, FTRL with properly tuned step sizes suffers at most $O(\sqrt{T})$ regret against Lipschitz continuous functions.⁷ When the loss functions are additionally strongly convex, FTRL suffers at most regret $O(\log T)$. In this section we describe one of our main results: the FTRL algorithm preserves these asymptotic regret guarantees in the relative setting.

Algorithm 1 Follow the Regularized Leader (FTRL) Algorithm

```

Compute  $x_1 \in \arg \min_{x \in \mathcal{X}} R(x)$ 
Set  $F_0 := 0$ 
for  $t = 1, 2, \dots$  do
  Observe  $f_t$  and suffer cost  $f_t(x_t)$ 
  Set  $F_t := F_{t-1} + f_t = \sum_{i=1}^t f_i$ 
  Compute  $x_{t+1} \in \arg \min_{x \in \mathcal{X}} (F_t(x) + \frac{1}{\eta_t} R(x))$ 

```

The usual first step in the analyses of FTRL algorithms is to use basic properties of the iterates (without relying on convexity) to bound the algorithm’s regret by easier-to-analyse terms. Such bounds are usually the sum of two terms: the “diameter” of the feasible set through the lens of the FTRL regularizer and a sum of the difference in “quality” between consecutive iterates. For a classic example, see [Shalev-Shwartz, 2011, Lemma 2.3]. For our analysis we shall use a slightly tighter bound given by the Strong FTRL Lemma due to McMahan [2017]. For the sake of completeness we give a proof of this lemma (and discuss its applications in the composite setting) in Appendix C.1.

Lemma 3.1. (Strong FTRL Lemma [McMahan, 2017]) Let $\{f_t\}_{t \geq 1}$ be a sequence of functions such that $f_t: \mathcal{X} \rightarrow \mathbb{R}$ for each $t \geq 1$. Let $\{\eta_t\}_{t \geq 1}$ be a positive non-increasing sequence. Let $R: \mathcal{X} \rightarrow \mathbb{R}$ be such that $\{x_t\}_{t \geq 1}$ given as in Algorithm 1 is properly defined. If $F_t: \mathcal{X} \rightarrow \mathbb{R}$ is defined as in Algorithm 1 for each $t \geq 1$, then,

$$\text{Regret}_T(z) \leq \sum_{t=0}^T \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) (R(z) - R(x_t)) + \sum_{t=1}^T (H_t(x_t) - H_t(x_{t+1})) \quad \forall T > 0,$$

where $\eta_0 := 1$, $\frac{1}{\eta_{-1}} := 0$, $x_0 := x_1$, and $H_t := F_t + \frac{1}{\eta_t} R$ for each $t \geq 1$.

3.1 Sublinear Regret with Relative Lipschitz Functions

In the following theorem we formally state our sublinear $O(\sqrt{T})$ regret bound of FTRL in T rounds in the setting where the cost functions are Lipschitz continuous relative to the regularizer function used in the FTRL method. The proof, which we defer to Appendix C.2, boils down to bounding the terms $H_t(x_t) - H_t(x_{t+1})$ from the Strong FTRL Lemma by (roughly) $L^2 \eta_{t-1} / 2$. We do so by combining the optimality conditions from the definition of the iterates in Algorithm 1 with the L -Lipschitz continuity relative to R of the loss functions.

Theorem 3.2. Let $\{x_t\}_{t \geq 1}$ be defined as in Algorithm 1 and suppose f_t is L -Lipschitz continuous relative to R for all $t \geq 1$. Let $z \in \mathcal{X}$ and let $K \in \mathbb{R}$ be such that $K \geq R(z) - R(x_1)$. Then,

$$\text{Regret}_T(z) \leq \frac{K}{\eta_T} + \sum_{t=1}^T \frac{L^2 \eta_{t-1}}{2}, \quad \forall T > 0.$$

In particular, if $\eta_t := \sqrt{K} / (L\sqrt{t+1})$ for each $t \geq 0$, then $\text{Regret}_T(z) \leq 2L\sqrt{K(T+1)}$.

⁷The big-O notation in this case hides constants that may depend on the dimension and other properties of the problem at hand. The best dependence on the Lipschitz constant and “distance to the comparison point” is usually achieved when the loss functions are Lipschitz continuous and the FTRL regularizer is strongly convex, both with respect to the same norm.

3.2 Logarithmic Regret with Relative Strongly Convex Functions

Hazan et al. [2007b] showed that if the cost functions are not only Lipschitz continuous but strongly convex as well, then the *follow the leader* (FTL) method—FTRL without any regularizer—attains logarithmic regret. Similarly, in this section we show that if the cost functions are relative Lipschitz continuous and relative strongly convex, both relative to the same fixed function, then FTL suffers regret at most logarithmic in the number of rounds. The proof of the next theorem is similar to the proof of Theorem 3.2 and is deferred to Appendix C.3.

Theorem 3.3. Let $\{x_t\}_{t \geq 1}$ be defined as in Algorithm 1 with $R := 0$. Assume that f_t is L -Lipschitz continuous and M -strongly convex relative to a differentiable convex function $h: \mathcal{D} \rightarrow \mathbb{R}$ for each $t \geq 1$. Then, for all $z \in \mathcal{X}$,

$$\text{Regret}_T(z) \leq \frac{L^2}{2M}(\log(T) + 1), \quad \forall T > 0.$$

One might wonder whether requiring both Lipschitz continuity and strong convexity relative to the same function is too restrictive. Indeed, let f be both L -Lipschitz continuous and M -strongly convex relative to R . Moreover, assume f is differentiable for the sake of simplicity. If $x^* \in \mathcal{X}$ is a minimizer of f over \mathcal{X} , then optimality conditions imply $-\nabla f(x^*) \in N_{\mathcal{X}}(x^*)$. Thus, by the definition of relative strong convexity,

$$f(y) - f(x^*) \geq \langle \nabla f(x^*), y - x^* \rangle + MD_R(y, x^*) \geq MD_R(y, x^*), \quad \forall y \in \mathcal{X}$$

At the same time, by relative Lipschitz continuity (see (2.3)) we have

$$f(y) - f(x^*) \leq L\sqrt{2D_R(x^*, y)}, \quad \forall y \in \mathcal{X}.$$

This means that f has a $\Omega(D_R(\cdot, x^*))$ lower-bound and a $O(\sqrt{D_R(x^*, \cdot)})$ upper-bound. If the Bregman divergence between y and x^* were to go to infinity as y ranges over \mathcal{X} , for example when $\mathcal{X} = \mathbb{R}^n$ and R is the squared ℓ_2 norm, then the lower-bound would eventually exceed the upper-bound on \mathcal{X} . Therefore, relative Lipschitz continuity and relative strong convexity can only coexist when \mathcal{X} and the Bregman divergence with respect to R of a minimizer and any point in \mathcal{X} are both bounded. Although this is a somewhat restrictive condition, classical logarithmic regret results such as the ones due to Hazan et al. [2007b] also only hold over bounded sets. Moreover, as the next example shows, there are cases where logarithmic regret is attainable but *do not* fit into classical logarithmic regret results.

Example (A class functions that are both relative Lipschitz continuous and relative strongly convex). Define $f := \frac{1}{p}\|\cdot\|_2^p$ for some $p \geq 2$ and suppose $\mathcal{X} = [-\alpha, \alpha]^n$ for some $\alpha > 0$. First, note that $\nabla f(x) = \|x\|_2^{p-2}x$ and $\nabla^2 f(x) = \|x\|_2^{p-2}I + (p-2)\|x\|_2^{p-4}xx^T$ for any $x \in \mathbb{R}^n$. By Proposition 5.1 in Lu [2019], f is 1-continuous relative to $R := \frac{1}{2p}\|\cdot\|_2^{2p}$ on \mathbb{R}^n since $\|\nabla f(x)\|_2^2 = \|x\|_2^{2(p-2)} \cdot \|x\|_2^2 = \|x\|_2^{2p-2}$. Moreover, to show that f is M -strongly convex relative to R , it suffices to show that $f - MR$ is convex (see Proposition 1.1 in Lu et al. [2018]). For any $M > 0$ and $x \in \mathbb{R}^n$ we have

$$\begin{aligned} \nabla^2 f(x) - M\nabla^2 R(x) &= \|x\|_2^{p-2}I + (p-2)\|x\|_2^{p-4}xx^T - M(\|x\|_2^{2p-2}I + (2p-2)\|x\|_2^{2p-4}xx^T), \\ &= \|x\|_2^{p-2}(1 - M\|x\|_2^p)I + \|x\|_2^{p-4}(p-2 - M(2p-2)\|x\|_2^p)xx^T, \\ &\succeq \|x\|_2^{p-4}(1 - M\|x\|_2^p + p-2 - M(2p-2)\|x\|_2^p)xx^T, \\ &= \|x\|_2^{p-4}(p-1 - M(2p-1)\|x\|_2^p)xx^T, \end{aligned}$$

where the only inequality follows since $\|x\|_2^2 I \succeq xx^T$ for any $x \in \mathbb{R}^n$. By setting $M := \frac{p-1}{(2p-1)(\sqrt{n}\alpha)^p}$ we have

$$p-1 - M(2p-1)\|x\|_2^p \geq p-1 - M(2p-1)(\sqrt{n}\alpha)^p = 0, \quad \forall x \in \mathcal{X} = [-\alpha, \alpha]^n.$$

Thus, we have $\nabla^2 f(x) - M\nabla^2 R(x) \succeq 0$. Therefore, $f - MR$ is convex, which implies that f is strongly convex relative to R on \mathcal{X} . Note that f is not classically strongly convex (that is, strongly convex with respect to the ℓ_2 norm) for $p \geq 4$. To see this, note that $\nabla^2 f(x) - MI$ is not positive semidefinite around 0 for any $M > 0$, and thus $f - M\|\cdot\|_2^2$ is not convex around 0 no matter how small we pick $M > 0$ to be.

4 Dual Averaging and Composite Loss Functions

FTRL is a cornerstone algorithm in OCO, but sometimes it is not practical. Each iterate requires *exact* minimization of the loss functions (plus the regularizer) which might not have always a closed form solution. A notable special case of FTRL that mitigates this problem is the (online) *dual averaging* (DA) method whose offline version is due to Nesterov [2009]. In each iteration, DA picks a point from \mathcal{X} that minimizes the sum of past subgradients (scaled by the step size) plus a FTRL regularizer R . Formally, for real convex functions $\{f_t\}_{t \geq 1}$ on \mathcal{X} , the online DA method computes iterates $\{x_t\}_{t \geq 1}$ such that

$$x_{t+1} \in \arg \min_{x \in \mathcal{X}} \left(\eta_t \sum_{i=1}^t \langle g_i, x \rangle + R(x) \right) \quad \forall t \geq 0, \quad (4.1)$$

where $g_t \in \partial f_t(x_t)$ for each $t \geq 1$.

Intuition. It is well-known that the DA algorithm reduces to FTRL applied to the linearized functions $\{\tilde{f}_t\}_{t \geq 1}$ given by $\tilde{f}_t := \langle g_t, \cdot \rangle$ for each $t \in \mathbb{N}$ (for details see Hazan [2016, Lemma 5.4]). This reduction obviously preserves the property of being Lipschitz continuous since the gradient of \tilde{f}_t is g_t everywhere. A natural idea would be to use this same reduction in the relative setting. Unfortunately, this reduction does not preserve the property of being relative Lipschitz! Luckily, our proof only requires a weaker condition: being “relative Lipschitz” at the particular point x_t . Namely, the relative L -Lipschitzness (see (2.3)) of f_t implies $\langle \nabla \tilde{f}_t(x_t), x_t - y \rangle = \langle g_t, x_t - y \rangle \leq L \sqrt{2D_R(y, x_t)}$ for all $y \in \mathcal{X}$. That is all we need for the proof of Theorem 3.2 to go through, although we did state the theorem with this exact condition for the sake of simplicity. This discussion leads to the following corollary of Theorem 3.2.

Corollary 4.1. Let $\{x_t\}_{t \geq 1}$ be defined as in (4.1) and suppose f_t is L -Lipschitz continuous relative to R for all $t \geq 1$. Let $z \in \mathcal{X}$ and let $K \in \mathbb{R}$ be such that $K \geq R(z) - R(x_1)$. If $\eta_t := \sqrt{2K}/(L\sqrt{t+1})$ for all $t \geq 1$, then $\text{Regret}_T(z) \leq 2L\sqrt{K(T+1)}$.

Another important consideration for applications is a variant of OCO in which the loss functions are composite [Duchi et al., 2010, Xiao, 2010]. More specifically, in this case we have a known “extra regularizer” Ψ , a (not necessarily differentiable) convex function, and add it to the loss functions. The goal is to induce some kind of structure in the iterates, such as adding ℓ_1 -regularization to promote sparsity. Note that OCO algorithms would still apply in this setting by replacing the loss functions f_t with $f_t + \Psi$ at each round t . However, in this case we are not exploiting the fact that the function Ψ is *known*. In the case of the relative setting, for example, it may be the case that the loss functions f_t are relative Lipschitz-continuous with respect to a certain function R , while Ψ is not. In Appendix D we extend the sublinear (composite) regret bound of Theorem 3.2 and show how this yields convergence bounds for regularized dual averaging [Xiao, 2010] in the relative setting.

5 Dual-Stabilized Online Mirror Descent

The mirror descent algorithm is a generalization of the classical gradient descent method that was first proposed by Nemirovsky and Yudin [1983]. A modern treatment was first given by Beck and Teboulle [2003]. The algorithm fits almost seamlessly into the OCO setting via a variant known as online mirror descent (OMD) (see [Hazan, 2016]). Recently, Orabona and Pál [2018] showed that OMD with a dynamic learning rate may suffer *linear* regret. (A dynamic learning rate is useful when we do not know the number of iterations ahead of time.) Moreover, this can happen even in simple and well-studied scenarios such as in the problem of prediction with expert advice, which corresponds to OMD equipped with negative entropy as a mirror map. In general, they showed that this may happen in cases where the Bregman divergence (with respect to the mirror map chosen) is *not* bounded over the entire feasible set. To resolve this issue, Fang et al. [2020] proposed a modified version of OMD called *dual-stabilized online mirror descent* (DS-OMD). In contrast to classical OMD, the regret bounds for the dual-stabilized version depend only on the Bregman divergence between the feasible set and the *initial iterate*.

We formally describe the DS-OMD method in Algorithm 2. Compared to OMD, DS-OMD adds an extra step in the dual space to mix the current dual iterate with the dual of the initial point. This step

at iteration t is controlled by a stabilization parameter γ_t and it can be seen as a way to “stabilize” the algorithm in the dual space. Throughout this section we closely follow the notation and assumptions of Bubeck [2015, Chapter 4]. We assume that we have a **mirror map** for \mathcal{X} , that is, a differentiable strictly-convex function $\Phi: \mathcal{D} \rightarrow \mathbb{R}$ for \mathcal{X} such that the gradient of Φ diverges on the boundary of \mathcal{D} , that is, $\lim_{x \rightarrow \partial \mathcal{D}} \|\nabla \Phi(x)\|_2 = \infty$ where $\partial \mathcal{D} := \mathcal{D} \setminus \mathcal{D}^\circ$. These conditions on the mirror map guarantee that the algorithm is well-defined (for example, they guarantee the existence and uniqueness of the last step of Algorithm 2).

Algorithm 2 Dual-Stabilized Online Mirror Descent

Input: Stabilization coefficient γ_t and an initial iterate $x_1 \in \mathcal{X}$.

for $t = 1, 2, \dots$ **do**

 Observe f_t and suffer cost $f_t(x_t)$

 Compute $g_t \in \partial f_t(x_t)$

$\hat{x}_t := \nabla \Phi(x_t)$

$\hat{w}_{t+1} := \hat{x}_t - \eta_t g_t$

$\hat{y}_{t+1} := \gamma_t \hat{w}_{t+1} + (1 - \gamma_t) \hat{x}_t$

$y_{t+1} := \nabla \Phi^*(\hat{y}_{t+1})$

 Compute $x_{t+1} \in \arg \min_{x \in \mathcal{X}} D_\Phi(x, y_{t+1}) = \Phi(x) - \Phi(y_{t+1}) - \langle \nabla \Phi(y_{t+1}), x - y_{t+1} \rangle$

5.1 Sublinear Regret with Relative Lipschitz Functions

In this section, we give a regret bound for DS-OMD when the cost functions are all Lipschitz continuous relative to the mirror map Φ . In this setting, if we set the stabilization coefficients to be $\gamma_t := \eta_{t+1}/\eta_t$ and step size $O(1/\sqrt{t})$, DS-OMD obtains sublinear regret. This is formally stated in the following theorem.

Theorem 5.1. Let $\{x_t\}_{t \geq 1}$ be defined as in Algorithm 2 with $\gamma_t := \eta_{t+1}/\eta_t$ for each $t \geq 1$. Assume that f_t is L -Lipschitz continuous relative to Φ for all $t \geq 1$. Let $z \in \mathcal{X}$ and $K \in \mathbb{R}$ be such that $K \geq D_\Phi(z, x_1)$. Then,

$$\text{Regret}_T(z) \leq \frac{K}{\eta_{T+1}} + \sum_{t=1}^T \frac{\eta_t L^2}{2}, \quad \forall T > 0.$$

In particular, if $\eta_t := \sqrt{K}/L\sqrt{t}$ for each $t \geq 1$, then $\text{Regret}_T(z) \leq 2L\sqrt{K(T+1)}$.

The proof is based on Theorem E.3, which gives an abstract regret upper bound for DS-OMD. Next we compute specific upper bounds of $D_\Phi(x_t, w_{t+1})$ for each $t \geq 1$ by relative Lipschitz continuity to make the abstract regret bound more specific. The whole proof of Theorem 5.1 is given in Appendix E.1.

If we set each f_t to be a fixed function f and take average of all iterates, then we get the following convergence rate for classical convex optimization as a corollary.

Corollary 5.2. Let Φ be a mirror map for \mathcal{X} and let $f: \mathcal{X} \rightarrow \mathbb{R}$ be a convex L -Lipschitz-continuous function relative to Φ . Let $\{x_t\}_{t \geq 1}$ be given as in Algorithm 2 with loss functions $f_t := f$, step sizes $\eta_t := \sqrt{K}/L\sqrt{t}$ for some $K \geq \sup_{z \in \mathcal{X}} D_\Phi(z, x_1)$, and stabilization parameter $\gamma_t := \eta_{t+1}/\eta_t$. If $x^* \in \mathcal{X}$ is a minimizer of f , then,

$$f\left(\frac{1}{T} \sum_{t=1}^T x_t\right) - f(x^*) \leq \frac{2L\sqrt{2K}}{\sqrt{T}}.$$

This recovers the same bound up to constant $4\sqrt{2}/3$ in Theorem 4.3 in Lu [2019], if we take $k = T - 1$ and $t_i = \frac{\sqrt{K}}{\sqrt{TL}}$ for $i \geq 0$ therein.

5.2 Logarithmic Regret with Relative Strongly Convex Functions

In Section 3.2 we showed that FTRL suffers at most logarithmic regret when the loss functions are Lipschitz continuous and strongly convex, both relative to the same fixed reference function.

Similarly, we show that OMD suffers at most logarithmic regret if we have Lipschitz continuity and strong convexity, both relative to the mirror map Φ . Interestingly, in this case the dual-stabilization step can be skipped (that is, we can use $\gamma_t := 1$ for all t) and Algorithm 2 boils down to classic OMD.

Theorem 5.3. Let $\{x_t\}_{t \geq 1}$ be given as in Algorithm 2 with $\gamma_t := 1$ for all $t \geq 1$. Assume that f_t is L -Lipschitz continuous and M -strongly convex relative to Φ for all $t \geq 1$. If $z \in \mathcal{X}$ and $\eta_t = \frac{1}{tM}$ for each $t \geq 1$, then,

$$\text{Regret}_T(z) \leq \frac{L^2}{2M}(\log T + 1), \quad \forall T > 0.$$

The proof involves modifications of Theorem 5.1 and is deferred to Appendix E.2.

5.3 Sublinear Regret with Composite Loss Functions

We can extend our regret bounds to the setting with composite cost functions with minor modifications to Algorithm 2. The classical version OMD adapted to this setting is due to Duchi et al. [2010] and is known by composite objective mirror descent (COMID). They showed that COMID generalizes much prior work like forward-backward splitting and derived new results on efficient matrix optimization with Schatten p -norms based on this framework. Details of the modification needed on Algorithm E.3 in this setting together with regret bounds can be found in Appendix E.3.

6 Conclusions and Discussion

In this paper we showed regret bounds for both FTRL and stabilized OMD in the relative setting proposed by Lu [2019]. All the results hold in the *anytime setting* in which we do not know the number of rounds/iterations beforehand. Additionally, we gave logarithmic regret bounds for both algorithms when the functions are relatively strongly convex, analogous to the results known in the classical setting. Finally, we extend our results to the setting of composite cost functions, which is pervasive in practice. These results open up the possibility of a new range of applications for OCO algorithms and may allow for new analysis for known problems with better dependence on the instance’s parameters.

At the moment there are at least two interesting directions for future research. The first would be to investigate the connections among the different notions of relative smoothness, Lipschitz continuity, and strong convexity in the literature. Another is to investigate systematic ways of choosing a regularizer/mirror map for any given optimization problem. The latter was already an interesting questions before notions of relative Lipschitz continuity and strong convexity were proposed, but these new ideas give more flexibility in the choice of a regularizer.

7 Statement of Broader Impact

In this paper we study the performance of online convex optimization algorithms when the functions are not necessarily Lipschitz continuous, a requirement in classical regret bounds. This opens up the range of applications for which we can use OCO with good guarantees and guides how such parameters such as regularizers/mirror maps and step sizes should be chosen. It is our hope that this aids practitioners to develop more efficient ways to optimize and train their current models. Furthermore, we hope theoreticians to be inspired to delve deep into the setting of non-smooth optimization beyond Lipschitz continuity. It not only opens up the range of applications, but sheds light onto the fundamental conditions on the cost functions and regularizers/mirror maps needed for OCO algorithms to have good guarantees. Due to the theoretical nature of this work, we do not see potentially bad societal or ethical impacts.

Acknowledgments

We would like to thank the three anonymous reviewers and the meta-reviewer for engaging with our work. Moreover, we are thankful for their useful suggestions regarding the logarithmic regret results and the relationship of relative Lipschitz continuity and Riemann-Lipschitz continuity [Antonakopoulos et al., 2020]. We are also thankful to Wu Lin for useful discussions during the development of this work. Finally, we are grateful to Francesco Orabona for identifying in the work of Hazan et al. [2007a] some relationship with our results and one of the first uses of relative strong convexity.

Funding Disclosure

This research was partially supported by NSERC Discovery Grants, Canada Research Chairs, the CIFAR Learning in Machines and Brains program, and the Canada CIFAR AI Chair Program.

References

- K. Antonakopoulos, E. V. Belmega, and P. Mertikopoulos. Online and stochastic optimization beyond lipschitz continuity: A riemannian approach. In *8th International Conference on Learning Representations, ICLR*, 2020.
- H. H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.
- A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.
- A. Ben-Tal and A. Nemirovski. *Lectures on modern convex optimization*. MPS/SIAM Series on Optimization. Society for Industrial and Applied Mathematics (SIAM), 2001.
- J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd. First order methods beyond convexity and lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM Journal on Optimization*, 28(3):2131–2151, 2018.
- S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.
- J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *COLT 2010*, pages 14–26. Omnipress, 2010.
- J. C. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, 2011.
- H. Fang, N. J. A. Harvey, V. S. Portella, and M. P. Friedlander. Online mirror descent and dual averaging: keeping pace in the dynamic case. 2020. URL <https://arxiv.org/abs/2006.02585>.
- T. Gao, S. Lu, J. Liu, and C. Chu. Randomized bregman coordinate descent methods for non-lipschitz optimization. *arXiv preprint arXiv:2001.05202*, 2020.
- B. Grimmer. Convergence rates for deterministic and stochastic subgradient methods without lipschitz continuity. *SIAM Journal on Optimization*, 29(2):1350–1365, 2019.
- F. Hanzely and P. Richtárik. Fastest rates for stochastic mirror descent methods. 2018. URL <http://arxiv.org/abs/1803.07374>.
- E. Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016. URL <http://ocobook.cs.princeton.edu/OC0book.pdf>.
- E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Mach. Learn.*, 69(2-3):169–192, 2007a.
- E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007b.
- H. Lu. “Relative continuity” for non-lipschitz nonsmooth convex optimization using stochastic (or deterministic) mirror descent. *Informs Journal on Optimization*, pages 265–352, 2019.
- H. Lu, R. M. Freund, and Y. Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.
- C. J. Maddison, D. Paulin, Y. W. Teh, B. O’Donoghue, and A. Doucet. Hamiltonian descent methods. *arXiv preprint arXiv:1809.05042*, 2018.

- H. B. McMahan. A survey of algorithms and analysis for adaptive online learning. *The Journal of Machine Learning Research*, 18(1):3117–3166, 2017.
- M. C. Mukkamala and P. Ochs. Beyond alternating updates for matrix factorization with inertial bregman proximal gradient algorithms. In *Advances in Neural Information Processing Systems*, pages 4268–4278, 2019.
- A. S. Nemirovsky and D. B. Yudin. Problem complexity and method efficiency in optimization. 1983.
- Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.
- F. Orabona and D. Pál. Scale-free online learning. *Theoretical Computer Science*, 716:50–69, 2018.
- R. T. Rockafellar. *Convex analysis*. Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ, 1997. ISBN 0-691-01586-4. Reprint of the 1970 original, Princeton Paperbacks.
- S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2011.
- Q. Van Nguyen. Forward-backward splitting with bregman distances. *Vietnam Journal of Mathematics*, 45(3):519–539, 2017.
- L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(Oct):2543–2596, 2010.
- M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003)*, pages 928–936, 2003.

A Relationship with Riemann-Lipschitz Continuity

Antonakopoulos et al. [2020] introduced the idea of *Riemann-Lipschitz* continuity (RLC). They show how FTRL and OMD can be used when the cost functions are all RLC in a way that guarantees $O(\sqrt{T})$ regret. In this section we shall discuss the relationship between these two generalizations of Lipschitz continuity. Ultimately, we will see that our results are at least as general but that further study into the relationship between these ideas is needed. We note that we will closely follow the notation of Antonakopoulos et al. [2020] and shall not discuss Riemannian metrics in full generality.

Let $G: \mathbb{R}^n \rightarrow \mathbb{R}^{n \times n}$ be such that $G(x)$ is a symmetric positive definite matrix for all $x \in \mathcal{X} \setminus \{0\}$ and $G(0)$ is symmetric positive semidefinite. Then the **Riemannian metric** (induced by G) is the collection of bilinear pairings $\{\langle \cdot, \cdot \rangle_x : x \in \mathcal{X}\}$ defined by

$$\langle y, z \rangle_x := y^\top G(x)z, \quad \forall x, y, z \in \mathcal{X}.$$

For conciseness, we shall denote the above metric induced by G simply as the metric G . Moreover, the local norm induced by such the metric G on $x \in \mathcal{X}$ is naturally given by

$$\|z\|_x := \sqrt{\langle z, G(x)z \rangle}, \quad \forall z \in \mathcal{X}.$$

Let us now give the definition of Riemann-Lipschitz continuity.

Definition A.1. Let $L > 0$. A function $f: \mathcal{X} \rightarrow \mathbb{R}$ is *L-Riemann-Lipschitz continuous* (RLC) relative to a Riemannian metric G if

$$|f(y) - f(x)| \leq L \cdot \text{dist}_G(x, y) \quad \forall x, y \in \mathcal{X},$$

where $\text{dist}_G(x, y)$ is the Riemannian distance^{††} between x and y induced by the Riemannian metric G .

The above definition is notably hard to work with. In the case of differentiable functions, RLC boils down to a much simpler and more intuitive condition.

Proposition A.2 ([Antonakopoulos et al., 2020, Proposition 1]). Suppose that $f: \mathcal{X} \rightarrow \mathbb{R}$ is differentiable. Then f is L -RLC if and only if

$$\|\text{grad } f(x)\|_x \leq L \quad \text{for all } x \in \mathcal{X}, \tag{A.1}$$

where^{‡‡} $\text{grad } f(x) := G(x)^{-1} \nabla f(x)$ is the Riemannian gradient of f at x with respect to the metric G .

Finally, Antonakopoulos et al. [2020] use the notion of a *strong convexity* of a closed convex function $R: \mathcal{X} \rightarrow \mathbb{R}$ with respect to a metric G . For the sake of conciseness and simplicity, we shall use the equivalent condition given by Antonakopoulos et al. [2020, Lemma 1] and assume that R is differentiable, but the arguments of this section hold even if R is a closed convex function with a continuous selection of subgradients. More specifically, a differentiable convex function R is *K-strongly convex* with respect to the metric G for $K > 0$ if

$$\frac{K}{2} \|x - y\|_x^2 \leq D_R(y, x), \quad \forall x, y \in \mathcal{X}.$$

We are now in place to discuss the relationship between the notions of relative Lipschitz continuity and RLC. First, one should note that Proposition A.2 requires differentiability to hold. Since the regret bounds in Antonakopoulos et al. [2020] rely on (A.1), they also rely on the cost functions being differentiable. Since most $O(\sqrt{T})$ regret bounds in the online convex optimization literature (as well as the regret bounds in this text) *do not* rely on differentiability of the cost functions, it would be interesting to investigate if differentiability of the cost functions is in fact needed for the regret bounds of Antonakopoulos et al. [2020] to hold. In particular, in a way similar to classic Lipschitz

^{*}Equal contributions.

^{††}We do not give here the full definition of a Riemannian metric as given by Antonakopoulos et al. [2020] since it will not be used in any of our discussions.

^{‡‡}Here we overlook the case when $x = 0$ (and, thus, when $G(x)$ is not necessarily invertible), for the sake of simplicity.

continuity, it might be the case that (A.1) holds for at least one subgradient (after transformation by the metric G) at each point $x \in \mathcal{X}$ in the non-differentiable case.

Assuming that the cost functions are indeed differentiable, we can show that relative Lipschitz continuity is at least as general as RLC. In the following proposition we show that if f is a RLC function with respect to a metric G and if we have a differentiable convex function R which is strongly convex w.r.t. G (which is used as a regularizer or a mirror map in FTRL and OMD), then f is Lipschitz continuous relative to R .

Proposition A.3. Let $f: \mathcal{X} \rightarrow \mathbb{R}$ be a differentiable convex function and let $R: \mathcal{X} \rightarrow \mathbb{R}$ be a differentiable convex function such that R is K -strongly convex with respect to the Riemannian metric G . If f is L -RLC with respect to G , then f is L' -Lipschitz continuous relative to R where we set $L' := L\sqrt{K/2}$.

Proof. Let $x \in \mathcal{X}$. First, note that

$$\begin{aligned} \|\text{grad } f(x)\|_x^2 &= \text{grad } f(x)^\top G(x) \text{grad } f(x) = \nabla f(x)^\top G(x)^{-1} G(x) G(x)^{-1} \nabla f(x) \\ &= \nabla f(x)^\top G(x)^{-1} \nabla f(x) = \|\nabla f(x)\|_{x,*}^2, \end{aligned}$$

where $\|\cdot\|_{x,*}$ is the dual norm of $\|\cdot\|_x$. Therefore, for any $y \in \mathcal{X}$,

$$\begin{aligned} \nabla f(x)^\top (x - y) &\leq \|\nabla f(x)\|_{x,*} \|x - y\|_x && \text{(by the definition of dual norm),} \\ &\leq L \|x - y\|_x, && \text{(by RLC),} \\ &\leq L \sqrt{\frac{K}{2}} D_R(y, x), && \text{(by strong convexity of } R \text{ w.r.t. } G). \quad \square \end{aligned}$$

The above proposition shows that Riemann-Lipschitz continuity (together with a strongly convex function with respect to the Riemannian metric) implies relative Lipschitz continuity. Thus, our regret bounds can be seen as generalizations of the regret bounds due to Antonakopoulos et al. [2020]. Moreover, the modularity of our proofs makes it easier to extend the results to the different settings (as demonstrated to the extension of some regret bounds to the composite setting as shown in Section 4, for example).

Regarding the implication in the other direction, that is, whether relative Lipschitz continuity implies Riemannian Lipschitz continuity with respect to some metric G , it is not clear if it holds in general. The problem is that we do not know a systematic way of obtaining a metric G given a function f Lipschitz continuous relative to a function R such that f is RLC with respect to G and R is strongly convex with respect to G . Still, in some examples such a metric G does seem to exist. It is not clear at the moment if both concepts of Lipschitz continuity are equivalent or not.

B Arithmetic Inequalities

Lemma B.1. Let $\{a_t\}_{t \geq 1}$ be a non-negative sequence with $a_1 > 0$. Then,

$$\sum_{t=1}^T \frac{a_t}{\sqrt{\sum_{i=1}^t a_i}} \leq 2 \sqrt{\sum_{t=1}^T a_t}, \quad \forall T \in \mathbb{N}.$$

Proof. The proof is by induction on T . The statement holds trivially for $T = 1$. Let $T > 1$ and define $s := \sum_{t=1}^T a_t$. By the induction hypothesis,

$$\sum_{t=1}^T \frac{a_t}{\sqrt{\sum_{i=1}^t a_i}} \leq 2 \sqrt{\sum_{t=1}^{T-1} a_t} + \frac{a_T}{\sqrt{\sum_{i=1}^T a_i}} = 2\sqrt{s - a_T} + \frac{a_T}{\sqrt{s}}.$$

Finally, note that

$$\begin{aligned} 2\sqrt{s - a_T} + \frac{a_T}{\sqrt{s}} \leq 2\sqrt{s} &\iff 2\sqrt{s(s - a_T)} \leq 2s - a_T \iff 4s(s - a_T) \leq (2s - a_T)^2, \\ &\iff 4s^2 - 4sa_T \leq 4s^2 - 4sa_T + a_T^2 \iff 0 \leq a_T^2. \quad \square \end{aligned}$$

C Proofs for Section 3

C.1 Strong FTRL Lemma

In this section we give a proof of Lemma 3.1 for completeness. We also show how the lemma can be used for the composite setting. For further discussions on the lemma and on FTRL, see the thorough survey of McMahan [2017].

Proof of Lemma 3.1. Fix $T > 0$. Define $r_t := (\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}})R$ for each $t \geq 0$ (recall that $\eta_0 := 1$ and $1/\eta_{-1} := 0$), define $h_t := r_t + f_t$ for each $t \geq 1$, and set $h_0 := r_0$. In this way, we have

$$\sum_{i=0}^t h_t = \sum_{i=1}^t f_t + \sum_{i=0}^t r_t = \sum_{i=1}^t f_t + \frac{1}{\eta_t}R = H_t, \quad \forall t \geq 0.$$

In particular,

$$x_t \in \arg \min_{x \in \mathcal{X}} H_{t-1}(x) = \arg \min_{x \in \mathcal{X}} \sum_{i=0}^{t-1} h_i(x), \quad \forall t \geq 0. \quad (\text{C.1})$$

Let us now bound the regret of the points x_1, \dots, x_T with respect to the functions h_1, \dots, h_T and to a comparison point $z \in \mathcal{X}$ (plus a $-h_0(z)$ term):

$$\begin{aligned} \sum_{t=1}^T (h_t(x_t) - h_t(z)) - h_0(z) &= \sum_{t=1}^T h_t(x_t) - H_T(z) = \sum_{t=1}^T (H_t(x_t) - H_{t-1}(x_t)) - H_T(z), \\ &\stackrel{(\text{C.1})}{\leq} \sum_{t=1}^T (H_t(x_t) - H_{t-1}(x_t)) - H_T(x_{T+1}), \\ &= \sum_{t=1}^T (H_t(x_t) - H_t(x_{t+1})) - H_0(x_1), \end{aligned}$$

where in the last equation we just re-indexed the summation, placing $H_{T+1}(x_{T+1})$ inside the summation, and leaving $H_0(x_1)$ out. Re-arranging the terms and using $H_0 = h_0 = r_0$ and $x_0 = x_1$ yield

$$\begin{aligned} \sum_{t=1}^T (f_t(x_t) + r_t(x_t) - f_t(z) - r_t(z)) &= \sum_{t=1}^T (h_t(x_t) - h_t(z)), \\ &\leq r_0(z) - r_0(x_0) + \sum_{t=1}^T (H_t(x_t) - H_t(x_{t+1})), \end{aligned}$$

which implies

$$\text{Regret}_T(z) = \sum_{t=1}^T (f_t(x_t) - f_t(z)) \leq \sum_{t=0}^T (r_t(z) - r_t(x_t)) + \sum_{t=1}^T (H_t(x_t) - H_t(x_{t+1})).$$

Since $r_t = (\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}})R$ for all $t \geq 0$, we have

$$\sum_{t=0}^T (r_t(z) - r_t(x_t)) = \sum_{t=0}^T \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) (R(z) - R(x_t)). \quad \square$$

For the composite setting (see Section D), we modify the definition of r_t for $t \geq 1$ (maintaining the definition of r_0) in the above proof for

$$r_t := \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) R + \Psi, \quad \forall t \geq 1.$$

In this case, we have

$$H_t = \sum_{i=1}^t f_t + \sum_{i=0}^t r_t = \sum_{i=1}^t f_t + \frac{1}{\eta_t}R + t\Psi.$$

Proceeding in the same way as in the proof of Lemma 3.1, we get

$$\begin{aligned} \sum_{t=1}^T (f_t(x_t) - f(z)) &\leq \sum_{t=0}^T \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) (R(z) - R(x_t)), \\ &\quad + \sum_{t=1}^T (\Psi(z) - \Psi(x_t)) + \sum_{t=1}^T (H_t(x_t) - H_t(x_{t+1})), \end{aligned}$$

Re-arranging yields

$$\text{Regret}_T^\Psi(z) \leq \sum_{t=0}^T \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) (R(z) - R(x_t)) + \sum_{t=1}^T (H_t(x_t) - H_t(x_{t+1})). \quad (\text{C.2})$$

C.2 Sublinear Regret with Relative Lipschitz Functions

With the Strong FTRL Lemma, to derive regret bounds we can focus on bounding the difference in cost between consecutive iterates. In this section we will prove the sublinear regret bound for FTRL from Theorem 3.2. In the next lemma we give a bound on these costs based on the Bregman divergence of the FTRL regularizer, this time relying on convexity (but not on much more). Loosely saying, the first claim of the next lemma follows from the optimality conditions of the iterates of FTRL and the second follows from the subgradient inequality.

Lemma C.1. Let $\{x_t\}_{t \geq 1}$ and $\{F_t\}_{t \geq 0}$ be defined as in Algorithm 1. Then, for each $t \in \mathbb{N}$ there is $p_t \in N_{\mathcal{X}}(x_t)$ such that $-p_t - \frac{1}{\eta_{t-1}} \nabla R(x_t) \in \partial F_{t-1}(x_t)$, where $\eta_0 \in \mathbb{R}$ can be any positive constant. Moreover, this implies

$$F_{t-1}(x_t) - F_{t-1}(x_{t+1}) \leq \frac{1}{\eta_{t-1}} (R(x_{t+1}) - R(x_t) - D_R(x_{t+1}, x_t)).$$

Proof. Let $t \geq 1$. By the definition of the FTRL algorithm, we have $x_t \in \arg \min_{x \in \mathcal{X}} (F_{t-1}(x) + \frac{1}{\eta_{t-1}} R(x))$. By the optimality conditions for convex programs, we have

$$\partial \left(F_{t-1} + \frac{1}{\eta_{t-1}} R \right) (x_t) \cap (-N_{\mathcal{X}}(x_t)) \neq \emptyset.$$

Since $\partial \left(F_{t-1} + \frac{1}{\eta_{t-1}} R \right) (x_t) = \partial F_{t-1}(x_t) + \frac{1}{\eta_{t-1}} \nabla R(x_t)$, the above shows there is $p_t \in N_{\mathcal{X}}(x_t)$ such that

$$-p_t - \frac{1}{\eta_{t-1}} \nabla R(x_t) \in \partial F_{t-1}(x_t).$$

Using the subgradient inequality (2.1) with the above subgradient yields,

$$\begin{aligned} &F_{t-1}(x_t) - F_{t-1}(x_{t+1}) \\ &\leq -\langle p_t, x_t - x_{t+1} \rangle - \frac{1}{\eta_{t-1}} \langle \nabla R(x_t), x_t - x_{t+1} \rangle, \\ &\leq -\frac{1}{\eta_{t-1}} \langle \nabla R(x_t), x_t - x_{t+1} \rangle \quad (\text{by the definition of normal cone}), \\ &= \frac{1}{\eta_{t-1}} (R(x_{t+1}) - R(x_t) - D_R(x_{t+1}, x_t)), \end{aligned}$$

where in the last equation we used that, by definition of the Bregman divergence, $D_R(x_{t+1}, x_t) = R(x_{t+1}) - R(x_t) - \langle \nabla R(x_t), x_{t+1} - x_t \rangle$ and, thus, $-\langle \nabla R(x_t), x_t - x_{t+1} \rangle = R(x_{t+1}) - R(x_t) - D_R(x_{t+1}, x_t)$. \square

Proof of Theorem 3.2. For each $t \geq 0$ let H_t be defined as in the Strong FTRL Lemma and fix $t \geq 0$. We have

$$H_t(x_t) - H_t(x_{t+1}) = F_t(x_t) - F_t(x_{t+1}) + \frac{1}{\eta_t} (R(x_t) - R(x_{t+1})). \quad (\text{C.3})$$

Using $F_t = F_{t-1} + f_t$ together with Lemma C.1 we have

$$\begin{aligned} F_t(x_t) - F_t(x_{t+1}) &= F_{t-1}(x_t) - F_{t-1}(x_{t+1}) + f_t(x_t) - f_t(x_{t+1}), \\ &\leq \frac{1}{\eta_{t-1}} (R(x_{t+1}) - R(x_t) - D_R(x_{t+1}, x_t)) + f_t(x_t) - f_t(x_{t+1}). \end{aligned}$$

Plugging the above inequality onto (C.3) yields

$$(C.3) \leq f_t(x_t) - f_t(x_{t+1}) - \frac{D_R(x_{t+1}, x_t)}{\eta_{t-1}} + \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)(R(x_t) - R(x_{t+1})). \quad (C.4)$$

Since f_t is L -relative Lipschitz continuous with respect to R , we apply (2.3) followed by the arithmetic-geometric mean inequality $\sqrt{\alpha\beta} \leq (\alpha + \beta)/2$ with $\alpha := L^2\eta_{t-1}$ and $\beta := 2D_R(x_{t+1}, x_t)/\eta_{t-1}$ to get

$$f_t(x_t) - f_t(x_{t+1}) - \frac{D_R(x_{t+1}, x_t)}{\eta_{t-1}} \stackrel{(2.3)}{\leq} L\sqrt{2D_R(x_{t+1}, x_t)} - \frac{D_R(x_{t+1}, x_t)}{\eta_{t-1}} \leq \frac{L^2\eta_{t-1}}{2}.$$

Applying the above on (C.4) yields

$$(C.4) \leq \frac{L^2\eta_{t-1}}{2} + \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)(R(x_t) - R(x_{t+1})).$$

Plugging the above inequality into the the Strong FTRL Lemma together with $R(x_1) \leq R(x_t)$ for each $t \geq 1$ (which follows by the definition of x_1) yields

$$\begin{aligned} \text{Regret}_T(z) &\leq \sum_{t=0}^T \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)(R(z) - R(x_t) + R(x_t) - R(x_{t+1})) + \sum_{t=1}^T \frac{L^2\eta_{t-1}}{2}, \\ &= \sum_{t=0}^T \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)(R(z) - R(x_{t+1})) + \sum_{t=1}^T \frac{L^2\eta_{t-1}}{2}, \\ &\leq \frac{1}{\eta_T}(R(z) - R(x_1)) + \sum_{t=1}^T \frac{L^2\eta_{t-1}}{2} \leq \frac{K}{\eta_T} + \sum_{t=1}^T \frac{L^2\eta_{t-1}}{2}. \end{aligned}$$

If we set $\eta_t := \sqrt{2K}/(L\sqrt{t+1})$ and since $\sum_{t=1}^T \frac{1}{\sqrt{t}} \leq 2\sqrt{T}$ by Lemma B.1 in Appendix B, then

$$\text{Regret}_T(z) \leq L\sqrt{K(T+1)} + \frac{L\sqrt{K}}{2} \sum_{t=1}^T \frac{1}{\sqrt{t}} \leq L\sqrt{K(T+1)} + L\sqrt{KT} \leq 2L\sqrt{K(T+1)}. \quad \square$$

C.3 Logarithmic Regret

The next lemma strengthens the bound from Lemma C.1 in the case where the loss functions are relative strongly convex with respect to a fixed reference function. We further simplify matters by taking $R = 0$, that is, regularization is not needed for FTRL in the relative strongly convex case.

Lemma C.2. Let $\{x_t\}_{t \geq 1}$ be defined as in Algorithm 1 with $R := 0$. Moreover, let $h: \mathcal{D} \rightarrow \mathbb{R}$ be a differentiable convex function such that f_t is M -strongly convex relative to h for each $t \geq 1$. Then, for all $T \geq 1$,

$$F_{t-1}(x_t) - F_{t-1}(x_{t+1}) \leq -(t-1)MD_h(x_{t+1}, x_t).$$

Proof. Let $t \geq 1$. Note that F_{t-1} is $(t-1)M$ -strongly convex relative to R since it is the sum of $t-1$ functions that are each M -strongly convex relative to R . Additionally, let $p_t \in N_{\mathcal{X}}(x_t)$ be as given by Lemma C.1. By this lemma we have $-p_t \in \partial F_{t-1}(x_t)$. Thus, using inequality (2.4) from the definition of relative strong convexity with this subgradient yields

$$F_{t-1}(x_t) - F_{t-1}(x_{t+1}) \leq -\langle p_t, x_t - x_{t+1} \rangle - (t-1)MD_h(x_{t+1}, x_t).$$

By the definition of normal cone we have $-\langle p_t, x_t - x_{t+1} \rangle = \langle p_t, x_{t+1} - x_t \rangle \leq 0$, which yields the desired inequality. \square

Proof of Theorem 3.3. For each $t \geq 0$ let $H_t: \mathcal{X} \rightarrow \mathbb{R}$ be defined as in the Strong FTRL Lemma and fix $t \geq 0$. Since $R = 0$, we have $H_t = F_t$. This together with Lemma C.2 yields

$$\begin{aligned} H_t(x_t) - H_t(x_{t+1}) &= F_t(x_t) - F_t(x_{t+1}) = F_{t-1}(x_t) - F_{t-1}(x_{t+1}) + f_t(x_t) - f_t(x_{t+1}), \\ &\leq -(t-1)MD_h(x_{t+1}, x_t) + f_t(x_t) - f_t(x_{t+1}). \end{aligned} \quad (C.5)$$

Let $g_t \in \partial f_t(x_t)$. Since f_t is L -Lipschitz continuous and M -strongly convex, both relative to h , we have

$$f_t(x_t) - f_t(x_{t+1}) \stackrel{(2.4)}{\leq} \langle g_t, x_t - x_{t+1} \rangle - MD_h(x_{t+1}, x_t) \stackrel{(2.3)}{\leq} L\sqrt{2D_R(x_{t+1}, x_t)} - MD_R(x_{t+1}, x_t).$$

Applying the above to (C.5) together with the fact that $\sqrt{\alpha\beta} \leq (\alpha + \beta)/2$ with $\alpha := L^2/(Mt)$ and $\beta := 2tMD_R(x_{t+1}, x_t)$ yields

$$H_t(x_t) - H_t(x_{t+1}) \leq L\sqrt{2D_R(x_{t+1}, x_t)} - tMD_R(x_{t+1}, x_t) \leq \frac{L^2}{2Mt}.$$

Finally, plugging the above inequality into the Strong FTRL Lemma (with $R = 0$) gives

$$\text{Regret}_T(z) \leq \sum_{t=0}^T (H_t(x_t) - H_t(x_{t+1})) \leq \frac{L^2}{2M} \sum_{t=1}^T \frac{1}{t} \leq \frac{L^2}{2M} (\log(T) + 1). \quad \square$$

D Sublinear Regret Bounds for FTRL with Composite Loss Functions

In this section we extend the results from Section 3 to the case where the loss functions are *composite*. Specifically, there is a known non-negative convex function $\Psi: \mathcal{X} \rightarrow \mathbb{R}_+$ (sometimes called *extra regularizer*) which is subdifferentiable on \mathcal{X} and at round t the loss function presented to the player is $f_t + \Psi$. Usually Ψ is a simple function which is easy to optimize over (such as the ℓ_1 -norm). Thus, although $f_t + \Psi$ might not preserve relative Lipschitz continuity of f_t , one might still hope to obtain good regret bounds in this case. We shall see that FTRL does not need any modifications to enjoy of good theoretical guarantees in this setting. Yet, its analysis in the composite case will allow us to derive regret bounds for the *regularized dual averaging* method due to Xiao [2010].

In the composite case we measure the performance of an OCO algorithm by its **composite regret** (against a point $z \in \mathcal{X}$) given by

$$\text{Regret}_T^\Psi(z) := \sum_{t=1}^T (f_t(x_t) + \Psi(x_t)) - \inf_{z \in \mathcal{X}} \sum_{t=1}^T (f_t(z) + \Psi(z)), \quad \forall T > 0. \quad (\text{D.1})$$

In the case of FTRL, practically no modifications to the algorithm are needed. Namely, the update of Algorithm 1 becomes

$$x_{t+1} \in \arg \min_{x \in \mathcal{X}} \left(\sum_{i=1}^t f_i(x) + t\Psi(x) + \frac{1}{\eta_t} R(x) \right), \quad \forall t \geq 0.$$

We do make the additional assumption that $\Psi(x_1) = 0$, that is, x_1 minimizes Ψ and the latter has minimum value of 0. In practice one has some control on Ψ , so this assumption is not too restrictive. The next theorem shows that we can recover the regret bound from Theorem 3.2 for the composite setting even if Ψ is not relative Lipschitz-continuous with respect to the FTRL regularizer.

Theorem D.1. Let $\Psi: \mathcal{X} \rightarrow \mathbb{R}_+$ be a nonnegative convex function such that $\{x_t\}_{t \geq 1}$ as given as in Algorithm 1 are such that $\Psi(x_1) = 0$. Assume that f_t is L -Lipschitz continuous relative to R for all $t \geq 1$. Let $z \in \mathcal{X}$ and $K \in \mathbb{R}$ be such that $K \geq R(z) - R(x_1)$. Additionally, assume $\Psi(x_1) = 0$. Then,

$$\text{Regret}_T^\Psi(z) \leq \frac{2K}{\eta_T} + \sum_{t=1}^T \frac{L^2 \eta_{t-1}}{2}, \quad \forall T > 0.$$

In particular, if $\eta_t := \sqrt{2K}/(L\sqrt{t+1})$ for each $t \geq 1$, then $\text{Regret}_T^\Psi(z) \leq 2L\sqrt{K(T+1)}$

The proof is largely identical to the proof of Theorem 3.2. One of the main differences in the analysis is the following version of Lemma C.1 tweaked for the composite setting. It follows by adding $(t-1)\Psi$ to F_{t-1} in the proof of the original lemma and using the properties of the subgradient. We give the full proof for the sake of completeness.

Lemma D.2. Let $\Psi: \mathcal{X} \rightarrow \mathbb{R}_+$ be a nonnegative convex function such that $\{x_t\}_{t \geq 1}$ as given as in Algorithm 1 are such that $\Psi(x_1) = 0$. Then, for each $t \in \mathbb{N}$ there is $p_t \in N_{\mathcal{X}}(x_t)$ such that

$$-p_t - \frac{1}{\eta_{t-1}} \nabla R(x_t) \in \partial(F_{t-1} + (t-1)\Psi)(x_t),$$

and the above implies

$$\begin{aligned} & F_{t-1}(x_t) - F_{t-1}(x_{t+1}) + (t-1)(\Psi(x_t) - \Psi(x_{t+1})) \\ & \leq \frac{1}{\eta_{t-1}}(R(x_{t+1}) - R(x_t) - D_R(x_{t+1}, x_t))(t-1). \end{aligned}$$

Proof. Let $t \geq 1$. By the definition of the FTRL algorithm, we have $x_t \in \arg \min_{x \in \mathcal{X}} (F_{t-1}(x) + (t-1)\Psi(x) + \frac{1}{\eta_{t-1}}R(x))$. By the optimality conditions for convex programs, we have

$$\partial\left(F_{t-1} + (t-1)\Psi(x) + \frac{1}{\eta_{t-1}}R\right)(x_t) \cap (-N_{\mathcal{X}}(x_t)) \neq \emptyset.$$

Since $\partial(F_{t-1} + (t-1)\Psi(x) + \frac{1}{\eta_{t-1}}R)(x_t) = \partial(F_{t-1} + (t-1)\Psi(x))(x_t) + \frac{1}{\eta_{t-1}}\nabla R(x_t)$, the above shows there is $p_t \in N_{\mathcal{X}}(x_t)$ such that

$$-p_t - \frac{1}{\eta_{t-1}}\nabla R(x_t) \in \partial(F_{t-1} + (t-1)\Psi(x))(x_t).$$

Using the subgradient inequality (2.1) with the above subgradient yields,

$$\begin{aligned} & F_{t-1}(x_t) + (t-1)\Psi(x_t) - F_{t-1}(x_{t+1}) - (t-1)\Psi(x_{t+1}) \\ & \leq -\langle p_t, x_t - x_{t+1} \rangle - \frac{1}{\eta_{t-1}}\langle \nabla R(x_t), x_t - x_{t+1} \rangle, \\ & \leq -\frac{1}{\eta_{t-1}}\langle \nabla R(x_t), x_t - x_{t+1} \rangle \quad \text{(by the definition of normal cone),} \\ & = \frac{1}{\eta_{t-1}}(R(x_{t+1}) - R(x_t) - D_R(x_{t+1}, x_t)), \end{aligned}$$

where in the last equation we used that, by definition of the Bregman divergence, $D_R(x_{t+1}, x_t) = R(x_{t+1}) - R(x_t) - \langle \nabla R(x_t), x_{t+1} - x_t \rangle$ and, thus, $-\langle \nabla R(x_t), x_t - x_{t+1} \rangle = R(x_{t+1}) - R(x_t) - D_R(x_{t+1}, x_t)$. \square

Now we are in position to prove Theorem D.1.

Proof of Theorem D.1. We proceed in a way extremely similar to the proof of Theorem 3.2, but in place of the standard FTRL Lemma we use its composite version as in (C.2).

For each $t \geq 0$ let H_t be define as in the (composite) Strong FTRL Lemma so that $H_t = \sum_{i=1}^t f_i + t\Psi + \frac{1}{\eta_t}R$ and fix $t \geq 0$. In this case we have

$$H_t(x_t) - H_t(x_{t+1}) = F_t(x_t) - F_t(x_{t+1}) + t(\Psi(x_t) - \Psi(x_{t+1})) + \frac{1}{\eta_t}(R(x_t) - R(x_{t+1})).$$

Using $F_t = F_{t-1} + f_t$ together with Lemma D.2 we have

$$\begin{aligned} & F_t(x_t) - F_t(x_{t+1}) + t(\Psi(x_t) - \Psi(x_{t+1})) \\ & \leq \frac{1}{\eta_{t-1}}(R(x_{t+1}) - R(x_t) - D_R(x_{t+1}, x_t)) + f_t(x_t) - f_t(x_{t+1}) + \Psi(x_t) - \Psi(x_{t+1}). \end{aligned}$$

Proceeding as in the proof of Theorem 3.2 (with the addition of a $\Psi(x_t) - \Psi(x_{t+1})$ term) we have

$$H_t(x_t) - H_t(x_{t+1}) \leq \frac{L^2\eta_{t-1}}{2} + \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)(R(x_t) - R(x_{t+1})) + \Psi(x_t) - \Psi(x_{t+1}).$$

When summing over $t \in \{1, \dots, T\}$, the terms $\Psi(x_t) - \Psi(x_{t+1})$ telescope so that, since x_1 minimizes Ψ , we have

$$\sum_{t=1}^T (\Psi(x_t) - \Psi(x_{t+1})) = \Psi(x_1) - \Psi(x_{T+1}) \leq 0.$$

Therefore, the remainder of the proof follows as in the proof of Theorem 3.2. \square

D.1 Regularized Dual Averaging

As previously discussed, applying OCO algorithms such as dual averaging in an out-of-the-box fashion when the loss functions are composite case does not exploit the structure of the extra-regularization given by Ψ and may have poor performance in practice. For example, McMahan [2017] shows that applying DA in the composite case with $\Psi := \|\cdot\|_1$ does not yield sparse solutions. Xiao [2010] proposed the *regularized dual averaging* (RDA) method to solve this issue. The algorithm is identical to DA but it *does not linearize* the function Ψ . Formally, the initial iterate x_1 is in $\arg \min_{x \in \mathcal{X}} (R(x))$ and is such that $\Psi(x_1) = 0$, that is, x_1 minimizes Ψ . For the following rounds, RDA computes

$$x_{t+1} \in \arg \min_{x \in \mathcal{X}} \left(\sum_{i=1}^t \langle g_i, x \rangle + t\Psi(x) + \frac{1}{\eta_t} R(x) \right) \quad \forall t \geq 1. \quad (\text{D.2})$$

With an argument analogous to the one made in Section 4, we can write RDA as an instance of FTRL (with composite loss functions) and obtain the following corollary of Theorem D.1.

Corollary D.3. Let $\Psi: \mathbb{R}^n \rightarrow \mathbb{R}_+$ be a nonnegative convex function. Let $\{x_t\}_{t \geq 1}$ be defined as in (D.2) and assume $\Psi(x_1) = 0$. Moreover, suppose f_t is L -Lipschitz continuous relative to R for all $t \geq 1$. Let $z \in \mathcal{X}$ and let $K \in \mathbb{R}$ be such that $K \geq R(z) - R(x_1)$. If $\eta_t := \sqrt{2K}/(L\sqrt{t+1})$ for all $t \geq 1$, then $\text{Regret}_T^\Psi(z) \leq 2L\sqrt{K(T+1)}$.

E Proofs for Section 5

In this section we give the missing proofs of Section 5. Throughout this section, let $\{x_t\}_{t \geq 1}$ and $\{\hat{w}_t\}_{t \geq 1}$ be defined as in Algorithm 2, and define

$$w_t := \nabla \Phi^*(\hat{w}_t), \quad \forall t \geq 1.$$

First, let us state inequality (4.9) and Claim 4.2 (without substituting exactly value of γ_t) from Fang et al. [2020] at the beginning, which will appear multiple times throughout this section, respectively as:

Claim E.1. If $\gamma_t = \eta_{t+1}/\eta_t \in (0, 1]$ for each $t \geq 1$, then

$$f_t(x_t) - f_t(z) \leq \frac{1}{\eta_t} (D_\Phi(x_t, w_{t+1}) - D_\Phi(z, w_{t+1}) + D_\Phi(z, x_t)).$$

Claim E.2. If $\gamma_t \in (0, 1]$ for all $t \geq 1$, then,

$$\begin{aligned} & \frac{1}{\eta_t} (D_\Phi(x_t, w_{t+1}) - D_\Phi(z, w_{t+1}) + D_\Phi(z, x_t)) \\ & \leq \frac{D_\Phi(x_t, w_{t+1})}{\eta_t} + \frac{1}{\eta_t} \left(\left(\frac{1}{\gamma_t} - 1 \right) D_\Phi(z, x_1) - \frac{1}{\gamma_t} D_\Phi(z, x_{t+1}) + D_\Phi(z, x_t) \right). \end{aligned}$$

E.1 Sublinear Regret for Relative Lipschitz Functions

In this subsection we prove sublinear regret for DS-OMD with relative Lipschitz continuous cost functions. First we use Theorem 4.1 in Fang et al. [2020]. This theorem is analogous to the bound given in the analysis of classic OMD given by Bubeck [2015, Theorem 4.2].

Theorem E.3 (Fang et al. [2020, Theorem 4.1]). If $\gamma_t := \eta_{t+1}/\eta_t$ for each $t \geq 1$, then

$$\text{Regret}_T(z) \leq \sum_{t=1}^T \frac{D_\Phi(x_t, w_{t+1})}{\eta_t} + \frac{D_\Phi(z, x_1)}{\eta_{T+1}}, \quad \forall T > 0.$$

Now we are ready to use Theorem E.3 to prove Theorem 5.1.

Proof of Theorem 5.1. We first need to bound the terms $D_\Phi(x_t, w_{t+1})$ for each $t \geq 1$. Fix $t \geq 1$. By the three-point identity for Bregman divergences (see (2.2)),

$$D_\Phi(x_t, w_{t+1}) = -D_\Phi(w_{t+1}, x_t) + \langle \nabla \Phi(x_t) - \nabla \Phi(w_{t+1}), x_t - w_{t+1} \rangle. \quad (\text{E.1})$$

From the definition of the iterates in Algorithm 2, we have $\eta_t g_t = \nabla \Phi(x_t) - \nabla \Phi(w_{t+1})$. Thus,

$$(E.1) = -D_\Phi(w_{t+1}, x_t) + \eta_t \langle g_t, x_t - w_{t+1} \rangle, \\ \stackrel{(2.3)}{\leq} -D_\Phi(w_{t+1}, x_t) + \eta_t L \sqrt{2D_\Phi(w_{t+1}, x_t)} \leq \frac{\eta_t^2 L^2}{2}, \quad (E.2)$$

where first inequality is from (2.3) (since f_t is Lipschitz continuous relative to Φ) and the second inequality comes from the fact that $\sqrt{\alpha\beta} \leq (\alpha + \beta)/2$ with $\alpha := \eta_t^2 L^2$ and $\beta := D_\Phi(w_{t+1}, x_t)$. Plugging the above in Theorem E.3, we get

$$\text{Regret}_T(z) \leq \sum_{t=1}^T \frac{\eta_t L^2}{2} + \frac{D_\Phi(z, x_1)}{\eta_{T+1}} \leq \sum_{t=1}^T \frac{\eta_t L^2}{2} + \frac{K}{\eta_{T+1}}.$$

Setting $\eta_t := \sqrt{K}/L\sqrt{t}$ for each $t \geq 1$ and by using Lemma B.1 from Appendix B we have

$$\text{Regret}_T(z) \leq \frac{L^2}{2} \cdot \frac{\sqrt{K}2\sqrt{T}}{L} + K \frac{L\sqrt{T+1}}{\sqrt{K}} \leq 2L\sqrt{K(T+1)}. \quad \square$$

E.2 Proof for Theorem 5.3

In this section we give a logarithmic regret bound for OMD the cost functions are when relative Lipschitz continuous and relative strongly convex, both relative to the mirror map. The first step in the proof is the following claim given by modifying Claims E.1 and E.2 and combining them together.

Claim E.4. Assume that $\gamma_t = 1$ for all $t \geq 1$, then

$$f_t(x_t) - f_t(z) \leq \frac{1}{\eta_t} (D_\Phi(x_t, w_{t+1}) - D_\Phi(z, w_{t+1}) + D_\Phi(z, x_t)) - MD_\Phi(z, x_t).$$

Proof of Claim E.4. This proof largely follows the structure of the proof of Claim E.1. First, instead of using subgradient inequality, we use the definition of relative strong convexity and get

$$f_t(x_t) - f_t(z) \leq \langle g_t, x_t - z \rangle - MD_\Phi(z, x_t).$$

By proceeding as in the proof of Claim E.1 but adding the extra term $-MD_\Phi(z, x_t)$ term we get

$$f_t(x_t) - f_t(z) \leq \frac{1}{\eta_t} (D_\Phi(x_t, w_{t+1}) - D_\Phi(z, w_{t+1}) + D_\Phi(z, x_t)) - MD_\Phi(z, x_t).$$

Then we apply Claim E.2 with $\gamma_t = 1$ to get the desired inequality. \square

The next step in the proof of the logarithmic regret bound is to sum Claim E.4 over t , yielding

$$\begin{aligned} & \sum_{t=1}^T (f_t(x_t) - f_t(z)) \\ & \leq \sum_{t=1}^T \frac{D_\Phi(x_t, w_{t+1})}{\eta_t} + \sum_{t=2}^T \left(\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) D_\Phi(z, x_t) - MD_\Phi(z, x_t) \right) \\ & \quad + \frac{1}{\eta_1} D_\Phi(z, x_1) - \frac{1}{\eta_T} D_\Phi(z, x_{T+1}) - MD_\Phi(z, x_1), \quad (\text{by Claim E.4}) \\ & \leq \sum_{t=1}^T \frac{D_\Phi(x_t, w_{t+1})}{\eta_t} + \sum_{t=2}^T \left(\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) D_\Phi(z, x_t) - MD_\Phi(z, x_t) \right). \quad (\eta_1 = 1/M) \end{aligned}$$

Since $\eta_t = \frac{1}{Mt}$, we have

$$\sum_{t=2}^T \left(\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) D_\Phi(z, x_t) - MD_\Phi(z, x_t) \right) = \sum_{i=2}^T \left(MD_\Phi(z, x_i) - MD_\Phi(z, x_i) \right) = 0.$$

We have already shown that $D_\Phi(x_t, w_{t+1}) \leq \frac{\eta_t^2 L^2}{2}$ in (E.2), so

$$\begin{aligned} \text{Regret}_T(z) &\leq \sum_{t=1}^T \frac{D_\Phi(x_t, w_{t+1})}{\eta_t} + \sum_{i=2}^T \left(\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) D_\Phi(z, x_t) - M D_\Phi(z, x_t) \right), \\ &\leq \sum_{t=1}^T \frac{\eta_t L^2}{2} = \frac{L^2}{2M} \sum_{t=1}^T \frac{1}{t} \leq \frac{L^2}{2M} (\log T + 1). \end{aligned}$$

The last step comes from upper bound of the harmonic series.

E.3 Sublinear Regret for DS-OMD with Extra Regularization

Following the notation from Appendix D, we let $\Psi: \mathcal{X} \rightarrow \mathbb{R}_+$ denote the extra regularizer, a non-negative convex function. We also assume Ψ is minimized at x_1 with value 0 and use composite regret to measure the performance. The only modification we need to make to Algorithm 2 is to change the projection step of the algorithm to

$$x_{t+1} = \arg \min_{x \in \mathbb{R}^n} (D_\Phi(x, y_{t+1}) + \eta_{t+1} \Psi(x)). \quad (\text{E.3})$$

Here we minimize over \mathbb{R}^n instead of over \mathcal{X} since we can introduce the constraint of the points lying in \mathcal{X} by adding to Ψ the indicator function of \mathcal{X} . That is, by adding to Ψ the function

$$\delta_{\mathcal{X}}(x) := \begin{cases} 0 & \text{if } x \in \mathcal{X}, \\ +\infty & \text{otherwise,} \end{cases} \quad \forall x \in \mathbb{R}^n.$$

In the remainder of this section we denote by $\Pi_{\eta_{t+1} \Psi}^\Phi(y_{t+1})$ the point computed by the right-hand side of (E.3). If we pick this projection coefficient α_t carefully, we can get $O(\sqrt{T})$ regret, as specified by the next theorem.

Theorem E.5. Let $\{x_t\}_{t \geq 1}$ be given as in Algorithm 2 with composite updates and with parameters $\gamma_t := \eta_{t+1}/\eta_t$ for each $t \geq 1$. Assume that $\Psi(x_1) = 0$ and that f_t is L -Lipschitz continuous relative to Φ for all $t \geq 1$. Let $z \in \mathcal{X}$ and $K \in \mathbb{R}$ be such that $K \geq D_\Phi(z, x_1)$. Then,

$$\text{Regret}_T^\Psi(z) \leq \sum_{t=1}^T \frac{\eta_t L^2}{2} + \frac{K}{\eta_{T+1}}, \quad \forall z \in \mathcal{X}, \forall T > 0.$$

In particular, for $\eta_t := \sqrt{K}/L\sqrt{t}$ for each $t \geq 1$, then $\text{Regret}_T^\Psi(z) \leq 2L\sqrt{K(T+1)}$.

The analysis hinges on the following generalization of [Bubeck, 2015, Lemma 4.1], which can be thought as a ‘‘pythagorean Theorem’’ for Bregman projections.

Lemma E.6. Let $x \in \mathbb{R}^n$, $y \in \mathcal{D}^\circ$, and set $\bar{y} := \Pi_{\alpha_t \Psi}^\Phi(y)$. If $\bar{y} \in \mathcal{D}^\circ$, then

$$D_\Phi(x, \bar{y}) + D_\Phi(\bar{y}, y) \leq D_\Phi(x, y) + \alpha_t (\Psi(x) - \Psi(\bar{y})).$$

Proof of Lemma E.6. By the optimality conditions of the projection, we have $\nabla \Phi(y) - \nabla \Phi(\bar{y}) \in \partial(\alpha_t \Psi)(\bar{y})$. Using the three-point identity of Bregman divergences (see (2.2)) and the subgradient inequality, we get

$$D_\Phi(x, \bar{y}) + D_\Phi(\bar{y}, y) - D_\Phi(x, y) = \langle \nabla \Phi(y) - \nabla \Phi(\bar{y}), x - \bar{y} \rangle \leq \alpha_t (\Psi(x) - \Psi(\bar{y})).$$

Rearranging yields the desired inequality. \square

We are now ready to prove Theorem E.5.

Proof of Theorem E.5. To prove the theorem, we just need to show that Theorem E.3 still holds (with respect to the composite regret) in the algorithm with composite projections. We modify Claims E.1 and E.2 to get the following claim.

Claim E.7.

$$\begin{aligned} &f_t(x_t) - f_t(z) \\ &\leq \frac{D_\Phi(x_t, w_{t+1})}{\eta_t} + \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) D_\Phi(z, x_1) + \frac{D_\Phi(z, x_t)}{\eta_t} - \frac{D_\Phi(z, x_{t+1})}{\eta_{t+1}} + (\Psi(z) - \Psi(x_{t+1})). \end{aligned}$$

Proof of Claim E.7. Claim E.1 gives us the following inequality:

$$f_t(x_t) - f_t(z) \leq \frac{1}{\eta_t}(D_\Phi(x_t, w_{t+1}) - D_\Phi(z, w_{t+1}) + D_\Phi(z, x_t)).$$

Then we just need to modify Claim E.2 to bound the right side of the above inequality. Using Lemma E.6, we have

$$D_\Phi(z, y_{t+1}) - D_\Phi(x_{t+1}, y_{t+1}) \geq D_\Phi(z, x_{t+1}) + \alpha_t(\Psi(x_{t+1}) - \Psi(z)).$$

Then we substitute the step $D_\Phi(z, y_{t+1}) - D_\Phi(x_{t+1}, y_{t+1}) \geq D_\Phi(z, x_{t+1})$ in the original proof of Claim E.2 in Fang et al. [2020] with the above inequality plus the extra regularization term and Claim E.7 follows. \square

Now the regret is bounded by

$$\begin{aligned} & \text{Regret}_T^\Psi(z) \\ &= \sum_{t=1}^T \left(f_t(x_t) + \Psi(x_t) - f_t(z) - \Psi(z) \right), \\ &= \sum_{t=1}^T \left(\left(f_t(x_t) - f_t(z) \right) + \left(\Psi(x_t) - \Psi(z) \right) \right), \\ &\leq \sum_{t=1}^T \frac{D_\Phi(x_t, w_{t+1})}{\eta_t} + \sup_{z \in \mathcal{X}} \frac{D_\Phi(z, x_1)}{\eta_{T+1}} + \sum_{t=1}^T (\Psi(x_t) - \Psi(x_{t+1})), \\ &= \sum_{t=1}^T \frac{D_\Phi(x_t, w_{t+1})}{\eta_t} + \sup_{z \in \mathcal{X}} \frac{D_\Phi(z, x_1)}{\eta_{T+1}} + \Psi(x_1) - \Psi(x_{T+1}), \\ &\leq \sum_{t=1}^T \frac{D_\Phi(x_t, w_{t+1})}{\eta_t} + \sup_{z \in \mathcal{X}} \frac{D_\Phi(z, x_1)}{\eta_{T+1}}. \end{aligned}$$

The first inequality follows Claim E.7 and the last step comes from the assumption that x_1 is the minimizer of Ψ . This shows Theorem E.3 holds as desired and then the proof of Theorem E.5 follows as in Appendix E.1. \square

Similarly, by setting all f_t to a fixed function f and taking average we get the following corollary.

Corollary E.8. Consider a convex function f and let x^* be a minimizer of f . Let Φ be a differentiable strictly convex mirror map such that $\mathcal{X} \subseteq \mathcal{D}^\circ$. Assume that f is L -Lipschitz continuous to Φ and there exists non-negative K such that $K \geq D_\Phi(x^*, x_1)$. Let $\{\eta_t\}_{t \geq 1}$ be a sequence of step sizes. If we pick step size $\eta_t = \frac{1}{\sqrt{t}}$, $\alpha_t = \eta_{t+1}$ and stabilization coefficient $\gamma_t = \eta_{t+1}/\eta_t$, then we have convergence rate

$$(f + \Psi) \left(\frac{1}{T} \sum_{t=1}^T x_t \right) - (f + \Psi)(x^*) \leq \frac{2L\sqrt{2K}}{\sqrt{T}}.$$