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

by

Yihan Zhou

B. Mathematics, University of Waterloo, 2018

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Computer Science)

The University of British Columbia
(Vancouver)

August 2020

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

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

submitted by **Yihan Zhou** in partial fulfillment of the requirements for the degree of **Master of Science** in **Computer Science**.

Examining Committee:

Nicholas J. A. Harvey, Computer Science
Supervisor

Mark Schmidt, Computer Science
Supervisory Committee Member

Abstract

Online convex optimization (OCO) is a powerful algorithmic framework that has extensive applications in different areas. Regret is a commonly-used measurement for the performance of algorithms in this framework. Lipschitz continuity of the cost functions is commonly assumed in order to obtain sublinear regret, that is to say, this condition is usually necessary for theoretical guarantees for good performances of OCO algorithms. Moreover, strong convexity of cost functions can sometimes give even better theoretical performance bounds, more specifically, logarithmic regret. 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 extend the results to algorithms with extra regularization such as regularized dual averaging.

Lay Summary

Machine learning involves processing data as a core step. In real life, data sometimes comes in one by one at each time not altogether. This means we need to process data one piece by one piece at each step and *online learning* studies this scenario. Due to its application potential, online learning has gained considerable popularity. Numerous algorithms have been designed to solve online learning problems in practice.

At the same time, along with applications, online learning theory has been developed. Such theory gives us more insights into understanding existing algorithms and designing new ones. It also gives performance guarantees, which can help us facilitate and choose algorithms. In this work we relax a classical presumption to extend the scope of existing theory. This could shed news lights on new directions in theory research and help invent novel algorithms.

Preface

This thesis is based on the homonymous unpublished work done with Victor Sanches Portella, Mark Schmidt and Nicholas J. A. Harvey. Chapter 3 and 4 consist of novel theoretical results. I contributed in the proof and analysis in these two chapters and wrote most of the first draft of the unpublished paper.

Table of Contents

Abstract . . . . **iii**

Lay Summary . . . . **iv**

Preface . . . . **v**

Table of Contents . . . . **vi**

Glossary . . . . **viii**

Acknowledgments . . . . **x**

致谢 . . . . **xi**

1 Introduction **1**

1.1 Lipschitzness and Strong Convexity 2

1.2 Follow The Regularized Leader 5

1.3 Mirror Descent 8

2 Related Work **11**

2.1 A New Descent Lemma 12

2.2 Relative Lipschitzness 14

2.3 Other Related Work 15

3 Follow the Regularized Leader **17**

3.1 Sublinear Regret with Relative Lipschitz Functions 19

3.1.1 Dual Averaging 22
3.2 Logarithmic Regret with Relative Strongly Convex Functions . . . 23
3.3 Sublinear Regret with Composite Loss Functions 26
3.3.1 Regularized Dual Averaging 29

4 Dual Stabilized-Online Mirror Descent 31
4.1 Sublinear Regret with Relative Lipschitz Functions 33
4.2 Logarithmic Regret with Relative Strongly Convex Functions . . . 34
4.3 Sublinear Regret for DS-OMD with Extra Regularization 36

5 Conclusion and Future Work 40
5.1 Conslusion 40
5.2 Future Work 40

Bibliography 42

A Supporting Materials 45
A.1 Arithmetic Inequalities 45

Glossary

DA dual averaging

DS-OMD dual-stabilized online mirror descent

FTL follow the leader

FTRL follow the regularized leader

OCO online convex optimization

OLO online linear optimization

OMD online mirror descent

PG proximal gradient descent

RDA regularized dual averaging

RLC Riemannian Lipschitz-continuity

SGD stochastic gradient descent

SVM support vector machine

Acknowledgments

I would like to thank two of my supervisors, Mark and Nick. They are both excellent researchers and superb lecturers. They have provided invaluable guidance on research to me during my Master period. They have given me valuable lessons outside academia as well. I am really fortunate to be supervised by them.

I would like to thank Aaron, Wilder, Cathy, Fred, Ben, Issam, Lironne, Michael, Alireza, Wu, Chris, Sikander and Victor. We share one supervisor and I view you as academic siblings. You inspired and motivated my research. I also want to thank my friends Devon, Jason, Taylor, David, Hu, Yuxi, Rui, Hieu, Yuchen, Garry and John-Jose for their caring and help. I would like to give a special acknowledgement to Victor for his contribution in the work leading to this thesis.

I am deeply grateful to my parents for their education and unconditional love for me. I am greatly indebted in Doug and Sharon for teaching me how to take care of myself. They are like my families in Canada.

致谢

感谢爸爸妈妈和其他家庭成员对我的关心和教导。你们对我的爱是我前进的动力。

感谢好友施昊辰一直以来陪伴在我身边，和你成为朋友是我一生的幸运。

感谢猫、狗和海豚、虎鲸等海洋哺乳动物。你们很可爱，你们让地球更加美丽。

此毕业论文完成于全球政治局势动荡时期，全球各地各种抗议活动风起云涌。感谢所有为反威权政府、反种族歧视、反性别歧视做出贡献的机构和个人，有你们这个世界更美好。感谢现实中和网络上志同道合的朋友，我们在这个糟糕的时代里抱团取暖，共同度过这个艰难的时刻。

Chapter 1

Introduction

An essential part of machine learning and artificial intelligence is data processing. Data comes in two forms: batch and stream. A batch of data is given at once at the beginning while a stream of data comes piece by piece. For example, a dataset is a batch of data and a video can be viewed as a stream of pictures. Online optimization is the field studying the case when data comes in a stream. This thesis will focuses on online optimization theory, and more specifically online convex optimization theory.

Online optimization has gained increasing interest among the machine learning community. There are two main reasons behind it. First, in many applications, data naturally come in streams. The epitome of this is real-time application. If we want to train a real-time predictor, like a predictor forecasting stock price or a detector capturing pedestrian in surveillance camera without a preprocessed dataset, we have to get data in real time. Such real-time data comes in streams and we cannot get future data in advance. Another reason that data may come in a stream is that sometimes it is computationally expensive or inefficient to process the whole dataset at one time, so we prefer to process data one by one. For this reason, in large scale machine learning, online optimization is prevalent.

We start with a brief introduction to the online convex optimization (OCO) framework. In online convex optimization, 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. For example, in online advertising, every time the algorithm pushes an advertisement to the audience and receives feedback. More precisely, in each round $t \in \mathbb{N}$, the player picks a point x_t from a fixed convex set $\mathscr{X} \subseteq \mathbb{R}^n$ and an adversary picks a convex function f_t possibly 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 [24], 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 [12, Chapter 3] and AdaGrad [8] is a foundational adaptive gradient descent method originally proposed in the OCO setting. The OCO framework can model a wide range of problems and Hazan [13, Chapter 1.2] illustrates some motivating examples in details. It is not hard to notice that measuring the performance of OCO algorithms simply by accumulated cost is meaningless since the adversary can make it arbitrarily large. Therefore, the notion of *regret* was proposed. It is the difference between the cost incurred to the player and a comparison point $z \in \mathscr{X} \subseteq \mathbb{R}^n$ (usually with minimum cumulative cost), that is to say,

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

If we fix the comparison point to be the point with minimum cumulative cost, regret measures how much worse the points picked by the algorithm compared to the best ones. Obviously, the goal of OCO is to minimize the regret by picking the best points that minimizes the cumulative costs.

1.1 Lipschitzness and Strong Convexity

Before we dive into classical OCO algorithms and their theoretical bounds, we need to mention several mathematical concepts. First we mention the use of some mathematical notations. Throughout this thesis, $\mathbb{R}^n$ denotes an n-dimensional real vector space endowed with an inner-product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$. We take $\mathscr{X} \subseteq \mathbb{R}^n$ to be a fixed closed convex set. The **dual norm** of $\|\cdot\|$ is defined by $\|x\|_* := \sup_{y \in \mathbb{R}^n \colon \|y\| \leq 1} \langle x, y \rangle$ for each $x \in \mathbb{R}^n$. The **normal cone** of $\mathscr{X}$ at a point $x \in \mathscr{X}$ is the set $N_{\mathscr{X}}(x) := \{ a \in \mathbb{R}^n : \langle a, z - x \rangle \leq 0 \text{ for all } z \in \mathscr{X} \}$.

Up to this point, the only constraint on the cost functions is convexity. However, generally OCO algorithms cannot have good performance guarantees over convex functions without extra assumptions. Let's consider the offline setting, where *Lipschitz smoothness* and *strong convexity* are two extra conditions we put on the objective function other than convexity to obtain convergence rate. Definitions of the two are stated below respectively.

Definition 1 (Lipschitz smoothness)**.** *A differentiable function $f\colon \mathscr{X} \to \mathbb{R}$ is L-Lipschitz smooth with respect to $\|\cdot\|$ on $\mathscr{X}' \subseteq \mathscr{X}$ for some $L > 0$ if*

$$|\nabla f(x) - \nabla f(y)| \leq L\|x-y\|, \qquad \forall x,y \in \mathscr{X}'.$$

By convention, $\nabla f(x)$ denotes the gradient of f at x.

Definition 2 (Strong convexity)**.** *A differentiable convex function $f\colon \mathscr{X} \to \mathbb{R}$ is M-strongly convex with respect to $\|\cdot\|$ on $\mathscr{X}' \subseteq \mathscr{X}$ for some $M > 0$ if*

$$f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle + \frac{M}{2}\|y-x\|^2, \qquad \forall x,y \in \mathscr{X}'.$$

If we combine convexity and Lipschitz smoothness, we get the following lemma [6, Lemma 3.4].

Lemma 1 (The Descent Lemma)**.** *If f is L-Lipschitz smooth on $\mathscr{X}' \subseteq \mathscr{X}$, then for any $x,y \in \mathscr{X}'$, one has*

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

Therefore, for differentiable convex functions, Lipschitz-smoothness gives a quadratic upper bound and strong convexity gives a quadratic lower bound. Both bounds, especially the lower bound, constrain the behaviour of the function in a desirable way. With such constraints, cost functions behave nicely and OCO algorithms can achieve sublinear regret.

The above definitions applied to functions that are continuously differentiable. For the sake of bookkeeping, in the later text when we take about differentiablity, we refer to continuous differentiability. Sometimes in offline convex optimization, we assume the differentiability of the objective function. On the other hand, such assumption is rare in the online setting. Indeed, differentiability is typically not assumed of cost functions in analysis of OCO algorithms due to its generality. For any convex function $f\colon \mathscr{X} \to \mathbb{R}$ 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*

$$f(z) \geq f(x) + \langle g, x - z \rangle, \qquad \forall z \in \mathbb{R}^n. \tag{1.1}$$

We denote by $\partial f(x)$ the set of all subgradients of f at x, called the **subdifferential** of f at x. For differentiable functions, the subdifferential at a certain point contains exactly one element, the gradient of the function at that point, subgradient is a more general concept than gradient and it can be seen as a proper surrogate for gradients of non-differentiable functions. Naturally, the definition of strong convexity can be extended to non-differentiable functions by replacing the gradient with a subgradient.

Definition 3 (Strong convexity for non-differentiable function)**.** *A convex function $f\colon \mathscr{X} \to \mathbb{R}$ is M-strongly convex with respect to $\|\cdot\|$ on $\mathscr{X}' \subseteq \mathscr{X}$ for some $M > 0$ if*

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

Again, this definition gives a quadratic lower bound for the function. On the other hand, the upper bound given by the Lipschitz smooth condition does not hold any more. Instead, we have the Lipschitz continuity condition.

Definition 4 (Lipschitz continuity)**.** *A function $f\colon \mathscr{X} \to \mathbb{R}$ is L-Lipschitz continuous with respect to $\|\cdot\|$ on $\mathscr{X}' \subseteq \mathscr{X}$ for some $L > 0$ if*

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

The definition means the growth of the function value is upper bounded by a constant multiplying the distance measured by a norm. More intuitively, we can say the function value will not change too abruptly and this property can be used to derive desirable regret bounds. Additionally, if f is convex, then the above definition implies[1] that $\|g\|_* \leq L$ for all $x \in \mathscr{X}$ and all $g \in \partial f(x)$, equivalently, the subgradient of the function is upper bounded by a constant [24, Lemma 2.6].

1.2 Follow The Regularized Leader

In this section we begin to introduce OCO algorithms. Since the cost functions are revealed after each round, a natural and simple algorithm will just try to minimize the cost incurred in all past rounds. This algorithm is called follow the leader (FTL). More specifically, in each round, FTL selects the point satisfying

$$x_t = \arg\min_{x \in \mathscr{X}} \sum_{i=1}^{t-1} f_i(x). \tag{1.2}$$

FTL uses the past accumulative costs to estimate the accumulative cost up to this round, so it implicitly assumes that the cost function in this round does not vary too much from the ones from the previous functions. However, this assumption may not hold. Here we give a counter example.

Example 1. *Let $\mathscr{X} = [-1,1]$ and $f_t(x) = z_t x + 1$. Define losses*

$$z_t = \begin{cases} -0.5 & t = 1, \\ 1 & t \text{ even}, \\ -1 & t \text{ odd}. \end{cases}$$

FTL suffers cost at least $2T - 1.5$ but the point $x = 0$ suffers cost T, thus the regret between points picked by FTL and the best points are at least $T - 1.5$.

At this point, it is not clear how good the regret $T - 1.5$ is. Later we will prove that regret in the order of $O(T)$ is considered to be bad in OCO for linear cost functions with bounded coefficients since some algorithms can achieve strictly better regret with similar computation cost. The weakness of FTL is the explicit assumption that the cost functions are rather "stable", which means they do not change too much from round to round. Therefore, a fix of FTL is to try to stabilize

[1] On the boundary of $\mathscr{X}$ 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 [4, §5.3]

these cost functions by adding a convex regularizer R. We assume there exists a minimizer for R. The resulting algorithm is called follow the regularized leader (FTRL), described as below.

Algorithm 1 Follow the Regularized Leader (FTRL) Algorithm

```
Compute $x_1 \in \arg\min_{x\in\mathscr{X}} R(x)$
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 $x_{t+1} \in \arg\min_{x\in\mathscr{X}} \left(F_t(x) + \frac{1}{\eta_t} R(x)\right)$
```

We start to analyse regret of FTRL. Because the main goal of this section is to give readers a taste of OCO algorithms, we will prove the regret upper bound in the simplest setting. To be more specific, we will assume $\mathscr{X} = \mathbb{R}^n$ and all cost functions are linear with bounded coefficients. This is called online linear optimization (OLO). Formally, this assumption means $f_t(x) = \langle z_t, x\rangle$ and $\|z_t\| \le L$ for all $t \ge 1$. Furthermore, we assume that the time horizon T is known so we can use a constant predetermined step size. In Chapter 3 we will prove a much more general result in a powerful proof framework.

The first step of the analysis of FTRL is to relate the regret to the cumulative difference between the cost of x_t and x_{t+1}. The following lemma is called the FTRL lemma.

Lemma 2 (FTRL Lemma)**.** *Let $x_1, x_2, \cdots$ be the sequence of points produced by* FTRL*. Then, for all $z \in \mathscr{X}$ we have*

$$Regret_T(z) = \sum_{t=1}^{T} (f_t(x_t) - f_t(z)) \le \frac{1}{\eta_T} R(z) + \sum_{t=1}^{T} (f_t(x_t) - f_t(x_{t+1})).$$

We will prove a stronger version of this lemma called strong FTRL lemma later. For a simpler proof of this lemma and more details about the difference between the stronger version and the standard version, we refer readers to McMahan [18, Lemma 13]. If we choose the regularizer to be $R(x) = \frac{1}{2}\|x\|_2^2$ and set step size $\eta_t = \eta$, we have the following regret bound for FTRL.

Theorem 3. *Consider running* FTRL *on a sequence of linear functions $f_t(x) = \langle z_t, x\rangle$ for all t with $\mathscr{X} = \mathbb{R}^n$, and with the regularizer $R(x) = \frac{1}{2}\|x\|_2^2$ and step size $\eta_t = \eta$ for $t \geq 1$. Then for all $z \in \mathscr{X}$, we have*

$$Regret_T(z) \leq \frac{1}{2\eta}\|z\|_2^2 + \eta\sum_{t=1}^{T}\|z_t\|_2^2,$$

In particular, if $\|z\|_2 \leq K$ and each f_t is L-Lipschiz continuous, then by setting $\eta = \frac{K}{L\sqrt{2T}}$, we obtain

$$Regret_T(z) \leq KL\sqrt{2T}.$$

Proof. The proof follows Shalev-Shwartz [24, Theorem 2.4]. By first-order optimality condition, we can compute x_{t+1} by

$$0 = \sum_{i=1}^{t} z_i + \frac{1}{\eta}x_{t+1}.$$

This is equivalent to

$$w_{t+1} = -\eta\sum_{i=1}^{t} z_i = w_t - \eta z_t. \tag{1.3}$$

Using Lemma 2 and (1.3), we have

$$\begin{aligned}
\text{Regret}_T(z) &\leq \frac{1}{\eta}R(z) + \sum_{t=1}^{T}(f_t(x_t) - f_t(x_{t+1})),\\
&\leq \frac{1}{2\eta}\|z\|_2^2 + \sum_{t=1}^{T}\langle x_t - x_{t+1}, z_t\rangle,\\
&= \frac{1}{2\eta}\|z\|_2^2 + \eta\sum_{t=1}^{T}\|z_t\|_2^2.
\end{aligned}$$

The first equality comes from the definition of R and the second inequality is the subgradient inequality. Lipschitz continuity implies $\|z_t\|_2^2 \leq L^2$ for every $t \geq 1$. By setting the step size η to be $\frac{K}{L\sqrt{2T}}$ and using $\|z\|_2 \leq K$, we get the desired bound as below.

$$\begin{aligned}\text{Regret}_T(z) &\leq \frac{1}{2\eta}K^2 + \eta\sum_{t=1}^{T}\|z_t\|_2^2,\\ &= KL\sqrt{2T}.\end{aligned}$$

□

With some hindsight, we realize that the regret obtained by FTL in Example 1 is indeed suboptimal and FTRL is superior in this case. Actually, this sublinear $O(\sqrt{T})$ regret matches the lower bound of OLO by a constant factor [22, Theorem 5]. Therefore, FTRL is optimal for linear cost functions. Since OLO is incorporated in OCO, this $O(\sqrt{T})$ is also lower bound for OCO. Chapter 3 shows the $O(\sqrt{T})$ regret holds for FTRL in a more general setting than OLO (relative Lipschitz continuous cost functions), so the optimality of FTRL holds in this more general setting as well.

If in addition we assume that all the cost functions are strongly convex, we will have regret bound $O(\log T)$ in terms of rounds T [14]. Again the details and the proof will be deferred to Chapter 3.2 when more general results are proved.

1.3 Mirror Descent

The mirror descent algorithm is a generalization of the classical gradient descent method that was first proposed by Nemirovsky and Yudin [20] with a modern treatment first given by Beck and Teboulle [3]. Moreover, the algorithm almost seamlessly fits into the OCO setting (see [12]), and its online version is known as online mirror descent (OMD). Throughout this thesis we follow Bubeck [6]'s notations and assumptions on mirror descent. We assume that we have a **mirror map** for $\mathscr{X}$, that is, a differentiable strictly-convex function $\Phi\colon \mathscr{D} \to \mathbb{R}$ for $\mathscr{X}$ such that the gradient of Φ diverges on the boundary of $\mathscr{D}$, that is, $\lim_{x\to\partial\mathscr{D}}\|\nabla\Phi(x)\|_2 = \infty$ and $\partial\mathscr{D} := \mathscr{D}\setminus\mathscr{D}^\circ$. Here $\mathscr{D}^\circ$ denotes the interior of $\mathscr{D}$. These presumed conditions of the mirror map gives a unique solution to the projection step of MD (last step in Algorithm 2). The mirror map function in OMD is sometimes called a regularizer function, just as in FTRL. OMD is described in Algorithm 2.

Algorithm 2 Online Mirror Descent

```
Require: An initial iterate $x_1 \in \mathcal{X}$.
  for $t = 1, 2, \ldots$ do
      Observe $f_t$ and suffer cost $f_t(x_t)$
      Compute $g_t \in \partial f_t(x_t)$
      $\hat{x}_t := \nabla\Phi(x_t)$
      $\hat{y}_{t+1} := \hat{x}_t - \eta_t g_t$
      $y_{t+1} := \nabla\Phi^*(\hat{y}_{t+1})$
      Compute $x_{t+1} \in \arg\min_{x \in \mathcal{X}} \Phi(x) - \Phi(y_{t+1}) - \langle \nabla\Phi(y_{t+1}), x - y_{t+1} \rangle$
```

The notation $\nabla\Phi^*$ denotes the **convex conjugate** of $\nabla\Phi$. Formally, for a function $f\colon \mathcal{X} \to \mathbb{R}$, the convex conjugate $f^*(y) = \sup_{x \in \mathcal{X}} \{\langle y, x \rangle - f(x)\}$ for every $y \in \mathcal{X}^*$, where $\mathcal{X}^*$ is the dual space of $\mathcal{X}$.

The motivation for mirror descent comes from looking at the gradient of a function f as the representation[2] of its derivative. The derivative is a functional on $\mathbb{R}^n$ because it is a linear function that receives a direction $d \in \mathbb{R}^n$ and evaluates the derivative in the direction d). If we were to perform gradient descent in a space where such a representation—the gradients—might not exist, such as in some Banach spaces, the gradient step would not make sense: it tries to subtract from a point in the "primal space" $\mathbb{R}^n$ an element from the the space of functionals on $\mathbb{R}^n$, known as the *dual space* of $\mathbb{R}^n$. A way to generalize gradient descent for it to make sense from this perspective is to have a strictly convex function $\Phi\colon \mathbb{R}^n \to \mathbb{R}$, a *mirror map*, to bridge the primal and dual spaces. As an example, note that by defining $\Phi := \frac{1}{2}\|\cdot\|_2^2$ we have $\nabla\Phi(x) = x$ for any $x \in \mathbb{R}^n$. In this case, the classical gradient step can be seen as subtracting the derivative of the objective function f from the derivative of the mirror map at the current iterate. Interestingly, we may pick mirror maps different from the squared Euclidean norm, yielding different algorithms with different performance guarantees. Thus, even though we do not actually need to worry about derivative representation in $\mathbb{R}^n$, this general perspective on gradient descent is useful in the design of efficient optimization methods. For a more detailed discussion of mirror descent and its intuition, see Bubeck [6].

OMD describes many algorithms. For example, if we take the mirror map to be

[2] Representation given by the Riesz Representation Theorem.

half of the ℓ_2 norm, both $\nabla\Phi$ and $\nabla^*\Phi$ are the identity function and the projection step becomes

$$\begin{aligned} x_{t+1} &\in \arg\min_{x\in\mathscr{X}} \Phi(x) - \Phi(y_{t+1}) - \langle \nabla\Phi(y_{t+1}), x - y_{t+1}\rangle, \\ &= \arg\min_{x\in\mathscr{X}} \frac{1}{2}\|x\|_2^2 - \frac{1}{2}\|x_t - \eta_t g_t\|_2^2 - \langle x_t - \eta_t g_t, x - (x_t - \eta_t g_t)\rangle, \\ &= \arg\min_{x\in\mathscr{X}} \frac{1}{2}\|x - (x_t - \eta_t g_t)\|_2^2. \end{aligned}$$

Then OMD recovers the classical projected subgradient descent method. To see more relevant applications of OMD, read Hazan [13, Chapter 5.4] and Bubeck [6, Chapter 4.3]. Similar to FTRL, classical results shows that if the cost functions f_t are Lipchitz continuous and we pick the mirror map to be strongly convex, we have $O(\sqrt{T})$ regret in terms of rounds T [13, Theorem 5.6], [6, Theorem 4.2]. In Chapter 3.2, we will prove a more general result on a variant version of OMD.

Chapter 2

Related Work

In the previous chapter, we introduced some classical OCO algorithms and their regret bounds. Despite that the classical regret bounds can be applied to various popular cost functions, it necessitates the Lipschitz continuity assumption of the cost functions. However, some cost functions may not satisfy the Lipschitz continuity condition, for example, the loss function for support vector machines, $f(x) = \max\{0, 1 - y_i x^T w_i\} + \frac{\lambda}{2}\|x\|_2^2$. The squared ℓ_2 norm term is not Lipschitz continuous thus classical regret bounds do not apply to it. This problem motivates a line of work that relax the Lipschitzness condition in both offline and online convex optimization settings. In this chapter we will talk about these work. Before diving into details, we need to familiarize ourselves with an important mathematical notion, the **Bregman divergence**.

Definition 5. *Let $R\colon \mathscr{D} \to \mathbb{R}$ be a convex function such that it is differentiable in $\mathscr{D}^{\mathrm{o}} := \operatorname{int}\mathscr{D}$ and such that we have $\mathscr{X} \subseteq \mathscr{D}^{\mathrm{o}}$. The Bregman divergence (with respect to R) is given by*

$$D_R(x,y) := R(x) - R(y) - \langle \nabla R(y), x - y \rangle, \qquad \forall x \in \mathscr{D}, y \in \mathscr{D}^{\mathrm{o}}.$$

More intuitively, the Bregman divergence measures the distance between two points through the lens of some reference function R. One nuance of this definition is that it is not symmetric, in general, $D_R(x,y) \neq D_R(y,x)$. A noteworthy example of Bregman divergence is the ℓ_2 distance between two points $D_R(x,y) = \frac{1}{2}\|x-y\|_2^2$

if we take R to be $\frac{1}{2}\|\cdot\|_2^2$. An interesting and useful identity regarding Bregman divergences, sometimes called the *three-point identity* [6], is

Lemma 4 (Three-point identity)**.**

$$D_R(x,y) + D_R(z,x) - D_R(z,y) = \langle \nabla R(x) - \nabla R(y), x - z \rangle, \qquad \forall z \in \mathscr{D}, \forall x,y \in \mathscr{D}^{\mathrm{o}}.$$

This identity can be simply proved by expanding the left side by the definition of Bregman divergence.

We collect here some additional notation and assumptions used throughout the thesis.[1] First, $\mathscr{X} \subseteq \mathbb{R}^n$ denotes a closed convex set and $\{f_t\}_{t\geq 1}$ denotes a sequence of convex functions such that $f_t \colon \mathscr{X} \to \mathbb{R}$ is subdifferentiable[2] on $\mathscr{X}$ 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, $\mathscr{D} \subseteq \mathbb{R}^n$ denotes a convex set with non-empty interior $\mathscr{D}^\circ := \mathrm{int}(\mathscr{D})$ such that $\mathscr{X} \subseteq \mathscr{D}^\circ$. This latter set will be the domain of the regularizer for FTRL and of the mirror map for OMD. Namely, in Chapter 3 we denote by $R \colon \mathscr{D} \to \mathbb{R}$ the *regularizer* of FTRL, a convex function which is differentiable on $\mathscr{D}^\circ$. In Chapter 4 we denote by $\Phi \colon \mathscr{D} \to \mathbb{R}$ the *mirror map* of online mirror descent.

2.1 A New Descent Lemma

We start by recalling the convex optimization problem with composite objective. Given a closed convex set C with nonempty interior, the problem is

$$\inf\{\Phi(x) := f(x) + g(x) : x \in C\},$$

where f and g are proper, convex and lower semicontinuous, with g continuously differentiable on the interior of its domain. Here f is called the proximal function. It can be seen as a regularizer that we want to minimize directly, for example, the ℓ_2 norm. The proximal gradient descent (PG) method is a classical first-order

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

[2] This is not too restrictive since convex functions are subdifferentiable on the relative interior of their domains [23, Theorem 23.4].

algorithm for solving this problem, the main step is

$$x_{t+1} = \arg\min\{g(x_t) + \langle \nabla g(x_t), u - x_t\rangle + \frac{1}{2\lambda}\|u - x_t\|^2 + f(u) : u \in C\},$$

where $\lambda > 0$ is the step size here. This key step tries to minimize a quadratic upper bound (for $\lambda < 1/L$ where L is the Lipschitz smoothness parameter) of the differentiable function g plus the proximal function f. It is not hard for us to relate this step to the descent lemma (1) since it also provides a quadratic upper bound. Actually the descent lemma is a crucial pillar in convergence analysis of PG. Bauschke et al. [2] took a new look at PG and relaxed the descent lemma to the following assumption. Let f be a convex and h be a Legendre function, which means h is essentially smooth and strictly convex in the interior of its domain. Let $L > 0$, we require

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + LD_h(y,x). \tag{2.1}$$

This new assumption substitutes the norm in the classical descent lemma with the Bregman divergence. As we discussed before, the Bregman divergence can be seen as a generalization of ℓ_2 norm, so we can see this assumption as a generalization of the classical descent lemma. They proposed the NoLips algorithm to solve the composite objective convex optimization problem. The main step is

$$x_{t+1} = \arg\min\{g(x_t) + \langle \nabla g(x_t), u - x_t\rangle + \frac{1}{2\lambda}D_h(u, x_t) + f(u) : u \in C\},$$

where h is the chosen Legendre function. Similarly, the norm in classical PG got replaced by the Bregman divergence. Therefore, NoLips is a generalization of the classical PG. Convergence analysis of classical PG needs the classical descent lemma. This further requires Lipschitz smoothness of the differentiable part g. Analogously, convergence analysis of NoLips onlys needs the new descent condition (2.1). Therefore, we are granted the freedom to choose a Legendre function h to satisfy (2.1) in order to make NoLips converge fast. This makes NoLips flexible and extends the classical convergence bound for PG like algorithms.

It is worth mentioning that Van Nguyen [25] independently developed similar ideas for analyzing the convergence of a Bregman proximal gradient applied to the convex composite model in Banach spaces. Bolte et al. [5] extended the framework of Bauschke et al. [2] to the non-convex setting.

2.2 Relative Lipschitzness

Along this direction, Lu et al. [16] further formalized and simplified this “beyond Lipschitz smoothness” idea and proposed the notion of relative Lipschitz smoothness.

Definition 6 (Relative Lipschitz smoothness [16])**.** *A differentiable convex function $f\colon \mathscr{X} \to \mathbb{R}$ is **L-Lipschitz smooth** relative to R for some $L \geq 0$ if*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + LD_R(y,x), \qquad \forall y,x \in \mathscr{X}, \forall g \in \partial f(x). \tag{2.2}$$

This relative Lipschitz smoothness notion is essentially equivalent to the condition (2.1) in Bauschke et al. [2], but Bauschke et al. [2] has the extra requirement that the reference function needs to be Legendre. Following the same logic, they also proposed the definition of relative strong convexity.

Definition 7 (Relative strong convexity [16])**.** *A convex function $f\colon \mathscr{X} \to \mathbb{R}$ is **M-strongly convex** relative to R if*

$$f(y) \geq f(x) + \langle g, y - x \rangle + MD_R(y,x), \qquad \forall y,x \in \mathscr{X}, \forall g \in \partial f(x). \tag{2.3}$$

Proposition 1.1 and Proposition 1.2 in Lu et al. [16] categorize equivalent definitions for relative Lipschitz smoothness/strong convexity and present some of their properties. Lu et al. [16] reproved the linear convergence of NoLips (without the proximal function) in a simpler style, where they called the algorithm primal gradient descent. They also proved the linear convergence of dual averaging (DA) under the relative Lipschitz smooth and relative strong convexity condition, an algorithm invented by Nesterov [21]. DA uses the averaged linearization over past iterates as an estimate to the current function. We will formally introduce it in Chapter 3.1.1. Application-wise, Lu et al. [16] proved their new theory can be used to solve the D-optimal design problem. Later, Mukkamala and Ochs [19] modified the NoLips algorithm by adding momentum to it and used it to solve the non-convex matrix factorization problem.

Aside from the above work, in a related paper Lu [15] extended this "beyond Lipschitzness" work to the non-differentiable convex optimization. He proposed the notion of relative Lipschitz continuity, which can be seen as a generalization of its classical counterpart.

Definition 8 (Relative Lipschitz continuity)**.** *A convex function $f\colon \mathscr{X} \to \mathbb{R}$ is* ***L-Lipschitz continuous*** *relative to R if*

$$\|g\|_* \leq \frac{L\sqrt{2D_R(y,x)}}{\|y-x\|}, \qquad \forall x,y \in \mathscr{X} \textit{ with } x \neq y, \forall g \in \partial f(x).$$

Again, if we pick the norm to be ℓ_2 norm and R to be $\frac{1}{2}\|\cdot\|$, this relative Lipschitz continuity recovers the classical definition. A drawback of this definition is it is norm dependent, thus, we proposed a slight modification of the definition.

Definition 9 (modified Relative Lipschitz continuity)**.** *A convex function $f\colon \mathscr{X} \to \mathbb{R}$ is* ***L-Lipschitz continuous*** *relative to R if*

$$\langle g, x-y\rangle \leq L\sqrt{2D_R(y,x)}, \qquad \forall x,y \in \mathscr{X}. \tag{2.4}$$

By definition of dual norm, $\langle g, x-y\rangle \leq \|g\|_*\|y-x\|$ for all $g \in \partial f(x)$, so this modified definition is more general. When we mention relative Lipschitz continuity in the later text, we mean this modified version. Lu [15] proved the $O(1/\sqrt{T})$ convergence rate of deterministic and stochastic mirror descent, given that the objective function is relative Lipschitz continuous to the mirror map. A faster $O(1/T)$ convergence rate was also proved if in addition the objective fucntion satisfies relative strong convexity with respect to the mirror map.

2.3 Other Related Work

Antonakopoulos et al. [1] generalized the Lipschitz continuity condition from the perspective of Riemannian geometry. They proposed the notion of Riemannian Lipschitz-continuity (RLC).

Definition 10. *We say that $f : \mathscr{X} \to \mathbb{R}$ is Riemann-Lipschitz continuous relative* *to a Riemann metric g if*

$$|f(y) - f(x)| \leq G dist_g(x,y) \quad \textit{for some } G \geq 0 \textit{ and all } x, y \in \mathscr{X},$$

where $\mathrm{dist}_g(x,y)$ is the Riemann distance between x and y induced by the Riemann metric g.

We note that RLC in Antonakopoulos et al. [1] is closely related to relative Lipschitz continuity. If we assume differentiablity of the function, by definition of RLC and Cauchy-Schwartz inequality, RLC implies relative Lipschitz continuity. We do not know if the reverse direction holds or the relationships of these two notions in the non-differentiable case due to the hardness to find a proper Riemannian metric or compute the Riemannian distance. In the paper, they analyzed how OCO algorithms perform in this setting and showed $O(\sqrt{T})$ regret bounds for both FTRL and OMD with RLC and differentiable cost functions in the online settings.

More recently, Gao et al. [10] analysed the coordinate descent method with composite Lipschitz smooth objectives and Grimmer [11] showed how projected subgradient descent method enjoys a $O(1/\sqrt{T})$ convergence rate without Lipschitz continuity if one has some control on the norm of the subgradients. Maddison et al. [17] relaxed the Lipschitz smoothness condition by proposing a new family of optimization methods motivated from physics, to be more specific, the conformal Hamiltonian dynamics.

Chapter 3

Follow the Regularized Leader

In this section we prove regret bounds of relative Lipschitz cost functions. Besides that, we also prove logarithmic regret if the cost functions are additionally relatively strongly convex and extend such results to the composite cost functions setting (defined in Chapter 2.1). These results give generalized regret bounds in OCO. Before giving details of our theorems and proofs, we want to introduce the Strong FTRL Lemma, which along with its variants will be a pillar in our regret analysis.

Lemma 5. *(Strong* FTRL *Lemma [18]) Let $\{f_t\}_{t\geq 1}$ be a sequence of functions such that $f_t \colon \mathscr{X} \to \mathbb{R}$ for each $t \geq 1$. Let $\{\eta_t\}_{t\geq 1}$ be a positive non-increasing sequence. Let $R \colon \mathscr{X} \to \mathbb{R}$ be such that $\{x_t\}_{t\geq 1}$ given as in Algorithm 1 is properly defined. If $F_t \colon \mathscr{X} \to \mathbb{R}$ is defined as in Algorithm 1 for each $t \geq 1$, then,*

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

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 $t \geq 1$.

Below we give the proof of this lemma. 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 [18].

Proof of Lemma 5. 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

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

In particular,

$$x_t \in \arg\min_{x \in \mathscr{X}} H_{t-1}(x) = \arg\min_{x \in \mathscr{X}} \sum_{i=0}^{t-1} h_i(x), \qquad \forall t \geq 0. \tag{3.1}$$

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

$$\begin{aligned}\sum_{t=1}^{T} (h_t(x_t) - h_t(z)) - h_0(z) = \sum_{t=1}^{T} h_t(x_t) - H_T(z) &= \sum_{t=1}^{T} (H_t(x_t) - H_{t-1}(x_t)) - H_T(z) \\ &\overset{(3.1)}{\leq} \sum_{t=1}^{T} (H_t(x_t) - H_{t-1}(x_t)) - H_T(x_{T+1}) \\ &= \sum_{t=1}^{T} (H_t(x_t) - H_t(x_{t+1})) - H_0(x_1),\end{aligned}$$

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

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

For the composite setting with an extra convex regularizer Ψ (see Chapter 3.3 for details), we modify the definition of r_t for $t \geq 1$ (maintaining the definition of r_0) in the above proof for

$$r_t \coloneqq \Big(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\Big)R + \Psi, \qquad \forall t \geq 1.$$

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 5, we get

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

Re-arranging yields

$$\mathrm{Regret}_T^{\Psi}(z) \leq \sum_{t=0}^{T}\Big(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\Big)(R(z) - R(x_t)) + \sum_{t=1}^{T}(H_t(x_t) - H_t(x_{t+1})). \tag{3.2}$$

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 boils down to bounding the terms $H_t(x_t) - H(x_{t+1})$ 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 6. *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 $z \in \mathscr{X}$ and let $K \in \mathbb{R}$ be such that $K \geq R(z) - R(x_1)$. Then,*

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

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

We notice that here the step size is adaptive. It does not depend on the horizon T but the current iteration t. In this way, the regret bound can hold for any horizon T without knowing it in advance. However, the step size has a dependence on the radius K and relative Lipschitz continuity parameter L. In practice we may not know these parameters and we may need to search for the proper step size.

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 6. 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 7. *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 $p_t \in N_{\mathscr{X}}(x_t)$ 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*

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

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

$$\partial\Big(F_{t-1} + \tfrac{1}{\eta_{t-1}}R\Big)(x_t) \cap (-N_{\mathscr{X}}(x_t)) \neq \varnothing.$$

Since $\partial\Big(F_{t-1} + \frac{1}{\eta_{t-1}}R\Big)(x_t) = \partial F_{t-1}(x_t) + \frac{1}{\eta_{t-1}}\nabla R(x_t)$, the above shows there is $p_t \in$

$N_{\mathscr{X}}(x_t)$ such that

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

Using the subgradient inequality (1.1) with the above subgradient yields,

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

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, $-\langle \nabla R(x_t), x_t - x_{t+1} \rangle = R(x_{t+1}) - R(x_t) - D_R(x_{t+1}, x_t)$. □

Proof of Theorem 6. For each $t \geq 0$ let H_t be defined as in the Strong FTRL Lemma and fix $t \geq 0$. We have

$$H_t(x_t) - H_t(x_{t+1}) = F_t(x_t) - F_t(x_{t+1}) + \frac{1}{\eta_t}(R(x_t) - R(x_{t+1})). \tag{3.3}$$

Using $F_t = F_{t-1} + f_t$ together with Lemma 7 we have

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

Plugging the above inequality into (3.3) yields

$$(3.3) \leq f_t(x_t) - f_t(x_{t+1}) - \frac{D_R(x_{t+1}, x_t)}{\eta_{t-1}} + \Big(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \Big)(R(x_t) - R(x_{t+1})). \tag{3.4}$$

Since f_t is L-relative Lipschitz continuous with respect to R, we apply subgradient inequality with (2.4) followed by the 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

$$\begin{aligned}
f_t(x_t) - f_t(x_{t+1}) - \frac{D_R(x_{t+1},x_t)}{\eta_{t-1}} &\leq \langle g_t, x_t - x_{t+1}\rangle - \frac{D_R(x_{t+1},x_t)}{\eta_{t-1}}, \qquad \text{(subgradient inequality)} \\
&\overset{(2.4)}{\leq} L\sqrt{2D_R(x_{t+1},x_t)} - \frac{D_R(x_{t+1},x_t)}{\eta_{t-1}}, \\
&\leq \frac{L^2\eta_{t-1}}{2}.
\end{aligned}$$

Applying the above on (3.4) yields

$$(3.4) \leq \frac{L^2\eta_{t-1}}{2} + \Big(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\Big)(R(x_t) - R(x_{t+1})).$$

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

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

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 25 in Appendix A.1, then

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

□

3.1.1 Dual Averaging

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 variation of FTRL that mitigates this problem is the online DA method whose offline version is due to

Nesterov [21]. In each iteration, DA picks a point from $\mathscr{X}$ 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 $\mathscr{X}$, the online DA method computes iterates $\{x_t\}_{t\geq 1}$ such that

$$x_{t+1} \in \operatorname*{arg\,min}_{x\in\mathscr{X}} \Big(\eta_t \sum_{i=1}^{t} \langle g_i, x\rangle + R(x) \Big) \qquad \forall t \geq 0, \tag{3.5}$$

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

It is well-known that the DA algorithm reduces to FTRL applied to the linearized functions $\{\tilde{t}_t\}_{t\geq 1}$ given by $\tilde{t}_t := \langle g_t, \cdot\rangle$ for each $t \in \mathbb{N}$ (for details see Hazan [12, Lemma 5.4]). This reduction obviously preserves the property of being Lipschitz continuous since the gradient of $\tilde{t}_t$ 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 (2.4)) of f_t implies $\langle \nabla \tilde{t}_t(x_t), x_t - y\rangle = \langle g_t, x_t - y\rangle \leq L\sqrt{2D_R(y,x_t)}$ for all $y \in \mathscr{X}$. That is all we need for the proof of Theorem 6 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 6.

Corollary 8. *Let $\{x_t\}_{t\geq 1}$ be defined as in (3.5) and suppose f_t is L-Lipschitz continuous relative to R for all $t \geq 1$. Let $z \in \mathscr{X}$ 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 $\mathrm{Regret}_T(z) \leq 2L\sqrt{K(T+1)}$.*

3.2 Logarithmic Regret with Relative Strongly Convex Functions

Hazan et al. [14] showed that if the cost functions are not only Lipschitz continuous but strongly convex as well, then FTL 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. A formal theorem is given below.

Theorem 9. *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 $h\colon \mathscr{D} \to \mathbb{R}$ for each $t \geq 1$. Then, for all $z \in \mathscr{X}$,*

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

The next lemma strengthens the bound from Lemma 7 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 10. *Let $\{x_t\}_{t\geq 1}$ be defined as in Algorithm 1 with $R := 0$. Moreover, let $h\colon \mathscr{D} \to \mathbb{R}$ 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 $p_t \in N_{\mathscr{X}}(x_t)$ be as given by Lemma 7. By this lemma we have $-p_t \in \partial F_{t-1}(x_t)$. Thus, using inequality (2.3) 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. □

Proof of Theorem 9. For each $t \geq 0$ let $H_t\colon \mathscr{X} \to \mathbb{R}$ 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 10 yields

$$\begin{aligned} H_t(x_t) - H_t(x_{t+1}) = F_t(x_t) - F_t(x_{t+1}) &= F_{t-1}(x_t) - F_{t-1}(x_{t+1}) + f_t(x_t) - f_t(x_{t+1}) \\ &\leq -(t-1)MD_h(x_{t+1}, x_t) + f_t(x_t) - f_t(x_{t+1}). \end{aligned} \tag{3.6}$$

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

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

Applying the above to (3.6) 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

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

By the first order optimization condition and the definition of relative strong convexity, for any $y \in \mathscr{X}$ and a minimizer x^* of f, we have

$$f(y) - f(x^*) \geq MD_R(y, x^*).$$

At the same time, by definition of relative Lipschitz continuity, we have

$$f(y) - f(x^*) \leq L\sqrt{2D_R(x^*, y)}.$$

This means that if the function satisfies relative Lipschitz continuity and relative strong convexity with the same reference function at the same time, it has a lower bound of order $O(D_R(y, x^*))$ and an upper bound of order $O(\sqrt{D_R(x^*, y)})$. If the Bregman divergence between y and x^* goes to infinity, the lower bound will eventually exceed the upper bound and these two conditions cannot hold concurrently. Therefore, relative Lipschitz continuity and relative strong convexity can only co-exist in a bounded domain where the Bregman divergence does not explode.

3.3 Sublinear Regret with Composite Loss Functions

An important consideration for applications is a variant of OCO in which the loss functions are composite [7, 26]. 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 this subsection we extend the results from Chapter 3.1 to the case where the loss functions are *composite*. Specifically, there is a known non-negative convex function $\Psi\colon \mathscr{X} \to \mathbb{R}_+$ (sometimes called *extra regularizer*) which is subdifferentiable on $\mathscr{X}$ 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 [26].

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

$$\mathrm{Regret}_T^{\Psi}(z) \coloneqq \sum_{t=1}^{T} (f_t(x_t) + \Psi(x_t)) - \inf_{z \in \mathscr{X}} \sum_{t=1}^{T} (f_t(z) + \Psi(z)), \qquad \forall T > 0. \quad (3.7)$$

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

$$x_{t+1} \in \operatorname*{arg\,min}_{x \in \mathscr{X}} \Big(\sum_{i=1}^{t} f_i(x) + t\Psi(x) + \frac{1}{\eta_t} R(x) \Big), \qquad \forall t \geq 0.$$

We do make the additional assumption that $\Psi(x_1) = 0$, that is, x_1 minimizes Ψ and tha 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 6 for the composite setting even if Ψ is not relative Lipschitz-continuous with respect to the FTRL regularizer.

Theorem 11. *Let $\Psi\colon \mathscr{X} \to \mathbb{R}_+$ 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 $z \in \mathscr{X}$ and $K \in \mathbb{R}$ be such that $K \geq R(z) - R(x_1)$. Additionally, assume $\Psi(x_1) = 0$. Then,*

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

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

The proof is largely identical to the proof of Theorem 6. One of the main differences in the analysis is the following version of Lemma 7 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 12. *Let $\Psi\colon \mathscr{X} \to \mathbb{R}_+$ 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 $p_t \in N_{\mathscr{X}}(x_t)$ such that*

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

and the above implies

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

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

$$\partial\Big(F_{t-1} + (t-1)\Psi(x) + \tfrac{1}{\eta_{t-1}} R\Big)(x_t) \cap (-N_{\mathscr{X}}(x_t)) \neq \varnothing.$$

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 $p_t \in N_{\mathscr{X}}(x_t)$ 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 (1.1) with the above subgradient yields,

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

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, $-\langle \nabla R(x_t), x_t - x_{t+1}\rangle = R(x_{t+1}) - R(x_t) - D_R(x_{t+1}, x_t)$. □

Now we are in position to prove Theorem 11.

Proof of Theorem 11. We proceed in a way extremely similar to the proof of Theorem 6, but in place of the standard FTRL Lemma we use its composite version as in (3.2).

For each $t \geq 0$ let H_t be define das 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 12 we have

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

Proceeding as in the proof of Theorem 6 (with the addition of a $\Psi(x_t) - \Psi(x_{t+1})$ term) we have

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

When summing over $t \in \{1, \ldots, T\}$, 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 6. □

3.3.1 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 [18] shows that applying DA in the composite case with $\Psi := \|\cdot\|_1$ does not yield sparse solutions. Xiao [26] 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 $\arg\min_{x\in\mathscr{X}}(R(x)$ and is such that $\Psi(x_1) = 0$, that is, x_1 minimizes Ψ. For the following rounds, RDA computes

$$x_{t+1} \in \arg\min_{x\in\mathscr{X}}\Big(\sum_{i=1}^{t}\langle g_i, x\rangle + t\Psi(x) + \frac{1}{\eta_t}R(x)\Big) \qquad \forall t \geq 1. \tag{3.8}$$

With an argument analogous to the one made in Chapter 3.1.1, we can write RDA as an instance of FTRL with composite loss functions and obtain the following corollary of Theorem 11.

Corollary 13. *Let* $\Psi\colon \mathbb{R}^n \to \mathbb{R}_+$ *be a nonnegative convex function. Let* $\{x_t\}_{t\geq 1}$ *be defined as in* (3.8) *and assume* $\Psi(x_1) = 0$. *Moreover, suppose* f_t *is L-Lipschitz continuous relative to R for all* $t \geq 1$. *Let* $z \in \mathscr{X}$ *and let* $K \in \mathbb{R}$ *be such that* $K \geq R(z) - R(x_1)$. *If* $\eta_t \coloneqq \sqrt{2K}/(L\sqrt{t+1})$ *for all* $t \geq 1$, *then* $\mathit{Regret}_T^{\Psi}(z) \leq 2L\sqrt{K(T+1)}$.

Chapter 4

Dual Stabilized-Online Mirror Descent

We introduced classical mirror descent in Chapter 1.3. Mirror descent is a generalization of standard gradient method. It describes many popular algorithms like the projected gradient descent method and the exponentiated gradient descent method. Its convergence rate can have a better dependence on the dimensionality of some problems than normal gradient descent method [6, Chapter 4]. Recently, Orabona and Pál [22] showed that OMD with a dynamic learning rate may suffer *linear* regret. (A dynamic learning rate is useful when we do not known 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. [9] 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 3. 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. The proof of this chapter will build upon the framework in Fang et al. [9]. Thus, let us state inequality (4.9) and Claim 4.2 (without substituting exactly value of γ_t) from Fang et al. [9] at the beginning, which will appear multiple times throughout this chapter, respectively as:

Algorithm 3 Dual-Stabilized Online Mirror Descent

```
Require: Stabilization coefficient $\gamma_t$ and an initial iterate $x_1 \in \mathscr{X}$.
  for $t = 1, 2, \ldots$ do
      Observe $f_t$ and suffer cost $f_t(x_t)$
      Compute $g_t \in \partial f_t(x_t)$
      $\hat{x}_t := \nabla\Phi(x_t)$
      $\hat{w}_{t+1} := \hat{x}_t - \eta_t g_t$
      $\hat{y}_{t+1} := \gamma_t \hat{w}_{t+1} + (1-\gamma_t)\hat{x}_1$
      $y_{t+1} := \nabla\Phi^*(\hat{y}_{t+1})$
      Compute $x_{t+1} \in \arg\min_{x\in\mathscr{X}} D_\Phi(x, y_{t+1}) = \Phi(x) - \Phi(y_{t+1}) - \langle \nabla\Phi(y_{t+1}), x - y_{t+1}\rangle$
```

Claim 14. *If $\gamma_t = \eta_{t+1}/\eta_t \in (0,1]$ for each $t \geq 1$, then*

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

Claim 15. *If $\gamma_t \in (0,1]$ for all $t \geq 1$, then,*

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

Throughout this chapter, let $\{x_t\}_{t\geq 1}$ and $\{\hat{w}_t\}_{t\geq 1}$ be defined as in Algorithm 3, and define

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

4.1 Sublinear Regret with Relative Lipschitz Functions

In this subchapter, 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 16. *Let $\{x_t\}_{t\geq 1}$ be defined as in Algorithm 3 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 $z \in \mathscr{X}$ and $K \in \mathbb{R}$ be such that $K \geq D_\Phi(z, x_1)$. Then,*

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

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

To prove the above theorem, we need to use Theorem 4.1 in Fang et al. [9]. This theorem is analogous to the bound given in the analysis of classic OMD given by Bubeck [6, Theorem 4.2].

Theorem 17 (Fang et al. [9, Theorem 4.1])**.** *If $\gamma_t := \eta_{t+1}/\eta_t$ for each $t \geq 1$, then*

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

Now we are ready to use Theorem 17 to prove Theorem 16.

Proof of Theorem 16. 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 (4)),

$$D_\Phi(x_t, w_{t+1}) = -D_\Phi(w_{t+1}, x_t) + \langle \nabla\Phi(x_t) - \nabla\Phi(w_{t+1}), x_t - w_{t+1} \rangle. \tag{4.1}$$

From the definition of the iterates in Algorithm 3, we have $\eta_t g_t = \nabla\Phi(x_t) - \nabla\Phi(w_{t+1})$. Thus,

$$\begin{aligned}(4.1) &= -D_\Phi(w_{t+1}, x_t) + \eta_t \langle g_t, x_t - w_{t+1} \rangle \\ &\overset{(2.4)}{\leq} -D_\Phi(w_{t+1}, x_t) + \eta_t L \sqrt{2D_\Phi(w_{t+1}, x_t)} \leq \frac{\eta_t^2 L^2}{2},\end{aligned} \tag{4.2}$$

where first inequality is from (2.4) (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 17, we get

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

Setting $\eta_t := \sqrt{K}/L\sqrt{t}$ for each $t \geq 1$ and by using Lemma 25 from Appendix A.1we have

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

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 18. *Let Φ be a mirror map for $\mathcal{X}$ and let $f \colon \mathcal{X} \to \mathbb{R}$ be a convex L-Lipschitz-continuous function relative to Φ. Let $\{x_t\}_{t\geq 1}$ be given as in Algorithm 3 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 [15], if we take $k = T-1$ and $t_i = \frac{\sqrt{K}}{\sqrt{T}L}$ for $i \geq 0$ therein.

4.2 Logarithmic Regret with Relative Strongly Convex Functions

In Chapter 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 3 boils down to classic

OMD.

Theorem 19. *Let $\{x_t\}_{t\geq 1}$ be given as in Algorithm 3 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 $z \in \mathscr{X}$ and $\eta_t = \frac{1}{tM}$ for each $t \geq 1$, then,*

$$Regret_T(z) \leq \frac{L^2}{2M}(\log T + 1), \qquad \forall T > 0.$$

The first step in the proof is the following claim given by modifying Claims 14 and 15 and combining them together.

Claim 20. *Assume that $\gamma_t = 1$ for all $t \geq 1$, then*

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

Proof of Claim 20. This proof largely follows the structure of the proof of Claim 14. 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 14 but adding the extra term $-MD_\Phi(z, x_t)$ term we get

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

Then we apply Claim 15 with $\gamma_t = 1$ to get the desired inequality. □

The next step in the proof of the logarithmic regret bound is to sum Claim 20 over

t, yielding

$$\begin{aligned}
&\sum_{t=1}^{T} \big(f_t(x_t) - f_t(z)\big) \\
&\leq \sum_{t=1}^{T} \frac{D_\Phi(x_t, w_{t+1})}{\eta_t} + \sum_{t=2}^{T} \left(\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) D_\Phi(z, x_t) - M D_\Phi(z, x_t) \right) \\
&+ \frac{1}{\eta_1} D_\Phi(z, x_1) - \frac{1}{\eta_T} D_\Phi(z, x_{T+1}) - M D_\Phi(z, x_1), \qquad \text{(by Claim 20)} \\
&\leq \sum_{t=1}^{T} \frac{D_\Phi(x_t, w_{t+1})}{\eta_t} + \sum_{t=2}^{T} \left(\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) D_\Phi(z, x_t) - M D_\Phi(z, x_t) \right). \qquad (\eta_1 = 1/M)
\end{aligned}$$

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

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

We have already shown that $D_\Phi(x_t, w_{t+1}) \leq \frac{\eta_t^2 L^2}{2}$ in (4.2), so

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

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

4.3 Sublinear Regret for DS-OMD with Extra Regularization

Following the notation from Chapter 3.3, we let $\Psi\colon \mathscr{X} \to \mathbb{R}_+$ denote the extra regularizer, a nonnegative 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 3 is to change the projection step of the algorithm to

$$x_{t+1} = \operatorname*{arg\,min}_{x \in \mathbb{R}^n} \big(D_\Phi\big(x, y_{t+1}\big) + \eta_{t+1} \Psi(x) \big). \tag{4.3}$$

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

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

In the remainder of this section we denote by $\Pi^{\Phi}_{\eta_{t+1}\Psi}(y_{t+1})$ the point computed by the right-hand side of (4.3). If we pick this projection coefficient α_t carefully, we can get $O(\sqrt{T})$ regret, as specified by the next theorem.

Theorem 21. *Let $\{x_t\}_{t\geq 1}$ be given as in Algorithm 3 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 $z \in \mathscr{X}$ and $K \in \mathbb{R}$ be such that $K \geq D_\Phi(z, x_1)$. Then,*

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

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

The analysis hinges on the following generalization of [6, Lemma 4.1], which can be thought as a "pythagorean Theorem" for Bregman projections.

Lemma 22. *Let $x \in \mathbb{R}^n$, $y \in \mathscr{D}^\circ$, and set $\bar{y} := \Pi^{\Phi}_{\alpha_t\Psi}(y)$. If $\bar{y} \in \mathscr{D}^\circ$, 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 22. 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 (4)) 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. □

We are now ready to prove Theorem 21.

Proof of Theorem 21. To prove the theorem, we just need to show that Theorem 17 still holds (with respect to the composite regret) in the algorithm with composite projections. We modify Claims 14 and 15 to get the following claim.

Claim 23.

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

Proof of Claim 23. Claim 14 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 15 to bound the right side of the above inequality. Using Lemma 22, 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 15 in Fang et al. [9] with the above inequality plus the extra regularization term and Claim 23 follows. □

Now the regret is bounded by

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

The first inequality follows Claim 23 and the last step comes from the assumption that x_1 is the minimizer of Ψ. This shows Theorem 17 holds as desired and then the proof of Theorem 21 follows as in Chapter 4.1. □

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

Corollary 24. *Consider a convex function f and let x^* be a minimizer of f. Let Φ be a differentiable strictly convex mirror map such that $\mathscr{X}\subseteq\mathscr{D}^\circ$. Assume that f is L-Lipschitz continuous to Φ and there exists non-negative K such that $K\ge D_\Phi(x^*,x_1)$. Let $\{\eta_t\}_{t\ge1}$ 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*

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

Chapter 5

Conclusion and Future Work

5.1 Conslusion

In this thesis we begin with an introduction of OCO framework and discussion of the upper bounds of two main OCO algorithms, FTRL and OMD. We then extend the classical sublinear regret bound to the more generalized relative setting proposed by Lu [15]. The results hold in the *anytime setting* in which we do not know the number of rounds/iterations beforehand. The main contribution of this work is that 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 novel 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.

5.2 Future Work

There are a few interesting directions of future work. The first would be to investigate the connections among the different notions of relative smoothness, Lipschitz continuity, and strong convexity in the literature. By doing this, we can acquire deeper understanding of these conditions giving favourable regret bounds.

Another direction is to investigate systematic ways of choosing a regularizer/mirror map for any given optimization problem. This could lead to novel algorithms of important applications.

Bibliography

[1] K. Antonakopoulos, E. V. Belmega, and P. Mertikopoulos. Online and stochastic optimization beyond lipschitz continuity: A riemannian approach. 2020. → pages 15, 16

[2] 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. → pages 13, 14

[3] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003. → page 8

[4] A. Ben-Tal and A. Nemirovski. *Lectures on modern convex optimization*. MPS/SIAM Series on Optimization. Society for Industrial and Applied Mathematics (SIAM), 2001. → page 5

[5] 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. → page 14

[6] S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015. → pages 3, 8, 9, 10, 12, 31, 33, 37

[7] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *COLT 2010*, pages 14–26. Omnipress, 2010. → page 26

[8] 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. → page 2

[9] 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. → pages 31, 32, 33, 38

[10] T. Gao, S. Lu, J. Liu, and C. Chu. Randomized bregman coordinate descent methods for non-lipschitz optimization. *arXiv preprint arXiv:2001.05202*, 2020. → page 16

[11] B. Grimmer. Convergence rates for deterministic and stochastic subgradient methods without lipschitz continuity. *SIAM Journal on Optimization*, 29(2): 1350–1365, 2019. → page 16

[12] E. Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016. URL http://ocobook.cs.princeton.edu/OCObook.pdf. → pages 2, 8, 23

[13] E. Hazan. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*, 2019. → pages 2, 10

[14] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007. → pages 8, 23

[15] H. Lu. "Relative continuity" for non-lipschitz nonsmooth convex optimization using stochastic (or deterministic) mirror descent. *Informs Journal on Optimization*, pages 265–352, 2019. → pages 15, 34, 40

[16] 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. → page 14

[17] C. J. Maddison, D. Paulin, Y. W. Teh, B. O'Donoghue, and A. Doucet. Hamiltonian descent methods. *arXiv preprint arXiv:1809.05042*, 2018. → page 16

[18] H. B. McMahan. A survey of algorithms and analysis for adaptive online learning. *The Journal of Machine Learning Research*, 18(1):3117–3166, 2017. → pages 6, 17, 29

[19] 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. → page 14

[20] A. S. Nemirovsky and D. B. Yudin. Problem complexity and method efficiency in optimization. 1983. → page 8

[21] Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009. → pages 14, 23

[22] F. Orabona and D. Pál. Scale-free online learning. *Theoretical Computer Science*, 716:50–69, 2018. → pages 8, 31

[23] 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. → page 12

[24] S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2011. → pages 2, 5, 7

[25] Q. Van Nguyen. Forward-backward splitting with bregman distances. *Vietnam Journal of Mathematics*, 45(3):519–539, 2017. → page 13

[26] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(Oct): 2543–2596, 2010. → pages 26, 29

Appendix A

Supporting Materials

A.1 Arithmetic Inequalities

Lemma 25. *Let* $\{a_t\}_{t\geq 1}$ *be a non-negative sequence with* $a_1 > 0$*. Then,*

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

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

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

□