# Adapting to Non-stationarity in Online Learning

by

Andrew Jacobsen

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Computing Science

University of Alberta

## Abstract

Over the last decade, machine learning (ML) has lead to advances in many fields, such as computer vision, online decision-making, robotics, natural language processing, and many others. The algorithms driving these successes typically have one or more user-specified free variables called *hyperparameters*, or simply *parameters*, which must be set prior to running the algorithm. These parameters can have a significant effect on an algorithm's performance in practice, and setting them optimally requires problem-dependent knowledge that the practitioner does not typically have access to, such as statistics of the underlying data-generating process. Common practice is to empirically "tune" the hyperparameters for a specific problem setting by repeatedly guessing values, observing the resulting performance, and adjusting the hyperparameter values accordingly. However, this practice is ultimately heuristic in nature and generally fails to provide meaningful performance guarantees, especially if the problem may change or drift over time.

The aim of this thesis is to develop learning algorithms which make meaningful performance guarantees in the absence prior knowledge, obviating the need to tune hyperparameters entirely. We focus in particular on developing algorithms which are suitable for *non-stationary* problem settings, in which the unknown problem solution may change arbitrarily over time. We study this problem through the lens of online learning, a framework used to model learning from a stream of data. We present the first online learning algorithms that achieve optimal performance guarantees in the complete absence of prior knowledge about the problem solution, even if it changes over time. We achieve this feat in the standard setting of Lipschitz losses, as well as under a relaxation of the Lipschitz condition which allows for unbounded losses, leading to novel results for stationary problem settings and saddle-point optimization as well. Our efforts culminate in a *universal* algorithm for online linear regression, which requires *no prior knowledge of any kind* to make optimal performance guarantees, even in the face of non-stationary data.

# Preface

This thesis is based off of a series of publications developed in collaboration with Ashok Cutkosky. Ashok assisted with editing each of the papers, occasional bug fixes, and contributed intellectually via weekly discussions. Everything included in this thesis is otherwise my own original work.

Chapters 4 and 6, and Section 9.1, as well as their corresponding appendices, are based on Jacobsen and Cutkosky (2022). The mirror descent equality in Appendix A.1 is a straight-forward generalization of the one from Jacobsen and Cutkosky (2022) to account for an additional "post-hoc adjustment" that we occasionally leverage. Chapter 7, Section 9.2, and their corresponding appendices, are based on Jacobsen and Cutkosky (2023). Chapter 10 and its appendices are based on Jacobsen and Cutkosky (2024).

*Don't quote me in your PhD thesis.*

– Cameron Linke, 2019.

# Acknowledgements

I'd like to thank Martha White for the incredible level of freedom I was entrusted with during my PhD. This freedom allowed me the time and space to approach things on precisely my own terms, and to develop into an independent researcher with unique perspectives in my field. I'd also like to thank Martha White, Adam White, Martin Mueller, and Alex Brown for all of the guidance during the early stages of my academic career.

There are many incredible people in the University of Alberta community who have contributed to the researcher that I developed into today by inspiring or helping me in one way or another. Some of these people include: Csaba Szepesevari, Matthew Schlegel, Andy Patterson, Cam Linke, Roshan Shariff, and Kris De Asis.

Most of all, I owe a huge debt of gratitude to Ashok Cutkosky. Ashok had absolutely no good reason to work with me when we met. I was not some promising up-and-comer in the online learning community and have never had a particularly strong background in rigorous mathematics. I had no publications in online learning, and my knowledge of online learning was completely self-taught and full of holes. I was literally just some random guy who started emailing him and was excited about the work he was publishing. Despite this, Ashok agreed to let me visit his lab for a summer, and the guidance I received during this time completely changed the trajectory of my academic career, enabling me to develop a unique and versatile approach to online learning which lead to the the exciting results that make up this thesis. I now confidently consider my self an expert in this topic, and I don't think it would have ever happened were it not for this initial act of kindness and blind faith from Ashok. I hope that I can one day pay this forward to my own students, and have the same impact on their development as researchers that Ashok has had on mine.

# Contents

# Chapter 1

# Introduction

Over the last decade, machine learning (ML) has lead to advances in many fields, such as computer vision (LeCun, Bengio, and Hinton 2015), online decision-making (Mnih et al. 2015; Silver et al. 2016; Abbeel et al. 2007; Ng et al. 2006), robotics (Lillicrap et al. 2015), natural language processing (Bahdanau, Cho, and Bengio 2014), and many others. The algorithms driving the successes in ML typically have one or more user-specified free variables called *hyperparameters*, or simply *parameters*, which must be set prior to running the algorithm. These parameters can have a significant effect on an algorithm's performance in practice, and setting them optimally generally requires problem-dependent *a priori* knowledge that the practitioner does not have access to. In practice, it is common to empirically "tune" the hyperparameters for a specific problem setting by repeatedly guessing values, observing the resulting performance, and adjusting the hyperparameters accordingly. However, this practice is ultimately *heuristic* in nature — there are typically no guarantees that a performant hyperparameter setting will be identified, or even what parameter ranges one should search over. Even if one chooses to accept these heuristic tuning procedures as a necessary evil, the use of free hyperparameters in algorithm design leads to several significant impediments to progress.

First, hyperparameters tend to exhibit ***high sensitivity to problem-dependent quantities*** — that is, the hyperparameters may have to be re-tuned if certain aspects of the problem change. This is a major stumbling block for the application of ML in real-world settings, as real-world problems can change over time. For example, an autonomous robot operating in the real world will be subject to the daily wear-and-tear of its hardware components; this can result in inconsistencies in the measurements used to inform decision-making, and as a result may require different hyperparameter settings to compensate for this additional uncertainty in its measurements. More generally, environmental changes can occur suddenly and without warning in the real world, requiring that the agent be able to adapt to these new conditions without being given the opportunity to re-calibrate its hyperparameters. In any such situations, traditional algorithms could exhibit unexpected or even unsafe behavior in deployment if the new conditions are sufficiently different from those expected *a priori* by the human designer.

Second, even in more "well-behaved" problem settings, one is still left to face the reality that ***parameter tuning is often infeasible in real-world domains***. Unlike in simulation domains, where the learner can experience a large number of examples in short periods of time, real-world applications are often limited by physical constraints. In applications such as robotics, for example, actions can take orders of magnitude longer to execute than they would in simulation domains since the actions correspond to real, physical movements. This can make thorough hyperparameter tuning impossible or prohibitively expensive in practice.

Finally, in addition to being a poor use of a highly-trained expert's time, ***this tuning process can be incredibly inefficient and expensive***. Each day *thousands* of hours of computation is spent by researchers and practitioners tuning these hyperparameters. Experiments can be run hundreds of times tuning the parameters of a *single algorithm*, and a thorough experiment typically requires tuning *multiple* hyperparameters of *multiple* algorithms. Not only is this a wasteful use of high-performance computing (HPC) facilities, which come at a high cost to operate and maintain, but it raises valid concerns about the environmental sustainability of ML research. Indeed, HPC facilities require massive amounts of energy resources to operate, and power is in fact one of the main operational expenses of these facilities (Couchman et al. 2015). Thus, any progress in reducing the need for such excessive parameter tuning could have significant impact on the efficiency, the cost, and the sustainability of research in ML.

This thesis is dedicated to the design of algorithms that achieve *provable* performance guarantees under minimal assumptions/prior knowledge, *without tuning any hyperparameters, whatsoever.* A major focus of this work is to develop algorithms which achieve these goals even in the face of problems in which may change in arbitrary and unpredictable ways over time, a property which we will broadly refer to as *non-stationarity.* This is in itself a very challenging type of adaptivity to achieve, and requires that we develop exceptionally strong tools and methodologies to achieve our goal. As such, as a result of our development we are able to make several substantial contributions advancing the state-of-the-art in online learning for both stationary and non-stationary problem settings alike.

## 1.1 Outline and Contributions

The remainder of this document is organized as follows.

**Part I: Foundations.** This thesis formalizes learning in the online learning framework, which is a framework for designing and analyzing algorithms that learn *incrementally* from a stream of data. In Part I we introduce the framework and provide a brief overview of of the hyperparameter-free algorithms that have emerged from this framework in recent years (Chapter 2), as well as common strategies for designing these algorithms (Section 2.2.1). In Chapter 4 we introduce our own approach and the algorithmic framework that will be used to design every algorithm in this thesis: the Centered Mirror Descent framework (Jacobsen and Cutkosky 2022).

**Part II: Adaptivity in Stationary Settings.** Part II is dedicated to hyperparameter-free learning in stationary settings. As a warm-up, in Chapter 6 we first apply our framework from Chapter 4 to design several new parameter-free algorithms in settings with *bounded gradients* (Lipschitz losses), achieving several novel results (Sections 6.3 and 6.4) and improving the guarantees of existing approaches (Sections 6.1 and 6.2).

In Chapter 7 we turn our attention to a problem setting in which the losses and gradients may potentially be *unbounded*. We provide the first parameter-free algorithms for online learning which achieve meaningful regret guarantees for non-Lipschitz losses in unbounded domains. We also provide a matching lower bound demonstrating that our result is unimprovable without further assumptions. Then, in Section 7.2 we use this approach to provide the first parameter-free algorithms for saddle-point optimization which converge in duality-gap without assuming strong convexity or bounded decision sets. This result provides as a special case algorithms for *bilinearly-coupled* saddle-point problems, which capture many notable problem settings, such as off-policy policy evaluation in reinforcement learning, quadratic games, and regularized empirical risk minimization (Du et al. 2022).

**Part III: Adapting to Non-stationarity.** The final part of the thesis is dedicated to hyperparameter-free learning in the face of non-stationarity. In Chapter 9, we provide the first algorithms for online learning which achieve meaningful guarantees in the absence of *all* assumptions on the problem "solution". In particular, our algorithms automatically adapt to notions of complexity of *any* benchmark sequence of decisions, which may be stationary, non-stationary, and could at any point be arbitrarily "far away" from the learner's own decisions. We begin by providing an algorithm for the setting of Lipschitz losses in Section 9.1. Then, as in Part II, we extend our result to the unbounded loss setting, and provide a matching lower bound for this new setting (Section 9.2).

In Chapter 10, we shift our focus to the related problem setting of online linear regression. In this setting, we develop algorithms that are not only hyperparameter free but *universal*: they utilize no instance-dependent prior knowledge *of any kind* yet still automatically adapt to natural notions of "difficulty" of any given problem instance without any hyperparameter tuning. We provide a matching lower bound demonstrating that our result is unimprovable in general. We also provide a simple extension of our result which makes a matching guarantee on *all intervals of time simultaneously.* Our result is the first instance of such "all-intervals" guarantees (called *strongly-adaptive* guarantees in the online learning literature) being achieved without any boundedness assumptions.

## 1.2 Notations

Throughout this document we will use the following common notations. We denote $[N] = \{1, \ldots, N\}$, $\mathbb{N} = \{0, 1, \ldots\}$, and $\mathbf{1}_N$ is the $N$-dimensional vector of ones. The $N$-dimensional simplex is denoted $\Delta_N$. The indicator function $\mathbb{I}_W(\cdot) = \mathbb{I}(\cdot \in W)$ is the function such that $\mathbb{I}_W(w) = 0$ if $w \in W$ and $\mathbb{I}_W(w) = \infty$ otherwise. For any sequence $a_1, a_2, \ldots$, we denote $a_{\max} = \max_t |a_t|$. Positive thresholding is denoted by $[\cdot]_+ = \max\{\cdot, 0\}$. We denote $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. We use the shorthand $\text{Clip}_{[a,b]}(y) = (y \vee a) \wedge b$ and the compressed sum notations $g_{i:j} = \sum_{t=i}^{j} g_t$ and $\|g\|_{a:b}^2 = \sum_{t=a}^{b} \|g_t\|^2$. For brevity, we occasionally abuse notation by letting $\nabla f(x)$ denote an arbitrary element of $\partial f(x)$. The Bregman divergence *w.r.t.* a differentiable function $\psi$ is $D_\psi(x|y) = \psi(x) - \psi(y) - \langle \nabla \psi(y), x - y \rangle$. Given a positive definite matrix $M$, the weighted norm *w.r.t.* $M$ is $\|w\|_M = \sqrt{\langle w, Mw \rangle}$. When unspecified, $\|\cdot\|$ is assumed to be the Euclidean norm. The notation $O(\cdot)$ hides constants, $\widehat{O}(\cdot)$ hides constants and log(log) terms, and $\widetilde{O}(\cdot)$ hides up-to log factors.

# Part I

# Foundations

# Chapter 2

# Online Learning

This thesis studies learning through the lens of the online learning framework, which is an elegant framework for analyzing and designing algorithms which learn *incrementally* from a stream of data (Zinkevich 2003; Nicolo Cesa-Bianchi and Lugosi 2006; Shalev-Shwartz and Singer 2007; McMahan 2017; Orabona 2019). This chapter provides a broad overview of the results and techniques in online learning, and reviews some of the ubiquitous design philosophies used in problems of this nature.

In the online learning framework, learning is formalized as a game played between a learner and an adversary (sometimes alternatively referred to as *nature* or *the environment*). On each round of the game, the learner makes a choice $w_t \in W$ from some convex *decision set* $W \subseteq \mathbb{R}^d$, then the adversary reveals a *loss function* $\ell_t : W \to \mathbb{R}$ and the learner pays a penalty of $\ell_t(w_t)$.

**Algorithm 1:** Online Learning Protocol

```
for $t = 1 : T$ do
    Learner plays $w_t \in W \subseteq \mathbb{R}^d$
    Adversary reveals loss function $\ell_t : W \to \mathbb{R}$
    Learner suffers a loss of $\ell_t(w_t)$
end
```

The standard performance metric in this setting is *regret* — the total loss of the learner compared against the total loss of some fixed *benchmark* $u \in W$, called the *comparator*:

$$R_T(u) = \sum_{t=1}^{T} \ell_t(w_t) - \ell_t(u).$$

Intuitively, one may think of the benchmark $u \in W$ as being a "batch" solution: if the losses $\ell_1, \ldots, \ell_T$ were given up front, one could choose a point of best-fit to make the function $w \mapsto \frac{1}{T}\sum_{t=1}^{T} \ell_t(w)$ small. Hence, we are typically interested in online algorithms which guarantee *sublinear* regret

($\lim_{T\to\infty} R_T(u)/T = 0$), as they perform *as well on average* as being able to choose with perfect hindsight.

The benefit of the preceeding formulation is that we avoid making any particular assumptions on the process generating the loss functions until they're relevant or necessary, yet we are still able to model a wide variety of problems by introducing constraints on the loss functions and adversary. In this thesis, a consistent limitation we'll impose is that the losses $\ell_t$ are *convex.*

**Definition 2.0.1.** Let $W$ be a convex subset of a real vector space. Then $\ell : W \to \mathbb{R}$ is convex if

$$\ell(\alpha x + (1-\alpha)y) \le \alpha\ell(x) + (1-\alpha)\ell(y)$$

for any $x, y \in W$ and $\alpha \in [0, 1]$.

For our purposes, the important property possessed by convex functions is that they can be lower-bounded by a first-order approximation given by its *subgradients.*

**Definition 2.0.2.** Let $\ell : W \to \mathbb{R}$ be a convex function. A subgradient of $\ell$ at $x \in W$ is any vector $g \in W^*$ satisfying

$$\ell(x) \ge \ell(y) + \langle g, x - y\rangle, \quad \forall y \in W$$

The set of all subgradients at $x \in W$ is the *subdifferential* of $\ell$ at $x$ and is denoted $\partial\ell(x)$. This property lets us bound the regret above by the regret against *linear* losses: letting $g_t \in \partial\ell_t(w_t)$ for all $t$, we can write

$$R_T(u) = \sum_{t=1}^{T} \ell_t(w_t) - \ell_t(u) \le \sum_{t=1}^{T} \langle g_t, w_t - u\rangle . \tag{2.1}$$

In this way, any algorithm $\mathcal{A}$ guaranteeing sublinear regret on *linear* losses can be used to achieve sublinear regret on *convex* losses as well — we simply choose any $g_t \in \partial\ell_t(w_t)$ and pass $\mathcal{A}$ the linearized losses $x \mapsto \langle g_t, x\rangle$. This reduction lets us focus our attention on designing algorithms for *online linear optimization.* Throughout this thesis, we let will frequently use $g_t$ to denote an arbitrary element of $\partial\ell_t(w_t)$. We will also occasionally write $\nabla\ell_t(w_t)$ to denote an arbitrary element of $\partial\ell_t(w_t)$ when appropriate (*e.g.*, to emphasize the dependence on $w_t$).

The second common assumption we will make is that the losses $\ell_t$ have *bounded subgradients* (*i.e.* the losses are Lipschitz).

**Definition 2.0.3.** A function $\ell : W \to \mathbb{R}$ is $G$-Lipschitz *w.r.t.* $\|\cdot\|$ if for every $x, y \in W$,

$$|f(x) - f(y)| \le G\,\|x - y\| .$$

If $\ell$ is convex, then $\ell$ is $G$-Lipschitz if and only if for every $x \in W$ and $g \in \partial\ell(x)$, it holds that $\|g\| \le G$. For the rest of this chapter we will assume that the losses are $G$-Lipschitz for simplicity, but we will relax this assumption later in Chapter 7.

Finally, we will occasionally consider losses which are *smooth*, meaning that they can be upper-bounded by a quadratic approximation.

**Definition 2.0.4.** A convex function $\ell : W \to \mathbb{R}$ is $L$-smooth *w.r.t.* $\|\cdot\|$ if for every $x, y \in W$,

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

For a detailed introduction to properties of smooth losses, we recommend X. Zhou (2018).

## 2.1 Minimizing Regret

To help build intuitions and get a feel for what kinds of guarantees we should expect in this problem setting, let us review the regret guarantee of the quintessential online learning algorithm: online (sub)gradient descent. Let $W$ be a convex set in $\mathbb{R}^d$, and consider linear losses $\ell_t(w) = \langle g_t, w\rangle$. Starting from $w_1 \in W$, on each round update

$$w_{t+1} = \Pi_W(w_t - \eta g_t),$$

where $\eta > 0$ and $\Pi_W(x) = \arg\min_{w \in W} \|w - x\|$ is the projection of $x$ onto $W$. Let us assume for simplicity that $W = \mathbb{R}^d$, so that $w_{t+1} = w_t - \eta g_t$ on each round. Now, for any $u \in W$ we may begin by investigating how $w_{t+1}$ relates to $u$ over time: in particular, observe that

$$\|w_{t+1} - u\|^2 = \|w_t - \eta g_t - u\|^2 = \|w_t - u\|^2 + \eta^2\|g_t\|^2 - 2\eta\langle g_t, w_t - u\rangle ,$$

hence, re-arranging, we have

$$\langle g_t, w_t - u\rangle = \frac{\|w_t - u\|^2 - \|w_{t+1} - u\|^2}{2\eta} + \frac{\eta}{2}\|g_t\|^2 ,$$

so summing over $t \in [T]$ we find that the regret is precisely

$$\begin{aligned} R_T(u) = \sum_{t=1}^T \langle g_t, w_t - u \rangle &= \sum_{t=1}^T \frac{\|w_t - u\|^2 - \|w_{t+1} - u\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^T \|g_t\|^2 \\ &= \frac{\|u - w_1\|^2 - \|u - w_{T+1}\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^T \|g_t\|^2 \\ &\le \frac{\|u - w_1\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^T \|g_t\|^2 . \end{aligned}$$

The step-size $\eta$ which minimizes the RHS above is $\eta = \frac{\|u - w_1\|}{\sqrt{\sum_{t=1}^T \|g_t\|^2}}$, which would yield

$$R_T(u) \le \|u - w_1\| \sqrt{\sum_{t=1}^T \|g_t\|^2}.$$

Moreover, a modest generalization of this argument shows that the same result also holds in constrained settings, where $W \subset \mathbb{R}^d$. The proof is standard in the literature (see, *e.g.*, Orabona 2019, Theorem 2.13).

**Proposition 2.1.1.** *Let $\ell_1, \ldots, \ell_T$ be arbitrary convex loss functions. Let $W$ be a convex set in $\mathbb{R}^d$, $w_1 \in W$, and set $w_{t+1} = \Pi_W(w_t - \eta g_t)$ for some $\eta > 0$ and $g_t \in \partial \ell_t(w_t)$. Then for any $u \in W$,*

$$R_T(u) \le \frac{\|w_1 - u\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^T \|g_t\|^2 .$$

*Moreover, setting $\eta = \eta^* := \frac{\|w_1 - u\|}{\sqrt{\sum_{t=1}^T \|g_t\|^2}}$ guarantees*

$$R_T(u) \le \|w_1 - u\| \sqrt{\sum_{t=1}^T \|g_t\|^2}. \tag{2.2}$$

This result has several desirable features:

1. It is essentially tight (nearly a regret *equality* in the unconstrained OLO setting, in fact!), so we can not hope to get any significantly tighter result than Equation (2.2). Indeed, as we will see in Section 2.2, this bound is actually *too* good to be achieved without access to rather strong prior knowledge.

2. It is *worst-case optimal*: in a domain $W$ with $\sup_{x,y \in W} \|x - y\| \le D$, it can be shown that any algorithm must incur at least $R_T(u) \ge \Omega(DG\sqrt{T})$ regret in the worst case (See *e.g.* Orabona and Pál (2018) Theorem 5; Orabona (2019) Theorem 5.1; Hazan (2019) Theorem 3.2). Equation (2.2) always matches this lower bound in the worst-case.

3. It is *data-dependent*: the bound automatically adapts to the "easiness" of the problem. If the gradients are "small" or the learner has sufficient prior knowledge to be able to choose $w_1$ reasonably "close" to $u$, regret can automatically be much smaller than the worst-case bound.

4. It holds even in *unbounded domains*[1]: nowhere in the analysis is it necessary to assume that the decision set is bounded (*e.g.* $\sup_{x,y\in W} \|x-y\| \le D$ for some $D$). Instead, the bound is *adaptive* to the initial distance from the comparator $\|u-w_1\|$.

5. It is *dimension-free*: there is no explicit penalty related to the dimension of the space $W$, so the performance naturally scales to high-dimensional data.

Of course, we can not actually implement this algorithm in practice — setting $\eta^* = \frac{\|w_1-u\|}{\sqrt{\sum_{t=1}^T \|g_t\|^2}}$ would require *a priori* knowledge of the comparator $u$ as well as all of the future subgradients $g_t$. Because of this, in practice $\eta$ is left as a free variable referred to as a *hyperparameter*, and its value is empirically "tuned" by running the algorithm with many different values for $\eta$ and measuring the performance. Not only is this an error-prone and computationally expensive process, but we will also tend to lose the worst-case robustness that made the algorithm interesting in the first place! Instead, over the last decade there has been a concerted effort to develop algorithms which *adapt* to these unknown quantities on-the-fly. In the following section we will review basic results and strategies related to these adaptive online learning algorithms.

## 2.2 Adaptivity in Online Learning

Motivated by the discussion in the previous section, a natural question is whether it's possible to achieve regret of the form $R_T(u) \le O\big(\|w_1-u\|\sqrt{\sum_{t=1}^T \|g_t\|^2}\big)$, *without* prescient knowledge of $g_1,\ldots,g_T$ or the comparator $u\in W$. In other words, is it possible to *adapt* to the unknown quantities $\|w_1-u\|$ and $\sqrt{\sum_{t=1}^T \|g_t\|^2}$ *on-the-fly*, without tuning hyperparameters. It turns out there are several different frontiers of adaptivity, characterized by different kinds of prior knowledge that the learner might have access to: typically that the subgradients are uniformly bounded by some $\mathfrak{G}_T \ge \max_t \|g_t\|$, that the losses map to a bounded range $\ell_t : W \to [a,b]$, that the learner has prior knowledge of a $D \ge \|u-w_1\|$ (usually by assuming that the domain is bounded with $D \ge \sup_{x,y\in W} \|x-y\|$), or some combination of these conditions.

In the simplest case, when one has access to both a bound $D \ge \|u-w_1\|$ and a bound on the subgradients $\mathfrak{G}_T \ge \|g_t\|$ for all $t$, it is well-known that adaptivity to $\sqrt{\sum_{t=1}^T \|g_t\|^2}$ can be achieved up to constant factors by simply using the best approximation to the optimal $\eta^* = \|u-w_1\|/\sqrt{\sum_{t=1}^T \|g_t\|^2}$

[1] Note that constrained does not imply bounded, so this point is relevant even in the constrained setting. To see why, consider an algorithm with domain $W = \{(x,y)\in\mathbb{R}^2 : y \ge x^2\}$, *i.e.*, the domain is the epigraph of a parabola. This is a convex, constrained domain in which $\sup_{x,y\in W} \|x-y\| = \infty$.

that one has access to on each round: $\eta_t = D/\sqrt{\mathfrak{G}_T^2 + \sum_{s=1}^{t-1}\|g_s\|^2}$.[2] This is the essence of the AdaGrad algorithm (McMahan and M. J. Streeter 2010; J. Duchi, Hazan, and Singer 2011). When $D$ is available but not the Lipschitz bound $\mathfrak{G}_T$, it is still possible to match this guarantee up to constant factors (for instance, by instead setting $\eta_t = D/\sqrt{\sum_{s=1}^{t-1}\|g_t\|^2}$), in which case the algorithm is said to be *Lipschitz adaptive* or *scale-free* (Orabona and Pál 2018; Mayo, Hadiji, and Erven 2022; Cutkosky 2019a).[3] When the losses are $L$-smooth, these bounds can be improved to $R_T(u) \le O\Big(LD^2 + D\sqrt{L\sum_{t=1}^T \ell_t(u)}\Big)$ — referred to as an $L^*$ bound or *small loss* bound — though the works which achieve these guarantees still require one or more of the following assumptions: prior knowledge of $\mathfrak{G}_T$, that $\ell_t$ has bounded range (known in advance), prior knowledge of a lower bound $\ell_t^* \le \ell_t(w)$ for all $w \in W$, additional structural assumptions such as strong convexity or exp-concavity, or by assuming the losses take some specific form such as the square loss (Nicolo Cesa-Bianchi, Long, and Manfred K Warmuth 1996; Jyrki Kivinen and Manfred K Warmuth 1997; Srebro, Sridharan, and Tewari 2010; Orabona, Nicolo Cesa-Bianchi, and Gentile 2012).

If a bound $\mathfrak{G}_T \ge \max_t \|g_t\|$ is known but not the bound $D \ge \|u - w_1\|$, the situation gets significantly trickier. The essential difficulty is that without prior knowledge of how far we started from the comparator, the learner's iterates $w_t$ could at any point be arbitrarily far away from the benchmark $u$, leading to high regret. As such, the learner must take great care to control $\|w_t\|$ in such a way that it is *adaptive* to this unknown unknown initial distance comparator norm $\|u - w_1\|$. Without prior-knowledge of this gap, AdaGrad and its variants can never guarantee the optimal dependence on $\|u - w_1\|$, even with some clever hyperparameter tuning. In fact, *no algorithm* can guarantee regret $R_T(u) \le \|u - w_1\|\sqrt{\sum_{t=1}^T \|g_t\|^2}$ without prior knowledge of $u$: it turns out that in the setting of $\mathfrak{G}_T$-Lipschitz losses and unbounded $W$, the worst-case regret of any algorithm is at least

$$R_T(u) \ge \Omega\left(\|u - w_1\|\,\mathfrak{G}_T\sqrt{T\log\left(\|u - w_1\|\sqrt{T} + 1\right)}\right) \tag{2.3}$$

in the worst-case (Mcmahan and M. Streeter 2012, Theorem 7; Orabona 2013, Theorem 2). Hence, there is an additional cost of at least $\Omega\left(\sqrt{\log\left(\|w_1 - u\|\sqrt{T} + 1\right)}\right)$ associated with adaptivity to $\|u - w_1\|$. For instance, a standard result in this setting is

$$R_T(u) \le O\left(\|u - w_1\|\sqrt{\sum_{t=1}^T \|g_t\|^2 \log\left(\|u - w_1\|\sqrt{T} + 1\right)} + \mathfrak{G}_T\|u - w_1\|\log\left(\|u - w_1\|\sqrt{T} + 1\right)\right) \tag{2.4}$$

[2] Bounds which scale with the adaptive $\sqrt{\sum_{t=1}^T \|g_t\|^2}$ instead of the pessimistic $G\sqrt{T}$ are occasionally referred to as "second-order adaptive", owing to the squared dependence on the gradient norms.

[3] Note that *scale-free* is actually a stronger notion than just Lipschitz adaptivity, in that the regret of a scale-free algorithm depends *only* on $\max_t \|g_t\|$, rather than the potentially pessimistic upper-bound $\mathfrak{G}_T$. However, for our purposes drawing this distinction is not necessary since all of the Lipschitz adaptive algorithms we discuss in this thesis will also be scale-free.

which exhibits the adaptivity to both $\|u-w_1\|$ and $\sqrt{\sum_{t=1}^T \|g_t\|^2}$, and matches Equation (2.3) up to logarithmic terms (Mcmahan and M. Streeter 2012; McMahan and Orabona 2014; Orabona 2013; Orabona and Pál 2016; Cutkosky and Orabona 2018; Hoeven 2019). Algorithms which guarantee regret matching the lower bound up to logarithmic terms are called "comparator-adaptive", or "parameter-free",[4] owing to the fact that they optimally adapt to both unknown quantities $\|u-w_1\|$ and $\sqrt{\sum_{t=1}^T \|g_t\|^2}$ simultaneously on-the-fly, and hence require no offline hyperparameter tuning to achieve near-optimal regret. Some works also consider the weaker bounds of the form $R_T(u) \le O\left(\|u-w_1\| \mathfrak{G}_T \sqrt{T \log\left(\|u-w_1\|\sqrt{T}+1\right)}\right)$ to be parameter-free, allowing the $\sqrt{\sum_{t=1}^T \|g_t\|^2}$ term to degrade to the worst-case $\mathfrak{G}_T\sqrt{T}$. In either case, note that the key property that distinguishes the parameter-free bound is that the regret against $w_1$ is constant:

$$R_T(w_1) \le \widetilde{O}\left(\|w_1 - w_1\| \sqrt{\sum_{t=1}^T \|g_t\|^2}\right) = O(1).$$

The first results to avoid *both* the bounded domain and bounded gradient assumptions have only been achieved in recent years. Cutkosky (2019a) develops an algorithm which achieves

$$R_T(u) \le \widetilde{O}\left(\|u-w_1\| \sqrt{\sum_{t=1}^T \|g_t\|^2 \log\left(\|u-w_1\|\sqrt{T}+1\right)} + \mathfrak{G}_T \|u-w_1\|^3\right),$$

and Mhammedi and Koolen (2020) shows that the additional cubic penalty is unavoidable while maintaining a $\widetilde{O}\left(\|u-w_1\|\mathfrak{G}_T\sqrt{T}\right)$ dependence. Alternatively, Orabona and Pál (2018) show that $R_T(u) \le O(\|u-w_1\|^2 \sqrt{\sum_{t=1}^T \|g_t\|^2})$ can be attained without prior knowledge of $\mathfrak{G}_T$ in an unbounded domain, avoiding the cubic penalty in exchange for a horizon-dependent quadratic penalty. Works such as Mayo, Hadiji, and Erven (2022) and Kempka, Kotlowski, and Manfred K. Warmuth (2019) show that the cubic penalty can be avoided in certain special cases such as regression-type losses.

Note that adaptivity to $\|u-w_1\|$ is closely related to the problem of *unconstrained* online learning in general. The reason being that in the unconstrained setting, there is never a constant $D$ such that $D \ge \sup_{x,y\in W}\|x-y\|$, so bounds which scale with $\|u-w_1\|$ are the only real option. As such, throughout this thesis we will primarily focus on the unconstrained setting with $W = \mathbb{R}^d$ for simplicity, though the results presented here can be easily generalized to constrained settings by accepting some notational and proof bloat (see Remark A.1.2).

*Remark* 2.2.1. For brevity, we will frequently adopt the common convention that $w_1 = \mathbf{0}$ (particularly in Parts II and III of the thesis), in which case parameter-free regret is characterized by the property $R_T(\mathbf{0}) \le O(1)$ and features bounds scaling with $\|u\|$ instead of $\|u-w_1\|$. This assumption is without

[4] Note that throughout the machine learning literature, the decision variable $w_t \in W$ is often referred to as a "parameter vector", while in our context "parameter-free" refers to *hyper*parameters.

loss of generality since one could otherwise perform a translation of the coordinate system.

### 2.2.1 Principles for Adaptive Algorithm Design

In the previous section we saw there were two main types of adaptivity in the general online convex optimization setting: adaptivity to $\sqrt{\sum_{t=1}^T \|g_t\|^2}$ and adaptivity to $\|u - w_1\|$. The former is fairly straight-forward to understand: we observe $g_t$ after each round, so we can reasonably "approximate" the optimal step-size by running (sub)gradient descent with an adaptive step-size $\eta_t \propto 1/\sqrt{\sum_{s=1}^{t-1} \|g_s\|^2}$. The adaptivity to $\|u - w_1\|$ is much less obvious at a glance; we know that the bound we're shooting for must have an additional multiplicative penalty of $O\left(\sqrt{\log\left(\|u - w_1\|\sqrt{T} + 1\right)}\right)$ in it, but where does it come from? How do we design a strategy that achieves such a bound? In this section we provide some of the key insights and approaches to designing algorithms which attain bounds of the form Equation (2.4). Throughout our exposition here we will assume for simplicity that we are in the unconstrained setting with $W = \mathbb{R}^d$. The goal here is to provide some of the broad-strokes of the main approaches so that the reader has a high-level perspective on how to go about designing these algorithms. Seeing the approaches here will also later help illustrate the need for a new approach, which will be introduced in Chapter 4.

One of the foundational observations leading to comparator-adaptive guarantees is a certain *reward-regret duality*, which tells us that designing an algorithm which guarantees $R_T(u) \le B_T(u)$ for some function $B_T : W \to \mathbb{R}$ is equivalent to designing an algorithm which guarantees $-\sum_{t=1}^T \langle g_t, w_t\rangle \ge B_T^*\left(-\sum_{t=1}^T g_t\right)$, where $B_T^*$ is the Fenchel conjugate of $B_T$, defined as

$$B_T^*(\theta) = \sup_w \langle \theta, w\rangle - B_T(w).$$

In particular, to achieve the optimal parameter-free bound, we would want to consider $B_T(u) = O\left(\|u - w_1\|\sqrt{\sum_{t=1}^T \|g_t\|^2 \log\left(\|u - w_1\|\sqrt{T} + 1\right)}\right)$ and its corresponding Fenchel conjugate $B_T^*$. The following theorem is a standard starting point for many works which develop parameter-free bounds (McMahan and Orabona 2014; Orabona and Pál 2016; Cutkosky and Orabona 2018; Cutkosky and Sarlos 2019; Mhammedi and Koolen 2020; Hoeven 2019; Jun and Orabona 2019). We provide a basic proof for convenience to the reader.

**Theorem 2.2.2.** *Let $B_T : W \to \mathbb{R}$ be a convex function. An algorithm guarantees*

$$R_T(u) \le B_T(u), \qquad \forall u \in W$$

*if and only if it guarantees*

$$-\sum_{t=1}^T \langle g_t, w_t\rangle \ge B_T^*\left(-\sum_{t=1}^T g_t\right), \qquad \forall g_1, \dots, g_T. \tag{2.5}$$

*Proof.* Suppose that for any $u \in W$, $\mathcal{A}$ guarantees $R_T(u) = \sum_{t=1}^T \langle g_t, w_t - u\rangle \leq B_T(u)$. Then rearranging, we equivalently have

$$\left\langle -\sum_{t=1}^T g_t, u\right\rangle - B_T(u) \leq -\sum_{t=1}^T \langle g_t, w_t\rangle,$$

and since this holds for any $u \in W$, it must hold for the one which tightens the bound:

$$\sup_u \left\langle -\sum_{t=1}^T g_t, u\right\rangle - B_T(u) \leq -\sum_{t=1}^T \langle g_t, w_t\rangle, \quad \textit{i.e.,} \quad B_T^*\left(\sum_{t=1}^T g_t\right) \leq -\sum_{t=1}^T \langle g_t, w_t\rangle.$$

For the other direction, suppose that $-\sum_{t=1}^T \langle g_t, w_t\rangle \geq B_T^*\left(-\sum_{t=1}^T g_t\right)$. Then we immediately have that

$$R_T(u) = \sum_{t=1}^T \langle g_t, w_t\rangle - \sum_{t=1}^T \langle g_t, u\rangle \leq -B_T^*\left(-\sum_{t=1}^T g_t\right) + \left\langle -\sum_{t=1}^T g_t, u\right\rangle \leq \sup_\theta \langle \theta, u\rangle - B_T^*(\theta) = B_T(u).$$

□

The value of this theorem is that it has shown us an *equivalent* condition to $R_T(u) \leq B_T(u)$ which *does not depend on the unobserved quantity $u$*. That is, by considering instead the equivalent condition $-\sum_{t=1}^T \langle g_t, w_t\rangle \geq B_T^*\left(-\sum_{t=1}^T g_t\right)$ we remove the comparator completely from our objective. Moreover, the new condition depends only on the gradients, which we do eventually observe, making Equation (2.5) appealing from an algorithm design perspective. Let us consider a few common approaches for designing an algorithm which guarantees Equation (2.5).

**Potential-based Arguments.** The idea with this approach is as follows. We want to design an algorithm which guarantees $-\sum_{t=1}^T \langle g_t, w_t\rangle \geq B_T^*\left(-\sum_{t=1}^T g_t\right)$. To this end, let $B_1^*, \ldots, B_{T-1}^*$ be an arbitrary sequence of functions (which the designer will eventually choose) and define the "potential" at time $t$ to be $\Phi_t = \sum_{s=1}^t \langle g_s, w_s\rangle + B_t^*\left(-\sum_{s=1}^t g_s\right)$ and $\Phi_0 = 0$. Then, if we could ensure that this potential is *non-increasing* (via our choices of $w_t$ and $B_t^*$), we would have

$$\Phi_T = \sum_{t=1}^T \langle g_t, w_t\rangle + B_T^*\left(-\sum_{t=1}^T g_t\right) \leq \Phi_{T-1} \leq \ldots \leq \Phi_0 = 0$$

hence,

$$B_T^*\left(-\sum_{t=1}^T g_t\right) \leq -\sum_{t=1}^T \langle g_t, w_t\rangle,$$

and so via Theorem 2.2.2 we will have $R_T(u) \leq B_T(u)$. To ensure non-increasing potential, we need only select $w_t$ and $B_t^*$ in such a way that

$$\Phi_t - \Phi_{t-1} = \langle g_t, w_t \rangle + B_t^* \left(-g_{1:t}\right) - B_{t-1}^* \left(-g_{1:t-1}\right) \le 0$$

on each round. Unfortunately, this is often easier said than done, but there is nonetheless a clear sequence of steps that the algorithm designer can take: define the potential $\Phi_t$ and ensure the sequence is decreasing by choosing $w_t$ and $B_t^*$ appropriately. See Abernethy et al. 2014; Hoeven 2019; Cutkosky and Sarlos 2019; Kempka, Kotlowski, and Manfred K. Warmuth 2019; Mhammedi and Koolen 2020; Orabona and Pál 2021 for examples using a potential-based approach to designing adaptive algorithms.

The main issue with the potential-based approach is that it is in some sense *too* general, in that it does not restrict the designer enough. Because of this, the potential-based approach often often requires a good deal of cleverness on the part of the designer to choose the $B_t^*$ and and $w_t$ appropriately. The ideal framework from an algorithm design perspective should instead *naturally guide the designer* towards the right choices by introducing natural restrictions/limitations, pruning the space of possible design choices without significantly limiting the power of the framework. The following approaches can all be loosely considered to be particular restrictions of the general potential-based approach.

**Coin-Betting.** Coin-betting is arguably the most well-known framework for designing parameter-free algorithms, and can be seen as a particular form of the potential-based approach which introduces the restriction that the designer need only choose a "betting fraction" on each round, and prescribes a choice of $w_t$ based on this betting fraction (Orabona and Pál 2016; Cutkosky and Orabona 2018; Jun, Orabona, et al. 2017; Orabona 2019). The idea is that if we define $\text{Wealth}_t = -\sum_{s=1}^t \langle g_s, w_s \rangle$, then on any time $t$ if we set $w_t = \text{Wealth}_{t-1}\beta_t$ for some $\beta_t \in \left\{\beta \in \mathbb{R}^d : \|\beta\| < 1\right\}$, we'd have

$$\text{Wealth}_t = \text{Wealth}_{t-1} - \langle g_t, w_t \rangle = \text{Wealth}_{t-1} - \text{Wealth}_{t-1} \langle g_t, \beta_t \rangle = \text{Wealth}_{t-1} \left(1 - \langle g_t, \beta_t \rangle\right).$$

So suppose that the desired wealth bound holds at time $t-1$: $\text{Wealth}_{t-1} \ge B_{t-1}^* \left(-g_{1:t-1}\right)$. Then

$$\text{Wealth}_t = \text{Wealth}_{t-1} \left(1 - \langle g_t, \beta_t \rangle\right) \ge B_{t-1}^* \left(-g_{1:t-1}\right) \left(1 - \langle g_t, \beta_t \rangle\right),$$

so if we can set $\beta_t$ in such a way that $B_{t-1}^* \left(-g_{1:t-1}\right) \left(1 - \langle g_t, \beta_t \rangle\right) \ge B_t^* \left(-g_{1:t}\right)$, we'd have

$$\text{Wealth}_t = -\sum_{s=1}^t \langle g_s, w_s \rangle \ge B_t^* \left(-g_{1:t}\right).$$

Formalizing this more rigorously as an induction argument, one can use this approach to ensure that $\text{Wealth}_t \ge B_t^*(-g_{1:t})$ for any $t$, so that $\text{Wealth}_T = -\sum_{t=1}^T \langle g_t, w_t \rangle \ge B_T^*(-g_{1:T})$ and hence $R_T(u) \le$

$B_T(u)$ via Theorem 2.2.2. Orabona and Pál 2016 provides a set fairly general conditions on $B_t^*$ that are sufficient to ensure that the induction step goes through.

**Follow the Regularized Leader (FTRL).** Instead of explicitly using the reward-regret duality, throughout this thesis we will take an FTRL-based perspective. The key observation is that it's not actually necessary to go through any fancy reward-regret duality to get a comparator-adaptive guarantee; we can use the same FTRL tools that are ubiquitious throughout online learning. In particular, on each round FTRL chooses

$$w_t = \arg\min_{w \in W} \left\langle \sum_{\tau=1}^{t-1} g_\tau, w \right\rangle + \psi_t(w) =: \arg\min_{w \in W} F_t(w),$$

where $\psi_t : W \to \mathbb{R}$ is a convex regularizer. Then, via the well-known regret guarantee for FTRL (see, *e.g.*, Orabona 2019, Theorem 7.1), we have

$$R_T(u) \le \psi_T(u) + \sum_{t=1}^{T} \underbrace{F_t(w_t) - F_{t+1}(w_{t+1}) + \langle g_t, w_t \rangle}_{=:\delta_t}.$$

and so, if we would like to guarantee $R_T(u) \le B_T(u)$ up to constants, all we have to do is design a sequence of regularizers $\psi_1, \ldots, \psi_T$ such that $\psi_T(u) \approx B_T(u)$ and that the latter terms sum to a constant, $\sum_{t=1}^{T} \delta_t \le O(1)$.

**Mirror Descent.** A closely-related approach to FTRL is mirror descent. The typical mirror descent update is of the form

$$w_{t+1} = \arg\min_{w \in W} \langle g_t, w \rangle + D_{\psi_t}(w|w_t),$$

where $\psi_t : W \to \mathbb{R}$ is a convex regularizer and $D_{\psi_t}(x|y) = \psi_t(x) - \psi_t(y) - \langle \nabla\psi_t(y), x - y \rangle$ is the Bregman divergence *w.r.t.* $\psi_t$ between $x, y \in W$. Setting $\psi_t(w) = \frac{1}{2\eta}\|w\|^2$ leads to the standard (projected) subgradient descent update, so mirror descent can be seen as generalizing gradient descent to different parameter-space geometries, represented by different choices of $\psi_t$.

The design principles of mirror descent are similar to those of FTRL: we still want to choose the regularizers in such a way that $\psi_T(u) \approx B_T(u)$, while also ensuring certain stability terms sum to some small constant. However, as we will elaborate in Chapter 4, mirror descent on its own is not suitable for designing parameter-free algorithms. The issue is that mirror descent is fundamentally unstable in unbounded settings, and in particular it is possible to show that mirror descent with a time-varying regularizer can incur linear regret (Orabona and Pál 2018). In Chapter 4 we present a framework that employs a generalization of the mirror descent perspective to incorporate the stability properties of FTRL. We will then use this approach to design every algorithm featured in this thesis. First, we take a short detour to to introduce some natural notions of *non-stationarity* for the online learning setting, which will motivate the development of a new approach which goes beyond the limitations imposed by reward-regret duality and FTRL.

# Chapter 3

# Learning in Dynamic Environments

In the previous chapter we discussed a class of algorithms which adapt to both the unknown initial distance $\|w_1 - u\|$ and the gradients $\sqrt{\sum_{t=1}^T \|g_t\|^2}$ simultaneously to achieve an optimal adaptive regret bound of $R_T(u) \le O\Big( \|u - w_1\| \sqrt{\sum_{t=1}^T \|g_t\|^2 \log\left(\|u - w_1\| \sqrt{T} + 1\right)}\Big)$. However, like any measure of performance, regret is only meaningful insofar as it captures some notion of "goodness" that we actually care about. In many problems of practical interest, competing against any *fixed* comparator $u \in W$ can fail to be meaningful, particularly when modelling problems with a *time-varying* or *non-stationary* solution.

As a simple illustrative example, consider a simple 1-dimensional prediction problem in which the objective is to predict a response variable $y_t \in \mathbb{R}$ before it is observed. A simple way to model this problem is as an online learning problem with losses that capture prediction error, such as $\ell_t(w) = \frac{1}{2}(y_t - w)^2$. On one hand, if

$$y_t = \mu + \varepsilon_t,$$

for some $\mu \in \mathbb{R}$, and mean-zero noise $\varepsilon_t$, then clearly the fixed comparator $u = \mu$ would provide a meaningful performance baseline. On the other hand, suppose instead that the mean of $y_t$ is drifting over time, according to an unknown dynamical system:

$$\begin{aligned} \mu_t &= F_t(\mu_{t-1}) \\ y_t &= \mu_t + \varepsilon_t, \end{aligned}$$

where $F_t$ is an unknown transfer function and $\mu_0$ is arbitrary. In this case, even under relatively simple time-varying dynamics such as $F_t(\mu) = \mu + \delta_t$ for zero-mean noise $\delta_t$, no fixed comparator $u$ will provide reasonable predictions of $y_t$ across time; instead, our baseline ought to somehow "track" $y_t$ as its distribution changes over time. More generally, to meaningfully model learning in a dynamically changing environment, we need to consider stronger notions of regret which better capture the dynamic nature of the problem. In this chapter, we review the two main notions of non-stationarity studied in online learning, dynamic regret and strongly-adaptive regret.

## 3.1 Dynamic Regret

The most straight-forward way to strengthen the notion of regret is to instead measure performance relative to a *sequence* of comparators $\boldsymbol{u} = (u_1, \ldots, u_T)$, leading to *dynamic* regret:

$$R_T(\boldsymbol{u}) = \sum_{t=1}^{T} \ell_t(w_t) - \ell_t(u_t). \tag{3.1}$$

Dynamic regret is more appropriate for true streaming settings in which data might drift over time, wherein a fixed benchmark $u$ may be an excessively weak baseline. However, it is also more demanding than the previous definition (called *static* regret, when the distinction matters). Indeed, without further assumptions on the losses it can be shown that there exist sequences of losses and comparator for which the dynamic regret is vacuous in the worst-case, $R_T(\boldsymbol{u}) \geq \Omega(T)$ (T. Yang et al. 2016). Intuitively, the issue is that the comparators $\boldsymbol{u}$ can be chosen to "overfit" to the sequence of losses so that the benchmark performance $\sum_{t=1}^{T} \ell_t(u_t)$ is too difficult to compete against. Thus, ideally we ought to somehow distinguish such sequences from the ones which are actually useful for us to reason about.

Recall that in the static regret case, the optimal regret scaled with $\|u - w_1\|$. In some sense, we can think of this quantity as a measure of "complexity" of the benchmark point $u \in W$: difficult-to-compete-with benchmarks are those that are very different from our initial preconceptions, represented by a large $\|u - w_1\|$, while easier benchmarks are those which are close to $w_1$. Likewise, one might expect that dynamic regret should also account for some notion of of "complexity" of the comparator sequence, so as to account for the "difficulty" of the sequence we're up against. One natural measure of this complexity is the *path-length*:

$$P(\boldsymbol{u}) = \sum_{t=2}^{T} \|u_t - u_{t-1}\|. \tag{3.2}$$

When the comparator sequence is clear from context, we will use the short-hand $P_T = P(\boldsymbol{u})$.

In the setting of $G$-Lipschitz losses and a bounded domain of radius $D = \max_{x,y \in W} \|x - y\|$, T. Yang et al. 2016 show that in the specific case where the comparator sequence is the sequence of local minimizers, $\boldsymbol{u}^* = (u_1^*, \ldots, u_T^*)$ with $u_t^* = \arg\min_{w \in W} \ell_t(w)$, a simple greedy strategy that plays $w_t = \arg\min_{w \in W} \ell_{t-1}(w)$ guarantees $R_T(\boldsymbol{u}^*) \leq O(GP(\boldsymbol{u}^*))$, and prove a matching lowerbound. However, as alluded to above, we may not always care about this specific comparator sequence: in many instances the sequence $\boldsymbol{u}^*$ can "overfit" to the losses, resulting in small cumulative loss but a long path-length, making the $R_T(\boldsymbol{u}^*) \leq O(P(\boldsymbol{u}^*))$ bound vacuous. For instance, let $W = [-1, 1]$ and $\ell_t(w) = \frac{1}{2}(w - \varepsilon_t)^2$ for $\varepsilon_t = (-1)^t$, then the path-length is clearly $P(\boldsymbol{u}^*) = \sum_{t=2}^T |u_t - u_{t-1}| = \Omega(T)$ and so the greedy strategy incurs linear regret. On the other hand, were we instead to compare against the optimal *fixed* comparator $u_t = \frac{1}{T}\sum_{s=1}^T \varepsilon_s =: \bar{u}$ for all $t$, then dynamic regret reduces to static regret and on the very same problem projected gradient descent with step-size $\eta = \frac{1}{G\sqrt{T}}$ will guarantee sublinear regret $R_T(\bar{\boldsymbol{u}}) \leq O(G\sqrt{T})$.

What the above discussion suggests is that making meaningful dynamic regret guarantees involves competing against a sequence $\boldsymbol{u}$ which somehow favorably strikes a trade-off between the path-length $P_T$ and the cumulative loss $\sum_{t=1}^T \ell_t(u_t)$. Yet this trade-off inevitably depends on the specific sequence of losses $\ell_t$, which are unknown to the learner *a priori*. An ideal strategy should thus instead make guarantees *w.r.t. arbitrary sequences* $\boldsymbol{u}$, not just for specific sequences such as $\boldsymbol{u}^*$ or $\bar{\boldsymbol{u}}$.

Interestingly, the question of how to compete with an arbitrary sequence $\boldsymbol{u}$ was introduced all the way back in the seminal work of Zinkevich (2003), considered by many to be one of the first works to popularize the online learning framework. Zinkevich shows that in a bounded domain with $G$-Lipschitz losses, projected gradient descent with step-size $\eta$ guarantees

$$R_T(\boldsymbol{u}) \leq O\left(\frac{D^2 + DP_T}{2\eta} + \frac{\eta}{2}\sum_{t=1}^T \|g_t\|^2\right) \tag{3.3}$$

Hence, if $P_T$ and $\sum_{t=1}^T \|g_t\|^2$ were known *a priori*, the learner could set the step-size to $\eta^* = \sqrt{\frac{D^2 + DP_T}{\sum_{t=1}^T \|g_t\|^2}}$ to get

$$R_T(\boldsymbol{u}) \leq \sqrt{(D^2 + DP_T)\sum_{t=1}^T \|g_t\|^2}, \tag{3.4}$$

which was later shown to be optimal via a matching lowerbound of $R_T(u) \geq \Omega\left(G\sqrt{(D^2 + DP_T)T}\right)$ due to L. Zhang, S. Lu, and Z.-H. Zhou (2018).

Clearly one can not actually choose the step-size $\eta^*$ that yields this bound in practice, but it turns out that one can match the optimal regret up to logarithmic terms using a simple mixture-of-experts approach. The idea is rather simple: we can run several instances of projected gradient descent in parallel, each using a different step-size $\eta_i$, and use an experts algorithm to combine their predictions. By using a carefully selected grid of $O(\log(T))$ different step-sizes $\eta_i$ and suitable experts algorithm, one can guarantee regret matching the optimal bound Equation (3.4) up to an additional factor of $O\left(GD\sqrt{T}\log(\log(T))\right)$. This is the essence of the ADER algorithm of L. Zhang, S. Lu, and Z.-H. Zhou (2018).

## 3.2 Strongly-adaptive Regret

At the beginning of this chapter we motivated the need for a stronger notion of regret by reasoning that in certain dynamic/non-stationary environments, a fixed comparator can be a weak baseline when it fails to "track" some time-varying statistic of the losses. Dynamic regret addresses this issue by comparing the learner's performance against that of an arbitrary solution *trajectory* $\boldsymbol{u} = (u_1, \ldots, u_T)$.

Another way to think of the issue is as a matter of insufficient resolution. That is, while a fixed comparator may be a weak baseline over the *entire* interval, it can still be a good baseline over *smaller subsets* of the loss sequence. Intuitively, by "zooming in" to smaller subsets of the sequence one could in theory partition $[1, T]$ into subintervals over which the local statistics of the losses are approximately fixed, so that a fixed comparator captures a *temporally-local* optimum. Then a strong baseline would be to insist that the learner achieves low static regret on each of these subintervals. The difficulty with this is that what constitutes a "reasonable" partition of $[1, T]$ will depend on the particular statistics of the particular loss sequence. To avoid making assumptions about the lengths and locations of these subintervals we can instead insist that the learner achieve low regret on *every sub-interval simultaneously.* In particular, an algorithm is called *strongly-adaptive* if it achieves static regret which is minimax optimal up to logarithmic terms on every sub-interval $[a, b] \subseteq [1, T]$ (Daniely, Gonen, and Shalev-Shwartz 2015).

These sorts of all-intervals guarantees were originally studied in the context of portfolio selection, under the assumption of exp-concave losses (Hazan and Comandur Seshadhri 2007; Hazan and C. Seshadhri 2009). Somewhat recently there has been a renewed interest in strongly adaptive guarantees, with Daniely, Gonen, and Shalev-Shwartz 2015 being the first to obtain a strongly-adaptive guarantee for general convex functions. In particular, for $G$-Lipschitz convex functions and domain $W$ of radius $D$, they derive an algorithm which guarantees

$$R_{[a,b]}(u) \leq O\left(DG\sqrt{b-a}\log(b+1)\right), \qquad \forall [a,b] \subseteq [1,T],$$

and the $\log(b+1)$ was later improved to $\sqrt{\log(b+1)}$ by Jun, Orabona, et al. (2017) by leveraging parameter-free algorithms. Cutkosky 2020 further refines the strongly-adaptive guarantee to yield near-optimal *dynamic* regret over each interval:

$$R_{[a,b]}(\boldsymbol{u}) \leq \widetilde{O}\left(\sqrt{(D^2 + DP_{[a,b]}) \sum_{t\in[a,b]} \|g_t\|^2}\right), \qquad \forall [a,b] \subseteq [1,T]$$

where $P_{[a,b]} = \sum_{t=a+1}^{b} \|u_t - u_{t-1}\|$. This is clearly the strongest type of guarantee, since it captures the optimal dynamic regret as the special case $[a,b] = [1,T]$ as well as as the strongly-adaptive guarantee $R_{[a,b]}(u) \leq \widetilde{O}\left(GD\sqrt{b-a}\right)$ by setting the comparator $u_a = \ldots = u_b = u$.

A natural question is whether a parameter-free analogue of these bounds might be attainable, avoiding the factors of $D$ by instead adapting to the comparator norm. Unfortunately, parameter-free guarantees appear to be incompatible with these all-intervals style guarantees. To see why, notice that for all intervals $[a, b]$ of some fixed length $\tau = b - a$, we would require $R_{[a,b]}(\mathbf{0}) = \sum_{t=a}^{b} \langle g_t, w_t \rangle \leq O(1)$, from which it can be shown that $\|w_t\| \leq O(2^\tau)$ (see, *e.g.*, J. Zhang and Cutkosky 2022, Lemma 8). Yet clearly for large enough $T$ we can not simultaneously guarantee $R_{[1,T]}(u) \leq O\Big(\|u\| G\sqrt{T \log\big(\|u\|\sqrt{T}\big)}\Big)$ for all $u \in \mathbb{R}^d$, since via reward-regret duality this entails competing against a comparator $u \in \mathbb{R}^d$ with $\|u\| = O\Big(\exp(T)/\sqrt{T}\Big)$ in the worst-case[1] which can be made arbitrarily large relative to the fixed $O(2^\tau)$. Hence, even in the best possible scenario where $w_t$ precisely aligns with $u$ on all rounds, the regret on the interval $[1, T]$ could still be very large simply due to the difference in magnitude between $w_t$ and $u$.

Interestingly, in Chapter 10 we will see that in the specific setting of online linear regression, it is actually possible to achieve the stronger all-intervals *dynamic* regret guarantees even in unbounded domains with unbounded losses, without tuning any hyperparameters. Note that this does not contradict the reasoning above because the algorithms in that setting instead guarantee $R_T(\mathbf{0}) = O(\log(T))$, rather than $R_T(\mathbf{0}) = O(1)$, so even though these algorithms make very strong guarantees, they are not considered "parameter-free" in the sense discussed in Section 2.2.

[1] *i.e.*, the comparator which tightens the regret inequality satisfies $\sum_{t=1}^{T} \langle g_t, w_t \rangle \leq \sum_{t=1}^{T} \langle g_t, u \rangle + \psi_T(u) = \min_{u^* \in W} \sum_{t=1}^{T} \langle g_t, u^* \rangle + \psi_T(u^*)$. For the usual comparator-adaptive guarantees this comparator can be as large as $\|u^*\| = \|\nabla \psi_T^*(-g_{1:T})\| = O(\exp(T)/\sqrt{T})$.

# Chapter 4

# Centered Mirror Descent

In order to design algorithms which make strong guarantees under minimal assumptions, we will require a great deal of flexibility in terms of algorithm design. In this chapter we introduce our framework and key technical tools which will be used to design every algorithm in the thesis, which we refer to as *Centered Mirror Descent*.

Before getting into the details of our approach, let us motivate why the existing approaches are not sufficient for our purposes. Recall from Section 2.2.1 that one of the key design principles behind comparator-adaptive algorithms is the reward-regret duality, which states that guaranteeing $R_T(u) \le B_T(u)$ is equivalent to guaranteeing $-\sum_{t=1}^T \langle g_t, w_t \rangle \ge B_T^*(-\sum_{t=1}^T g_t)$. In fact, it will be instructive to recall the reasoning connecting the regret upper bound to the wealth lower bound: suppose that we wish to guarantee static regret of $R_T(u) = \sum_{t=1}^T \langle g_t, w_t - u \rangle \le B_T(u)$ for all $u \in \mathbb{R}^d$. Since this must hold for *any* $u \in \mathbb{R}^d$, it must hold for the $u$ which tightens the inequality:

$$\sup_u R_T(u) - B_T(u) = \sum_{t=1}^T \langle g_t, w_t \rangle + \sup_u \left\langle -\sum_{t=1}^T g_t, u \right\rangle - B_T(u) = \sum_{t=1}^T \langle g_t, w_t \rangle + B_T^*\Big( -\sum_{t=1}^T g_t \Big).$$

so re-arranging we have $-\sum_{t=1}^T \langle g_t, w_t \rangle \ge B_T^*\Big( \sum_{t=1}^T g_t \Big)$. Crucially, this latter condition does not depend on the unknown comparator, making it more amenable to algorithm design. However, notice that the assumption of a *fixed* comparator $u \in \mathbb{R}^d$ was crucial for the above argument to work. It is unclear in general what the analogue of this argument should be for *dynamic* regret, where we instead have a *sequence* of comparators. Similarly, the FTRL-based approach to regret minimization is strongly tied to competing with a fixed comparator, and devising dynamic regret guarantees for FTRL is non-trivial in general. Indeed, we show in Section 4.1 of Jacobsen and Cutkosky (2022) that vanilla FTRL algorithms are not capable of guaranteeing sublinear dynamic regret in general.

On the other hand, it is well-known that there exists a natural connection between mirror descent

```
Algorithm 2: (Generalized) Centered Mirror Descent
1  Input: $\psi_1 : W \to \mathbb{R}$, $\mathcal{M}_1 : W \to W$
2  Initialize: $w_1 = \arg\min_{w \in W} \psi_1(w)$, $\widetilde{w}_1 = \mathcal{M}_1(w_1)$
3  for $t = 1 : T$ do
4      Play $\widetilde{w}_t \in W$, observe loss function $\ell_t : W \to \mathbb{R}$

5      Choose regularizer $\psi_{t+1}$ and composite penalty $\varphi_t$
6      Define $\Delta_t(w) = D_{\psi_{t+1}}(w|w_1) - D_{\psi_t}(w|w_1)$
7      Update $w_{t+1} = \arg\min_{w \in W} \ell_t(w) + D_{\psi_t}(w|\widetilde{w}_t) + (\Delta_t + \varphi_t)(w)$

8      Choose mapping $\mathcal{M}_{t+1} : W \to W$
9      Update $\widetilde{w}_{t+1} = \mathcal{M}_{t+1}(w_{t+1})$
10 end
```

algorithms and dynamic regret. For instance, in a bounded domain with $\sup_{x,y \in W} \|x - y\| \leq D$, subgradient descent[1] with a fixed step-size $\eta$ guarantees dynamic regret (see, *e.g.*, L. Zhang, S. Lu, and Z.-H. Zhou (2018))

$$R_T(\boldsymbol{u}) = \sum_{t=1}^{T} \langle g_t, w_t - u_t \rangle \leq \frac{D^2 + DP_T}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|g_t\|^2 ,$$

where we recall from Chapter 3 that $P_T = \sum_{t=2}^{T} \|u_t - u_{t-1}\|$ is the path-length of the comparator sequence. Observe that this differs from the usual static regret bound from Section 2.1 by only one term, $\frac{D}{\eta} P_T$, which naturally introduces the path-length of the comparator sequence into the bound. Moreover, optimally tuning $\eta^* = \sqrt{\frac{D^2 + DP_T}{\sum_{t=1}^{T} \|g_t\|^2}}$ yields the optimal dynamic regret bound:

$$R_T(\boldsymbol{u}) \leq \sqrt{(D^2 + DP_T) \sum_{t=1}^{T} \|g_t\|^2}.$$

This suggests that a mirror descent-based approach might be naturally well-suited to designing algorithms for dynamic regret. At the same time, we noted in Section 2.2.1 that the vanilla mirror descent algorithm can be fundamentally unstable in unbounded domains. The key to our approach is to remedy this stability issue by incorporating an additional penalty into the mirror descent update, which helps the algorithm behave more similarly to an FTRL algorithm while still maintaining the natural connection to dynamic regret inherent to mirror descent algorithms.

The algorithm is shown in Algorithm 2. Here we present a modest generalization of the original framework due to Jacobsen and Cutkosky (2023) which incorporates a post-hoc adjustment into the update (shown in blue in lines 8-9). For ease of exposition, we first discuss the core algorithm

[1] Recall that subgradient descent is a special case of mirror descent, obtained by setting $\psi(w) = \frac{1}{2\eta} \|w\|^2$

which skips the adjustment step (*i.e.* setting $\mathcal{M}_t(w) = w$ for all $t$), and will discuss the implications of the post-hoc adjustment separately in Section 4.1.

The base algorithm is a particular instance of composite mirror descent (John C Duchi et al. 2010), which is a mirror descent update that adds an auxiliary penalty $\phi_t(w)$ to the loss function $\ell_t(w)$. Typically, $\phi_t(w)$ is a composite loss function which enforces some additional desirable properties of the solution, such as sparsity. In contrast, we will use these terms $\phi_t(w)$ as a crucial stabilizing quantity in our algorithms. This composite term is composed of two parts, $\Delta_t(w)$ and $\varphi_t(w)$, with the distinguishing feature of our approach being the $\Delta_t(w) = D_{\psi_{t+1}}(w|w_1) - D_{\psi_t}(w|w_1)$.

To see what this term $\Delta_t(w)$ contributes, assume $\ell_t(w) = \langle g_t, w\rangle$ for some $g_t \in \mathbb{R}^d$ and suppose we set $\psi_t$ and $w_1$ such that $\min_w \psi_t(w) = \psi_t(w_1) = 0$ for all $t$, $\varphi_t(w) \equiv 0$, and $\mathcal{M}_t(w) = w$. From the first-order optimality condition $w_{t+1} = \arg\min_{w\in\mathbb{R}^d} \langle g_t, w_t\rangle + D_{\psi_t}(w|w_t) + \Delta_t(w)$, we find that $\nabla\psi_{t+1}(w_{t+1}) = \nabla\psi_t(w_t) - g_t$, so unrolling the recursion and solving for $w_{t+1}$ yields $w_{t+1} = \nabla\psi^*_{t+1}(-g_{1:t})$, where $\psi^*_{t+1}$ is the Fenchel conjugate of $\psi_{t+1}$. This latter expression is equivalent to the *Follow-the-Regularized-Leader* (FTRL) update $w_{t+1} = \arg\min_{w\in\mathbb{R}^d} \langle g_{1:t}, w\rangle + \psi_{t+1}(w)$ (McMahan 2017). Moreover, in the constrained setting, letting $\psi_{t+1,W}(w)$ denote the restriction of $\psi_{t+1}$ to constraint set $W$, Algorithm 2 captures both the "greedy projection" update $w_{t+1} = \nabla\psi^*_{t+1,W}(\nabla\psi_t(w_t) - g_t)$ and the "lazy projection" update $w_{t+1} = \nabla\psi^*_{t+1,W}(-g_{1:t})$ by adding the indicator function $\mathbb{I}_W(w)$ to the $\varphi_t$ terms or to the $\psi_t$ terms respectively. Hence, including $\Delta_t(w)$ in Algorithm 2 is a way to incorporate some properties of FTRL into a mirror descent framework.

In the unconstrained setting, the function $\Delta_t(w)$ is in fact a critical stabilizing quantity in the update. Indeed, Orabona and Pál (2018) showed that adaptive mirror descent algorithms can incur *linear* regret in settings where the divergence $D_{\psi_t}(\cdot|\cdot)$ may be unbounded. The issue is that vanilla mirror descent does not properly account for changes in the regularizer $\psi_t$, allowing the iterates $w_t$ to travel away from their initial position $w_1$ too quickly. Algorithm 2 fixes this by adding a corrective penalty $\Delta_t(w)$ related to how much $\psi_t$ has changed between rounds. Since this penalty acts to bias the iterates back towards some central reference point $w_1$, we refer to Algorithm 2 as *Centered Mirror Descent*.

Our approach is similar to *dual-stabilized mirror descent* (DS-MD), proposed by Fang et al. (2020), which employs the update $w_{t+1} = \arg\min_{w\in\mathbb{R}^d} \gamma_t\big(\langle \eta_t g_t, w\rangle + D_\psi(w|w_t)\big) + (1-\gamma_t)D_\psi(w|w_1)$ for scalars $\gamma_t \in (0,1)$. This prevents the iterates $w_t$ from moving too far from $w_1$ by decaying the dual representation of $w_t$ towards that of $w_1$. The DS-MD approach considers only $\psi_t$ of the form $\psi_t = \frac{\psi}{\eta_t}$ for a fixed $\psi$, whereas Centered Mirror Descent applies more generally to $\psi_t$. This property is crucial for our purposes, as the $\psi_t$s we employ cannot be captured by a linear scaling of a fixed underlying $\psi$. One could view our approach as a generalization of Fang et al. (2020) that easily captures a variety of applications, such as dynamic regret, composite losses, and implicit updates. The following lemma provides a generic regret decomposition that we'll use throughout this work.

**Lemma 4.0.1.** ***(Centered Mirror Descent Lemma)*** *Let $\psi_t(\cdot)$ be an arbitrary sequence of differentiable non-negative convex functions, and assume that $w_1 \in \arg\min_{w\in\mathbb{R}^d} \psi_t(w)$ for all $t$. Let $\varphi_t(\cdot)$ be an arbitrary sequence of sub-differentiable non-negative convex functions. Then for any $u_1, \ldots, u_T$, Algorithm 2 guarantees*

$$R_T(\boldsymbol{u}) \le \psi_{T+1}(u_T) + \sum_{t=1}^{T} \varphi_t(u_t) + \sum_{t=2}^{T} \underbrace{\langle \nabla\psi_t(w_t) - \nabla\psi_t(w_1), u_{t-1} - u_t\rangle}_{=:\mathcal{P}_t}$$
$$+ \sum_{t=1}^{T} \underbrace{\langle g_t, w_t - w_{t+1}\rangle - D_{\psi_t}(w_{t+1}|w_t) - (\Delta_t + \varphi_t)(w_{t+1})}_{=:\delta_t}, \tag{4.1}$$

*where $g_t \in \partial\ell_t(w_t)$.*

Proof of this Lemma can be found in Appendix A.2.1. The proof follows as a special case of a regret *equality* which we derive in Appendix A.1. To build intuition for how to use the Lemma, consider the static regret of Algorithm 2 with $\varphi_t(w) \equiv 0$ and $\mathcal{M}_t(w) = w$ for all $t$. In this case, Equation (4.1) becomes $R_T(u) \le \psi_{T+1}(u) + \sum_{t=1}^T \delta_t$. Now, to guarantee a parameter-free bound of the form $R_T(u) \le \widetilde{O}\big(\|u\|\sqrt{T}\big)$ for all $u$, a natural approach is to set $\psi_{T+1}(u) = \widetilde{O}\big(\|u\|\sqrt{T}\big)$, and then focus our efforts on controlling the stability terms $\sum_{t=1}^T \delta_t$. To this end, the following Lemma (proven in Appendix A.2.2) provides a set of simple conditions for bounding an expression closely related to $\delta_t$:

**Lemma 4.0.2.** ***(Stability Lemma)*** *Let $\psi_t(w) = \Psi_t(\|w\|)$ where $\Psi_t : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ is a convex function satisfying $\Psi_t'(x) \ge 0$, $\Psi_t''(x) \ge 0$, and $\Psi_t'''(x) \le 0$ for all $x \ge 0$. Let $c > 0$, $G_{\max} \ge 0$, $G_{\max} \ge G_t$, and assume that there exists an $\mathring{x}_t \ge 0$ and $1/G_{\max}$-Lipschitz convex function $\eta_t : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ satisfying $\eta_t(0) = 0$ such that $|\Psi_t'''(x)| \le \frac{2\eta_t'(x)}{(c+1)^2}\Psi_t''(x)^2$ for all $x \ge \mathring{x}_t$. Then for any $w_{t+1}, w_t \in W$,*

$$\widehat{\delta}_t \overset{def}{=} cG_t\|w_t - w_{t+1}\| - D_{\psi_t}(w_{t+1}|w_t) - \eta_t(\|w_{t+1}\|)G_t^2 \le \frac{(c+1)^2G_t^2}{2\Psi_t''(\mathring{x}_t)}$$

We will generally apply this lemma in the context of $G$-Lipschitz losses, in which case we can set $G_{\max} = G$ and $G_t = \|g_t\|$. The more general conditions in terms of a $G_t \le G_{\max}$ will become relevant in Chapter 7 when we consider a generalization of the usual Lipschitz assumption which captures quadratic losses.

To see the utility of Lemma 4.0.2, let $\phi_t(w) = (\Delta_t + \varphi_t)(w)$ and observe that the only difference between the $\delta_t$ of Lemma 4.0.1 and the $\widehat{\delta}_t$ of Lemma 4.0.2 is that in the former has a $-\phi_t(w_{t+1})$ where the latter has a $-\eta_t(\|w_{t+1}\|)\|g_t\|^2$. Our approach throughout this work will be to design he components of $\phi_t(w)$ to satisfy $\phi_t(w) \ge \eta_t(\|w\|)\|g_t\|^2$ for all $w \in \mathbb{R}^d$ so that $\delta_t \le \widehat{\delta}_t$, and then apply the stability lemma to get $\sum_{t=1}^T \delta_t \le \sum_{t=1}^T \widehat{\delta}_t \le O\left(\sum_{t=1}^T \frac{\|g_t\|^2}{\Psi_t''(x_0)}\right)$. Then, we design $\Psi_t(\cdot)$ to ensure this summation sums to a constant, leading to small regret.

In the sections to follow we will see several examples of $\psi_t$ which meet the conditions of the stability lemma, but for concreteness let us consider as a simple demonstration the fixed function $\psi_t(w) = \Psi(\|w\|) = 2\int_0^{\|w\|} \frac{\log(x/\eta+1)}{\eta}dx$ where $\eta \le \frac{1}{G}$. Careful calculation shows that $\Psi(\cdot)$ satisfies the conditions of Lemma 4.0.2 with $\eta_t(x) = \eta x$. Hence, $\widehat{\delta}_t \le \frac{2\|g_t\|^2}{\Psi_t''(0)} = 2\eta^2 \|g_t\|^2$. Now, we wish to achieve $\phi_t(w_{t+1}) \ge \eta_t(\|w_{t+1}\|)\|g_t\|^2$. This is easily accomplished by setting $\varphi_t(w) = 2\eta^2\|g_t\|^2\|w\|$. Thus, setting $\eta = O(1/\sqrt{T})$ yields $\sum_{t=1}^T \delta_t \le \sum_{t=1}^T \widehat{\delta}_t = \sum_{t=1}^T 2\eta^2\|g_t\|^2 \le O(1)$ so that overall we would achieve a regret of $\widetilde{O}(\|u\|\sqrt{T})$.

This example demonstrates the purpose of $\varphi_t$ in the update. When $\Delta_t(w_{t+1}) \ge \eta_t(\|w_{t+1}\|)\|g_t\|^2$, we already obtain $\delta_t \le \widehat{\delta}_t$. However, this identity may be false (as in the previous example) or difficult to prove.[2] In such cases, we include a small additional $\varphi_t$ term to easily ensure the desired bounds. In fact, this strategy can be viewed as generalizing a certain "correction" term which appears in the experts literature (e.g. Steinhardt and Liang 2014; L. Chen, Luo, and Wei 2021), but to our knowledge is not typically employed in the general online linear optimization setting.

## 4.1 Incorporating Post-hoc Adjustments

Now that we have a feel for how to use the base algorithm, let us consider a modest generalization of which on each round makes an additional *post-hoc* adjustment to the choices of the base algorithm through the use of an arbitrary mapping $\mathcal{M}_t : W \to W$. Algorithms of this form have been studied in prior works such as Gyorgy and Szepesvari (2016) and Hall and Willett (2016), wherein $\mathcal{M}_t$ is interpreted as a dynamical model that the learner has access to. In this work, we will typically use $\mathcal{M}_t$ as a convenient way to formulate algorithms such as fixed-share, which can be interpreted as "mixing in" the uniform distribution to the outputs of the Hedge algorithm to ensure that the output iterates are bounded away from zero (Nicolo Cesa-Bianchi, Gaillard, et al. 2012). Note that the algorithm with no post-hoc adjustment can be interpreted as the special case in which $\mathcal{M}_t$ is the identity mapping $\mathcal{M}_t(w) = w$ for all $t$. When $\mathcal{M}_t$ is not explicitly stated, it is assumed to be the identity mapping and we will write $\widetilde{w}_t = w_t$.

The following lemma provides a regret template for the general algorithm. Observe that several of the key terms related to the algorithm's stability replace the mirror descent iterates $w_t$ with the adjusted iterates $\widetilde{w}_t = \mathcal{M}_t(w_t)$; this property is particularly useful when the regularizer becomes unstable at some $w_0 \in W$, in which case we can use $\mathcal{M}_t$ to bound iterates chosen by the mirror descent update away from that point. The trade-off is that we must ensure that the new penalty terms $\xi_t = D_{\psi_{t+1}}(u_t|\widetilde{w}_{t+1}) - D_{\psi_{t+1}}(u_t|w_{t+1})$ are not too large, which places an implicit restriction on how much we can adjust the iterates via $\mathcal{M}_t$. Proof of the lemma can be found in Appendix A.2.3.

[2] Proving an analogous identity is the principle technical challenge in deriving FTRL-based parameter-free algorithms.

**Lemma 4.1.1.** *For all $t$ let $\psi_t : W \to \mathbb{R}$ be differentiable convex functions, $\varphi_t : W \to \mathbb{R}$ be subdifferentiable convex functions, and let $\mathcal{M}_t : W \to W$ be arbitrary mappings. Then for any sequence $\boldsymbol{u} = (u_1, \ldots, u_T)$ in $W$, Algorithm 2 guarantees*

$$\begin{aligned} R_T(\boldsymbol{u}) \le\ & D_{\psi_{T+1}}(u_T|\widetilde{w}_1) - D_{\psi_{T+1}}(u_T|\widetilde{w}_{T+1}) + \sum_{t=1}^{T} \varphi_t(u_t) \\ & + \sum_{t=2}^{T} \underbrace{\langle \nabla\psi_t(\widetilde{w}_t) - \nabla\psi_t(\widetilde{w}_1), u_{t-1} - u_t\rangle}_{=:\mathcal{P}_t} + \sum_{t=1}^{T} \underbrace{D_{\psi_{t+1}}(u_t|\widetilde{w}_{t+1}) - D_{\psi_{t+1}}(u_t|w_{t+1})}_{\xi_t} \\ & + \sum_{t=1}^{T} \underbrace{\langle g_t, \widetilde{w}_t - w_{t+1}\rangle - D_{\psi_t}(w_{t+1}|\widetilde{w}_t) - (\Delta_t + \varphi_t)(w_{t+1})}_{=:\delta_t}, \end{aligned}$$

*where $g_t \in \partial \ell_t(\widetilde{w}_t)$.*

We will occasionally use a linearization of the composite penalties $w \mapsto \langle \nabla\varphi_t(\widetilde{w}_t), w\rangle$ for $\nabla\varphi_t(\widetilde{w}_t) \in \partial\varphi_t(\widetilde{w}_t)$. The main reason for doing so is that it can lead to updates with simpler closed-form expressions. The drawback is that doing this generally leads to slightly worse constants in the regret bound.

**Lemma 4.1.2.** *Under the same conditions as Lemma 4.1.1, let $\nabla\varphi_t(\widetilde{w}_t) \in \partial\varphi_t(\widetilde{w}_t)$ and suppose we replace $\varphi_t$ with its linearization $\langle \nabla\varphi_t(\widetilde{w}_t), w\rangle$. Then Algorithm 2 guarantees*

$$\begin{aligned} R_T(\boldsymbol{u}) \le\ & D_{\psi_{T+1}}(u_T|\widetilde{w}_1) - D_{\psi_{T+1}}(u_T|\widetilde{w}_{T+1}) + \sum_{t=1}^{T} \varphi_t(u_t) + \mathcal{P}_{2:T} + \xi_{1:T} \\ & + \sum_{t=1}^{T} \underbrace{\langle g_t + \nabla\varphi_t(\widetilde{w}_t), \widetilde{w}_t - w_{t+1}\rangle - D_{\psi_t}(w_{t+1}|\widetilde{w}_t) - \Delta_t(w_{t+1}) - \varphi_t(\widetilde{w}_t)}_{=:\delta_t}, \end{aligned}$$

*where $g_t \in \partial \ell_t(\widetilde{w}_t)$.*

The proof is immediate by observing that

$$\begin{aligned} \sum_{t=1}^{T} \ell_t(\widetilde{w}_t) - \ell_t(u_t) \le \sum_{t=1}^{T} \langle g_t, \widetilde{w}_t - u_t\rangle &= \sum_{t=1}^{T} \langle g_t, \widetilde{w}_t - u_t\rangle \pm [\varphi_t(\widetilde{w}_t) - \varphi_t(u_t)] \\ &\le \langle g_t + \nabla\varphi_t(\widetilde{w}_t), \widetilde{w}_t - u_t\rangle + \sum_{t=1}^{T} \varphi_t(u_t) - \varphi_t(\widetilde{w}_t) \end{aligned}$$

and then applying Lemma 4.1.1 with losses $w \mapsto \langle g_t + \nabla\varphi_t(\widetilde{w}_t), w\rangle$ and composite penalty $\varphi_t(w) \equiv 0$.

## 4.2 Conclusions

In this chapter we introduced our main framework for designing and analyzing OCO algorithms. Our approach maintains the natural connection to dynamic regret provided by a mirror descent algorithm, while incorporating the desirable stability properties of an FTRL method. In the remaining chapters, we will leverage these properties to design novel algorithms for both static and dynamic regret alike.

# Part II

# Adaptivity in Stationary Settings

# Chapter 5

# Overview of Part II

In Part II of this thesis, we begin by applying our framework from Chapter 4 to design parameter-free algorithms for *static* regret. As discussed in Chapter 3, static regret is a special case of the more general notion of dynamic regret, so in this part of the thesis we are considering strictly easier problem settings than what we've set out to solve in this thesis. However, starting with static regret allows us to first showcase how to use our framework and build up a set of tools and intuitions for designing comparator-adaptive algorithms without the additional complexity of a time-varying comparator. Moreover, we are able to highlight the utility of our approach by developing several novel parameter-free guarantees as well as improvements to existing results. Notably, in this part of the thesis we develop the first parameter-free algorithms that can be applied in settings with non-Lipschitz losses, such as quadratic and logistic losses. This part of the thesis is organized as follows.

**Learning with Lipschitz Losses.** We begin in Chapter 6 in the setting of $G$-Lipschitz losses and unbounded domain $W$. This is the standard setting in which parameter-free algorithms are studied. The results in this chapter serve as a nice warm-up demonstrating how to accomplish various forms of adaptivity in our framework. We first show that our approach to algorithm design presented in Chapter 4 allows us to improve upon existing results in the literature.

- In Section 6.1, we construct a parameter-free which attains the optimal parameter-free rate, achieving complete second-order adaptivity to the gradients

  $$R_T(u) \le \widehat{O}\left(G\epsilon + \|u\|\sqrt{\|g\|_{1:T}^2 \log\left(\frac{\|u\|\sqrt{\|g\|_{1:T}^2}}{\epsilon G} + 1\right)}\right).$$

  Note that bound is *fully* second-order adaptive, meaning that *nowhere* in the bound does the pessimistic upper bound $G\sqrt{T} \geq \sqrt{\|g\|^2_{1:T}}$ appear. Prior works have only achieved this property while maintaining the optimal $\log\left(\frac{\|u\|\sqrt{\|g\|^2_{1:T}}}{G\epsilon}+1\right)$ dependence by resorting to the doubling trick (Cutkosky and Sarlos 2019).

- In Section 6.2, we consider the Lipschitz-adaptive setting in which the learner does not have *a priori* access to an upperbound $G \geq \|g_t\|$, and instead has to estimate $G_t = \max_{s\leq t}\|g_s\|$ on-the-fly. The state-of-the-art result in this setting is the FreeGrad algorithm of Mhammedi and Koolen 2020, which requires a doubling-like restart strategy to ensure non-vacuous regret. We use our framework to design a scale-free algorithm which avoids resorting to restarts and even modestly improves the regret guarantee of FreeGrad. These improvements fall seamlessly out of our mirror-descent-based approach.

On top of these improvements to existing results, we also develop *new* types of parameter-free guarantees. In particular,

- In Section 6.3, we design an algorithm which adapts to the *gradient variability*, guaranteeing

$$R_T(u) \leq \widehat{O}\left(\|u\|\sqrt{\sum_{t=1}^T \|\nabla\ell_t(w_t) - \nabla\ell_{t-1}(w_t)\|^2 \log\left(\frac{\|u\|\sqrt{\sum_{t=1}^T \|\nabla\ell_t(w_t) - \nabla\ell_{t-1}(w_t)\|^2}}{G\epsilon}+1\right)}\right).$$

  Notice that this bound can automatically be much smaller than the standard parameter-free bound which scales with $\sqrt{\sum_{t=1}^T \|g_t\|^2}$ in any problem where the losses are "slowly varying". We achieve this result by leveraging modern extensions of mirror descent (*implicit* and *optimistic* updates) which have no obvious analogue in existing approaches to parameter-free learning such as coin-betting.

- An alternative parameter-free bound that has received recent interest takes the form

$$R_T(u) \leq O\left(\epsilon\sqrt{\|g\|^2_{1:T}} + \|u\|\sqrt{\|g\|^2_{1:T}\log\left(\|u\|/\epsilon+1\right)}\right),$$

  sacrificing larger regret at the origin to remove the horizon-dependent penalty $\sqrt{\|g\|^2_{1:T}}$ from the logarithm. In Section 6.4 we show that not only does our approach allow us to immediately derive bounds of this form, but we can in fact achieve any intermediate guarantee on a spectrum between this and the usual parameter-free bound, leading to a novel frontier of parameter-free guarantees. We provide analogous results for our scale-free and implicit/optimistic algorithm as well.

**Learning with Unbounded Losses.** The main contributions of Part II of this thesis are presented in Chapter 7. We consider a relaxation of the standard Lipschitz assumption which captures many standard but non-Lipschitz loss functions as special cases, such as the square loss and logistic loss. Then, under this assumption,

- We design an algorithm which achieves regret guarantees of the form

  $$R_T(u) \le \widehat{O}\left(\|u\| G\sqrt{T\log\left(\|u\|\sqrt{T}/\epsilon + 1\right)} + L\|u\|^2\sqrt{T}\right),$$

  where $L$ is a scaling factor related to non-Lipschitzness of the losses. Not only does our result provide a strict generalization of the usual parameter-free bound, but it is the first result in OCO to achieve non-trivial regret guarantees in a setting where both the domain and the losses may be unbounded.

- We provide a lower bound demonstrating that our result is unimprovable in general.

- As an application of our result, in Section 7.2 we are able to design the first algorithms for saddle-point optimization which converge in duality gap in unbounded domains without curvature assumptions such as strong convexity. As a special case, our results can be applied to *bilinearly coupled* saddle-point problems, a very common type of problem which captures many settings of practical interest, such as off-policy policy evaluation in reinforcement learning (see Section 7.2.1).

# Chapter 6

# Lipschitz Losses

In this chapter, we consider online learning in unbounded domains $W \subseteq \mathbb{R}^d$ with $G$-Lipschitz convex losses, satisfying $\|\nabla \ell_t(w)\| \leq G$ for any $\nabla \ell_t(w) \in \partial \ell_t(w)$ and $w \in W$. We begin our study by limiting our scope to *static* regret, representing problem settings with stationary dynamics. For simplicity, in this chapter we assume $W = \mathbb{R}^d$.

## 6.1 Parameter-free Learning

As a warm-up, we first use our framework developed in Chapter 4 to construct a parameter-free algorithm which achieves the optimal static regret, matching the lower bound in Equation (2.3). Pseudocode for the algorithm characterized in the following theorem is provided in Algorithm 3 for convenience.

**Theorem 6.1.1.** *Let $\ell_1, \ldots, \ell_T$ be $G$-Lipschitz convex functions and $g_t \in \partial \ell_t(w_t)$ for all $t$. Let $\epsilon > 0$, $V_t = 4G^2 + \|g\|^2_{1:t-1}$, $\alpha_t = \frac{\epsilon G}{\sqrt{V_t}\log^2(V_t/G^2)}$, and set $\psi_t(w) = 3\int_0^{\|w\|} \min_{\eta \leq \frac{1}{G}} \left[\frac{\log(x/\alpha_t+1)}{\eta} + \eta V_t\right] dx$. Then for all $u \in \mathbb{R}^d$, Algorithm 2 guarantees*

$$R_T(u) \leq \widehat{O}\left(G\epsilon + \|u\|\left[\sqrt{\|g\|^2_{1:T}\log\left(\frac{\|u\|\sqrt{\|g\|^2_{1:T}}}{\epsilon G}+1\right)} \vee G\log\left(\frac{\|u\|\sqrt{\|g\|^2_{1:T}}}{\epsilon G}+1\right)\right]\right)$$

*where $\widehat{O}(\cdot)$ hides constant and* $\log(\log)$ *factors (but not* $\log$ *factors).*

The full proof can be found in Appendix B.1.1, along with derivation of the update formula shown in Algorithm 3. It follows the intuition developed in the Chapter 4: Lemma 4.0.1 implies $R_T(u) \leq \psi_{T+1}(u) + \sum_{t=1}^T \delta_t$. Then, we show that $\psi_t$ satisfies the conditions of Lemma 4.0.2 while the

```
Algorithm 3: Parameter-free Learning via Centered Mirror Descent
Input: Lipschitz bound G, Value $\epsilon > 0$
Initialize: $V_1 = 4G^2$, $w_1 = \mathbf{0}$, $\theta_1 = \mathbf{0}$
for $t = 1 : T$ do
    Play $w_t$, receive subgradient $g_t$
    Set $\theta_{t+1} = \theta_t - g_t$, $V_{t+1} = V_t + \|g_t\|^2$, $\alpha_{t+1} = \frac{\epsilon G}{\sqrt{V_{t+1}}\log^2(V_{t+1}/G^2)}$, and define
        $f_{t+1}(\theta) = \begin{cases} \frac{\|\theta\|^2}{36V_{t+1}} & \text{if } \|\theta\| \le \frac{6V_{t+1}}{G} \\ \frac{\|\theta\|}{3G} - \frac{V_{t+1}}{G^2} & \text{otherwise} \end{cases}$
    Update $w_{t+1} = \frac{\alpha_{t+1}\theta_{t+1}}{\|\theta_{t+1}\|}\left[\exp\left(f_{t+1}(\theta_{t+1})\right) - 1\right]$
end
```

growth rate $\Delta_t(w)$ ensures that $\delta_t \le \widehat{\delta}_t$, so that $\delta_t \le \widehat{\delta}_t \le O\left(\frac{2\|g_t\|^2}{\Psi''_t(\alpha_t)}\right) \le O(\frac{\alpha_t\|g_t\|^2}{\sqrt{V_t}})$. Finally, we choose $\alpha_t$ small enough to ensure $\sum_{t=1}^{T}\delta_t \le O(1)$.

Treating log(log) terms as effectively constant, the bound in Theorem 6.1.1 achieves the "ideal" dependence on $\|g\|^2_{1:T}$ in the logarithmic factors. Indeed, given oracle access to $\|u\|$ and $\|g\|^2_{1:T}$, we could set $\epsilon = O\Big(\frac{\|u\|\sqrt{\|g\|^2_{1:T}}}{G}\Big)$, causing all the log terms to disappear from the bound and leaving only $R_T(u) \le \widehat{O}\big(\,\|u\|\sqrt{\|g\|^2_{1:T}}\big)$, which matches the optimal rate that vanilla gradient descent would achieve with oracle tuning up to a log(log) factor. Prior works typically do not have this property, failing to avoid additional log penalties even *with* oracle tuning of $\epsilon$. One exception we are aware of is Cutkosky and Sarlos (2019, Appendix C.1), which requires resorting to the doubling trick, and partial exceptions include McMahan and Orabona (2014) and Jun and Orabona (2019), which fail to maintain complete second-order adaptivity, incurring worst-case dependencies $G^2T \ge \|g\|^2_{1:T}$ in the bound.

## 6.2 Lipschitz Adaptivity and Scale-free Learning

The algorithm in the previous section requires *a priori* knowledge of the Lipschitz constant $G$ to run. This is unfortunate, as such knowledge may not be known in practice. An ideal algorithm would instead *adapt* to an unknown Lipschitz constant $G$ on-the-fly, while still maintaining $R_T(u) \le \widetilde{O}\left(\|u\|\,G\sqrt{T}\right)$ static regret. Unfortunately, various lower bounds (e.g. Cutkosky and Boahen (2017) and Mhammedi and Koolen (2020)) show that this goal is not achievable in general. Nevertheless, it is possible to make significant partial progress.

One simple approach to this problem, suggested by Cutkosky (2019a), is the following reduction based on a gradient-clipping approach. First, we design an algorithm $\mathcal{A}$ which achieves suitable regret when given prescient "hints" $h_t$ satisfying $h_t \ge \|g_t\|$ at the start of round $t$. In practice, we obviously can not provide such hints because we have not yet observed $g_t$, so instead we pass our best estimate, $h_t = \max_{s<t} \|g_s\|$. Then, we simply pass $\mathcal{A}$ *clipped* subgradients $\overline{g}_t = g_t \min\left\{1, \frac{h_t}{\|g_t\|}\right\}$, which ensures that $h_t \geq \|\overline{g}_t\|$, so that the hint given to $\mathcal{A}$ is never incorrect. Finally, the outputs $w_t$ of $\mathcal{A}$ are constrained to lie in the domains $W_t = \left\{w \in \mathbb{R}^d : \|w\| \leq \sqrt{\sum_{s=1}^{t-1} \|g_s\| / G_s}\right\}$ where $G_t = \max_{\tau \leq t} \|g_\tau\|$. Cutkosky 2019a showed that this approach ensures $R_T(u) \leq R_T^{\mathcal{A}}(u) + G_T \|u\| + G_T \sqrt{\sum_{t=1}^T \|g_t\| / G_t} + G_T \|u\|^3$, where $R_T^{\mathcal{A}}(u)$ is the regret of $\mathcal{A}$ on the losses $\overline{g}_t$, and Mhammedi and Koolen (2020) showed that these additive penalties are unimprovable, so this bound captures the best-possible compromise.

While this hint-based strategy can be used to mitigate the problem of an unknown Lipschitz constant $G$, a truly ideal algorithm would be *scale-free*. That is, the algorithm's outputs $w_t$ are invariant to any constant rescaling of the gradients $g_t \mapsto c g_t$ for all $t$. Scale-free regret bounds scale with the maximal subgradient *encountered* $G_T = \max_{t \leq T} \|g_t\|$, while non-scale free bounds typically depend on some user-specified estimate of $G_T$ and may perform much worse if this estimate is very poor. Mhammedi and Koolen (2020) used the approach proposed by Cutkosky (2019a) to develop FreeGrad, the first parameter-free and scale-free algorithm.

FreeGrad unfortunately suffers from an analytical difficulty called the *range-ratio* problem. Briefly, the range-ratio problem occurs when $h_T/h_1$ (called the range-ratio) is very large: in principle if we set $h_1 = \|g_1\|$, then this quantity could grow arbitrarily large, and so even logarithmic dependencies can make the regret bound vacuous. In order to circumvent this difficulty, Mhammedi and Koolen 2020 utilize a doubling-based scheme, restarting FreeGrad whenever a particular technical condition was met. While such restart strategies only lose a constant factor in the regret in theory, they are unsatisfying: scale-free updates are motivated by potential practical performance benefits, yet any algorithm which is forced to restart from scratch several times during deployment is unlikely to achieve high performance in practice. The following theorem, proven in Appendix B.1.3, characterizes a new base algorithm that, when combined with the reduction of Cutkosky (2019a), generates a scale-free algorithm which avoids the range-ratio problem without resorting to restarts. Our approach employs a simple analysis which follows easily using the tools developed in Chapter 4, enabling us to achieve tighter logarithmic factors than FreeGrad.

**Theorem 6.2.1.** *Let $\ell_1, \ldots, \ell_T$ be convex functions and $g_t \in \partial \ell_t(w_t)$ for all $t$. Let $h_1 \leq \ldots \leq h_T$ be a sequence of hints such that $h_t \geq \|g_t\|$, and assume that $h_t$ is provided at the start of each round $t$. Set $\psi_t(w) = 3\int_0^{\|w\|} \min_{\eta \leq \frac{1}{h_t}} \left[\frac{\log(x/\alpha_t + 1)}{\eta} + \eta V_t\right] dx$ where $V_t = 4h_t^2 + \|g\|_{1:t-1}^2$, $\alpha_t = \frac{\epsilon}{\sqrt{B_t}\log^2(B_t)}$, $B_t = 4\sum_{s=1}^t \left(4 + \sum_{s'=1}^{s-1} \frac{\|g_{s'}\|^2}{h_{s'}^2}\right)$, and $\epsilon > 0$. Then for all $u \in \mathbb{R}^d$, Algorithm 2 guarantees*

$$R_T(u) \leq \widehat{O}\left(\epsilon h_T + \|u\| \left[\sqrt{\|g\|_{1:T}^2 \log\left(\frac{\|u\| \sqrt{B_{T+1}}}{\epsilon} + 1\right)} \vee h_T \log\left(\frac{\|u\| \sqrt{B_{T+1}}}{\epsilon} + 1\right)\right]\right)$$

*where $\widehat{O}(\cdot)$ hides constant and* $\log(\log)$ *factors*

**Algorithm 4:** Unbounded, Scale-Free, Lipschitz Adaptivity

```
Initialize: $w_1 = \mathbf{0}$, $h_1 = 0$, $G_0 = 0$, $\widetilde{b}_1 = 4$, $\widetilde{B}_1 = 4\widetilde{b}_1$, $\widetilde{\theta}_1 = \mathbf{0}$
for $t = 1 : T$ do
    Define $D_t = \sqrt{\sum_{s=1}^{t-1} \frac{\|g_s\|}{G_s}}$ and $W_t = \left\{w \in \mathbb{R}^d : \|w\| \le D_t\right\}$
    Play $\widehat{w}_t = \Pi_{W_t}(w_t) = w_t \min\left\{1, \frac{D_t}{\|w_t\|}\right\}$
    Receive subgradient $g_t$

    Set $\overline{g}_t = g_t \min\left\{1, \frac{h_t}{\|g_t\|}\right\}$, $G_t = \max\left\{\|g_t\|, G_{t-1}\right\}$, and $h_{t+1} = G_t$
    Set $\widetilde{\ell}_t(w) = \frac{1}{2}\langle \overline{g}_t, w\rangle + \frac{1}{2}\|\overline{g}_t\| \max\left\{0, \|w_t\| - D_t\right\}$ and compute $\widetilde{g}_t \in \partial\widetilde{\ell}_t(w_t)$

    Set $\widetilde{\theta}_{t+1} = \widetilde{\theta}_t - \widetilde{g}_t$, $\widetilde{V}_{t+1} = 4h_{t+1}^2 + \|\widetilde{g}\|_{1:t}^2$, $\widetilde{b}_{t+1} = \widetilde{b}_t + \frac{\|\widetilde{g}_t\|^2}{h_t^2}$, $\widetilde{B}_{t+1} = \widetilde{B}_t + 4\widetilde{b}_t$, and
        $\widetilde{\alpha}_{t+1} = \frac{\epsilon}{\sqrt{\widetilde{B}_{t+1}}\log^2(\widetilde{B}_{t+1})}$
    Define $f_{t+1}(\theta) = \begin{cases} \frac{\|\theta\|^2}{36\widetilde{V}_{t+1}} & \text{if } \|\theta\| \le \frac{6\widetilde{V}_{t+1}}{h_{t+1}} \\ \frac{\|\theta\|}{3h_{t+1}} - \frac{\widetilde{V}_{t+1}}{h_{t+1}^2} & \text{otherwise} \end{cases}$
    Update $w_{t+1} = \frac{\widetilde{\alpha}_{t+1}\widetilde{\theta}_{t+1}}{\|\widetilde{\theta}_{t+1}\|}\left[\exp\left(f_{t+1}\left(\widetilde{\theta}_{t+1}\right)\right) - 1\right]$
end
```

The proof of this Theorem follows the strategy of previous sections: from Lemma 4.0.1 we have $R_T(u) \le \psi_{T+1}(u) + \sum_{t=1}^T \delta_t$. To bound $\sum_{t=1}^T \delta_t$, we apply Lemma 4.0.2 and show that the growth rate $\Delta_t(w)$ is sufficiently large to ensure $\sum_{t=1}^T \delta_t \le \sum_{t=1}^T \frac{2\|g_t\|^2}{\Psi_t''(x_0)}$ for some small $x_0$. The main subtlety compared to Theorem 6.1.1 is the influence of the terms $B_t$.

The terms $B_t$ are carefully chosen to address the range-ratio problem in an *online* fashion: we show that $\sqrt{B_t}$ upper bounds the quantity $h_t/h_{\tau_t}$, where starting from $\tau_1 = 1$, the variable $\tau_t$ roughly tracks the most-recent round $t$ where the ratio $h_t/h_{\tau_{t-1}}$ exceeds a threshold analogous to the one used by FreeGrad to trigger restarts. That is, $B_t$ enacts a kind of "soft restarting" by shrinking $w_t$ according to the restarting threshold, just as setting a learning rate of $1/\sqrt{t}$ in online gradient descent can be viewed as a "soft restart" in contrast to the standard doubling trick. It is quite possible that FreeGrad could be similarly modified to avoid restarts by incorporating $B_t$ directly, but this is difficult to verify due to highly non-trivial polynomial expressions appearing in the analysis.

The full pseudocode for our Scale-free, Lipschitz adaptive algorithm for unbounded domains is given in Algorithm 4. The update equation is derived in a similar manner to the algorithm in Section 6.1. The implementation can be understood as the Leashed meta-algorithm of Cutkosky (2019a) with an instance of the algorithm specified in Theorem 6.2.1 as the base algorithm. The corresponding regret guarantee is immediate using Theorem 6.2.1 along with the with the aforementioned reductions (Cutkosky 2019a, Theorem 3).

**Algorithm 5:** Implicit-Optimistic Centered Mirror Descent

```
Input: Initial regularizer $\psi_1 : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$, initial $\widehat{\ell}_1(\cdot)$
Initialize: $x_1 = \arg\min_x \psi_1(x)$, $w_1 = \arg\min_w \widehat{\ell}_1(w) + D_{\psi_1}(w|x_1)$
for $t = 1 : T$ do
    Play $w_t$, observe loss function $\ell_t(\cdot)$
    Set $g_t \in \partial \ell_t(w_t)$
    Choose functions $\psi_{t+1}$, $\widehat{\ell}_{t+1}$, and define $\Delta_t(w) = D_{\psi_{t+1}}(w|x_1) - D_{\psi_t}(w|x_1)$
    Update $x_{t+1} = \arg\min_x \langle g_t, x\rangle + D_{\psi_t}(x|x_t) + \Delta_t(x)$
           $w_{t+1} = \arg\min_w \widehat{\ell}_{t+1}(w) + D_{\psi_{t+1}}(w|x_{t+1})$
end
```

**Corollary 6.2.2.** *For any $u \in \mathbb{R}^d$, Algorithm 4 guarantees*

$$R_T(u) \leq \widehat{O}\Bigg(\epsilon G_T + \|u\|\left[\sqrt{V_{T+1}\log\left(\frac{\|u\|\sqrt{B_{T+1}}}{\epsilon}+1\right)} \vee G_T\log\left(\frac{\|u\|\sqrt{B_{T+1}}}{\epsilon}+1\right)\right]$$

$$+ G_T\|u\|^3 + G_T\|u\| + G_T\sqrt{\sum_{t=1}^{T}\frac{\|g_t\|}{G_t}}\Bigg),$$

*where $G_T = \max_{\tau\leq T}\|g_\tau\|$ and $B_{T+1} = 4\sum_{t=1}^{T+1}\left(4 + \sum_{s=1}^{t-1}\frac{\|g_s\|^2}{h_s^2}\right)$ and $\widehat{O}(\cdot)$ hides constant and* $\log(\log)$ *factors.*

## 6.3 Adapting to Gradient Variability

A useful consequence of our mirror descent formulation is that we can easily incorporate the entire loss function $\ell_t(\cdot)$ rather than the linear proxy $w \mapsto \langle \nabla \ell_t(w_t), w\rangle$ used in the usual mirror descent update. Mirror descent updates incorporating $\ell_t(\cdot)$ in their update are called *implicit*, because setting $w_{t+1} = \arg\min_{w\in\mathbb{R}^d}\ell_t(w) + D_\psi(w|w_t)$ leads to an equation of the form $w_{t+1} = \nabla\psi^*(\nabla\psi(w_t) - \nabla\ell_t(w_{t+1}))$, which must be solved for $w_{t+1}$ to obtain the update.

Implicit updates are appealing in practice because they enable one to more directly incorporate known properties of the loss functions or additional modeling assumptions to improve convergence rates (Asi and John C. Duchi 2019). Moreover, in bounded domains there may be advantages even without any additional assumptions on the loss functions. Indeed, Campolongo and Orabona (2020) recently developed an implicit mirror descent which guarantees $R_T(u) \leq O\Big(\min\Big\{\sqrt{\|g\|_{1:T}^2}, \mathcal{V}_T\Big\}\Big)$ where $\mathcal{V}_T = \sum_{t=2}^T \sup_{x\in\mathcal{X}} \ell_t(x) - \ell_{t-1}(x)$ is the *temporal variability* of the loss sequence. This bound has the appealing property that $R_T(u) \leq O(1)$ when the loss functions are fixed $\ell_t(\cdot) = \ell(\cdot)$.

In this section we leverage our mirror descent formulation to incorporate an additional implicit update on each step to guarantee $R_T(u) \leq \widetilde{O}\Big(\|u\|\sqrt{\sum_{t=1}^T\|\nabla\ell_t(w_t) - \nabla\ell_{t-1}(w_t)\|^2}\Big)$, which can be significantly smaller than the usual $R_T(u) \leq \widetilde{O}\Big(\|u\|\sqrt{\sum_{t=1}^T\|\nabla\ell_t(w_t)\|^2}\Big)$ bound when the loss functions are "slowly moving". Similar to Campolongo and Orabona (2020), this bound guarantees that $R_T(u) \le O(1)$ when the loss functions are fixed, yet our result holds even in unconstrained domains. In fact, in the setting of Lipschitz losses in unconstrained domains, the quantity $\sqrt{\sum_{t=1}^{T} \|\nabla \ell_t(w_t) - \nabla \ell_{t-1}(w_t)\|^2}$ is perhaps a more suitable way to achieve this property, since in unbounded domains $\mathcal{V}_T$ is typically infinite unless $\ell_t - \ell_{t-1}$ is constant.

The only prior method we are aware of to incorporate implicit updates into parameter-free learning was recently developed by K. Chen, Cutkosky, and Orabona 2022. They propose an interesting new regret decomposition and apply it to develop closed-form implicit updates for truncated linear losses. We adopt different goals: without attempting to build efficient closed-form updates, we consider general loss functions and show that implicit updates fall easily out of our mirror-descent formulation.

Our algorithm is derived as a special case of the algorithm shown in Algorithm 5, which can be understood as an instance of centered mirror descent with an additional *optimistic* step on each round. The optimistic step leverages an arbitrary guess $\widehat{\ell}_{t+1}(\cdot)$ about what the next loss function will be. Intuitively, if the learner could deduce the trajectory of the loss functions, they'd be able to "think ahead" and play a point $w_{t+1}$ for which the next loss $\ell_{t+1}(\cdot)$ is minimized. The following theorem provides an algorithm which guarantees $R_T(u) \le \widetilde{O}\Big( \|u\| \sqrt{\sum_{t=1}^{T} \|\nabla \ell_t(w_t) - \nabla \widehat{\ell}_t(w_t)\|^2}\Big)$ using an arbitrary sequence of optimistic guesses $\widehat{\ell}_t(\cdot)$.

**Theorem 6.3.1.** *Let $\ell_1, \dots, \ell_T$ and $\widehat{\ell}_1, \dots, \widehat{\ell}_T$ be G-Lipschitz convex functions. For all $t$, let $\psi_t(w) = 3\int_0^{\|w\|} \min_{\eta \le \frac{1}{2G}} \Big[\frac{\log(x/\widehat{\alpha}_t + 1)}{\eta} + \eta \widehat{V}_t\Big] dx$, where $\widehat{V}_t = 16G^2 + \sum_{s=1}^{t-1} \|\nabla \ell_s(w_s) - \nabla \widehat{\ell}_s(w_s)\|^2$, $\widehat{\alpha}_t = \frac{\epsilon G}{\sqrt{\widehat{V}_t} \log^2(\widehat{V}_t / G^2)}$, and $\epsilon > 0$. Then for all $u \in \mathbb{R}^d$, Algorithm 5 guarantees*

$$R_T(u) \le \widehat{O}\left( \epsilon G + \|u\| \left[ \sqrt{\widehat{V}_{T+1} \log\left( \frac{\|u\| \sqrt{\widehat{V}_{T+1}}}{G\epsilon} + 1 \right)} \vee G \log\left( \frac{\|u\| \sqrt{\widehat{V}_{T+1}}}{G\epsilon} + 1 \right) \right] \right),$$

*where $\widehat{O}(\cdot)$ hides constant and* $\log(\log)$ *factors.*

The proof is similar to the proof of Theorem 6.1.1, with some tweaks to account for the optimistic step, and is deferred to Appendix B.1.2. As an immediate corollary, we have that by setting $\widehat{\ell}_{t+1}(w) = \ell_t(w)$, the regret is bounded as $R_T(u) \le \widetilde{O}\Big( \|u\| \sqrt{\sum_{t=1}^{T} \|\nabla \ell_t(w_t) - \nabla \ell_{t-1}(w_t)\|^2}\Big)$. Bounds of this form have previously only been obtained in bounded domains (Zhao et al. 2020).

Note that our algorithm only makes use of an implicit update during the optimistic step. One could also implement an implicit update in the primary update, but it is unclear what concrete improvements this would yield in the regret bound in the unbounded setting. We leave this as an exciting direction for future work.

## 6.4 Trade-offs in the Horizon Dependence

In the preceeding sections, we focused primarily on standard parameter-free guarantees of the form

$$R_T(u) \le \widetilde{O}\left(G\epsilon + \|u\|\sqrt{\|g\|_{1:T}^2 \log\left(\frac{\|u\|\sqrt{\|g\|_{1:T}^2}}{G\epsilon} + 1\right)}\right). \tag{6.1}$$

However, recently there has been interest in a variant of Equation (6.1) that scales instead as

$$R_T(u) \le \widetilde{O}\left(\epsilon\sqrt{\|g\|_{1:T}^2} + \|u\|\sqrt{\|g\|_{1:T}^2 \log\left(\frac{\|u\|}{\epsilon} + 1\right)}\right) \tag{6.2}$$

which captures the optimal *asymptotic* dependence on the variance terms $\|g\|_{1:T}^2$ by moving them out of the logarithm (Z. Zhang, Cutkosky, and I. Paschalidis 2022b; Z. Zhang, Cutkosky, and I. Paschalidis 2022a; Z. Zhang, H. Yang, et al. 2023; Z. Zhang, Cutkosky, and Y. Paschalidis 2023). It is easy to see that non-adaptive forms of Equation (6.2) can be achieved using the usual parameter-free guarantee, Equation (6.1), by setting $\epsilon = \sqrt{T}$. The result can likewise also be achieved in a horizon-independent manner by applying the doubling trick. The first work to achieve guarantees of the form $R_T(u) \le O\big(G\epsilon\sqrt{T} + G\|u\|\sqrt{T\log(\|u\|/\epsilon + 1)}\big)$ *without* resorting to the doubling trick was Z. Zhang, Cutkosky, and I. Paschalidis 2022b, using a novel approach based on discretizing the dynamics of a continuous-time potential function. The fully-adaptive bound shown in Equation (6.2) was then later achieved by Z. Zhang, H. Yang, et al. 2023 by using an improved discretization strategy.

Inspired by these works, in this section we show that bounds in the form of Equation (6.2) can also be attained in a straight-forward manner using our mirror descent framework. Interestingly, each of our static regret algorithms attain a bound analogous to Equation (6.2) by simply setting $\alpha_t = \epsilon$ for all $t$. The following proposition shows the core argument in the context of our scale-free algorithm in Section 6.2. Analogous results hold for the parameter-free algorithm in Section 6.1 and the optimistic algorithm in Section 6.3 using an identical argument.

**Proposition 6.4.1.** *Under the same assumptions as Theorem 6.2.1, suppose we instead set $\alpha_t = \epsilon$ for all $t$. Then*

$$R_T(u) \le O\left(\epsilon\sqrt{\|g\|_{1:T}^2} + \|u\|\left[\sqrt{\|g\|_{1:T}^2 \log\left(\frac{\|u\|}{\epsilon} + 1\right)} \vee h_T \log\left(\frac{\|u\|}{\epsilon} + 1\right)\right]\right).$$

*Proof.* Following the same arguments as Theorem 6.2.1 and recalling that $V_t = 4h_t^2 + \|g\|_{1:t-1}^2 \ge \|g\|_{1:t}^2$, we can bound

$$R_T(u) \le \psi_{T+1}(u) + \sum_{t=1}^{T} \frac{2\alpha_t \left\|g_t\right\|^2}{\sqrt{V_t}} \le \psi_{T+1}(u) + 2\epsilon \sum_{t=1}^{T} \frac{\left\|g_t\right\|^2}{\sqrt{\left\|g\right\|_{1:t}^2}} \le \psi_{T+1}(u) + 4\epsilon\sqrt{\left\|g\right\|_{1:T}^2},$$

where the last line invokes Lemma A.3.3. The result then follows by using the same argument as Theorem 6.2.1 to bound $\psi_{T+1}(u) \le O\Big( \left\|u\right\| \Big[\sqrt{\left\|g\right\|_{1:T}^2 \log\left(\left\|u\right\|/\epsilon + 1\right)} \vee h_T \log\left(\left\|u\right\|/\epsilon + 1\right) \Big]\Big)$. □

Interestingly, as observed by Z. Zhang, H. Yang, et al. 2023, the scale-free guarantees in the form of Equation (6.2) naturally avoid the range-ratio problem. Indeed, Proposition 6.4.1 requires neither the restarting strategy of FreeGrad nor the soft-restarting scheme of our scale-free algorithm in Theorem 6.2.1. This is because in order to achieve a scale-free version of the standard parameter-free guarantee (Equation (6.1)), we must balance out the gradient "units" in the logarithm term of the regularizer, and this unit-balancing is what gives rise to the range-ratio problem. By instead setting $\alpha_t = \epsilon$, no such unit correction is needed and the range-ratio problem is naturally avoided.

More generally, it is possible to achieve any of the intermediate results between the two types of parameter-free guarantee using a similar argument to Proposition 6.4.1. The following theorem provides a result analogous to Theorem 6.1.1, and is proven in Appendix B.1.4. An equivalent result also holds for our optimistic algorithm in Section 6.3, which we defer to Appendix B.1.4.

**Theorem 6.4.2.** *Under the same assumptions as Theorem 6.1.1, let $\rho \in [0, \frac{1}{2})$ and suppose we set $\alpha_t = \epsilon G^{2\rho}/V_t^{\rho}$ for all $t$. Then for all $u \in \mathbb{R}^d$, Algorithm 2 guarantees*

$$R_T(u) \le O\left( \frac{\epsilon G^{2\rho}}{1-2\rho} V_{T+1}^{\frac{1}{2}-\rho} + \left\|u\right\| \left[ \sqrt{V_{T+1} \log\left( \frac{\left\|u\right\| V_{T+1}^{\rho}}{\epsilon G^{2\rho}} + 1 \right)} \vee G \log\left( \frac{\left\|u\right\| V_{T+1}^{\rho}}{\epsilon G^{2\rho}} + 1 \right) \right] \right),$$

*where $V_{T+1} \le O(\left\|g\right\|_{1:T}^2)$.*

Hence, at $\rho = 0$ we have the bound matching the bound in Z. Zhang, H. Yang, et al. 2023 up to constant terms, and as $\rho \to \frac{1}{2}$ we move toward the usual parameter-free bound, Equation (6.1). It should be noted that this result is complimentary to Theorem 6.1.1, rather than a generalization: the leading term blows up as we approach $\rho = \frac{1}{2}$. This is unsurprising, as the $\log\log(T)$ penalty incurred by setting $\alpha_t = \frac{\epsilon}{\sqrt{V_t}\log^2(V_t/G^2)}$ in Theorem 6.1.1 is necessary — without this $\log\log(T)$ dependence, it would be possible to use the regret guarantee to contradict the Law of Iterated Logarithm. Indeed, there are well-known connections between regret guarantees and concentration inequalities (Rakhlin and Sridharan 2017), and the regret guarantees of parameter-free algorithms in particular can be used to derive tight concentration inequalities matching the Law of Iterated Logarithm (see, *e.g.*, Orabona and Jun 2023).

A similar result can also be shown our scale-free algorithm (Proof in Appendix B.1.4).

**Theorem 6.4.3.** *Under the same assumptions as Theorem 6.2.1, let* $\rho \in [0, \frac{1}{2})$ *and suppose we set* $B_t^\rho = \left(4\sum_{s=1}^t \left[2^{\frac{1}{\rho}} + \sum_{s'=1}^{s-1} \frac{\|g_{s'}\|^2}{h_{s'}^2}\right]\right)^\rho$ *and* $\alpha_t = \epsilon / B_t^\rho$ *for all* $t$.[1] *Then for all* $u \in \mathbb{R}^d$*, Algorithm 2 guarantees*

$$R_T(u) \leq O\left(\frac{\epsilon h_T^{2\rho}}{1-2\rho} V_{T+1}^{\frac{1}{2}-\rho} + \|u\| \left[\sqrt{V_{T+1} \log\left(\frac{\|u\|\, B_{T+1}^\rho}{\epsilon} + 1\right)} \vee h_T \log\left(\frac{\|u\|\, B_{T+1}^\rho}{\epsilon} + 1\right)\right]\right)$$

*where and* $V_{T+1} \leq O(\|g\|_{1:T}^2)$.

## 6.5 Conclusions

In this chapter, we used our framework developed in Chapter 4 to design new parameter-free algorithms in the standard setting of $G$-Lipschitz losses. Our approach allows us to streamline existing results by avoiding applications of the doubling trick or related restart strategies, while also enabling improvements to the existing bounds (Sections 6.1 and 6.2). Moreover, our approach naturally leads to new parameter-free guarantees that could not be easily obtained using previous frameworks for parameter-free online learning such as coin-betting (Sections 6.3 and 6.4). In the next chapter, we will extend these techniques to relax the restrictive Lipschitz assumption, enabling parameter-free learning in several natural problem settings which have unbounded domains and losses.

[1] Note that $\lim_{\rho\to 0} B_t^\rho = 2$, so for $\rho = 0$ we allow an abuse of notation by letting $B_t^\rho := 2$ to avoid specifying separate cases.

# Chapter 7

# Beyond Lipschitz Losses

Up to this point, our discussion has revolved around parameter-free algorithms for Lipschitz loss functions — losses $\ell$ such that for all $w \in W$ and $g_w \in \partial\ell(w)$, $\|g_w\| \le G$ for some $G$. This is unfortunate because some of the most pervasive loss functions in machine learning are in fact *not* Lipschitz on an unbounded domain. The obvious example being prediction error type losses such as $\ell_t(w) = \frac{1}{2}\|y_t - w\|^2$ for some target vector $y_t$: clearly for this loss we have $\|\nabla\ell_t(w)\| = \|y_t - w\|$, which can be arbitrarily large when $\|w\|$ is unbounded, so $\ell_t$ is not Lipschitz.

Another use-case where the Lipschitz assumption tends to fail is in saddle-point optimization. Let $\mathcal{L} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a convex-concave function and consider the following min-max problem:

$$\min_{x^* \in \mathcal{X}} \max_{y^* \in \mathcal{Y}} \mathcal{L}(x^*, y^*).$$

A common approach to such problems is to reduce this problem into an online convex optimization problem with $g_t = (g_t^x, -g_t^y)^\top$ where $g_t^x \in \partial_x\mathcal{L}(x_t, y_t)$ and $g_t^y \in \partial_y\mathcal{L}(x_t, y_t)$ (see Section 7.2 for more details). In this setting, $\|g_t\|$ will tend to be unbounded if $\mathcal{X}$ and/or $\mathcal{Y}$ are unbounded. The reason being that most interesting saddle-point problem will contain some coupling between the variables $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. For instance, let $B$ be an arbitrary *coupling matrix* and consider a simple bilinear optimization problem such as $\mathcal{L}(x, y) = \langle x, By\rangle + f(x) - g(y)$ where $f(x)$ and $g(x)$ are Lipschitz convex functions. Here we have $\nabla_x\mathcal{L}(x, y) = By + \nabla f(x)$ — which can grow arbitrarily large for unbounded $\mathcal{Y}$ — and likewise $\nabla_y\mathcal{L}(x, y) = B^\top x - \nabla g(y)$ can grow without bound for unbounded $\mathcal{X}$.

Our goal in this chapter is to develop algorithms which achieve favorable regret guarantees when *both* the domain $W$ and the losses may be unbounded.

## 7.1 Online Learning with Quadratically Bounded Losses

In an unbounded domain with unbounded losses, it will generally be impossible to avoid linear regret without *some* additional assumptions. Intuitively, what's missing in this problem is a frame-of-reference for the magnitude of a given loss. In the Lipschitz or bounded-range settings, the learner always has a frame-of-reference for the worst-case loss they might encounter. In contrast, without these assumptions, hindsight becomes the only frame-of-reference, and the adversary can exploit this to "trick" the learner into playing too greedily or too conservatively.

To make the problem tractable, yet still allowing the losses to have unbounded range and subgradients, we assume that the subgradients are bounded for all $t$ at *some* reference point $w_0$, but may become arbitrarily large away from $w_0$. This effectively gives the learner access to an *a priori* frame-of-reference for loss magnitudes, yet still captures many problem settings where the losses can become arbitrarily large in an unbounded domain.

**Definition 7.1.1.** Let $(W, \|\cdot\|)$ be a normed space. A function $\ell : W \to \mathbb{R}$ is $(G, L)$-quadratically bounded *w.r.t.* $\|\cdot\|$ at $w_0$ if for any $w \in W$ and $\nabla\ell(w) \in \partial\ell(w)$ it holds that

$$\|\nabla\ell(w)\| \le G + L\,\|w - w_0\|\,. \tag{7.1}$$

Note that Definition 7.1.1 is a strict generalization of the standard Lipschitz condition: any $G$-Lipschitz function is $(G, 0)$-quadratically bounded. The definition also captures $L$-smooth functions as a special case, since any $L$-smooth function is $(\|\partial\ell_t(w_0)\|, L)$-quadratically bounded at $w_0$. However, in general a function satisfying the quadratically bounded property need not be smooth. As a simple illustration, note that if $f(w)$ is an $L$-smooth and $(G, L)$-quadratically bounded function, then $f(w) + c\|w\|$ will be $(G + c, L)$ quadratically bounded but non-smooth. For the remainder of the paper we assume without loss of generality that $w_0 = \mathbf{0}$ and $\|\cdot\|$ is the Euclidean norm.

The quadratically bounded assumption was initially studied in the context of stochastic optimization by Telgarsky (2022), where it was sufficient to attain convergence in several settings of practical relevance. In this work, we show that it is also sufficient to achieve sublinear regret even in *adversarial* problem settings. We will in fact take it one step further and consider a stronger Online Linear Optimization (OLO) version of the problem. We say that a *sequence* $\{g_t\}$ is $(G_t, L_t)$-quadratically bounded *w.r.t.* $\{w_t\}$ if for every $t$ we have $\|g_t\| \le G_t + L_t\|w_t\|$. Then using the standard reduction from OCO to OLO, for any sequence of $(G_t, L_t)$-quadratically bounded convex functions we have the following regret upper bound:

$$R_T(u) = \sum_{t=1}^{T} \ell_t(w_t) - \ell_t(u) \le \sum_{t=1}^{T} \langle g_t, w_t - u \rangle\,,$$

where $g_t \in \partial\ell_t(w_t)$ and $\{g_t\}$ is a $(G_t, L_t)$-quadratically bounded sequence *w.r.t.* $\{w_t\}$. Hence, one can solve OCO problems involving quadratically bounded losses using any OLO algorithm that achieves sublinear regret against sequences $\{g_t\}$ that are quadratically bounded *w.r.t.* its outputs $\{w_t\}$. Note that this is potentially a more difficult problem, as it gives the adversary freedom to impose severe penalties whenever the learner plays large $w_t$, yet this effect is experienced asymmetrically by the comparator: the comparator can have large norm and not necessarily experience large losses unless $u$ is aligned with $g_t$ *and* the learner plays a point $\|w_t\| \propto \|u\|$. We refer to this harder problem setting as the QB-OLO setting, and QB-OCO for the setting where adversary must play $\ell_t$ satisfying Definition 7.1.1.

Surprisingly, it turns out that it is possible to achieve sublinear regret even in the QB-OLO setting. The following theorem provides an algorithm which achieves sublinear regret and requires no instance-specific hyperparameter tuning. Proof can be found in Appendix B.2.1.

**Theorem 7.1.2.** *Let $\mathcal{A}$ be an online learning algorithm and let $w_t \in W$ its output on round $t$. Let $\{g_t\}$ be a $(G_t, L_t)$-quadratically bounded sequence w.r.t. $\{w_t\}$, where $G_t \in [0, G_{\max}]$ and $L_t \in [0, L_{\max}]$ for all $t$. Let $\epsilon > 0$, $V_{t+1} = 4G_{\max}^2 + G_{1:t}^2$, $\rho_{t+1} = \frac{1}{\sqrt{L_{\max}^2 + L_{1:t}^2}}$, $\alpha_{t+1} = \frac{\epsilon G_{\max}}{\sqrt{V_{t+1}}\log^2(V_{t+1}/G_{\max}^2)}$. Denote $\Psi_t(w) = 3\int_0^{\|w\|} \min_{\eta \le \frac{1}{G_{\max}}} \left[\frac{\log(x/\alpha_t + 1)}{\eta} + \eta V_t\right] dx$ and set*

$$\psi_t(w) = \Psi_t(w) + \frac{2}{\rho_t}\|w\|^2, \quad \varphi_t(w) = \frac{L_t^2}{2\sqrt{L_{1:t}^2}}\|w\|^2.$$

*Then for any $u \in W$, Algorithm 6 guarantees*

$$R_T(u) \le \widehat{O}\left(\epsilon G_{\max} + \|u\|\sqrt{G_{1:T}^2 \log\left(\frac{\|u\|\sqrt{G_{1:T}^2}}{\epsilon G_{\max}} + 1\right)} + \|u\|^2\sqrt{L_{1:T}^2}\right),$$

*where $\widehat{O}(\cdot)$ hides constant and* $\log\log$ *terms.*

Let us briefly develop some intuition for how the above result is constructed. Algorithm 6 is an instance of Centered Mirror Descent (Algorithm 2), applied with a *linear* composite penalties $w \mapsto \langle \nabla\varphi_t(w_t), w\rangle$, which by Lemma 4.1.2 admits a generic regret guarantee of the form $R_T(u) \le \psi_T(u) + \sum_{t=1}^T \varphi_t(u) + \sum_{t=1}^T \delta_t$, where the $\delta_t$ are similar to the "stability" terms encountered in vanilla Mirror Descent, but with certain additional negative terms $\Delta_t$ and $\varphi_t$:

$$\delta_t \le O\Big(\langle g_t, w_t - w_{t+1}\rangle - D_{\psi_t}(w_{t+1}|w_t) - \Delta_t(w_{t+1}) - \varphi_t(w_t)\Big).$$

It's easily verified that that the $\psi_{T+1}(u) + \sum_{t=1}^T \varphi_t(u)$ terms in Theorem 7.1.2 match the terms in the stated upper bound, so the main difficulty is making sure that the stability terms $\sum_{t=1}^T \delta_t$

**Algorithm 6:** Algorithm for Quadratically Bounded Losses

```
Input: $\psi_1 : W \to \mathbb{R}_{\geq 0}$ with $\min_{w \in W} \psi_1(w) = 0$, $G_{\max}$ and $L_{\max}$
Initialize: $w_1 = \arg\min_{w \in W} \psi_1(w)$
for $t = 1 : T$ do
    Play $w_t$, observe $g_t \in \partial \ell_t(w_t)$
    Choose $G_t$ and $L_t$ satisfying $\|g_t\| \leq G_t + L_t \|w_t\|$
    Choose functions $\psi_{t+1}$, $\varphi_t$
    Set $\nabla \varphi_t(w_t) \in \partial \varphi_t(w_t)$ and $\widetilde{g}_t = g_t + \nabla \varphi_t(w_t)$
    Set $\Delta_t(w) = \psi_{t+1}(w) - \psi_t(w)$
    Update
        $$w_{t+1} = \arg\min_{w \in W} \langle \widetilde{g}_t, w \rangle + D_{\psi_t}(w | w_t) + \Delta_t(w)$$
end
```

disappear. Crucially, because $\{g_t\}$ is a $(G_t, L_t)$-quadratically bounded sequence *w.r.t.* $\{w_t\}$, we have $\|g_t\| \leq G_t + L_t \|w_t\|$. The utility of this is that we can design *separate regularizers* control the "Lipschitz part" $G_t$ and the "non-Lipschitz part" $L_t \|w_t\|$. In particular, using a similar argument to the one in Chapter 6, by setting $\Psi_t(w) = O\left(G_{\max} \|w\| \sqrt{T \log\left(\|w\| \sqrt{T}/\epsilon\right)}\right)$ we can ensure that the Lipschitz part of the bound is well-controlled:

$$\sum_{t=1}^{T} G_t \|w_t - w_{t+1}\| - D_{\Psi_t}(w_{t+1}|w_t) - \Delta_t(w_{t+1}) \leq O(1).$$

However, in general this $\Psi_t$ is not strong enough to control the non-Lipschitz part $L_t \|w_t\|$. Instead, for this term we use $\Phi_t(w) = O\left(L_{\max} \sqrt{T} \|w\|^2\right)$, and then using standard arguments for Mirror Descent with a strongly convex regularizer, it can be shown that

$$L_t \|w_t\| \|w_t - w_{t+1}\| - D_{\Phi_t}(w_{t+1}|w_t) - \varphi_t(w_t) \leq O\left(\frac{L_t \|w_t\|^2}{\sqrt{T}} - \varphi_t(w_t)\right) \leq 0$$

by choosing $\varphi_t(w_t) = O\left(\frac{L_t \|w_t\|^2}{\sqrt{T}}\right)$.

Note that in the setting of $G$-Lipschitz losses we have $L_{\max} = 0$ and hence set $G_t = \|g_t\|$, so the bound reduces to the usual parameter-free guarantee $R_T(u) \leq \widehat{O}\left(\|u\| \sqrt{\sum_{t=1}^{T} \|g_t\|^2 \log\left(\frac{\|u\|\sqrt{T}}{\epsilon} + 1\right)}\right)$ studied in the previous chapter, which is known to be optimal up to constant and log(log) terms (Mcmahan and M. Streeter 2012; Orabona 2013). On the other hand, if $L_{\max} > 0$ the algorithm can choose any $G_t \leq G_{\max}$ and $L_t \leq L_{\max}$ such that $G_t + L_t \|w_t\| \geq \|g_t\|$. Ideally these factors should be chosen to be tight — that is, to minimize $G + L \|w_t\|$ subject to the constraints $\{G \leq G_{\max}, L \leq L_{\max}, G + L \|w_t\| \geq \|g_t\|\}$. However, there may be many such $(G, L)$ satisfying these conditions, and in general it is unclear whether there exists a general-purpose strategy to choose among them without further assumptions. Indeed, Theorem 7.1.2 suggests that when $\|u\|$ is very large, we'd prefer to set the $L_t$'s smaller at the expense of large $G_t$'s, and vice-versa when $\|u\|$ is sufficiently small, so optimally trading off $G_t$ and $L_t$ would require some prior knowledge about $\|u\|$.

Nevertheless, there are many situations in which one can choose $(G_t, L_t)$ pairs along some pareto-frontier. As an illustrative example, consider an online regression setting in which $\ell_t(w) = \frac{1}{2}(y_t - \langle x_t, w\rangle)^2$ for some target variable $y_t \in \mathbb{R}$ and feature vector $x_t \in \mathbb{R}^d$. Observe that $\nabla\ell_t(w_t) = -(y_t - \langle x_t, w_t\rangle)x_t$, so setting $G_t = |y_t| \|x_t\|$ and $L_t = |\langle x_t, w_t / \|w_t\|\rangle| \|x_t\|$, we have

$$\|\nabla\ell_t(w_t)\| = \|(y_t - \langle x_t, w_t\rangle)x_t\| \le G_t + L_t \|w_t\|,$$

so $\{\nabla\ell_t(w_t)\}$ is a $(G_t, L_t)$-quadratically bounded sequence *w.r.t.* $\{w_t\}$, and Theorem 7.1.2 quarantees regret scaling as

$$\widetilde{O}\left(\|u\|\sqrt{\sum_{t=1}^{T} y_t^2 \|x_t\|^2} + \|u\|^2\sqrt{\sum_{t=1}^{T}\left\langle x_t, \frac{w_t}{\|w_t\|}\right\rangle^2 \|x_t\|^2}\right),$$

which is more adaptive to sequence of observed feature vectors $x_t$ and targets $y_t$ than the worst-case bound of $R_T(u) \le \widetilde{O}\left(\|u\| |y_{\max}| \|x_{\max}\| \sqrt{T} + \|u\|^2 \|x_{\max}\|^2 \sqrt{T}\right)$.

Finally, notice that for $L_{\max} > 0$ Algorithm 6 suffers an additional $O(L_{\max} \|u\|^2 \sqrt{T})$ penalty which is not present in the Lipschitz losses setting. The following theorem demonstrates that this penalty is in fact unavoidable in the QB-OLO setting. Proof can be found in Appendix B.2.1.

**Theorem 7.1.3.** *Let $\mathcal{A}$ be an algorithm defined over $\mathbb{R}^2$ and let $w_t$ denote the output of $\mathcal{A}$ on round $t$. Let $\epsilon > 0$ and suppose $\mathcal{A}$ guarantees $R_T(\mathbf{0}) \le \epsilon$ against any quadratically bounded sequence $\{g_t\}$. Then for any $T \ge 1$, $G > 0$ and $L \ge 0$ there exists a sequence $g_1, \ldots, g_T$ satisfying $\|g_t\| \le G + L\|w_t\|$ and a comparator $u \in \mathbb{R}^2$ such that*

$$R_T(u) \ge \Omega\left(G\|u\|\sqrt{T\log\left(\|u\|\sqrt{T}/\epsilon\right)} \vee L\|u\|^2\sqrt{T}\right).$$

*Remark* 7.1.4. An alternative way to approach online learning in our problem setting would be to apply an algorithm which is both comparator-adaptive and Lipschitz-adaptive, since these algorithms do not require an *a priori* upper bound on $\|u\|$ nor on $\|g_t\|$. Theorem 7.1.3 demonstrates that this approach would be sub-optimal in our setting. Indeed, Mhammedi and Koolen (2020) show that without prior knowledge of a Lipschitz bound, there is an unavoidable $O(\|u\|^3 \max_{t\le T} \|g_t\|)$ penalty associated with comparator-norm adaptivity, which can lead to a sub-optimal $O(\|u\|^3 L \max_t \|w_t\|) \ge O(L\|u\|^3\sqrt{T})$ dependence in our problem setting.

**Algorithm 7:** Saddle-point Reduction

```
Input Domain $W = \mathcal{X} \times \mathcal{Y}$, OLO Algorithm $\mathcal{A}$
for $t = 1 : T$ do
    Get $w_t = (x_t, y_t) \in W$ from $\mathcal{A}$
    Receive $g_t^x \in \partial_x \mathcal{L}(x_t, y_t)$ and $g_t^y \in \partial_y[-\mathcal{L}(x_t, y_t)]$
    Send $g_t = (g_t^x, g_t^y)$ to $\mathcal{A}$ as the $t^{\text{th}}$ subgradient
end
Return $\overline{w}_T = \left(\frac{\sum_{t=1}^T x_t}{T}, \frac{\sum_{t=1}^T y_t}{T}\right)$
```

## 7.2 Unconstrained Saddle-point Optimization

As a result of the algorithm in the previous section, we are immediately able to produce a novel algorithm for saddle-point optimization in unbounded domains. Consider the following convex-concave saddle-point problem:

$$\inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \mathcal{L}(x, y) \tag{7.2}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are convex sets, $\mathcal{L}(\cdot, y)$ is convex for all $y \in \mathcal{Y}$, and $\mathcal{L}(x, \cdot)$ is concave for all $x \in \mathcal{X}$. Solutions to Equation (7.2) are captured by the notion of a *saddle-point*: a point $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ is called saddle-point of $\mathcal{L}$ if for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$ it satisfies

$$\mathcal{L}(x^*, y) \leq \mathcal{L}(x^*, y^*) \leq \mathcal{L}(x, y^*).$$

When such a point exists, it satisfies $\mathcal{L}(x^*, y^*) = \inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \mathcal{L}(x, y) = \sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} \mathcal{L}(x, y)$. Quality of a candidate solution $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is commonly measured in terms of the *duality gap*:

$$G(x, y) \overset{\text{def}}{=} \sup_{y^* \in \mathcal{Y}} \mathcal{L}(x, y^*) - \inf_{x^* \in \mathcal{X}} \mathcal{L}(x^*, y).$$

It can be shown that the duality gap is non-negative, and that any $(x, y) \in \mathcal{X} \times \mathcal{Y}$ such that $G(x, y) = 0$ must be a saddle-point (Boyd and Vandenberghe 2004). Fortunately, any such gap is easily controlled using an online learning algorithm via the well-known reduction to Online Linear Optimization (OLO) shown in Algorithm 7. We provide a simple proof in our notation for convience to the reader, though we emphasize that this is a very common lemma in the saddle-point literature.

**Lemma 7.2.1.** *For any $\mathring{w} = (\mathring{x}, \mathring{y}) \in \mathcal{X} \times \mathcal{Y}$, Algorithm 7 guarantees*

$$\mathcal{L}(\overline{x}_T, \mathring{y}) - \mathcal{L}(\mathring{x}, \overline{y}_T) \leq \frac{\sum_{t=1}^T \langle g_t, w_t - \mathring{w} \rangle}{T} = \frac{R_T^{\mathcal{A}}(\mathring{w})}{T}.$$

*Proof.* To see why this is true, observe that by convexity of $x \mapsto \mathcal{L}(x, y)$ and $y \mapsto -\mathcal{L}(x, y)$, we can apply Jensen's inequality in both arguments to get:

$$\mathcal{L}(\overline{x}_T, \mathring{y}) - \mathcal{L}(\mathring{x}, \overline{y}_T) \le \frac{1}{T}\left[\sum_{t=1}^{T} \mathcal{L}(x_t, \mathring{y}) - \mathcal{L}(\mathring{x}, y_t)\right]$$

now add and subtract $\mathcal{L}(x_t, y_t)$:

$$= \frac{\sum_{t=1}^{T} \mathcal{L}(x_t, y_t) - \mathcal{L}(\mathring{x}, y_t) - \mathcal{L}(x_t, y_t) + \mathcal{L}(x_t, \mathring{y})}{T}$$

let $g_t^x \in \partial_x \mathcal{L}(x_t, y_t)$ and $g_t^y \in \partial_y[-\mathcal{L}(x_t, y_t)]$ and again use convexity to upper bound both difference terms:

$$\le \frac{\sum_{t=1}^{T} \langle g_t^x, x_t - \mathring{x}\rangle + \langle g_t^y, y_t - \mathring{y}\rangle}{T}$$

and now define $w_t = (x_t, y_t)$, $\mathring{w} = (\mathring{x}, \mathring{y})$, and $g_t = (g_t^x, g_t^y)$ to complete the proof:

$$= \frac{\sum_{t=1}^{T} \langle g_t, w_t - \mathring{w}\rangle}{T} = \frac{R_T^{\mathcal{A}}(\mathring{w})}{T}.$$

□

Thus in order to control the duality gap $G(x, y)$, it suffices to provide any OLO algorithm that achieves sublinear regret under the given assumptions.

The only existing work to achieve a comparator-adaptive convergence guarantee for the duality gap in general saddle-point problems is Liu and Orabona (2022). Their approach does indeed guarantee a rate of the form $G(\overline{x}_T, \overline{y}_T) \le \frac{R_T^{\mathcal{A}}(w^*)}{T} \le \widetilde{O}\left(\frac{G\|w^*\|}{\sqrt{T}}\right)$ under the assumption that the $\mathcal{L}(\cdot, \cdot)$ is $G$-Lipschitz in both arguments, which is justified by assuming that $\mathcal{X}$ and $\mathcal{Y}$ are bounded domains. However, generally saddle-point problems can have some coupling between the $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, leading to factors of $\|x\|$ and $\|y\|$ showing up in both $\|\nabla_x \mathcal{L}(x, y)\|$ and $\|\nabla_y \mathcal{L}(x, y)\|$. Thus, even in a bounded domain a bound of the form $R_T^{\mathcal{A}}(w^*) \le \widetilde{O}\left(\|w^*\| G\sqrt{T}\right)$ actually still falls short of being fully comparator-adaptive because the Lipschitz constant $G$ is subtly hiding factors of $D_{\mathcal{X}} = \max_{x, x' \in \mathcal{X}} \|x - x'\|$ and $D_{\mathcal{Y}} = \max_{y, y' \in \mathcal{Y}} \|y - y'\|$. See Section 7.2.1 for a more explicit example of this issue.

On the other hand, there are many interesting problems in which $\mathcal{L}(\cdot, \cdot)$ is quadratically bounded in both arguments, which will enable us to immediately apply Algorithm 6 to the linear losses $g_t = (g_t^x, g_t^y)$ as described above. In particular, we have the following:

**Proposition 7.2.2.** *Suppose that for all $\widetilde{y} \in \mathcal{Y}$, the function $x \mapsto \mathcal{L}(x,\widetilde{y})$ is $(G_x + L_{xy}\|\widetilde{y}\|, L_{xx})$-quadratically bounded, and for all $\widetilde{x} \in \mathcal{X}$ the function $y \mapsto -\mathcal{L}(\widetilde{x},y)$ is $(G_y + L_{yx}\|\widetilde{x}\|, L_{yy})$-quadratically bounded. Let $g_t^x \in \partial_x\mathcal{L}(x_t,y_t)$ and $g_t^y \in \partial_y[-\mathcal{L}(x_t,y_t)]$, and set $g_t = (g_t^x, g_t^y)$. Then $\{g_t\}$ is a $(G_w, L_w)$-quadratically bounded sequence w.r.t. norm $\|(x,y)\| = \sqrt{\|x\|^2 + \|y\|^2}$, where $G_w \le O\left(\sqrt{G_x^2 + G_y^2}\right)$ and $L_w \le O\left(\sqrt{L_{xx}^2 + L_{yy}^2 + L_{xy}^2 + L_{yx}^2}\right)$.*

*Proof.* Let $(x,y) \in W$. For $g_x \in \partial_x\mathcal{L}(x,y)$ observe that

$$\begin{aligned}\|g_x\|^2 &\le (G_x + L_{xy}\|y\| + L_{xx}\|x\|)^2\\ &\le 5\left(G_x^2 + L_{xy}^2\|y\|^2 + L_{xx}^2\|x\|^2\right),\end{aligned}$$

where the first line uses the assumption that $x \mapsto \mathcal{L}(x,y)$ is $(G_x + L_{xy}\|y\|, L_{xx})$ quadratically bounded for any $y \in \mathcal{Y}$ and the last line uses $(a+b+c)^2 \le 5a^2 + 5b^2 + 5c^2$. Likewise,

$$\|g_y\|^2 \le 5\left(G_y^2 + L_{yx}^2\|x\|^2 + L_{yy}^2\|y\|^2\right),$$

and so overall, letting $g_w = (g_x, g_y)$ we have

$$\begin{aligned}\|g_w\| &= \sqrt{\|g_x\|^2 + \|g_y\|^2}\\ &\overset{(\star)}{\le} \underbrace{\sqrt{5}\sqrt{G_x^2 + G_y^2}}_{=:G_w} + \underbrace{\sqrt{5}\sqrt{L_{xx}^2 + L_{yy}^2 + L_{xy}^2 + L_{yx}^2}}_{L_w}\sqrt{\|x\|^2 + \|y\|^2}\\ &= G_w + L_w\|w\|\end{aligned}$$

where $(\star)$ uses $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ for $x, y \ge 0$. □

Hence, with this in hand we can use Algorithm 6 to guarantee that for any $\mathring{w} = (\mathring{x}, \mathring{y}) \in W$,

$$\begin{aligned}\mathcal{L}(\overline{x}_T, \mathring{y}) - \mathcal{L}(\mathring{x}, \overline{y}_T) &\overset{\textit{Lemma } 7.2.1}{\le} \frac{R_T^{\mathcal{A}}(\mathring{w})}{T}\\ &\overset{\textit{Theorem } 7.1.2}{\le} \widetilde{O}\left(\frac{G_w\|\mathring{w}\| + L_w\|\mathring{w}\|^2}{\sqrt{T}}\right),\end{aligned}$$

which is indeed fully adaptive to comparator $\mathring{w}$.

It is important to note that our results in this section are made possible because our algorithm works even in the more difficult QB-OLO setting. It may be possible to get a similar result by using two QB-OCO algorithms designed for quadratically bounded functions $\ell_t$, though it it seems more challenging. In particular, letting $\ell_t^x(\cdot) = \mathcal{L}(\cdot, y_t)$ and $\ell_t^y(\cdot) = -\mathcal{L}(x_t, \cdot)$, one might instead run separate algorithms against the quadratically bounded loss sequences $\ell_t^x$ and $\ell_t^y$. However, now both algorithms need to very carefully regularize their iterates such that the gradients received by the other algorithm are never too large, since $\|\nabla\ell_t^x(x_t)\|$ may can contain factors of $\|y_t\|$ and $\|\nabla\ell_t^y(y_t)\|$ can contain factors of $\|x_t\|$. Hence careful coordination between the two algorithms will be required. The upshot is that by using un-linearized losses $\ell_t^x$ and $\ell_t^y$ it may be possible to get faster rates in some settings by better accounting for local curvature. We leave this as an exciting direction for future investigation.

### 7.2.1 Example: Bilinearly-coupled saddle-points

Before moving on, let us make things less abstract with a simple example. Consider a *bilinearly-coupled saddle-point* problem of the form

$$\mathcal{L}(x,y) = F_x(x) + H(x,y) - F_y(y) \tag{7.3}$$

where $F_x$ and $F_y$ are convex and $(\widetilde{G}_x,\widetilde{L}_x)$ and $(\widetilde{G}_y,\widetilde{L}_y)$-quadratically bounded respectively, and $H(x,y) = \langle x, By\rangle - \langle u_x, x\rangle + \langle u_y, y\rangle$ for some *coupling matrix* $B$ and vectors $u_x$, $u_y$. This problem captures several notable problem settings, such as minimizing the mean-squared projected bellman error for off-policy policy evaluation in reinforcement learning, quadratic games, and regularized empirical risk minimization (Du et al. 2022). The following proposition demonstrates that these problems do indeed satisfy the conditions of Proposition 7.2.2.

**Proposition 7.2.3.** *Equation (7.3) satisfies the assumptions of Proposition 7.2.2 with* $G_x = \widetilde{G}_x + \|u_x\|$*,* $L_{xx} = \widetilde{L}_x$*,* $L_{xy} = \|B\|_{op}$*,* $G_y = \widetilde{G}_y + \|u_y\|$*,* $L_{yy} = \widetilde{L}_y$*, and* $L_{yx} = \|B^\top\|_{op}$*.*

*Proof.* Observe that for any $(x,y) \in \mathcal{X} \times \mathcal{Y}$ and $g_x \in \partial_x\mathcal{L}(x,y)$, we have

$$\begin{aligned}\|g_x\| &= \|\nabla F_x(x) + By - u_x\| \\ &\le \|\nabla F_x(x)\| + \|B\|_{\text{op}}\|y\| + \|u_x\| \\ &\le \widetilde{G}_x + \|u_x\| + \widetilde{L}_x\|x\| + \|B\|_{\text{op}}\|y\|,\end{aligned}$$

where $\nabla F_x(x) \in \partial F_x(x)$ and $\|B\|_{\text{op}}$ denotes the operator norm $\|B\|_{\text{op}} = \sup_{x:\|x\|=1}\|Bx\|$. Likewise,

$$\|g_y\| \le \widetilde{G}_y + \|u_y\| + \widetilde{L}_y\|y\| + \left\|B^\top\right\|_{\text{op}}\|x\|.$$

Hence, $\mathcal{L}(\cdot,\cdot)$ satisfies the assumptions of Proposition 7.2.2 with $G_x = \widetilde{G}_x + \|u_x\|$, $L_{xx} = \widetilde{L}_x$, $L_{xy} = \|B\|_{\text{op}}$, $G_y = \widetilde{G}_y + \|u_y\|$, $L_{yy} = \widetilde{L}_y$, and $L_{yx} = \|B^\top\|_{\text{op}}$. □

We note that this specific example is mainly for illustrative purposes — in many instances of Equation (7.3) the functions $F_x$ and $F_y$ satisfy stronger curvature assumptions than used here, and our approach would be improved by more explicitly leveraging these assumptions when they hold. Nevertheless, our approach here does have a few key benefits: first, we naturally attain convergence in duality gap with an explicit dependence on the comparator, whereas prior works generally only attain a bound of this form making stronger assumptions such as strong convexity or one of the boundedness assumptions we're seeking to avoid (Liu and Orabona 2022; Du et al. 2022; Ibrahim et al. 2020; Azizian et al. 2020). Second, our approach can be applied under fairly weak assumptions: $\mathcal{L}(\cdot, \cdot)$ need not be Lipschitz, strongly-convex, nor smooth in either argument, and we do not require $\mathcal{X} \times \mathcal{Y}$ to be a bounded domain.

## 7.3 Conclusions

In this chapter, we developed the first parameter-free guarantees for a setting in which both the domain and the losses may be unbounded. Our results generalize the standard parameter-free bounds, and our lower bound demonstrates that the additional penalties incurred are unavoidable without further assumptions. As an application of our results, we develop novel saddle-point optimization algorithms which converge in duality gap even in unbounded decision sets. We obtain as a special case parameter-free algorithms for bilinearly-coupled saddle-point problems, which capture many problems of practical interest, such as minimizing the mean-squared projected bellman error for off-policy policy evaluation in reinforcement learning.

# Part III

# Adapting to Non-stationarity

# Chapter 8

# Overview of Part III

Now that we've had a taste of how to use our techniques, in Part III of this thesis we turn to the challenging problem of competing with an arbitrary *sequence* of comparators $\boldsymbol{u} = (u_1, \ldots, u_T)$, and develop algorithms which adapt to the comparator sequence without tuning any hyperparameters. Notably, these are the first algorithms in online learning that require *absolutely no prior knowledge about the comparator sequence.*

To understand this claim, note that prior works which study dynamic regret only do so in the bounded domain setting (Zinkevich 2003; T. Yang et al. 2016; Hall and Willett 2016; Gyorgy and Szepesvari 2016; L. Zhang, S. Lu, and Z.-H. Zhou 2018); this amounts to having strong prior knowledge of a radius $D$ for which $\|u_t - w_1\| \le D$ for all $t$. Alternatively, prior works which make meaningful guarantees in unbounded domains (*i.e.* the standard parameter-free online learning literature) only study static regret (Mcmahan and M. Streeter 2012; McMahan and Orabona 2014; Orabona and Pál 2016; Cutkosky and Orabona 2018); this amounts amounts to the strong prior knowledge that the comparator does not vary with time. In contrast, each of our results in this part of the thesis hold regardless of how large $\max_t \|u_t - w_1\|$ may be *and* each result achieves near-optimal performance against both fixed and time-varying comparators. In this sense, truly nothing about the comparator sequence needs to be known *a priori* in order to implement these algorithms or achieve their associated performance guarantees. Not only are these the first results to fully remove prior knowledge of the comparator sequence in online learning, we are able to extend these novelties outside the standard setting of Lipschitz losses, achieving optimal dynamic regret guarantees in two settings in which both the domain and the losses are unbounded.

This part of the thesis is composed of two main chapters: In Chapter 9 we revisit the Lipschitz losses and quadratically bounded losses problem settings from Part II. Then, in Chapter 10 we consider the related problem setting of online linear regression.

**Adapting to Non-stationarity with Lipschitz Losses.** We begin with the Lipschitz loss setting in Section 9.1. Our contributions in this setting are as follows:

- We develop an algorithm which guarantees dynamic regret

  $$R_T(\boldsymbol{u}) \le \widetilde{O}\left(\sqrt{(M + P_T)\sum_{t=1}^{T}\|g_t\|^2\|u_t - w_1\|}\right),$$

  where $M = \max_t \|u_t - w_1\|$ and $P_T = \sum_{t=2}^{T}\|u_t - u_{t-1}\|$. This result is the first non-trivial dynamic regret guarantee of any kind in unbounded domains, and the result matches the minimax optimal guarantee from the bounded domain setting up to logarithmic terms. Our result also naturally attains a new form of *per-comparator adaptivity* $\sum_{t=1}^{T}\|g_t\|^2\|u_t - w_1\|$, in which the variance penalty $\|g_t\|^2$ is removed on all rounds where our initial guess $w_1$ matches the comparator $u_t$. Even in bounded domains, prior works instead scale with the significantly worse $\max_{x,y\in W}\|x - y\|\sum_{t=1}^{T}\|g_t\|^2$.

- We additionally provide two useful dynamic regret reductions. In Section 9.1.1, we show that if one is willing to forego the per-comparator adaptivity above, $O(\sqrt{(M^2 + MP_T)\sum_{t=1}^{T}\|g_t\|^2})$ can be achieved using dynamic regret algorithms for bounded domains via a straight-forward application of the 1-dimensional reduction of Cutkosky and Orabona (2018). This bound still matches the minimax optimal guarantee from the bounded domain setting up to logarithmic terms. Then, in Section 9.1.2, we provide a simple reduction which reduces the per-round computation from $O(d\log(T))$ down to $O(d)$ *on average*, while still maintaining the optimal $\widetilde{O}\left(\sqrt{(M^2 + MP_T)\sum_{t=1}^{T}\|g_t\|^2}\right)$ guarantee up to poly-log terms.

**Adapting to Non-stationarity with Unbounded Losses.** Next, returning to the setting of quadratically bounded losses in Section 9.2, we make the following contributions:

- We design an algorithm which guarantees dynamic regret

  $$R_T(\boldsymbol{u}) \le \widetilde{O}\left(\sqrt{(M^2 + MP_T)\sum_{t=1}^{T}\left[G_t^2 + ML_t^2\right]}\right),$$

  where $M = \max_t \|u_t - w_1\|$, and $G_t$ and $L_t$ are constants satisfying $\|\nabla\ell_t(w)\| \le G_t + L_t\|w\|$ for any $\nabla\ell_t(w) \in \partial\ell_t(w)$ and $w \in W$. When the losses are $L_t$-smooth, the bound automatically improves to an $L^*$ bound of the form

  $$R_T(\boldsymbol{u}) \le \widetilde{O}\left(\sqrt{(M^2 + MP_T)\sum_{t=1}^{T}L_t\left[\ell_t(u_t) - \ell_t^*\right]}\right),$$

where $\ell_t^* = \min_w \ell_t(w)$. These are the first non-trivial dynamic regret guarantees to hold in settings where both the domain and the losses can be unbounded.

- We provide a matching lower bound demonstrating that these results are unimprovable in general.

**Online Prediction in the Complete Absence of Prior Knowledge.** Finally, in Chapter 10, we turn our attention to the related problem setting of online linear regression. This is an online learning problem with losses $\ell_t(w) = \frac{1}{2}(y_t - \langle x_t, w\rangle)^2$, where $y_t \in \mathbb{R}$ and $x_t \in \mathbb{R}^d$ is a vector of *features* which are observed at the beginning of the round. We develop the first algorithms for online linear regression that require *absolutely no prior knowledge* about the data stream, yet still make strong performance guarantees. In particular, our contributions are as follows:

- We show that even in the absence of any boundedness assumptions, a discounted variant of the Vovk-Azoury-Warmuth (VAW) forecaster with a well-chosen discount factor achieves dynamic regret $R_T(u_1, \ldots, u_T) \le O\Big(d \log(T) \vee \sqrt{dP_T^\gamma(\boldsymbol{u})T}\Big)$, where $P_T^\gamma(\boldsymbol{u})$ is a measure of variability of the comparator sequence (*i.e.* the magnitude of $P_T^\gamma(\boldsymbol{u})$ is related to how drastically the comparator changes over time). We also obtain *small-loss* guarantees of the form $R_T(\boldsymbol{u}) \le O\Big(d \log(T) \vee \sqrt{dP_T^\gamma(\boldsymbol{u}) \sum_{t=1}^T \ell_t(u_t)}\Big)$, so that the algorithm will automatically perform better on "easy" data where the comparator has low loss.
- We provide a matching lower bound of the form $R_T(\boldsymbol{u}) \ge \Omega\Big(d \log(T) \vee \sqrt{dTP_T^\gamma(\boldsymbol{u})}\Big)$, demonstrating optimality of the discounted VAW forecaster.
- We show that the discount factors required to obtain the results in the first point can be learned on-the-fly, leading to algorithms that make guarantees matching our lower bound. Moreover, we show how to extend our approach to achieve bounds of a similar form over *every sub-interval* $[a, b] \subseteq [1, T]$ *simultaneously.* This is a significantly stronger form of adaptivity which has previously only been achieved in bounded domains with Lipschitz losses. Our results are in fact the first strongly-adaptive guarantees to be achieved in the absence of boundedness assumptions.

# Chapter 9

# Non-stationarity in Online Learning

In this chapter we develop parameter-free dynamic regret guarantees in both the Lipschitz and quadratically bounded settings studied in Chapters 6 and 7. Our goal is to develop guarantees which are fully adaptive to arbitrary comparator *sequences*, requiring no prior knowledge about the magnitude or variability of the sequence, while still making near-optimal guarantees without any hyperparameter tuning.

Prior works studying dynamic regret guarantees for general OCO assume both bounded domains $W$ and bounded subgradients (Lipschitz losses). In this chapter, we develop the first dynamic regret guarantees for unbounded domains, under the standard Lipschitz assumption (Section 9.1). Our algorithm guarantees $R_T(\boldsymbol{u}) \le \widetilde{O}(\sqrt{(M^2 + MP_T)\sum_{t=1}^T \|g_t\|^2})$, where $M = \max_t \|u_t - w_1\|$, matching the minimax optimal guarantee (Equation (3.4)) from the easier bounded domain setting up to logarithmic terms. Then, in Section 9.2 we consider the $(G, L)$-quadratically bounded setting, in which both the domain and the losses may be unbounded. We develop an algorithm guaranteeing $\widetilde{O}(G\sqrt{MP_T T} + LM^{3/2}\sqrt{P_T T})$ dynamic regret, and provide a lower bound showing that this result is unimprovable without further assumptions.

## 9.1 Lipschitz Losses

We begin our study of dynamic regret guarantees in the $G$-Lipschitz loss setting. To get a feel for what we should expect in this setting, let us recall the results from the simpler bounded domain setting, in which $\sup_{x,y\in W} \|x - y\| \le D$. The minimax optimal dynamic regret in this setting is $R_T(\boldsymbol{u}) \le G\sqrt{(D^2 + DP_T)T}$, where $P_T = \sum_{t=2}^T \|u_t - u_{t-1}\|$ is the path-length of the comparator sequence (L. Zhang, S. Lu, and Z.-H. Zhou 2018). A matching guarantee can be made using an "online hyperparameter tuning" argument. The idea is simple: vanilla gradient descent with a fixed step-size $\eta$ guarantees (see, *e.g.*, Lemma 4.0.1)

$$R_T(\boldsymbol{u}) \leq \frac{D^2 + \sum_{t=2}^T \|w_t - w_1\| \, \|u_{t-1} - u_t\|}{2\eta} + \frac{\eta}{2}\sum_{t=1}^T \|g_t\|^2 \leq \frac{D^2 + DP_T}{2\eta} + \frac{\eta}{2}\sum_{t=1}^T \|g_t\|^2$$

where $g_t \in \partial \ell_t(w_t)$, and the final inequality uses the bounded domain assumption to bound $\|w_t - w_1\| \leq D$. Observe that the optimal $\eta$ would be $\eta^* = \sqrt{\frac{D^2 + DP_T}{\sum_{t=1}^T \|g_t\|^2}}$, which would yield regret which matches the minimax optimal bound in the worst-case:

$$R_T(\boldsymbol{u}) \leq \sqrt{(D^2 + DP_T)\sum_{t=1}^T \|g_t\|^2} \leq G\sqrt{(D^2 + DP_T)T}.$$

The approach of L. Zhang, S. Lu, and Z.-H. Zhou (2018) is to simply run many instances of gradient descent in parallel with different step-sizes, and combine their predictions using a mixture-of-experts approach. In particular, let $\eta_1, \ldots, \eta_N$ be a collection of step-sizes, and for each $i \in [N]$, let $\mathcal{A}_{\eta_i}$ denote an instance of gradient descent with step-size $\eta_i$ and let $w_t^{\eta_i}$ denote the output of $\mathcal{A}_{\eta_i}$ on round $t$. Let $\mathcal{A}_{\text{Meta}}$ denote an experts algorithm (*e.g.*, Hedge) which outputs a distribution $p_t(\eta)$ over the outputs $\{w_t^{\eta_i}\}_{i=1}^N$ on each round. On round $t$, we play $w_t = \sum_{i \in [N]} p_t(\eta_i) w_t^{\eta_i}$ and observe $g_t \in \partial \ell_t(w_t)$. Then, we update our predictions by passing $g_t$ to each of the $\mathcal{A}_{\eta_i}$ as the $t^{\text{th}}$ subgradient, and passing $\widetilde{\ell}_t = \left(\langle g_t, w_t^{\eta_1}\rangle, \ldots, \langle g_t, w_t^{\eta_N}\rangle\right)^\top \in \mathbb{R}^N$ to $\mathcal{A}_{\text{Meta}}$ as the $t^{\text{th}}$ loss vector. Using a geometrically spaced grid of $O(\log(T))$ different step-sizes, this approach allows one to guarantee the optimal dynamic regret up to a $\log(\log(T))$ factor.

Now returning to the unconstrained setting, there are two places where the above argument will fail. First, we can no longer bound $\|w_t - w_1\| \leq D$, and we will generally fail to produce meaningful guarantees by naively bounding $\sum_{t=2}^T \|w_t - w_1\| \, \|u_t - u_{t-1}\| \leq \max_t \|w_t - w_1\| P_T$ since $\|w_t - w_1\|$ can be arbitrarily large in unbounded domains. Moreover, even if we could argue that this term is bounded by some comparator-dependent quantity such as $\|w_t - w_1\| \leq \max_t \|u_t - w_1\|$, the mixture-of-experts part of this argument will fail due to being passed losses of unknown scale. Indeed, the loss vectors $\widetilde{\ell}_t$ passed will have components $\langle g_t, w_t^{\eta_i}\rangle$, which could be arbitrarily large.

Our approach will instead be to leverage properties of parameter-free guarantees to avoid these difficulties. Using the tools developed in Chapter 4, we'll first derive an algorithm which, for any $\eta \leq \frac{1}{G}$, guarantees $R_T(u) \leq \widetilde{O}\Big(\frac{P_T + \max_t \|u_t\|}{\eta} + \eta \sum_{t=1}^T \|g_t\|^2 \|u_t\|\Big)$. The key observation is that this bound has the property that $R_T(\mathbf{0}) \leq O(1)$ for any $\eta$, which will allow us to combine algorithms using a simple iterate adding approach instead of relying on a mixture-of-experts approach (Cutkosky 2019b). Indeed, suppose we run an instance of this algorithm $\mathcal{A}_\eta$ for each $\eta$ in some set $\mathcal{S} = \left\{\eta \in \mathbb{R} : 0 < \eta \leq \frac{1}{G}\right\}$, and on each round we play $w_t = \sum_{\eta \in \mathcal{S}} w_t^\eta$ where $w_t^\eta$ is the output of $\mathcal{A}_\eta$. Then for any arbitrary $\widetilde{\eta} \in \mathcal{S}$, we can write $\langle g_t, w_t - u_t\rangle = \Big\langle g_t, w_t^{\widetilde{\eta}} - u_t\Big\rangle + \sum_{\eta \neq \widetilde{\eta}} \langle g_t, w_t^\eta\rangle$, so the regret is bounded

**Algorithm 8:** Dynamic Regret Algorithm

```
**Input**: Lipschitz bound $G$, value $\varepsilon > 0$, step-sizes $\mathcal{S} = \left\{\frac{2^k}{G\sqrt{T}} \wedge \frac{1}{G} : 1 \leq k \leq \lceil \log_2 \sqrt{T} \rceil\right\}$
**Initialize**: $\epsilon = \frac{\varepsilon}{|\mathcal{S}|} = \frac{\varepsilon}{\lceil \log_2(\sqrt{T}) \rceil}$, $V_1 = 4G^2$, $w_1^\eta = \mathbf{0}$ and $\theta_t^\eta = \mathbf{0}$ for each $\eta \in \mathcal{S}$
**for** $t = 1 : T$ **do**
    Play $w_t = \sum_{\eta \in \mathcal{S}} w_t^\eta$, receive subgradient $g_t$
    Update $V_{t+1} = V_t + \|g_t\|^2$ and $\alpha_{t+1} = \frac{\epsilon G^2}{V_{t+1} \log^2(V_{t+1}/G^2)}$
    **for** $\eta \in \mathcal{S}$ **do**
        Set $\theta_t^\eta = \frac{2 w_t^\eta \log(\|w_t^\eta\|/\alpha_t + 1)}{\eta \|w_t^\eta\|} - g_t$    (with $\theta_t^\eta = -g_t$ if $w_t^\eta = \mathbf{0}$)
        Update $w_{t+1}^\eta = \frac{\alpha_{t+1}\theta_t^\eta}{\|\theta_t^\eta\|}\left[\exp\left[\frac{\eta}{2}\max(\|\theta_t^\eta\| - 2\eta\|g_t\|^2, 0)\right] - 1\right]$
    **end**
**end**
```

as

$$R_T(\boldsymbol{u}) \leq \sum_{t=1}^{T} \left\langle g_t, w_t^{\widetilde{\eta}} - u_t \right\rangle + \sum_{\eta \neq \widetilde{\eta}} \left[ \sum_{t=1}^{T} \langle g_t, w_t^\eta \rangle \right] = R_T^{\mathcal{A}_{\widetilde{\eta}}}(\boldsymbol{u}) + \sum_{\eta \neq \widetilde{\eta}} R_T^{\mathcal{A}_\eta}(\mathbf{0}) \leq O(R_T^{\mathcal{A}_{\widetilde{\eta}}}(\boldsymbol{u}) + |\mathcal{S}|).$$

Since this holds for an *arbitrary* $\widetilde{\eta} \in \mathcal{S}$, it must hold for the $\eta \in \mathcal{S}$ for which $R_T^\eta(\boldsymbol{u})$ is smallest, so we need only ensure that there is *some* near-optimal $\eta \in \mathcal{S}$, and that $|\mathcal{S}|$ is not too large. The latter condition is easily accomplished by setting $\mathcal{S}$ to be a geometrically-spaced grid such that $|\mathcal{S}| \leq O(\log(T))$. The base algorithms $\mathcal{A}_\eta$ and their corresponding regret guarantee are given in the following proposition.

**Proposition 9.1.1.** *Let $\ell_1, \dots, \ell_T$ be $G$-Lipschitz convex functions and $g_t \in \partial \ell_t(w_t)$ for all $t$. Let $\epsilon > 0$, $V_t = 4G^2 + \|g\|_{1:t-1}^2$, $\alpha_t = \frac{\epsilon G^2}{V_t \log^2(V_t/G^2)}$, and set $\psi_t(w) = 2\int_0^{\|w\|} \frac{\log(x/\alpha_t + 1)}{\eta} dx$ and $\varphi_t(w) = 2\eta \|g_t\|^2 \|w\|$. Then for any $u_1, \dots, u_T$ in $\mathbb{R}^d$, Algorithm 2 guarantees*

$$R_T(\boldsymbol{u}) \leq \widehat{O}\left( \epsilon + \frac{(M + P_T)\left[\log\left(\frac{MT^2\|g\|_{1:T}^2}{\epsilon G^2} + 1\right) \vee 1\right]}{\eta} + \eta \sum_{t=1}^{T} \|g_t\|^2 \|u_t\| \right),$$

*where $M = \max_t \|u_t\|$ and $\widehat{O}(\cdot)$ hides constant and* $\log(\log)$ *factors.*

The proof can be found in Appendix C.1.1, and again follows the intuition in Chapter 4: we first apply Lemma 4.0.1 to get $R_T(\boldsymbol{u}) \leq \psi_{T+1}(u_T) + \sum_{t=2}^T \mathcal{P}_t + \sum_{t=1}^T \varphi_t(u_t) + \sum_{t=1}^T \delta_t$. Unlike in Section 6.1, the regularizer $\psi_t$ generally does not grow fast enough for $\Delta_t(w) = \psi_{t+1}(w) - \psi_t(w)$ to ensure that $\delta_t \leq \widehat{\delta}_t$. Instead, we include an additional composite regularizer $\varphi_t$ in the update and show that this now ensures $\delta_t \leq \widehat{\delta}_t$, so that by Lemma 4.0.2 we have $\delta_t \leq \widehat{\delta}_t \leq \frac{2\|g_t\|^2}{\Psi_t''(0)} \leq 2\eta\alpha_t \|g_t\|^2$. Then we choose $\alpha_t$ to be small enough to ensure that $\sum_{t=1}^T \delta_t \leq O(1)$. We also now need to control the additional terms associated with the time-varying comparator, $\mathcal{P}_t = \langle \nabla \psi_t(w_t), u_{t-1} - u_t \rangle$. To handle these, we again exploit $\varphi_t$: by increasing it slightly more, we can decrease $\delta_t$ enough to cancel out the part of $\mathcal{P}_t$ which depends on $\|w_t\|$, so that this (unbounded!) quantity does not appear in the regret bound.

With this result in hand, we proceed to "tune" the optimal step-size by simply adding the iterates of a collection of these simple learners $\mathcal{A}_\eta$, as discussed above. The full algorithm is given in Algorithm 8, and the overall regret guarantee is given in Theorem 9.1.2 (with proof in Appendix C.1.1).

**Theorem 9.1.2.** *For any $\varepsilon > 0$ and $u_1, \ldots, u_T$ in $\mathbb{R}^d$, Algorithm 8 guarantees*

$$R_T(\boldsymbol{u}) \le \widehat{O}\left(\varepsilon G + \sqrt{(M + P_T)\sum_{t=1}^{T} \|g_t\|^2 \|u_t\| \log\left(\frac{MT^2 \|g\|_{1:T}^2}{\varepsilon G^2} + 1\right)} + GP_T \log\left(\frac{MT^2 \|g\|_{1:T}^2}{\varepsilon G^2} + 1\right)\right)$$

*where $M = \max_t \|u_t\|$ and $\widehat{O}(\cdot)$ hides constant and* $\log(\log)$ *factors.*

The bound achieved by Algorithm 8 is the first dynamic regret guarantee of any kind in unbounded domains. Further, Theorem 9.1.2 exhibits a stronger *per-comparator adaptivity* than previously obtained by depending on the individual comparators $\|u_t\|$, in contrast to the $R_T(\boldsymbol{u}) \le \widetilde{O}\left(\sqrt{(M^2 + MP_T)\sum_{t=1}^{T} \|g_t\|^2}\right)$ rate attained by prior works in bounded domains (L. Zhang, S. Lu, and Z.-H. Zhou 2018; Jadbabaie et al. 2015).

To see why this per-comparator adaptivity is interesting, let us consider a learning scenario in which there is a nominal "default" decision $\overline{u}$ which we expect to perform well *most* of the time, but may perform poorly during certain rare or unpredictable events. One example of such a situation is when one has access to an batch of data collected *offline*, which we can leverage to fit a parameterized model $\mathcal{M}(\overline{u})$ to the data to use as a baseline predictor. Deploying such a model online can be dangerous in practice because there may be certain events that are poorly covered by our dataset, leading to unpredictable behavior from the model. In this context, we can think of $\overline{u}$ as the learned model parameters, and without loss of generality we can assume $\overline{u} = \mathbf{0}$ (since otherwise we could just translate the decision space to be centered at $\overline{u}$). In this context, Theorem 9.1.2 tells us that Algorithm 8 will accumulate *no regret* over any intervals where we would want to compare performance against the baseline model, and over any intervals $[a, b]$ where the model is a poor comparison we are still guaranteed to accumulate no more than a $\widetilde{O}\left(\sqrt{(M^2 + MP_{[a,b]}) \|g\|_{a:b}^2}\right)$ penalty, where $P_{[a,b]} = \sum_{t=a+1}^{b} \|u_t - u_{t-1}\|$ is the path-length of any other arbitrary sequence of comparators over the interval $[a, b]$.

The property in the preceeding discussion is similar to the notion of *strong adaptivity* in the *constrained* setting, in which an algorithm guarantees the optimal static regret over all sub-intervals of $[1, T]$ *simultaneously* (Daniely, Gonen, and Shalev-Shwartz 2015; Jun, Orabona, et al. 2017). One might wonder if instead we should hope for the natural analog in the unconstrained setting: $R_{[a,b]}(u) = \sum_{t=a}^{b} \langle g_t, w_t - u\rangle \le \widetilde{O}(\|u\|\sqrt{b-a})$ for all $[a, b]$. Unfortunately, this natural analog is likely unattainable. To see why, notice that for all intervals $[a, b]$ of some fixed length $\tau = b - a$, we would require $R_{[a,b]}(\mathbf{0}) = \sum_{t=a}^{b} \langle g_t, w_t\rangle \le O(1)$, suggesting that no $w_t$ can be larger than some fixed constant (dependent on $\tau$). Yet clearly for large enough $T$ this cannot be guaranteed while simultaneously guaranteeing $R_T(u) \le O\left(\|u\| G\sqrt{T \log(\|u\| GT)}\right)$ for all $u \in \mathbb{R}^d$, since via reward-regret duality this entails competing against a fixed comparator $u \in \mathbb{R}^d$ with $\|u\| = O\Big(\exp(T)/\sqrt{T}\Big)$ in the worst-case, which can get arbitrarily large as $T$ increases. For this reason, we consider Theorem 9.1.2 to be a suitable relaxation of the strongly adaptive guarantee for unbounded domains.

### 9.1.1 A Simple Reduction for Dynamic Regret in Unbounded Domains

Interestingly, if one is willing to forego adaptivity to the individual $\|u_t\|$ observed in the previous section, it turns out that a dynamic regret bound of $R_T(\boldsymbol{u}) \le \widetilde{O}\left(\sqrt{(M^2 + MP_T)\|g\|^2_{1:T}}\right)$ can be achieved very simply using a generalization of the one-dimensional reduction of Cutkosky and Orabona 2018 to dynamic regret. Note however, that this approach fails to achieve the improved per-comparator adaptivity observed in Section 9.1. The following lemma shows that achieving the $R_T(\boldsymbol{u}) \le \widetilde{O}\left(\sqrt{(M^2 + MP_T)\|g\|^2_{1:T}}\right)$ bound in an unconstrained domain is essentially no harder than achieving it in a bounded domain, so long as one has access to an algorithm guaranteeing parameter-free *static* regret.

**Algorithm 9:** One-dimensional Reduction (Cutkosky and Orabona 2018)

```
Input: 1D online learning algorithm $\mathcal{A}_{1\mathrm{D}}$, online learning algorithm $\mathcal{A}_S$ with domain equal
    to the unit-ball $S \subseteq \{x \in \mathbb{R}^d : \|x\| \le 1\}$
for $t = 1 : T$ do
    Get point $x_t \in S$ from $\mathcal{A}_S$
    Get point $\beta_t \in \mathbb{R}$ from $\mathcal{A}_{1\mathrm{D}}$
    Play point $w_t = \beta_t x_t \in \mathbb{R}^d$, receive subgradient $g_t$
    Send $\widehat{g}_t = \langle g_t, x_t \rangle$ to $\mathcal{A}_{1\mathrm{D}}$ as the $t^{\mathrm{th}}$ loss
    Send $g_t$ to $\mathcal{A}_S$ as the $t^{\mathrm{th}}$ loss
end
```

**Lemma 9.1.3.** *Suppose that $\mathcal{A}_S$ guarantees dynamic regret $R_T^{\mathcal{A}_S}(\boldsymbol{u})$ for any sequence $u_1, \ldots, u_T$ in the unit-ball $S = \{w \in \mathbb{R}^d : \|w\| \le 1\}$ and suppose $\mathcal{A}_{1D}$ obtains static regret $R_T^{\mathcal{A}_{1D}}(u)$ for any $u \in \mathbb{R}$. Then for any $u_1, \ldots, u_T$ in $\mathbb{R}^d$, Algorithm 9 guarantees*

$$R_T(\boldsymbol{u}) = R_T^{\mathcal{A}_{1D}}(M) + M R_T^{\mathcal{A}_S}\left(\frac{\boldsymbol{u}}{M}\right)$$

*where $M = \max_{t \le T} \|u_t\|$.*

*Proof.* the proof follows the same reasoning as in the static regret case (Cutkosky and Orabona

2018, Theorem 2):

$$
\begin{aligned}
R_T(\boldsymbol{u}) &= \sum_{t=1}^{T} \langle g_t, w_t - u_t \rangle = \sum_{t=1}^{T} \langle g_t, \beta_t x_t - u_t \rangle \\
&= \sum_{t=1}^{T} \langle g_t, x_t \rangle \beta_t + \Big[ \langle g_t, x_t \rangle M - \langle g_t, x_t \rangle M \Big] - \langle g_t, u_t \rangle \\
&= \sum_{t=1}^{T} \langle g_t, x_t \rangle \beta_t - \langle g_t, x_t \rangle M + \sum_{t=1}^{T} \langle g_t, x_t \rangle M - \langle g_t, u_t \rangle \\
&= \sum_{t=1}^{T} \widehat{g}_t (\beta_t - M) + M \sum_{t=1}^{T} \left\langle g_t, x_t - \frac{u_t}{M} \right\rangle = R_T^{\mathcal{A}_{1\mathrm{D}}}(M) + M R_T\left(\frac{\boldsymbol{u}}{M}\right)
\end{aligned}
$$

□

Using this, one could let $\mathcal{A}_{1\mathrm{D}}$ be any parameter-free algorithm and let $\mathcal{A}_S$ be any algorithm which achieves the desired dynamic regret on a bounded domain. For instance, to get the optimal $\sqrt{P_T}$ dependence we can choose $\mathcal{A}_S$ to be the Ader algorithm of L. Zhang, S. Lu, and Z.-H. Zhou (2018), which will guarantee $MR_T^{\mathcal{A}_S}(\frac{\boldsymbol{u}}{M}) \leq O\left(MG\sqrt{T\left(1+\frac{P_T}{M}\right)}\right) = O\left(G\sqrt{(M^2+MP_T)T}\right)$.

### 9.1.2 Amortized Computation for Dynamic Regret

All known algorithms which achieve the optimal $O(\sqrt{TP_T})$ dynamic regret follow a similar construction, in which several instances of a simple base algorithm $\mathcal{A}$ are run simultaneously and their outputs combined in a way that guarantees dynamic regret approximately equal to an instance with a near-optimal choice of step-size. Generally $O(\log(T))$ instances of $\mathcal{A}$ are required to ensure that one of them has a near-optimal choice of step-size. Assuming the base algorithm $\mathcal{A}$ uses $O(d)$ computation per round, the full algorithm then requires $O(d\log(T))$ computation per round. Ideally we'd like to avoid this $\log(T)$ overhead.

A simple way to combat this computational overhead is to only update the algorithm every $O(\log(T))$ rounds, so that the *amortized* computation per round is $O(d)$ on average. The following proposition shows that $R_T(\boldsymbol{u}) \leq O\left(\sqrt{TP_T}\right)$ can be maintained up to poly-logarithmic terms using only $O(d)$ per-round computation on average by updating only at the end of itervals $I_k$ of length $O(\log(T))$. Proof can be found in Appendix C.1.1.

**Algorithm 10:** Lazy Reduction for Amortized Computation

```
Input: Algorithm $\mathcal{A}$, Disjoint intervals $I_1, \dots, I_K$ such that $\cup_{k=1}^K I_k \supseteq [1, T]$
Get $w_1$ from $\mathcal{A}$
Set $k = 1$
for $t = 1 : T$ do
    Play $w_t$, observe loss $g_t$
    if $t + 1 \notin I_k$ then
        Send $\widetilde{g}_k = \sum_{s \in I_k} g_s$ to $\mathcal{A}$
        Update $k \leftarrow k + 1$
        Get $w_{t+1}$ from $\mathcal{A}$
    else
        Set $w_{t+1} = w_t$
    end
end
```

**Proposition 9.1.4.** *Suppose $\mathcal{A}$ is an online learning algorithm which guarantees*

$$R_T^{\mathcal{A}}(\boldsymbol{u}) \leq \widetilde{O}\left(\sqrt{(M^2 + MP_T)\sum_{t=1}^T \|g_t\|^2}\right),$$

*for all $u_1, \dots, u_T$ in $\mathbb{R}^d$ with $\max_{t \leq T} \|u_t\| \leq M$. Then for all $u_1, \dots, u_T$ in $\mathbb{R}^d$, Algorithm 10 guarantees*

$$R_T(\boldsymbol{u}) \leq \widetilde{O}\left(\max_{k \leq K} |I_k| \sqrt{(M^2 + MP_T)\|g\|_{1:T}^2}\right)$$

## 9.2 Unbounded Losses

In this section, we return to the quadratically bounded losses setting introduced in Chapter 7. Here we consider in particular the QB-OCO setting, in which the losses satisfy $\|\nabla \ell_t(w)\| \leq G_t + L_t \|w\|$ for some $0 \leq G_t \leq G_{\max}$ and $0 \leq L_t \leq L_{\max}$, rather than the more difficult QB-OLO setting. In contrast to previous sections, in this section we consider full-information feedback, in which the learner observes the function $\ell_t(\cdot)$ (as opposed to first-order feedback $\nabla \ell_t(w_t) \in \partial \ell_t(w_t)$).

In the static regret setting, we saw in Section 7.1 that to control the stability of the algorithm it was necessary to add an additional term $\Phi_t(w) = O\left(L_{\max}\sqrt{T}\|w\|^2\right)$ to the regularizer to help control the "non-Lipschitz" part of the loss. We will likewise need a stronger regularizer to control the gradients for dynamic regret, but now it will lead to new difficulties. To see why, consider the dynamic regret of gradient descent with a fixed step-size $\eta$. The regret can be bound (*e.g.* using

Lemma 4.0.1) as

$$R_T(\boldsymbol{u}) \le O\left(\frac{\|u_T\|^2 + \max_t \|w_t\| P_T}{2\eta} + \frac{\eta}{2}\sum_{t=1}^T \|g_t\|^2\right), \tag{9.1}$$

where $g_t \in \partial \ell_t(w_t)$ and $P_T = \sum_{t=2}^T \|u_t - u_{t-1}\|$. In a bounded domain of diameter $D$, we can bound $\|u_T\|^2 \le D^2$ and $\max_t \|w_t\| \le D$, and then by optimally tuning $\eta$ we get

$$R_T(\boldsymbol{u}) \le O\left(\sqrt{(D^2 + DP_T)\sum_{t=1}^T \|g_t\|^2}\right)$$

which is optimal in the Lipschitz loss setting (L. Zhang, S. Lu, and Z.-H. Zhou 2018). More generally, using Mirror Descent with regularizer $\psi(w)$, one can derive an analogous bound:

$$R_T(\boldsymbol{u}) \le O\left(\psi(u_T) + \max_t \|\nabla\psi(w_t)\| P_T + \sum_{t=1}^T \delta_t\right),$$

where $\delta_t$ are the stability terms discussed in Section 7.1. In an unbounded domain, in Chapter 7 we used a regularizer of the form $\psi(w) = O\left(\|w\| \log\left(\|w\| T\right)/\eta\right)$, which enabled us to bound $\sum_{t=1}^T \delta_t \le O(\eta)$, and moreover, we showed that $\max_t \|\nabla\psi_t(w_t)\|$ can be bound from above by $O\left(\log\left(MT/\epsilon\right)/\eta\right)$ after adding a composite penalty $\varphi_t(w) = \eta \|g_t\|^2 \|w\|$ to the update. Then optimally tuning $\eta$ lead to regret scaling as

$$R_T(\boldsymbol{u}) \le \widetilde{O}\left(\sqrt{(M^2 + MP_T)\sum_{t=1}^T \|g_t\|^2}\right), \tag{9.2}$$

where $M = \max_t \|u_t\|$, which matches the bound from the bounded-domain setting up to logarithmic terms.

In the quadratically bounded setting, the situation gets significantly more challenging. As in Section 7.1, we will need to include an $O(\|w\|^2/\eta)$ term in the regularizer $\psi_t$ in order to control the "non-Lipschitz" part of the loss function. However, as above this leads to coupling $\max_t \|\nabla\psi_t(w_t)\| P_t = \max_t \|w_t\| P_T/\eta$ in the dynamic regret, and the term $\max_t \|w_t\|$ is generally too large to cancel out with additional regularization as done in Chapter 7. Even more troubling is that our lower bound in Theorem 9.2.2 suggests that the ideal dependence would be $O(MP_T/\eta)$, which we can only hope to achieve by constraining $\|w_t\|$ to a ball of diameter proportional to $M = \max_t \|u_t\|$. Yet $M$ is unknown to the learner!

Luckily, hope is not all lost. Taking inspiration from Luo, M. Zhang, et al. (2022), we can still attain a bound similar to Equation (9.2) by tuning the diameter of an artificial domain constraint. The approach is as follows: for each $(\eta, D)$ in some set $\mathcal{S}$, we run an instance of gradient descent

**Algorithm 11:** Dynamic Regret Algorithm

```
Input: $G_{\max}$, $L_{\max}$, weights $\beta_1, \dots, \beta_T$ in $[0,1]$, hyperparameter set
  $\mathcal{S} = \left\{(\eta, D) : \eta \le \frac{1}{8L_{\max}}, D > 0\right\}$, $p_1 \in \Delta_{|\mathcal{S}|}$.
for $\tau = (\eta, D) \in \mathcal{S}$ do
    Initialize: $w_1^{(\tau)} = \mathbf{0}$, $q_1(\tau) = p_1(\tau)$
    Define $\mu_\tau = \frac{1}{2D(G_{\max} + D/\eta)}$
    Define $\psi_\tau(x) = \frac{9}{2\mu_\tau} \int_0^x \log(v)\, dv$
end
for $t = 1 : T$ do
    Play $w_t = \sum_{\tau \in \mathcal{S}} p_t(\tau) w_t^{(\tau)}$, observe loss $\ell_t : W \to \mathbb{R}$
    Choose any reference point $\widetilde{w}_t \in W$ s.t. $\|\widetilde{w}_t\| \le D_{\min}$
    for $\tau = (\eta, D) \in \mathcal{S}$ do
        Query $g_t^{(\tau)} \in \partial \ell_t(w_t^{(\tau)})$
        Set $w_{t+1}^{(\tau)} = \Pi_{\{w \in W : \|w\| \le D\}} \left(w_t^{(\tau)} - \eta(1 + 8\eta L_t) g_t^{(\tau)}\right)$
        Define $\widetilde{\ell}_{t,\tau} = \ell_t(w_t^{(\tau)}) - \ell_t(\widetilde{w}_t)$
    end
    Set $q_{t+1} = \operatorname*{argmin}_{q \in \Delta_{|\mathcal{S}|}} \sum_{\tau \in \mathcal{S}} (\widetilde{\ell}_{t\tau} + \mu_\tau \widetilde{\ell}^2_{t\tau}) q_\tau + D_{\psi_\tau}(q_\tau | p_{t\tau})$
    Set $p_{t+1} = (1 - \beta_t) q_{t+1} + \beta_t p_1$.
end
```

$\mathcal{A}(\eta, D)$ which uses step-size $\eta$ and projects to the set $W_D = \{w \in W : \|w\| \le D\}$. Then, using a carefully designed experts algorithm, it is possible to ensure that the overall regret of the algorithm scales roughly as $R_T(\boldsymbol{u}) \le \widetilde{O}\left(R_T^{\mathcal{A}(\eta,D)}(\boldsymbol{u})\right)$ for *any* $(\eta, D) \in \mathcal{S}$. Thus if we can ensure that there is *some* $(\eta, D) \in \mathcal{S}$ for which $D \approx M$ and $\eta$ is near-optimal, then we'll be able to achieve dynamic regret with the desired $MP_T$ dependence. The following theorem, proven in Appendix C.1.2, characterizes an algorithm which achieves dynamic regret analogous to the above bounds, and in Theorem 9.2.2 we show that this is indeed unimprovable. Notably, our result also *automatically* improves to a novel $L^*$ bound when the losses are smooth.

**Theorem 9.2.1.** *For all $t$ let $\ell_t : W \to \mathbb{R}$ be a $(G_t, L_t)$-quadratically bounded convex function with $G_t \in [0, G_{\max}]$ and $L_t \in [0, L_{\max}]$. Let $\epsilon > 0$, $\beta_t \le 1 - \exp(-1/T)$ for all $t$, and for any $i, j \ge 0$ let $D_j = \frac{\epsilon}{T}\left[2^j \wedge 2^T\right]$ and $\eta_i = \left[\frac{\epsilon 2^i}{8(G_{\max}+\epsilon L_{\max})T} \wedge \frac{1}{8L_{\max}}\right]$, and let $\mathcal{S} = \{(\eta_i, D_j) : i, j \ge 0\}$. For each $\tau = (\eta, D) \in \mathcal{S}$ let $\mu_\tau = \frac{1}{2D(G_{\max}+D/\eta)}$ and set $p_1(\tau) = \frac{\mu_\tau^2}{\sum_{\tilde{\tau}\in\mathcal{S}} \mu_{\tilde{\tau}}^2}$. Then for any $\boldsymbol{u} = (u_1, \dots, u_T)$ in $W$, Algorithm 11 guarantees*

$$R_T(\boldsymbol{u}) \le O\left( \left[(M+\epsilon)\Lambda_T^* + P_T\right]\mathfrak{G}_{\max} + \sqrt{(M^2\Lambda_T^* + MP_T)\mathcal{L}_T} \right).$$

*where $\Lambda_T^* \le O\left(\log\left(\frac{MT\log(T)}{\epsilon}\right) + \log\left(\log\left(\frac{G_{\max}}{\epsilon L_{\max}}\right)\right)\right)$, $\mathfrak{G}_{\max} = G_{\max} + ML_{\max}$, $\mathcal{L}_T \le \sum_{t=1}^T \left[G_t^2 + ML_t^2\right]$, $P_T = \sum_{t=2}^T \|u_t - u_{t-1}\|$, and $M = \max_t \|u_t\|$. Moreover, when the losses are $L_t$-smooth, the terms $\mathcal{L}_T$ automatically improve to*

$$\mathcal{L}_T \le \min\left\{ \sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t^*\right], \sum_{t=1}^T \left[G_t^2 + ML_t^2\right] \right\}$$

Hiding constants and logarithmic terms, the first bound is effectively

$$R_T(\boldsymbol{u}) \le \widetilde{O}\left( \sqrt{(M^2 + MP_T)\sum_{t=1}^T G_t^2 + L_t^2 M^2} \right).$$

Notice that our result again generalizes the bounds established in prior works. Unfortunately, the result is not a strict generalization as Theorem 9.2.1 requires $L_{\max} > 0$ for the hyperparameter set $\mathcal{S}$ to be finite. To achieve a strict generalization, one can simply define a procedure which runs the algorithm Theorem 9.1.2 when $L_{\max} = 0$ and Algorithm 11 otherwise; this is possible because $L_{\max}$ must be provided as input to the algorithm. Notably, the algorithm of Section 9.1 does not use the aforementioned domain tuning trick and requires significantly less per-round computation as a result ($O(d\log(T))$ vs. $O(dT\log(T))$). We leave open the question of whether the exists a unifying analysis for $L_{\max} = 0$ and $L_{\max} > 0$, and whether the per-round computation can be improved.

As in Section 7.1, we again observe an additional penalty associated with non-Lipschitzness, this time on the order of $\widetilde{O}\left(M^{3/2}\sqrt{(M+P_T)\sum_{t=1}^T L_t^2}\right)$. The following theorem shows that these penalties are unavoidable in general (proof in Appendix C.1).

**Theorem 9.2.2.** *For any $M > 0$ there is a sequence of $(G, L)$-quadratically bounded functions with $\frac{G}{L} \le M$ such that for any $\gamma \in [0, \frac{1}{2}]$,*

$$R_T(\boldsymbol{u}) \ge \Omega\left(GM^{1-\gamma}\left[P_T T\right]^\gamma + LM^{2-\gamma}\left[P_T T\right]^\gamma\right).$$

*where $P_T = \sum_{t=2}^T \|u_t - u_{t-1}\|$ and $M \ge \max_t \|u_t\|$.*

Notice that with $\gamma = \frac{1}{2}$, we have $R_T(\boldsymbol{u}) \ge \Omega\left(G\sqrt{MP_T T} + LM^{3/2}\sqrt{P_T T}\right)$, matching our upper bound in Theorem 9.2.1 up to logarithmic terms. On the otherhand, for $\gamma = 0$ we have $R_T(\boldsymbol{u}) \geq GM + LM^2$, suggesting that the lower-order leading terms of our upper bound are also necessary. We also note that the assumption $G/L \leq M$ is without loss of generality: when $G/L \geq M$ one can construct a sequence of $(G+LM)$-Lipschitz losses according to existing lower bounds to show that

$$R_T(\boldsymbol{u}) \geq \Omega\left((G+LM)\sqrt{MP_TT}\right) = \Omega\left(G\sqrt{MP_TT} + LM^{3/2}\sqrt{P_TT}\right).$$

Interestingly, when the losses are smooth, the bound in Theorem 9.2.1 has the appealing property that it automatically improves to an $L^*$ bound of the form

$$R_T(\boldsymbol{u}) \leq \widetilde{O}\left(\sqrt{(M^2 + MP_T)\sum_{t=1}^{T} L_t\left[\ell_t(u_t) - \ell_t^*\right]}\right),$$

which matches bounds established in the Lipschitz and bounded domain setting up to logarithmic penalties (Zhao et al. 2020). This is the first $L^*$ bound that we are aware of to be achieved in an unbounded domain for general smooth losses without a Lipschitz or bounded-range assumption. Moreover, our bound features improved adaptivity to the *individial* $L_t$'s, scaling as $\sum_{t=1}^{T} L_t\left[\ell_t(u_t) - \ell_t^*\right]$ instead of the usual $L_{\max}\sum_{t=1}^{T}\ell_t(u_t) - \ell_t^*$ achieved by prior works (Srebro, Sridharan, and Tewari 2010; Orabona, Nicolo Cesa-Bianchi, and Gentile 2012; Zhao et al. 2020).

On the other hand, our upper bound bound contains terms of the form $\frac{G_{\max}}{L_{\max}\epsilon}$. Such ratios are unappealing in general because $G_{\max}$ and $L_{\max}$ are not under our control — it's possible for this ratio to be arbitrarily large. Fortunately, this ratio only shows up only in doubly-logarithmic terms, and hence these penalties can be regarded as effectively constant as far as the regret bound is concerned.

A more pressing issue is that the ratio $\frac{G_{\max}}{\epsilon L_{\max}}$ shows up in the number of experts. That is, setting $\mathcal{S}$ as in Theorem 9.2.1 requires a collection of $O(T\log_2(\sqrt{T}) + T\log_2\left(G_{\max}/L_{\max}\epsilon\right))$ experts, so in practice we can only tolerate $G_{\max}/L_{\max}\epsilon \leq \text{poly}(T)$ without increasing the (already quite high!) order of computation. We note that any algorithm that guarantees $R_T(\mathbf{0}) \leq G_{\max}\epsilon$ cannot hope to ensure non-vacuous regret when $G_{\max}/L_{\max}\epsilon > T$ anyways, so this seems to be a fundamental restriction in this setting. Nevertheless, the following result shows that if we know *a priori* that the losses will be smooth, then we can avoid this $\log\left(\log\left(\frac{G_{\max}}{L_{\max}\epsilon}\right)\right)$ penalty entirely and reduce the number of experts to $T\log_2(\sqrt{T})$ by instead setting $\eta_{\min} \propto \frac{1}{L_{\max}\sqrt{T}}$. Proof can be found in Appendix C.1.2.

**Theorem 9.2.3.** *For all $t$ let $\ell_t : W \to \mathbb{R}$ be $(G_t, L_t)$-quadratically bounded and $L_t$-smooth convex function with $G_t \in [0, G_{\max}]$ and $L_t \in [0, L_{\max}]$. Let $\epsilon > 0$ and for any $i, j \geq 0$ let $D_j = \frac{\epsilon}{\sqrt{T}}\left[2^j \wedge 2^T\right]$ and $\eta_i = \frac{1}{8L_{\max}\sqrt{T}}\left[2^i \wedge \sqrt{T}\right]$, and let $\mathcal{S} = \{(\eta_i, D_j) : i, j \geq 0\}$. Then for any $\boldsymbol{u} = (u_1, \ldots, u_T)$ in $W$, Algorithm 11 guarantees*

$$R_T(\boldsymbol{u}) \leq O\Bigg((M+\epsilon)\Big(L_{\max}P_T + \mathfrak{G}_{\max}\Lambda_T^*\Big) + \sqrt{\sum_{t=1}^{T}\left[\ell_t(u_t) - \ell_t^*\right]^2} + \sqrt{(M^2\Lambda_T^* + MP_T)\sum_{t=1}^{T} L_t\left[\ell_t(u_t) - \ell_t^*\right]}\Bigg),$$

*where* $\Lambda_T^* \leq O\left(\log\left(\frac{M\sqrt{T}\log\left(\sqrt{T}\right)}{\epsilon}\right)\right)$, $\mathfrak{G}_{\max} = G_{\max} + ML_{\max}$, $M = \max_t \|u_t\|$, *and* $P_T = \sum_{t=2}^{T}\|u_t - u_{t-1}\|$.

## 9.3 Conclusions

In this chapter, we developed parameter-free algorithms for dynamic regret. Our results in Section 9.1 are the first dynamic regret guarantees of any kind for unbounded domains, and likewise our results in Section 9.2 are the first for unbounded domains and losses. In both cases, our results automatically obtain near-optimal guarantees in *both* stationary and non-stationary settings using no instance-specific hyperparameter tuning, and require no prior knowledge of the magnitude of the comparator sequence. Thus, our algorithms are the first to completely remove *a priori* knowledge of the comparator sequence in general OCO.

An important open question is whether it is possible to reduce the computational or memory overhead of the algorithms designed here. All known algorithms which obtain an $O(\sqrt{P_T T})$ dependence, including our own, require at least $O(d\log(T))$ per-round computation and memory. This sort of horizon-dependent complexity can be prohibitively expensive for many settings in which one is concerned with adapting to non-stationarity, such as in continual learning settings. It may be possible to construct the set of step-sizes in a more on-the-fly manner, beginning with a small number of step-sizes and selectively adding more based on some empirical measure of performance. This way it may be possible to be both dynamic and efficient in *most* problems, and computationally expensive only in unreasonably hard problems. We leave these as exciting directions for future investigation.

# Chapter 10

# Non-stationarity in Online Linear Regression

This chapter presents new techniques and analyses for online linear regression, a variant of the classic least-squares regression problem tailored to streaming data (Azoury and Manfred K Warmuth 2001; Vovk 2001; Orabona, Crammer, and Nicolò Cesa-Bianchi 2015; Foster, Kale, and Karloff 2016). Formally, consider $T$ rounds of interaction between a learner and an environment, in which learner's objective is to accurately predict some observable target signal $y_t \in \mathbb{R}$ before it's revealed. On each round, a vector of *features* $x_t \in \mathbb{R}^d$ is first revealed, representing the context of the environment at the start of the round, and the learner predicts $\widehat{y}_t = \langle x_t, w_t \rangle$ by means of a weight vector $w_t \in \mathbb{R}^d$. The signal $y_t \in \mathbb{R}^d$ is then observed, and the learner incurs a loss proportional to the prediction error, $\ell_t(w_t) = \frac{1}{2}(y_t - \langle x_t, w_t \rangle)^2$. Since $w_t$ is allowed to depend on $x_t$, this protocol is sometimes referred to as *improper* online regression, as the learner is able to make predictions outside of the class of linear models. Indeed, since $x_t$ is revealed *before* the learner must make their prediction, it is always possible to make predictions $\widehat{y}_t = f_t(x_t)$ for any arbitrary transformation $f_t : \mathbb{R}^d \to \mathbb{R}$, for instance by setting $w_t = f_t(x_t) x_t / \|x_t\|^2$.

As in previous chapters, the classical measure of the learner's performance in this setting is *regret*, the cumulative prediction error relative to some fixed benchmark point $u \in \mathbb{R}^d$:

$$R_T(u) = \sum_{t=1}^{T} \ell_t(w_t) - \ell_t(u).$$

Notice that this performance measure can only properly reflect prediction accuracy when there exists a *fixed* $u \in \mathbb{R}^d$ which predicts well on average. For example, this may occur when when the $(x_t, y_t)$ pairs are all generated *i.i.d.* from some well-behaved distribution. However, in many true streaming settings the data-generating distribution may change over time due to changes in the environment. *Dynamic* regret attempts to model such settings by comparing against a *sequence* of comparators $\boldsymbol{u} = (u_1, \ldots, u_T)$:

$$R_T(\boldsymbol{u}) = \sum_{t=1}^{T} \ell_t(w_t) - \ell_t(u_t).$$

Notice that dynamic regret captures the usual notion of regret (referred to as *static* regret) as a special case by setting $u_1 = \ldots = u_T$. Our goal in this chapter is to make *favorable dynamic regret guarantees even in the complete absence of any prior knowledge of the underlying data-generating process.* Naturally, because such an algorithm leverages no prior knowledge, it necessarily must be adaptive to all problem-dependent quantities without requiring any instance-specific hyperparameter tuning.

**Related Works.** Despite being a well-studied problem setting, there are no prior works which approach online linear regression with sufficient generality to be considered free from prior knowledge. The closest works to our own are Vovk (2001), Azoury and Manfred K Warmuth (2001), Orabona, Crammer, and Nicolò Cesa-Bianchi (2015), and Mayo, Hadiji, and Erven (2022), each of which consider the same improper online learning setting as this work and present algorithms that can be run in an unbounded domain (hence requiring no prior knowledge about the comparator) and without any prior knowledge of the data stream. Yet these works provide guarantees that only hold for *static* regret—the *dynamic* regret of the algorithms in these works may be arbitrarily bad. In this sense, deploying any such algorithm implicitly requires rather strong prior knowledge: that the data-generating distribution is not changing over time.

A closely related problem setting which does account for potential non-stationarity is the classic *filtering* problem (Kalman 1960; Simon 2006; Kozdoba et al. 2019; Hazan and Singh 2022). This problem setting assumes that the $y_t$ are generated from a dynamical system of a specific form, and seeks to estimate the hidden state of the system. Thus, these works revolve around strong structural assumptions about the data-generating process from the outset. Similarly, there is a large literature on *adaptive* filtering which seeks to solve the filtering problem without *a priori* knowledge of the system (J. Kivinen, M. Warmuth, and Hassibi 2006; Hazan, Singh, and C. Zhang 2017; Hazan, Lee, et al. 2018; Rashidinejad, Jiao, and Russell 2020; Tsiamis and Pappas 2022; Ghai et al. 2020), though these works still implicitly require prior knowledge that the underlying dynamical system is from some specific class, as any performance guarantees may otherwise fail to hold.

Alternatively, there are several related problem settings that one might hope to leverage results from, but these all inevitably require additional assumptions of some form to be applied to the online linear regression problem. For instance, many prior works develop algorithms for general online regression settings that capture linear regression as a special case (Orabona, Crammer, and Nicolò Cesa-Bianchi 2015; Luo, Agarwal, et al. 2016; Kotłowski 2017; Kempka, Kotlowski, and

Manfred K. Warmuth 2019; Mhammedi and Koolen 2020). Even more generally, one might hope to approach online linear regression via reduction to a more general online convex optimization setting (L. Zhang, S. Lu, and Z.-H. Zhou 2018; Yuan and Lamperski 2019; Zhao et al. 2020; Baby, Hasson, and Y. Wang 2021; Baby and Y.-X. Wang 2021; Luo, M. Zhang, et al. 2022; Jacobsen and Cutkosky 2022; Z. Zhang, Cutkosky, and Y. Paschalidis 2023; Zhao et al. 2024). Unfortunately, all of these works require additional boundedness assumptions on the losses such as Lipschitzness or exp-concavity, both of which require a bounded domain in the context of losses $\ell_t(w) = \frac{1}{2}(y_t - \langle x_t, w\rangle)^2$. Yet assuming a bounded domain amounts amounts to having strong prior knowledge that the comparator sequence $\boldsymbol{u} = (u_1, \ldots, u_T)$ lies entirely within some bounded subset $W \subset \mathbb{R}^d$, which must be known and accounted for *a priori* for the guarantees to hold.

One recent exception to the limitations mentioned above is our contribution in the previous section, Section 9.2. Recall that in that section, we developed algorithms which can be applied to any loss functions satisfying $\|\nabla \ell_t(w)\| \le G_t + L_t \|w\|$ for some non-negative constants $G_t$ and $L_t$, and hence could be applied in the improper online regression setting by setting $G_t = |y_t| \|x_t\|$ and $L_t = \|x_t\|^2$. The algorithm in that section achieves a dynamic regret guarantee on the order of $O(M^{3/2}\sqrt{P_T T})$ where $M = \max_t \|u_t\|$ and $P_T = \sum_{t=2}^{T} \|u_t - u_{t-1}\|$. However, the approach still requires prior knowledge of $G_{\max} \ge G_t$ and $L_{\max} \ge L_t$ for all $t$ (and hence is not prior-knowledge-free), and later in this chapter we provide a lower bound demonstrating that the $M^{3/2}\sqrt{T}$ dependence is overly pessimistic in the improper online regression setting. Moreover the algorithm from that section requires $O(dT \log(T))$ per-round computation, making it inappropriate for many of the long-running problems where non-stationarity naturally emerges due to subtle changes in the environment over time.

## 10.1 The Vovk-Azoury-Warmuth Forecaster

In the context of *static* regret, it is well known that the optimal strategy in our improper online linear regression setting is the Vovk-Azoury-Warmuth (VAW) forecaster, discovered independently by Azoury and Manfred K Warmuth (2001) and Vovk (2001). On each round, the standard VAW forecaster sets

$$w_t = \left(\lambda I + \sum_{s=1}^{t} x_s x_s^\top\right)^{-1} \sum_{s=1}^{t-1} y_s x_s. \tag{10.1}$$

The VAW forecaster is well-known for the following regret guarantee (Azoury and Manfred K Warmuth 2001; Vovk 2001; Orabona, Crammer, and Nicolò Cesa-Bianchi 2015).

**Theorem 10.1.1.** *For any $u \in \mathbb{R}^d$ and any sequences $(y_t)_{t=1}^T$ in $\mathbb{R}$ and $(x_t)_{t=1}^T$ in $\mathbb{R}^d$, the VAW forecaster guarantees*

$$R_T(u) \le \frac{\lambda}{2}\|u\|_2^2 + \frac{d \max_t y_t^2}{2} \log\left(1 + \frac{\sum_{t=1}^T \|x_t\|_2^2}{\lambda d}\right),$$

Let us briefly pause to appreciate some of the subtleties of this result, as it represents a very high standard of excellence in online learning. First, note that the result holds using *no prior knowledge about the data* — there are no underlying assumptions about how the features $x_t$ or the targets $y_t$ are distributed, the algorithm requires no specific statistics or bounds such as $|y_t| \le Y$ or $\|x_t\| \le X$, and the algorithm works in an unbounded domain — a relative rarity in adversarial settings. Yet despite this incredible degree of generality, the VAW forecaster boasts a strong *logarithmic* regret guarantee, which can be shown to be optimal up to constant factors (See, *e.g.*, Nicolo Cesa-Bianchi and Lugosi (2006, Theorem 11.9)). Thus, the VAW forecaster achieves a harmony between theory and practice which is quite rare in online learning, requiring no problem-specific information or assumptions while still guaranteeing optimal regret.

However, a major caveat to the above discussion is that these favorable properties hold only within the context of *static* regret. The *dynamic* regret of the VAW forecaster can be arbitrarily bad in general. To see why, let us consider the simple case where $d = 1$ and $x_t = 1$ for all $t$. In this case, the VAW forecaster predicts $\widehat{y}_t = x_t w_t = (\lambda + t)^{-1} \sum_{s=1}^{t-1} y_s$, which approximates an empirical average of the targets observed up to round $t$. It is easy to see that any such prediction strategy can fail when competing with a time-varying comparator. For instance, if the first $T/2$ targets are $-1$ but the second half are $+1$, the VAW forecaster will quickly converge to predicting $-1$ in the first $T/2$ rounds, but will be unable to quickly adapt after the change in the latter $T/2$ rounds, leading to linear regret overall. In this sense, the VAW forecaster actually *implicitly requires quite strong prior knowledge* about the data: that it is, in some sense, *stationary.* Because of this, its predictions can *not* be trusted in the absence of prior knowledge, but rather only when the practitioner knows they are dealing with data that can be reasonably predicted using only a single fixed hypothesis $u \in \mathbb{R}^d$. In the next section, we will see that this issue can be alleviated by incorporating a suitable recency bias to the statistics of the VAW forecaster.

## 10.2 Dynamic Regret via Discounting

Despite making strong static regret guarantees, we saw in the previous section that the standard VAW forecaster may fail to attain low regret when competing against a time-varying comparator. Loosely speaking, the problem is that the VAW forecaster treats all time-steps as equally important.

Indeed, it can be shown that VAW forecaster can be understood as updating

$$w_t = \arg\min_{w\in\mathbb{R}^d} \frac{1}{2} \|w\|_{\Lambda_t}^2 + \sum_{s=1}^{t-1} \ell_s(w),$$

where $\Lambda_t = \lambda I + x_t x_t^\top$.[1] The latter term $\sum_{s=1}^{t-1} \ell_s(w)$ forces the VAW forecaster to choose a $w$ which balances all of the losses encountered so-far. Yet in dynamic scenarios, the losses that contain the most-relevant information for predicting $y_t$ are typically the ones that have been observed the most recently. In order to more closely track these recently-observed losses, we make two modifications to the VAW forecaster. First, we incorporate a *forgetting* or *discount* factor $\gamma$ in to the algorithm's statistics, placing less emphasis on losses observed far in the past. Second, we allow the update to additionally make use of a sequence of "predicted labels" or "hints" $\widetilde{y}_t$ that are available before we commit to $\widehat{y}_t$. Intuitively, we would like our algorithm to do better when $\widetilde{y}_t = y_t$. Later, we will provide some concrete ways to set $\widetilde{y}_t$ that yield strong regret bounds.

The variant of the VAW forecaster described above is provided concretely in Algorithm 12. Observe that by unrolling the recursions for $\theta_t$ and $\Sigma_t$, the update $w_t = \Sigma_t^{-1}[\widetilde{y}_t x_t + \gamma\theta_t]$ can be written in closed-form as

$$w_t = \left(\gamma^t \lambda I + \sum_{s=1}^{t} \gamma^{t-s} x_s x_s^\top\right)^{-1} \left[\widetilde{y}_t x_t + \gamma \sum_{s=1}^{t-1} \gamma^{t-1-s} y_s x_s\right].$$

By setting $\gamma = 1$ and $\widetilde{y}_t = 0$, the update precisely reduces to Equation (10.1), so the discounted VAW forecaster is a strict generalization of the standard VAW forecaster. Likewise, the following theorem shows that Algorithm 12 obtains a regret guarantee which captures Theorem 10.1.1 as a special case. Proof can be found in Appendix C.2.1.

**Theorem 10.2.1.** *Let $\lambda > 0$ and $\gamma \in (0,1]$. Then for any sequence $\boldsymbol{u} = (u_1, \ldots, u_T)$ in $\mathbb{R}^d$, Algorithm 12 guarantees*

$$\begin{aligned} R_T(\boldsymbol{u}) \le \frac{\gamma\lambda}{2} \|u_1\|_2^2 + \frac{d}{2} \max_t (y_t - \widetilde{y}_t)^2 \log\left(1 + \frac{\sum_{t=1}^T \gamma^{T-t} \|x_t\|_2^2}{\lambda d}\right) \\ + \gamma \sum_{t=1}^{T-1} \left[F_t^\gamma(u_{t+1}) - F_t^\gamma(u_t)\right] + \frac{d}{2} \log(1/\gamma) \sum_{t=1}^{T} (y_t - \widetilde{y}_t)^2 \end{aligned}$$

*where $F_t^\gamma(w) = \gamma^t \frac{\lambda}{2} \|w\|_2^2 + \sum_{s=1}^t \gamma^{t-s} \ell_s(w)$.*

The regret decomposition obtained in Theorem 10.2.1 is appealing for two reasons. First, it captures Theorem 10.1.1 as a special case: setting $\gamma = 1$, $\widetilde{y}_t = 0$, and $u_1 = \ldots = u_T = u$, the last two

[1] The equivalence to Equation (10.1) is readily checked via the first-order optimality condition, though this claim can also be derived as a special case of a more general claim Proposition C.2.1 provided in the appendix.

**Algorithm 12:** Discounted VAW Forecaster

```
Input: $\lambda > 0$, $\gamma \in (0, 1]$
Initialize: $w_1 = \mathbf{0}$, $\Sigma_0 = \lambda I$, $\theta_1 = \mathbf{0}$
for $t = 1 : T$ do
    Receive features $x_t \in \mathbb{R}^d$
    Set $\Sigma_t = x_t x_t^\top + \gamma \Sigma_{t-1}$, choose $\widetilde{y}_t \in \mathbb{R}$
    Update $w_t = \Sigma_t^{-1}\left[\widetilde{y}_t x_t + \gamma \theta_t\right]$

    Predict $\langle x_t, w_t \rangle$ and observe $y_t$
    Incur loss $\ell_t(w_t) = \frac{1}{2}(y_t - \langle x_t, w_t \rangle)^2$
    Set $\theta_{t+1} = y_t x_t + \gamma \theta_t$
end
```

terms of the bound evaluate to zero, so the regret is bounded by $\frac{\lambda}{2}\|u\|_2^2 + \frac{d}{2}\max_t y_t^2 \log\left(1 + \frac{\sum_{t=1}^T \|x_t\|_2^2}{\lambda d}\right)$, which is precisely the guarantee promised by Theorem 10.1.1. Second, the decomposition displays a clean separation of concerns. The terms in the first line are the unavoidable penalties associated with *static* regret, which are of course also unavoidable here in the more general dynamic regret setting. In the second line, any penalties incurred as a result of a changing comparator sequence are captured entirely by the *variability term* $\gamma \sum_{t=1}^T F_t^\gamma(u_{t+1}) - F_t^\gamma(u_t)$, while the term $d \log(1/\gamma) \sum_{t=1}^T \frac{1}{2}(y_t - \widetilde{y}_t)^2$ represents a *stability penalty* incurred due to discounting.

Intuitively, the terms in the second line represent a tracking/stability trade-off: against a volatile comparator sequence, we would ideally like to set the discount factor $\gamma$ to be small to control the variability penalty, yet this will come at the expense of increasing the stability penalty $d \log(1/\gamma) \sum_{t=1}^T \frac{1}{2}(y_t - \widetilde{y}_t)^2$. In its current form, however, this trade-off is still a bit mysterious. The variability term $\gamma \sum_{t=1}^{T-1} F_t^\gamma(u_{t+1}) - F_t^\gamma(u_t)$ is not necessarily monotonic as a function of $\gamma$ nor is it necessarily positive, making it difficult to meaningfully analyze or understand how it relates to the stability penalty $\frac{d}{2}\log(1/\gamma)\sum_{t=1}^T (y_t - \widetilde{y}_t)^2$. If we instead consider a modest upper bound on these terms we can reveal a more explicit trade-off. We provide proof of a slightly more general statement of the following lemma in Appendix C.2.1.

**Lemma 10.2.2.** *(simplified) Let $\ell_0, \ell_1, \ldots, \ell_T$ be arbitrary non-negative functions, $\gamma \in (0, 1)$, and $F_t^\gamma(w) = \sum_{s=0}^t \gamma^{t-s}\ell_s(w)$. For all t, define*

$$\bar{d}_t^\gamma(u, v) = \sum_{s=0}^t \frac{\gamma^{t-s}}{\sum_{s'=0}^t \gamma^{t-s'}}\left[\ell_s(u) - \ell_s(v)\right]_+$$

*and $P_T^\gamma(\boldsymbol{u}) = \sum_{t=1}^{T-1} \bar{d}_t^\gamma(u_{t+1}, u_t)$. Then for any $V_T \geq 0$,*

$$\gamma \sum_{t=1}^{T-1}\left[F_t^\gamma(u_{t+1}) - F_t^\gamma(u_t)\right] + \log\left(\frac{1}{\gamma}\right) V_T \leq \frac{\gamma}{1-\gamma} P_T^\gamma(\boldsymbol{u}) + \frac{1-\gamma}{\gamma} V_T$$

The lemma bounds the variability term $\gamma\sum_{t=1}^{T-1}\left[F_t^\gamma(u_{t+1}) - F_t^\gamma(u_t)\right]$ from Theorem 10.2.1 in terms of a new one $P_T^\gamma(\boldsymbol{u})$. To understand this new measure of variability, for each $t$ let us first define a $\gamma$-exponentially-decaying distribution over time-steps $s \le t$ as $p_t^\gamma(s) = \frac{\gamma^{t-s}}{\sum_{s'=0}^{t}\gamma^{t-s'}}$. Then, given $\gamma$ we can express $P_T^\gamma(\boldsymbol{u})$ as

$$\begin{aligned}
P_T^\gamma(\boldsymbol{u}) &= \sum_{t=1}^{T-1} \overbrace{\sum_{s=0}^{t} p_t^\gamma(s)[\ell_s(u_{t+1}) - \ell_s(u_t)]_+}^{\bar{d}_t^\gamma(u_{t+1},u_t)} \\
&= \sum_{t=1}^{T-1} \mathbb{E}_{s\sim p_t^\gamma}\Big[(\ell_s(u_{t+1}) - \ell_s(u_t))_+\Big],
\end{aligned}$$

so each term of $P_T^\gamma(\boldsymbol{u})$ is a measure of how different the prediction errors of $u_t$ and $u_{t+1}$ are on average across “recent” losses. The quantity $P_T^\gamma(\boldsymbol{u})$ can also be naively related to the more common measure of variability — the path-length $P_T^{\|\cdot\|} = \sum_{t=1}^{T-1}\|u_t - u_{t+1}\|$ — as follows:

$$\begin{aligned}
P_T^\gamma(\boldsymbol{u}) &\le \sum_{t=1}^{T-1} \max_s \|\nabla\ell_s(u_{t+1})\|\|u_t - u_{t+1}\| \\
&\le \max_{t,s}\|\nabla\ell_s(u_t)\| P_T^{\|\cdot\|} \le O\left(\max_t \|u_t\| \, P_T^{\|\cdot\|}\right).
\end{aligned}$$

Thus, $P_T^\gamma(\boldsymbol{u})$ is proportional to the usual path-length. Note that a multiplicative penalty of $\max_t\|u_t\|$ is the same worst-case penalty that appears in prior works, even in bounded domains (L. Zhang, S. Lu, and Z.-H. Zhou 2018; Jacobsen and Cutkosky 2022; Z. Zhang, Cutkosky, and Y. Paschalidis 2023; Zhao et al. 2024).

Letting $\eta = \frac{\gamma}{1-\gamma}$, Lemma 10.2.2 tells us that that latter terms of Theorem 10.2.1 are bounded by

$$\eta P_T^\gamma(\boldsymbol{u}) + \frac{d}{2\eta}\sum_{t=1}^{T}(y_t - \widetilde{y}_t)^2,$$

a trade-off which can be optimized by choosing $\eta = \sqrt{\frac{\frac{d}{2}\sum_{t=1}^{T}(y_t - \widetilde{y}_t)^2}{P_T^\gamma(\boldsymbol{u})}}$ to get

$$\eta P_T^\gamma(\boldsymbol{u}) + \frac{d}{2\eta}\sum_{t=1}^{T}(y_t - \widetilde{y}_t)^2 = 2\sqrt{dP_T^\gamma(\boldsymbol{u})\sum_{t=1}^{T}\frac{1}{2}(y_t - \widetilde{y}_t)^2}.$$

This is very promising; as we will see in Section 10.2.2, a penalty of this form is unavoidable in general. Plugging this choice of $\eta$ back into $\eta = \frac{\gamma}{1-\gamma}$ and solving for $\gamma$, we find that the ideal choice of discount factor would be a $\gamma \in [0,1]$ satisfying

$$\gamma = \frac{\sqrt{\frac{d}{2}\sum_{t=1}^{T}(y_t - \widetilde{y}_t)^2}}{\sqrt{\frac{d}{2}\sum_{t=1}^{T}(y_t - \widetilde{y}_t)^2} + \sqrt{P_T^{\gamma}(\boldsymbol{u})}}.$$

Notice in particular that $\gamma$ appears on both sides of the expression, and solving for this $\gamma$ explicitly is non-trivial in general. Nonetheless, the following theorem shows that a discount factor satisfying the above expression always exists, and if it could somehow be provided to the discounted VAW forecaster we would achieve dynamic regret matching the lower bound in Section 10.2.2. Proof can be found in Appendix C.2.1.

**Theorem 10.2.3.** *For any sequences $y_1, \dots, y_T$ and $\widetilde{y}_1, \dots, \widetilde{y}_T$ in $\mathbb{R}$ and any sequence $\boldsymbol{u} = (u_1, \dots, u_T)$ in $\mathbb{R}^d$, there is a discount factor $\gamma^* \in [0,1]$ satisfying*

$$\gamma^* = \frac{\sqrt{\frac{d}{2}\sum_{t=1}^{T}(y_t - \widetilde{y}_t)^2}}{\sqrt{\frac{d}{2}\sum_{t=1}^{T}(y_t - \widetilde{y}_t)^2} + \sqrt{P_T^{\gamma^*}(\boldsymbol{u})}} \tag{10.2}$$

*with which the regret of Algorithm 12 is bounded above by*

$$R_T(\boldsymbol{u}) \le O\left(d \max_t (y_t - \widetilde{y}_t)^2 \log(T) + \sqrt{dP_T^{\gamma^*}(\boldsymbol{u}) \sum_{t=1}^{T}(y_t - \widetilde{y}_t)^2}\right)$$

While this result is promising, it is important to note that it still falls short of our desired goal of prior-knowledge-free learning. Indeed, it seems that we require *exceptionally strong* prior knowledge to choose the prescribed discount factor $\gamma^*$ satisfying Equation (10.2). We will return to this issue in Section 10.3 to show that this discount factor can be learned on-the-fly, resulting in algorithms that *are* truly free of prior knowledge.

Interestingly, the discount factor $\gamma^*$ in Theorem 10.2.3 can help to shed some light on the variability measure $P_T^{\gamma^*}(\boldsymbol{u})$. Observe from the relation in Equation (10.2) that $\gamma^*$ can be near zero only when $P_T^{\gamma^*}(\boldsymbol{u})$ is very large relative to the stability penalty, and likewise, if $\gamma^*$ is near 1 then $P_T^{\gamma^*}(\boldsymbol{u})$ must be inconsequentially small. In this sense, the $P_T^{\gamma^*}(\boldsymbol{u})$ corresponding to small $\gamma^*$ can be regarded as the worst-case measures of variability. Yet as $\gamma^*$ approaches zero, $P_T^{\gamma^*}(\boldsymbol{u})$ approaches $\sum_{t=1}^{t-1}[\ell_t(u_{t+1}) - \ell_t(u_t)]_+$, which can be naturally related other standard measures of variability. Indeed, this penalty is similar in spirit to the temporal variability $\sum_{t=1}^{T-1}|\ell_{t+1}(u_t) - \ell_t(u_t)|$ studied in works such as Campolongo and Orabona (2021) and Besbes, Gur, and Zeevi (2015), and can be related to the path-length $\sum_{t=1}^{T-1}\|u_t - u_{t+1}\|$ via convexity of $\ell_t$. In this sense, $P_T^{\gamma^*}(\boldsymbol{u})$ can be thought of as a relaxation of the more common measures of variability.

### 10.2.1 Small-loss Bounds via Self-confident Predictions

In the previous section, we saw that the discounted VAW forecaster can achieve regret scaling as $O\left(\sqrt{dP_T^{\gamma^*}(\boldsymbol{u})\sum_{t=1}^T(y_t-\widetilde{y}_t)^2}\right)$, where $\widetilde{y}_t \in \mathbb{R}$ is an arbitrary "hint" available before observing the true $y_t$. One particularly interesting option is to use *the learner's own prediction* as a hint, $\widetilde{y}_t = \langle x_t, w_t\rangle$. The reasoning is that any learner achieving low dynamic regret must be predicting $y_t$ reasonably well on average, so their own predictions would naturally make for reasonable predicted labels $\widetilde{y}_t$. Concretely, observe that by choosing $\widetilde{y}_t = \langle x_t, w_t\rangle$ we would have $\sum_{t=1}^T(y_t-\widetilde{y}_t)^2 = \sum_{t=1}^T(y_t-\langle x_t, w_t\rangle)^2 = 2\sum_{t=1}^T \ell_t(w_t)$, and hence for some $\gamma \in [0,1]$ the guarantee in Theorem 10.2.3 would scale as

$$R_T(\boldsymbol{u}) = \sum_{t=1}^T \ell_t(w_t) - \ell_t(u_t) \le \widetilde{O}\left(\sqrt{dP_T^\gamma(\boldsymbol{u})\sum_{t=1}^T \ell_t(w_t)}\right),$$

where the $\widetilde{O}(\cdot)$ hides the logarithmic factor. Now notice that $\sum_{t=1}^T \ell_t(w_t)$ appears on both sides of this inequality. Solving for $\sum_{t=1}^T \ell_t(w_t)$, we find that $\sqrt{\sum_{t=1}^T \ell_t(w_t)} \le \widetilde{O}\left(\sqrt{dP_T^\gamma(\boldsymbol{u})} + \sqrt{\sum_{t=1}^T \ell_t(u_t)}\right)$, so plugging this back into the regret bound we have

$$R_T(\boldsymbol{u}) \le \widetilde{O}\left(P_T^\gamma(\boldsymbol{u}) + \sqrt{P_T^\gamma(\boldsymbol{u})\sum_{t=1}^T \ell_t(u_t)}\right).$$

Bounds of this form, sometimes called *small-loss* or $L^*$ bounds, are highly desirable because they naturally adapt to the total loss of the comparator sequence, potentially leading to lower regret than more naive hint choices such as $\widetilde{y}_t = y_{t-1}$ or $\widetilde{y}_t = 0$.

Unfortunately, the above argument does not quite go through because the now the logarithmic penalty in Theorem 10.2.3 scales as $O\left(d\max_t(y_t-\widetilde{y}_t)^2\log(T)\right) = O\left(d\max_t \ell_t(w_t)\log(T)\right)$, and this $\max_t \ell_t(w_t)$ could be arbitrarily large. Fortunately, it turns out that this issue can be remedied by a simple trust-region argument. On each round, instead of directly using hints $\widetilde{y}_t = \langle x_t, w_t\rangle$, we can constrain these predictions to be close to some arbitrary reference point $y_t^{\text{Ref}}$. In particular, in Lemma C.2.7 we show by clipping the learner's predictions to a suitable interval centered at $y_t^{\text{Ref}}$ we can guarantee $(y_t-\widetilde{y}_t)^2 \le O\left(\max_t(y_t-y_t^{\text{Ref}})^2 \wedge \ell_t(w_t)\right)$. This gives us the best-of-both-worlds: a similar self-bounding argument to above still yields a small-loss penalty $O\left(\sqrt{dP_T^\gamma(\boldsymbol{u})\sum_{t=1}^T \ell_t(u_t)}\right)$, while the logarithmic penalty can be bounded as $O\left(d\max_t(y_t-y_t^{\text{Ref}})^2\log(T)\right) \le O(d\max_t y_t^2\log(T))$ by setting $y_t^{\text{Ref}} = y_{t-1}$ or $y_t^{\text{Ref}} = 0$. The following theorem follows this above argument through, demonstrating that the discounted VAW forecaster can achieve small-loss bounds when using a well-chosen discount factor.

**Theorem 10.2.4.** *Let $y_t^{Ref} \in \mathbb{R}$ be an arbitrary reference point and let $\mathcal{B}_t = [y_t^{Ref} - M_t, y_t^{Ref} + M_t]$ for $M_t = \max_{s<t} \left| y_s - y_s^{Ref} \right|$. Suppose that we apply Algorithm 12 with hints $\widetilde{y}_t = \text{Clip}_{\mathcal{B}_t}(\langle x_t, w_t \rangle)$. Then for any sequence of losses $\ell_1, \ldots, \ell_T$ and any sequence $\boldsymbol{u} = (u_1, \ldots, u_T)$ in $\mathbb{R}^d$, there is a $\gamma^\circ \in [0,1]$ satisfying*

$$\gamma^\circ = \frac{\sqrt{d \sum_{t=1}^T \ell_t(u_t)}}{\sqrt{d \sum_{t=1}^T \ell_t(u_t)} + \sqrt{P_T^{\gamma^\circ}(\boldsymbol{u})}}. \tag{10.3}$$

*Moreover, running Algorithm 12 with discount $\gamma^\circ \vee \gamma_{\min}$ for $\gamma_{\min} = \frac{2d}{2d+1}$ ensures*

$$R_T(\boldsymbol{u}) \le O\left( d P_T^{\gamma_{\min}}(\boldsymbol{u}) + d \max_t (y_t - y_t^{Ref})^2 \log(T) + \sqrt{d P_T^{\gamma^\circ}(\boldsymbol{u}) \sum_{t=1}^T \ell_t(u_t)} \right),$$

Notice that unlike the previous section, there are two different variability penalties, $P_T^{\gamma^\circ}(\boldsymbol{u})$ and $P_T^{\gamma_{\min}}(\boldsymbol{u})$. The first mirrors the measure encountered in the last section. The other, $P_T^{\gamma_{\min}}(\boldsymbol{u})$, is rather annoying; in high dimensions $\gamma_{\min} = \frac{2d}{2d+1}$ is generally quite large, so $P_T^{\gamma_{\min}}(\boldsymbol{u})$ may evaluate losses at irrelevant comparators that are far away in time. Nevertheless, notice that this term satisfies $P_T^{\gamma_{\min}}(\boldsymbol{u}) \le \sum_{t=1}^{T-1} \max_s [\ell_s(u_{t+1}) - \ell_s(u_t)]_+$, a penalty which we will show is unavoidable in general in Theorem 10.2.5.

### 10.2.2 Dimension-dependent Lower Bound

In this section, we show that the regret penalties observed in the previous sections are unavoidable without further assumptions. The following lower bound is proven in Appendix C.2.3.

**Theorem 10.2.5.** *For any $d, T \ge 1$ and $P, Y > 0$ such that $dP \le 2TY^2$, there is a sequence of losses $\ell_t(w) = \frac{1}{2}(y_t - \langle x_t, w \rangle)^2$ and a comparator sequence $\boldsymbol{u} = (u_1, \ldots, u_T)$ satisfying $\max_t |y_t| \le Y$ and $\sum_{t=1}^{T-1} \max_s [\ell_s(u_{t+1}) - \ell_s(u_t)]_+ \le P$ such that*

$$R_T(\boldsymbol{u}) \ge \Omega\left( dY^2 \log(T) + dP + \sqrt{dP \sum_{t=2}^T (y_t - y_{t-1})^2} \right).$$

The key observation is that there is always a sequence of losses such that $\sum_{t=1}^T \ell_t(u_t) = 0$ can be ensured using only $T/d$ different comparators. Indeed, letting the features $x_t$ cycle through the standard basis vectors, for any sub-interval $[s, s+d] \subseteq [1, T]$ we can choose a single $u \in \mathbb{R}^d$ such that $\langle x_t, u \rangle = y_t$ for each $t$ in the interval. Then by sampling the $y_t$ randomly from $\{-Y\sigma, Y\sigma\}$ for some $\sigma \in [0,1]$, we can ensure variability of at most $O(TY^2\sigma^2/d) \le P$ but regret of at least $\Omega(TY^2\sigma^2) \ge \Omega\left( \sqrt{dP \left[ \sum_{t=1}^T (y_t - y_{t-1})^2 \vee dP \right]} \right)$.

Note that the condition $dP \leq 2TY^2$ captures a natural restriction of the problem setting, in that for larger $P$ the vacuous lower bound $R_T(\boldsymbol{u}) \geq \Omega(TY^2)$ can be constructed. Indeed, in the boundary case where $dP = 2TY^2$, Theorem 10.2.5 tells us that there is a sequence such that $R_T(\boldsymbol{u}) \geq \Omega\left(\sqrt{dP\mathcal{V}_T}\right) = \Omega\left(dP\right) = \Omega\left(TY^2\right)$. Yet this bound is achieved against *any* comparator sequence by the algorithm that naively predicts $\mathbf{0}$ on every round: $R_T(\boldsymbol{u}) = \sum_{t=1}^T \ell_t(\mathbf{0}) - \ell_t(u_t) \leq \sum_{t=1}^T \frac{1}{2}y_t^2 \leq \frac{1}{2}TY^2$. Hence, no lower bound can exceed $\frac{1}{2}TY^2$, so it is sufficient to consider comparator sequences with variability bounded by $P \leq 2TY^2$.

If we instead consider a more restricted problem setting by assuming a bounded domain, then the losses $\ell_t(w) = \frac{1}{2}(y_t - \langle x_t, w\rangle)^2$ can be considered to be exp-concave. In this setting, Baby and Y.-X. Wang (2021) have shown a lower bound of

$$R_T(\boldsymbol{u}) \geq \Omega\left(Y^{4/3}d^{1/3}T^{1/3}C_T^{2/3}\right), \tag{10.4}$$

where $C_T = \sum_{t=1}^{T-1} \|u_t - u_{t-1}\|_1$. A natural question is whether similar results also hold in the unbounded setting, and how they compare to our lower bound in Theorem 10.2.5. Note that even in the exp-concave setting, the bound in Equation (10.4) is not necessarily tight. Indeed, Baby and Y.-X. Wang (2021) provide an algorithm which guarantees

$$R_T(\boldsymbol{u}) \leq \widetilde{O}(Y^{4/3}d^{3.5}T^{1/3}C_T^{2/3}),$$

which does not match the lower bound *w.r.t.* the dimension $d$. In contrast, our lower bound in Theorem 10.2.5 matches our upper bounds in all involved quantities (see Sections 10.2 and 10.3). Regardless, we also demonstrate in Appendix C.2 that the same $\widetilde{O}(Y^{4/3}d^{3.5}T^{1/3}C_T^{2/3})$ upper bound can be attained, even in unbounded domains, using the strongly-adaptive guarantees developed in Section 10.4.

## 10.3 Learning the Optimal Discount Factor

Recall that our goal from the outset has been to design algorithms that achieve favourable dynamic regret guarantees using *no prior knowledge*. To this end, we showed in Section 10.2 that the discounted VAW forecaster can achieve dynamic regret guarantees of the form

$$R_T(\boldsymbol{u}) \leq O\left(\sqrt{dP_T^{\gamma}(\boldsymbol{u})T} \vee d\log\left(T\right)\right)$$

where $P_T^{\gamma}(\boldsymbol{u})$ is a certain measure of variability of the comparator sequence, and in Section 10.2.2 we showed that these penalties are unavoidable in general. However, these results hold under the assumption that the learner chooses discount rates satisfying special conditions (Equations (10.2) and (10.3)), either of which would require exceptionally strong prior knowlege to ensure. Indeed, the learner would need to know the future! In order to achieve our goal of learning in the complete absence of prior knowledge, we need to ensure that the learner can adequately guess or learn these ideal discount factors on-the-fly.

A common way to achieve runtime parameter-tuning of this sort would be to run many instances of the algorithm for different choices of $\gamma$ in parallel, and combine the predictions using a suitable meta-algorithm. In particular, suppose we have a collection of algorithms $\mathcal{A}_1, \ldots, \mathcal{A}_N$ and on each round we can query each $\mathcal{A}_i$ for a prediction $y_t^{(i)} \in \mathbb{R}$. Moreover, suppose we have a meta-algorithm $\mathcal{A}_{\text{Meta}}$ which tells us how to combine these predictions by outputting a $p_t$ from the $N$-dimensional simplex $\Delta_N$. Then by predicting $\overline{y}_t = \sum_{i=1}^N p_{ti} y_t^{(i)}$, [2] for any benchmark sequence $\boldsymbol{u} = (u_1, \ldots, u_T)$ and any $j \in [N]$ we have

$$\begin{aligned} R_T(\boldsymbol{u}) &= \sum_{t=1}^T \ell_t(\overline{y}_t) - \ell_t(u_t) \\ &= \underbrace{\sum_{t=1}^T \ell_t(y_t^{(j)}) - \ell_t(u_t)}_{=:R_T^{\mathcal{A}_j}(\boldsymbol{u})} + \underbrace{\sum_{t=1}^T \ell_t(\overline{y}_t) - \ell_t(y_t^{(j)})}_{=:R_T^{\text{Meta}}(e_j)} \end{aligned}$$

where the last line observes that $y_t^{(j)} = \langle x_t, w_t^{(j)} \rangle$. Hence, we may achieve our goal if we can ensure 1) that there is a $j \in [N]$ such that $\mathcal{A}_j$ uses a near-optimal discount factor $\gamma_j$, and 2) we can provide a meta-algorithm which guarantees low regret $R_T^{\text{Meta}}(e_j)$. We first investigate the latter point, and return to the former in Theorems 10.3.2 and 10.3.3.

The obvious approach to bounding the meta-algorithm's regret would be to observe that the losses $\ell_t(\overline{y}_t) = \frac{1}{2}(y_t - \overline{y}_t)^2$ are $\alpha_t$-exp-concave for $\alpha_t = \frac{1}{2 \max_i \ell_t(y_t^{(i)})}$ (Lemma C.2.8), which will allow us to apply an instance of the fixed-share algorithm (Nicolo Cesa-Bianchi, Gaillard, et al. 2012) to get:

$$R_T^{\text{Meta}}(e_j) \le O\left(\frac{\log(NT)}{\alpha_{T+1}}\right) \le O\Big(\max_{t,i} \ell_t(y_t^{(i)}) \log(NT)\Big),$$

as shown in Theorem C.2.12. However, just like in Section 10.2.1, the term $\max_{t,i} \ell_t(y_t^{(i)})$ is hard to quantify and could be be arbitrarily large in general. Fortunately the very same clipping trick used in Section 10.2.1 also works here: instead of having the meta-algorithm combine the *raw* predictions $y_t^{(i)}$, we can simply clip the predictions to a trust-region around a given reference point $y_t^{\text{Ref}}$. In Lemma C.2.9 we show that the clipping strategy detailed in Algorithm 13 incurs only an additional

[2] Recall from the introduction that because the features $x_t$ are provided at the start of the round, we can work directly in the output space $\mathbb{R}$ if we so choose by setting $w_t = \overline{y}_t x_t / \|x_t\|^2$. Hence, given $\overline{y} \in \mathbb{R}$ we allow a slight abuse of notation by letting $\ell_t(\overline{y}) = \frac{1}{2}(y_t - \overline{y})^2$.

**Algorithm 13:** Range-clipped Meta-algorithm

```
Input: Online learning algorithms $\mathcal{A}_1, \dots, \mathcal{A}_N$, experts algorithm $\mathcal{A}_{\text{Meta}}$ over the simplex $\Delta_N$.
Initialize: $\mathcal{A}_{\text{Meta}}, \mathcal{A}_1, \dots, \mathcal{A}_N$, and set $M_1 = 0$
for $t = 1 : T$ do
    Receive features $x_t$
    Choose reference point $y_t^{\text{Ref}}$
    Define $\mathcal{B}_t = [y_t^{\text{Ref}} - M_t, y_t^{\text{Ref}} + M_t]$
    for $i = 1, \dots, N$ do
        Send $x_t$ to $\mathcal{A}_i$
        Get prediction $y_t^{(i)} = \langle x_t, w_t^{(i)} \rangle$ from $\mathcal{A}_i$
        Compute $\overline{y}_t^{(i)} = \text{Clip}_{\mathcal{B}_t}(y_t^{(i)})$
    end
    Get $p_t \in \Delta_N$ from $\mathcal{A}_{\text{Meta}}$
    Predict $\overline{y}_t = \sum_{i=1}^N p_{ti} \overline{y}_t^{(i)}$ and observe $y_t$
    Update $M_{t+1} = M_t \vee |y_t - y_t^{\text{Ref}}|$

    Send $\ell_t(w) = \frac{1}{2}(y_t - \langle x_t, w \rangle)^2$ to $\mathcal{A}_i \ \forall i$
    Send $\ell_t(\overline{y}_t^{(1)}), \dots, \ell_t(\overline{y}_t^{(N)})$ to $\mathcal{A}_{\text{Meta}}$
end
```

constant penalty in the regret. Then, using Lemma C.2.7, using these clipped predictions leads to

$$R_T^{\text{Meta}}(e_j) \leq O(\max_t (y_t - y_t^{\text{Ref}})^2 \log(NT)).$$

Note that a penalty of a similar order is already present in the regret of the VAW forecaster (*e.g.* Theorem 10.2.1) so this result will be sufficient for our purposes. Overall, the following theorem formalizes the argument described above. We provide a simplified statement here for brevity, but the full statement and its proof can be found in Appendix C.2.4.

**Theorem 10.3.1.** *(simplified) Let $\mathcal{A}_{Meta}$ be the instance of fixed-share characterized in Theorem C.2.12. Then for any sequence $\boldsymbol{u} = (u_1, \dots, u_T)$ in $\mathbb{R}$ and any $j \in [N]$, Algorithm 13 guarantees*

$$R_T(\boldsymbol{u}) \leq \widehat{O}\left(R_T^{\mathcal{A}_j}(\boldsymbol{u}) + \max_t \left(y_t - y_t^{Ref}\right)^2 \log(NT)\right),$$

*where $\widehat{O}(\cdot)$ hides* $\log\log$ *terms.*

A similar target-clipping strategy was recently used by Mayo, Hadiji, and Erven (2022) to prove a static regret result for scale-free unconstrained online regression. Theorem 10.3.1 generalizes their approach by clipping to a trust-region of an arbitrary center $y_t^{\text{Ref}} \in \mathbb{R}$, and offers a somewhat streamlined argument which does not appeal to probabilistic notions such as mixibility.

Finally, with Theorem 10.3.1 in hand, we can achieve our desired result by running Algorithm 13 with the base algorithms $\mathcal{A}_i$ being instances of the discounted VAW forecaster with different discount factors $\gamma$. The following theorems show that for a well-chosen set of discount factors, we can make guarantees that match the bounds attained under oracle tuning of $\gamma$ (Theorems 10.2.3 and 10.2.4), yet require no prior knowledge of any sort. Proofs can be found in Appendix C.2.4 respectively.

**Theorem 10.3.2.** *Let $b > 1$, $\eta_{\min} = 2d$, $\eta_{\max} = dT$, and for all $i \in \mathbb{N}$ let $\eta_i = \eta_{\min} b^i \wedge \eta_{\max}$, and construct the set of discount factors $\mathcal{S}_\gamma = \left\{\gamma_i = \frac{\eta_i}{1+\eta_i} : i \in \mathbb{N}\right\} \cup \{0\}$. For any $\gamma$ in $\mathcal{S}_\gamma$, let $\mathcal{A}_\gamma$ denote an instance of Algorithm 12 with discount $\gamma$.[3] Let $\mathcal{A}_{Meta}$ be an instance of the algorithm characterized in Theorem 10.3.1, and suppose we set $y_t^{Ref} = \widetilde{y}_t$ for all $t$. Then for any $\boldsymbol{u} = (u_1, \dots, u_T)$ in $\mathbb{R}^d$, Algorithm 13 guarantees*

$$R_T(\boldsymbol{u}) \le O\left(d \max_t (y_t - y_t^{Ref})^2 \log(T) + b\sqrt{dP_T^{\gamma^*}(\boldsymbol{u}) \sum_{t=1}^{T} (y_t - \widetilde{y}_t)^2}\right)$$

*where $\gamma^* \in [0, 1]$ satisfies Equation (10.2).*

**Theorem 10.3.3.** *Under the same conditions as Theorem 10.3.2, suppose each $\mathcal{A}_\gamma$ sets hints $\widetilde{y}_t = \overline{y}_t^\gamma = \mathrm{Clip}_{\mathcal{B}_t}(\langle, x_t, w_t^\gamma\rangle)$, where $\mathcal{B}_t = [y_t^{Ref} - M_t, y_t^{Ref} + M_t]$ and $M_t = \max_{s<t} \left|y_s - y_s^{Ref}\right|$. Then for any $\boldsymbol{u} = (u_1, \dots, u_T)$ in $\mathbb{R}^d$, Algorithm 13 guarantees*

$$R_T(\boldsymbol{u}) \le O\left(dP_T^{\gamma_{\min}}(\boldsymbol{u}) + d \max_t \left(y_t - y_t^{Ref}\right)^2 \log(T) + b\sqrt{dP_T^{\gamma^\circ}(\boldsymbol{u}) \sum_{t=1}^{T} \ell_t(u_t)}\right)$$

*where $\gamma_{\min} = \frac{2d}{2d+1}$ and $\gamma^\circ \in [0, 1]$ satisfies Equation (10.3).*

It is worth noting that Theorems 10.3.2 and 10.3.3 use knowledge of the horizon $T$ to construct the set of experts. All of our results extend immediately to the unknown $T$ setting as well via the standard doubling trick (Nicolo Cesa-Bianchi and Lugosi 2006), so for simplicity we treat $T$ as part of the problem setting rather than a potentially unknown property of the data. An interesting direction for future development would be to construct the set of experts in a more on-the-fly way, so as to avoid using the doubling trick to adapt to unknown $T$.

[3] For brevity, here we refer to an algorithm that directly predicts $\widetilde{y}_t$ on every round as being an instance of the discounted VAW forecaster with $\gamma = 0$. This terminology can be justified by Remark C.2.2, but for our purposes here it's sufficient to consider it convenient alias.

## 10.4 Strongly-Adaptive Guarantees

While our original goal was only to achieve dynamic regret guarantees in the absence of prior knowledge, it turns out that we can actually achieve an even stronger result: dynamic regret guarantees that hold over *every sub-interal* $[a,b] \subseteq [1,T]$ *simultaneously.* To our knowledge, *strongly-adaptive* guarantees of this sort have previously only been achieved under various boundedness assumptions (Baby, Hasson, and Y. Wang 2021; Baby and Y.-X. Wang 2022b; Baby and Y.-X. Wang 2022a; Jun, Orabona, et al. 2017; Cutkosky 2020; Daniely, Gonen, and Shalev-Shwartz 2015).

The results can be derived using the results in the previous section. As shown in Appendix C.2.4, for any $[s,\tau] \subseteq [1,T]$, $\boldsymbol{u} = (u_s, \ldots, u_\tau)$, and $\gamma \in \mathcal{S}_\gamma$, Algorithm 13 more generally guarantees that

$$R_{[s,\tau]}(\boldsymbol{u}) \le \widehat{O}\left(R^{\mathcal{A}_\gamma}_{[s,\tau]}(\boldsymbol{u}) + \max_t (y_t - y_t^{\text{Ref}})^2 \log(N\tau)\right),$$

where $R_{[s,\tau]}$ denotes the regret over sub-interval $[s,\tau] \subseteq [1,T]$. The only caveat is that the regret guarantees of the discounted VAW forecaster only hold when the algorithm *begins learning* on round $s$.[4] However, suppose that for each $s \in [1,T]$ and each $\gamma \in \mathcal{S}_\gamma$ we define an algorithm $\mathcal{A}_{\gamma,s}$ which uses discount $\gamma$ but begins learning at time $s$. Then for any $[s,\tau]$ Lemma C.2.10 implies that there is a $\mathcal{A}_{\gamma,s}$ such that $R^{\mathcal{A}_{\gamma,s}}_{[s,\tau]}(\boldsymbol{u}) \le O(d\max_t(y_t - y_t^{\text{Ref}})^2 \log(\tau - s) + b\sqrt{dP^{\gamma^*}_{[s,\tau]}(\boldsymbol{u})\sum_{t=s}^{\tau}(y_t - \widetilde{y}_t)^2})$. Plugging this back into the previous display and choosing $|\mathcal{S}_\gamma| \le O(\log(T))$, we have $N \le O(T\log(T))$ and an overall regret bound of

$$R_{[s,\tau]}(\boldsymbol{u}) \le \widehat{O}\left(d\max_t (y_t - \widetilde{y}_t)^2 \log(T) + b\sqrt{dP^{\gamma^*}_{[s,\tau]}(\boldsymbol{u})\sum_{t=s}^{\tau}(y_t - \widetilde{y}_t)^2}\right).$$

This is the essence of the Follow the Leading History algorithm of Hazan and Comandur Seshadhri (2007) and Hazan and C. Seshadhri (2009).

While the above approach leads to a strongly-adaptive guarantee, it would be excessively expensive in general, since we'd now have $O(T\log(T))$ total experts to update on every round. We may instead lower this to $O(\log^2(T))$ experts using the geometric covering intervals of Daniely, Gonen, and Shalev-Shwartz (2015) and Veness et al. (2013). The idea is as follows: instead of initializing a new instance of each $\mathcal{A}_\gamma$ on every round $s \in [T]$, we will construct a set of intervals $S$ such that any $[s,\tau] \subseteq [1,T]$ can be covered using only a small number of intervals from $S$. Then for each $\gamma \in \mathcal{S}_\gamma$ and each $I \in S$, we can define an instance of the discounted VAW forecaster $\mathcal{A}_{\gamma,I}$ which is run only during the interval $I$. The geometric covering intervals are constructed in such a way that 1) any round $t$ can fall into at most $O(\log(T))$ of the intervals, and 2) any $[s,\tau] \subseteq [1,T]$ can be

[4] More generally, it can be seen from the analysis that if the algorithm starts at time $t = 1$ and we try to bound the regret over $[s,\tau]$, then after telescoping the divergence terms we will end up with a non-trivial term $D_{\psi_s}(u_s|w_s)$ which is hard to quantify in general for $s > 1$ without further assumptions.

covered using only $O(\log(\tau - s))$ disjoint intervals from $S$. The first property ensures that there at most $O(\log^2(T))$ active experts on each round, while the second property implies that there is a disjoint set of intervals $I_1, \dots, I_K$ such that $R_{[s,\tau]}(\boldsymbol{u}) = \sum_{i=1}^K R_{I_i}(\boldsymbol{u})$, so bounding each of these using a similar argument to the above followed by an application of Cauchy-Schwarz inequality yields

$$R_{[s,\tau]}(\boldsymbol{u}) \le \widehat{O}\left(d \max_t (y_t - \widetilde{y}_t)^2 \log^2(T) + b\sqrt{dP_{[s,\tau]}^{\gamma^*}(\boldsymbol{u}) \sum_{t=s}^{\tau} (y_t - \widetilde{y}_t)^2}\right),$$

where $P_{[s,\tau]}(\boldsymbol{u})$ is the total variability over the intervals and we've used $K \log(T) \le O(\log^2(T))$. Hence, overall the penalty we incur for using the geometric covering is a modest increase from $\log(T)$ to $K \log(T) \le O(\log^2(T))$ in the leading term. Likewise, a similar argument holds for our small-loss bounds. We formalize these intuitions in the following theorem. Prof can be found in Appendix C.2.5.

**Theorem 10.4.1.** *Let $\mathcal{S}_\gamma$ be the set of discount factors defined in Theorem 10.3.2, let $S$ denote a set of geometric covering intervals over $[1, T]$, and for each $\gamma \in \mathcal{S}_\gamma$ and $I \in S$, let $\mathcal{A}_{\gamma,I}$ be an instance of Algorithm 12 using discount $\gamma$ and applied during interval $I$. Let $\mathcal{A}_{Meta}$ be an instance of the meta-algorithm characterized in Theorem 10.3.1. Then for any $[s, \tau] \subseteq [1, T]$, there is a set of disjoint intervals $I_1, \dots, I_K$ in $S$ such that $\cup_{i=1}^K I_i = [s, \tau]$, and moreover, for any $\boldsymbol{u} = (u_s, \dots, u_\tau)$ Algorithm 13 with $y_t^{Ref} = \widetilde{y}_t$ guarantees*

$$R_{[s,\tau]}(\boldsymbol{u}) \le \widehat{O}\left(d \max_t (y_t - y_t^{Ref})^2 \log^2(T) + b\sqrt{dP_{[s,\tau]}^{\gamma^*}(\boldsymbol{u}) \sum_{t \in [s,\tau]} (y_t - \widetilde{y}_t)^2}\right)$$

*where $P_{[s,\tau]}^{\gamma^*}(\boldsymbol{u}) = \sum_{i=1}^K P_{I_i}^{\gamma_i^*}(\boldsymbol{u})$ and each $\gamma_i^* \in [0,1]$ satisfies $\gamma_i^* = \frac{\sqrt{\frac{d}{2}\sum_{t \in I_i}(y_t - \widetilde{y}_t)^2}}{\sqrt{\frac{d}{2}\sum_{t \in I_i}(y_t - \widetilde{y}_t)^2} + \sqrt{P_{I_i}^{\gamma_i^*}(\boldsymbol{u})}}$.*
*If we instead suppose each $\mathcal{A}_{\gamma,I}$ sets hints as in Theorem 10.3.3, then for any $\boldsymbol{u} = (u_s, \dots, u_\tau)$ Algorithm 13 guarantees*

$$R_{[s,\tau]}(\boldsymbol{u}) \le \widehat{O}\left(dP_{[s,\tau]}^{\gamma_{\min}}(\boldsymbol{u}) + d \max_t (y_t - y_t^{Ref})^2 \log^2(T) + b\sqrt{dP_{[s,\tau]}^{\gamma^\circ}(\boldsymbol{u}) \sum_{t \in [s,\tau]} \ell_t(u_t)}\right)$$

*where $P_{[s,\tau]}^{\gamma^\circ}(\boldsymbol{u}) = \sum_{i=1}^K P_{I_i}^{\gamma_i^\circ}(\boldsymbol{u})$ and each $\gamma_i^\circ \in [0,1]$ satisfies $\gamma_i^\circ = \frac{\sqrt{d\sum_{t \in I_i}\ell_t(u_t)}}{\sqrt{d\sum_{t \in I_i}\ell_t(u_t)} + \sqrt{P_{I_i}^{\gamma_i^\circ}(\boldsymbol{u})}}$.*

## 10.5 Conclusion

In this chapter, we designed algorithms for online linear regression which achieve optimal dynamic regret guarantees, even in the absence of all prior knowledge. We developed a novel analysis of a discounted variant of the Vovk-Azoury-Warmuth forecaster, showing that it can guarantee dynamic regret of the form $R_T(\boldsymbol{u}) \leq O\left(d\log(T) \vee \sqrt{dP_T^\gamma(\boldsymbol{u})T}\right)$ when equipped with an appropriate discount factor (Section 10.2). We also provided a matching lower bound, demonstrating that these penalties are unavoidable in general (Section 10.2.2). We then showed that the ideal discount factors can be learned on-the-fly, resulting in algorithms that can be applied in the complete absence of prior knowledge yet still make optimal dynamic regret guarantees (Section 10.3) and strongly-adaptive guarantees (Section 10.4). These are the first algorithms for online linear regression that make meaningful guarantees without making assumptions of any kind on the underlying data.

As in the previous chapter, an important direction for future work is to reduce the computational complexity of the algorithms. Similar to the traditional VAW forecaster, the approach developed here can be infeasible for very high-dimensional features, requiring roughly $O(d^2\log(T))$ computation every round. The $d^2$ factor likely can be reduced by extending our analysis to use modern sketching techniques (Luo, Agarwal, et al. 2016), and the $\log(T)$ factor can possibly be reduced using similar techniques to the recent work of Z. Lu and Hazan (2022).

# Bibliography

Abbeel, Pieter, Adam Coates, Morgan Quigley, and Andrew Y Ng (2007). “An application of reinforcement learning to aerobatic helicopter flight.” In: *Advances in neural information processing systems* (cit. on p. 1).

Abernethy, Jacob, Chansoo Lee, Abhinav Sinha, and Ambuj Tewari (2014). “Online Linear Optimization via Smoothing.” In: *Proceedings of The 27th Conference on Learning Theory*. Ed. by Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvari. Vol. 35. Proceedings of Machine Learning Research. Barcelona, Spain: PMLR, pp. 807–823 (cit. on p. 15).

Asi, Hilal and John C. Duchi (2019). “Stochastic (Approximate) Proximal Point Methods: Convergence, Optimality, and Adaptivity.” In: *SIAM Journal on Optimization* 29.3, pp. 2257–2290 (cit. on p. 38).

Azizian, Waïss, Damien Scieur, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel (2020). “Accelerating Smooth Games by Manipulating Spectral Shapes.” In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by Silvia Chiappa and Roberto Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, pp. 1705–1715 (cit. on p. 52).

Azoury, Katy S and Manfred K Warmuth (2001). “Relative loss bounds for on-line density estimation with the exponential family of distributions.” In: *Machine learning* 43, pp. 211–246 (cit. on pp. 69–71, 167).

Baby, Dheeraj, Hilaf Hasson, and Yuyang Wang (2021). *Dynamic Regret for Strongly Adaptive Methods and Optimality of Online KRR*. arXiv: `2111.11550 [cs.LG]` (cit. on pp. 71, 83).

Baby, Dheeraj and Yu-Xiang Wang (2021). “Optimal Dynamic Regret in Exp-Concave Online Learning.” In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by Mikhail Belkin and Samory Kpotufe. Vol. 134. Proceedings of Machine Learning Research. PMLR, pp. 359–409 (cit. on pp. 71, 79, 196).

— (2022a). “Optimal Dynamic Regret in LQR Control.” In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., pp. 24879–24892 (cit. on p. 83).

Baby, Dheeraj and Yu-Xiang Wang (2022b). “Optimal Dynamic Regret in Proper Online Learning with Strongly Convex Losses and Beyond.” In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. Ed. by Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera. Vol. 151. Proceedings of Machine Learning Research. PMLR, pp. 1805–1845 (cit. on p. 83).

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014). “Neural machine translation by jointly learning to align and translate.” In: *arXiv preprint arXiv:1409.0473* (cit. on p. 1).

Besbes, Omar, Yonatan Gur, and Assaf Zeevi (2015). “Non-Stationary Stochastic Optimization.” In: *Operations Research* 63.5, pp. 1227–1244 (cit. on p. 76).

Boyd, Stephen P and Lieven Vandenberghe (2004). *Convex optimization*. Cambridge university press (cit. on p. 48).

Campolongo, Nicolò and Francesco Orabona (2020). “Temporal Variability in Implicit Online Learning.” In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., pp. 12377–12387 (cit. on pp. 38, 39).

— (2021). *A Closer Look at Temporal Variability in Dynamic Online Learning*. arXiv: `2102.07666 [cs.LG]` (cit. on p. 76).

Cesa-Bianchi, Nicolo, Pierre Gaillard, Gábor Lugosi, and Gilles Stoltz (2012). “Mirror descent meets fixed share (and feels no regret).” In: *Advances in Neural Information Processing Systems* 25 (cit. on pp. 27, 80, 197).

Cesa-Bianchi, Nicolo, Philip M Long, and Manfred K Warmuth (1996). “Worst-case quadratic loss bounds for prediction using linear functions and gradient descent.” In: *IEEE Transactions on Neural Networks* 7.3, pp. 604–619 (cit. on p. 11).

Cesa-Bianchi, Nicolo and Gábor Lugosi (2006). *Prediction, learning, and games*. Cambridge university press (cit. on pp. 6, 72, 82).

Chen, Keyi, Ashok Cutkosky, and Francesco Orabona (2022). “Implicit Parameter-free Online Learning with Truncated Linear Models.” In: *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*. Ed. by Sanjoy Dasgupta and Nika Haghtalab. Vol. 167. Proceedings of Machine Learning Research. PMLR, pp. 148–175 (cit. on p. 39).

Chen, Liyu, Haipeng Luo, and Chen-Yu Wei (2021). “Impossible Tuning Made Possible: A New Expert Algorithm and Its Applications.” In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by Mikhail Belkin and Samory Kpotufe. Vol. 134. Proceedings of Machine Learning Research. PMLR, pp. 1216–1259 (cit. on pp. 27, 130).

Couchman, Hugh, Robert Deupree, Ken Edgecombe, Wagdi Habashi, Richard Peltier, Jonathan Schaeffer, and Danial Senechal (2015). *A proposal to the Canada Foundation for Innovation – National Platforms Fund*. `https://www.computecanada.ca/wp-content/uploads/2015/02/NPF.pdf`. Accessed: August 01, 2020 (cit. on p. 2).

Cutkosky, Ashok (2019a). “Artificial Constraints and Hints for Unbounded Online Learning.” In: *Proceedings of the Thirty-Second Conference on Learning Theory*. Ed. by Alina Beygelzimer and Daniel Hsu. Vol. 99. Proceedings of Machine Learning Research. Phoenix, USA: PMLR, pp. 874–894 (cit. on pp. 11, 12, 35–37).

— (2019b). “Combining Online Learning Guarantees.” In: *Proceedings of the Thirty-Second Conference on Learning Theory*. Ed. by Alina Beygelzimer and Daniel Hsu. Vol. 99. Proceedings of Machine Learning Research. Phoenix, USA: PMLR, pp. 895–913 (cit. on pp. 58, 142).

— (2020). “Parameter-free, Dynamic, and Strongly-Adaptive Online Learning.” In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. Virtual: PMLR, pp. 2250–2259 (cit. on pp. 21, 83).

Cutkosky, Ashok and Kwabena Boahen (2017). “Online Learning Without Prior Information.” In: *Proceedings of the 2017 Conference on Learning Theory*. Ed. by Satyen Kale and Ohad Shamir. Vol. 65. Proceedings of Machine Learning Research. PMLR, pp. 643–677 (cit. on p. 35).

Cutkosky, Ashok and Francesco Orabona (2018). “Black-Box Reductions for Parameter-free Online Learning in Banach Spaces.” In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, pp. 1493–1529 (cit. on pp. 12, 13, 15, 54, 55, 61).

Cutkosky, Ashok and Tamas Sarlos (2019). “Matrix-Free Preconditioning in Online Learning.” In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 1455–1464 (cit. on pp. 13, 15, 32, 35).

Daniely, Amit, Alon Gonen, and Shai Shalev-Shwartz (2015). “Strongly Adaptive Online Learning.” In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, pp. 1405–1411 (cit. on pp. 21, 60, 83, 194).

Du, Simon S., Gauthier Gidel, Michael I. Jordan, and Chris Junchi Li (2022). *Optimal Extragradient-Based Bilinearly-Coupled Saddle-Point Optimization* (cit. on pp. 3, 51, 52).

Duchi, John, Elad Hazan, and Yoram Singer (2011). “Adaptive subgradient methods for online learning and stochastic optimization.” In: *Journal of machine learning research* 12.7 (cit. on p. 11).

Duchi, John C, Shai Shalev-Shwartz, Yoram Singer, and Ambuj Tewari (2010). “Composite Objective Mirror Descent.” In: *COLT*, pp. 14–26 (cit. on p. 25).

Fang, Huang, Nick Harvey, Victor Portella, and Michael Friedlander (2020). “Online mirror descent and dual averaging: keeping pace in the dynamic case.” In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 3008–3017 (cit. on p. 25).

Foster, Dean, Satyen Kale, and Howard Karloff (2016). “Online Sparse Linear Regression.” In: *29th Annual Conference on Learning Theory*. PMLR (cit. on p. 69).

Gaillard, Pierre, Sébastien Gerchinovitz, Malo Huard, and Gilles Stoltz (2019). “Uniform regret bounds over $\mathbb{R}^d$ for the sequential linear regression problem with the square loss.” In: *Proceedings of the 30th International Conference on Algorithmic Learning Theory*. Ed. by Aurélien Garivier and Satyen Kale. Vol. 98. Proceedings of Machine Learning Research. PMLR, pp. 404–432 (cit. on p. 182).

Ghai, Udaya, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang (2020). “No-Regret Prediction in Marginally Stable Systems.” In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by Jacob Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, pp. 1714–1757 (cit. on p. 70).

Gyorgy, Andras and Csaba Szepesvari (2016). “Shifting Regret, Mirror Descent, and Matrices.” In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by Maria Florina Balcan and Kilian Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, pp. 2943–2951 (cit. on pp. 27, 54).

Hall, Eric C. and Rebecca M. Willett (2016). *Online Optimization in Dynamic Environments*. arXiv: 1307.5944 [stat.ML] (cit. on pp. 27, 54).

Hazan, Elad (2019). “Introduction to Online Convex Optimization.” In: *CoRR* abs/1909.05207. arXiv: 1909.05207 (cit. on pp. 9, 197).

Hazan, Elad, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang (2018). “Spectral filtering for general linear dynamical systems.” In: *Advances in Neural Information Processing Systems* 31 (cit. on p. 70).

Hazan, Elad and C. Seshadhri (2009). “Efficient Learning Algorithms for Changing Environments.” In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML ’09. Montreal, Quebec, Canada: Association for Computing Machinery, pp. 393–400 (cit. on pp. 21, 83).

Hazan, Elad and Comandur Seshadhri (2007). “Adaptive algorithms for online decision problems.” In: *Electronic colloquium on computational complexity (ECCC)*. Vol. 14. 088 (cit. on pp. 21, 83).

Hazan, Elad and Karan Singh (2022). *Introduction to Online Nonstochastic Control* (cit. on p. 70).

Hazan, Elad, Karan Singh, and Cyril Zhang (2017). “Learning linear dynamical systems via spectral filtering.” In: *Advances in Neural Information Processing Systems* 30 (cit. on p. 70).

Hoeven, Dirk van der (2019). “User-Specified Local Differential Privacy in Unconstrained Adaptive Online Learning.” In: *NeurIPS*, pp. 14080–14089 (cit. on pp. 12, 13, 15).

Ibrahim, Adam, Waïss Azizian, Gauthier Gidel, and Ioannis Mitliagkas (2020). “Linear Lower Bounds and Conditioning of Differentiable Games.” In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 4583–4593 (cit. on p. 52).

Jacobsen, Andrew and Ashok Cutkosky (2022). “Parameter-free Mirror Descent.” In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Ed. by Po-Ling Loh and Maxim Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, pp. 4160–4211 (cit. on pp. iii, 3, 23, 71, 75, 102).

— (2023). “Unconstrained Online Learning with Unbounded Losses.” In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett. Vol. 202. Proceedings of Machine Learning Research. PMLR, pp. 14590–14630 (cit. on pp. iii, 24).

— (2024). “Online Linear Regression in Dynamic Environments via Discounting.” In: *Forty-first International Conference on Machine Learning* (cit. on p. iii).

Jadbabaie, Ali, Alexander Rakhlin, Shahin Shahrampour, and Karthik Sridharan (2015). “Online Optimization : Competing with Dynamic Comparators.” In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Guy Lebanon and S. V. N. Vishwanathan. Vol. 38. Proceedings of Machine Learning Research. San Diego, California, USA: PMLR, pp. 398–406 (cit. on p. 60).

Jun, Kwang-Sung and Francesco Orabona (2019). “Parameter-Free Online Convex Optimization with Sub-Exponential Noise.” In: *Proceedings of the Thirty-Second Conference on Learning Theory*. Ed. by Alina Beygelzimer and Daniel Hsu. Vol. 99. Proceedings of Machine Learning Research. Phoenix, USA: PMLR, pp. 1802–1823 (cit. on pp. 13, 35).

Jun, Kwang-Sung, Francesco Orabona, Stephen Wright, and Rebecca Willett (2017). “Improved Strongly Adaptive Online Learning using Coin Betting.” In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. Ed. by Aarti Singh and Jerry Zhu. Vol. 54. Proceedings of Machine Learning Research. PMLR, pp. 943–951 (cit. on pp. 15, 21, 60, 83).

Kalman, Rudolph Emil (1960). “A New Approach to Linear Filtering and Prediction Problems.” In: *Transactions of the ASME–Journal of Basic Engineering* 82.Series D, pp. 35–45 (cit. on p. 70).

Kempka, Michal, Wojciech Kotlowski, and Manfred K. Warmuth (2019). “Adaptive Scale-Invariant Online Algorithms for Learning Linear Models.” In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 3321–3330 (cit. on pp. 12, 15, 70).

Kivinen, J., M.K. Warmuth, and B. Hassibi (2006). “The p-norm generalization of the LMS algorithm for adaptive filtering.” In: *IEEE Transactions on Signal Processing* 54.5, pp. 1782–1793 (cit. on p. 70).

Kivinen, Jyrki and Manfred K Warmuth (1997). “Exponentiated gradient versus gradient descent for linear predictors.” In: *information and computation* 132.1, pp. 1–63 (cit. on p. 11).

Kotłowski, Wojciech (2017). “Scale-Invariant Unconstrained Online Learning.” In: *Proceedings of the 28th International Conference on Algorithmic Learning Theory*. Ed. by Steve Hanneke and Lev

Reyzin. Vol. 76. Proceedings of Machine Learning Research. Kyoto University, Kyoto, Japan: PMLR, pp. 412–433 (cit. on p. 70).

Kozdoba, Mark, Jakub Marecek, Tigran Tchrakian, and Shie Mannor (2019). "On-line learning of linear dynamical systems: Exponential forgetting in kalman filters." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01, pp. 4098–4105 (cit. on p. 70).

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning." In: *nature* 521.7553, pp. 436–444 (cit. on p. 1).

Lillicrap, Timothy P, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra (2015). "Continuous control with deep reinforcement learning." In: *arXiv preprint arXiv:1509.02971* (cit. on p. 1).

Liu, Mingrui and Francesco Orabona (2022). "On the Initialization for Convex-Concave Min-max Problems." In: *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*. Ed. by Sanjoy Dasgupta and Nika Haghtalab. Vol. 167. Proceedings of Machine Learning Research. PMLR, pp. 743–767 (cit. on pp. 49, 52).

Lu, Zhou and Elad Hazan (2022). *Efficient Adaptive Regret Minimization* (cit. on p. 85).

Luo, Haipeng, Alekh Agarwal, Nicolò Cesa-Bianchi, and John Langford (2016). "Efficient Second Order Online Learning by Sketching." In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Vol. 29. Curran Associates, Inc. (cit. on pp. 70, 85).

Luo, Haipeng, Mengxiao Zhang, Peng Zhao, and Zhi-Hua Zhou (2022). "Corralling a Larger Band of Bandits: A Case Study on Switching Regret for Linear Bandits." In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Ed. by Po-Ling Loh and Maxim Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, pp. 3635–3684 (cit. on pp. 64, 71).

Mayo, Jack J., Hedi Hadiji, and Tim van Erven (2022). "Scale-free Unconstrained Online Learning for Curved Losses." In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Ed. by Po-Ling Loh and Maxim Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, pp. 4464–4497 (cit. on pp. 11, 12, 70, 81, 182).

Mcmahan, Brendan and Matthew Streeter (2012). "No-Regret Algorithms for Unconstrained Online Convex Optimization." In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Vol. 25. Curran Associates, Inc. (cit. on pp. 11, 12, 46, 54, 130).

McMahan, H. Brendan (2017). "A Survey of Algorithms and Analysis for Adaptive Online Learning." In: *J. Mach. Learn. Res.* 18.1, pp. 3117–3166 (cit. on pp. 6, 25, 96, 97).

McMahan, H. Brendan and Francesco Orabona (2014). "Unconstrained Online Linear Learning in Hilbert Spaces: Minimax Algorithms and Normal Approximations." In: *Proceedings of The 27th Conference on Learning Theory*. Vol. 35. Proceedings of Machine Learning Research. Barcelona, Spain: PMLR, pp. 1020–1039 (cit. on pp. 12, 13, 35, 54).

McMahan, H. Brendan and Matthew J. Streeter (2010). “Adaptive Bound Optimization for Online Convex Optimization.” In: *CoRR* abs/1002.4908. arXiv: 1002.4908 (cit. on p. 11).

Mhammedi, Zakaria and Wouter M. Koolen (2020). “Lipschitz and Comparator-Norm Adaptivity in Online Learning.” In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by Jacob Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, pp. 2858–2887 (cit. on pp. 12, 13, 15, 32, 35, 36, 47, 71, 118).

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. (2015). “Human-level control through deep reinforcement learning.” In: *nature* 518.7540, pp. 529–533 (cit. on p. 1).

Ng, Andrew Y, Adam Coates, Mark Diel, Varun Ganapathi, Jamie Schulte, Ben Tse, Eric Berger, and Eric Liang (2006). “Autonomous inverted helicopter flight via reinforcement learning.” In: *Experimental robotics IX* (cit. on p. 1).

Orabona, Francesco (2013). “Dimension-Free Exponentiated Gradient.” In: *Advances in Neural Information Processing Systems*. Ed. by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger. Vol. 26. Curran Associates, Inc. (cit. on pp. 11, 12, 46).

— (2019). “A Modern Introduction to Online Learning.” In: *CoRR* abs/1912.13213. arXiv: 1912.13213 (cit. on pp. 6, 9, 15, 16, 97, 106).

Orabona, Francesco, Nicolo Cesa-Bianchi, and Claudio Gentile (2012). “Beyond Logarithmic Bounds in Online Learning.” In: *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Neil D. Lawrence and Mark Girolami. Vol. 22. Proceedings of Machine Learning Research. La Palma, Canary Islands: PMLR, pp. 823–831 (cit. on pp. 11, 67).

Orabona, Francesco, Koby Crammer, and Nicolò Cesa-Bianchi (2015). “A Generalized Online Mirror Descent with Applications to Classification and Regression.” In: *Mach. Learn.* 99.3, pp. 411–435 (cit. on pp. 69–71).

Orabona, Francesco and Kwang-Sung Jun (2023). “Tight Concentrations and Confidence Sequences from the Regret of Universal Portfolio.” In: *IEEE Transactions on Information Theory*, pp. 1–1 (cit. on p. 41).

Orabona, Francesco and Dávid Pál (2016). “Coin Betting and Parameter-Free Online Learning.” In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS’16. Barcelona, Spain: Curran Associates Inc., pp. 577–585 (cit. on pp. 12, 13, 15, 16, 54).

— (2018). “Scale-free online learning.” In: *Theoretical Computer Science* 716. Special Issue on ALT 2015, pp. 50–69 (cit. on pp. 9, 11, 12, 16, 25).

— (2021). “Parameter-free Stochastic Optimization of Variationally Coherent Functions.” In: arXiv: 2102.00236 [math.OC] (cit. on pp. 15, 104).

Rakhlin, Alexander and Karthik Sridharan (2017). “On Equivalence of Martingale Tail Bounds and Deterministic Regret Inequalities.” In: *Proceedings of the 2017 Conference on Learning Theory*.

Ed. by Satyen Kale and Ohad Shamir. Vol. 65. Proceedings of Machine Learning Research. PMLR, pp. 1704–1722 (cit. on p. 41).

Rashidinejad, Paria, Jiantao Jiao, and Stuart Russell (2020). "SLIP: Learning to predict in unknown dynamical systems with long-term memory." In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., pp. 5716–5728 (cit. on p. 70).

Shalev-Shwartz, Shai and Yoram Singer (2007). "Online learning: Theory, algorithms, and applications." In: (cit. on p. 6).

Silver, David, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. (2016). "Mastering the game of Go with deep neural networks and tree search." In: *nature* 529.7587, pp. 484–489 (cit. on p. 1).

Simon, Dan (2006). *Optimal state estimation: Kalman, H infinity, and nonlinear approaches*. John Wiley & Sons (cit. on p. 70).

Srebro, Nathan, Karthik Sridharan, and Ambuj Tewari (2010). "Smoothness, low noise and fast rates." In: *Advances in neural information processing systems* 23 (cit. on pp. 11, 67).

Steinhardt, J. and P. Liang (2014). "Adaptivity and Optimism: An Improved Exponentiated Gradient Algorithm." In: *International Conference on Machine Learning (ICML)* (cit. on p. 27).

Telgarsky, Matus (2022). *Stochastic linear optimization never overfits with quadratically-bounded losses on general data* (cit. on p. 44).

Tsiamis, Anastasios and George J Pappas (2022). "Online learning of the kalman filter with logarithmic regret." In: *IEEE Transactions on Automatic Control* (cit. on p. 70).

Veness, Joel, Martha White, Michael Bowling, and András György (2013). "Partition Tree Weighting." In: *2013 Data Compression Conference*, pp. 321–330 (cit. on p. 83).

Vovk, Volodya (2001). "Competitive On-line Statistics." In: *International Statistical Review* 69.2, pp. 213–248. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1751-5823.2001.tb00457.x` (cit. on pp. 69–71, 182).

Yang, Tianbao, Lijun Zhang, Rong Jin, and Jinfeng Yi (2016). "Tracking Slowly Moving Clairvoyant: Optimal Dynamic Regret of Online Learning with True and Noisy Gradient." In: *Proceedings of The 33rd International Conference on Machine Learning*. Vol. 48. Proceedings of Machine Learning Research. PMLR, pp. 449–457 (cit. on pp. 19, 54).

Yuan, Jianjun and Andrew G. Lamperski (2019). "Trading-Off Static and Dynamic Regret in Online Least-Squares and Beyond." In: *CoRR* abs/1909.03118. arXiv: `1909.03118` (cit. on p. 71).

Zhang, Jiujia and Ashok Cutkosky (2022). *Parameter-free Regret in High Probability with Heavy Tails* (cit. on p. 22).

Zhang, Lijun, Shiyin Lu, and Zhi-Hua Zhou (2018). "Adaptive online learning in dynamic environments." In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 1330–1340 (cit. on pp. 20, 24, 54, 57, 58, 60, 62, 64, 71, 75, 165).

Zhang, Zhiyu, Ashok Cutkosky, and Ioannis Paschalidis (2022a). "Adversarial Tracking Control via Strongly Adaptive Online Learning with Memory." In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. Ed. by Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera. Vol. 151. Proceedings of Machine Learning Research. PMLR, pp. 8458–8492 (cit. on p. 40).

— (2022b). "PDE-Based Optimal Strategy for Unconstrained Online Learning." In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato. Vol. 162. Proceedings of Machine Learning Research. PMLR, pp. 26085–26115 (cit. on p. 40).

Zhang, Zhiyu, Ashok Cutkosky, and Yannis Paschalidis (2023). "Unconstrained Dynamic Regret via Sparse Coding." In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc., pp. 74636–74670 (cit. on pp. 40, 71, 75).

Zhang, Zhiyu, Heng Yang, Ashok Cutkosky, and Ioannis Ch. Paschalidis (2023). *Improving Adaptive Online Learning Using Refined Discretization*. arXiv: `2309.16044 [cs.LG]` (cit. on pp. 40, 41).

Zhao, Peng, Yu-Jie Zhang, Lijun Zhang, and Zhi-Hua Zhou (2020). "Dynamic Regret of Convex and Smooth Functions." In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., pp. 12510–12520 (cit. on pp. 39, 67, 71).

— (2024). "Adaptivity and Non-stationarity: Problem-dependent Dynamic Regret for Online Convex Optimization." In: *Journal of Machine Learning Research* 25.98, pp. 1–52 (cit. on pp. 71, 75).

Zhou, Xingyu (2018). "On the fenchel duality between strong convexity and lipschitz continuous gradient." In: *arXiv preprint arXiv:1803.06573* (cit. on p. 8).

Zinkevich, Martin (2003). "Online convex programming and generalized infinitesimal gradient ascent." In: *Proceedings of the 20th international conference on machine learning (icml-03)*, pp. 928–936 (cit. on pp. 6, 20, 54).

# Appendices

# Appendix A

# Part I (Foundations)

## A.1 A Strong Mirror Descent Lemma

In this section we derive a regret template for Centered Mirror Descent which holds for arbitrary sequences of loss functions and choices of $\psi_t$ and $\varphi_t$. The result is analogous to the Strong FTRL Lemma of McMahan (2017), but applies to a sequence of comparators and is tailored to mirror descent-style analysis. Here we present a mild generalization of the lemma which incorporates the post-hoc adjustments $\widetilde{w}_t = \mathcal{M}_t(w_t)$.

In this section, the following short-hand notation will be convenient:

$$\widehat{D}_f(x, y; g_y) \overset{\text{def}}{=} f(x) - f(y) - \langle g_y, x - y \rangle .$$

where $f$ is a subdifferentiable function and $g_y$ is an arbitrary element of $\partial f(y)$. Note that when $f$ is differentiable, then $\partial f(y) = \{\nabla f(y)\}$, so the short-hand reduces to the standard bregman divergence. Moreover, observe that $\widehat{D}$ still satisfies the usual subgradient inequalities. For instance, if $f$ is convex, then for any $g_y \in \partial f(y)$ we have $\widehat{D}_f(x, y; g_y) \geq 0$.

The following lemma provides a powerful regret *equality* for the centered mirror descent framework, showing how each of the components of the update feature in the regret bound. We will use this lemma to derive many of our results as special cases.

**Lemma A.1.1.** ***(Strong Centered Mirror Descent Lemma)*** *For all $t$, let $\ell_t(\cdot)$ be a subdifferentiable function, $\varphi_t(\cdot)$ be a subdifferentiable non-negative function, and $\psi_t(\cdot)$ be a differentiable non-negative function. Define $\Delta_t(w) = D_{\psi_{t+1}}(w|w_1) - D_{\psi_t}(w|w_1)$, and*

$$w_{t+1} = \arg\min_{w\in\mathbb{R}^d} \ell_t(w) + D_{\psi_t}(w|w_t) + (\Delta_t + \varphi_t)(w).$$

*Then, for all $t$ there is some $\nabla\ell_t(w_{t+1}) \in \partial\ell_t(w_{t+1})$ and $\nabla\varphi_t(w_{t+1}) \in \partial\varphi_t(w_{t+1})$ such that $\nabla\ell_t(w_{t+1}) = \nabla\psi_t(w_t) - \nabla\psi_{t+1}(w_{t+1}) - \nabla\Delta_t(w_{t+1}) - \nabla\varphi_t(w_{t+1})$, and for any $u_1, \ldots, u_T$ in $\mathbb{R}^d$,*

$$\begin{aligned}
\sum_{t=1}^T \ell_t(\widetilde{w}_t) - \ell_t(u_t) = D_{\psi_{T+1}}(u_T|\widetilde{w}_1) - D_{\psi_{T+1}}(u_T|\widetilde{w}_{T+1}) + \sum_{t=1}^T \varphi_t(u_t)& \\
+ \sum_{t=1}^T \underbrace{D_{\psi_{t+1}}(u_t|\widetilde{w}_{t+1}) - D_{\psi_{t+1}}(u_t|w_{t+1})}_{=:\xi_t} + \sum_{t=2}^T \underbrace{\langle \nabla\psi_t(\widetilde{w}_t) - \nabla\psi_t(\widetilde{w}_1), u_{t-1} - u_t\rangle}_{=:\mathcal{P}_t}& \\
+ \sum_{t=1}^T \underbrace{\ell_t(\widetilde{w}_t) - \ell_t(w_{t+1}) - D_{\psi_t}(w_{t+1}|\widetilde{w}_t) - (\Delta_t + \varphi_t)(w_{t+1})}_{=:\delta_t}& \\
\sum_{t=1}^T \underbrace{-\widehat{D}_{\ell_t+\varphi_t}\Big(u_t, w_{t+1}; \nabla\ell_t(w_{t+1}) + \nabla\varphi_t(w_{t+1})\Big)}_{=:\mathcal{L}_t}&
\end{aligned}$$

Observe that the lemma holds even for non-convex losses; in this case we'll need to account for the fact that the terms $-\widehat{D}_{\ell_t}(u_t, w_{t+1}; \nabla\ell_t(w_{t+1}))$ may be positive and may require additional effort to control. When the losses are convex the terms $-\widehat{D}_{\ell_t}(u_t, w_{t+1}; \nabla\ell_t(w_{t+1}))$ can often be leveraged in useful ways, particularly when the $\ell_t$ have nice properties such as strong convexity or exp-concavity. We will typically only assume convexity of $\ell_t$ and drop these terms. Similarly, for simplicity we assume that $\varphi_t$ is convex so that we can bound $-\widehat{D}_{\varphi_t}(u_t, w_{t+1}; \nabla\varphi_t(w_{t+1})) \le 0$. It's possible that this term could also be leveraged in some useful way, but we do not investigate this in the current work.

*Remark* A.1.2. Note that Lemma A.1.1 captures constrained updates as a special case. In particular, it can be shown that the constrained update is equivalent to an unconstrained one which includes an indicator function in the composite penalty (see, *e.g.*, Orabona 2019, Theorem 6.3, McMahan 2017, Section 2.4), that is,

$$\begin{aligned}
&\arg\min_{w\in W} \ell_t(w) + D_{\psi_t}(w|w_t) + (\Delta_t + \varphi_t)(w) = \\
&\qquad \arg\min_{w\in\mathbb{R}^d} \ell_t(w) + D_{\psi_t}(w|w_t) + (\Delta_t + \varphi_t)(w) + \mathbb{I}(w \in W).
\end{aligned}$$

Moreover, it is easy to see that the same regret bound holds; from Lemma A.1.1 we can see that the effect of this addition in the regret guarantee is simply that $\sum_{t=1}^{T}\varphi_t(u_t)$ will contain factors of $\mathbb{I}(u_t \in W) = 0$ for sequences $\boldsymbol{u} = (u_1, \ldots, u_T)$ in $W$.

*Proof. (of Lemma A.1.1)*

First, observe that the existence of the specified $\nabla\ell_t(w_{t+1}) \in \partial\ell_t(w_{t+1})$ and $\nabla\varphi_t(w_{t+1}) \in \partial\varphi_t(w_{t+1})$ follows directly from the first order optimality conditions applied to the update $w_{t+1} = \arg\min_w \ell_t(w) + D_{\psi_t}(w|\widetilde{w}_t) + \Delta_t(w) + \varphi_t(w)$.

Thus, using the notation $\widehat{D}_f(x, y; g_y) = f(x) - f(y) - \langle g_y, x - y\rangle$ for $g_y \in \partial f(y)$, we can write

$$\begin{aligned}\sum_{t=1}^{T}\ell_t(\widetilde{w}_t) - \ell_t(u_t) &= \sum_{t=1}^{T}\ell_t(w_{t+1}) - \ell_t(u_t) + \sum_{t=1}^{T}\ell_t(\widetilde{w}_t) - \ell_t(w_{t+1})\\ &= \sum_{t=1}^{T}\langle\nabla\ell_t(w_{t+1}), w_{t+1} - u_t\rangle - \widehat{D}_{\ell_t}(u_t, w_{t+1}; \nabla\ell_t(w_{t+1})) + \sum_{t=1}^{T}\ell_t(\widetilde{w}_t) - \ell_t(w_{t+1})\end{aligned} \tag{A.1}$$

Further, again by first order optimality conditions, we have:

$$\nabla\ell_t(w_{t+1}) + \nabla\varphi_t(w_{t+1}) + \nabla\Delta_t(w_{t+1}) + \nabla\psi_t(w_{t+1}) - \nabla\psi_t(\widetilde{w}_t) = \mathbf{0},$$

so the first summation can be witten as

$$\begin{aligned}\sum_{t=1}^{T}\langle\nabla\ell_t(w_{t+1}), w_{t+1} - u_t\rangle &= \sum_{t=1}^{T}\langle\nabla\psi_t(\widetilde{w}_t) - \nabla\psi_t(w_{t+1}), w_{t+1} - u_t\rangle - \langle\nabla\Delta_t(w_{t+1}) + \nabla\varphi_t(w_{t+1}), w_{t+1} - u_t\rangle\\ &\overset{(a)}{=} \sum_{t=1}^{T} D_{\psi_t}(u_t|\widetilde{w}_t) - D_{\psi_t}(u_t|w_{t+1}) - D_{\psi_t}(w_{t+1}|\widetilde{w}_t)\\ &\quad + \sum_{t=1}^{T}\langle\nabla\Delta_t(w_{t+1}) + \nabla\varphi_t(w_{t+1}), u_t - w_{t+1}\rangle\\ &\overset{(b)}{=} \sum_{t=1}^{T} D_{\psi_t}(u_t|\widetilde{w}_t) - D_{\psi_t}(u_t|w_{t+1}) - D_{\psi_t}(w_{t+1}|\widetilde{w}_t)\\ &\quad + \sum_{t=1}^{T}\Delta_t(u_t) - \Delta_t(w_{t+1}) - D_{\Delta_t}(u_t|w_{t+1})\\ &\quad + \sum_{t=1}^{T}\varphi_t(u_t) - \varphi_t(w_{t+1}) - \widehat{D}_{\varphi_t}(u_t, w_{t+1}; \nabla\varphi_t(w_{t+1})),\end{aligned}$$

where $(a)$ uses the well-known three-point relation of Bregman divergences $\langle\nabla f(a) - \nabla f(b), b - c\rangle = D_f(c|a) - D_f(c|b) - D_f(b|a)$, and $(b)$ observes that $\langle\nabla f(y), y - x\rangle = f(y) - f(x) - \widehat{D}_f(x, y, \nabla f(y))$ for $\nabla f(y) \in \partial f(y)$, and that $\widehat{D}_{\Delta_t}(x, y; \nabla\Delta_t(y)) = D_{\Delta_t}(x|y)$ since $\Delta_t$ is a differentiable function.

Plugging this back into Equation (A.1), and re-arranging terms, we have

$$\begin{aligned}\sum_{t=1}^{T} \ell_t(\widetilde{w}_t) - \ell_t(u_t) &= \sum_{t=1}^{T} D_{\psi_t}(u_t|\widetilde{w}_t) - D_{\psi_t}(u_t|w_{t+1}) + \sum_{t=1}^{T} \Delta_t(u_t) - D_{\Delta_t}(u_t|w_{t+1}) \\ &\quad + \sum_{t=1}^{T} \varphi_t(u_t) + \sum_{t=1}^{T} \underbrace{\ell_t(\widetilde{w}_t) - \ell_t(w_{t+1}) - D_{\psi_t}(w_{t+1}|\widetilde{w}_t) - (\Delta_t + \varphi_t)(w_{t+1})}_{=:\delta_t} \\ &\quad + \sum_{t=1}^{T} \underbrace{-\widehat{D}_{\ell_t}(u_t, w_{t+1}; \nabla \ell_t(w_{t+1})) - \widehat{D}_{\varphi_t}(u_t, w_{t+1}; \nabla \varphi_t(w_{t+1}))}_{=:\mathcal{L}_t}, \end{aligned} \tag{A.2}$$

So it remains to study the terms in the first line. We have

$$\begin{aligned}&\sum_{t=1}^{T} D_{\psi_t}(u_t|\widetilde{w}_t) - D_{\psi_t}(u_t|w_{t+1}) + D_{\Delta_t}(u_t|w_{t+1}) + \Delta_t(u_t) \\ &= \sum_{t=1}^{T} D_{\psi_t}(u_t|\widetilde{w}_t) - D_{\psi_t}(u_t|w_{t+1}) + D_{\psi_{t+1}-\psi_t}(u_t|w_{t+1}) + D_{\psi_{t+1}}(u_t|\widetilde{w}_1) - D_{\psi_t}(u_t|\widetilde{w}_1) \\ &= \sum_{t=1}^{T} D_{\psi_t}(u_t|\widetilde{w}_t) - D_{\psi_{t+1}}(u_t|w_{t+1}) + D_{\psi_{t+1}}(u_t|\widetilde{w}_1) - D_{\psi_t}(u_t|\widetilde{w}_1)\end{aligned}$$

Add and subtract $D_{\psi_{t+1}}(u_t|\widetilde{w}_{t+1})$:

$$
\begin{aligned}
&= \sum_{t=1}^{T} \left[D_{\psi_t}(u_t|\widetilde{w}_t) - D_{\psi_{t+1}}(u_t|\widetilde{w}_{t+1})\right] + \underbrace{\left[D_{\psi_{t+1}}(u_t|\widetilde{w}_{t+1}) - D_{\psi_{t+1}}(u_t|w_{t+1})\right]}_{=:\xi_t} \\
&\qquad + \sum_{t=1}^{T} \left[D_{\psi_{t+1}}(u_t|\widetilde{w}_1) - D_{\psi_t}(u_t|\widetilde{w}_1)\right] \\
&= \xi_{1:T} + D_{\psi_1}(u_1|\widetilde{w}_1) - D_{\psi_{T+1}}(u_T|\widetilde{w}_{T+1}) + \sum_{t=2}^{T} D_{\psi_t}(u_t|\widetilde{w}_t) - D_{\psi_t}(u_{t-1}|\widetilde{w}_t) \\
&\qquad + D_{\psi_{T+1}}(u_T|\widetilde{w}_1) - D_{\psi_1}(u_1|\widetilde{w}_1) + \sum_{t=2}^{T} D_{\psi_t}(u_{t-1}|\widetilde{w}_1) - D_{\psi_t}(u_t|\widetilde{w}_1) \\
&= \xi_{1:T} + D_{\psi_{T+1}}(u_T|\widetilde{w}_1) - D_{\psi_{T+1}}(u_T|\widetilde{w}_{T+1}) \\
&\qquad + \sum_{t=2}^{T} \psi_t(u_t) - \psi_t(u_{t-1}) - \langle \nabla\psi_t(\widetilde{w}_t), u_t - u_{t-1}\rangle \\
&\qquad + \sum_{t=2}^{T} \psi_t(u_{t-1}) - \psi_t(u_t) - \langle \nabla\psi_t(\widetilde{w}_1), u_{t-1} - u_t\rangle \\
&= \xi_{1:T} + D_{\psi_{T+1}}(u_T|\widetilde{w}_1) - D_{\psi_{T+1}}(u_T|\widetilde{w}_{T+1}) \\
&\qquad \sum_{t=2}^{T} \underbrace{\langle \nabla\psi_t(\widetilde{w}_t) - \nabla\psi_t(\widetilde{w}_1), u_{t-1} - u_t\rangle}_{\mathcal{P}_t}.
\end{aligned}
$$

Plugging this into Equation (A.2) yields the stated result:

$$
\sum_{t=1}^{T} \ell_t(\widetilde{w}_t) - \ell_t(u_t) = D_{\psi_{T+1}}(u_T|\widetilde{w}_1) - D_{\psi_{T+1}}(u_T|\widetilde{w}_{T+1}) + \sum_{t=1}^{T} \varphi_t(u_t) + \xi_{1:T} + \mathcal{P}_{2:T} + \delta_{1:T} + \mathcal{L}_{1:T}
$$

□

## A.2 Proofs for Chapter 4 (Centered Mirror Descent)

### A.2.1 Proof of Lemma 4.0.1

**Lemma 4.0.1.** ***(Centered Mirror Descent Lemma)*** *Let $\psi_t(\cdot)$ be an arbitrary sequence of differentiable non-negative convex functions, and assume that $w_1 \in \arg\min_{w\in\mathbb{R}^d} \psi_t(w)$ for all $t$. Let $\varphi_t(\cdot)$ be an arbitrary sequence of sub-differentiable non-negative convex functions. Then for any $u_1, \ldots, u_T$, Algorithm 2 guarantees*

$$
\begin{aligned}
R_T(\boldsymbol{u}) \le \psi_{T+1}(u_T) + \sum_{t=1}^{T} \varphi_t(u_t) + \sum_{t=2}^{T} \underbrace{\langle \nabla\psi_t(w_t) - \nabla\psi_t(w_1), u_{t-1} - u_t \rangle}_{=:\mathcal{P}_t} \\
+ \sum_{t=1}^{T} \underbrace{\langle g_t, w_t - w_{t+1} \rangle - D_{\psi_t}(w_{t+1}|w_t) - (\Delta_t + \varphi_t)(w_{t+1})}_{=:\delta_t}, \qquad (4.1)
\end{aligned}
$$

*where $g_t \in \partial\ell_t(w_t)$.*

*Proof.* From Lemma A.1.1 with $\mathcal{M}_t(w) = w$ for all $t$, we have have $w_t = \widetilde{w}_t$ for all $t$ and

$$
\sum_{t=1}^{T} \ell_t(w_t) - \ell_t(u_t) = D_{\psi_{T+1}}(u_T|w_1) - D_{\psi_{T+1}}(u_T|w_{T+1}) + \sum_{t=1}^{T} \varphi_t(u_t) + \sum_{t=2}^{T} \mathcal{P}_t + \sum_{t=1}^{T} \delta_t + \sum_{t=1}^{T} \mathcal{L}_t,
$$

where

$$
\begin{aligned}
\mathcal{P}_t &= \langle \nabla\psi_t(w_t) - \nabla\psi_t(w_1), u_{t-1} - u_t \rangle \\
\delta_t &= \langle g_t, w_t - w_{t+1} \rangle - D_{\psi_t}(w_{t+1}|w_t) - (\Delta_t + \varphi_t)(w_{t+1}) \\
\mathcal{L}_t &= -\widehat{D}_{\ell_t}(u_t, w_{t+1}; \nabla\ell_t(w_{t+1})) - \widehat{D}_{\varphi_t}(u_t, w_{t+1}; \nabla\varphi_t(w_{t+1})),
\end{aligned}
$$

where $g_t \in \ell_t(w_t)$ and $\widehat{D}_f(x, y, g_y) = f(x) - f(y) - \langle g_y, x - y \rangle$ for subdifferentiable function $f$ and $g_y \in \partial f(y)$. Since $\ell_t(\cdot)$ and $\varphi_t(\cdot)$ are convex, for any $x, y \in \mathbb{R}^d$ we have $\widehat{D}_{\ell_t}(x, y; \nabla\ell_t(y)) \ge 0$ for any $\nabla\ell_t(y) \in \partial\ell_t(y)$ and $\widehat{D}_{\varphi_t}(x, y; \nabla\varphi_t(y)) \ge 0$ for any $\nabla\varphi_t(y) \in \partial\varphi_t(y)$, so $\sum_{t=1}^{T} \mathcal{L}_t \le 0$. Further, using the assumption that $w_1 \in \arg\min_{w\in\mathbb{R}^d} \psi_t(w)$ and $\psi_t(w) \ge 0$ for all $t$, we have that $\nabla\psi_t(w_1) = \mathbf{0}$ and $D_{\psi_t}(w|w_1) \le \psi_t(w)$ for any $w \in \mathbb{R}^d$. Using this along with the fact that Bregman divergences *w.r.t* convex functions are non-negative yields

$$
\sum_{t=1}^{T} \ell_t(w_t) - \ell_t(u_t) \le \psi_{T+1}(u_T) + \sum_{t=2}^{T} \langle \nabla\psi_t(w_t) - \nabla\psi_t(w_1), u_{t-1} - u_t \rangle + \sum_{t=1}^{T} \varphi_t(u_t) + \sum_{t=1}^{T} \delta_t
$$

□

### A.2.2 Proof of Lemma 4.0.2

**Lemma 4.0.2.** ***(Stability Lemma)*** *Let $\psi_t(w) = \Psi_t(\|w\|)$ where $\Psi_t : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a convex function satisfying $\Psi_t'(x) \geq 0$, $\Psi_t''(x) \geq 0$, and $\Psi_t'''(x) \leq 0$ for all $x \geq 0$. Let $c > 0$, $G_{\max} \geq 0$, $G_{\max} \geq G_t$, and assume that there exists an $\mathring{x}_t \geq 0$ and $1/G_{\max}$-Lipschitz convex function $\eta_t : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ satisfying $\eta_t(0) = 0$ such that $|\Psi_t'''(x)| \leq \frac{2\eta_t'(x)}{(c+1)^2}\Psi_t''(x)^2$ for all $x \geq \mathring{x}_t$. Then for any $w_{t+1}, w_t \in W$,*

$$\widehat{\delta}_t \overset{def}{=} cG_t \|w_t - w_{t+1}\| - D_{\psi_t}(w_{t+1}|w_t) - \eta_t(\|w_{t+1}\|)G_t^2 \leq \frac{(c+1)^2 G_t^2}{2\Psi_t''(\mathring{x}_t)}$$

*Proof.* The proof follows using similar arguments to Jacobsen and Cutkosky (2022) with a few minor adjustments to correct for the leading term $c$.

First, consider the case that the origin is contained in the line segment connecting $w_t$ and $w_{t+1}$. Then, there exists sequences $\widehat{w}_t^1, \widehat{w}_t^2 \ldots$ and $\widehat{w}_{t+1}^1, \widehat{w}_{t+1}^2 \ldots$ such that $\lim_{n\to\infty} \widehat{w}_t^n = w_t$, $\lim_{n\to\infty} \widehat{w}_{t+1}^n = w_{t+1}$ and 0 is not contained in the line segment connecting $\widehat{w}_t^n$ and $\widehat{w}_{t+1}^n$ for all $n$. Since $\psi$ is twice differentiable everywhere except the origin, if we define $\widehat{\delta}_t^n = G_t \|\widehat{w}_t^n - \widehat{w}_{t+1}^n\| - D_{\psi_t}(\widehat{w}_{t+1}^n|\widehat{w}_t^n) - \eta_t(\|\widehat{w}_{t+1}^n\|)G_t^2$, then $\lim_{n\to\infty} \widehat{\delta}_t^n = \widehat{\delta}_t$. Thus, it suffices to prove the result for the case that the origin is *not* contained in the line segment connecting $w_t$ and $w_{t+1}$. The rest of the proof considers exclusively this case.

For brevity denote $\widehat{\delta}_t \overset{\text{def}}{=} G_t \|w_t - w_{t+1}\| - D_{\psi_t}(w_{t+1}|w_t) - \eta_t(\|w_{t+1}\|) \|g_t\|^2$. Since the origin is not in the line segment connecting $w_t$ and $w_{t+1}$, $\psi_t$ is twice differentiable on this line segment. Thus, By Taylor's theorem, there is a $\widetilde{w}$ on the line connecting $w_t$ and $w_{t+1}$ such that

$$D_{\psi_t}(w_{t+1}|w_t) = \frac{1}{2}\|w_t - w_{t+1}\|^2_{\nabla^2\psi_t(\widetilde{w})} \geq \frac{1}{2}\|w_t - w_{t+1}\|^2 \Psi_t''(\|\widetilde{w}\|)$$

where the last line observes $\psi_t(w) = \Psi_t(\|w\|)$ and uses the regularity assumptions $\Psi_t'''(x) \leq 0$, and $\Psi_t'(x) \geq 0$ for $x \geq 0$ to apply Lemma A.3.2. Hence,

$$\begin{aligned}
\widehat{\delta}_t &= cG_t \|w_t - w_{t+1}\| - D_{\psi_t}(w_{t+1}|w_t) - \eta_t(\|w_{t+1}\|)G_t^2 \\
&\leq cG_t \|w_t - w_{t+1}\| - \frac{1}{2}\|w_t - w_{t+1}\|^2 \Psi_t''(\|\widetilde{w}\|) - \eta_t(\|w_{t+1}\|)G_t^2 \\
&\overset{(a)}{\leq} cG_t \|w_t - w_{t+1}\| - \frac{1}{2}\|w_t - w_{t+1}\|^2 \Psi_t''(\|\widetilde{w}\|) - \eta_t(\|\widetilde{w}\|)G_t^2 + \eta_t'(\|\widetilde{w}\|)G_t^2\|w_{t+1} - \widetilde{w}\| \\
&\overset{(b)}{\leq} (c+1)G_t \|w_t - w_{t+1}\| - \frac{1}{2}\|w_t - w_{t+1}\|^2 \Psi_t''(\|\widetilde{w}\|) - \eta_t(\|\widetilde{w}\|)G_t^2 \\
&\overset{(c)}{\leq} \frac{(c+1)^2 G_t^2}{2\Psi_t''(\|\widetilde{w}\|)} - \eta_t(\|\widetilde{w}\|)G_t^2
\end{aligned}$$

where $(a)$ uses convexity of $\eta_t(x)$, $(b)$ uses the Lipschitz assumption $\eta_t'(\|\widetilde{w}\|) \le 1/G_t$ and the fact that $\|\widetilde{w} - w_t\| \le \|w_{t+1} - w_t\|$ for any $\widetilde{w}$ on the line connecting $w_t$ and $w_{t+1}$, and $(c)$ uses Fenchel-Young inequality. If $\|\widetilde{w}\| \le \mathring{x}_t$, then we have

$$\frac{(c+1)^2 G_t^2}{2\Psi_t''(\|\widetilde{w}\|)} - \eta_t(\|\widetilde{w}\|)G_t^2 \le \frac{(c+1)^2 G_t^2}{2\Psi_t''(\mathring{x}_t)},$$

which follows from the fact that $\Psi_t'''(x) \le 0$ implies $\Psi_t''(x)$ is non-increasing in $x$, and hence $\Psi_t''(\|\widetilde{w}\|) \ge \Psi_t''(\mathring{x}_t)$. Otherwise, if $\|\widetilde{w}\| \ge \mathring{x}_t$, we have by assumption that $\frac{|\Psi_t'''(x)|}{\Psi_t''(x)^2} = \frac{-\Psi_t'''(x)}{\Psi_t''(x)^2} = \frac{d}{dx}\frac{1}{\Psi_t''(x)} \le \frac{2\eta_t'(x)}{(c+1)^2}$ for any $x \ge \mathring{x}_t$, so integrating from $\mathring{x}_t$ to $\|\widetilde{w}\|$ we have

$$\frac{1}{\Psi_t''(\|\widetilde{w}\|)} - \frac{1}{\Psi_t''(\mathring{x}_t)} \le \frac{2}{(c+1)^2}\int_{\mathring{x}_t}^{\|\widetilde{w}\|} \eta_t'(x)dx,$$

so:

$$\begin{aligned}\frac{1}{\Psi_t''(\|\widetilde{w}\|)} &\le \frac{1}{\Psi_t''(\mathring{x}_t)} + \frac{2}{(c+1)^2}\int_{\mathring{x}_t}^{\|\widetilde{w}\|} \eta_t'(x)dx\\ &\le \frac{1}{\Psi_t''(\mathring{x}_t)} + \frac{2}{(c+1)^2}\int_{0}^{\|\widetilde{w}\|} \eta_t'(x)dx\\ &= \frac{1}{\Psi_t''(\mathring{x}_t)} + \frac{2\eta_t(\|\widetilde{w}\|)}{(c+1)^2},\end{aligned}$$

and hence,

$$\begin{aligned}\frac{(c+1)^2 G_t^2}{2\Psi_t''(\|\widetilde{w}\|)} - \eta_t(\|\widetilde{w}\|)G_t^2 &\le \frac{(c+1)^2 G_t^2}{2\Psi_t''(\mathring{x}_t)} + \frac{(c+1)^2 G_t^2}{2}\frac{2}{(c+1)^2}\eta_t(\|\widetilde{w}\|) - \eta_t(\|\widetilde{w}\|)G_t^2\\ &= \frac{(c+1)^2 G_t^2}{2\Psi_t''(\mathring{x}_t)} + \eta_t(\|\widetilde{w}\|)G_t^2 - \eta_t(\|\widetilde{w}\|)G_t^2\\ &= \frac{(c+1)^2 G_t^2}{2\Psi_t''(\mathring{x}_t)},\end{aligned}$$

so in either case we have

$$\begin{aligned}\widehat{\delta}_t &= G_t\|w_t - w_{t+1}\| - D_{\psi_t}(w_{t+1}|w_t) - \eta_t(\|w_{t+1}\|)G_t^2\\ &\le \frac{(c+1)^2 G_t^2}{2\Psi_t''(\mathring{x}_t)}.\end{aligned}$$

□

### A.2.3 Proof of Lemma 4.1.1

**Lemma 4.1.1.** *For all $t$ let $\psi_t : W \to \mathbb{R}$ be differentiable convex functions, $\varphi_t : W \to \mathbb{R}$ be subdifferentiable convex functions, and let $\mathcal{M}_t : W \to W$ be arbitrary mappings. Then for any sequence $\boldsymbol{u} = (u_1, \ldots, u_T)$ in $W$, Algorithm 2 guarantees*

$$
\begin{aligned}
R_T(\boldsymbol{u}) \le\ & D_{\psi_{T+1}}(u_T|\widetilde{w}_1) - D_{\psi_{T+1}}(u_T|\widetilde{w}_{T+1}) + \sum_{t=1}^{T} \varphi_t(u_t) \\
& + \sum_{t=2}^{T} \underbrace{\langle \nabla\psi_t(\widetilde{w}_t) - \nabla\psi_t(\widetilde{w}_1), u_{t-1} - u_t\rangle}_{=:\mathcal{P}_t} + \sum_{t=1}^{T} \underbrace{D_{\psi_{t+1}}(u_t|\widetilde{w}_{t+1}) - D_{\psi_{t+1}}(u_t|w_{t+1})}_{\xi_t} \\
& + \sum_{t=1}^{T} \underbrace{\langle g_t, \widetilde{w}_t - w_{t+1}\rangle - D_{\psi_t}(w_{t+1}|\widetilde{w}_t) - (\Delta_t + \varphi_t)(w_{t+1})}_{=:\delta_t},
\end{aligned}
$$

*where $g_t \in \partial\ell_t(\widetilde{w}_t)$.*

*Proof.* The proof follows immediately from the general regret equality of Lemma A.1.1 by bounding $\ell_t(\widetilde{w}_t) - \ell_t(w_{t+1}) \le \langle g_t, \widetilde{w}_t - w_{t+1}\rangle$ for $g_t \in \partial\ell_t(\widetilde{w}_t)$ and observing that the terms $\sum_{t=1}^{T} - \le 0$ for convex $\ell_t$ and $\varphi_t$. Note that the result is valid even in constrained settings by including the indicator function $\mathbb{I}(w \in W)$ in $\varphi_t$, as discussed in Remark A.1.2. □

## A.3 Supporting Lemmas

In this section we collect the miscellaneous supporting lemmas used in our proofs.

**Lemma A.3.1.** *(Orabona and Pál 2021, Lemma 23) Let $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}$ be defined as $g(x) = f(\|x\|)$. If $f$ is twice differentiable at $\|x\|$ and $\|x\| > 0$ then*

$$
\min\left\{g''(\|x\|), \frac{g'(\|x\|)}{\|x\|}\right\} I \preceq \nabla^2 g(x) \preceq \max\left\{g''(\|x\|), \frac{g'(\|x\|)}{\|x\|}\right\} I
$$

**Lemma A.3.2.** *Under the same assumptions as Lemma A.3.1, further suppose that $f'(x)$ is concave and non-negative. If $f$ is twice-differentiable at $\|x\|$ and $\|x\| > 0$, then*

$$
\nabla^2 g(x) \succeq f''(\|x\|) I
$$

*Proof.* Apply Lemma A.3.1,

$$\nabla^2 g(x) \geq I \min\left\{f''(\|x\|), \frac{f'(\|x\|)}{\|x\|}\right\},$$

and use the fact that $f'(x)$ is concave and $f'(x) \geq 0$ to bound

$$\frac{f'(\|x\|)}{\|x\|} \geq \frac{f'(0) + f''(\|x\|)(\|x\| - 0)}{\|x\|} \geq f''(\|x\|).$$

□

The following lemma is common in adaptive online learning and provided for completeness.

**Lemma A.3.3.** *Let $a_1, \dots, a_T$ be arbitrary non-negative numbers in $\mathbb{R}$. Then*

$$\sqrt{\sum_{t=1}^T a_t} \leq \sum_{t=1}^T \frac{a_t}{\sqrt{\sum_{s=1}^t a_s}} \leq 2\sqrt{\sum_{t=1}^T a_t}$$

*Proof.* By concavity of $x \mapsto \sqrt{x}$, we have

$$\sqrt{a_{1:t}} - \sqrt{a_{1:t-1}} \geq \frac{a_t}{2\sqrt{a_{1:t}}},$$

so summing over $t$ and observing the resulting telescoping sum yields

$$\sum_{t=1}^T \frac{a_t}{\sqrt{a_{1:t}}} \leq 2\sum_{t=1}^T \sqrt{a_{1:t}} - \sqrt{a_{1:t-1}} = 2\sqrt{a_{1:T}}.$$

For the lower bound, observe that

$$\sum_{t=1}^T \frac{a_t}{\sqrt{a_{1:t}}} \geq \sum_{t=1}^T \frac{a_t}{\sqrt{a_{1:T}}} = \frac{a_{1:T}}{\sqrt{a_{1:T}}} = \sqrt{a_{1:T}}$$

□

The following lemma shows that we can bound sums of the form $\sum_{t=1}^T \frac{\|g_t\|^2}{\|g\|_{1:t}^2 \log^2(\|g\|_{1:t}^2/G^2)}$ by a constant.

**Lemma A.3.4.** *Let $V_t \geq 4G^2 + \sum_{s=1}^{t-1} \|g_s\|^2$ where $G \geq \|g_t\|$ for all $t$. Then*

$$\sum_{t=1}^T \frac{\|g_t\|^2}{V_t \log^2(V_t/G^2)} \leq 2$$

*Proof.* Let $c \geq 4$ and $V_t = cG^2 + \|g\|_{1:t-1}^2$. We apply the integral bound $\sum_{t=1}^T a_t f(\sum_{i=0}^t a_t) \leq \int_{a_0}^{\sum_{s=0}^t a_s} f(x) dx$ for non-increasing $f$ (Orabona 2019, Lemma 4.13) to get

$$\begin{aligned}
\sum_{t=1}^T \frac{\|g_t\|^2}{V_t \log^2(V_t/G^2)} &\leq \sum_{t=1}^T \frac{\|g_t\|^2}{\left((c-1)G^2 + \|g\|_{1:t}^2\right) \log^2\left(\frac{(c-1)G^2 + \|g\|_{1:t}^2}{G^2}\right)} \\
&\leq \int_{(c-1)G^2}^{(c-1)G^2 + \|g\|_{1:T}^2} \frac{1}{x \log^2(x/G^2)} dx = \frac{-2}{\log(x/G^2)}\bigg|_{x=(c-1)G^2}^{(c-1)G^2 + \|g\|_{1:T}^2} \\
&\leq \frac{2}{\log(c-1)} \leq 2,
\end{aligned}$$

where the last line uses $\log(c-1) \geq \log(3) \geq 1$.

□

# Appendix B

# Part II (Adaptivity in Stationary Settings)

## B.1 Details for Chapter 6

### B.1.1 Proofs for Section 6.1 (Parameter-free Learning)

**Proof of Theorem 6.1.1**

**Theorem 6.1.1.** *Let $\ell_1, \ldots, \ell_T$ be $G$-Lipschitz convex functions and $g_t \in \partial\ell_t(w_t)$ for all $t$. Let $\epsilon > 0$, $k \geq 3$, $V_t = 4G^2 + \|g\|^2_{1:t-1}$, and $\alpha_t = \frac{\epsilon G}{\sqrt{V_t}\log^2(V_t/G^2)}$. For all $t$, set*

$$\psi_t(w) = k\int_0^{\|w\|} \min_{\eta \leq 1/G}\left[\frac{\log(x/\alpha_t + 1)}{\eta} + \eta V_t\right]dx.$$

*Then for all $u \in \mathbb{R}^d$, Algorithm 3 guarantees*

$$R_T(u) \leq 4G\epsilon + 2k\|u\|\max\left\{\sqrt{V_{T+1}\log(\|u\|/\alpha_{T+1} + 1)}, G\log(\|u\|/\alpha_{T+1} + 1)\right\}$$

*Proof.* First, let us derive the update formula, which can be seen in Algorithm 3. By first-order optimality conditions for $w_{t+1} = \arg\min_{w \in \mathbb{R}^d} \langle g_t, w\rangle + D_{\psi_t}(w|w_t) + \Delta_t(w)$ we have:

$$g_t + \nabla\psi_t(w_{t+1}) - \nabla\psi_t(w_t) + \nabla\Delta_t(w_{t+1}) = \mathbf{0}$$

Expanding the definition of $\Delta_t(w) = \psi_{t+1}(w) - \psi_t(w)$, we obtain:

$$g_t + \nabla\psi_{t+1}(w_{t+1}) - \nabla\psi_t(w_t) = \mathbf{0},$$

and unrolling the recursion we have

$$\nabla\psi_{t+1}(w_{t+1}) = \nabla\psi_t(w_t) - g_t = \nabla\psi_{t-1}(w_{t-1}) - g_{t-1} - g_t = \ldots = -g_{1:t}.$$

Inspecting the equation for $\psi_{t+1}$ then yields:

$$\frac{w_{t+1}}{\|w_{t+1}\|}\Psi'_{t+1}(\|w_{t+1}\|) = -g_{1:t}$$

where we define the function

$$\begin{aligned}\Psi'_{t+1}(x) &= k\min_{\eta\le 1/G}\left[\frac{\log\left(x/\alpha_{t+1}+1\right)}{\eta} + \eta V_{t+1}\right]\\ &= \begin{cases} 2k\sqrt{V_{t+1}\log\left(x/\alpha_{t+1}+1\right)} & \text{if } G\sqrt{\log\left(x/\alpha_{t+1}+1\right)} \le \sqrt{V_{t+1}}\\ kG\log\left(x/\alpha_{t+1}+1\right) + \frac{kV_{t+1}}{G} & \text{otherwise.}\end{cases}\end{aligned}$$

From this, we immediately see that $w_{t+1} = x\frac{-g_{1:t}}{\|g_{1:t}\|}$ for some constant $x$ that satisfies:

$$\Psi'_{t+1}(x) = \|g_{1:t}\|$$

Now we see that one of two cases occurs: either

$$\Psi'_{t+1}(x) = 2k\sqrt{V_{t+1}\log\left(x/\alpha_{t+1}+1\right)},$$

which holds when $\frac{1}{G} \ge \sqrt{\log\left(x/\alpha_{t+1}+1\right)/V_{t+1}}$, or alternatively we have

$$\Psi'_{t+1}(x) = kG\log\left(x/\alpha_{t+1}+1\right) + \frac{kV_{t+1}}{G}$$

which holds when $\frac{1}{G} \le \sqrt{\log\left(x/\alpha_{t+1}+1\right)/V_{t+1}}$. Observe that at the boundary value where $\frac{1}{G} = \sqrt{\log\left(x/\alpha_{t+1}+1\right)/V_{t+1}}$ we have

$$\Psi'_{t+1}(x) = 2k\sqrt{V_{t+1}\log\left(x/\alpha_{t+1}+1\right)} = \frac{2kV_{t+1}}{G}.$$

Using this, we consider two cases. First, if $\|g_{1:t}\| \le \frac{2kV_{t+1}}{G}$, then we have

$$\begin{aligned} 2k\sqrt{V_{t+1}\log\left(\|w_{t+1}\|/\alpha_{t+1}+1\right)} &= \|g_{1:t}\| \\ \|w_{t+1}\| &= \alpha_{t+1}\left[\exp\left(\frac{\|g_{1:t}\|^2}{4k^2V_{t+1}}\right)-1\right]. \end{aligned}$$

On the other hand, if $\|g_{1:t}\| \ge \frac{2kV_{t+1}}{G}$ then

$$\begin{aligned} kG\log\left(\|w_{t+1}\|/\alpha_{t+1}+1\right)+\frac{kV_{t+1}}{G} &= \|g_{1:t}\| \\ \|w_{t+1}\| &= \alpha_{t+1}\left[\exp\left(\frac{\|g_{1:t}\|}{kG}-\frac{V_{t+1}}{G^2}\right)-1\right]. \end{aligned}$$

Putting these cases together yields the update described in Algorithm 3 (with $k = 3$, which is important later in the regret analysis).

Now, we concentrate on proving the regret bound.

For brevity we define the function $F_t(x) = \log\left(x/\alpha_t+1\right)$. Recall that we have set $\Psi_t'(x) = k\min_{\eta\le 1/G}\left[\frac{F_t(x)}{\eta}+\eta V_t\right]$ so that $\Psi_t(x) = k\int_0^x \min_{\eta\le 1/G}\left[\frac{F_t(z)}{\eta}+\eta V_t\right]dz$ and $\psi_t(w) = \Psi_t(\|w\|)$, and $\phi_t(w) = \Delta_t(w) = \Psi_{t+1}(\|w\|)-\Psi_t(\|w\|)$. We have by Lemma 4.0.1 that

$$\begin{aligned} R_T(u) &\le \psi_{T+1}(u)+\sum_{t=1}^T\delta_t \\ &\overset{(a)}{\le} \|u\|\,\Psi_{T+1}'(\|u\|)+\sum_{t=1}^T\delta_t \\ &\overset{(b)}{\le} 2k\|u\|\max\left\{\sqrt{V_{T+1}\log\left(\|u\|/\alpha_{T+1}+1\right)}, G\log\left(\|u\|/\alpha_{T+1}+1\right)\right\}+\sum_{t=1}^T\delta_t \end{aligned}$$

where $(a)$ observes that $\Psi_{T+1}'(x)$ is non-decreasing in $x$, so

$$\psi_{T+1}(u) = \int_0^{\|u\|}\Psi_{T+1}'(x)dx \le \int_0^{\|u\|}dx\Psi_{T+1}'(\|u\|) = \|u\|\,\Psi_t'(\|u\|),$$

and $(b)$ observes that $V_t/G \le GF_t(x)$ whenever $\Psi_t'(x) = kGF_t(x)+\frac{kV_t}{G}$ and hence

$$\begin{aligned} \Psi_{T+1}'(\|u\|) &= \begin{cases} 2k\sqrt{V_{T+1}F_{T+1}(\|u\|)} & \text{if } G\sqrt{F_{T+1}(\|u\|)} \le \sqrt{V_{T+1}} \\ kGF_{T+1}(\|u\|)+\frac{kG}{V_{T+1}} & \text{otherwise} \end{cases} \\ &\le \begin{cases} 2k\sqrt{V_{T+1}F_{T+1}(\|u\|)} & \text{if } G\sqrt{F_{T+1}(\|u\|)} \le \sqrt{V_{T+1}} \\ 2kGF_{T+1}(\|u\|) & \text{otherwise} \end{cases} \\ &= 2k\max\left\{\sqrt{V_{T+1}F_{T+1}(\|u\|)}, GF_{T+1}(\|u\|)\right\}. \end{aligned}$$

Thus, we need only bound the stability terms $\sum_{t=1}^{T} \delta_t$, which we will handle using the Stability Lemma (Lemma 4.0.2).

For any $x > 0$, we have

$$\Psi_t'(x) = \begin{cases} 2k\sqrt{V_t F_t(x)} & \text{if } G\sqrt{F_t(x)} \le \sqrt{V_t} \\ kGF_t(x) + \frac{kV_t}{G} & \text{otherwise} \end{cases}$$

$$\Psi_t''(x) = \begin{cases} \frac{k\sqrt{V_t}}{(x+\alpha_t)\sqrt{F_t(x)}} & \text{if } G\sqrt{F_t(x)} \le \sqrt{V_t} \\ \frac{kG}{x+\alpha_t} & \text{otherwise} \end{cases}$$

$$\Psi_t'''(x) = \begin{cases} \frac{-k\sqrt{V_t}(1+2F_t(x))}{2(x+\alpha_t)^2 F_t(x)^{3/2}} & \text{if } G\sqrt{F_t(x)} \le \sqrt{V_t} \\ \frac{-kG}{(x+\alpha_t)^2} & \text{otherwise} \end{cases}.$$

Clearly $\Psi_t(x) \ge 0$, $\Psi_t'(x) \ge 0$, $\Psi_t''(x) \ge 0$, $\Psi_t'''(x) \le 0$ for all $x > 0$. Moreover, observe that for any $x > \alpha_t(e-1) =: x_0$, we have

$$\begin{aligned} \frac{|\Psi_t'''(x)|}{\Psi_t''(x)^2} &= \begin{cases} \frac{k\sqrt{V_t}(1+2F_t(x))}{2(x+\alpha_t)^2 F_t(x)^{3/2}} \frac{(x+\alpha_t)^2 F_t(x)}{k^2 V_t} & \text{if } G\sqrt{F_t(x)} \le \sqrt{V_t} \\ \frac{kG}{(x+\alpha_t)^2} \frac{(x+\alpha_t)^2}{k^2 G^2} & \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{1}{2k\sqrt{V_t}} \left( \frac{1}{\sqrt{F_t(x)}} + 2\sqrt{F_t(x)} \right) & \text{if } G\sqrt{F_t(x)} \le \sqrt{V_t} \\ \frac{1}{kG} & \text{otherwise} \end{cases} \end{aligned}$$

Now, since $x > \alpha_t(e-1)$, we have $F_t(x) > 1$ so that $\frac{1}{\sqrt{F_t(x)}} \le \sqrt{F_t(x)}$. Thus:

$$\begin{aligned} &\le \begin{cases} \frac{3}{2k}\sqrt{\frac{F_t(x)}{V_t}} & \text{if } G\sqrt{F_t(x)} \le \sqrt{V_t} \\ \frac{1}{kG} & \text{otherwise} \end{cases} \\ &\le \frac{1}{2} \min\left\{ \sqrt{\frac{F_t(x)}{V_t}}, \frac{1}{G} \right\} = \frac{1}{2}\eta_t'(x), \end{aligned}$$

where the last line defines $\eta_t(x) = \int_0^x \min\left\{ \sqrt{\frac{F_t(v)}{V_t}}, \frac{1}{G} \right\} dv$ and uses $k \ge 3$. We also have $\eta_t'(x) = \min\left\{ \sqrt{\frac{F_t(x)}{V_t}}, \frac{1}{G} \right\} \le \frac{1}{G}$, and $\eta_t'(x)$ is monotonic, so $\eta_t(x)$ is convex and $1/G$ Lipschitz. Hence, by Lemma 4.0.2 we have

$$\widehat{\delta}_t = \langle g_t, w_t - w_{t+1} \rangle - D_{\psi_t}(w_{t+1}|w_t) - \eta_t(\|w_{t+1}\|)\|g_t\|^2 \le \frac{2\|g_t\|^2}{\Psi_t''(x_0)} \tag{B.1}$$

with $x_0 = \alpha_t(e-1)$.

Next, we want to show that $\phi_t(w) = \Delta_t(w) \geq \eta_t(\|w\|)\|g_t\|^2$, so that $\delta_t \leq \widehat{\delta}_t$. To this end, let $x > 0$ and observe that for $\alpha_{t+1} \leq \alpha_t$, we have $F_{t+1}(x) = \log(x/\alpha_{t+1}+1) \geq \log(x/\alpha_t+1) = F_t(x)$, so

$$\begin{aligned}\Psi'_{t+1}(x) - \Psi'_t(x) &= k\min_{\eta\leq\frac{1}{G}}\left[\frac{F_{t+1}(x)}{\eta}+\eta V_{t+1}\right] - k\min_{\eta\leq\frac{1}{G}}\left[\frac{F_t(x)}{\eta}+\eta V_t\right]\\ &\geq k\min_{\eta\leq\frac{1}{G}}\left[\frac{F_t(x)}{\eta}+\eta V_{t+1}\right] - k\min_{\eta\leq\frac{1}{G}}\left[\frac{F_t(x)}{\eta}+\eta V_t\right],\end{aligned}$$

and using the fact that for any $\eta \leq 1/G$ we can bound $\frac{F_t(x)}{\eta} + \eta V_{t+1} = \frac{F_t(x)}{\eta} + \eta V_t + \eta\|g_t\|^2 \geq \min_{\eta^*\leq 1/G}\left[\frac{F_t(x)}{\eta^*} + \eta^* V_t\right] + \eta\|g_t\|^2$, we have

$$\begin{aligned}&\geq k\|g_t\|^2\min\left\{\sqrt{\frac{F_t(x)}{V_{t+1}}}, \frac{1}{G}\right\} + k\min_{\eta\leq\frac{1}{G}}\left[\frac{F_t(x)}{\eta}+\eta V_t\right] - k\min_{\eta\leq\frac{1}{G}}\left[\frac{F_t(x)}{\eta}+\eta V_t\right]\\ &= k\|g_t\|^2\min\left\{\sqrt{\frac{F_t(x)}{V_{t+1}}}, \frac{1}{G}\right\} \geq \frac{k}{\sqrt{2}}\|g_t\|^2\min\left\{\sqrt{\frac{F_t(x)}{V_t}}, \frac{1}{G}\right\} \geq \|g_t\|^2\eta'_t(x),\end{aligned}$$

where the last line uses $k \geq 3$ and $\frac{1}{V_t} = \frac{1}{V_{t+1}}\frac{V_{t+1}}{V_t} = \frac{1}{V_{t+1}}\left(1+\|g_t\|^2/V_t\right) \leq \frac{2}{V_{t+1}}$ for $V_t \geq \|g_t\|^2$. From this, we immediately have

$$\Delta_t(w) = \int_0^{\|w\|}\Psi'_{t+1}(x) - \Psi'_t(x)dx \geq \|g_t\|^2\int_0^{\|w\|}\eta'_t(x)dx = \eta_t(\|w\|)\|g_t\|^2,$$

and hence

$$\begin{aligned}\delta_t &= \langle g_t, w_t - w_{t+1}\rangle - D_{\psi_t}(w_{t+1}|w_t) - \Delta_t(w_{t+1})\\ &\leq \langle g_t, w_t - w_{t+1}\rangle - D_{\psi_t}(w_{t+1}|w_t) - \eta_t(\|w_{t+1}\|)\|g_t\|^2 = \widehat{\delta}_t \leq \frac{2\|g_t\|^2}{\Psi''_t(x_0)}\end{aligned}$$

for $x_0 = \alpha_t(e-1)$ via Equation (B.1). Summing over $t$ then yields

$$\begin{aligned}\sum_{t=1}^T\delta_t &\leq \sum_{t=1}^T\frac{2\|g_t\|^2}{\Psi''_t(\alpha_t(e-1))} \leq \sum_{t=1}^T\frac{2e\alpha_t}{k}\|g_t\|^2\sqrt{\frac{F_t(\alpha_t(e-1))}{V_t}}\\ &\leq \sum_{t=1}^T\frac{2e\alpha_t}{k}\|g_t\|^2\frac{1}{\sqrt{V_t}} \leq \sum_{t=1}^T\frac{6}{k}\frac{\alpha_t\|g_t\|^2}{\sqrt{V_t}}\\ &\overset{(a)}{\leq} 2G\epsilon\sum_{t=1}^T\frac{\|g_t\|^2}{V_t\log^2(V_t/G^2)}\\ &\overset{(b)}{\leq} 4G\epsilon\end{aligned}$$

where $(a)$ chooses $\alpha_t = \frac{\epsilon G}{\sqrt{V_t}\log^2(V_t/G^2)}$ and recalls $k \geq 3$, and $(b)$ recalls $V_t = 4G^2 + \|g\|^2_{1:t-1}$ and uses

Lemma A.3.4 to bound $\sum_{t=1}^{T} \frac{\|g_t\|^2}{V_t \log^2(V_t/G^2)} \le 2$. Returning to our regret bound we have

$$
\begin{aligned}
R_T(u) &\le 2k\left\|u\right\| \max\left\{\sqrt{V_{T+1}\log\left(\left\|u\right\|/\alpha_{T+1}+1\right)}, G\log\left(\left\|u\right\|/\alpha_{T+1}+1\right)\right\} + \sum_{t=1}^{T}\delta_t \\
&\le 4G\epsilon + 2k\left\|u\right\| \max\left\{\sqrt{V_{T+1}\log\left(\left\|u\right\|/\alpha_{T+1}+1\right)}, G\log\left(\left\|u\right\|/\alpha_{T+1}+1\right)\right\} \\
&\le \widehat{O}\left(G\epsilon + \left\|u\right\|\left[\sqrt{\left\|g\right\|_{1:T}^2 \log\left(\frac{\left\|u\right\|\sqrt{\left\|g\right\|_{1:T}^2}}{\epsilon G}+1\right)} \vee G\log\left(\frac{\left\|u\right\|\sqrt{\left\|g\right\|_{1:T}^2}}{\epsilon G}+1\right)\right]\right)
\end{aligned}
$$

□

### B.1.2 Proofs for Section 6.3 (Adapting to Gradient Variability)

#### Proof of Theorem 6.3.1

**Theorem 6.3.1.** *Let $\ell_1, \dots, \ell_T$ and $\widehat{\ell}_1, \dots, \widehat{\ell}_T$ be $G$-Lipschitz convex functions. Let $\epsilon > 0$, $k \ge 3$, and for all $t$ set $\widehat{V}_t = 16G^2 + \sum_{s=1}^{t-1}\left\|\nabla\ell_s(w_s) - \nabla\widehat{\ell}_s(w_s)\right\|^2$, $\widehat{\alpha}_t = \frac{\epsilon G}{\sqrt{\widehat{V}_t}\log^2(\widehat{V}_t/G^2)}$, and*

$$
\psi_t(w) = k\int_0^{\|w\|} \min_{\eta \le \frac{1}{2G}}\left[\frac{\log\left(x/\widehat{\alpha}_t+1\right)}{\eta} + \eta\widehat{V}_t\right]dx.
$$

*Then for all $u \in \mathbb{R}^d$, Algorithm 5 guarantees*

$$
R_T(u) \le 4\epsilon G + 2k\left\|u\right\| \max\left\{\sqrt{\widehat{V}_t\log\left(\left\|u\right\|/\widehat{\alpha}_{T+1}+1\right)}, 2G\log\left(\left\|u\right\|/\widehat{\alpha}_{T+1}+1\right)\right\}
$$

*Proof.* The proof follows similar steps to Theorem 6.1.1. Let $g_t \in \ell_t(w_t)$ and let $h_t \in \partial\widehat{\ell}_t(w_t)$ be the subgradient of $\widehat{\ell}_t(w_t)$ for which the first-order optimality condition $h_t + \nabla\psi_t(w_t) - \nabla\psi_t(x_t) = \mathbf{0}$ holds. Then

$$
\begin{aligned}
\sum_{t=1}^{T}\langle g_t, w_t - u\rangle &= \sum_{t=1}^{T}\langle g_t, x_{t+1} - u\rangle + \langle g_t, w_t - x_{t+1}\rangle \\
&= \sum_{t=1}^{T}\langle g_t, x_{t+1} - u\rangle + \langle h_t, w_t - x_{t+1}\rangle + \langle g_t - h_t, w_t - x_{t+1}\rangle\,.
\end{aligned}
$$

Following the same steps as Lemma 4.0.1 we have

$$\begin{aligned}\sum_{t=1}^{T}\langle g_t, x_{t+1}-u\rangle &\leq D_{\psi_{T+1}}(u|x_1)-D_{\psi_{T+1}}(u|x_{T+1})+\sum_{t=1}^{T}-D_{\psi_t}(x_{t+1}|x_t)-\phi_t(x_{t+1})\\ &\leq \psi_{T+1}(u)+\sum_{t=1}^{T}-D_{\psi_t}(x_{t+1}|x_t)-\phi_t(w_{t+1}),\end{aligned}$$

where the last line observes $\arg\min_{x\in\mathbb{R}^d}\psi_{T+1}(x)=\psi_{T+1}(x_1)=0$, so $D_{\psi_{T+1}}(u|x_1)=\psi_{T+1}(u)$ and $-D_{\psi_{T+1}}(u|x_{T+1})\leq 0$. Similarly, from the first-order optimality condition for $w_t$ we have

$$\begin{aligned}\sum_{t=1}^{T}\langle h_t, w_t-x_{t+1}\rangle &= \sum_{t=1}^{T}\langle\nabla\psi_t(w_t)-\nabla\psi_t(x_t), w_t-x_{t+1}\rangle\\ &= \sum_{t=1}^{T}D_{\psi_t}(x_{t+1}|x_t)-D_{\psi_t}(x_{t+1}|w_t)\underbrace{-D_{\psi_t}(w_t|x_t)}_{\leq 0}\\ &\leq \sum_{t=1}^{T}D_{\psi_t}(x_{t+1}|x_t)-D_{\psi_t}(x_{t+1}|w_t)\end{aligned}$$

where the second line applies the three-point relation for Bregman divergences:

$$\langle\nabla f(y)-\nabla f(x), x-z\rangle = D_f(z|y)-D_f(z|x)-D_f(x|y).$$

Combining these two observations yields

$$R_T(u)\leq\psi_{T+1}(u)+\sum_{t=1}^{T}\underbrace{\langle g_t-h_t, w_t-x_{t+1}\rangle-D_{\psi_t}(x_{t+1}|w_t)-\phi_t(x_{t+1})}_{=:\delta_t}$$

To bound $\delta_t$, define $\widehat{g}_t=\nabla\ell_t(w_t)-\nabla\widehat{\ell}_t(w_t)$, $\widehat{G}=2G$, $\widehat{V}_t=4\widehat{G}^2+\sum_{s=1}^{t-1}\|\widehat{g}_s\|^2$, $\widehat{\alpha}_t=\frac{\epsilon G}{\sqrt{\widehat{V}_t}\log^2(\widehat{V}_t/G^2)}$, and observe that $\psi_t(w)=k\int_0^{\|w\|}\min_{\eta\leq 1/\widehat{G}}\left[\frac{\log(x/\widehat{\alpha}_t+1)}{\eta}+\eta\widehat{V}_t\right]dx$ is equivalent to the regularizer from Theorem 6.1.1. Hence, borrowing the arguments of Theorem 6.1.1, we can bound $\sum_{t=1}^{T}\delta_t\leq 4\epsilon G$. Returning to our regret bound, we have

$$\begin{aligned}R_T(u)&\leq\psi_{T+1}(u)+4\epsilon G\overset{(a)}{\leq}4\epsilon G+\|u\|\,\Psi'_{T+1}(\|u\|)\\ &\overset{(b)}{\leq}4\epsilon G+2k\|u\|\max\left\{\sqrt{\widehat{V}_t\log\left(\|u\|/\widehat{\alpha}_{T+1}+1\right)}, 2G\log\left(\|u\|/\widehat{\alpha}_{T+1}+1\right)\right\}\end{aligned}$$

where $(a)$ defines

$$\begin{aligned}\Psi'_{T+1}(x) &= k \min_{\eta\le 1/2G}\left[\frac{\log\left(x/\widehat{\alpha}_{T+1}+1\right)}{\eta}+\eta\widehat{V}_{T+1}\right]\\ &= \begin{cases} 2k\sqrt{\widehat{V}_{T+1}\log\left(x/\widehat{\alpha}_{T+1}+1\right)} & \text{if } 2G\sqrt{\log\left(x/\widehat{\alpha}_{T+1}+1\right)}\le\sqrt{\widehat{V}_{T+1}}\\ 2kG\log\left(x/\widehat{\alpha}_{T+1}+1\right)+\frac{k\widehat{V}_{T+1}}{2G} & \text{otherwise}\end{cases}\end{aligned}$$

and observes that $\psi_{T+1}(u)=\int_0^{\|u\|}\Psi'_t(x)dx\le\|u\|\,\Psi'_t(\|u\|)$ since $\Psi'_t$ is non-decreasing in its argument, and $(b)$ observes that the case $\Psi'_t(x)=2kG\log\left(x/\widehat{\alpha}_{T+1}+1\right)+\frac{k\widehat{V}_{T+1}}{2G}$, coincides with $\widehat{V}_{T+1}/2G\le\sqrt{\widehat{V}_{T+1}\log\left(x/\widehat{\alpha}_{T+1}+1\right)}\le 2G\log\left(x/\widehat{\alpha}_{T+1}+1\right)$, so

$$\begin{aligned}\Psi'_{T+1}(x) &\le \begin{cases} 2k\sqrt{\widehat{V}_{T+1}\log\left(x/\widehat{\alpha}_{T+1}+1\right)} & \text{if } 2G\sqrt{\log\left(x/\widehat{\alpha}_{T+1}+1\right)}\le\sqrt{\widehat{V}_{T+1}}\\ 4kG\log\left(x/\widehat{\alpha}_{T+1}+1\right) & \text{otherwise}\end{cases}\\ &= 2k\max\left\{\sqrt{\widehat{V}_{T+1}\log\left(x/\widehat{\alpha}_{T+1}+1\right)},2G\log\left(x/\widehat{\alpha}_{T+1}+1\right)\right\}\end{aligned}$$

□

### B.1.3 Proofs for Section 6.2 (Lipschitz Adaptivity and Scale-free Learning)

**Proof of Theorem 6.2.1**

The complete theorem is stated below.

**Theorem 6.2.1.** *Let $\ell_1,\ldots,\ell_T$ be $G$-Lipschitz convex functions and $g_t\in\partial\ell_t(w_t)$ for all $t$. Let $h_1\le\ldots\le h_T$ be a sequence of hints such that $h_t\ge\|g_t\|$, and assume that $h_t$ is provided at the start of each round $t$. Let $\epsilon>0$, $k\ge 3$, $V_t=4h_t^2+\|g\|^2_{1:t-1}$, $B_t=4\sum_{s=1}^t\left(4+\sum_{s'=1}^{s-1}\frac{\|g_{s'}\|^2}{h_{s'}^2}\right)$, $\alpha_t=\frac{\epsilon}{\sqrt{B_t}\log^2(B_t)}$, and set*

$$\psi_t(w)=k\int_0^{\|w\|}\min_{\eta\le\frac{1}{h_t}}\left[\frac{\log\left(x/\alpha_t+1\right)}{\eta}+\eta V_t\right]dx.$$

*Then for all $u\in\mathbb{R}^d$, Algorithm 2 guarantees*

$$\begin{aligned}R_T(u)\le 4\epsilon h_T+2k\|u\|\max\Bigg\{&\sqrt{V_{T+1}\log\left(\frac{\|u\|\sqrt{B_{T+1}}\log^2\left(B_{T+1}\right)}{\epsilon}+1\right)},\\ &h_T\log\left(\frac{\|u\|\sqrt{B_{T+1}}\log^2\left(B_{T+1}\right)}{\epsilon}+1\right)\Bigg\}\end{aligned}$$

*Proof.* The proof follows similar steps to Theorem 6.1.1. We have via Lemma 4.0.1 that

$$R_T(u) \le \psi_{T+1}(u) + \sum_{t=1}^{T} \underbrace{\langle g_t, w_t - w_{t+1}\rangle - D_{\psi_t}(w_{t+1}|w_t) - \Delta_t(w_{t+1})}_{=:\delta_t},$$

so the main challenge is to bound the stability terms $\sum_{t=1}^{T} \delta_t$, which we focus on first.

Let $F_t(w) = \log(x/\alpha_t + 1)$ and define

$$\Psi_t(x) = k \int_0^x \min_{\eta \le 1/h_t} \left[ \frac{F_t(x)}{\eta} + \eta V_t \right] dx,$$

so that $\psi_t(w) = \Psi_t(\|w\|)$, and observe that

$$\begin{aligned}
\Psi_t'(x) &= k \min_{1/h_t} \left[ \frac{F_t(x)}{\eta} + \eta V_t \right] \\
&= \begin{cases} 2k\sqrt{V_t F_t(x)} & \text{if } h_t\sqrt{F_t(x)} \le \sqrt{V_t} \\ k h_t F_t(x) + \frac{kV_t}{h_t} & \text{otherwise} \end{cases} \\
\Psi_t''(x) &= \begin{cases} \frac{k}{x+\alpha_t}\sqrt{\frac{V_t}{F_t(x)}} & \text{if } h_t\sqrt{F_t(x)} \le \sqrt{V_t} \\ \frac{k h_t}{x+\alpha_t} & \text{otherwise} \end{cases} \\
\Psi_t'''(x) &= \begin{cases} \frac{-k\sqrt{V_t}(1+2F_t(x))}{2(x+\alpha_t)^2 F_t(x)^{3/2}} & \text{if } h_t\sqrt{F_t(x)} \le \sqrt{V_t} \\ \frac{-k h_t}{(x+\alpha_t)^2} & \text{otherwise.} \end{cases}
\end{aligned}$$

Hence, $\Psi_t(x) \ge 0$, $\Psi_t'(x) \ge 0$, $\Psi_t''(x) \ge 0$, and $\Psi_t'''(x) \le 0$ for all $x > 0$. Moreover, for any $x > \alpha_t(e-1) \overset{\text{def}}{=} x_0$ we have

$$\begin{aligned}
-\frac{\Psi_t'''(x)}{\Psi_t''(x)^2} &= \begin{cases} \frac{k\sqrt{V_t}(1+2F_t(x))}{2(x+\alpha_t)F_t(x)^{3/2}} \frac{(x+\alpha_t)^2 F_t(x)}{k^2 V_t} & \text{if } h_t\sqrt{F_t(x)} \le \sqrt{V_t} \\ \frac{k h_t}{(x+\alpha_t)^2} \frac{(x+\alpha_t)^2}{k^2 h_t^2} & \text{otherwise} \end{cases} \\
&\le \begin{cases} \frac{1}{2k\sqrt{V_t}} \left( \frac{1}{\sqrt{F_t(x)}} + 2\sqrt{F_t(x)} \right) & \text{if } h_t\sqrt{F_t(x)} \le \sqrt{V_t} \\ \frac{1}{k h_t} & \text{otherwise,} \end{cases}
\end{aligned}$$

and since $x > x_0$, we have $\sqrt{F_t(x)} > 1$ and $\frac{1}{\sqrt{F_t(x)}} \le \sqrt{F_t(x)}$, hence

$$\begin{aligned}
&\le \begin{cases} \frac{3}{2k}\sqrt{\frac{F_t(x)}{V_t}} & \text{if } h_t\sqrt{F_t(x)} \le \sqrt{V_t} \\ \frac{1}{kh_t} & \text{otherwise} \end{cases} \\
&\le \frac{1}{2}\min\left\{\sqrt{\frac{F_t(x)}{V_t}}, \frac{1}{h_t}\right\} \\
&= \frac{1}{2}\eta_t'(x),
\end{aligned}$$

for $k \ge 3$ and $\eta_t(x) = \int_0^{\|w\|} \min\left\{\sqrt{\frac{F_t(x)}{V_t}}, \frac{1}{h_t}\right\} dx$. Notice that $\eta_t$ is convex and $1/h_t$ Lipschitz with $h_t \ge \|g_t\|$. Hence, $\Psi_t$ satisfies the conditions of Lemma 4.0.2 with $x_0 = \alpha_t(e-1)$, so

$$\widehat{\delta_t} = \langle g_t, w_t - w_{t+1}\rangle - D_{\psi_t}(w_{t+1}|w_t) - \eta_t(\|w_{t+1}\|)\|g_t\|^2 \le \frac{2\|g_t\|^2}{\Psi_t''(x_0)}. \tag{B.2}$$

Next, we want to show that $\delta_t \le \widehat{\delta_t}$, which will follow if we can show that $\Delta_t(w) \ge \eta_t(\|w\|)\|g_t\|^2$ for any $w$. Observe that for any $x > 0$ we have

$$\begin{aligned}
\Psi_{t+1}'(x) - \Psi_t'(x) &= k \min_{\eta \le \frac{1}{h_{t+1}}}\left[\frac{F_{t+1}(x)}{\eta} + \eta V_{t+1}\right] - k\min_{\eta \le \frac{1}{h_t}}\left[\frac{F_t(x)}{\eta} + \eta V_t\right] \\
&\ge k\min_{\eta \le \frac{1}{h_t}}\left[\frac{F_{t+1}(x)}{\eta} + \eta V_{t+1}\right] - k\min_{\eta \le \frac{1}{h_t}}\left[\frac{F_t(x)}{\eta} + \eta V_t\right].
\end{aligned}$$

Now observe that for any $\eta \le 1/h_t$, if we define $\Delta_h = 4h_{t+1}^2 - 4h_t^2$, it holds that $\frac{F_{t+1}(x)}{\eta} + \eta V_{t+1} \ge \frac{F_{t+1}(x)}{\eta} + \eta V_t + \eta(\Delta_h + \|g_t\|^2) \ge \min_{\eta^* \le 1/h_t}\left[\frac{F_{t+1}(x)}{\eta^*} + \eta^* V_t\right] + \eta(\Delta_h + \|g_t\|^2)$, which yields

$$\begin{aligned}
&\ge k(\Delta_h + \|g_t\|^2)\min\left\{\sqrt{\frac{F_{t+1}(x)}{V_{t+1}}}, \frac{1}{h_t}\right\} + k\min_{\eta \le \frac{1}{h_t}}\left[\frac{F_{t+1}(x)}{\eta} + \eta V_t\right] - k\min_{\eta \le \frac{1}{h_t}}\left[\frac{F_t(x)}{\eta} + \eta V_t\right] \\
&\overset{(a)}{\ge} k(\Delta_h + \|g_t\|^2)\min\left\{\sqrt{\frac{F_t(x)}{V_{t+1}}}, \frac{1}{h_t}\right\} + k\min_{\eta \le \frac{1}{h_t}}\left[\frac{F_t(x)}{\eta} + \eta V_t\right] - k\min_{\eta \le \frac{1}{h_t}}\left[\frac{F_t(x)}{\eta} + \eta V_t\right] \\
&= k(\Delta_h + \|g_t\|^2)\min\left\{\sqrt{\frac{F_t(x)}{V_{t+1}}}, \frac{1}{h_t}\right\} \overset{(b)}{\ge} \frac{k}{\sqrt{2}}\|g_t\|^2 \min\left\{\sqrt{\frac{F_t(x)}{V_t}}, \frac{1}{h_t}\right\} \\
&\ge \|g_t\|^2 \min\left\{\sqrt{\frac{F_t(x)}{V_t}}, \frac{1}{h_t}\right\} = \eta_t'(x)\|g_t\|^2,
\end{aligned}$$

where $(a)$ uses that $\alpha_{t+1} \le \alpha_t$ so $F_{t+1}(x) = \log(x/\alpha_{t+1} + 1) \ge \log(x/\alpha_t + 1) = F_t(x)$. For inequality

$(b)$, observe that:

$$\begin{aligned}
\frac{\Delta_h + \|g_t\|^2}{\sqrt{V_{t+1}}} &= \frac{\Delta_h + \|g_t\|^2}{\sqrt{V_t + \Delta_h + \|g_t\|^2}} \\
&\geq \inf_{\Delta_h} \frac{\Delta_h + \|g_t\|^2}{\sqrt{V_t + \Delta_h + \|g_t\|^2}} \\
&\geq \frac{\|g_t\|^2}{\sqrt{V_t + \|g_t\|^2}} \\
&= \frac{\|g_t\|^2}{\sqrt{V_t}} \sqrt{\frac{V_t}{V_t + \|g_t\|^2}}
\end{aligned}$$

Observing that $\|g_t\| \leq h_t$ and $V_t \geq 4h_t^2$:

$$\geq \frac{\|g_t\|^2}{\sqrt{2V_t}}$$

From this we immediately have

$$\Delta_t(w) = \int_0^{\|w\|} \Psi'_{t+1}(x) - \Psi'_t(x) dx \geq \|g_t\|^2 \int_0^{\|w\|} \eta'_t(x) dx = \eta_t(\|w\|) \|g_t\|^2 ,$$

so combining this with Equation (B.2), we have

$$\begin{aligned}
\delta_t &= \langle g_t, w_t - w_{t+1} \rangle - D_{\psi_t}(w_{t+1}|w_t) - \Delta_t(w_{t+1}) \\
&\leq \langle g_t, w_t - w_{t+1} \rangle - D_{\psi_t}(w_{t+1}|w_t) - \eta_t(\|w_{t+1}\|) \|g_t\|^2 \\
&= \widehat{\delta}_t \leq \frac{2\|g_t\|^2}{\Psi''_t(x_0)} = \frac{2\alpha_t e \|g_t\|^2}{k\sqrt{V_t}} \leq \frac{2\alpha_t \|g_t\|^2}{\sqrt{V_t}}
\end{aligned}$$

for $k \geq 3$. Returning to our regret bound, we have

$$\begin{aligned}
R_T(u) &\leq \psi_{T+1}(u) + \sum_{t=1}^T \delta_t \leq \psi_{T+1}(u) + 2\sum_{t=1}^T \frac{\alpha_t \|g_t\|^2}{\sqrt{V_t}} \\
&\overset{(a)}{\leq} \|u\| \Psi'_{T+1}(\|u\|) + 2\sum_{t=1}^T \frac{\alpha_t \|g_t\|^2}{\sqrt{V_t}} \\
&\overset{(b)}{\leq} 2k\|u\| \max\left\{\sqrt{V_{T+1}\log\left(\|u\|/\alpha_{T+1} + 1\right)}, h_{T+1}\log\left(\|u\|/\alpha_{T+1} + 1\right)\right\} \\
&\quad + 2\sum_{t=1}^T \frac{\alpha_t \|g_t\|^2}{\sqrt{V_t}} \qquad \text{(B.3)}
\end{aligned}$$

where $(a)$ observes that $\Psi_t'(x)$ is increasing in $x$, so

$$\psi_{T+1}(u) = \int_0^{\|u\|} \Psi_{T+1}'(x)dx \le \Psi_t'(\|u\|) \int_0^{\|u\|} dx = \|u\| \, \Psi_t'(\|u\|),$$

and $(b)$ observes that the case $\Psi_t'(x) = kh_tF_t(x) + \frac{kV_t}{h_t}$ coincides with $\frac{V_t}{h_t} \le h_tF_t(x)$, so

$$\begin{aligned}\Psi_{T+1}'(\|u\|) &= \begin{cases} 2k\sqrt{V_{T+1}F_{T+1}(\|u\|)} & \text{if } h_{T+1}\sqrt{F_{T+1}(\|u\|)} \le \sqrt{V_{T+1}} \\ kh_{T+1}F_{T+1}(\|u\|) + \frac{kV_{T+1}}{h_{T+1}} & \text{otherwise} \end{cases} \\ &\le \begin{cases} 2k\sqrt{V_{T+1}F_{T+1}(\|u\|)} & \text{if } h_{T+1}\sqrt{F_{T+1}(\|u\|)} \le \sqrt{V_{T+1}} \\ 2kh_{T+1}F_{T+1}(\|u\|) & \text{otherwise} \end{cases} \\ &= 2k \max\left\{\sqrt{V_{T+1}F_{T+1}(\|u\|)}, h_{T+1}F_{T+1}(\|u\|)\right\}.\end{aligned}$$

Note that the regret does not depend on $g_{T+1}$, so without loss of generality we can assume $g_{T+1} = g_T$ and hence $h_{T+1} = h_T$. Finally, Lemma B.1.1 bounds $2\sum_{t=1}^T \frac{\alpha_t\|g_t\|^2}{\sqrt{V_t}} \le 4\epsilon h_T$, so plugging this into Equation (B.3) yields the stated result. Notice that Lemma B.1.1 is responsible for removing the "range-ratio" problem addressed by Mhammedi and Koolen 2020 via a doubling-like restart scheme. □

**Lemma B.1.1.** *Let $c \ge 4$, $V_t = ch_t^2 + \|g\|_{1:t-1}^2$, $B_t = c\sum_{s=1}^t \left(4 + \sum_{s'=1}^{s-1} \frac{\|g_{s'}\|^2}{h_{s'}^2}\right)$ and set $\alpha_t = \frac{\epsilon}{\sqrt{B_t}\log^2(B_t)}$. Then*

$$\sum_{t=1}^T \frac{\alpha_t \|g_t\|^2}{\sqrt{V_t}} \le 2\epsilon h_T.$$

*Proof.* Define $\tau_1 = 1$ and $\tau_t = \max\left\{t' : t' \le t \text{ and } \sum_{s=1}^{t'-1} \frac{\|g_s\|^2}{h_s^2} + 4 < \frac{h_{t'}^2}{h_{\tau_{t'-1}}^2}\right\}$ for $t > 1$. Then, we partition $[1, T]$ into the disjoint intervals $[1, T] = \mathcal{I}_1 \cup \ldots \cup \mathcal{I}_N$ over which $\tau_t$ is fixed. Denote $\mathcal{I}_j = [\widetilde{\tau}_j, \widetilde{\tau}_{j+1} - 1]$ where $\widetilde{\tau}_1 = 1$, $\widetilde{\tau}_{N+1} = T + 1$, and $\widetilde{\tau}_j = \min\{t > \widetilde{\tau}_{j-1} : \tau_t > \tau_{t-1}\}$ for $j \in [2, N]$. Observe that by definition, $\tau_t = \widetilde{\tau}_j$ for all $t \in \mathcal{I}_j$. Further, for all $j$ and $t \in \mathcal{I}_j$, we have either $t = \widetilde{\tau}_j$ or $\tau_{t-1} = \widetilde{\tau}_j < t$, so that:

$$\frac{h_t^2}{h_{\widetilde{\tau}_j}^2} \le 4 + \sum_{s=1}^{t-1} \frac{\|g_s\|^2}{h_s^2}$$

Now, we show that $V_{t+1}/h_{\tau_{t+1}^2} \le B_{t+1}$. Notice that if $t$ is the last round of an interval $\mathcal{I}_k$, then $t + 1$ would be the start of the next epoch so $h_{\tau_{t+1}} = h_{t+1}$ and $V_{t+1}/h_{\tau_{t+1}}^2 = V_{t+1}/h_{t+1}^2 \le B_{t+1}$ (since

$c \geq 1$). Otherwise, $t+1$ occurs before the end of interval $\mathcal{I}_k$ so

$$\begin{aligned}\frac{V_{t+1}}{h^2_{\tau_{t+1}}} = \frac{ch^2_{t+1} + \|g\|^2_{1:t}}{h^2_{\widetilde{\tau}_k}} &\leq c\frac{h^2_{t+1}}{h^2_{\widetilde{\tau}_k}} + \sum_{j=1}^{k} \sum_{\substack{s\in\mathcal{I}_j \\ s\leq t}} \frac{\|g_s\|^2}{h^2_{\widetilde{\tau}_j}} \\ &\leq c\frac{h^2_{t+1}}{h^2_{\widetilde{\tau}_k}} + \sum_{j=1}^{k} \sum_{\substack{s\in\mathcal{I}_j \\ s\leq t}} \frac{h^2_s}{h^2_{\widetilde{\tau}_j}}\end{aligned}$$

Now, apply the definition of $\widetilde{\tau}_j$ to get:

$$\begin{aligned}&\leq c\left(4 + \sum_{s=1}^{t} \frac{\|g_s\|^2}{h^2_s}\right) + \sum_{j=1}^{k} \sum_{\substack{s\in\mathcal{I}_j \\ s\leq t}} \left(4 + \sum_{s'=1}^{s-1} \frac{\|g_{s'}\|^2}{h^2_{s'}}\right) \\ &\leq c\left(4 + \sum_{s=1}^{t} \frac{\|g_s\|^2}{h^2_s}\right) + \sum_{s=1}^{t} \left(4 + \sum_{s'=1}^{s-1} \frac{\|g_{s'}\|^2}{h^2_{s'}}\right) \\ &\leq c\sum_{s=1}^{t+1} \left(4 + \sum_{s'=1}^{s-1} \frac{\|g_{s'}\|^2}{h^2_{s'}}\right) = B_{t+1}.\end{aligned}$$

Now, using this we have that $\alpha_t = \frac{\varepsilon}{\sqrt{B_t}\log^2(B_t)} \leq \frac{\varepsilon h_{\tau_t}}{\sqrt{V_t}\log^2\left(V_t/h_{\tau_t^2}\right)}$ and thus

$$\begin{aligned}\sum_{t=1}^{T} \frac{\alpha_t \|g_t\|^2}{\sqrt{V_t}} = \sum_{j=1}^{N}\sum_{t\in\mathcal{I}_j} \frac{\alpha_t \|g_t\|^2}{\sqrt{V_t}} &= \epsilon \sum_{j=1}^{N}\sum_{t\in\mathcal{I}_j} \frac{\|g_t\|^2}{\sqrt{V_t}\sqrt{B_t}\log^2(B_t)} \leq \epsilon \sum_{j=1}^{N}\sum_{t\in\mathcal{I}_j} h_{\tau_t} \frac{\|g_t\|^2}{V_t \log^2\left(V_t/h^2_{\tau_t}\right)} \\ &= \epsilon \sum_{j=1}^{N} h_{\widetilde{\tau}_j} \sum_{t\in\mathcal{I}_j} \frac{\|g_t\|^2}{\left(ch^2_t + \|g\|^2_{1:t-1}\right)\log^2\left(\frac{ch^2_t + \|g\|^2_{1:t-1}}{h^2_{\tau_t}}\right)} \\ &\leq \epsilon \sum_{j=1}^{N} h_{\widetilde{\tau}_j} \sum_{t\in\mathcal{I}_j} \frac{\|g_t\|^2}{\left((c-1)h^2_{\widetilde{\tau}_j} + \|g\|^2_{1:t}\right)\log^2\left(\frac{(c-1)h^2_{\widetilde{\tau}_j} + \|g\|^2_{1:t}}{h^2_{\widetilde{\tau}_j}}\right)} \\ &\leq \epsilon \sum_{j=1}^{N} h_{\widetilde{\tau}_j} \int_{(c-1)h^2_{\widetilde{\tau}_j}}^{(c-1)h^2_{\widetilde{\tau}_j} + \|g\|^2_{1:t}} \frac{1}{x\log^2(x/h^2_{\widetilde{\tau}_j})} dx \\ &= \epsilon \sum_{j=1}^{N} h_{\widetilde{\tau}_j} \frac{-1}{\log\left(x/h^2_{\widetilde{\tau}_j}\right)} \Bigg|_{x=(c-1)h^2_{\widetilde{\tau}_j}}^{(c-1)h^2_{\widetilde{\tau}_j} + \|g\|^2_{1:t}} \leq \frac{\epsilon}{\log(c-1)} \sum_{j=1}^{N} h_{\widetilde{\tau}_j}.\end{aligned}$$

Notice that each interval begins when $\frac{h^2_{\widetilde{\tau}_j}}{h^2_{\widetilde{\tau}_{j-1}}} > \sum_{s=1}^{t-1} \frac{\|g_s\|^2}{h^2_s} + 4 > 4$, so $h_{\widetilde{\tau}_j} > 2h_{\widetilde{\tau}_{j-1}}$ and hence

$$\frac{\epsilon}{\log(c-1)} \sum_{j=1}^{N} h_{\widetilde{\tau}_j} \leq \frac{\epsilon}{\log(c-1)} \sum_{j=0}^{N-1} \frac{1}{2^j} h_{\widetilde{\tau}_N} \leq \frac{2\epsilon h_T}{\log(c-1)} \leq 2\epsilon h_T,$$

for $c > 4$. □

### B.1.4 Proofs for Section 6.4 (Trade-offs in the Horizon Dependence)

#### Proof of Theorem 6.4.2

**Theorem 6.4.2.** *Under the same assumptions as Theorem 6.1.1, let $\rho \in [0, \frac{1}{2})$ and suppose we set $\alpha_t = \epsilon G^{2\rho} / V_t^{\rho}$ for all $t$. Then for all $u \in \mathbb{R}^d$, Algorithm 2 guarantees*

$$R_T(u) \le O\left( \frac{\epsilon G^{2\rho}}{1 - 2\rho} V_{T+1}^{\frac{1}{2} - \rho} + \|u\| \left[ \sqrt{V_{T+1} \log\left( \frac{\|u\| V_{T+1}^{\rho}}{\epsilon G^{2\rho}} + 1 \right)} \vee G \log\left( \frac{\|u\| V_{T+1}^{\rho}}{\epsilon G^{2\rho}} + 1 \right) \right] \right),$$

*where $V_{T+1} \le O(\|g\|_{1:T}^2)$.*

*Proof.* We will prove the result with $\alpha_t = \epsilon G^{1-2\beta} / V_t^{\frac{1}{2} - \beta}$ for $\beta \in (0, \frac{1}{2}]$ and then conclude by choosing $\beta = \frac{1}{2} - \rho$ for $\rho \in [0, \frac{1}{2})$. Following the same steps as Theorem 6.1.1, we have

$$\begin{aligned} R_T(u) &\le \psi_{T+1}(u) + \sum_{t=1}^{T} \delta_t \\ &\le \psi_{T+1}(u) + \sum_{t=1}^{T} \frac{2\alpha_t \|g_t\|^2}{\sqrt{V_t}} \end{aligned}$$

and substituting $\alpha_t = \epsilon G^{1-2\beta} / V_t^{\frac{1}{2} - \beta}$,

$$\begin{aligned} &= \psi_{T+1}(u) + \sum_{t=1}^{T} \frac{2\epsilon G^{1-2\beta} \|g_t\|^2}{V_t^{1-\beta}} \\ &\le \psi_{T+1}(u) + 2\epsilon G^{1-2\beta} \sum_{t=1}^{T} \frac{\|g_t\|^2}{(\|g\|_{1:t}^2)^{1-\beta}} \end{aligned}$$

where we've used $V_t = 4G^2 + \|g\|_{1:t-1}^2 \ge \|g\|_{1:t}^2$. Moreover, by concavity of $x \mapsto x^{\beta}$ for $\beta < 1$ we have

$(\|g\|_{1:t}^2)^\beta - (\|g\|_{1:t-1}^2)^\beta \geq \frac{\beta\|g_t\|^2}{(\|g\|_{1:t}^2)^{1-\beta}}$, and hence $\sum_{t=1}^T \frac{\|g_t\|^2}{(\|g\|_{1:t}^2)^{1-\beta}} \leq \frac{1}{\beta}(\|g\|_{1:T}^2)^\beta$, giving an overall bound of

$$\begin{aligned}
R_T(u) &\leq \psi_{T+1}(u) + \frac{2\epsilon G^{1-2\beta}}{\beta}(\|g\|_{1:T}^2)^\beta \\
&\leq k\|u\|\left[\sqrt{V_{T+1}\log\left(\frac{\|u\|}{\alpha_{T+1}}+1\right)+1} \vee G\log\left(\frac{\|u\|}{\alpha_{T+1}}+1\right)\right] + \frac{2\epsilon G^{1-2\beta}}{\beta}(V_{T+1})^\beta \\
&\leq O\left(\frac{\epsilon}{\beta}G^{1-2\beta}V_{T+1}^\beta + \|u\|\left[\sqrt{V_{T+1}\log\left(\frac{\|u\|V_{T+1}^{\frac{1}{2}-\beta}}{\epsilon G^{1-2\beta}}+1\right)} \vee G\log\left(\frac{\|u\|V_{T+1}^{\frac{1}{2}-\beta}}{\epsilon G^{1-2\beta}}+1\right)\right]\right)
\end{aligned}$$

where the first inequality bounds $\psi_{T+1}(u)$ using the same argument as Theorem 6.1.1. Substituting $\beta = \frac{1}{2} - \rho$ gives the stated result. □

### Proof of Theorem 6.4.3

**Theorem 6.4.3.** *Under the same assumptions as Theorem 6.2.1, let $\rho \in [0, \frac{1}{2})$ and suppose we set $B_t^\rho = \left(4\sum_{s=1}^t\left[2^{\frac{1}{\rho}} + \sum_{s'=1}^{s-1}\frac{\|g_{s'}\|^2}{h_{s'}^2}\right]\right)^\rho$ and $\alpha_t = \epsilon/B_t^\rho$ for all $t$.*[1] *Then for all $u \in \mathbb{R}^d$, Algorithm 2 guarantees*

$$R_T(u) \leq O\left(\frac{\epsilon h_T^{2\rho}}{1-2\rho}V_{T+1}^{\frac{1}{2}-\rho} + \|u\|\left[\sqrt{V_{T+1}\log\left(\frac{\|u\|B_{T+1}^\rho}{\epsilon}+1\right)} \vee h_T\log\left(\frac{\|u\|B_{T+1}^\rho}{\epsilon}+1\right)\right]\right)$$

*where and $V_{T+1} \leq O(\|g\|_{1:T}^2)$.*

*Proof.* First note that $\lim_{\rho\to 0} B_t^\rho = \lim_{\rho\to 0}\left[4\sum_{s=1}^t\left[2^{\frac{1}{\rho}} + \sum_{s'=1}^{s-1}\frac{\|g_s\|^2}{h_{s^2}}\right]\right]^\rho = 2$, so for the case $\rho = 0$ we let $B_t^\rho = 2$ for all $t$. Then following the same argument as Proposition 6.4.1 we get

$$R_T(u) \leq O\left(\epsilon\sqrt{\|g\|_{1:T}^2} + \|u\|\left[\sqrt{\|g\|_{1:T}^2\log\left(\frac{\|u\|}{\epsilon}+1\right)} \vee h_T\log\left(\frac{\|u\|}{\epsilon}+1\right)\right]\right).$$

Next, we consider the case $\rho > 0$. Similar to Theorem 6.4.2, we prove the result with $B_t = 4\sum_{s=1}^t\left[2^{\frac{2}{1-2\beta}} + \sum_{s'=1}^{s-1}\frac{\|g_s\|^2}{h_{s^2}}\right]$ and $\alpha_t = \epsilon/B_t^{\frac{1}{2}-\beta}$ for $\beta \in (0, \frac{1}{2})$, and then substitute $\rho = \frac{1}{2} - \beta$ to complete the result. Following the same arguments as Theorem 6.2.1, we have

$$R_T(u) \leq 2k\|u\|\left[\sqrt{V_{T+1}\log(\|u\|/\alpha_{T+1}+1)} \vee h_T\log(\|u\|/\alpha_{T+1}+1)\right] + 2\sum_{t=1}^T\frac{\alpha_t\|g_t\|^2}{\sqrt{V_t}}$$

[1] Note that $\lim_{\rho\to 0} B_t^\rho = 2$, so for $\rho = 0$ we allow an abuse of notation by letting $B_t^\rho := 2$ to avoid specifying separate cases.

where $V_t = 4h_t^2 + \|g\|^2_{1:t-1}$ and $h_t \geq \|g_t\|$ for all $t$.

Next, we follow the same argument as Lemma B.1.1. Define $\tau_1 = 1$ and for $t > 1$ define $\tau_t = \max\left\{t' : t' \leq t \text{ and } \sum_{s=1}^{t'-1} \frac{\|g_s\|^2}{h_s^2} + 2^{\frac{2}{1-2\beta}} < \frac{h_{t'}^2}{h_{\tau_{t'-1}}^2}\right\}$. Then, we partition $[1, T]$ into the disjoint intervals $[1, T] = \mathcal{I}_1 \cup \ldots \cup \mathcal{I}_N$ over which $\tau_t$ is fixed. Denote $\mathcal{I}_j = [\widetilde{\tau}_j, \widetilde{\tau}_{j+1} - 1]$ where $\widetilde{\tau}_1 = 1$, $\widetilde{\tau}_{N+1} = T + 1$, and $\widetilde{\tau}_j = \min\{t > \widetilde{\tau}_{j-1} : \tau_t > \tau_{t-1}\}$ for $j \in [2, N]$. Using the same argument as Lemma B.1.1, it holds that $B_{t+1} = 4\sum_{s=1}^{t+1}\left[2^{\frac{2}{1-2\beta}} + \sum_{s'=1}^{s-1} \frac{\|g_{s'}\|^2}{h_{s'}^2}\right] \geq \frac{V_{t+1}}{h^2_{\tau_{t+1}}}$, so plugging this in above we have

$$\begin{aligned}
\sum_{t=1}^{T} \frac{\alpha_t \|g_t\|^2}{\sqrt{V_t}} &= \epsilon \sum_{j=1}^{N} \sum_{t \in \mathcal{I}_j} \frac{\|g_t\|^2}{B_t^{\frac{1}{2}-\beta} \sqrt{V_t}} \\
&\leq \epsilon \sum_{j=1}^{N} h_{\widetilde{\tau}_j}^{1-2\beta} \sum_{t \in \mathcal{I}_j} \frac{\|g_t\|^2}{V_t^{1-\beta}} \\
&\overset{(*)}{\leq} \epsilon \sum_{j=1}^{N} h_{\widetilde{\tau}_j}^{1-2\beta} \frac{1}{\beta} (\sum_{t \in \mathcal{I}_j} \|g_t\|^2)^{\beta} \\
&\leq \frac{\epsilon}{\beta} (\|g\|^2_{1:T})^{\beta} \sum_{j=1}^{N} h_{\widetilde{\tau}_j}^{1-2\beta},
\end{aligned}$$

where $(*)$ bounds $\sum_{t \in \mathcal{I}_j} \frac{\|g_t\|^2}{V_t^{1-\beta}}$ using the same argument as Theorem 6.4.2. Then, since each interval begins when $h^2_{\widetilde{\tau}_j}/h^2_{\widetilde{\tau}_{j-1}} \geq \sum_{s=1}^{t-1} \frac{\|g_s\|^2}{h_s^2} + 2^{\frac{2}{1-2\beta}} \geq 2^{\frac{2}{1-2\beta}}$, we have $h_{\widetilde{\tau}_j}^{1-2\beta} \geq 2h_{\widetilde{\tau}_{j-1}}^{1-2\beta}$, so

$$\begin{aligned}
\sum_{t=1}^{T} \frac{\alpha_t \|g_t\|^2}{\sqrt{V_t}} &\leq \frac{\epsilon}{\beta} (\|g\|^2_{1:T})^{\beta} \sum_{j=1}^{N} h_{\widetilde{\tau}_j}^{1-2\beta} \\
&\leq \frac{\epsilon}{\beta} (\|g\|^2_{1:T})^{\beta} \sum_{j=0}^{N-1} h_{\widetilde{\tau}_N}^{1-2\beta} \frac{1}{2^j} \\
&\leq \frac{\epsilon}{\beta} h_T^{1-2\beta} (\|g\|^2_{1:T})^{\beta} \sum_{j=0}^{N-1} \frac{1}{2^j} \\
&\leq \frac{2\epsilon}{\beta} h_T^{1-2\beta} (\|g\|^2_{1:T})^{\beta}
\end{aligned}$$

Plugging this back in above and substituting $\beta = \frac{1}{2} - \rho$, we have

$$\begin{aligned}
R_T(u) &\leq 2k\|u\| \left[\sqrt{V_{T+1} \log(\|u\|/\alpha_{T+1} + 1)} \vee h_T \log(\|u\|/\alpha_{T+1} + 1)\right] + \frac{2\epsilon}{\beta} h_T^{1-2\beta} (\|g\|^2_{1:T})^{\beta} \\
&\leq O\left(\frac{\epsilon h_T^{2\rho}}{1-2\rho} V_{T+1}^{\frac{1}{2}-\rho} + \|u\| \left[\sqrt{V_{T+1} \log\left(\frac{\|u\| B_{T+1}^{\rho}}{\epsilon} + 1\right)} \vee h_T \log\left(\frac{\|u\| B_{T+1}^{\rho}}{\epsilon} + 1\right)\right]\right),
\end{aligned}$$

and $B_{T+1}^{\rho} = \left(4\sum_{s=1}^{T+1}\left[2^{\frac{1}{\rho}} + \sum_{s'=1}^{s-1} \frac{\|g_{s'}\|^2}{h_{s'}^2}\right]\right)^{\rho}$. □

### Optimistic Trade-offs in the Horizon

A result analogous to Theorem 6.4.2 can be shown for our optimistic algorithm as well, and is stated here for completeness. Formal proof is omitted since it follows the same argument as Theorem 6.4.2 with only superficial modification: following the same steps as Theorem 6.3.1, we have

$$\begin{aligned}R_T(u) &\le 2k\left\|u\right\|\left[\sqrt{\widehat{V}_{T+1}\log\left(\left\|u\right\|/\widehat{\alpha}_{T+1}+1\right)}\vee 2G\log\left(\left\|u\right\|/\widehat{\alpha}_{T+1}+1\right)\right]\\ &\quad+\sum_{t=1}^{T}\frac{2\widehat{\alpha}_t\left\|\nabla\ell_t(w_t)-\nabla\widehat{\ell}_t(w_t)\right\|^2}{\sqrt{\widehat{V}_t}},\end{aligned}$$

where $\widehat{V}_t = 16G^2 + \sum_{s=1}^{t-1}\left\|\nabla\ell_s(w_s)-\nabla\widehat{\ell}_s(w_s)\right\|^2$. Now follow the same arguments as Theorem 6.4.2 to prove the following result.

**Theorem B.1.2.** *Under the same assumptions as Theorem 6.1.1, let $\rho\in\left[0,\frac{1}{2}\right)$ and suppose we set $\alpha_t = \epsilon G^{2\rho}/\widehat{V}_t^{\rho}$ for all $t$. Then for all $u\in\mathbb{R}^d$, Algorithm 2 guarantees*

$$R_T(u)\le O\left(\frac{\epsilon G^{2\rho}}{1-2\rho}\widehat{V}_{T+1}^{\frac{1}{2}-\rho}+\left\|u\right\|\left[\sqrt{\widehat{V}_{T+1}\log\left(\frac{\left\|u\right\|\widehat{V}_{T+1}^{\rho}}{\epsilon G^{2\rho}}+1\right)}\vee G\log\left(\frac{\left\|u\right\|\widehat{V}_{T+1}^{\rho}}{\epsilon G^{2\rho}}+1\right)\right]\right),$$

*where $\widehat{V}_{T+1} = 16G^2+\sum_{t=1}^{T}\left\|\nabla\ell_t(w_t)-\nabla\widehat{\ell}_t(w_t)\right\|^2$*

## B.2 Details for Chapter 7

### B.2.1 Proofs for Section 7.1 (Online Learning with Quadratically Bounded Losses)

**Proof of Theorem 7.1.2**

**Theorem 7.1.2.** *Let $\mathcal{A}$ be an online learning algorithm and let $w_t \in W$ be its output on round $t$. Let $\{g_t\}$ be a $(G_t, L_t)$-quadratically bounded sequence w.r.t. $\{w_t\}$, where $G_t \in [0, G_{\max}]$ and $L_t \in [0, L_{\max}]$ for all $t$. Let $\epsilon > 0$, $k \ge 3$, $\kappa \ge 4$, $c \ge 4$, $V_{t+1} = cG_{\max}^2 + G_{1:t}^2$, $\rho_{t+1} = \frac{1}{\sqrt{L_{\max}^2 + L_{1:t}^2}}$, $\alpha_{t+1} = \frac{\sqrt{V_{t+1}}\log^2(V_{t+1}/G_{\max}^2)}{\epsilon G_{\max}}$, and set*

$$\psi_t(w) = k\int_0^{\|w\|} \min_{\eta \le \frac{1}{G_{\max}}} \left[\frac{\log(x/\alpha_t + 1)}{\eta} + \eta V_t\right] dx + \frac{\kappa \|w\|^2}{2\rho_t} \qquad \text{and} \qquad \varphi_t(w) = \frac{L_t^2}{2\sqrt{L_{1:t}^2}} \|w\|^2 .$$

*Then for any $u \in W$, Algorithm 6 guarantees*

$$R_T(u) \le 2\epsilon G_{\max} + \kappa \|u\|^2 \sqrt{L_{\max}^2 + L_{1:T}^2} + 2k\|u\| \max\left\{\sqrt{V_{T+1} F_{T+1}(\|u\|)}, G_{\max} F_{T+1}(\|u\|)\right\}$$

*where $F_{T+1}(\|u\|) = \log(\|u\|/\alpha_{T+1} + 1)$.*

*Proof.* We can assume without loss of generality that $\mathbf{0} \in W$, since we could otherwise just perform a coordinate translation. Hence, we have $w_1 = \arg\min_{w \in W} \psi_1(w) = \mathbf{0}$, and it is easily seen that for any $w \in W$ we'll have $D_{\psi_t}(w|w_1) = D_{\psi_t}(w|\mathbf{0}) = \psi_t(w)$.

First apply Lemma 4.1.2 with $\mathcal{M}_t(w) = w$ and $\varphi_t(w) = \frac{L_t^2}{2\sqrt{L_{1:t}^2}} \|w\|^2$ to get

$$\begin{aligned}\sum_{t=1}^T \langle g_t, w_t - u\rangle &\le D_{\psi_{T+1}}(u|w_1) + \varphi_{1:T}(u) + \sum_{t=1}^T \underbrace{\langle g_t + \nabla\varphi_t(w_t), w_t - w_{t+1}\rangle - D_{\psi_t}(w_{t+1}|w_t) - \Delta_t(w_{t+1}) - \varphi_t(w_t)}_{=:\delta_t} \\ &\le \psi_{T+1}(u) + \varphi_{1:T}(u) + \delta_{1:T}.\end{aligned}$$

Let us first bound the leading term $\psi_{T+1}(u)$. For brevity, denote $F_t(x) = \log(x/\alpha_t + 1)$ and let $\Psi_t(\|w\|) = \int_0^{\|w\|} \min_{\eta \le 1/G_{\max}} \left[\frac{F_t(x)}{\eta} + \eta V_t\right] dx$ and $\Phi_t(\|w\|) = \frac{\kappa}{2\rho_t}\|w\|^2$, so that $\psi_t(w) = \Psi_t(\|w\|) + \Phi_t(\|w\|)$. Then

$$\begin{aligned}\psi_{T+1}(u) &= k\int_0^{\|u\|} \Psi_{T+1}'(x) dx + \frac{\kappa}{2\rho_t}\|u\|^2 \\ &\le k\|u\| \Psi_{T+1}'(\|u\|) + \frac{\kappa}{2}\|u\|^2 \sqrt{L_{\max}^2 + L_{1:T}^2}.\end{aligned}$$

Moreover,

$$
\begin{aligned}
\Psi_t'(\|u\|) &= k \min_{\eta\leq 1/G_{\max}}\left[\frac{F_t(\|u\|)}{\eta}+\eta V_t\right]\\
&= \begin{cases} 2k\sqrt{V_tF_t(\|u\|)} & \text{if } G_{\max}\sqrt{F_t(\|u\|)}\leq\sqrt{V_t}\\ kG_{\max}F_t(\|u\|)+k\frac{V_t}{G_{\max}} & \text{otherwise}\end{cases}\\
&\overset{(*)}{\leq} \begin{cases} 2k\sqrt{V_tF_t(\|u\|)} & \text{if } G_{\max}\sqrt{F_t(\|u\|)}\leq\sqrt{V_t}\\ 2kG_{\max}F_t(\|u\|) & \text{otherwise}\end{cases}\\
&= 2k\max\left\{\sqrt{V_tF_t(\|u\|)}, G_{\max}F_t(\|u\|)\right\}.
\end{aligned}
$$

where $(*)$ observes that $V_t/G_{\max}\leq G_{\max}F_t(x)$ whenever $\Psi_t'(x)=kG_{\max}F_t(x)+kV_t/G_{\max}$. Next, using Lemma A.3.3 we have

$$
\varphi_{1:T}(u)=\frac{1}{2}\|u\|^2\sum_{t=1}^T\frac{L_t^2}{\sqrt{L_{1:t}^2}}\leq\|u\|^2\sqrt{L_{1:T}^2},
$$

so overall we have

$$
\sum_{t=1}^T\langle g_t, w_t-u\rangle\leq 2k\|u\|\max\left\{\sqrt{V_{T+1}F_{T+1}(\|u\|)}, G_{\max}F_{T+1}(\|u\|)\right\}+\frac{\kappa}{2}\|u\|^2\sqrt{L_{\max}^2+L_{1:T}^2}+\|u\|^2\sqrt{L_{1:T}^2}+\delta_{1:T} \tag{B.4}
$$

We conclude by bounding the stability terms $\delta_{1:T}$. Recall that

$$
\delta_t=\langle g_t+\nabla\varphi_t(w_t), w_t-w_{t+1}\rangle-D_{\psi_t}(w_{t+1}|w_t)-\Delta_t(w_{t+1})-\varphi_t(w_t),
$$

where $\Delta_t(w)=\psi_{t+1}(w)-\psi_t(w)$. We first separate into terms related to the $G_t$'s and terms related to the $L_t$'s:

$$
\begin{aligned}
\delta_t &\leq (\|g_t\|+\|\nabla\varphi_t(w_t)\|)\|w_t-w_{t+1}\|\\
&\quad -D_{\psi_t}(w_{t+1}|w_t)-\Delta_t(w_{t+1})-\varphi_t(w_t)\\
&\leq G_t\|w_t-w_{t+1}\|-D_{\Psi_t}(w_{t+1}|w_t)-\Delta_t(w_{t+1})\\
&\quad +2L_t\|w_t\|\|w_t-w_{t+1}\|-D_{\Phi_t}(w_{t+1}|w_t)-\varphi_t(w_t),
\end{aligned}
$$

where we slightly abuse notations $D_{\Psi_t}$ and $D_{\Phi_t}$ to denote the Bregman divergences *w.r.t.* the function $w\mapsto\Psi_t(\|w\|)$ and $w\mapsto\Phi_t(\|w\|)$. In the second line, observe that $\Phi_t(\|w\|)=\frac{\kappa}{2\rho_t}\|w\|^2$ is $\frac{\kappa}{\rho_t}$ strongly convex, so $D_{\Phi_t}(w_{t+1}|w_t)\geq\frac{\kappa}{2\rho_t}\|w_{t+1}-w_t\|^2$ and an application of Fenchel-Young inequality yields

$$\begin{aligned}
2L_t\left\|w_t\right\|\left\|w_t-w_{t+1}\right\|-D_{\Phi_t}(w_{t+1}|w_t)-\varphi_t(w_t) &\le 2L_t\left\|w_t\right\|\left\|w_t-w_{t+1}\right\|-\frac{\kappa}{2\rho_t}\left\|w_{t+1}-w_t\right\|^2-\varphi_t(w_t)\\
&\le \frac{4\rho_t L_t^2\left\|w_t\right\|^2}{2\kappa}-\varphi_t(w_t)\\
&= \frac{2L_t^2\left\|w_t\right\|^2}{\kappa\sqrt{L_{\max}+L_{1:t-1}^2}}-\frac{L_t^2}{2\sqrt{L_{1:t}^2}}\left\|w_t\right\|^2\\
&\le \frac{2L_t^2\left\|w_t\right\|^2}{\kappa\sqrt{L_{1:t}^2}}-\frac{L_t^2}{2\sqrt{L_{1:t}^2}}\left\|w_t\right\|^2\\
&\le 0
\end{aligned}$$

for $\kappa \ge 4$. Hence,

$$\delta_t \le G_t\left\|w_t-w_{t+1}\right\|-D_{\Psi_t}(w_{t+1}|w_t)-\Delta_t(w_{t+1}),$$

which we will bound by showing that $\Delta_t(w) \ge \eta_t(w)G_t^2$ for some suitable $G_t$-Lipschitz convex function $\eta_t$ and then invoking Lemma 4.0.2. To this end, observe that

$$\begin{aligned}
\Delta_t(w) &= \psi_{t+1}(w)-\psi_t(w)\\
&= \underbrace{\Psi_{t+1}(\|w\|)-\Psi_t(\|w\|)}_{=:\Delta_t^{\Psi}(w)}+\underbrace{\Phi_{t+1}(\|w\|)-\Phi_t(\|w\|)}_{=:\Delta_t^{\Phi}(w)}\\
&\ge \Delta_t^{\Psi}(w).
\end{aligned}$$

Moreover, writing $\Delta_t^{\Psi}(w)=\Psi_{t+1}(\|w\|)-\Psi_t(\|w\|)=\int_0^{\|w\|}\Psi_{t+1}'(x)-\Psi_t'(x)dx$, we have

$$\begin{aligned}
\Psi_{t+1}'(x)-\Psi_t'(x) &= k\min_{\eta\le 1/G_{\max}}\left[\frac{F_{t+1}(x)}{\eta}+\eta V_{t+1}\right]-k\min_{\eta\le 1/G_{\max}}\left[\frac{F_t(x)}{\eta}+\eta V_t\right]\\
&\ge k\min_{\eta\le 1/G_{\max}}\left[\frac{F_t(x)}{\eta}+\eta V_{t+1}\right]-k\min_{\eta\le 1/G_{\max}}\left[\frac{F_t(x)}{\eta}+\eta V_t\right]
\end{aligned}$$

and using the fact that for any $\eta \le 1/G_{\max}$, we can bound $\frac{F_t(x)}{\eta}+\eta V_t+\eta G_t^2 \ge \min_{\eta^*\le 1/G}\left[\frac{F_t(x)}{\eta^*}+\eta^* V_t\right]+\eta G_t^2$, we have

$$\begin{aligned}
&\ge k\min_{\eta\le 1/G_{\max}}\left[\frac{F_t(x)}{\eta}+\eta V_t\right]-k\min_{\eta\le 1/G_{\max}}\left[\frac{F_t(x)}{\eta}+\eta V_t\right]+kG_t^2\min\left\{\sqrt{\frac{F_t(x)}{V_{t+1}}},\frac{1}{G_{\max}}\right\}\\
&\ge kG_t^2\min\left\{\sqrt{\frac{F_t(x)}{2V_t}},\frac{1}{G_{\max}}\right\}\ge G_t^2\min\left\{\sqrt{\frac{F_t(x)}{V_t}},\frac{1}{G_{\max}}\right\},
\end{aligned}$$

where the last line observes that $\frac{1}{V_t} = \frac{1}{V_{t+1}}\frac{V_{t+1}}{V_t} = \frac{1}{V_{t+1}}\left(1 + \frac{G_t^2}{V_t}\right) \le \frac{2}{V_{t+1}}$ for $V_t \ge G_t^2$ and recalls $k \ge 3$. Defining $\eta_t(\|w\|) = \int_0^{\|w\|} \min\left\{\sqrt{\frac{F_t(x)}{V_t}}, \frac{1}{G_{\max}}\right\} dx$, we then immediately have:

$$\Delta_t^{\Psi}(\|w\|) \ge G_t^2 \int_0^{\|w\|} \min\left\{\sqrt{\frac{F_t(x)}{V_t}}, \frac{1}{G_{\max}}\right\} dx = \eta_t(\|w\|)G_t^2.$$

Hence:

$$\begin{aligned} \delta_t &\le G_t \|w_t - w_{t+1}\| - D_{\psi_t}(w_{t+1}|w_t) - \Delta_t(w_{t+1}) \\ &\le G_t \|w_t - w_{t+1}\| - D_{\psi_t}(w_t|w_{t+1}) - \eta_t(\|w_{t+1}\|)G_t^2 \end{aligned} \tag{B.5}$$

Finally, we conclude by showing that $\psi_t$ satisfies the assumptions of Lemma 4.0.2 *w.r.t.* this function $\eta_t$.

We can write

$$\begin{aligned} \Psi_t(x) &= k\int_0^x \min_{\eta \le 1/G_{\max}} \left[\frac{F_t(v)}{\eta} + \eta V_t\right] dv \\ &= k\int_0^x \max\left\{2\sqrt{V_t F_t(v)}, G_{\max}F_t(v) + \frac{V_t}{G_{\max}}\right\} dv \end{aligned}$$

and so for any $x > 0$ we have

$$\begin{aligned} \Psi_t'(x) &= \begin{cases} 2k\sqrt{V_t F_t(x)} & \text{if } G_{\max}\sqrt{F_t(x)} \le \sqrt{V_t} \\ kG_{\max}F_t(x) + \frac{kV_t}{G_{\max}} & \text{otherwise} \end{cases} \\ \Psi_t''(x) &= \begin{cases} \frac{k\sqrt{V_t}}{(x+\alpha_t)\sqrt{F_t(x)}} & \text{if } G_{\max}\sqrt{F_t(x)} \le \sqrt{V_t} \\ \frac{kG_{\max}}{x+\alpha_t} & \text{otherwise} \end{cases} \\ \Psi_t'''(x) &= \begin{cases} \frac{-k\sqrt{V_t}(1+2F_t(x))}{2(x+\alpha_t)^2F_t(x)^{3/2}} & \text{if } G_{\max}\sqrt{F_t(x)} \le \sqrt{V_t} \\ \frac{-kG_{\max}}{(x+\alpha_t)^2} & \text{otherwise} \end{cases}. \end{aligned}$$

Clearly, we have $\Psi_t(x) \ge 0$, $\Psi_t'(x) \ge 0$, $\Psi_t''(x) \ge 0$, and $\Psi_t'''(x) \le 0$ for all $x > 0$. Moreover, for any $x \ge \alpha_t(e-1) =: \mathring{x}_t$, we have

$$\begin{aligned} \frac{|\Psi_t'''(x)|}{\Psi_t''(x)^2} &= \begin{cases} \frac{k\sqrt{V_t}(1+2F_t(x))}{2(x+\alpha_t)^2F_t(x)^{3/2}}\frac{(x+\alpha_t)^2F_t(x)}{k^2V_t} & \text{if } G_{\max}\sqrt{F_t(x)} \le \sqrt{V_t} \\ \frac{kG_{\max}}{(x+\alpha_t)^2}\frac{(x+\alpha_t)^2}{k^2G_{\max}^2} & \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{1}{2k\sqrt{V_t}}\left(\frac{1}{\sqrt{F_t(x)}} + 2\sqrt{F_t(x)}\right) & \text{if } G_{\max}\sqrt{F_t(x)} \le \sqrt{V_t} \\ \frac{1}{kG_{\max}} & \text{otherwise} \end{cases} \end{aligned}$$

and since $x > \alpha_t(e-1)$, we have $F_t(x) > 1$ and hence $\frac{1}{\sqrt{F_t(x)}} \le \sqrt{F_t(x)}$:

$$
\begin{aligned}
&\le \begin{cases} \frac{3\sqrt{F_t(x)}}{2k\sqrt{V_t}} & \text{if } G_{\max}\sqrt{F_t(x)} \le \sqrt{V_t} \\ \frac{1}{kG_{\max}} & \text{otherwise} \end{cases} \\
&\le \frac{1}{2}\min\left\{\sqrt{\frac{F_t(x)}{V_t}}, \frac{1}{G_{\max}}\right\} = \frac{1}{2}\eta_t'(x),
\end{aligned}
$$

where the last line recalls $\eta_t(x) = \int_0^x \min\left\{\sqrt{\frac{F_t(v)}{V_t}}, \frac{1}{G_{\max}}\right\} dv$ and chooses $k \ge 3$. Further, observe that $\eta_t(x)$ is convex and $\eta_t'(x) \le \frac{1}{G_{\max}}$, hence $\frac{1}{G_{\max}}$-Lipschitz. Thus, $\Psi_t$ satisfies the conditions of Lemma 4.0.2 with $\eta_t(x) = \int_0^x \min\left\{\sqrt{\frac{F_t(x)}{V_t}}, \frac{1}{G_{\max}}\right\} dx$ and $\mathring{x}_t = \alpha_t(e-1)$, so summing Equation (B.5) over all $t$, we have

$$
\begin{aligned}
\sum_{t=1}^T \delta_t &\le \sum_{t=1}^T G_t \left\| w_t - w_{t+1} \right\| - D_{\Psi_t}(w_{t+1}|w_t) - \eta_t(\left\| w_{t+1} \right\|)G_t^2 \\
&\le \sum_{t=1}^T \frac{2G_t^2}{\Psi_t''(\mathring{x}_t)} = \sum_{t=1}^T \frac{2G_t^2}{k\sqrt{V_t}}(\left\| \mathring{x}_t \right\| + \alpha_t) \\
&\le \sum_{t=1}^T \frac{2e\alpha_t G_t^2}{k\sqrt{V_t}} \le \sum_{t=1}^T 2\frac{\alpha_t G_t^2}{\sqrt{V_t}}
\end{aligned}
$$

where the last line bounds $e/k \le 3/k \le 1$ for $k \ge 3$. Next, substitute $\alpha_t = \frac{\epsilon G_{\max}}{\sqrt{V_t}\log^2(V_t/G_{\max})}$ to bound

$$
\begin{aligned}
\sum_{t=1}^T \delta_t &\le 2\epsilon G_{\max} \sum_{t=1}^T \frac{G_t^2}{V_t \log^2\left(V_t/G_{\max}^2\right)} \\
&\le 2\epsilon G_{\max} \sum_{t=1}^T \frac{G_t^2}{\left((c-1)G_{\max}^2 + G_{1:t}^2\right)\log^2\left(\frac{(c-1)G_{\max}^2 + G_{1:t}^2}{G_{\max}^2}\right)} \\
&\le 2\epsilon G_{\max} \int_{(c-1)G_{\max}^2}^{(c-1)G_{\max}^2 + G_{1:T}^2} \frac{1}{x\log^2(x/G_{\max}^2)} dx \\
&= 2\epsilon G_{\max} \frac{1}{\log(x/G_{\max}^2)}\Bigg|_{(c-1)G_{\max}^2}^{(c-1)G_{\max}^2 + G_{1:T}^2} \\
&\le \frac{2\epsilon G_{\max}}{\log\left(c-1\right)} \le 2\epsilon G_{\max},
\end{aligned}
$$

for $c \geq 4$. Finally, plugging this back into Equation (B.4) yields

$$\begin{aligned}
\sum_{t=1}^{T} \langle g_t, w_t - u\rangle &\leq 2k \left\| u \right\| \max\left\{\sqrt{V_{T+1}F_{T+1}(\left\| u \right\|)}, G_{\max}F_{T+1}(\left\| u \right\|)\right\} \\
&\quad + \frac{\kappa}{2}\left\| u \right\|^2 \sqrt{L_{\max}^2 + L_{1:T}^2} + \left\| u \right\|^2 \sqrt{L_{1:T}^2} + \delta_{1:T} \\
&\leq 2k \left\| u \right\| \max\left\{\sqrt{V_{T+1}F_{T+1}(\left\| u \right\|)}, G_{\max}F_{T+1}(\left\| u \right\|)\right\} \\
&\quad + \frac{\kappa}{2}\left\| u \right\|^2 \sqrt{L_{\max}^2 + L_{1:T}^2} + \left\| u \right\|^2 \sqrt{L_{1:T}^2} + 2\epsilon G_{\max} \\
&\leq 2\epsilon G_{\max} + \kappa \left\| u \right\|^2 \sqrt{L_{\max}^2 + L_{1:T}^2} \\
&\quad 2k \left\| u \right\| \max\left\{\sqrt{V_{T+1}F_{T+1}(\left\| u \right\|)}, G_{\max}F_{T+1}(\left\| u \right\|)\right\}
\end{aligned}$$

□

### Proof of Theorem 7.1.3

**Theorem 7.1.3.** *Let $\mathcal{A}$ be an algorithm defined over $\mathbb{R}^2$ and let $w_t$ denote the output of $\mathcal{A}$ on round $t$. Let $\epsilon > 0$ and suppose $\mathcal{A}$ guarantees $R_T(\mathbf{0}) \leq \epsilon$ against any quadratically bounded sequence $\{g_t\}$. Then for any $T \geq 1$, $G > 0$ and $L \geq 0$ there exists a sequence $g_1, \ldots, g_T$ satisfying $\left\| g_t \right\| \leq G + L\left\| w_t \right\|$ and a comparator $u \in \mathbb{R}^2$ such that*

$$R_T(u) \geq \Omega\left(G\left\| u \right\| \sqrt{T \log\left(\left\| u \right\| \sqrt{T}/\epsilon\right)} \vee L\left\| u \right\|^2 \sqrt{T}\right).$$

*Proof.* Let $w_t \in \mathbb{R}^2$ be the output of algorithm $\mathcal{A}$ at time $t$. Consider sequences $g_1, \ldots, g_T$ where $g_t \in \left\{\begin{pmatrix} -G \\ L\left\| w_t \right\| \end{pmatrix}, \begin{pmatrix} -G \\ -L\left\| w_t \right\| \end{pmatrix}\right\}$, and define the randomized sequence $\widetilde{g}_t = \begin{pmatrix} -G \\ -\varepsilon_t L\left\| w_t \right\| \end{pmatrix}$ where $\varepsilon_t$ are independent random signs. Consider the worst-case regret against a comparator constrained to an

$\ell_\infty$ ball of radius $U$:

$$\begin{aligned}
\sup_{g_1,\ldots,g_T} R_T &= \sup_{g_1,\ldots,g_T} \sum_{t=1}^{T} \langle g_t, w_t\rangle - \min_{u:\|u\|_\infty \le U} \sum_{t=1}^{T} \langle g_t, u\rangle \\
&\ge \mathbb{E}_{\varepsilon_1,\ldots,\varepsilon_T}\left[\sum_{t=1}^{T} \langle \widetilde{g}_t, w_t\rangle - \min_{u:\|u\|_\infty \le U} \sum_{t=1}^{T} \langle \widetilde{g}_t, u\rangle\right] \\
&\ge \mathbb{E}_{\varepsilon_1,\ldots,\varepsilon_T}\left[-\sum_{t=1}^{T} G\|w_t\| - \min_{u:\|u\|_\infty \le U} \sum_{t=1}^{T} -Gu_1 - u_2\varepsilon_t L\|w_t\|\right] \\
&= \mathbb{E}_{\varepsilon_1,\ldots,\varepsilon_T}\left[-G\sum_{t=1}^{T} \|w_t\| + GTU + \max_{|u_2|\le U} u_2 L \sum_{t=1}^{T} \varepsilon_t \|w_t\|\right] \\
&= \mathbb{E}_{\varepsilon_1,\ldots,\varepsilon_T}\left[GTU + UL\left|\sum_{t=1}^{T} \varepsilon_t \|w_t\|\right| - G\sum_{t=1}^{T} \|w_t\|\right] \\
&\overset{(a)}{\ge} \mathbb{E}_{\varepsilon_1,\ldots,\varepsilon_T}\left[GTU + \frac{UL}{\sqrt{2}}\sqrt{\sum_{t=1}^{T} \|w_t\|^2} - G\sum_{t=1}^{T} \|w_t\|\right] \\
&\overset{(b)}{\ge} \mathbb{E}_{\varepsilon_1,\ldots,\varepsilon_T}\left[GTU + \frac{UL}{\sqrt{2}}\sqrt{\sum_{t=1}^{T} \|w_t\|^2} - G\sqrt{T\sum_{t=1}^{T} \|w_t\|^2}\right]
\end{aligned}$$

where $(a)$ applies Khintchine inequality, $(b)$ applies Cauchy-Schwarz inequality, and choosing $U = \frac{G}{L}\sqrt{2T}$ we have

$$= GTU = \frac{L}{\sqrt{2}} U^2 \sqrt{T} = \frac{L\|u\|^2\sqrt{T}}{2\sqrt{2}},$$

where the final equality bounds $\|u\|^2 = u_1^2 + u_2^2 \le 2U^2$. Hence, there exists a sequence of $g_t$ which incurs at least $\Omega(L\|u\|^2\sqrt{T})$ regret. Moreover, for any algorithm which guarantees $R_T(\mathbf{0}) \le \epsilon$, there exists a sequence $g_1, \ldots, g_T$ with $\|g_t\| \le G$ for all $t$ such that for any $T$ and $u$, $R_T(u) \ge \frac{G}{3\sqrt{2}}\|u\|\sqrt{T\log\left(\|u\|\sqrt{T}/\sqrt{2}\epsilon\right)}$ (Mcmahan and M. Streeter 2012, Theorem 8). Thus, taking the worst of these two sequences yields

$$\sup_{g_1,\ldots,g_T} R_T \ge \max\left\{\frac{G}{3\sqrt{2}}\|u\|\sqrt{T\log\left(\|u\|\sqrt{T}/\sqrt{2}\epsilon\right)}, \frac{L\|u\|^2\sqrt{T}}{2\sqrt{2}}\right\}$$

□

### B.2.2 Multi-scale Experts Algorithm

For completeness, in this section we provide a multi-scale experts algorithm which achieves the bound required for our dynamic regret algorithm in Section 9.2. Our approach is inspired by the Multi-scale Multiplicative-weight with Correction (MsMwC) algorithm of L. Chen, Luo, and

**Algorithm 14:** Multi-scale Fixed-share

```
Input: $p_1 \in \Delta_N \cap (0,1]^N$, $\mu_1, \dots, \mu_N$ in $\mathbb{R}_{>0}$, $k > 0$, weights $\beta_1, \dots, \beta_T$ in $[0,1]$
Initialize: $q_1 = p_1$
Define $\psi_i(x) = \frac{k}{\mu_i}\int_0^x \log(v)\,dv$ for $i \in [N]$
for $t = 1 : T$ do
    Play $p_t \in \Delta_N$, receive loss $\widetilde{\ell}_t \in \mathbb{R}^N$
    Update $q_{t+1} = \arg\min_{q \in \Delta_N} \sum_{i=1}^N (\widetilde{\ell}_{ti} + \mu_i \widetilde{\ell}_{ti}^2) q_i + D_{\psi_i}(q_i|p_{ti})$
    Set $p_{t+1} = (1-\beta_t) q_{t+1} + \beta_t p_1$
end
```

Wei (2021), but formulated as a fixed-share update instead of an update on a "clipped" simplex $\widetilde{\Delta}_N = \Delta_N \cap [\beta, 1]^N$. The MsMwC algorithm provides a guarantee analogous to the following theorem, but formulating it as a fixed-share update will allow us a bit more modularity when constructing our dynamic regret algorithm in Appendix C.1.2, which requires several rather delicate conditions to come together in the right way.

**Theorem B.2.1.** *Let $k \ge \frac{9}{2}$ and assume $\mu_1, \dots, \mu_N$ satisfy $\mu_i \widetilde{\ell}_{ti} \le 1$ for all $t \in [T]$ and $i \in [N]$. Then for any $u \in \Delta_N$, Algorithm 14 guarantees*

$$\sum_{t=1}^T \left\langle \widetilde{\ell}_t, p_t - u \right\rangle \le \sum_{i=1}^N u_i \left[ \frac{k\left[\log(u_i/p_{1i}) + \sum_{t=1}^T \log\left(\frac{1}{1-\beta_t}\right)\right]}{\mu_i} + \mu_i \sum_{t=1}^T \widetilde{\ell}_{ti}^2 \right] + k(1+\beta_{1:T}) \sum_{i=1}^N \frac{p_{1i}}{\mu_i}.$$

*Moreover, for $\beta_t \le 1 - \exp\left(-\frac{1}{T}\right)$,*

$$\sum_{t=1}^T \left\langle \widetilde{\ell}_t, p_t - u \right\rangle \le \sum_{i=1}^N u_i \left[ \frac{k\left[\log(u_i/p_{1i}) + 1\right]}{\mu_i} + \mu_i \sum_{t=1}^T \widetilde{\ell}_{ti}^2 \right] + 2k \sum_{i=1}^N \frac{p_{1i}}{\mu_i}$$

*Proof.* The described algorithm is an instance of Algorithm 2 applied to the simplex $\Delta_N$ with $\varphi_t(p) = \sum_{i=1}^N \mu_i \widetilde{\ell}_{ti}^2 p_i$, and $\mathcal{M}_{t+1}(p) = (1-\beta_t)p + \beta_t p_1$. Applying Lemma 4.1.2:

$$\sum_{t=1}^T \left\langle \widetilde{\ell}_t, p_t - u \right\rangle \le D_\psi(u|p_1) - D_\psi(u|p_{T+1}) + \varphi_{1:T}(u) + \xi_{1:T} + \delta_{1:T},$$

where

$$\begin{aligned} \xi_t &= D_\psi(u|p_{t+1}) - D_\psi(u|q_{t+1}) \\ \delta_t &= \left\langle \widetilde{\ell}_t + \nabla \varphi_t(p_t), p_t - q_{t+1} \right\rangle - D_\psi(q_{t+1}|p_t) - \varphi_t(p_t). \end{aligned}$$

Observe that for any $u$, $p$, and $q$ in $\Delta_N$ we can write

$$
\begin{aligned}
D_\psi(u|p) - D_\psi(u|q) &= \sum_{i=1}^{N} \frac{k}{\mu_i} \left[ u_i \log\left(u_i/p_i\right) - u_i + p_i \right] - \sum_{i=1}^{N} \frac{k}{\mu_i} \left[ u_i \log\left(u_i/q_i\right) - u_i + q_i \right] \\
&= \sum_{i=1}^{N} \frac{k}{\mu_i} \left[ u_i \log\left(q_i/p_i\right) + p_i - q_i \right],
\end{aligned}
$$

so we have

$$
\begin{aligned}
D_\psi(u|p_1) - D_\psi(u|p_{T+1}) &= k \sum_{i=1}^{N} \frac{u_i \log\left(p_{T+1,i}/p_{1i}\right) + p_{1i} - p_{T+1,i}}{\mu_i} \\
&\leq k \sum_{i=1}^{N} \sup_{p \geq 0} \frac{u_i \log\left(p/p_{1i}\right) + p_{1i} - p}{\mu_i} \\
&= k \sum_{i=1}^{N} \frac{u_i \log\left(u_i/p_{1i}\right) + p_{1i} - u_i}{\mu_i} \\
&\leq k \sum_{i=1}^{N} \frac{u_i \log\left(u_i/p_{1i}\right)}{\mu_i} + k \sum_{i=1}^{N} \frac{p_{1i}}{\mu_i}
\end{aligned}
$$

and

$$
\begin{aligned}
\sum_{t=1}^{T} \xi_t &= \sum_{t=1}^{T} D_\psi(u|p_{t+1}) - D_\psi(u|q_{t+1}) \\
&= k \sum_{t=1}^{T} \sum_{i=1}^{N} \frac{u_i \log\left(q_{t+1,i}/p_{t+1,i}\right) + p_{t+1,i} - q_{t+1,i}}{\mu_i} \\
&= k \sum_{t=1}^{T} \sum_{i=1}^{N} \frac{u_i \log\left(\frac{q_{t+1,i}}{(1-\beta_t) q_{t+1,i} + \beta_t q_{1,i}}\right)}{\mu_i} \\
&\quad + k \sum_{t=1}^{T} \sum_{i=1}^{N} \frac{(1-\beta_t) q_{t+1,i} + \beta_t q_{1,i} - q_{t+1,i}}{\mu_i} \\
&= k \sum_{t=1}^{T} \sum_{i=1}^{N} \frac{u_i}{\mu_i} \log\left(\frac{q_{t+1,i}}{(1-\beta_t) q_{t+1,i} + \beta_t q_{1,i}}\right) \\
&\quad + k \sum_{t=1}^{T} \sum_{i=1}^{N} \frac{\beta_t (q_{1,i} - q_{t+1,i})}{\mu_i} \\
&\leq k \sum_{t=1}^{T} \sum_{i=1}^{N} \frac{u_i}{\mu_i} \log\left(\frac{1}{1-\beta_t}\right) + \frac{\beta_t q_{1i}}{\mu_i} \\
&= k \sum_{i=1}^{N} \frac{u_i}{\mu_i} \sum_{t=1}^{T} \log\left(\frac{1}{1-\beta_t}\right) + k \beta_{1:T} \sum_{i=1}^{N} \frac{p_{1i}}{\mu_i},
\end{aligned}
$$

where the last line recalls $p_1 = q_1$. Plugging these bounds back into the above regret bound yields

$$\begin{aligned}\sum_{t=1}^{T}\left\langle\widetilde{\ell}_t, p_t - u\right\rangle &\leq k\sum_{i=1}^{N}\left[\frac{u_i\log\left(u_i/p_{1i}\right)}{\mu_i} + (1+\beta_{1:T})\frac{p_{1i}}{\mu_i} + \frac{u_i}{\mu_i}\sum_{t=1}^{T}\log\left(\frac{1}{1-\beta_t}\right)\right] + \varphi_{1:T}(u) + \delta_{1:T}\\ &= \sum_{i=1}^{N}u_i\left[\frac{k\left[\log\left(u_i/p_{1i}\right) + \sum_{t=1}^{T}\log\left(\frac{1}{1-\beta_t}\right)\right]}{\mu_i} + \mu_i\sum_{t=1}^{T}\widetilde{\ell}_{ti}^2\right] + k(1+\beta_{1:T})\sum_{i=1}^{N}\frac{p_{1i}}{\mu_i}\\ &\quad + \sum_{i=1}^{N}\sum_{t=1}^{T}\underbrace{(\widetilde{\ell}_{ti} + \mu_i\widetilde{\ell}_{ti}^2)(p_{ti} - q_{t+1,i}) - D_{\psi_i}(q_{t+1,i}|p_{t,i}) - \mu_i\widetilde{\ell}_{ti}^2 p_{ti}}_{=:\delta_{ti}},\end{aligned} \tag{B.6}$$

where the last line recalls $\delta_t = \left\langle\widetilde{\ell}_t + \nabla\varphi_t(p_t), p_t - q_t\right\rangle - D_\psi(q_{t+1}|p_t) - \varphi_t(p_t)$, $\varphi_t(p) = \sum_{i=1}^{N}\mu_i\widetilde{\ell}_{ti}^2 p_i$, and denotes $\psi_i(p) = \frac{k}{\mu_i}\int_0^p \log(x)\,dx$ so that $\psi(p) = \sum_{i=1}^{N}\psi_i(p_i)$. We next focus our attention on the terms in the last line, $\delta_{ti}$.

Note that by construction, we have $p_{ti} = (1-\beta_t)q_{ti} + \beta_t q_{1i} \geq \beta_t q_{1i} > 0$ for all $i$. Thus, $\psi_i(p) = \frac{k}{\mu_i}\int_0^p \log(v)\,dv$ is twice differentiable everywhere on the line connecting $p_{ti}$ and $q_{t+1,i}$ for any $i$ with $q_{t+1,i} > 0$. For any such $i$, we have via Taylor's theorem that there exists a $\widetilde{p}_i$ on the line connecting $p_{ti}$ and $q_{t+1,i}$ such that

$$D_{\psi_i}(q_{t+1,i}|p_{ti}) \geq \frac{1}{2}(p_{ti} - q_{t+1,i})^2\psi_i''(\widetilde{p}_i) = \frac{1}{2}\frac{(p_{ti} - q_{t+1,i})^2 k}{\mu_i\widetilde{p}_i}$$

so using this with the assumption that $\mu_i\left|\widetilde{\ell}_{ti}\right| \leq 1$, we have

$$\begin{aligned}\delta_{ti} &\leq \left|\widetilde{\ell}_{ti} + \mu_i\widetilde{\ell}_{ti}^2\right|\left|p_{ti} - q_{t+1,i}\right| - \frac{1}{2}\frac{(p_{ti} - q_{t+1,i})^2 k}{\mu_i\widetilde{p}_i} - \mu_i\widetilde{\ell}_{ti}^2 p_{ti}\\ &\leq 2\left|\widetilde{\ell}_{ti}\right|\left|p_{ti} - q_{t+1,i}\right| - \frac{1}{2}\frac{(p_{ti} - q_{t+1,i})^2 k}{\mu_i\widetilde{p}_i} - \mu_i\widetilde{\ell}_{ti}^2\widetilde{p}_i + \mu_i\widetilde{\ell}_{ti}^2\left|p_{ti} - \widetilde{p}_i\right|\\ &\overset{(a)}{\leq} 3\left|\widetilde{\ell}_{ti}\right|\left|p_{ti} - q_{t+1,i}\right| - \frac{1}{2}\frac{(p_{ti} - q_{t+1,i})^2 k}{\mu_i\widetilde{p}_i} - \mu_i\widetilde{\ell}_{ti}^2\widetilde{p}_i\\ &\leq \frac{9}{2k}\mu_i\left|\widetilde{\ell}_{ti}\right|^2\widetilde{p}_i - \mu_i\widetilde{\ell}_{ti}^2\widetilde{p}_i\\ &\overset{(b)}{\leq} 0,\end{aligned}$$

where $(a)$ uses $\left|\widetilde{p}_i - p_{ti}\right| \leq \left|q_{t+1,i} - p_{ti}\right|$ for any $\widetilde{p}_i$ on the line connecting $q_{t+1,i}$ and $p_{ti}$ and $(b)$ chooses

$k \geq \frac{9}{2}$. Similarly, for any $i$ for which $q_{t+1,i} = 0$ we have

$$\begin{aligned}\delta_{ti} &= (\widetilde{\ell}_{ti} + \mu_i \widetilde{\ell}_{ti}^2)p_{ti} - D_{\psi_i}(0|p_{ti}) - \mu_i \widetilde{\ell}_{ti}^2 p_{ti} \\ &\leq \widetilde{\ell}_{ti} p_{ti} - \frac{p_{ti}}{\mu_i} \\ &\leq \frac{p_{ti}}{\mu_i} - \frac{p_{ti}}{\mu_i} \leq 0,\end{aligned}$$

where the last line again uses $\mu_i \left|\widetilde{\ell}_{ti}\right| \leq 1$. Thus, in either case we have $\delta_{ti} \leq 0$. Plugging this into Equation (B.6) reveals the first statement of the theorem:

$$\sum_{t=1}^{T} \left\langle \widetilde{\ell}_t, p_t - u \right\rangle \leq \sum_{i=1}^{N} u_i \left[ \frac{k\left[\log(u_i/p_{1i}) + \sum_{t=1}^{T} \log\left(\frac{1}{1-\beta_t}\right)\right]}{\mu_i} + \mu_i \sum_{t=1}^{T} \widetilde{\ell}_{ti}^2 \right] + k(1 + \beta_{1:T}) \sum_{i=1}^{N} \frac{p_{1i}}{\mu_i}.$$

For the second statement of the theorem, observe that $\beta_t \leq 1 - \exp(-1/T) \leq \frac{1}{T}$, so $\beta_{1:T} \leq 1$, and likewise $\log\left(\frac{1}{1-\beta_t}\right) = \log(\exp(1/T)) = \frac{1}{T}$, so $\sum_{t=1}^{T} \log\left(\frac{1}{1-\beta_t}\right) \leq 1$. Hence, the previous display is bounded as

$$\sum_{t=1}^{T} \left\langle \widetilde{\ell}_t, p_t - u \right\rangle \leq \sum_{i=1}^{N} u_i \left[ \frac{k\left[\log(u_i/p_{1i}) + 1\right]}{\mu_i} + \mu_i \sum_{t=1}^{T} \widetilde{\ell}_{ti}^2 \right] + 2k \sum_{i=1}^{N} \frac{p_{1i}}{\mu_i}$$

□

# Appendix C

# Part III (Adapting to Non-stationarity)

## C.1 Details for Chapter 9

### C.1.1 Proofs for Section 9.1 (Lipschitz Losses)

**Proof of Proposition 9.1.1**

We break the proof of Proposition 9.1.1 into parts; we first derive a partial result in Proposition C.1.1, and then make particular choices for the unspecified parameters $\alpha_t$ and $b_t$.

**Proposition C.1.1.** *$(\alpha_t)_{t=1}^T$ be a non-increasing sequence and consider Algorithm 2 with*

$$\begin{aligned}\psi_t(w) &= 2\int_0^{\|w\|}\frac{\log\left(x/\alpha_t+1\right)}{\eta}dx\\ \varphi_t(w) &= \left(\eta\left\|g_t\right\|^2+b_t\right)\|w\|,\end{aligned}$$

*where $b_t \geq 0$ and $\eta \leq \frac{1}{G}$. Then for all $u_1,\ldots,u_T$ in $\mathbb{R}^d$, Algorithm 2 guarantees*

$$\begin{aligned}R_T(\boldsymbol{u}) \leq \frac{2M\log\left(M/\alpha_{T+1}+1\right)}{\eta} &+ \sum_{t=1}^{T-1}\left[\frac{2\left\|u_{t+1}-u_t\right\|\log\left(\left\|w_{t+1}\right\|/\alpha_{t+1}+1\right)}{\eta} - b_t\left\|w_{t+1}\right\|\right]\\ &+\sum_{t=1}^{T}\left(\eta\left\|g_t\right\|^2+b_t\right)\left\|u_t\right\|+\eta\sum_{t=1}^{T}\alpha_t\left\|g_t\right\|^2,\end{aligned}$$

*where $M = \max_t \|u_t\|$.*

*Proof.* Using Lemma 4.0.1 we have

$$\begin{aligned}
R_T(\boldsymbol{u}) &\le \psi_{T+1}(u_T) + \sum_{t=1}^{T-1} \rho_t + \sum_{t=1}^{T} \varphi_t(u_t) + \sum_{t=1}^{T} \delta_t \\
&\le \frac{2\left\|u_T\right\| \log\left(\left\|u_T\right\|/\alpha_{T+1} + 1\right)}{\eta} + \sum_{t=1}^{T-1} \rho_t + \sum_{t=1}^{T} \varphi_t(u_t) + \sum_{t=1}^{T} \delta_t \\
&\le \frac{2M \log\left(M/\alpha_{T+1} + 1\right)}{\eta} + \sum_{t=1}^{T-1} \rho_t + \sum_{t=1}^{T} \varphi_t(u_t) + \sum_{t=1}^{T} \delta_t
\end{aligned}$$

where $M = \max_{t\le T}\left\|u_t\right\|$ and

$$\begin{aligned}
\sum_{t=1}^{T-1} \rho_t &= \sum_{t=1}^{T-1} \langle \nabla\psi_{t+1}(w_{t+1}), u_t - u_{t+1}\rangle \le \sum_{t=1}^{T-1} \left\|\nabla\psi_{t+1}(w_{t+1})\right\| \left\|u_t - u_{t+1}\right\| \\
&= 2\sum_{t=1}^{T-1} \frac{\log\left(\left\|w_{t+1}\right\|/\alpha_{t+1} + 1\right)}{\eta} \left\|u_t - u_{t+1}\right\| \\
\sum_{t=1}^{T} \delta_t &= \sum_{t=1}^{T} \langle g_t, w_t - w_{t+1}\rangle - D_{\psi_t}(w_{t+1}|w_t) - \phi_t(w_{t+1}) \\
&= \sum_{t=1}^{T} \langle g_t, w_t - w_{t+1}\rangle - D_{\psi_t}(w_{t+1}|w_t) - \Delta_t(w_{t+1}) - \varphi_t(w_{t+1})
\end{aligned}$$

First consider the terms $\sum_{t=1}^{T} \delta_t$. Since $(\alpha_t)_{t=1}^{T}$ is a non-increasing sequence, we have $\Delta_t(w_{t+1}) = \psi_{t+1}(w_{t+1}) - \psi_t(w_{t+1}) \ge 0$ and

$$\begin{aligned}
\delta_t &= \langle g_t, w_t - w_{t+1}\rangle - D_{\psi_t}(w_{t+1}|w_t) - \Delta_t(w_{t+1}) - \varphi_t(w_{t+1}) \\
&\le \langle g_t, w_t - w_{t+1}\rangle - D_{\psi_t}(w_{t+1}|w_t) - \varphi_t(w_{t+1}) \\
&= \langle g_t, w_t - w_{t+1}\rangle - D_{\psi_t}(w_{t+1}|w_t) - \eta\left\|g_t\right\|^2\left\|w_{t+1}\right\| - b_t\left\|w_{t+1}\right\|.
\end{aligned}$$

We proceed by showing that the regularizers $\psi_t(\cdot)$ satisfy the conditions of Lemma 4.0.2. we have $\psi_t(w) = \Psi_t(\left\|w\right\|) = 2\int_0^{\left\|w\right\|} \frac{\log(x/\alpha_t+1)}{\eta}dx$ and

$$\Psi_t'(x) = 2\frac{\log\left(x/\alpha_t + 1\right)}{\eta}, \qquad \Psi_t''(x) = \frac{2}{\eta\left(x + \alpha_t\right)}, \qquad \Psi_t'''(x) = \frac{-2}{\eta\left(x + \alpha_t\right)^2},$$

so $\Psi_t(x) \ge 0$, $\Psi_t'(x) \ge 0$, $\Psi_t''(x) \ge 0$, and $\Psi_t'''(x) \le 0$ for all $x > 0$. Moreover,

$$\frac{-\Psi_t'''(x)}{\Psi_t''(x)^2} = \frac{2}{\eta(x+\alpha_t)^2}\frac{\eta^2(x+\alpha_t)^2}{2^2} = \frac{\eta}{2},$$

so assuming $\eta \le \frac{1}{G}$ and letting $\eta_t(\left\|w\right\|) = \eta\left\|w\right\|$, we have $|\Psi_t'''(x)| \le \frac{\eta_t'(x)}{2}\Psi_t''(x)^2$ for all $x > 0$, and

$\eta_t(x)$ is a $1/G$ Lipschitz convex function. Hence, using Lemma 4.0.2 we have

$$\begin{aligned}
\delta_t &\le \langle g_t, w_t - w_{t+1}\rangle - D_{\psi_t}(w_{t+1}|w_t) - \eta \|g_t\|^2 \|w_{t+1}\| - b_t \|w_{t+1}\| \\
&= \langle g_t, w_t - w_{t+1}\rangle - D_{\psi_t}(w_{t+1}|w_t) - \eta_t(\|w_{t+1}\|) \|g_t\|^2 - b_t \|w_{t+1}\| \\
&\le \frac{2\|g_t\|^2}{\Psi_t''(0)} - b_t \|w_{t+1}\| = \eta\alpha_t \|g_t\|^2 - b_t \|w_{t+1}\| .
\end{aligned}$$

Plugging this back into the full regret bound we have

$$\begin{aligned}
R_T(\boldsymbol{u}) &\le \frac{2M\log\left(M/\alpha_{T+1}+1\right)}{\eta} + 2\sum_{t=1}^{T-1} \frac{\|u_t - u_{t+1}\| \log\left(\|w_{t+1}\|/\alpha_{t+1}+1\right)}{\eta} + \sum_{t=1}^{T} \varphi_t(u_t) \\
&\quad + \sum_{t=1}^{T} \eta\alpha_t \|g_t\|^2 - b_t \|w_{t+1}\| \\
&= \frac{2M\log\left(M/\alpha_{T+1}+1\right)}{\eta} + \sum_{t=1}^{T-1} \left[\frac{2\|u_{t+1} - u_t\| \log\left(\|w_{t+1}\|/\alpha_{t+1}+1\right)}{\eta} - b_t \|w_{t+1}\|\right] \\
&\quad + \sum_{t=1}^{T} \left(\eta\|g_t\|^2 + b_t\right)\|u_t\| + \eta\sum_{t=1}^{T} \alpha_t \|g_t\|^2 .
\end{aligned}$$

□

With this result in hand, we prove Proposition 9.1.1 by choosing values $\alpha_t = \frac{\epsilon G^2}{V_t \log^2(V_t/G^2)}$ and $b_t = \eta\|g_t\|^2$. The full version of the result is given below.

**Proposition 9.1.1.** *Let $\ell_1, \dots, \ell_T$ be G-Lipschitz convex functions and $g_t \in \partial\ell_t(w_t)$ for all $t$. Let $\epsilon > 0$, $V_t = 4G^2 + \|g\|_{1:t-1}^2$, and $\alpha_t = \frac{\epsilon G^2}{V_t \log^2(V_t/G^2)}$. For all $t$, set $\psi_t(w) = 2\int_0^{\|w\|} \frac{\log(x/\alpha_t+1)}{\eta} dx$, and $\varphi_t(w) = 2\eta\|g_t\|^2 \|w\|$. Then after each round Algorithm 2 updates*

$$\begin{aligned}
\theta_t &= \nabla\psi_t(w_t) - g_t \\
w_{t+1} &= \frac{\alpha_{t+1}\theta_t}{\|\theta_t\|}\left[\exp\left[\frac{\eta}{2}\max\left(\|\theta_t\| - 2\eta\|g_t\|^2, 0\right)\right] - 1\right]
\end{aligned}$$

*where we we define $C\frac{x}{\|x\|} = \mathbf{0}$ for all $C$ when $x = \mathbf{0}$. Moreover, for any $u_1, \dots, u_T$ in $\mathbb{R}^d$, Algorithm 2 guarantees*

$$R_T(\boldsymbol{u}) \le 2\epsilon G + \frac{4\left(M + P_T\right)\left[\log\left(\frac{9MT^2}{4\alpha_{T+1}} + 1\right) \vee 1\right]}{\eta} + 2\eta\sum_{t=1}^{T} \|g_t\|^2 \|u_t\| .$$

*where $M = \max_t \|u_t\|$.*

*Proof.* First, we will verify the update equation, and then show the regret bound. To compute the update, observe that from the first-order optimality conditions, there is some $\nabla\phi_t(w_{t+1}) \in \partial\phi_t(w_{t+1})$ such that

$$g_t + \nabla\psi_t(w_{t+1}) - \nabla\psi_t(w_t) + \nabla\phi_t(w_{t+1}) = \mathbf{0}$$

Now, notice that we can write $\nabla\phi_t(w_{t+1}) = \nabla\psi_{t+1}(w_{t+1}) - \nabla\psi_t(w_{t+1}) + \nabla\varphi_t(w_{t+1})$ for some $\nabla\varphi_t(w_{t+1}) \in \partial\varphi_t(w_{t+1})$. Thus, we have:

$$g_t + \nabla\psi_{t+1}(w_{t+1}) - \nabla\psi_t(w_t) + \nabla\varphi_t(w_{t+1}) = \mathbf{0}$$

Moreover, *any* value for $w_{t+1}$ such that there is a $\varphi_t(w_{t+1}) \in \partial\varphi_t(w_{t+1})$ satisfying the above condition is valid solution to the mirror descent update. We justify our update equation in two cases.

First, consider the case $\max(\|\theta_t\| - 2\eta\|g_t\|, 0) = 0$, In this case, the update equation suggests $w_{t+1} = \mathbf{0}$. To justify this, notice that $\partial\varphi_t(\mathbf{0})$ consists of all vectors of norm at most $2\eta\|g_t\|^2$. Further, $\nabla\psi_{t+1}(\mathbf{0}) = \mathbf{0}$. Thus, whenever $\max(\|\theta_t\| - 2\eta\|g_t\|, 0) = 0$, we can set $w_{t+1} = \mathbf{0}$ as described by our update.

Now, let us suppose $\max(\|\theta_t\| - 2\eta\|g_t\|, 0) = \|\theta_t\| - 2\eta\|g_t\| > 0$. Note that this implies $\theta_t \neq \mathbf{0}$, and the update equation sets $w_{t+1} \neq \mathbf{0}$. In the case $w_{t+1} \neq \mathbf{0}$, $\varphi_t(w_{t+1})$ is differentiable so that $\varphi_t(w_{t+1}) = 2\eta\|g_t\|^2 \frac{w_{t+1}}{\|w_{t+1}\|}$. Thus, we need to establish that indeed a non-zero $w_{t+1}$ given by the update equation is a solution to the optimality condition:

$$g_t + \nabla\psi_{t+1}(w_{t+1}) - \nabla\psi_t(w_t) + 2\eta\|g_t\|^2 \frac{w_{t+1}}{\|w_{t+1}\|} = \mathbf{0}.$$

Writing $\psi_t(w) = \Psi_t(\|w\|) = \int_0^{\|w\|} \Psi_t'(x)dx$, we have $\nabla\psi_{t+1}(w_{t+1}) = \frac{w_{t+1}}{\|w_{t+1}\|}\Psi_{t+1}'(\|w_{t+1}\|)$ (where we define $\frac{w_{t+1}}{\|w_{t+1}\|} \cdot 0 = \mathbf{0}$) and hence the optimality condition can be re-written:

$$\frac{w_{t+1}}{\|w_{t+1}\|}\left[\Psi_{t+1}'(\|w_{t+1}\|) + 2\eta\|g_t\|^2\right] = \nabla\psi_t(w_t) - g_t = \theta_t$$

Now we need only verify that our expression $w_{t+1} = \frac{\alpha_{t+1}\theta_t}{\|\theta_t\|}\left[\exp\left[\frac{\eta}{2}(\|\theta_t\| - 2\eta\|g_t\|)\right] - 1\right]$ satisfies this condition. Fortunately, this is easily checked by observing the stated update satisfies:

$$\Psi_{t+1}'(\|w_{t+1}\|) = \frac{2}{\eta}\log(\|w_{t+1}\|/\alpha_{t+1} + 1) = \|\theta_t\| - 2\eta\|g_t\|^2.$$

Turning now to the regret, we begin by replacing the comparator sequence with an auxiliary sequence $\widehat{u}_1, \ldots, \widehat{u}_T$ to be determined later. This alternative sequence will eventually be designed to have some useful stability properties while still being "close" to the real sequence $u_1, \dots, u_T$:

$$\begin{aligned} R_T(\boldsymbol{u}) &= \sum_{t=1}^{T} \langle g_t, w_t - u_t \rangle = \sum_{t=1}^{T} \langle g_t, w_t - \widehat{u}_t \rangle + \sum_{t=1}^{T} \langle g_t, \widehat{u}_t - u_t \rangle \\ &\leq R_T(\widehat{\boldsymbol{u}}) + \sum_{t=1}^{T} \|g_t\| \, \|\widehat{u}_t - u_t\| \end{aligned}$$

The first term is bounded via Proposition C.1.1 as

$$\begin{aligned} R_T(\widehat{\boldsymbol{u}}) \leq{}& \frac{2\widehat{M} \log\left(\widehat{M}/\alpha_{T+1} + 1\right)}{\eta} + 2\eta \sum_{t=1}^{T} \|g_t\|^2 \|\widehat{u}_t\| + \eta \sum_{t=1}^{T} \alpha_t \|g_t\|^2 \\ &+ \sum_{t=1}^{T-1} \left[ \frac{2 \|\widehat{u}_t - \widehat{u}_{t+1}\| \log\left(\|w_{t+1}\| / \alpha_{t+1} + 1\right)}{\eta} - \eta \|g_t\|^2 \|w_{t+1}\| \right] \end{aligned}$$

where $\widehat{M} = \max_{t \leq T} \|\widehat{u}_t\|$. We focus first on bounding the sum in the second line. To do so, we first provide the definition of $\widehat{u}_t$:

$$\text{Let } \mathcal{T} > 0 \text{ and set } \widehat{u}_T = u_T \text{ and } \widehat{u}_t = \begin{cases} u_t & \text{if } \|g_t\| \geq \mathcal{T} \\ \widehat{u}_{t+1} & \text{otherwise} \end{cases} \text{ for } t < T.$$

Hence, by definition we have $\|\widehat{u}_t - \widehat{u}_{t+1}\| = 0$ whenever $\|g_t\| \leq \mathcal{T}$, so

$$\begin{aligned} &\sum_{t=1}^{T-1} \left[ \frac{2 \|\widehat{u}_t - \widehat{u}_{t+1}\| \log\left(\|w_{t+1}\| / \alpha_{t+1} + 1\right)}{\eta} - \eta \|g_t\|^2 \|w_{t+1}\| \right] \\ &\quad \leq \sum_{t : \|g_t\| \geq \mathcal{T}} \left[ \frac{2 \|\widehat{u}_t - \widehat{u}_{t+1}\| \log\left(\|w_{t+1}\| / \alpha_{t+1} + 1\right)}{\eta} - \eta \mathcal{T}^2 \|w_{t+1}\| \right] \\ &\quad \leq \sum_{t : \|g_t\| \geq \mathcal{T}} \left[ \sup_{X \geq 0} \frac{2 \|\widehat{u}_t - \widehat{u}_{t+1}\| \log\left(X / \alpha_{t+1} + 1\right)}{\eta} - \eta \mathcal{T}^2 X \right] \\ &\quad \overset{(*)}{\leq} \sum_{t : \|g_t\| \geq \mathcal{T}} \frac{2 \|\widehat{u}_t - \widehat{u}_{t+1}\| \log\left(\frac{2\|\widehat{u}_t - \widehat{u}_{t+1}\|}{\alpha_{t+1} \eta^2 \mathcal{T}^2}\right)}{\eta} \end{aligned}$$

where $(*)$ observes that either the max is obtained at $X = 0$, for which $\sup_{X \geq 0} \frac{2\|\widehat{u}_t - \widehat{u}_{t+1}\| \log(X/\alpha_{t+1} + 1)}{\eta} - \eta \mathcal{T}^2 X = 0$, and otherwise the max is obtained at $X = \frac{2\|\widehat{u}_{t+1} - u_t\|}{\eta^2 \mathcal{T}^2} - \alpha_{t+1} > 0$, which leads to an upperbound of

$$\sup_{X \geq 0} \frac{2 \|\widehat{u}_t - \widehat{u}_{t+1}\| \log\left(X / \alpha_{t+1} + 1\right)}{\eta} - \eta \mathcal{T}^2 X \leq \frac{2 \|\widehat{u}_t - \widehat{u}_{t+1}\| \log\left(\frac{2\|\widehat{u}_{t+1} - \widehat{u}_t\|}{\alpha_{t+1} \eta^2 \mathcal{T}^2}\right)}{\eta}$$

in both cases. Moreover, for any $t$ such that $\|g_t\| \geq \mathcal{T}$ let $t'$ denote the smallest index greater than

$t$ for which $\|g_{t'}\| \geq \mathcal{T}$; then by triangle inequality we have $\|\widehat{u}_t - \widehat{u}_{t+1}\| = \|u_t - u_{t'}\| \leq \sum_{s=t}^{t'} \|u_s - u_{s+1}\|$ and

$$\begin{aligned}\sum_{t:\|g_t\|\geq\mathcal{T}} \frac{2\|\widehat{u}_t - \widehat{u}_{t+1}\| \log\left(\frac{2\|\widehat{u}_t - \widehat{u}_{t+1}\|}{\alpha_{t+1}\eta^2\mathcal{T}^2}\right)}{\eta} &\leq \sum_{t:\|g_t\|\geq\mathcal{T}} \frac{\sum_{s=t}^{t'} 2\|u_s - u_{s+1}\| \log\left(\frac{4\widehat{M}}{\alpha_{T+1}\eta^2\mathcal{T}^2}\right)}{\eta} \\ &= \frac{2P_T \log\left(\frac{4\widehat{M}}{\alpha_{T+1}\eta^2\mathcal{T}^2}\right)}{\eta}.\end{aligned}$$

Returning to the regret against the auxiliary comparator sequence we have

$$\begin{aligned}R_T(\widehat{\boldsymbol{u}}) &\leq \frac{2\widehat{M}\log\left(\widehat{M}/\alpha_{T+1}+1\right)}{\eta} + 2\eta\sum_{t=1}^T \|g_t\|^2 \|\widehat{u}_t\| + \eta\sum_{t=1}^T \alpha_t \|g_t\|^2 + \frac{2P_T\log\left(\frac{4\widehat{M}}{\alpha_{t+1}\eta^2\mathcal{T}^2}\right)}{\eta} \\ &\overset{(a)}{\leq} \frac{2M\log\left(M/\alpha_{T+1}+1\right) + 2P_T\log\left(\frac{4M}{\alpha_{t+1}\eta^2\mathcal{T}^2}\right)}{\eta} + 2\eta\sum_{t=1}^T \|g_t\|^2 \|\widehat{u}_t\| + \eta\sum_{t=1}^T \alpha_t \|g_t\|^2 \\ &\overset{(b)}{\leq} \frac{2M\log\left(M/\alpha_{T+1}+1\right) + 2P_T\log\left(\frac{4M}{\alpha_{t+1}\eta^2\mathcal{T}^2}\right)}{\eta} + \eta\sum_{t=1}^T \alpha_t \|g_t\|^2 \\ &\quad + 2\eta\sum_{t=1}^T \|g_t\|^2 \|u_t\| + 2\sum_{t=1}^T \|g_t\| \|\widehat{u}_t - u_t\| \\ &\overset{(c)}{\leq} \frac{2M\log\left(M/\alpha_{T+1}+1\right) + 2P_T\log\left(\frac{4M}{\alpha_{t+1}\eta^2\mathcal{T}^2}\right)}{\eta} + 2\epsilon G \\ &\quad + 2\eta\sum_{t=1}^T \|g_t\|^2 \|u_t\| + 2\sum_{t=1}^T \|g_t\| \|\widehat{u}_t - u_t\|,\end{aligned}$$

where $(a)$ observes that $\widehat{M} = \max_{t\leq T}\|\widehat{u}_t\| \leq \max_{t\leq T}\|u_t\| = M$ and $(b)$ recalls $\eta \leq \frac{1}{G}$ and uses $\eta\|g_t\|^2\|\widehat{u}_t\| \leq \eta\|g_t\|^2(\|u_t - \widehat{u}_t\| + \|u_t\|) \leq \eta\|g_t\|^2\|u_t\| + \|g_t\|\|u_t - \widehat{u}_t\|$, and $(c)$ chooses $\alpha_t = \frac{\epsilon G^2}{V_t\log^2(V_t/G^2)}$ for $V_t = 4G^2 + \|g\|_{1:t-1}^2$ and applies Lemma A.3.4 to bound

$$\begin{aligned}\eta\sum_{t=1}^T \alpha_t\|g_t\|^2 &= \eta\epsilon G^2\sum_{t=1}^T \frac{\|g_t\|^2}{V_t\log^2\left(V_t/G^2\right)} \\ &\leq 2\eta\epsilon G^2 \leq 2\epsilon G\end{aligned}$$

Returning now to the full regret bound and recalling $\widehat{u}_t = u_t$ whenever $\|g_t\| \geq \mathcal{T}$ and $\widehat{u}_t = \widehat{u}_{t+1}$

otherwise, we have

$$
\begin{aligned}
R_T(\boldsymbol{u}) &\leq R_T(\widehat{\boldsymbol{u}}) + \sum_{t=1}^{T} \|g_t\| \, \|\widehat{u}_t - u_t\| \\
&\leq 2\epsilon G + \frac{2M\log\left(M/\alpha_{T+1} + 1\right) + 2P_T \log\left(\frac{4M}{\alpha_{t+1}\eta^2\mathcal{T}^2}\right)}{\eta} \\
&\quad + 2\eta\sum_{t=1}^{T} \|g_t\|^2 \|u_t\| + 3\sum_{t=1}^{T} \|g_t\| \, \|\widehat{u}_t - u_t\| \\
&\leq 2\epsilon G + \frac{2M\log\left(M/\alpha_{T+1} + 1\right) + 2P_T \log\left(\frac{4M}{\alpha_{t+1}\eta^2\mathcal{T}^2}\right)}{\eta} \\
&\quad + 2\eta\sum_{t=1}^{T} \|g_t\|^2 \|u_t\| + 3\mathcal{T} \sum_{t:\|g_t\|\leq\mathcal{T}} \|\widehat{u}_{t+1} - u_t\| \\
&\overset{(a)}{\leq} 2\epsilon G + \frac{2M\log\left(M/\alpha_{T+1} + 1\right) + 2P_T \log\left(\frac{4M}{\alpha_{t+1}\eta^2\mathcal{T}^2}\right)}{\eta} \\
&\quad + 2\eta\sum_{t=1}^{T} \|g_t\|^2 \|u_t\| + 3\mathcal{T}TP_T.
\end{aligned}
$$

where $(a)$ uses the fact that $\widehat{u}_{t+1} = u_{t'}$ for *some* $t' \geq t$, so that $\|\widehat{u}_{t+1} - u_t\| \leq \sum_{s=1}^{t'-1} \|u_{s+1} - u_s\| \leq P_T$. Since this bound holds for an arbitrary $\mathcal{T} > 0$ we are free to choose a $\mathcal{T}$ which tightens the upperbound, such as $\mathcal{T} = \frac{4}{3\eta T}$:

$$
\begin{aligned}
R_T(\boldsymbol{u}) &\leq \inf_{\mathcal{T}>0} 2\epsilon G + \frac{2M\log\left(M/\alpha_{T+1} + 1\right) + 2P_T \log\left(\frac{4M}{\alpha_{t+1}\eta^2\mathcal{T}^2}\right)}{\eta} \\
&\quad + 2\eta\sum_{t=1}^{T} \|g_t\|^2 \|u_t\| + \mathcal{T}3TP_T \\
&\leq \frac{2M\log\left(M/\alpha_{T+1} + 1\right) + 2P_T \left(\log\left(\frac{9MT^2}{4\alpha_{t+1}}\right) + 2\right)}{\eta} \\
&\quad + 2\epsilon G + 2\eta\sum_{t=1}^{T} \|g_t\|^2 \|u_t\| \\
&\leq 2\epsilon G + \frac{4\left(M + P_T\right)\left\{\log\left(\frac{9MT^2}{4\alpha_{T+1}} + 1\right) \vee 1\right\}}{\eta} + 2\eta\sum_{t=1}^{T} \|g_t\|^2 \|u_t\|.
\end{aligned}
$$

□

### Proof of Theorem 9.1.2

The full statement of the theorem is given below.

**Theorem 9.1.2.** *For any $\varepsilon > 0$ and $u_1, \dots, u_T$ in $\mathbb{R}^d$, Algorithm 8 guarantees*

$$R_T(\boldsymbol{u}) \le 2\varepsilon G + 6\sqrt{2(M+P_T)\left[\log\left(\frac{9M\Lambda_T}{4\varepsilon}+1\right)\vee 1\right]\sum_{t=1}^{T}\|g_t\|^2\|u_t\|}$$
$$+ 4G(M+P_T)\left[\log\left(\frac{9M\Lambda_T}{4\varepsilon}+1\right)\vee 1\right].$$

*where* $\Lambda_T = T^2\left(4+\frac{\|g\|_{1:T}^2}{G^2}\right)\log^2\left(4+\frac{\|g\|_{1:T}^2}{G^2}\right)\lceil\log_2(\sqrt{T})\rceil \le O\left(T^3\log^3(T)\right)$ *and* $M = \max_t \|u_t\|$.

*Proof.* Let $\mathcal{A}_\eta$ denote an instance of the algorithm in Proposition 9.1.1, $w_t^\eta$ denote its iterates, and let $R_T^{\mathcal{A}_\eta}(\boldsymbol{u})$ denote the dynamic regret of $\mathcal{A}_\eta$. From Proposition 9.1.1, we have that for any $\eta \le \frac{1}{G}$,

$$R_T^{\mathcal{A}_\eta}(\boldsymbol{u}) \le 2\epsilon G + \frac{4\left(M+P_T\right)\left[\log\left(\frac{9MT^2}{4\alpha_{T+1}}+1\right)\vee 1\right]}{\eta} + 2\eta\sum_{t=1}^{T}\|g_t\|^2\|u_t\|,$$

where $\alpha_{T+1} = \frac{\epsilon G^2}{V_{T+1}\log^2(V_{T+1}/G^2)}$ and $V_{T+1} = 4G^2 + \|g\|_{1:T}^2$, $M = \max_{t\le T}\|u_t\|$, $P_T = \sum_{t=2}^{T}\|u_t - u_{t-1}\|$, and $\epsilon > 0$. The stepsize which minimizes the right-hand side of the inequality is

$$\eta^* = \min\left\{\sqrt{\frac{2(M+P_T)\left[\log\left(\frac{9MT^2}{4\alpha_{T+1}}+1\right)\vee 1\right]}{\sum_{t=1}^{T}\|g_t\|^2\|u_t\|}}, \frac{1}{G}\right\},$$

for which we have

$$R_T^{\mathcal{A}_{\eta^*}}(\boldsymbol{u}) \le 2\epsilon G + 4\sqrt{2(M+P_T)\left[\log\left(\frac{9MT^2}{4\alpha_{T+1}}+1\right)\vee 1\right]\sum_{t=1}^{T}\|g_t\|^2\|u_t\|}$$
$$+ 2G(M+P_T)\left[\log\left(\frac{9MT^2}{4\alpha_{T+1}}+1\right)\vee 1\right].$$

In what follows, we will match this bound up to constant factors using the iterate adding approach proposed by Cutkosky 2019b.

Suppose that we have a collection of step-sizes $\mathcal{S} = \left\{\eta \in \mathbb{R} : 0 < \eta \le \frac{1}{G}\right\}$ and suppose that on each round we play $w_t = \sum_{\eta \in \mathcal{S}} w_t^\eta$ where $w_t^\eta$ is the output of $\mathcal{A}_\eta$. Then for any $\widetilde{\eta} \in \mathcal{S}$ we can write

$$\begin{aligned}
R_T(\boldsymbol{u}) &= \sum_{t=1}^T \langle g_t, w_t - u_t \rangle = \sum_{t=1}^T \left\langle g_t, \sum_{\eta \in \mathcal{S}} w_t^\eta - u_t \right\rangle \\
&= \sum_{t=1}^T \left\langle g_t, w_t^{\widetilde{\eta}} - u_t \right\rangle + \sum_{\eta \neq \widetilde{\eta} \in \mathcal{S}} \sum_{t=1}^T \langle g_t, w_t^\eta - \boldsymbol{0} \rangle \\
&= R_T^{\mathcal{A}_{\widetilde{\eta}}}(\boldsymbol{u}) + \sum_{\eta \neq \widetilde{\eta} \in \mathcal{S}} R_T^{\mathcal{A}_\eta}(\boldsymbol{0}) \\
&\leq R_T^{\mathcal{A}_{\widetilde{\eta}}}(\boldsymbol{u}) + 2\epsilon G(|\mathcal{S}| - 1). \qquad \text{(C.1)}
\end{aligned}$$

Notice that since this holds for any $\widetilde{\eta} \in \mathcal{S}$, it holds for the one with the lowest dynamic regret, hence

$$R_T(\boldsymbol{u}) \leq 2\epsilon G(|\mathcal{S}| - 1) + \min_{\eta \in \mathcal{S}} R_T^{\mathcal{A}_\eta}(\boldsymbol{u}).$$

Thus, we need only ensure that there is *some* $\eta \in \mathcal{S}$ which is close to the optimal $\eta^*$. It is easy to see that

$$\eta^* = \min\left\{ \sqrt{\frac{2(M + P_T)\left[\log\left(\frac{9MT^2}{4\alpha_{T+1}} + 1\right) \vee 1\right]}{\sum_{t=1}^T \|g_t\|^2 \|u_t\|}}, \frac{1}{G} \right\} \implies \frac{2}{G\sqrt{T}} \leq \eta^* \leq \frac{1}{G},$$

so if we let $\mathcal{S} = \left\{ \frac{2^k}{G\sqrt{T}} \wedge \frac{1}{G} : 1 \leq k \leq \lceil \log_2\left(\sqrt{T}\right) \rceil \right\}$, we'll have

$$\eta_{\min} = \frac{2}{G\sqrt{T}} \leq \eta^* \leq \frac{1}{G} = \eta_{\max},$$

where $\eta_{\min}$ and $\eta_{\max}$ are the smallest and largest step-sizes in $\mathcal{S}$ respectively. Hence, there must be an $\eta_k \in \mathcal{S}$ such that $\eta_k \leq \eta^* \leq \eta_{k+1} \leq 2\eta_k$. Using $\widetilde{\eta} = \eta_k$ in Equation (C.1) yields

$$\begin{aligned}
R_T(\boldsymbol{u}) &\leq 2\epsilon G(|\mathcal{S}| - 1) + R_T^{\mathcal{A}_{\eta_k}}(\boldsymbol{u}) \\
&\leq 2\epsilon G|\mathcal{S}| + \frac{4(M + P_T)\left[\log\left(\frac{9MT^2}{4\alpha_{T+1}} + 1\right) \vee 1\right]}{\eta_k} + 2\eta_k \sum_{t=1}^T \|g_t\|^2 \|u_t\| \\
&\leq 2\epsilon G|\mathcal{S}| + \frac{8(M + P_T)\left[\log\left(\frac{9MT^2}{4\alpha_{T+1}} + 1\right) \vee 1\right]}{\eta^*} + 2\eta^* \sum_{t=1}^T \|g_t\|^2 \|u_t\| \\
&= 2\epsilon G|\mathcal{S}| + 6\sqrt{2(M + P_T)\left[\log\left(\frac{9MT^2}{4\alpha_{T+1}} + 1\right) \vee 1\right] \sum_{t=1}^T \|g_t\|^2 \|u_t\|} \\
&\quad + 4G(M + P_T)\left[\log\left(\frac{9MT^2}{4\alpha_{T+1}} + 1\right) \vee 1\right].
\end{aligned}$$

The result then follows by choosing $\epsilon = \frac{\varepsilon}{\lceil \log_2(\sqrt{T})\rceil} \leq \frac{\varepsilon}{|\mathcal{S}|}$. □

### Proof of Proposition 9.1.4

**Proposition 9.1.4.** *Suppose $\mathcal{A}$ is an online learning algorithm which guarantees*

$$R_T^{\mathcal{A}}(\boldsymbol{u}) \leq \widetilde{O}\left(\sqrt{(M^2 + MP_T)\sum_{t=1}^{T}\|g_t\|^2}\right),$$

*for all $u_1, \ldots, u_T$ in $\mathbb{R}^d$ with $\max_{t\leq T}\|u_t\| \leq M$. Then for all $u_1, \ldots, u_T$ in $\mathbb{R}^d$, Algorithm 10 guarantees*

$$R_T(\boldsymbol{u}) \leq \widetilde{O}\left(\max_{k\leq K}|I_k|\sqrt{(M^2 + MP_T)\|g\|_{1:T}^2}\right)$$

*Proof.* First observe that for any interval $I = [a, b]$, we have

$$\sum_{t\in I}\langle g_t, w_t - u_t\rangle = \sum_{t\in I}\langle g_t, w_t - u_b\rangle + \sum_{t\in I}\langle g_t, u_b - u_t\rangle,$$

and bound the second sum as

$$\begin{aligned}
\sum_{t\in I}\left\langle g_t, \sum_{s=t+1}^{b} u_s - u_{s-1}\right\rangle &= \sum_{t=a}^{b}\sum_{s=t+1}^{b}\langle g_t, u_s - u_{s-1}\rangle \\
&= \sum_{s=a+1}^{b}\langle g_{a:s-1}, u_s - u_{s-1}\rangle \leq \sqrt{\sum_{s=a+1}^{b}\|g_{a:s-1}\|^2\sum_{t=a+1}^{b}\|u_t - u_{t-1}\|^2} \\
&\leq \sqrt{\left(\sum_{t=a+1}^{b}\|g_t\|^2 + \sum_{t=a+1}^{b}\sum_{t'\neq t}^{b}\|g_t\|\|g_{t'}\|\right)S_I} \\
&\leq \sqrt{\left(\sum_{t=a+1}^{b}\|g_t\|^2 + \max_{s\in[a,b]}\|g_s\|^2|I|^2\right)S_I} \\
&\leq \sqrt{2\|g\|_{a+1:b}^2|I|^2 S_I} = \sqrt{2\|g\|_{a+1:b}^2 S_I}|I|.
\end{aligned}$$

where $S_I = \sum_{t=a+1}^{b}\|u_t - u_{t-1}\|^2$. Thus, denoting $I_1 = [1, \tau_1]$, $I_2 = [\tau_1 + 1, \tau_2], \ldots, I_K = [\tau_{K-1} + 1, \tau_K]$, we can bound

$$\begin{aligned}\sum_{t=1}^{T}\langle g_t, w_t - u_t\rangle &= \sum_{k=1}^{K}\sum_{t\in I_k}\langle g_t, w_t - u_t\rangle = \sum_{k=1}^{K}\sum_{t\in I_k}\langle g_t, w_t - u_{\tau_k}\rangle + \langle g_t, u_{\tau_k} - u_t\rangle\\ &\le \sum_{k=1}^{K}\sum_{t\in I_k}\langle g_t, w_t - u_{\tau_k}\rangle + \sum_{k=1}^{K}\sqrt{2S_{I_k}\left\|g\right\|_{t\in I_k}^2}|I_k|\\ &\le \sum_{k=1}^{K}\left\langle\sum_{t\in I_k} g_t, w_{\tau_k} - u_{\tau_k}\right\rangle + \sqrt{2S_T\left\|g\right\|_{1:T}^2}\max_{k\le K}|I_k|\end{aligned}$$

where the last line observes that $w_t$ is fixed within each interval. From the regret guarantee of algorithm $\mathcal{A}$ we have

$$\begin{aligned}\sum_{k=1}^{K}\left\langle\sum_{t\in I_k} g_t, w_{\tau_k} - u_{\tau_k}\right\rangle &= \sum_{k=1}^{K}\langle\widetilde{g}_{\tau_k}, w_{\tau_k} - u_{\tau_k}\rangle \le \widetilde{O}\left(\sqrt{(M^2 + M\widehat{P}_K)\left\|\widetilde{g}\right\|_{1:K}^2}\right)\\ &\le \widetilde{O}\left(\max_{k\le K}|I_k|\sqrt{2(M^2 + MP_T)\left\|g\right\|_{1:T}^2}\right),\end{aligned}$$

where the first line defines $\widehat{P}_K = \sum_{k=2}^{K}\left\|u_{\tau_k} - u_{\tau_{k-1}}\right\|$ and the last line observes $\widehat{P}_K \le P_T$. Hence,

$$\sum_{t=1}^{T}\langle g_t, w_t - u_t\rangle \le \widetilde{O}\left(\max_{k\le K}|I_k|\left(\sqrt{2(M^2 + MP_T)\left\|g\right\|_{1:T}^2} + \sqrt{2S_T\left\|g\right\|_{1:T}^2}\right)\right).$$

The stated bound follows by observing that $S_T \le MP_T \le M^2 + MP_T$ and hiding constants.

□

### C.1.2 Proofs for Section 9.2 (Unbounded Losses)

The main objective of this section is to prove Theorems 9.2.1 and 9.2.3. At a high level, the strategy is simple: we run several instances of projected gradient descent, each with a different restricted domain $W_D = \{w \in W : \|w\| \le D\}$ and stepsize $\eta$, and then use a particular experts algorithm to combine them. We first assemble a collection of core lemmas that provide the regret of the base algorithm (Lemma C.1.2), the regret of Algorithm 11 in terms of the regret of any of the base algorithms (Lemma C.1.3), as well as some utility lemmas (Lemmas C.1.4 to C.1.7) to help tame some unwieldy algebraic expressions and case work. We then prove the main results Theorems 9.2.1 and 9.2.3 in Appendix C.1.2 respectively. Finally, we prove our lowerbound Theorem 9.2.2 in Appendix C.1.

The base algorithms that we combine are instances of (projected) online gradient descent with an additional bias term added to the update. The following lemma provides the regret template for this algorithm.

**Lemma C.1.2.** *For all $t$ let $\ell_t : W \to \mathbb{R}$ be convex. Let $K \geq 1$, $L_t \geq 0$, and $K\eta L_t \leq 1$ for all $t$. Let $W_D = \{w \in W : \|w\| \leq D\}$, $w_1 = \mathbf{0}$, and on each round update $w_{t+1} = \Pi_{w \in W_D}(w_t - \eta(1 + K\eta L_t)g_t)$, where $g_t \in \partial\ell_t(w_t)$. Then for any $\boldsymbol{u} = (u_1, \ldots, u_T)$ in $W_D$,*

$$R_T(\boldsymbol{u}) \leq \frac{\|u_T\|^2 + 2DP_T}{2\eta} + K\eta\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t(w_t)\right] + 2\eta\sum_{t=1}^T \|g_t\|^2$$

*where $P_T = \sum_{t=2}^T \|u_t - u_{t-1}\|$.*

*Proof.* The result follows easily using existing analyses. For instance, the update can be seen as an instance of Algorithm 2 with $\psi_t(w) = \frac{1}{2\eta}\|w\|^2$, $\phi_t(w) = K\eta L_t \langle g_t, w\rangle$ for $g_t \in \partial\ell_t(w_t)$, domain $W_D = \{w \in W : \|w\| \leq D\}$, and $\mathcal{M}_t(w) = w$ for all $t$. Letting $w_1 = \mathbf{0}$ and applying Lemma 4.1.2, we have:

$$\begin{aligned}
R_T(\boldsymbol{u}) &\leq \psi_{T+1}(u_T) + \sum_{t=2}^T \langle \nabla\psi_t(w_t), u_{t-1} - u_t\rangle + K\eta\sum_{t=1}^T L_t\ell_t(u_t) \\
&\quad + \sum_{t=1}^T \langle g_t + K\eta L_t g_t, w_t - w_{t+1}\rangle - D_{\psi_t}(w_{t+1}|w_t) - K\eta L_t\ell_t(w_t) \\
&\leq \frac{\|u_T\|^2}{2\eta} + \sum_{t=2} \frac{D}{\eta}\|u_t - u_{t-1}\| + K\eta\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t(w_t)\right] \\
&\quad + \sum_{t=1}^T (1 + K\eta L_t)\langle g_t, w_t - w_{t+1}\rangle - D_{\psi_t}(w_{t+1}|w_t) \\
&\overset{(a)}{\leq} \frac{\|u_T\|^2 + 2DP_T}{2\eta} + K\eta\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t(w_t)\right] \\
&\quad + \sum_{t=1}^T (1 + K\eta L_t)\langle g_t, w_t - w_{t+1}\rangle - \frac{\|w_{t+1} - w_t\|^2}{2\eta} \\
&\overset{(b)}{\leq} \frac{\|u_T\|^2 + 2DP_T}{2\eta} + K\eta\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t(w_t)\right] + \frac{\eta}{2}\sum_{t=1}^T (1 + K\eta L_t)^2\|g_t\|^2 \\
&\overset{(c)}{\leq} \frac{\|u_T\|^2 + 2DP_T}{2\eta} + K\eta\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t(w_t)\right] + 2\eta\sum_{t=1}^T \|g_t\|^2
\end{aligned}$$

the $(a)$ observes that $D_{\psi_t}(w_{t+1}|w_t) \geq \frac{\|w_{t+1} - w_t\|^2}{2\eta}$ by $\frac{1}{\eta}$-strong convexity of $\psi$, $(b)$ is Fenchel-Young inequality, and $(c)$ uses $K\eta L_t \leq 1$.

□

The following lemma provides a generic regret bound for Algorithm 11. The take-away is that the regret will scale with the regret of any of the experts up to two extra terms $C_{\mathcal{S}}$ and $\Lambda_T(\eta, D)$, which we will later ensure are small.

**Lemma C.1.3.** *For any $\tau = (\eta, D) \in \mathcal{S}$ with $\eta \le \frac{1}{KL_{\max}}$ and sequence $\boldsymbol{u} = (u_1, \dots, u_T)$ in $W$ satisfying $\|u_t\| \le D$ for all $t$, Algorithm 11 guarantees*

$$R_T(\boldsymbol{u}) \le 2kC_{\mathcal{S}} + 2kDG_{\max}\Lambda_T(\tau) + \frac{\|u_T\|^2 + 2DP_T + 4kD^2\Lambda_T(\tau)}{2\eta}$$
$$+ K\eta \sum_{t=1}^T L_t \left[\ell_t(u_t) - \ell_t(w_t^{(\tau)})\right] + 4\eta \sum_{t=1}^T \left\|g_t^{(\tau)}\right\|^2$$

*where $k \ge 9/2$ and*

$$C_{\mathcal{S}} \overset{def}{=} \frac{\sum_{\widetilde{\tau} \in \mathcal{S}} \mu_{\widetilde{\tau}}}{\sum_{\widetilde{\tau} \in \mathcal{S}} \mu_{\widetilde{\tau}}^2}, \qquad \Lambda_T(\tau) \overset{def}{=} \log\left(\frac{\sum_{\widetilde{\tau} \in \mathcal{S}} \mu_{\widetilde{\tau}}^2}{\mu_\tau^2}\right) + 1.$$

*Proof.* Let $\tau = (\eta, D) \in \mathcal{S}$ and let $\mathcal{A}_\tau$ denote an algorithm which after each round updates its parameters using

$$w_{t+1}^{(\tau)} = \Pi_{w \in W : \|w\| \le D}\left(w_t^{(\tau)} - \eta(1 + K\eta L_t)g_t^{(\tau)}\right)$$

for $g_t^{(\tau)} \in \partial \ell_t(w_t^{(\tau)})$. Algorithm 11 is constructed as a collection of algorithms $\mathcal{A}_\tau$, with an multi-scale experts algorithm (Algorithm 14) to combine their predictions. First, observe that the regret decomposes into the regret of any expert $\mathcal{A}_\tau$ plus the regret of the experts algorithm relative to expert $\mathcal{A}_\tau$:

$$\begin{aligned}
R_T(\boldsymbol{u}) &= \sum_{t=1}^T \ell_t(w_t) - \ell_t(u_t) \\
&= \underbrace{\sum_{t=1}^T \ell_t\left(w_t^{(\tau)}\right) - \ell_t(u_t)}_{=:R_T^{\mathcal{A}_\tau}(\boldsymbol{u})} + \sum_{t=1}^T \ell_t(w_t) - \ell_t\left(w_t^{(\tau)}\right) \\
&= R_T^{\mathcal{A}_\tau}(\boldsymbol{u}) + \sum_{t=1}^T \ell_t\left(\sum_{\widetilde{\tau} \in \mathcal{S}} p_t(\widetilde{\tau})\ell_t(w_t^{(\widetilde{\tau})})\right) - \ell_t\left(w_t^{(\tau)}\right)
\end{aligned}$$

and by convexity of $\ell_t$ and Jensen's inequality:

$$\begin{aligned}
&\leq R_T^{\mathcal{A}_\tau}(\boldsymbol{u}) + \sum_{t=1}^{T}\left[\sum_{\widetilde{\tau}\in\mathcal{S}} p_t(\widetilde{\tau})\ell_t\left(w_t^{(\widetilde{\tau})}\right)\right] - \ell_t\left(w_t^{(\tau)}\right) \\
&= R_T^{\mathcal{A}_\tau}(\boldsymbol{u}) + \sum_{t=1}^{T}\sum_{\widetilde{\tau}\in\mathcal{S}} \ell_t\left(w_t^{(\widetilde{\tau})}\right)\left[p_t(\widetilde{\tau}) - \mathbf{1}\left\{\tau = \widetilde{\tau}\right\}\right] \\
&\overset{(a)}{=} R_T^{\mathcal{A}_\tau}(\boldsymbol{u}) \\
&\qquad + \sum_{t=1}^{T}\sum_{\widetilde{\tau}\in\mathcal{S}}\left[\ell_t\left(w_t^{(\widetilde{\tau})}\right) - \ell_t(\widetilde{w}_t)\right]\left[p_t(\widetilde{\tau}) - p_\tau^*(\widetilde{\tau})\right] \\
&\qquad + \sum_{t=1}^{T}\sum_{\widetilde{\tau}\in\mathcal{S}} \ell_t(\widetilde{w}_t)(p_t(\widetilde{\tau}) - p_\tau^*(\widetilde{\tau})) \\
&\overset{(b)}{=} R_T^{\mathcal{A}_\tau}(\boldsymbol{u}) + \sum_{t=1}^{T}\sum_{\widetilde{\tau}\in\mathcal{S}}\left[\ell_t\left(w_t^{(\widetilde{\tau})}\right) - \ell_t(\widetilde{w}_t)\right]\left[p_t(\widetilde{\tau}) - p_\tau^*(\widetilde{\tau})\right] \\
&\overset{(c)}{=} R_T^{\mathcal{A}_\tau}(\boldsymbol{u}) + \underbrace{\sum_{t=1}^{T}\left\langle\widetilde{\ell}_t, p_t - p_\tau^*\right\rangle}_{=:R_T^{\mathrm{Meta}}(p_\tau^*)} \\
&= R_T^{\mathcal{A}_\tau}(\boldsymbol{u}) + R_T^{\mathrm{Meta}}(p_\tau^*), \qquad\qquad (\mathrm{C.2})
\end{aligned}$$

where $\widetilde{w}_t$ is an arbitrary reference point with $\|\widetilde{w}_t\| \leq D_{\min}$ (and hence is in the domain of all of the experts $\mathcal{A}_\tau$), $(a)$ defines $p_\tau^*(\widetilde{\tau}) = 1$ if $\widetilde{\tau} = \tau$ and 0 otherwise, $(b)$ observes that $\sum_{\widetilde{\tau}\in\mathcal{S}} \ell_t(\widetilde{w}_t)(p_t(\widetilde{\tau}) - p_\tau^*(\widetilde{\tau})) = \ell_t(\widetilde{w}_t)\sum_{\widetilde{\tau}\in\mathcal{S}} p_t(\widetilde{\tau}) - p_\tau^*(\widetilde{\tau}) = 0$, and $(c)$ defines $\widetilde{\ell}_t \in \mathbb{R}^{|\mathcal{S}|}$ with $\widetilde{\ell}_{t,\tau} = \ell_t(w_t^{(\tau)}) - \ell_t(\widetilde{w}_t)$.

Now for any $\tau = (\eta, D) \in \mathcal{S}$ with $D \geq \max_t \|u_t\|$, we have via Lemma C.1.2 that

$$R_T^{\mathcal{A}_\tau}(u) \leq \frac{\|u_T\|^2 + 2DP_T}{2\eta} + K\eta\sum_{t=1}^{T} L_t\left[\ell_t(u_t) - \ell_t(w_t^{(\tau)})\right] + 2\eta\sum_{t=1}^{T}\left\|g_t^{(\tau)}\right\|^2. \qquad (\mathrm{C.3})$$

To bound $R_T^{\mathrm{Meta}}(p_\tau^*)$, observe that for any $\widetilde{\tau} = (\widetilde{\eta}, \widetilde{D})$, we have

$$\begin{aligned}
\widetilde{\ell}_{t,\widetilde{\tau}} = \ell_t(w_t^{(\widetilde{\tau})}) - \ell_t(\widetilde{w}_t) &\leq \left\|\nabla\ell_t(w_t^{(\widetilde{\tau})})\right\|\left\|w_t^{(\widetilde{\tau})} - \widetilde{w}_t\right\| \\
&\leq \left(G_{\max} + L_{\max}\widetilde{D}\right)2\widetilde{D},
\end{aligned}$$

and so with $\mu_{\widetilde{\tau}} = \frac{1}{2\widetilde{D}(G_{\max}+\widetilde{D}/\widetilde{\eta})}$ and $\widetilde{\eta} \leq \frac{1}{KL_{\max}} \leq \frac{1}{L_{\max}}$ we have

$$\begin{aligned}
\mu_{\widetilde{\tau}}\ell_{t,\widetilde{\tau}} &\leq \frac{1}{2\widetilde{D}\left(G_{\max} + \widetilde{D}/\widetilde{\eta}\right)} 2\widetilde{D}\left(G_{\max} + L_{\max}\widetilde{D}\right) \\
&\leq \frac{1}{\left(G_{\max} + L_{\max}\widetilde{D}\right)}\left(G_{\max} + L_{\max}\widetilde{D}\right) \\
&= 1,
\end{aligned}$$

so these choices meet the assumptions of Theorem B.2.1 and we have:

$$R_T^{\mathrm{Meta}}(p_\tau^*) \leq \sum_{\widetilde{\tau}\in\mathcal{S}} p_\tau^*(\widetilde{\tau})\left[\frac{k\left[\log\left(p_\tau^*(\widetilde{\tau})/p_{1\widetilde{\tau}}\right) + 1\right]}{\mu_{\widetilde{\tau}}} + \mu_{\widetilde{\tau}}\sum_{t=1}^{T}\widetilde{\ell}_{t\widetilde{\tau}}^2\right] + 2k\sum_{\widetilde{\tau}\in\mathcal{S}}\frac{p_{1\widetilde{\tau}}}{\mu_{\widetilde{\tau}}}$$

for $k \geq 9/2$. Recalling that $p_\tau^*(\widetilde{\tau}) = 1$ when $\widetilde{\tau} = \tau$ and 0 otherwise and that $\tau = (D, \eta)$, the first sum is bound as

$$\begin{aligned}
\frac{k\left[\log\left(p_\tau^*(\tau)/p_{1\tau}\right) + 1\right]}{\mu_\tau} + \mu_\tau\sum_{t=1}^{T}\widetilde{\ell}_{t\tau}^2 &= 2kD\left(G_{\max} + \frac{D}{\eta}\right)\left[\log\left(1/p_{1\tau}\right) + 1\right] + \frac{\eta}{2D\left(G_{\max}\eta + D\right)}\sum_{t=1}^{T}\widetilde{\ell}_{t,\tau}^2 \\
&\leq 2kD\left(G_{\max} + \frac{D}{\eta}\right)\left[\log\left(1/p_{1\tau}\right) + 1\right] + \frac{\eta}{2D^2}\sum_{t=1}^{T}\left\|\nabla\ell_t(w_t^{(\tau)})\right\|^2 4D^2 \\
&= 2kD\left(G_{\max} + \frac{D}{\eta}\right)\left[\log\left(1/p_{1\tau}\right) + 1\right] + 2\eta\sum_{t=1}^{T}\left\|g_t^{(\tau)}\right\|^2,
\end{aligned}$$

and so with $p_{1,\tau} = \frac{\mu_\tau^2}{\sum_{\widetilde{\tau}\in\mathcal{S}}\mu_{\widetilde{\tau}}^2}$, we have

$$R_T^{\mathrm{Meta}}(p_\tau^*) \leq 2kD\left(G_{\max} + \frac{D}{\eta}\right)\left[\log\left(\frac{\sum_{\widetilde{\tau}\in\mathcal{S}}\mu_{\widetilde{\tau}}^2}{\mu_\tau^2}\right) + 1\right] + 2\eta\sum_{t=1}^{T}\left\|g_t^{(\tau)}\right\|^2 + 2k\sum_{\widetilde{\tau}\in\mathcal{S}}\frac{\mu_{\widetilde{\tau}}}{\sum_{\widetilde{\tau}\in\mathcal{S}}\mu_{\widetilde{\tau}}^2}.$$

Combining this with Equations (C.2) and (C.3) yields the stated result:

$$\begin{aligned}
R_T(\boldsymbol{u}) \leq{}& \frac{\left\|u_T\right\|^2 + 2DP_T}{2\eta} + K\eta\sum_{t=1}^{T}L_t\left[\ell_t(u_t) - \ell_t(w_t^{(\tau)})\right] + 4\eta\sum_{t=1}^{T}\left\|g_t^{(\tau)}\right\|^2 \\
&+ 2kD\left(G_{\max} + \frac{D}{\eta}\right)\left[\log\left(\frac{\sum_{\widetilde{\tau}\in\mathcal{S}}\mu_{\widetilde{\tau}}^2}{\mu_\tau}\right) + 1\right] + 2k\frac{\sum_{\widetilde{\tau}\in\mathcal{S}}\mu_{\widetilde{\tau}}}{\sum_{\widetilde{\tau}\in\mathcal{S}}\mu_{\widetilde{\tau}}^2} \\
={}& 2kC_{\mathcal{S}} + 2kDG_{\max}\Lambda_T(\tau) + \frac{\left\|u_T\right\|^2 + 2DP_T + 4kD^2\Lambda_T(\tau)}{2\eta} \\
&+ K\eta\sum_{t=1}^{T}L_t\left[\ell_t(u_t) - \ell_t(w_t^{(\tau)})\right] + 4\eta\sum_{t=1}^{T}\left\|g_t^{(\tau)}\right\|^2
\end{aligned}$$

where the last line defines the shorthand notations

$$C_{\mathcal{S}} \overset{\text{def}}{=} \frac{\sum_{\widetilde{\tau} \in \mathcal{S}} \mu_{\widetilde{\tau}}}{\sum_{\widetilde{\tau} \in \mathcal{S}} \mu_{\widetilde{\tau}}^2} \qquad \text{and} \qquad \Lambda_T(\tau) \overset{\text{def}}{=} \log\left(\frac{\sum_{\widetilde{\tau} \in \mathcal{S}} \mu_{\widetilde{\tau}}^2}{\mu_\tau^2}\right) + 1.$$

□

Next, we provide bounds on the terms terms $C_{\mathcal{S}}$ and $\Lambda_T$ in terms of the hyperparameter ranges $[\eta_{\min}, \eta_{\max}]$ and $[D_{\min}, D_{\max}]$ that the meta-algorithm tunes the hyperparameters over.

**Lemma C.1.4.** *Let $0 < \eta_{\min} \le \eta_{\max}$, $0 < D_{\min} \le D_{\max}$, and define the hyperparameter set $\mathcal{S} = \mathcal{S}_\eta \times \mathcal{S}_D$ for $\mathcal{S}_\eta = \left\{\eta_i = \left[\eta_{\min} 2^i \wedge \eta_{\max}\right] : i \ge 0\right\}$ and $\mathcal{S}_D = \left\{D_j = \left[D_{\min} 2^j \wedge D_{\max}\right] : j \ge 0\right\}$. For each $\tau = (\eta, D) \in \mathcal{S}$, let $\mu_\tau = \frac{1}{2D(G_{\max} + D/\eta)}$. Then*

$$C_{\mathcal{S}} \overset{def}{=} \frac{\sum_{\tau \in \mathcal{S}} \mu_\tau}{\sum_{\tau \in \mathcal{S}} \mu_\tau^2} \le 2\sqrt{T} D_{\min} \left(G_{\max} + \frac{D_{\min}}{\eta_{\max}}\right)$$

*and for any $\tau \in \mathcal{S}$,*

$$\Lambda_T(\tau) \overset{def}{=} \log\left(\frac{\sum_{\widetilde{\tau} \in \mathcal{S}} \mu_{\widetilde{\tau}}^2}{\mu_\tau^2}\right) + 1 \le \log\left(\frac{24\eta_{\max}^2 D^4}{\eta_{\min}^2 D_{\min}^4} \wedge \frac{6\left|\mathcal{S}_\eta\right| D^2}{D_{\min}^2}\right) + 1$$

*Proof.* For the first statement, we have

$$\begin{aligned} C_{\mathcal{S}} = \frac{\sum_{\widetilde{\tau} \in \mathcal{S}} \mu_{\widetilde{\tau}}}{\sum_{\widetilde{\tau} \in \mathcal{S}} \mu_{\widetilde{\tau}}^2} &\le \sqrt{\frac{T}{\sum_{\widetilde{\tau} \in \mathcal{S}} \mu_{\widetilde{\tau}}^2}} \le \sqrt{\frac{T}{\mu_{(\eta_{\max}, D_{\min})}^2}} \\ &= \sqrt{T(2D_{\min})^2 \left(G_{\max} + D_{\min}/\eta_{\max}\right)^2} \\ &= 2\sqrt{T} D_{\min} \left(G_{\max} + \frac{D_{\min}}{\eta_{\max}}\right) \end{aligned}$$

where the first inequality applies Cauchy-Schwarz inequality. Moreover, for any $\tau = (\eta, D) \in \mathcal{S}$ we have

$$\begin{aligned}
\frac{\sum_{\widetilde{\tau}\in\mathcal{S}}\mu_{\widetilde{\tau}}^2}{\mu_\tau^2} &= \frac{1}{\mu_{(\eta,D)}^2}\left[\sum_{(\eta_i,D_j)\in\mathcal{S}}\frac{1}{(2D_j)^2\left[G_{\max}+D_j/\eta_i\right]^2}\right] \\
&\leq \frac{1}{4\mu_{(\eta,D)}^2}\sum_{(\eta_i,D_j)\in\mathcal{S}}\frac{\eta_i^2}{D_j^4} \\
&= \frac{1}{4\mu_{(\eta,D)}^2}\sum_{(\eta_i,D_j)\in\mathcal{S}}\frac{2^{2i}\eta_{\min}^2}{D_{\min}^4 2^{4j}} \\
&= \frac{\eta_{\min}^2}{4\mu_{(\eta,D)}^2 D_{\min}^4}\sum_{i=0}^{\lceil\log_2(\eta_{\max}/\eta_{\min})\rceil}\sum_{j=0}^{\lceil\log_2(D_{\max}/D_{\min})\rceil}\frac{2^{2i}}{2^{4j}} \\
&\leq \frac{\eta_{\min}^2}{4\mu_{(\eta,D)}^2 D_{\min}^4}\frac{2^{2\lceil\log_2(\eta_{\max}/\eta_{\min})\rceil+2}-1}{3}\frac{1}{1-\frac{1}{16}} \\
&\leq \frac{\eta_{\min}^2}{4\mu_{(\eta,D)}^2 D_{\min}^4}\frac{2^{\log_2(\eta_{\max}^2/\eta_{\min}^2)+4}-1}{3}\frac{16}{15} \\
&\leq \frac{\eta_{\min}^2}{4\mu_{(\eta,D)}^2 D_{\min}^4}16\frac{\eta_{\max}^2}{\eta_{\min}^2}\frac{16}{45} \\
&\leq \frac{6\eta_{\max}^2}{4\mu_{(\eta,D)}^2 D_{\min}^4} = \frac{3\eta_{\max}^2}{2\mu_{(\eta,D)}^2 D_{\min}^4}
\end{aligned}$$

At the same time, we can also bound this term as

$$\begin{aligned}
\frac{\sum_{\widetilde{\tau}\in\mathcal{S}}\mu_{\widetilde{\tau}}^2}{\mu_\tau^2} &= \frac{1}{\mu_{(\eta,D)}^2}\left[\sum_{(\eta_i,D_j)\in\mathcal{S}}\frac{1}{(2D_j)^2\left[G_{\max}+D_j/\eta_i\right]^2}\right] \\
&\leq \frac{1}{4\mu_{(\eta,D)}^2}\sum_{(\eta_i,D_j)\in\mathcal{S}}\frac{1}{D_j^2 G_{\max}^2} \\
&\leq \frac{|\mathcal{S}_\eta|}{4\mu_{(\eta,D)}^2 G_{\max}^2}\sum_{j=0}^{\lceil\log_2(D_{\max}/D_{\min})\rceil}\frac{1}{D_{\min}^2 2^{2j}} \\
&\leq \frac{|\mathcal{S}_\eta|}{4\mu_{(\eta,D)}^2 G_{\max}^2 D_{\min}^2}\frac{1}{1-\frac{1}{4}} \\
&\leq \frac{4\,|\mathcal{S}_\eta|}{4\cdot 3\mu_{(\eta,D)}^2 G_{\max}^2 D_{\min}^2} = \frac{|\mathcal{S}_\eta|}{3\mu_{(\eta,D)}^2 G_{\max}^2 D_{\min}^2}
\end{aligned}$$

Hence,

$$\begin{aligned}
\Lambda_T(\eta, D) &= \log\left(\frac{\sum_{\widetilde{\tau}} \mu_{\widetilde{\tau}}}{\mu_{\tau}^2}\right) + 1 \\
&\leq \log\left(\left[\frac{3\eta_{\max}^2}{2D_{\min}^2} \wedge \frac{|\mathcal{S}_\eta|}{3G_{\max}^2}\right] \frac{1}{\mu_{(\eta,D)}^2 D_{\min}^2}\right) + 1 \\
&= \log\left(\left[\frac{3\eta_{\max}^2}{2D_{\min}^2} \wedge \frac{|\mathcal{S}_\eta|}{3G_{\max}^2}\right] \frac{(2D)^2 \left[G_{\max} + D/\eta\right]^2}{D_{\min}^2}\right) + 1.
\end{aligned}$$

Now if $G_{\max} \leq D/\eta$, we have

$$\begin{aligned}
\Lambda_T(\eta, D) &\leq \log\left(\frac{3 \cdot 4 \cdot \eta_{\max}^2 D^2 \left[G_{\max} + D/\eta\right]^2}{2D_{\min}^4}\right) + 1 \\
&\leq \log\left(\frac{6\eta_{\max}^2 D^2 \cdot (2D/\eta)^2}{D_{\min}^4}\right) + 1 \\
&\leq \log\left(\frac{24\eta_{\max}^2 D^4}{\eta_{\min}^2 D_{\min}^4}\right) + 1
\end{aligned}$$

and otherwise

$$\begin{aligned}
\Lambda_T(\eta, D) &\leq \log\left(\frac{4\left|\mathcal{S}_\eta\right| D^2 \left[G_{\max} + D/\eta\right]^2}{3G_{\max}^2 D_{\min}^2}\right) + 1 \\
&\leq \log\left(\frac{6\left|\mathcal{S}_\eta\right| D^2 G_{\max}^2}{G_{\max}^2 D_{\min}^2}\right) + 1 \\
&= \log\left(\frac{6\left|\mathcal{S}_\eta\right| D^2}{D_{\min}^2}\right) + 1.
\end{aligned}$$

Thus, we can bound

$$\Lambda_T(\eta, D) \leq \log\left(\frac{24\eta_{\max}^2 D^4}{\eta_{\min}^2 D_{\min}^4} \wedge \frac{6\left|\mathcal{S}_\eta\right| D^2}{D_{\min}^2}\right) + 1$$

□

Lemma C.1.5 provides a simple but tedius calculation which we will use a few times in the proof of Theorem 9.2.1.

**Lemma C.1.5.** *Let $\ell_t$ be $(G_t, L_t)$-quadratically bounded, $c_1, c_2 \geq 0$, $u, w \in W$, and $g_t \in \partial\ell_t(w)$. Assume $\|w\| \leq D$ and $\|u\| \leq D$. Then*

$$c_1 L_t \left[\ell_t(u) - \ell_t(w)\right] + c_2 \left\|g_t\right\|^2 \leq 3(c_1 + c_2)\left(G_t^2 + L_t^2 D^2\right)$$

*Proof.* Since $\ell_t$ is $(G_t, L_t)$-quadratically bounded, and $g_t \in \partial\ell_t(w)$ where $\|w\| \leq D$ we have

$$\|g_t\|^2 \leq (G_t + L_t\|w\|)^2 \leq 2G_t^2 + 2L_t^2\|w\|^2 \leq 2G_t^2 + 2L_t^2D^2.$$

Moreover, letting $\nabla\ell_t(u) \in \partial\ell_t(u)$ and $\|u\| \leq D$ we have

$$\begin{aligned}
L_t(\ell_t(u) - \ell_t(w)) &\leq L_t\|\nabla\ell_t(u)\|\|u - w\| \\
&\leq 2DL_t\|\nabla\ell_t(u)\| \\
&\leq 2DL_t(G_t + L_tD) \\
&= 2DL_tG_t + 2L_t^2D^2 \\
&\leq G_t^2 + L_t^2D^2 + 2L_t^2D^2 \\
&= G_t^2 + 3L_t^2D^2.
\end{aligned}$$

Thus,

$$\begin{aligned}
c_1L_t(\ell_t(u) - \ell_t(w)) + c_2\|g_t\|^2 &\leq (c_1 + 2c_2)G_t^2 + (3c_1 + 2c_2)L_t^2D^2 \\
&\leq 3(c_1 + c_2)\left(G_t^2 + L_t^2D^2\right)
\end{aligned}$$

□

Lastly, we provide two lemmas which let us assume that there is a $\tau = (\eta, D) \in \mathcal{S}$ for which $\frac{1}{2}D \leq M = \max_t\|u_t\| \leq D$ by showing that the regret is trivially well-controlled whenever $M$ is "too big" (Lemma C.1.6) or "too small" (Lemma C.1.7).

**Lemma C.1.6.** *For all $t$ let $\ell_t$ be a $(G_t, L_t)$-quadratically bounded convex function for $G_t \in [0, G_{\max}]$ and $L_t \in [0, L_{\max}]$. Let $\varepsilon > 0$, $D_{\max} = \varepsilon 2^T$, and let $\boldsymbol{u} = (u_1, \ldots, u_T)$ be an arbitrary sequence in $W$ such that $M := \max_t\|u_t\| \geq D_{\max}$. Then for any $w_1, \ldots, w_T$ with $\|w_t\| \leq D_{\max}$,*

$$\sum_{t=1}^{T}\ell_t(w_t) - \ell_t(u_t) \leq 2\left(G_{\max}M + L_{\max}M^2\right)\log_2\left(\frac{M}{\varepsilon}\right).$$

*Proof.* Let $g_t \in \partial\ell_t(w_t)$ and observe that

$$\begin{aligned}
\sum_{t=1}^{T} \ell_t(w_t) - \ell_t(u_t) &\leq \sum_{t=1}^{T} \|g_t\| \, \|w_t - u_t\| \\
&\leq \sum_{t=1}^{T} \|g_t\| \left(D_{\max} + \|u_t\|\right) \\
&\leq 2M \sum_{t=1}^{T} \|g_t\| \\
&\leq 2M \left(G_{\max} + L_{\max} D_{\max}\right) T \\
&\leq 2M \left(G_{\max} + L_{\max} M\right) T \\
&\leq 2\left(G_{\max} M + L_{\max} M^2\right) \log_2\left(\frac{M}{\varepsilon}\right),
\end{aligned}$$

where the last line uses $M \geq \varepsilon 2^T \implies T \leq \log_2\left(\frac{M}{\varepsilon}\right)$. □

**Lemma C.1.7.** *For all $t$ let $\ell_t$ be a $(G_t, L_t)$-quadratically bounded convex function for $G_t \in [0, G_{\max}]$ and $L_t \in [0, L_{\max}]$. Let $\varepsilon > 0$, $D_{\min} = \frac{\varepsilon}{T}$, $\eta_{\max} = \frac{1}{KL_{\max}}$, and $\eta_{\min} = \frac{\epsilon}{K(G_{\max}+\epsilon L_{\max})T}$. Let $w_t \in W$ be the outputs of the algorithm characterized in Lemma C.1.3 with $\eta = \eta_{\min}$ and $D = D_{\min}$, and let $\boldsymbol{u} = (u_1, \ldots, u_T)$ be an arbitrary sequence in $W$ with $M = \max_t \|u_t\| \leq D_{\min}$. Then*

$$R_T(\boldsymbol{u}) \leq \left(G_{\max} + \epsilon L_{\max}\right)\left[K(M + P_T) + \epsilon \mathcal{C}_T\right]$$

*where* $\mathcal{C}_T \leq O\left(\frac{\log\left(\log\left(\frac{G_{\max}}{\epsilon L_{\max}}\right)\right)}{T}\right)$.

*Proof.* For $M \leq D_{\min}$, we can apply Lemma C.1.3 with $\tau = (\eta_{\min}, D_{\min})$ to get

$$\begin{aligned}
R_T(\boldsymbol{u}) \leq 2kC_{\mathcal{S}} + 2kD_{\min}G_{\max}\Lambda_T(\tau) + \frac{\|u_T\|^2 + 2D_{\min}P_T + 4kD_{\min}^2\Lambda_T(\tau)}{2\eta_{\min}} \\
+ K\eta_{\min}\sum_{t=1}^{T} L_t\left[\ell_t(u_t) - \ell_t(w_t^{(\tau)})\right] + 4\eta_{\min}\sum_{t=1}^{T}\left\|g_t^{(\tau)}\right\|^2,
\end{aligned}$$

where $\mu_{\widetilde{\tau}} = \frac{1}{2\widetilde{D}(G_{\max}+\widetilde{D}/\widetilde{\eta})}$ for any $\widetilde{\tau} = (\widetilde{\eta}, \widetilde{D}) \in \mathcal{S}$, $k \geq 9/2$, and

$$C_{\mathcal{S}} \overset{\text{def}}{=} \frac{\sum_{\widetilde{\tau}\in\mathcal{S}} \mu_{\widetilde{\tau}}}{\sum_{\widetilde{\tau}\in\mathcal{S}} \mu_{\widetilde{\tau}}^2}, \qquad \Lambda_T(\tau) = \Lambda_T(\eta_{\min}, D_{\min}) \overset{\text{def}}{=} \log\left(\frac{\sum_{\widetilde{\tau}\in\mathcal{S}} \mu_{\widetilde{\tau}}^2}{\mu_{(\eta_{\min},D_{\min})^2}}\right) + 1.$$

Observe that with $M = \max_t \|u_t\| \le D_{\min}$ and $\frac{D_{\min}}{\eta_{\min}} = K(G_{\max} + \epsilon L_{\max})$, we have

$$\begin{aligned}\frac{\|u_T\|^2 + 2D_{\min}P_T + 4kD_{\min}^2\Lambda_T(\tau)}{2\eta_{\min}} &\le \frac{D_{\min}}{\eta_{\min}}\frac{1}{2}\left(\|u_T\| + 2P_T + 4kD_{\min}\Lambda_T(\tau)\right)\\ &= \frac{1}{2}K(G_{\max} + \epsilon L_{\max})\left(\|u_T\| + 2P_T + 4k\frac{\epsilon\Lambda_T(\tau)}{T}\right).\end{aligned}$$

Moreover, by Lemma C.1.5 we have

$$\begin{aligned}\sum_{t=1}^{T} K\eta_{\min}L_t\left[\ell_t(u_t) - \ell_t(w_t^{(\tau)})\right] + 4\eta_{\min}\left\|g_t^{(\tau)}\right\|^2 &\le \eta_{\min}\sum_{t=1}^{T} 3(K+4)\left(G_t^2 + L_t^2D_{\min}^2\right)\\ &\le 3(K+4)\frac{\epsilon\left(G_{\max}^2 + L_{\max}^2D_{\min}^2\right)}{K(G_{\max} + \epsilon L_{\max})}\\ &\le \frac{3(K+4)}{K}\left(\epsilon G_{\max} + \frac{\epsilon^2 L_{\max}}{T^2}\right)\end{aligned}$$

Plugging in the previous two displays back into the full regret bound yields

$$\begin{aligned}R_T(\boldsymbol{u}) &\le 2kC_{\mathcal{S}} + 2k\epsilon G_{\max}\frac{\Lambda_T(\tau)}{T} + \frac{1}{2}K(G_{\max} + \epsilon L_{\max})\left(\|u_T\| + 2P_T + 4k\frac{\epsilon\Lambda_T(\tau)}{T}\right)\\ &\quad + \frac{3(K+4)}{K}\left[\epsilon G_{\max} + \frac{L_{\max}\epsilon^2}{T^2}\right]\\ &\le 2kC_{\mathcal{S}} + \epsilon G_{\max}\left[\frac{3(K+4)}{K} + \frac{(K+1)2k\Lambda_T(\tau)}{T}\right] + \epsilon^2 L_{\max}\left[\frac{3(K+4)}{KT^2} + \frac{2kK\Lambda_T(\tau)}{T}\right]\\ &\quad + K(G_{\max} + \epsilon L_{\max})\left[M + P_T\right].\end{aligned}$$

Finally, Lemma C.1.4 bounds

$$\begin{aligned}2kC_{\mathcal{S}} &\le 2k\cdot 2\sqrt{T}D_{\min}\left(G_{\max} + \frac{D_{\min}}{\eta_{\max}}\right)\\ &\le 4k\sqrt{T}\frac{\epsilon}{T}\left[G_{\max} + \frac{K\epsilon L_{\max}}{T}\right]\\ &\le \frac{4k\left(\epsilon G_{\max} + K\epsilon^2 L_{\max}/T\right)}{\sqrt{T}}\end{aligned}$$

and

$$\begin{aligned}\Lambda_T(\eta_{\min}, D_{\min}) &\le \log\left(6\left|\mathcal{S}_\eta\right|\right) + 1 \le \log\left(\left|\mathcal{S}_\eta\right|\right) + 3\\ &\le \log\left(\left\lceil \log_2\left(\frac{TG_{\max}}{\epsilon L_{\max}}\right)\right\rceil + 1\right) + 3\\ &\le \log\left(\log_2\left(\frac{TG_{\max}}{\epsilon L_{\max}}\right) + 2\right) + 3\end{aligned}$$

Plugging these back in above:

$$\begin{aligned}R_T(\boldsymbol{u}) &\le \epsilon G_{\max}(K+4)\left[\frac{3}{K} + \frac{k}{\sqrt{T}} + \frac{2k\Lambda_T(\eta_{\min}, D_{\min})}{T}\right]\\ &\quad + \epsilon^2 L_{\max}(K+4)\left[\frac{3}{KT^2} + \frac{4k}{T^{3/2}} + \frac{2k\Lambda_T(\eta_{\min}, D_{\min})}{T}\right]\\ &\quad + K(G_{\max} + \epsilon L_{\max})\left[M + P_T\right]\\ &\le \mathcal{C}_T\left(\epsilon G_{\max} + \epsilon^2 L_{\max}\right) + K(G_{\max} + \epsilon L_{\max})\left[M + P_T\right]\\ &= (G_{\max} + \epsilon L_{\max})\left[K(M + P_T) + \epsilon \mathcal{C}_T\right]\end{aligned}$$

where

$$\begin{aligned}\mathcal{C}_T &\le (K+4)\left(\frac{3}{K} + \frac{4k}{\sqrt{T}} + \frac{2k\left(\log\left(\log_2\left(\frac{TG_{\max}}{\epsilon L_{\max}}\right) + 2\right) + 3\right)}{T}\right)\\ &\le O\left(\frac{\log\left(\log\left(\frac{G_{\max}}{\epsilon L_{\max}}\right)\right)}{T}\right)\end{aligned}$$

□

**Proof of Theorem 9.2.1**

**Theorem 9.2.1.** *For all $t$ let $\ell_t : W \to \mathbb{R}$ be a $(G_t, L_t)$-quadratically bounded convex function with $G_t \in [0, G_{\max}]$ and $L_t \in [0, L_{\max}]$. Let $\epsilon > 0$, $K \geq 8$, $\beta_t = 1 - \exp(-1/T)$ for all $t$, and for any $i, j \geq 0$ let $D_j = \frac{\epsilon}{T}\left[2^j \wedge 2^T\right]$ and $\eta_i = \left[\frac{\epsilon 2^i}{K(G_{\max}+\epsilon L_{\max})T} \wedge \frac{1}{KL_{\max}}\right]$, and let $\mathcal{S} = \{(\eta_i, D_j) : i, j \geq 0\}$. For each $\tau = (\eta, D) \in \mathcal{S}$ let $\mu_\tau = \frac{1}{2D(G_{\max}+D/\eta)}$, and set $p_1(\tau) = \frac{\mu_\tau^2}{\sum_{\tilde{\tau}\in\mathcal{S}} \mu_{\tilde{\tau}}^2}$. Then for any $\boldsymbol{u} = (u_1, \ldots, u_T)$ in $W$, Algorithm 11 guarantees*

$$R_T(\boldsymbol{u}) \leq O\left(\left[G_{\max} + (M+\epsilon)L_{\max}\right]\left[(M+\epsilon)\Lambda_T^* + P_T\right] + \sqrt{(M^2\Lambda_T^* + MP_T)\sum_{t=1}^T G_t^2 + L_t^2 M^2,}\right).$$

*where $P_T = \sum_{t=2}^T \|u_t - u_{t-1}\|$, $M = \max_t \|u_t\|$, and $\Lambda_T^* \leq O\left(\log\left(\frac{MT\log(T)}{\epsilon}\right) + \log\left(\log\left(\frac{G_{\max}}{\epsilon L_{\max}}\right)\right)\right)$. Moreover, when the losses are $L_t$-smooth, the bound automatically improves to*

$$R_T(\boldsymbol{u}) \leq O\Bigg(\left[G_{\max} + (M+\epsilon)L_{\max}\right]\left[(M+\epsilon)\Lambda_T^* + P_T\right] + \sqrt{(M^2\Lambda_T^* + MP_T)\left[\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t^*\right] \wedge \sum_{t=1}^T G_t^2 + L_t^2 M^2\right]}\Bigg).$$

*Proof.* First observe that we can assume that there is a $\tau = (\eta, D) \in \mathcal{S}$ for which $D \geq \max_t \|u_t\| = M$, since otherwise using Lemma C.1.6 with $\varepsilon = \frac{\epsilon}{T}$ the regret is bounded as

$$R_T(\boldsymbol{u}) \leq 2M\left(G_{\max} + ML_{\max}\right)\log\left(\frac{MT}{\epsilon}\right). \tag{C.4}$$

Likewise, if $M \leq D_{\min}$ then by Lemma C.1.7 we have

$$R_T(\boldsymbol{u}) \leq (G_{\max} + L_{\max}\epsilon)\left[K(M + P_T) + \epsilon\mathcal{C}_T\right], \tag{C.5}$$

where $\mathcal{C}_T \leq O\left(\frac{\log\left(\log\left(\frac{G_{\max}}{\epsilon L_{\max}}\right)\right)}{T}\right)$. Otherwise, we have $M \in [D_{\min}, D_{\max}]$, in which case there is a $D_j = \frac{\epsilon 2^j}{T}$ for which $D_j \geq M \geq D_{j-1} = \frac{1}{2}D_j$, so for any $\tau = (\eta, D_j) \in \mathcal{S}$ we can apply Lemma C.1.3 to get

$$\begin{aligned}R_T(\boldsymbol{u}) \leq 2kC_{\mathcal{S}} + 2kD_jG_{\max}\Lambda_T(\tau) + \frac{\|u_T\|^2 + 2D_jP_T + 4kD_j^2\Lambda_T(\tau)}{2\eta} \\ + K\eta\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t(w_t^{(\tau)})\right] + 4\eta\sum_{t=1}^T \left\|g_t^{(\tau)}\right\|^2,\end{aligned}$$

where $g_t^{(\tau)} \in \partial \ell_t(w_t^{(\tau)})$, $P_T = \sum_{t=2}^T \|u_t - u_{t-1}\|$, and

$$\begin{aligned}
C_{\mathcal{S}} &= \frac{\sum_{\widetilde{\tau} \in \mathcal{S}} \mu_{\widetilde{\tau}}}{\sum_{\widetilde{\tau} \in \mathcal{S}} \mu_{\widetilde{\tau}}} \\
\Lambda_T(\eta, D_j) &= \log\left(\frac{\sum_{\widetilde{\tau} \in \mathcal{S}} \mu_{\widetilde{\tau}}^2}{\mu_{(\eta, D_j)}^2}\right) + 1 \\
&= \log\left(D_j^2 \left[G_{\max} + \frac{D_j}{\eta}\right]^2 \sum_{\widetilde{\tau} \in \mathcal{S}} \mu_{\widetilde{\tau}}^2\right) + 1 \\
&\leq \log\left((2M)^2 \left[G_{\max} + \frac{2M}{\eta_{\min}}\right]^2 \sum_{\widetilde{\tau} \in \mathcal{S}} \mu_{\widetilde{\tau}}^2\right) + 1 \\
&= \Lambda_T(\eta_{\min}, 2M).
\end{aligned}$$

Thus, bounding $D_j \leq 2M$ and denoting $\Omega_T := \sum_{t=1}^T KL_t \left[\ell_t(u_t) - \ell_t(w_t^{(\tau)})\right] + 4 \left\|g_t^{(\tau)}\right\|^2$, we have:

$$\begin{aligned}
R_T(\boldsymbol{u}) \leq 2kC_{\mathcal{S}} + 4kMG_{\max}\Lambda_T(\eta_{\min}, 2M) + \frac{M^2\left(1 + 16k\Lambda_T(\eta_{\min}, 2M)\right) + 4MP_T}{2\eta} \\
\underbrace{\eta \sum_{t=1}^T \left[KL_t\left[\ell_t(u_t) - \ell_t(w_t^{(\tau)})\right] + 4\left\|g_t^{(\tau)}\right\|^2\right]}_{=:\Omega_T}.
\end{aligned} \tag{C.6}$$

Next, we show that there is an $\eta$ for which the above expression is well-controlled.

Observe that choosing $\eta$ optimally in Equation (C.6) would yield

$$\eta^* = \sqrt{\frac{M^2(1 + 16k\Lambda_T(\eta_{\min}, 2M)) + 4MP_T}{2\Omega_T}}.$$

If $\eta^* \geq \eta_{\max}$, then choosing $\eta = \eta_{\max}$ yields

$$\begin{aligned}
R_T(\boldsymbol{u}) &\leq 2kC_{\mathcal{S}} + 4kMG_{\max}\Lambda_T(\eta_{\min}, 2M) + \frac{M^2\left(1 + 16k\Lambda_T(\eta_{\min}, 2M)\right) + 4MP_T}{2\eta_{\max}} + \eta^*\Omega_T \\
&= 2kC_{\mathcal{S}} + 4kMG_{\max}\Lambda_T(\eta_{\min}, 2M) + \frac{KL_{\max}}{2}\left[M^2(1 + 16k\Lambda_T(\eta_{\min}, 2M)) + 4MP_T\right] \\
&\quad + \sqrt{\frac{1}{2}\left[M^2\left(1 + 16k\Lambda_T(\eta_{\min}, 2M)\right) + 4MP_T\right]\Omega_T}.
\end{aligned} \tag{C.7}$$

Similarly, if $\eta^* \le \eta_{\min}$, then choosing $\eta = \eta_{\min}$ yields

$$\begin{aligned}
R_T(\boldsymbol{u}) &\le 2kC_{\mathcal{S}} + 4kMG_{\max}\Lambda_T(\eta_{\min}, 2M) + \frac{M^2\left(1 + 16k\Lambda_T(\eta_{\min}, 2M)\right) + 4MP_T}{2\eta^*} + \eta_{\min}\Omega_T \\
&= 2kC_{\mathcal{S}} + 4kMG_{\max}\Lambda_T(\eta_{\min}, 2M) + \sqrt{\frac{1}{2}\left[M^2\left(1 + 16k\Lambda_T(\eta_{\min}, 2M)\right) + 4MP_T\right]\Omega_T} \\
&\quad + \frac{\epsilon\Omega_T}{K\left(G_{\max} + \epsilon L_{\max}\right)T}.
\end{aligned}$$

Observe that by Lemma C.1.5, we have

$$\begin{aligned}
\Omega_T &= \sum_{t=1}^{T} KL_t\left[\ell_t(u_t) - \ell_t(w_t^{(\tau)})\right] + 4\left\|g_t^{(\tau)}\right\|^2 \\
&\le \sum_{t=1}^{T} 3(K+4)\left(G_{\max}^2 + L_{\max}^2 D_j^2\right) \\
&\le 3(K+4)\left(G_{\max}^2 + 4M^2 L_{\max}^2\right)T.
\end{aligned}$$

Thus

$$\begin{aligned}
\frac{\epsilon\Omega_T}{K\left(G_{\max} + \epsilon L_{\max}\right)T} &\le \frac{\epsilon \cdot 3(K+4)\left(G_{\max}^2 + 4M^2 L_{\max}^2\right)T}{K\left(G_{\max} + \epsilon L_{\max}\right)T} \\
&\le \frac{3(K+4)}{K}\left(\epsilon G_{\max} + 4M^2 L_{\max}\right) \\
&\le (K+4)\left(\epsilon G_{\max} + 4M^2 L_{\max}\right)
\end{aligned}$$

for $K \ge 3$. so overall when $\eta^* \le \eta_{\min}$ the regret can be bounded as

$$\begin{aligned}
R_T(\boldsymbol{u}) &\le 2kC_{\mathcal{S}} + 4kMG_{\max}\Lambda_T(\eta_{\min}, 2M) + \sqrt{\frac{1}{2}\left[M^2\left(1 + 16k\Lambda_T(\eta_{\min}, 2M)\right) + 4MP_T\right]\Omega_T} \\
&\quad + (K+4)\epsilon G_{\max} + 4(K+4)M^2 L_{\max}. \qquad \text{(C.8)}
\end{aligned}$$

Finally, if $\eta^* \in [\eta_{\min}, \eta_{\max}]$, then there is an $\eta_i = \frac{2^i \epsilon}{K(G_{\max} + \epsilon L_{\max})T}$ such that $\eta_i \le \eta^* \le \eta_{i+1} = 2\eta_i$, so choosing $\eta = \eta_i$ Equation (C.6) is bounded by

$$\begin{aligned}
R_T(\boldsymbol{u}) &\le 2kC_{\mathcal{S}} + 4kMG_{\max}\Lambda_T(\eta_{\min}, 2M) + \frac{M^2\left(1 + 16k\Lambda_T(\eta_{\min}, 2M)\right) + 4MP_T}{\eta^*} + \eta^*\Omega_T \\
&\le 2kC_{\mathcal{S}} + 4kMG_{\max}\Lambda_T(\eta_{\min}, 2M) + 3\sqrt{\frac{1}{2}\left[M^2\left(1 + 16k\Lambda_T(\eta_{\min}, 2M)\right) + 4MP_T\right]\Omega_T}. \qquad \text{(C.9)}
\end{aligned}$$

Now combining Equations (C.4), (C.5) and (C.7) to (C.9), we have

$$
\begin{aligned}
R_T(\boldsymbol{u}) \leq 2M\left(G_{\max} + ML_{\max}\right)\log\left(\frac{MT}{\epsilon}\right) \\
&+ \left(G_{\max} + L_{\max}\epsilon\right)\left[K(M + P_T) + \epsilon\mathcal{C}_T\right] \\
&+ 2kC_{\mathcal{S}} + 4kMG_{\max}\Lambda_T(\eta_{\min}, 2M) \\
&+ 3\sqrt{\frac{1}{2}\left[M^2\left(1 + 16k\Lambda_T(\eta_{\min}, 2M)\right) + 4MP_T\right]\Omega_T} \\
&+ (K + 4)\epsilon G_{\max} + 4(K + 4)M^2L_{\max} \\
&+ \frac{KL_{\max}}{2}\left[M^2(1 + 16k\Lambda_T(\eta_{\min}, 2M)) + 4MP_T\right].
\end{aligned}
$$

From Lemma C.1.4 we have

$$
\begin{aligned}
C_{\mathcal{S}} &\leq 2\sqrt{T}D_{\min}\left(G_{\max} + \frac{D_{\min}}{\eta_{\max}}\right) \\
&\leq \frac{2K\left(\epsilon G_{\max} + \epsilon^2 L_{\max}\right)}{\sqrt{T}} \\
\Lambda_T(\eta_{\min}, 2M)& \\
&\leq \log\left(\frac{6\left|\mathcal{S}\right|(2M)^2}{D_{\min}^2}\right) + 1 \\
&\leq \log\left(\frac{24M^2T^2\left(\left\lceil\log_2\left(\frac{TG_{\max}}{\epsilon L_{\max}}\right)\right\rceil + 1\right)}{\epsilon^2}\right) + 1 \\
&\leq 2\log\left(\frac{5MT}{\epsilon}\right) + \log\left(\log_2\left(\frac{TG_{\max}}{\epsilon L_{\max}}\right) + 2\right) + 1
\end{aligned}
$$

Hence, hiding constants we may write

$$
R_T(\boldsymbol{u}) \leq O\left(G_{\max}((M + \epsilon)\Lambda_T^* + P_T) + L_{\max}\left[(M + \epsilon)^2\Lambda_T^* + (M + \epsilon)P_T\right] + \sqrt{(M^2\Lambda_T^* + MP_T)\Omega_T},\right).
$$

where $\Lambda_T^* \leq O\left(\log\left(\frac{MT}{\epsilon}\right) + \log\left(\log\left(\frac{TG_{\max}}{\epsilon L_{\max}}\right)\right)\right) \leq O\left(\log\left(\frac{MT\log(T)}{\epsilon}\right) + \log\left(\log\left(\frac{G_{\max}}{\epsilon L_{\max}}\right)\right)\right)$. Finally, the proof is completed by observing that if the $\ell_t$ are $L_t$-smooth, then using the self-bounding property we have $\left\|g_t^{(\tau)}\right\|^2 \le 2L_t\left(\ell_t(w_t^{(\tau)}) - \ell_t^*\right)$ for $\ell_t^* = \min_{w\in W}\ell_t(w)$, and thus

$$\begin{aligned}\Omega_T &= \sum_{t=1}^T KL_t\left[\ell_t(u_t) - \ell_t(w_t^{(\tau)})\right] + 4\sum_{t=1}^T \left\|g_t^{(\tau)}\right\|^2\\ &\le \sum_{t=1}^T KL_t\left[\ell_t(u_t) - \ell_t(w_t^{(\tau)})\right] + 8\sum_{t=1}^T L_t\left[\ell_t(w_t^{(\tau)}) - \ell_t^*\right]\\ &\le \sum_{t=1}^T KL_t\left[\ell_t(u_t) - \ell_t^*\right]\end{aligned}$$

where the second-to-last line chooses $K \ge 8$, and simultaneously we have using Lemma C.1.5 that

$$\begin{aligned}\Omega_T &\le 3(K+4)\sum_{t=1}^T\left[G_t^2 + L_t^2 D_j^2\right]\\ &\le 3(K+4)\sum_{t=1}^T\left[G_t^2 + 4L_t^2 M^2\right],\end{aligned}$$

and so we have $\Omega_T \le O\left(\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t^*\right] \wedge \sum_{t=1}^T G_t^2 + L_t^2 M^2\right)$. □

### Proof of Theorem 9.2.3

**Theorem 9.2.3.** *For all $t$ let $\ell_t : W \to \mathbb{R}$ be $(G_t, L_t)$-quadratically bounded and $L_t$-smooth convex function with $G_t \in [0, G_{\max}]$ and $L_t \in [0, L_{\max}]$. Let $\epsilon > 0$, $K \ge 8$, and for any $i, j \ge 0$ let $D_j = \frac{\epsilon}{\sqrt{T}}\left[2^j \wedge 2^T\right]$ and $\eta_i = \frac{1}{KL_{\max}\sqrt{T}}\left[2^i \wedge \sqrt{T}\right]$, and let $\mathcal{S} = \{(\eta_i, D_j) : i, j \ge 0\}$. Then for any $\boldsymbol{u} = (u_1, \ldots, u_T)$ in $W$, Algorithm 11 guarantees*

$$\begin{aligned}R_T(\boldsymbol{u}) \le O\Bigg(& G_{\max}(M+\epsilon)\Lambda_T^* + L_{\max}(M+\epsilon)^2\Lambda_T^* + L_{\max}(M+\epsilon)P_T\\ &+ \sqrt{\sum_{t=1}^T\left[\ell_t(u_t) - \ell_t^*\right]^2} + \sqrt{(M^2\Lambda_T^* + MP_T)\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t^*\right]}\Bigg),\end{aligned}$$

*where $M = \max_t \|u_t\|$, $P_T = \sum_{t=2}^T \|u_t - u_{t-1}\|$, and $\Lambda_T^* \le O\left(\log\left(\frac{M\sqrt{T}\log\left(\sqrt{T}\right)}{\epsilon}\right)\right)$.*

*Proof.* By Lemma C.1.6, we can assume that there is a $\tau = (\eta, D) \in \mathcal{S}$ for which $D \ge \max_t \|u_t\| = M$, since otherwise the regret is bounded as

$$R_T(\boldsymbol{u}) \le 2M\left(G_{\max} + ML_{\max}\right)\log\left(\frac{M\sqrt{T}}{\epsilon}\right). \tag{C.10}$$

Hence, we can assume there is a $(\eta, D) \in \mathcal{S}$ which has $M \le D$. For any such $(\eta, D) \in \mathcal{S}$, we can apply Lemma C.1.3 to get

$$\begin{aligned}R_T(\boldsymbol{u}) &\le 2kC_{\mathcal{S}} + 2kDG_{\max}\Lambda_T(\tau) + \frac{\|u_T\|^2 + 2DP_T + 4kD^2\Lambda_T(\tau)}{2\eta}\\&\quad + K\eta\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t(w_t^{(\tau)})\right] + 4\eta\sum_{t=1}^T \left\|g_t^{(\tau)}\right\|^2,\end{aligned}$$

where $g_t^{(\tau)} \in \partial\ell_t(w_t^{(\tau)})$, $P_T = \sum_{t=2}^T \|u_t - u_{t-1}\|$ and

$$C_{\mathcal{S}} \overset{\text{def}}{=} \frac{\sum_{\widetilde{\tau}\in\mathcal{S}}\mu_{\widetilde{\tau}}}{\sum_{\widetilde{\tau}\in\mathcal{S}}\mu_{\widetilde{\tau}}^2}, \qquad \Lambda_T(\tau) \overset{\text{def}}{=} \log\left(\frac{\sum_{\widetilde{\tau}\in\mathcal{S}}\mu_{\widetilde{\tau}}^2}{\mu_\tau^2}\right) + 1,$$

where for any $\widetilde{\tau} = (\widetilde{D},\widetilde{\eta}) \in \mathcal{S}$ we define $\mu_{(\widetilde{\eta},\widetilde{D})} = \frac{1}{2\widetilde{D}(G_{\max}+\widetilde{D}/\widetilde{\eta})}$. Using the self-bounding property of smooth functions, for any $g_t^{(\tau)} \in \partial\ell_t(w_t^{(\tau)})$ we have $\left\|g_t^{(\tau)}\right\|^2 \le 2L_t\left[\ell_t(w_t^{(\tau)}) - \ell_t^*\right]$ for $\ell_t^* = \arg\min_{w\in W}\ell_t(w)$, so the last line is bound as

$$K\eta\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t(w_t^{(\tau)})\right] + 8\eta\sum_{t=1}^T L_t\left[\ell_t(w_t^{(\tau)}) - \ell_t^*\right] \le K\eta\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t^*\right]$$

for $K \ge 8$. Hence,

$$\begin{aligned}R_T(\boldsymbol{u}) &\le 2kC_{\mathcal{S}} + 2kDG_{\max}\Lambda_T(\tau) + \frac{\|u_T\|^2 + 2DP_T + 4kD^2\Lambda_T(\tau)}{2\eta}\\&\quad + K\eta\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t^*\right]\end{aligned} \tag{C.11}$$

Now suppose that $M \le D_{\min}$, then choosing $\tau = \tau_{\min} = (\eta_{\min}, D_{\min})$ we would have

$$\begin{aligned}R_T(\boldsymbol{u}) &\le 2kC_{\mathcal{S}} + 2kD_{\min}G_{\max}\Lambda_T(\tau_{\min}) + \frac{\|u_T\|^2 + 2D_{\min}P_T + 4kD_{\min}^2\Lambda_T(\tau_{\min})}{2\eta_{\min}}\\&\quad + K\eta_{\min}\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t^*\right]\\&\le 2kC_{\mathcal{S}} + 2kD_{\min}G_{\max}\Lambda_T(\tau_{\min}) + \frac{D_{\min}}{\eta_{\min}}\frac{1}{2}\left(M + 2P_T + 4kD_{\min}\Lambda_T(\tau_{\min})\right)\\&\quad + \frac{1}{\sqrt{T}L_{\max}}\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t^*\right]\\&\le 2kC_{\mathcal{S}} + 2kG_{\max}\frac{\epsilon\Lambda_T(\tau_{\min})}{\sqrt{T}} + K\epsilon L_{\max}\left(M + P_T + 2k\frac{\epsilon\Lambda_T(\tau_{\min})}{\sqrt{T}}\right)\\&\quad + \sqrt{\sum_{t=1}^T\left[\ell_t(u_t) - \ell_t^*\right]^2}\end{aligned} \tag{C.12}$$

where the last line applies Cauchy-Schwarz inequality, observes that $D_{\min}/\eta_{\min} = K\epsilon L_{\max}$, and recalls $D_{\min} = \frac{\epsilon}{\sqrt{T}}$. Finally, assume that $M \in [D_{\min}, D_{\max}]$, then there is a $D_j = \frac{\epsilon 2^j}{\sqrt{T}}$ for which $D_j \geq M \geq D_{j-1} = \frac{1}{2}D_j$. Then, choosing $\tau = (\eta, D_j)$, Equation (C.11) yields

$$\begin{aligned} R_T(\boldsymbol{u}) &\leq 2kC_{\mathcal{S}} + 4kMG_{\max}\Lambda_T(\eta_{\min}, 2M) + \frac{M^2 + 4MP_T + 16kM^2\Lambda_T(\eta_{\min}, 2M)}{2\eta} \\ &\quad + K\eta\sum_{t=1}^{T} L_t\left[\ell_t(u_t) - \ell_t^*\right] \end{aligned} \tag{C.13}$$

where we've observed that

$$\begin{aligned} \Lambda(\eta, D_j) &= \log\left(\frac{\sum_{\widetilde{\tau}\in\mathcal{S}} \mu_{\widetilde{\tau}}^2}{\mu_{(\eta, D_j)}}\right) + 1 \\ &= \log\left(\sum_{\widetilde{\tau}\in\mathcal{S}} \mu_{\widetilde{\tau}}^2 D_j^2\left[G_{\max} + D_j/\eta\right]^2\right) + 1 \\ &\leq \log\left(\sum_{\widetilde{\tau}\in\mathcal{S}} \mu_{\widetilde{\tau}}^2 (2M)^2\left[G_{\max} + 2M/\eta\right]^2\right) + 1 \\ &= \Lambda_T(\eta, 2M) \end{aligned}$$

so it remains to show that there is an $\eta$ that favorably balances the last two terms of Equation (C.13).

Observe that the optimal choice for $\eta$ would be

$$\eta^* = \sqrt{\frac{M^2(1 + 16k\Lambda_T(\eta_{\min}, 2M)) + 4MP_T}{2K\sum_{t=1}^{T} L_t\left[\ell_t(u_t) - \ell_t^*\right]}}.$$

If $\eta^* \leq \eta_{\min}$ then choosing $\eta = \eta_{\min}$ we have

$$\begin{aligned} R_T(\boldsymbol{u}) &\leq 2kC_{\mathcal{S}} + 4kMG_{\max}\Lambda_T(\eta_{\min}, 2M) + \frac{M^2(1 + 16k\Lambda_T(\eta_{\min}, 2M) + 4MP_T}{2\eta^*} \\ &\quad + K\eta_{\min}\sum_{t=1}^{T} L_t\left[\ell_t(u_t) - \ell_t^*\right] \\ &\leq 2kC_{\mathcal{S}} + 4kMG_{\max}\Lambda_T(\eta_{\min}, 2M) + \sqrt{\frac{K}{2}\left(M^2(1 + 16k\Lambda_T(\eta_{\min}, 2M)) + 4MP_T\right)\Omega_T} \\ &\quad + \sqrt{\sum_{t=1}^{T}\left[\ell_t(u_t) - \ell_t^*\right]^2}, \end{aligned} \tag{C.14}$$

where the last line defines the short-hand notation $\Omega_T = \sum_{t=1}^{T} L_t\left[\ell_t(u_t) - \ell_t^*\right]$ and uses Cauchy-Schwarz inequality to bound $K\eta_{\min}\sum_{t=1}^{T} L_t\left[\ell_t(u_t) - \ell_t^*\right] \leq \sqrt{\sum_{t=1}^{T}\left[\ell_t(u_t) - \ell_t^*\right]^2}$. Likewise, if $\eta^* \geq$

$\eta_{\max}$ then by choosing $\eta = \eta_{\max}$ we have via Equation (C.13) that

$$\begin{aligned}
R_T(\boldsymbol{u}) &\le 2kC_{\mathcal{S}} + 4kMG_{\max}\Lambda_T(\eta_{\min}, 2M) + \frac{M^2 + 4MP_T + 16kM^2\Lambda_T(\eta_{\min}, 2M)}{2\eta_{\max}} \\
&\quad + K\eta^* \sum_{t=1}^{T} L_t\left[\ell_t(u_t) - \ell_t^*\right] \\
&= 2kC_{\mathcal{S}} + 4kMG_{\max}\Lambda_T(\eta_{\min}, 2M) + KL_{\max}\left(M^2(1 + 8k\Lambda_T(\eta_{\min}, 2M)) + 2MP_T\right) \\
&\quad + \sqrt{\frac{K}{2}\left(M^2(1 + 16k\Lambda_T(\eta_{\min}, 2M)) + 4MP_T\right)\Omega_T} \qquad \text{(C.15)}
\end{aligned}$$

Finally, if $\eta^* \in [\eta_{\min}, \eta_{\max}]$ then there is an $\eta_i = \frac{2^i}{\epsilon L_{\max}\sqrt{T}}$ for which $\eta_i \le \eta^* \le \eta_{i+1} = 2\eta_i$, so Equation (C.13) is gives us

$$R_T(\boldsymbol{u}) \le 2kC_{\mathcal{S}} + 4kMG_{\max}\Lambda_T(\eta_{\min}, 2M) + 3\sqrt{\frac{K}{2}\left(M^2(1 + 16k\Lambda_T(\eta_{\min}, 2M)) + 4MP_T\right)\Omega_T} \qquad \text{(C.16)}$$

Finally, combining Equations (C.10), (C.12) and (C.14) to (C.16), we have

$$\begin{aligned}
R_T(\boldsymbol{u}) \le 2kC_{\mathcal{S}} &+ 4kG_{\max}\left[M\Lambda_T(\eta_{\min}, 2M) + \frac{\epsilon\Lambda_T(\tau_{\min})}{\sqrt{T}}\right] \\
&+ 2M\left(G_{\max} + ML_{\max}\right)\log\left(\frac{M\sqrt{T}}{\epsilon}\right) \\
&+ K\epsilon L_{\max}\left(M + P_T + 2k\frac{\epsilon\Lambda_T(\tau_{\min})}{\sqrt{T}}\right) \\
&+ KL_{\max}\left(M^2(1 + 8k\Lambda_T(\eta_{\min}, 2M)) + 2MP_T\right) \\
&+ \sqrt{\sum_{t=1}^{T}\left[\ell_t(u_t) - \ell_t^*\right]^2} \\
&+ 3\sqrt{\frac{K}{2}\left(M^2(1 + 16k\Lambda_T(\eta_{\min}, 2M)) + 4MP_T\right)\Omega_T}
\end{aligned}$$

Lastly, note that by Lemma C.1.4, we have

$$\begin{aligned}
C_{\mathcal{S}} &\leq 2\sqrt{T}D_{\min}\left(G_{\max} + \frac{D_{\min}}{\eta_{\max}}\right)\\
&= 2\epsilon G_{\max} + \frac{2K\epsilon^2 L_{\max}}{\sqrt{T}}\\
\Lambda_T(\tau) &\leq \log\left(\frac{24\eta_{\max}^2 D^4}{\eta_{\min}^2 D_{\min}^4} \wedge \frac{6\left|\mathcal{S}_\eta\right| D^2}{D_{\min}^2}\right) + 1\\
&\leq \log\left(\frac{6\left|\mathcal{S}_\eta\right| D^2}{D_{\min}^2}\right) + 1
\end{aligned}$$

so we have the following bounds:

$$\begin{aligned}
\Lambda_T(\eta_{\min}, D_{\min}) &\leq \log\left(6\log_2\left(\lceil\log_2(\sqrt{T})\rceil + 1\right)\right) \leq O\left(\log(\log(\sqrt{T}))\right)\\
\Lambda_T(\eta_{\min}, 2M) &\leq \log\left(\frac{24TM^2\log_2\left(\lceil\log_2(\sqrt{T})\rceil + 1\right)}{\epsilon^2}\right) \leq O\left(\log\left(M\sqrt{T}\log\left(\sqrt{T}\right)/\epsilon\right)\right).
\end{aligned}$$

Overall, the dynamic regret is bounded as

$$\begin{aligned}
R_T(\boldsymbol{u}) \leq O\Bigg(& G_{\max}(M+\epsilon)\Lambda_T^*\\
&+ L_{\max}(M+\epsilon)^2\Lambda_T^* + L_{\max}(M+\epsilon)P_T\\
&+ \sqrt{\sum_{t=1}^T\left[\ell_t(u_t) - \ell_t^*\right]^2}\\
&+ \sqrt{\left(M^2\Lambda_T^* + MP_T\right)\Omega_T}\Bigg)
\end{aligned}$$

where $\Lambda_T^* \leq O\left(\log\left(\frac{M\sqrt{T}\log\left(\sqrt{T}\right)}{\epsilon}\right) + \log\left(\log\left(\sqrt{T}\right)\right)\right) \leq O\left(\log\left(\frac{M\sqrt{T}\log\left(\sqrt{T}\right)}{\epsilon}\right)\right)$. □

### Proof of Theorem 9.2.2

We focus on the case where $G/L \leq M$, since otherwise when $G/L \geq M$ the loss function $\ell_t(w) = (\frac{1}{2}G + \frac{1}{2}LM)\xi_t w$ for $\xi_t \in \{-1,1\}$ satisfies $|\ell_t'(w)| = \frac{1}{2}(G + LM) \leq G$ for any $w \in W$, so $\ell_t$ is $G$-Lipschitz. Hence, existing lower bounds tell us that there exists a sequence $\xi_t \in [-1,1]$ such that $R_T(\boldsymbol{u}) \geq \Omega\big(G\sqrt{MP_T T}\big) \geq \Omega\big(\frac{1}{2}(G+LM)\sqrt{MP_T T}\big) = \Omega\big(\frac{1}{2}G\sqrt{MP_T T} + \frac{1}{2}LM^{3/2}\sqrt{P_T T}\big)$ where $M = \max_t \|u_t\|$ and $P_T = \sum_{t=2}^T \|u_t - u_{t-1}\|$ (L. Zhang, S. Lu, and Z.-H. Zhou 2018).

**Theorem 9.2.2.** *For any $M > 0$ there is a sequence of $(G, L)$-quadratically bounded functions with $\frac{G}{L} \le M$ such that for any $\gamma \in [0, \frac{1}{2}]$,*

$$R_T(\boldsymbol{u}) \ge \frac{G}{4} M^{1-\gamma} \left[P_T T\right]^{\gamma} + \frac{L}{8} M^{2-\gamma} \left[P_T T\right]^{\gamma}.$$

*where $P_T = \sum_{t=2}^{T} \|u_t - u_{t-1}\|$ and $M \ge \max_t \|u_t\|$.*

*Proof.* On each round $t$, we can always find a $u_t$ such that $u_t \perp w_t$. Let $\|u_t\| := \sigma \le M$ for some $\sigma$ to be decided. Let $G > 0$, $L \ge 0$ such that $G/L \le \sigma$, let $\xi_t = \frac{u_t}{\|u_t\|}$, and on each round set

$$\ell_t(w) = -\frac{1}{2} G \langle \xi_t, w \rangle + \frac{L}{4} (\sigma - \langle \xi_t, w \rangle)^2.$$

Observe that these losses are $(\widetilde{G}, \widetilde{L})$ quadratically bounded with $\widetilde{G} = \frac{1}{2} G + \frac{1}{2} \sigma L$ and $\widetilde{L} = L$, and $\widetilde{G}/\widetilde{L} \le \sigma \le M$ as required. Since $w_t \perp \xi_t$ and $\langle \xi_t, u_t \rangle = \|u_t\| = \sigma$, we have

$$R_T(\boldsymbol{u}) = \sum_{t=1}^{T} \ell_t(w_t) - \ell_t(u_t) \ge \frac{1}{2} G \sigma T + \frac{L}{4} T \sigma^2.$$

Note also that the path-length of this comparator sequence is bounded as

$$P_T = \sum_{t=2}^{T} \|u_t - u_{t-1}\| \le 2\sigma T.$$

Now for $\mu \in [0, 1/2]$ set $\sigma = M T^{-\mu}$, then the path-length is bounded as

$$P_T \le 2 M T^{1-\mu}$$

and the regret is bounded below by

$$\frac{1}{2} G M T^{1-\mu} + \frac{L}{4} T^{1-2\mu} M^2.$$

Now set $\gamma = \frac{1-2\mu}{2-\mu} \in [0, \frac{1}{2}]$ and consider the second term:

$$\begin{aligned}
\frac{L}{4} T^{1-2\mu} M^2 &= \frac{L}{4} (M T^{1-\mu})^{\gamma} (M T^{1-\mu})^{1-\gamma} T^{-\mu} M \\
&\ge \frac{L}{4 \cdot 2^{\gamma}} (P_T)^{\gamma} (M T^{1-\mu})^{1-\gamma} T^{-\mu} M \\
&= \frac{L}{8} M^{2-\gamma} P_T^{\gamma} T^{(1-\mu)(1-\gamma)-\mu} \\
&= \frac{L}{8} M^{2-\gamma} \left[P_T T\right]^{\gamma}
\end{aligned}$$

where the last line observes $\gamma = \frac{1-2\mu}{2-\mu} \in [0, \frac{1}{2}]$, so that $(1-\mu)(1-\gamma) - \mu = \gamma$. Similarly,

$$\begin{aligned}\frac{1}{2}GMT^{1-\mu} &= \frac{1}{2}G(MT^{1-\mu})^{\gamma}(MT^{1-\mu})^{1-\gamma} \geq \frac{1}{2\cdot 2^{\gamma}}GM^{1-\gamma}(P_T)^{\gamma}T^{(1-\mu)(1-\gamma)}\\ &\geq \frac{1}{4}GM^{1-\gamma}(P_TT)^{\gamma},\end{aligned}$$

so we have

$$R_T(\boldsymbol{u}) \geq \frac{G}{4}M^{1-\gamma}\left[P_TT\right]^{\gamma} + \frac{L}{8}M^{2-\gamma}\left[P_TT\right]^{\gamma}.$$

□

## C.2 Details for Chapter 10

### C.2.1 Proofs for Section 10.2 (Dynamic Regret via Discounting)

**Equivalence to FTRL and Mirror Descent**

We accomplish our analysis of the discounted VAW forecaster using the equivalence in the following proposition, proving both optimistic FTRL and and optimistic mirror descent interpretations of the discounted VAW forecaster. Equation (C.18) is perhaps the most natural interpretation of the update: it says that the discounted VAW forecaster chooses the $w$ which minimizes the *discounted sum* $h_t(w) + \gamma\ell_{t-1}(w) + \gamma^2\ell_{t-2}(w) + \ldots$, thus placing greater emphasis on the most-recent losses and the hint function $h_t(w)$. However, it is not at all obvious how to analyze the dynamic regret of the discounted VAW forecaster when interpreted in this FTRL-like form. Rather, the key to our results in this work is to instead approach the analysis through the lens of the mirror descent update (Equation (C.19)). Interestingly, a similar mirror descent interpretation was used in the seminal work of Azoury and Manfred K Warmuth (2001), though they did not account for an arbitrary $\widetilde{y}_t$ and they did not refer to the algorithm in terms of mirror descent.

**Proposition C.2.1.** *(Discounted VAW Forecaster) Let $\gamma \in (0,1]$, $\lambda > 0$, $\widetilde{y}_1 = 0$, and $\widetilde{y}_t \in \mathbb{R}$ for $t > 1$. Define $h_t(w) = \frac{1}{2}(\widetilde{y}_t - \langle x_t, w\rangle)^2$ and $\ell_0(w) = \frac{\lambda}{2}\|w\|_2^2$. Recursively define $\Sigma_t = x_tx_t^\top + \gamma\Sigma_{t-1}$ starting from $\Sigma_0 = \lambda I$, let $\psi_t(w) = \frac{1}{2}\|w\|_{\Sigma_t}^2$ and set $w_1 = \arg\min_{w\in\mathbb{R}^d}\psi_1(w) = \mathbf{0}$. Then the following are equivalent*

$$\Sigma_t^{-1}\Big[\widetilde{y}_tx_t + \gamma\sum_{s=1}^{t-1}\gamma^{t-1-s}y_sx_s\Big] \tag{C.17}$$

$$\arg\min_{w\in\mathbb{R}^d} h_t(w) + \gamma\sum_{s=0}^{t-1}\gamma^{t-1-s}\ell_s(s) \tag{C.18}$$

$$\arg\min_{w\in\mathbb{R}^d}(\gamma\ell_{t-1} - \gamma h_{t-1} + h_t)(w) + \gamma D_{\psi_{t-1}}(w|w_{t-1}) \tag{C.19}$$

*Remark* C.2.2. Note that with $\gamma = 0$, Equations (C.18) and (C.19) prescribe choosing any $w_t$ satisfying $\langle w_t, x_t\rangle = \widetilde{y}_t$. The choice is not unique, but nevertheless it will often be convenient to refer to an algorithm which greedily predicts $\widetilde{y}_t$ on each round as an instance of Algorithm 12 with $\gamma = 0$.

*Proof.* The result follows by showing that Equations (C.18) and (C.19) are both equivalent to Equation (C.17). First consider the former, Equation (C.18). From the first-order optimality condition we have

$$\begin{aligned}\mathbf{0} &= \nabla h_t(w_t) + \gamma \sum_{s=0}^{t-1} \gamma^{t-1-s}\nabla \ell_s(w_t)\\ &= -(\widetilde{y}_t - \langle x_t, w_t\rangle)x_t - \gamma\sum_{s=1}^{t-1}\gamma^{t-1-s}(y_s - \langle x_s, w_t\rangle)x_s + \gamma^t\lambda w_t,\end{aligned}$$

where the last line recalls that we defined $\ell_0(w) = \frac{\lambda}{2}\|w\|_2^2$. Hence,

$$\begin{aligned}\left(\gamma^t\lambda I + \sum_{s=1}^{t}\gamma^{t-s}x_s x_s^\top\right)w_t &= \widetilde{y}_t x_t + \sum_{s=1}^{t}\gamma^{t-s}y_s x_s\\ \implies w_t &= \left(\gamma^t\lambda I + \sum_{s=1}^{t}\gamma^{t-s}x_s x_s^\top\right)^{-1}\left[\widetilde{y}_t x_t + \gamma\sum_{s=1}^{t-1}\gamma^{t-1-s}y_s x_s\right]\\ &= \Sigma_t^{-1}\left[\widetilde{y}_t x_t + \gamma\sum_{s=1}^{t-1}\gamma^{t-1-s}y_s x_s\right],\end{aligned}$$

where the last line can be seen by unrolling the recursion for $\Sigma_t$.

Likewise, consider Equation (C.19). From the first-order optimality condition $w_t = \arg\min_{w\in\mathbb{R}^d}(\gamma\ell_{t-1} - \gamma h_{t-1} + h_t)(w) + \gamma D_{\psi_{t-1}}(w|w_t)$, we have

$$\begin{aligned}\mathbf{0} &= \gamma(\nabla\ell_{t-1}(w_t) - \nabla h_{t-1}(w_t)) + \nabla h_t(w_t) + \gamma\left[\nabla\psi_{t-1}(w_t) - \nabla\psi_{t-1}(w_{t-1})\right]\\ &= -\gamma y_{t-1}x_{t-1} + \gamma\widetilde{y}_{t-1}x_{t-1} - \widetilde{y}_t x_t + x_t x_t^\top w_t + \gamma\Sigma_{t-1}w_t - \gamma\Sigma_{t-1}w_{t-1}\\ &= -\gamma y_{t-1}x_{t-1} + \gamma\widetilde{y}_{t-1}x_{t-1} - \widetilde{y}_t x_t + \Sigma_t w_t - \gamma\Sigma_{t-1}w_{t-1},\end{aligned}$$

where the last line observes that $\Sigma_t = x_t x_t^\top + \gamma\Sigma_{t-1}$ by construction. Hence, re-arranging we have

$$\Sigma_t w_t = \widetilde{y}_t x_t + \gamma y_{t-1}x_{t-1} - \gamma\widetilde{y}_{t-1}x_{t-1} + \gamma\Sigma_{t-1}w_{t-1}$$

and unrolling the recursion:

$$\begin{aligned}
&= \widetilde{y}_t x_t + \gamma y_{t-1} x_{t-1} - \gamma \widetilde{y}_{t-1} x_{t-1} + \gamma \left[\widetilde{y}_{t-1} x_{t-1} + \gamma y_{t-2} x_{t-2} - \gamma \widetilde{y}_{t-2} x_t + \gamma \Sigma_{t-2} w_{t-2}\right] \\
&= \widetilde{y}_t x_t + \gamma y_{t-1} x_{t-1} + \gamma^2 y_{t-2} x_{t-2} - \gamma^2 \widetilde{y}_{t-2} x_{t-2} + \gamma^2 \Sigma_{t-2} w_{t-2} \\
&= \dots \\
&= \widetilde{y}_t x_t - \gamma^{t-1} \widetilde{y}_1 x_1 + \gamma \sum_{s=1}^{t-1} \gamma^{t-1-s} y_s x_s \\
&= \widetilde{y}_t x_t + \gamma \sum_{s=1}^{t-1} \gamma^{t-1-s} y_s x_s,
\end{aligned}$$

for $\widetilde{y}_1 = 0$. Hence, applying $\Sigma_t^{-1}$ to both sides we have

$$w_t = \Sigma_t^{-1} \left[ \widetilde{y}_t x_t + \gamma \sum_{s=1}^{t-1} \gamma^{t-1-s} y_s x_s \right]$$

□

### Proof of Theorem 10.2.1

**Theorem 10.2.1.** *Let $\lambda > 0$ and $\gamma \in (0, 1]$. Then for any sequence $\boldsymbol{u} = (u_1, \dots, u_T)$ in $\mathbb{R}^d$, Algorithm 12 guarantees*

$$\begin{aligned}
R_T(\boldsymbol{u}) \le \frac{\gamma\lambda}{2} \|u_1\|_2^2 &+ \frac{d}{2} \max_t (y_t - \widetilde{y}_t)^2 \log\left(1 + \frac{\sum_{t=1}^T \gamma^{T-t} \|x_t\|_2^2}{\lambda d}\right) \\
&+ \gamma \sum_{t=1}^{T-1} \left[F_t^\gamma(u_{t+1}) - F_t^\gamma(u_t)\right] + \frac{d}{2} \log(1/\gamma) \sum_{t=1}^{T} (y_t - \widetilde{y}_t)^2
\end{aligned}$$

*where $F_t^\gamma(w) = \gamma^t \frac{\lambda}{2} \|w\|_2^2 + \sum_{s=1}^t \gamma^{t-s} \ell_s(w)$.*

*Proof.* Begin by applying the regret template provided by Lemma C.2.3:

$$R_T(\boldsymbol{u}) \le \sum_{t=1}^T D_{\psi_t}(u_t | w_t) - D_{\psi_{t+1}}(u_t | w_{t+1}) + \sum_{t=1}^T h_{t+1}(u_t) - h_t(u_t) + \frac{1}{2} \sum_{t=1}^T (y_t - \widetilde{y}_t)^2 \|x_t\|_{\Sigma_t^{-1}}^2,$$

bound the first two summations using Lemma C.2.4:

$$\le \frac{\gamma\lambda}{2} \|u_1\|_2^2 + h_{T+1}(u_T) + \gamma \sum_{t=1}^{T-1} \left[F_t^\gamma(u_{t+1}) - F_t^\gamma(u_t)\right] + \frac{1}{2} \sum_{t=1}^T (y_t - \widetilde{y}_t)^2 \|x_t\|_{\Sigma_t^{-1}}^2,$$

and apply a discounted variant of the log-determinant lemma (Lemma C.2.15) to bound the final summation:

$$\le \frac{\gamma\lambda}{2}\left\|u_1\right\|_2^2 + h_{T+1}(u_T) + \frac{d}{2}\max_t (y_t - \widetilde{y}_t)^2 \log\left(1 + \frac{\sum_{t=1}^T \gamma^{T-t}\left\|x_t\right\|_2^2}{\lambda d}\right)$$

$$+ \gamma \sum_{t=1}^{T-1}\left[F_t^\gamma(u_{t+1}) - F_t^\gamma(u_t)\right] + \frac{d}{2}\log\left(1/\gamma\right)\sum_{t=1}^T (y_t - \widetilde{y}_t)^2$$

Finally, since the regret does not depend on $h_{T+1}(\cdot)$ we may set $h_{T+1}(\cdot) \equiv 0$ in the analysis and hide constants to arrive at the stated bound. □

### Proof of Lemma C.2.3

The following lemma provides the base regret decomposition that we use as a jumping-off point to prove Theorem 10.2.1. The result follows using mostly standard mirror descent analysis, though with a bit of additional care to handle issues related to the discounted regularizer.

**Lemma C.2.3.** *Let $\gamma \in (0, 1]$. Then for any sequence $\boldsymbol{u} = (u_1, \ldots, u_T)$ in $\mathbb{R}^d$, Algorithm 12 guarantees*

$$R_T(\boldsymbol{u}) \le \sum_{t=1}^T D_{\psi_t}(u_t|w_t) - D_{\psi_{t+1}}(u_t|w_{t+1})$$

$$+ \sum_{t=1}^T h_{t+1}(u_t) - h_t(u_t)$$

$$+ \sum_{t=1}^T \frac{1}{2}(y_t - \widetilde{y}_t)^2 \left\|x_t\right\|_{\Sigma_t^{-1}}^2$$

*Proof.* We will proceed following a mirror-descent-based analysis, and thus begin by exposing the terms $(\gamma\ell_t - \gamma h_t + h_{t+1})(w_{t+1})$ observed in the mirror-descent interpretation of the update (Equation (C.19)):

$$\begin{aligned}
R_T(\boldsymbol{u}) &= \sum_{t=1}^{T} \ell_t(w_t) - \ell_t(u_t) \\
&= \sum_{t=1}^{T} \gamma \left[\ell_t(w_t) - \ell_t(u_t)\right] + (1-\gamma) \sum_{t=1}^{T} \ell_t(w_t) - \ell_t(u_t) \\
&= \sum_{t=1}^{T} \gamma \left[(\ell_t - h_t)(w_t) - (\ell_t - h_t)(u_t)\right] + \sum_{t=1}^{T} \gamma h_t(w_t) - \gamma h_t(u_t) \\
&\qquad + (1-\gamma) \sum_{t=1}^{T} \ell_t(w_t) - \ell_t(u_t) \\
&= \sum_{t=1}^{T} \gamma \left[(\ell_t - h_t)(w_{t+1}) - (\ell_t - h_t)(u_t)\right] + \sum_{t=1}^{T} \gamma h_t(w_t) - \gamma h_t(u_t) \\
&\qquad + \gamma \sum_{t=1}^{T} (\ell_t - h_t)(w_t) - (\ell_t - h_t)(w_{t+1}) \\
&\qquad + (1-\gamma) \sum_{t=1}^{T} \ell_t(w_t) - \ell_t(u_t) \\
&= \sum_{t=1}^{T} (\gamma \ell_t - \gamma h_t + h_{t+1})(w_{t+1}) - (\gamma \ell_t - \gamma h_t + h_{t+1})(u_t) \\
&\qquad + \sum_{t=1}^{T} \gamma h_t(w_t) - h_{t+1}(w_{t+1}) + \sum_{t=1}^{T} h_{t+1}(u_t) - \gamma h_t(u_t) \\
&\qquad + \gamma \sum_{t=1}^{T} (\ell_t - h_t)(w_t) - (\ell_t - h_t)(w_{t+1}) \\
&\qquad + (1-\gamma) \sum_{t=1}^{T} \ell_t(w_t) - \ell_t(u_t)
\end{aligned}$$

Re-arranging factors of $\gamma$ from the second-line and observing that $\sum_{t=1}^{T} h_t(w_t) - h_{t+1}(w_{t+1}) = h_1(w_1) - h_{T+1}(w_{T+1})$:

$$\begin{aligned}
&= \sum_{t=1}^{T} (\gamma\ell_t - \gamma h_t + h_{t+1})(w_{t+1}) - (\gamma\ell_t - \gamma h_t + h_{t+1})(u_t) \\
&\quad + \sum_{t=1}^{T} h_t(w_t) - h_{t+1}(w_{t+1}) + \sum_{t=1}^{T} -(1-\gamma)h_t(w_t) + (1-\gamma)h_t(u_t) + \sum_{t=1}^{T} h_{t+1}(u_t) - h_t(u_t) \\
&\quad + \gamma \sum_{t=1}^{T} (\ell_t - h_t)(w_t) - (\ell_t - h_t)(w_{t+1}) \\
&\quad + (1-\gamma) \sum_{t=1}^{T} \ell_t(w_t) - \ell_t(u_t) \\
&= \sum_{t=1}^{T} (\gamma\ell_t - \gamma h_t + h_{t+1})(w_{t+1}) - (\gamma\ell_t - \gamma h_t + h_{t+1})(u_t) \\
&\quad + h_1(w_1) - h_{T+1}(w_{T+1}) + \sum_{t=1}^{T} h_{t+1}(u_t) - h_t(u_t) \\
&\quad + \gamma \sum_{t=1}^{T} (\ell_t - h_t)(w_t) - (\ell_t - h_t)(w_{t+1}) \\
&\quad + (1-\gamma) \sum_{t=1}^{T} (\ell_t - h_t)(w_t) - (\ell_t - h_t)(u_t) \qquad \text{(C.20)}
\end{aligned}$$

Moreover, from the first-order optimality condition $w_{t+1} = \arg\min_{w\in\mathbb{R}^d} (\gamma\ell_t - \gamma h_t + h_{t+1})(w) + \gamma D_{\psi_t}(w|w_t)$, we have

$$\langle \nabla(\gamma\ell_t - \gamma h_t + h_{t+1})(w_{t+1}) + \gamma\nabla\psi_t(w_{t+1}) - \gamma\nabla\psi_t(w_t), w_{t+1} - u_t \rangle \le 0$$

so re-arranging:

$$\begin{aligned}
\langle \nabla(\gamma\ell_t - \gamma h_t + h_{t+1})(w_{t+1}), w_{t+1} - u_t \rangle &\le \gamma \langle \nabla\psi_t(w_t) - \nabla\psi_t(w_{t+1}), w_{t+1} - u_t \rangle \\
&= \gamma D_{\psi_t}(u_t|w_t) - \gamma D_{\psi_t}(u_t|w_{t+1}) - \gamma D_{\psi_t}(w_{t+1}|w_t),
\end{aligned}$$

where the last line uses the three-point relation for bregman divergences, $\langle \nabla f(w) - \nabla f(w'), w' - u \rangle =$

$D_f(u|w) - D_f(u|w') - D_f(w'|w)$. Thus,

$$\begin{aligned}
&\sum_{t=1}^{T}(\gamma\ell_t - \gamma h_t + h_{t+1})(w_{t+1}) - (\gamma\ell_t - \gamma h_t + h_{t+1})(u_t)\\
&\quad\overset{(a)}{=} \sum_{t=1}^{T}\langle\nabla(\gamma\ell_t - \gamma h_t + h_{t+1})(w_{t+1}), w_{t+1} - u_t\rangle - D_{\gamma\ell_t-\gamma h_t+h_{t+1}}(u_t|w_{t+1})\\
&\quad\le \sum_{t=1}^{T}\gamma D_{\psi_t}(u_t|w_t) - \gamma D_{\psi_t}(u_t|w_{t+1}) - \gamma D_{\psi_t}(w_{t+1}|w_t) - D_{\gamma\ell_t-\gamma h_t+h_{t+1}}(u_t|w_{t+1})\\
&\quad\overset{(b)}{=} \sum_{t=1}^{T}\gamma D_{\psi_t}(u_t|w_t) - \gamma D_{\psi_t}(u_t|w_{t+1}) - D_{h_{t+1}}(u_t|w_{t+1}) - \gamma D_{\psi_t}(w_{t+1}|w_t)\\
&\quad\overset{(c)}{=} \sum_{t=1}^{T}\gamma D_{\psi_t}(u_t|w_t) - D_{\psi_{t+1}}(u_t|w_{t+1}) - \gamma D_{\psi_t}(w_{t+1}|w_t)\\
&\quad= \sum_{t=1}^{T}D_{\psi_t}(u_t|w_t) - D_{\psi_{t+1}}(u_t|w_{t+1}) - (1-\gamma)D_{\psi_t}(u_t|w_t) - \gamma D_{\psi_t}(w_{t+1}|w_t),
\end{aligned}$$

where $(a)$ uses the definition of Bregman divergence to re-write $f(w) - f(u) = \langle\nabla f(w), w - u\rangle - D_f(u|w)$, $(b)$ observes that $\gamma(\ell_t - h_t)(w) = \gamma\left(\frac{1}{2}y_t^2 - \frac{1}{2}\widetilde{y}_t^2 + (\widetilde{y}_t - y_t)\langle x_t, w\rangle\right)$, so $D_{\gamma\ell_t-\gamma h_t+h_{t+1}}(\cdot|\cdot) = D_{h_{t+1}}(\cdot|\cdot)$ due to the invariance of Bregman divergences to linear terms, and $(c)$ recalls that $\Sigma_{t+1} = x_{t+1}x_{t+1}^\top + \gamma\Sigma_t$ so that overall we have:

$$\begin{aligned}
\gamma D_{\psi_t}(u_t|w_{t+1}) + D_{h_{t+1}}(u_t|w_{t+1}) &= \frac{\gamma}{2}\|u_t - w_{t+1}\|_{\Sigma_t}^2 + \frac{1}{2}\langle x_{t+1}, u_t - w_{t+1}\rangle^2\\
&= \frac{1}{2}\|u_t - w_{t+1}\|_{\Sigma_{t+1}}^2\\
&= D_{\psi_{t+1}}(u_t|w_{t+1}).
\end{aligned}$$

Plugging this back into Equation (C.20), we have

$$\begin{aligned}
R_T(\boldsymbol{u}) \le \sum_{t=1}^{T}&D_{\psi_t}(u_t|w_t) - D_{\psi_{t+1}}(u_t|w_{t+1})\\
&+ h_1(w_1) - h_{T+1}(w_{T+1}) + \sum_{t=1}^{T}h_{t+1}(u_t) - h_t(u_t)\\
&+ \gamma\sum_{t=1}^{T}(\ell_t - h_t)(w_t) - (\ell_t - h_t)(w_{t+1}) - D_{\psi_t}(w_{t+1}|w_t)\\
&+ (1-\gamma)\sum_{t=1}^{T}(\ell_t - h_t)(w_t) - (\ell_t - h_t)(u_t) - D_{\psi_t}(u_t|w_{t+1}).
\end{aligned}$$

Finally, observe that for any $u, v \in \mathbb{R}^d$, $(\ell_t - h_t)(u) - (\ell_t - h_t)(v) = (\widetilde{y}_t - y_t)\langle x_t, u - v\rangle$, so an application of Fenchel-Young inequality yields

$$\begin{aligned}(\ell_t - h_t)(u) - (\ell_t - h_t)(v) - D_{\psi_t}(v|u) &= (\widetilde{y}_t - y_t)\langle x_t, u - v\rangle - \frac{1}{2}\|u - v\|^2_{\Sigma_t}\\ &\leq \frac{1}{2}(y_t - \widetilde{y}_t)^2 \|x_t\|^2_{\Sigma_t^{-1}}.\end{aligned}$$

Applying this in the last two lines of the previous display yields

$$\begin{aligned}R_T(\boldsymbol{u}) \leq &\sum_{t=1}^{T} D_{\psi_t}(u_t|w_t) - D_{\psi_{t+1}}(u_t|w_{t+1})\\ &\quad \underbrace{h_1(w_1) - h_{T+1}(w_{T+1})}_{\leq 0} + \sum_{t=1}^{T} h_{t+1}(u_t) - h_t(u_t)\\ &\quad \gamma\sum_{t=1}^{T}\frac{1}{2}(y_t - \widetilde{y}_t)^2\|x_t\|^2_{\Sigma_t^{-1}} + (1-\gamma)\sum_{t=1}^{T}\frac{1}{2}(y_t - \widetilde{y}_t)^2\|x_t\|^2_{\Sigma_t^{-1}}\\ \leq &\sum_{t=1}^{T} D_{\psi_t}(u_t|w_t) - D_{\psi_{t+1}}(u_t|w_{t+1})\\ &\quad + \sum_{t=1}^{T} h_{t+1}(u_t) - h_t(u_t)\\ &\quad + \sum_{t=1}^{T}\frac{1}{2}(y_t - \widetilde{y}_t)^2\|x_t\|^2_{\Sigma_t^{-1}}\end{aligned}$$

□

### Proof of Lemma C.2.4

The following lemma bounds the sum of divergence terms. Intuitively, the goal here is to remove all instances of $w_t$ from the analysis, since in an unbounded domain any terms depending on $w_t$ will be hard to quantify and could be arbitrarily large in general. Lemma C.2.4 shows how get rid of the $w_t$-dependent terms left in the bound from Lemma C.2.3, such that only dependencies on the comparators $u_t$ remain.

**Lemma C.2.4.** *Under the same conditions as Lemma C.2.3,*

$$\sum_{t=1}^{T} D_{\psi_t}(u_t|w_t) - D_{\psi_{t+1}}(u_t|w_{t+1}) + \sum_{t=1}^{T} h_{t+1}(u_t) - h_t(u_t) \leq \frac{\gamma\lambda}{2}\|u_1\|_2^2 + h_{T+1}(u_T) + \gamma\sum_{t=1}^{T-1} F_t^\gamma(u_{t+1}) - F_t^\gamma(u_t).$$

*where* $F_t^\gamma(w) = \sum_{s=0}^{t}\gamma^{t-s}\ell_s(w)$.

*Proof.* Observe that by Lemma C.2.14 we have $D_{\ell_t}(u|v) = \frac{1}{2}\langle x_t, u - v\rangle^2 = D_{h_t}(u|v)$ for any $u, v \in W$. Hence, letting $F_t^\gamma(w) = \sum_{s=0}^{t}\gamma^{t-s}\ell_s(w)$ and $\widehat{F}_t^\gamma(w) = h_t(w) + \gamma F_{t-1}^\gamma(w)$, and recalling $\psi_t(w) =$

$\frac{1}{2}\left\|w\right\|_{\Sigma_t}^2 = \frac{\gamma^t\lambda}{2}\left\|w\right\|_2^2 + \frac{1}{2}\sum_{s=1}^t \gamma^{t-s}\left\langle x_s, w\right\rangle^2$, we have $D_{\psi_t}(u|v) = D_{\widehat{F}_t^\gamma}(u|v)$ for any $u, v \in\in \mathbb{R}^d$. Thus:

$$
\begin{aligned}
&\sum_{t=1}^T D_{\psi_t}(u_t|w_t) - D_{\psi_{t+1}}(u_t|w_{t+1}) \\
&\qquad= D_{\psi_1}(u_1|w_1) - D_{\psi_{T+1}}(u_T|w_{T+1}) + \sum_{t=2}^T D_{\psi_t}(u_t|w_t) - D_{\psi_t}(u_{t-1}|w_t) \\
&\qquad= D_{\psi_1}(u_1|w_1) - D_{\psi_{T+1}}(u_T|w_{T+1}) + \sum_{t=2}^T D_{\widehat{F}_t^\gamma}(u_t|w_t) - D_{\widehat{F}_t^\gamma}(u_{t-1}|w_t) \\
&\qquad= D_{\psi_1}(u_1|w_1) - D_{\psi_{T+1}}(u_T|w_{T+1}) + \sum_{t=2}^T \widehat{F}_t^\gamma(u_t) - \widehat{F}_t^\gamma(u_{t-1}) - \left\langle \nabla\widehat{F}_t^\gamma(w_t), u_t - u_{t-1}\right\rangle.
\end{aligned}
$$

Moreover, by Proposition C.2.1 we have

$$
w_t = \arg\min_{w\in\mathbb{R}^d} h_t(w) + \gamma\sum_{s=0}^{t-1}\gamma^{t-1-s}\ell_s(w) = \arg\min_{w\in\mathbb{R}^d}\widehat{F}_t^\gamma(w),
$$

hence by convexity of $\widehat{F}_t^\gamma$ and the first-order optimality condition we have $\nabla\widehat{F}_t^\gamma(w_t) = \mathbf{0}$, so overall we have

$$
\begin{aligned}
&\sum_{t=1}^T D_{\psi_t}(u_t|w_t) - D_{\psi_t}(u_t|w_{t+1}) + \sum_{t=1}^T h_{t+1}(u_t) - h_t(u_t) \\
&\quad= D_{\psi_1}(u_1|w_1) - D_{\psi_{T+1}}(u_T|w_{T+1}) + \sum_{t=2}^T \widehat{F}_t^\gamma(u_t) - \widehat{F}_t^\gamma(u_{t-1}) + \sum_{t=1}^T h_{t+1}(u_t) - h_t(u_t) \\
&\quad= D_{\psi_1}(u_1|w_1) - D_{\psi_{T+1}}(u_T|w_{T+1}) + \sum_{t=2}^T\left[h_t(u_t) - h_t(u_{t-1}) + \gamma F_{t-1}^\gamma(u_t) - \gamma F_{t-1}^\gamma(u_{t-1})\right] + \sum_{t=1}^T h_{t+1}(u_t) - h_t(u_t) \\
&\quad= D_{\psi_1}(u_1|w_1) - D_{\psi_{T+1}}(u_T|w_{T+1}) + \gamma\sum_{t=1}^{T-1} F_t^\gamma(u_{t+1}) - F_t^\gamma(u_t) + \sum_{t=2}^T h_{t+1}(u_t) - h_t(u_{t-1}) + h_2(u_1) - h_1(u_1) \\
&\quad= D_{\psi_1}(u_1|w_1) - D_{\psi_{T+1}}(u_T|w_{T+1}) + h_{T+1}(u_T) - h_1(u_1) + \gamma\sum_{t=1}^{T-1} F_t^\gamma(u_{t+1}) - F_t^\gamma(u_t).
\end{aligned}
$$

Finally, observe that with $w_1 = \mathbf{0}$ and $\widetilde{y}_1 = 0$ we have

$$
D_{\psi_1}(u_1|w_1) = \psi_1(u_1) - \psi_1(\mathbf{0}) - \left\langle\nabla\psi_1(\mathbf{0}), u_1\right\rangle = h_1(u_1) + \gamma\ell_0(u_1) = h_1(u_1) + \frac{\gamma\lambda}{2}\left\|u_1\right\|_2^2
$$

so we can express the bound as the bound as

$$
\begin{aligned}
&\sum_{t=1}^T D_{\psi_t}(u_t|w_t) - D_{\psi_t}(u_t|w_{t+1}) + \sum_{t=1}^T h_{t+1}(u_t) - h_t(u_t) \\
&\qquad\le \frac{\gamma\lambda}{2}\left\|u_1\right\|_2^2 + h_{T+1}(u_T) + \gamma\sum_{t=1}^{T-1} F_t^\gamma(u_{t+1}) - F_t^\gamma(u_t).
\end{aligned}
$$

□

**Proof of Lemma 10.2.2**

The following lemma bounds the variability and stability terms from Theorem 10.2.1 to expose a more explicit trade-off in terms of the discount factor $\gamma$.

**Lemma 10.2.2.** *Let $\ell_0, \ell_1, \dots, \ell_T$ be arbitrary non-negative functions, $0 < \gamma \leq \beta < 1$, and $F_t^\gamma(w) = \sum_{s=0}^t \gamma^{t-s} \ell_s(w)$. For all $t$, define*

$$\bar{d}_t^\beta(u, v) = \frac{1}{\sum_{s=0}^t \beta^{t-s}} \sum_{s=0}^t \beta^{t-s} \left[\ell_s(u) - \ell_s(v)\right]_+$$

*and let $P_T^\beta(\boldsymbol{u}) = \sum_{t=1}^{T-1} \bar{d}_t^\beta(u_{t+1}, u_t)$. Then for any $V_T \geq 0$,*

$$\gamma \sum_{t=1}^{T-1} \left[F_t^\gamma(u_{t+1}) - F_t^\gamma(u_t)\right] + \log\left(\frac{1}{\gamma}\right) V_T \leq \frac{\beta}{1-\beta} P_T^\beta(\boldsymbol{u}) + \frac{1-\gamma}{\gamma} V_T$$

*Proof.* The first summation can be bounded as

$$\begin{aligned}
\gamma \sum_{t=1}^{T-1} \left[F_t^\gamma(u_t) - F_t^\gamma(u_{t-1})\right] &= \gamma \sum_{t=1}^{T-1} \sum_{s=0}^t \gamma^{t-s} \left[\ell_s(u_{t+1}) - \ell_s(u_t)\right] \\
&\leq \gamma \sum_{t=1}^{T-1} \sum_{s=0}^t \gamma^{t-s} \left[\ell_s(u_{t+1}) - \ell_s(u_t)\right]_+ \\
&\leq \beta \sum_{t=1}^{T-1} \sum_{s=0}^t \frac{\sum_{s'=0}^t \beta^{t-s'}}{\sum_{s'=0}^t \beta^{t-s'}} \beta^{t-s} \left[\ell_s(u_{t+1}) - \ell_s(u_t)\right]_+ \\
&\leq \frac{\beta}{1-\beta} \sum_{t=1}^{T-1} \sum_{s=0}^t \frac{\beta^{t-s}}{\sum_{s'=0}^t \beta^{t-s'}} \left[\ell_s(u_{t+1}) - \ell_s(u_t)\right]_+ \\
&= \frac{\beta}{1-\beta} P_T^\beta(\boldsymbol{u}),
\end{aligned}$$

where the last inequality uses $\sum_{s=0}^t \beta^{t-s} = \frac{1-\beta^{t+1}}{1-\beta} \leq \frac{1}{1-\beta}$. Using this along with the elementary inequality $\log(x) \leq x - 1$, for any $V_T \geq 0$ we have

$$\begin{aligned}
\gamma \sum_{t=1}^{T-1} \left[F_t^\gamma(u_t) - F_t^\gamma(u_{t-1})\right] + \log\left(\frac{1}{\gamma}\right) V_T &\leq \frac{\beta}{1-\beta} P_T^\beta(\boldsymbol{u}) + \left(\frac{1}{\gamma} - 1\right) V_T \\
&= \frac{\beta}{1-\beta} P_T^\beta(\boldsymbol{u}) + \frac{1-\gamma}{\gamma} V_T
\end{aligned}$$

□

### Existence of a Good Discount Factor

The following lemma establishes the existence of a discount factor that will lead to favorable tuning of the $\gamma$-dependent terms in Lemma 10.2.2.

**Lemma C.2.5.** *Let $\ell_0, \ell_1, \dots$ be arbitrary non-negative functions, $V_T \geq 0$, and denote $\bar{d}_t^{\gamma}(u,v) = \frac{\sum_{s=0}^{t} \gamma^{t-s}[\ell_s(u) - \ell_s(v)]_+}{\sum_{s=0}^{t} \gamma^{t-s}}$ for $\gamma \in [0,1]$, and $P_T^{\gamma}(\boldsymbol{u}) = \sum_{t=1}^{T-1} \bar{d}_t^{\gamma}(u_{t+1}, u_t)$. Then there is a $\gamma^* \in [0,1]$ such that*

$$\gamma^* = \frac{\sqrt{V_T}}{\sqrt{V_T} + \sqrt{P_T^{\gamma^*}(\boldsymbol{u})}}.$$

*Proof.* First, notice that that any such $\gamma$ with the stated property must be in $[0,1]$ since

$$0 \leq \frac{\sqrt{V_T}}{\sqrt{V_T} + \sqrt{P_T^{\gamma}(\boldsymbol{u})}} \leq \frac{\sqrt{V_T}}{\sqrt{V_T}} = 1.$$

Next, observe that the condition can be equivalently expressed as follows:

$$\begin{aligned}
& \gamma = \frac{\sqrt{V_T}}{\sqrt{V_T} + \sqrt{P_T^{\gamma}(\boldsymbol{u})}} \\
\iff & \sqrt{V_T}(1-\gamma) = \gamma\sqrt{P_T^{\gamma}(\boldsymbol{u})} \\
& = \gamma\sqrt{\sum_{t=1}^{T-1}\sum_{s=0}^{t} \frac{\gamma^{t-s}}{\sum_{s=0}^{t}\gamma^{t-s}} \left[\ell_s(u_{t+1}) - \ell_s(u_t)\right]_+} \\
& = \gamma\sqrt{\sum_{t=1}^{T-1}\sum_{s=0}^{t} \frac{\gamma^{t-s}}{1-\gamma^{t+1}} (1-\gamma) \left[\ell_s(u_{t+1}) - \ell_s(u_t)\right]_+} \\
\iff & \sqrt{V_T(1-\gamma)} = \gamma\sqrt{\sum_{t=1}^{T-1}\sum_{s=0}^{t} \frac{\gamma^{t-s}}{1-\gamma^{t+1}} \left[\ell_s(u_{t+1}) - \ell_s(u_t)\right]_+}.
\end{aligned}$$

The quantity on the LHS begins at $\sqrt{V_T}$ (for $\gamma = 0$) and then decreases to 0 as a function of $\gamma$. Likewise, the RHS begins at 0 (for $\gamma = 0$) and increases as a function of $\gamma$, approaching $\infty$ as $\gamma \to 1$. Hence, there must be some $\gamma \in [0,1]$ at which the two lines cross, and hence a $\gamma \in [0,1]$ which satisfies the above relation, so there is a $\gamma \in [0,1]$ such that

$$\gamma = \frac{\sqrt{V_T}}{\sqrt{V_T} + \sqrt{P_T^{\gamma}(\boldsymbol{u})}}.$$

□

### Proof of Theorem 10.2.3

Now combining everything we've seen in the previous sections, we can easily prove the following bound for the discounted VAW forecaster under *oracle tuning* of the discount factor.

**Theorem 10.2.3.** *For any sequences $y_1, \dots, y_T$ and $\widetilde{y}_1, \dots, \widetilde{y}_T$ in $\mathbb{R}$ and any sequence $\boldsymbol{u} = (u_1, \dots, u_T)$ in $\mathbb{R}^d$, there is a discount factor $\gamma^* \in [0,1]$ satisfying*

$$\gamma^* = \frac{\sqrt{\frac{d}{2}\sum_{t=1}^T (y_t - \widetilde{y}_t)^2}}{\sqrt{\frac{d}{2}\sum_{t=1}^T (y_t - \widetilde{y}_t)^2} + \sqrt{P_T^{\gamma^*}(\boldsymbol{u})}} \tag{10.2}$$

*with which the regret of Algorithm 12 is bounded above by*

$$R_T(\boldsymbol{u}) \le O\left( d \max_t (y_t - \widetilde{y}_t)^2 \log(T) + \sqrt{d P_T^{\gamma^*}(\boldsymbol{u}) \sum_{t=1}^T (y_t - \widetilde{y}_t)^2} \right)$$

*Proof.* Lemma C.2.5 shows that for any sequence $\boldsymbol{u} = (u_1, \dots, u_T)$, there is a $\gamma^* \in [0,1]$ such that

$$\gamma^* = \frac{\sqrt{d\sum_{t=1}^T \frac{1}{2}(y_t - \widetilde{y}_t)^2}}{\sqrt{d\sum_{t=1}^T \frac{1}{2}(y_t - \widetilde{y}_t)^2} + \sqrt{P_T^{\gamma^*}(\boldsymbol{u})}},$$

so choosing $\gamma = \gamma^*$ and applying Theorem 10.2.1, we have

$$\begin{aligned}
R_T(\boldsymbol{u}) &\le \frac{\gamma^* \lambda}{2} \|u_1\|_2^2 + \frac{d}{2} \max_t (y_t - \widetilde{y}_t)^2 \log\left(1 + \frac{\sum_{t=1}^T \|x_t\|_2^2}{\lambda d}\right) \\
&\quad + \gamma^* \sum_{t=1}^{T-1} \left[ F_t^{\gamma^*}(u_{t+1}) - F_t^{\gamma^*}(u_t) \right] + \frac{d}{2} \log(1/\gamma^*) \sum_{t=1}^T (y_t - \widetilde{y}_t)^2 \\
&\overset{(*)}{\le} \frac{\lambda}{2} \|u_1\|_2^2 + \frac{d}{2} \max_t (y_t - \widetilde{y}_t)^2 \log\left(1 + \frac{\sum_{t=1}^T \|x_t\|_2^2}{\lambda d}\right) \\
&\quad + \frac{\gamma^*}{1-\gamma^*} P_T^{\gamma^*}(\boldsymbol{u}) + \frac{1-\gamma^*}{\gamma^*} \frac{d}{2} \sum_{t=1}^T (y_t - \widetilde{y}_t)^2 \\
&= \frac{\lambda}{2} \|u_1\|_2^2 + \frac{d}{2} \max_t (y_t - \widetilde{y}_t)^2 \log\left(1 + \frac{\sum_{t=1}^T \|x_t\|_2^2}{\lambda d}\right) + \sqrt{2d P_T^{\gamma^*}(\boldsymbol{u}) \sum_{t=1}^T (y_t - \widetilde{y}_t)^2}
\end{aligned}$$

where $(*)$ uses Lemma 10.2.2 (with $\beta = \gamma = \gamma^*$). The stated result follows by hiding lower-order terms. □

### C.2.2 Proofs for Section 10.2.1 (Small-loss Bounds via Self-confident Predictions)

#### Proof of Theorem 10.2.4

We split the proof of Theorem 10.2.4 into two parts. The following lemma, proven in Appendix C.2.2, first derives an initial regret template that does most of the heavy lifting. We will later re-use this template in the proof of Theorem 10.3.3 to avoid repeating the argument. The high-level intuition is that choosing hints $\widetilde{y}_t \approx \langle x_t, w_t\rangle$ leads to $\sum_{t=1}^T (y_t - \widetilde{y}_t)^2 \approx \sum_{t=1}^T \ell_t(w_t)$, which leads to a self-bounding argument that lets us replace $\sum_{t=1}^T (y_t - \widetilde{y}_t)^2$ with $\sum_{t=1}^T \ell_t(u_t)$ in the regret bound. We defer proof of the lemma to the next subsection, Appendix C.2.2.

**Lemma C.2.6.** *Let $y_t^{Ref} \in \mathbb{R}$ be an arbitrary reference point, available at the start of round $t$, and let $\mathcal{B}_t = \left\{y \in \mathbb{R} : y_t^{Ref} - M_t \le y \le y_t^{Ref} + M_t\right\}$ for $M_t = \max_{s<t} \left|y_s - y_s^{Ref}\right|$. Suppose that we apply Algorithm 12 with hints $\widetilde{y}_t = \overline{y}_t := \mathrm{Clip}_{\mathcal{B}_t}(\langle x_t, w_t\rangle)$. Then for any sequence $\boldsymbol{u} = (u_1, \ldots, u_T)$ in $\mathbb{R}^d$ and any $\gamma, \beta \in [0,1]$ such that $\beta \ge \gamma \ge \gamma_{\min} = \frac{2d}{2d+1}$,*

$$R_T(\boldsymbol{u}) \le \gamma\lambda \|u_1\|_2^2 + 4d \max_t (y_t - y_t^{Ref})^2 \log\left(1 + \frac{\sum_{t=1}^T \gamma^{T-t} \|x_t\|_2^2}{\lambda d}\right)$$
$$+ 2\frac{\beta}{1-\beta} P_T^{\beta}(\boldsymbol{u}) + \frac{1-\gamma}{\gamma} 2d \sum_{t=1}^T \ell_t(u_t)$$

Now using this template, Theorem 10.2.4 is easily proven by plugging in the stated discount factor $\gamma = \gamma^\circ \vee \gamma_{\min}$

**Theorem 10.2.4.** *Let $y_t^{Ref} \in \mathbb{R}$ be an arbitrary reference point and let $\mathcal{B}_t = [y_t^{Ref} - M_t, y_t^{Ref} + M_t]$ for $M_t = \max_{s<t} \left|y_s - y_s^{Ref}\right|$. Suppose that we apply Algorithm 12 with hints $\widetilde{y}_t = \mathrm{Clip}_{\mathcal{B}_t}(\langle x_t, w_t\rangle)$. Then for any sequence of losses $\ell_1, \ldots, \ell_T$ and any sequence $\boldsymbol{u} = (u_1, \ldots, u_T)$ in $\mathbb{R}^d$, there is a $\gamma^\circ \in [0,1]$ satisfying*

$$\gamma^\circ = \frac{\sqrt{d \sum_{t=1}^T \ell_t(u_t)}}{\sqrt{d \sum_{t=1}^T \ell_t(u_t)} + \sqrt{P_T^{\gamma^\circ}(\boldsymbol{u})}}. \tag{10.3}$$

*Moreover, running Algorithm 12 with discount $\gamma^\circ \vee \gamma_{\min}$ for $\gamma_{\min} = \frac{2d}{2d+1}$ ensures*

$$R_T(\boldsymbol{u}) \le O\left(d P_T^{\gamma_{\min}}(\boldsymbol{u}) + d \max_t (y_t - y_t^{Ref})^2 \log(T) + \sqrt{d P_T^{\gamma^\circ}(\boldsymbol{u}) \sum_{t=1}^T \ell_t(u_t)}\right),$$

*Proof.* By Lemma C.2.6 (with $\beta = \gamma$), for any $\gamma \geq \gamma_{\min} = \frac{2d}{2d+1}$, we have

$$R_T(\boldsymbol{u}) \leq \gamma\lambda \left\|u_1\right\|_2^2 + 4d \max_t (y_t - y_t^{\mathrm{Ref}})^2 \log\left(1 + \frac{\sum_{t=1}^T \gamma^{T-t} \left\|x_t\right\|_2^2}{\lambda d}\right)$$
$$+ 2\frac{\gamma}{1-\gamma} P_T^{\gamma}(\boldsymbol{u}) + \frac{1-\gamma}{\gamma} 2d \sum_{t=1}^T \ell_t(u_t).$$

Now by Lemma C.2.5, there is a $\gamma^\circ \in [0,1]$ satisfying $\gamma^\circ = \frac{\sqrt{d\sum_{t=1}^T \ell_t(u_t)}}{\sqrt{d\sum_{t=1}^T \ell_t(u_t)} + \sqrt{P_T^{\gamma^\circ}(\boldsymbol{u})}}$. If $\gamma^\circ \geq \gamma_{\min}$, then for $\gamma = \gamma^\circ \vee \gamma_{\min}$, the terms in the second line reduce to

$$2\frac{\gamma^\circ}{1-\gamma^\circ} P_T^{\gamma^\circ}(\boldsymbol{u}) + \frac{1-\gamma^\circ}{\gamma^\circ} 2d \sum_{t=1}^T \ell_t(u_t) = 4\sqrt{dP_T^{\gamma^\circ}(\boldsymbol{u}) \sum_{t=1}^T \ell_t(u_t)},$$

and otherwise for $\gamma^\circ \leq \gamma_{\min}$ we have

$$\begin{aligned}2\frac{\gamma_{\min}}{1-\gamma_{\min}} P_T^{\gamma_{\min}}(\boldsymbol{u}) + \frac{1-\gamma_{\min}}{\gamma_{\min}} 2d \sum_{t=1}^T \ell_t(u_t) &\leq 2\frac{\gamma_{\min}}{1-\gamma_{\min}} P_T^{\gamma_{\min}}(\boldsymbol{u}) + \frac{1-\gamma^\circ}{\gamma^\circ} 2d \sum_{t=1}^T \ell_t(u_t) \\ &\leq 4dP_T^{\gamma_{\min}}(\boldsymbol{u}) + 2\sqrt{dP_T^{\gamma^\circ}(\boldsymbol{u}) \sum_{t=1}^T \ell_t(u_t)},\end{aligned}$$

so combining these two bounds and plugging back into the regret bound above, we have

$$\begin{aligned}R_T(\boldsymbol{u}) &\leq \gamma\lambda \left\|u_1\right\|_2^2 + 4d \max_t (y_t - y_t^{\mathrm{Ref}})^2 \log\left(1 + \frac{\sum_{t=1}^T \gamma^{T-t} \left\|x_t\right\|_2^2}{\lambda d}\right) \\ &\quad + 4dP_T^{\gamma_{\min}}(\boldsymbol{u}) + 4\sqrt{dP_T^{\gamma^\circ}(\boldsymbol{u}) \sum_{t=1}^T \ell_t(u_t)} \\ &\leq O\left(dP_T^{\gamma_{\min}}(\boldsymbol{u}) + d \max_t (y_t - y_t^{\mathrm{Ref}})^2 \log(T) + \sqrt{dP_T^{\gamma^\circ}(\boldsymbol{u}) \sum_{t=1}^T \ell_t(u_t).}\right)\end{aligned}$$

□

### Proof of Lemma C.2.6

**Lemma C.2.6.** *Let $y_t^{Ref} \in \mathbb{R}$ be an arbitrary reference point, available at the start of round $t$, and let $\mathcal{B}_t = \left\{y \in \mathbb{R} : y_t^{Ref} - M_t \le y \le y_t^{Ref} + M_t\right\}$ for $M_t = \max_{s<t} \left|y_s - y_s^{Ref}\right|$. Suppose that we apply Algorithm 12 with hints $\widetilde{y}_t = \overline{y}_t := \mathrm{Clip}_{\mathcal{B}_t}(\langle x_t, w_t\rangle)$. Then for any sequence $\boldsymbol{u} = (u_1, \dots, u_T)$ in $\mathbb{R}^d$ and any $\gamma, \beta \in [0,1]$ such that $\beta \ge \gamma \ge \gamma_{\min} = \frac{2d}{2d+1}$,*

$$R_T(\boldsymbol{u}) \le \gamma\lambda \left\|u_1\right\|_2^2 + 4d \max_t (y_t - y_t^{Ref})^2 \log\left(1 + \frac{\sum_{t=1}^T \gamma^{T-t} \left\|x_t\right\|_2^2}{\lambda d}\right) + 2\frac{\beta}{1-\beta} P_T^\beta(\boldsymbol{u}) + \frac{1-\gamma}{\gamma} 2d \sum_{t=1}^T \ell_t(u_t)$$

*Proof.* Applying Theorem 10.2.1 followed by Lemma 10.2.2, for any $\gamma \in (0,1]$ and $\beta \ge \gamma$ we have

$$\begin{aligned}
R_T^{\mathcal{A}_\gamma}(\boldsymbol{u}) &\le \frac{\gamma\lambda}{2} \left\|u_1\right\|_2^2 + \frac{d}{2} \max_t (y_t - \overline{y}_t)^2 \log\left(1 + \frac{\sum_{t=1}^T \gamma^{T-t} \left\|x_t\right\|_2^2}{\lambda d}\right) \\
&\quad + \gamma \sum_{t=1}^{T-1} \left[F_t^\gamma(u_{t+1}) - F_t^\gamma(u_t)\right] + \frac{d}{2} \log(1/\gamma) \sum_{t=1}^T (y_t - \overline{y}_t)^2 \\
&\le \frac{\gamma\lambda}{2} \left\|u_1\right\|_2^2 + \frac{d}{2} \max_t (y_t - \overline{y}_t)^2 \log\left(1 + \frac{\sum_{t=1}^T \gamma^{T-t} \left\|x_t\right\|_2^2}{\lambda d}\right) \\
&\quad + \frac{\beta}{1-\beta} P_T^\beta(\boldsymbol{u}) + \frac{1-\gamma}{\gamma} \frac{d}{2} \sum_{t=1}^T (y_t - \overline{y}_t)^2,
\end{aligned}$$

Using Lemma C.2.7 we have

$$\sum_{t=1}^T (y_t - \overline{y}_t)^2 \le \sum_{t=1}^T \left[M_{t+1}^2 - M_t^2 + 2\ell_t(w_t)\right] \le M_{T+1}^2 + 2\sum_{t=1}^T \ell_t(w_t),$$

so for any $\gamma \ge \frac{2d}{2d+1}$, we have

$$\begin{aligned}
\frac{1-\gamma}{\gamma} \frac{d}{2} \sum_{t=1}^T (y_t - \overline{y}_t)^2 &\le \frac{1-\gamma}{\gamma} d \left[\frac{1}{2} M_{T+1}^2 + \sum_{t=1}^T \ell_t(w_t)\right] \\
&= \frac{1-\gamma}{\gamma} d \left[\frac{1}{2} M_{T+1}^2 + \sum_{t=1}^T \ell_t(w_t) - \ell_t(u_t) + \sum_{t=1}^T \ell_t(u_t)\right] \\
&\le \frac{1}{4} M_{T+1}^2 + \frac{1}{2} \sum_{t=1}^T \ell_t(w_t) - \ell_t(u_t) + \frac{1-\gamma}{\gamma} d \sum_{t=1}^T \ell_t(u_t),
\end{aligned}$$

where the final inequality uses $\gamma \ge \frac{2d}{2d+1} \implies \frac{1-\gamma}{\gamma} \le \frac{1}{2d}$ and bounds $\frac{1-\gamma}{\gamma} d \sum_{t=1}^T \ell_t(w_t) - \ell_t(u_t) \le \frac{1}{2} \sum_{t=1}^T \ell_t(w_t) - \ell_t(u_t)$ (assuming $\sum_{t=1}^T \ell_t(w_t) - \ell_t(u_t) \ge 0$, which can be assumed without loss of generality since otherwise the stated bound trivially holds). Plugging this back into the regret bound and re-arranging terms, we have

$$\begin{aligned}
R_T(\boldsymbol{u}) &\leq \frac{\gamma\lambda}{2}\left\|u_1\right\|_2^2 + \frac{d}{2}\max_t(y_t - \overline{y}_t)^2\log\left(1 + \frac{\sum_{t=1}^T \gamma^{T-t}\left\|x_t\right\|_2^2}{\lambda d}\right)\\
&\quad + \frac{\gamma}{1-\gamma}P_T^{\gamma}(\boldsymbol{u}) + \frac{1}{2}R_T(\boldsymbol{u}) + \frac{1-\gamma}{\gamma}\sum_{t=1}^T \ell_t(u_t)\\
\implies R_T(\boldsymbol{u}) &\leq \gamma\lambda\left\|u_1\right\|_2^2 + 4d\max_t(y_t - y_t^{\text{Ref}})^2\log\left(1 + \frac{\sum_{t=1}^T \gamma^{T-t}\left\|x_t\right\|_2^2}{\lambda d}\right)\\
&\quad + 2\frac{\beta}{1-\beta}P_T^{\beta}(\boldsymbol{u}) + \frac{1-\gamma}{\gamma}2d\sum_{t=1}^T \ell_t(u_t),
\end{aligned}$$

where we've bounded $\max_t(y_t - \overline{y}_t)^2 \leq 4M_{T+1} = 4\max_t(y_t - y_t^{\text{Ref}})^2$ using Lemma C.2.7. □

### C.2.3 Proofs for Section 10.2.2 (Dimension-dependent Lower Bound)

**Proof of Theorem 10.2.5**

**Theorem 10.2.5.** *For any $d, T \geq 1$ and $P, Y > 0$ such that $dP \leq 2TY^2$, there is a sequence of losses $\ell_t(w) = \frac{1}{2}(y_t - \langle x_t, w\rangle)^2$ and a comparator sequence $\boldsymbol{u} = (u_1, \ldots, u_T)$ satisfying $\max_t|y_t| \leq Y$ and $\sum_{t=1}^{T-1}\max_s\left[\ell_s(u_{t+1}) - \ell_s(u_t)\right]_+ \leq P$ such that*

$$R_T(\boldsymbol{u}) \geq \Omega\left(dY^2\log(T) + dP + \sqrt{dP\sum_{t=2}^T (y_t - y_{t-1})^2}\right).$$

*Proof.* First notice that the trivial comparator sequence with $u_1 = \ldots = u_T$ always satisfies

$$\sum_{t=2}^T \max_s\left[\ell_s(u_{t+1}) - \ell_s(u_t)\right]_+ = 0 \leq P,$$

so we can always lower-bound the dynamic regret using the well-known lower bound for the static regret in this setting (see, *e.g.*, Vovk (2001), Gaillard et al. (2019), and Mayo, Hadiji, and Erven (2022)). In particular, for any $u \in W$ we have

$$\sup_{y_1,\ldots,y_T} R_T(u) \geq \Omega\left(dY^2\log(T)\right) \tag{C.21}$$

Next, let $\sigma \in [0, 1]$ and let $\sigma_1, \ldots, \sigma_t$ be a sequence of iid random variables drawn uniformly from $\{-\sigma, \sigma\}$, and let $y_t = Y\sigma_t$. Choose feature vectors $x_t$ which cycle through the standard basis vectors (*e.g.* define $\iota(t) = t \pmod{d} + 1$ and let $x_t = e_{\iota(t)}$). Now observe that the comparator sequence can always exactly fit a sequence $y_1, \ldots, y_T$ by setting $u_t$ to satisfy $\langle x_t, u_t\rangle = u_{t,\iota(t)} = y_t$. In particular, by letting $\widetilde{u}_1 = (y_1, \ldots, y_d)$, $\widetilde{u}_2 = (y_{d+1}, \ldots, y_{2d}), \ldots, \widetilde{u}_{\lceil T/d \rceil} = (y_{\lceil T/d \rceil+1}, \ldots, y_T, 0, 0, \ldots)$ we can set $u_t = \widetilde{u}_{\lceil t/d \rceil}$ to guarantee $\langle x_t, u_t \rangle = y_t$ on all rounds, while only changing the comparator $\lceil T/d \rceil$ times at most. From this, we have the following initial bound on the regret:

$$\begin{aligned}
\sup_{y_1,\ldots,y_T} R_T(\boldsymbol{u}) &\geq \mathbb{E}_{\boldsymbol{y}}\left[\sum_{t=1}^{T} \ell_t(w_t) - \ell_t(u_t)\right] \\
&\geq \mathbb{E}_{\boldsymbol{y}}\left[\frac{1}{2}y_t^2 + \frac{1}{2}\langle x_t, w_t \rangle^2 + y_t \langle x_t, w_t \rangle\right] \\
&\geq \frac{1}{2}\sigma^2 Y^2 T, \qquad (C.22)
\end{aligned}$$

where the last line uses $y_t^2 = Y^2\sigma^2$ and $\mathbb{E}[y_t] = 0$. Moreover, since the comparator changes only every $d$ rounds, the variability is bounded as

$$\sum_{t=1}^{T-1} \max_s [\ell_s(u_{t+1}) - \ell_s(u_t)]_+ \leq \sum_{i=1}^{\lceil T/d \rceil - 1} \max_s [\ell_s(\widetilde{u}_{i+1}) - \ell_s(\widetilde{u}_i)]_+ .$$

Observe that $\ell_s(\widetilde{u}_{i+1}) - \ell_s(\widetilde{u}_i)$ can only be positive when $\langle x_s, \widetilde{u}_i \rangle = y_s$ and $\langle x_s, \widetilde{u}_{i+1} \rangle = -y_s$, hence

$$\begin{aligned}
\sum_{t=1}^{T-1} \max_s [\ell_s(u_{t+1}) - \ell_s(u_t)]_+ &\leq \sum_{i=1}^{\lceil T/d \rceil - 1} \max_s [\ell_s(\widetilde{u}_{i+1}) - \ell_s(\widetilde{u}_i)]_+ \\
&\leq \sum_{i=1}^{\lceil T/d \rceil - 1} \frac{1}{2}(y_s - (-y_s))^2 \\
&\leq \frac{2TY^2}{d}\sigma^2.
\end{aligned}$$

Hence, setting $\sigma = \sqrt{\frac{dP}{2TY^2}} \leq 1$ ensures $\sum_{t=1}^{T-1} \max_s [\ell_s(u_{t+1}) - \ell_s(u_t)]_+ \leq \frac{2TY^2}{d}\sigma^2 \leq P$, and the regret is bounded below by

$$\sup_{y_1,\ldots,y_T} R_T(\boldsymbol{u}) \geq \frac{1}{2}\sigma^2 Y^2 T = \frac{1}{4}dP,$$

which we can further lower bound as:

$$\begin{aligned}
&= \frac{1}{4}\sqrt{dP \cdot dP} \geq \frac{1}{4}\sqrt{dP \cdot d \sum_{t=1}^{T-1} \max_s [\ell_s(u_{t+1}) - \ell_s(u_t)]_+} \\
&\geq \frac{1}{4}\sqrt{dP \sum_{t=1}^{T-1} [\ell_t(u_{t+1}) - \ell_t(u_t)]_+} = \Omega\left(\sqrt{dP \sum_{t=2}^{T} \frac{1}{2}(y_t - y_{t-1})^2}\right). \qquad (C.23)
\end{aligned}$$

Taken together with Equation (C.21), we have

$$R_T(\boldsymbol{u}) \geq \Omega\left(dY^2 \log(T) \vee \sqrt{dP\mathcal{V}_T}\right)$$

where $\mathcal{V}_T = dP \vee \sum_{t=2}^{T} \frac{1}{2}(y_t - y_{t-1})^2$.

□

### C.2.4 Proofs for Section 10.3 (Learning the Optimal Discount Factor)

**Proof of Lemma C.2.7**

The following lemma shows that by clipping our predictions to some crude "trust-region", the loss of the clipped prediction is at worst portional to the maximal deviation of the true $y_t$ from the trust region. Intuitively, we can think of $y^{\text{Ref}}$ as being some data-dependent but already-observed quantity, such as $y_{t-1}$.

**Lemma C.2.7.** *Define $M_t = \max_{s<t} \left|y_s - y_s^{Ref}\right|$, $\mathcal{B}_t = \left\{x \in \mathbb{R} : y_t^{Ref} - M_t \leq x \leq y_t^{Ref} + M_t\right\}$, and let $\overline{y}_t = \text{Clip}_{\mathcal{B}_t}(\langle x_t, w_t\rangle)$ for some $w_t \in \mathbb{R}^d$. Then for any $t$ we have*

$$(y_t - \overline{y}_t)^2 \leq \min\left\{4M_{t+1}^2, 2\ell_t(w_t) + M_{t+1}^2 - M_t^2\right\}$$

*Proof.* First, observe that we always have

$$(y_t - \overline{y}_t)^2 = \left(y_t - y_t^{\text{Ref}} + y_t^{\text{Ref}} - \overline{y}_t\right)^2 \leq 2\left(y_t - y_t^{\text{Ref}}\right)^2 + 2\left(y_t^{\text{Ref}} - \overline{y}_t\right)^2 \leq 2M_{t+1}^2 + 2M_t^2 \leq 4M_{t+1}^2.$$

Next, observe that if $\langle x_t, w_t\rangle = \overline{y}_t$, then we trivially have $(y_t - \overline{y}_t)^2 = (y_t - \langle x_t, w_t\rangle)^2 = 2\ell_t(w_t)$. Otherwise, when $\langle x_t, w_t\rangle \neq \overline{y}_t$, we have clipped $\overline{y}_t$ to be a distance of $M_t$ away from $y_t^{\text{Ref}}$ and there are two cases to consider. If $\text{Sgn}\left(\overline{y}_t - y_t^{\text{Ref}}\right) \neq \text{Sgn}\left(y_t - y_t^{\text{Ref}}\right)$, then the clipping operation $\overline{y}_t = \text{Clip}_{\mathcal{B}_t}(\langle x_t, w_t\rangle)$ moves us closer to $y_t$, hence $|y_t - \overline{y}_t| \leq |y_t - \langle x_t, w_t\rangle|$. If $\text{Sgn}\left(\overline{y}_t - y_t^{\text{Ref}}\right) = \text{Sgn}\left(y_t - y_t^{\text{Ref}}\right)$, then we precisely have $|y_t - \overline{y}_t| = M_{t+1} - M_t$ when $y_t \notin \mathcal{B}_t$ and $|y_t - \overline{y}_t| \leq |y_t - \langle x_t, w_t\rangle|$ when $y_t \in \mathcal{B}_t$. Hence, combining these cases we have

$$(y_t - \overline{y}_t)^2 \leq (y_t - \langle w_t, x_t\rangle)^2 + (M_{t+1} - M_t)^2 \leq 2\ell_t(w_t) + M_{t+1}^2 - M_t^2,$$

where we have used $(u - l)^2 \leq u^2 - l^2$ for $u \geq l \geq 0$. Hence, combining with the first display we have

$$(y_t - \overline{y}_t)^2 \leq \min\left\{4M_{t+1}^2, M_{t+1}^2 - M_t^2 + 2\ell_t(w_t)\right\}.$$

□

### Proof of Lemma C.2.8

The following lemma shows the following important property of the meta-learner's losses: they are $\alpha_t$-exp-concave with $\alpha_t = \frac{1}{2\max_i \ell_t(y_t^{(i)})}$ in the domain $\widehat{\mathcal{Y}}_t = \left\{y = \sum_{i=1}^N p_i y_t^{(i)} : \sum_{i=1}^N p_i = 1\right\}$.

**Lemma C.2.8.** *Let $y^{(1)}, \dots, y^{(N)}$ be arbitrary real numbers and let $\widehat{\mathcal{Y}} = \left\{\overline{y} = \sum_{i=1}^N p_i y^{(i)} : p \in \Delta_N\right\}$. Then $\ell_t(\overline{y}) = \frac{1}{2}(y_t - \overline{y})^2$ is $\alpha_t$-Exp-Concave on $\widehat{\mathcal{Y}}$ for $\alpha_t \le \frac{1}{2\max_i \ell_t(y^{(i)})}$.*

*Proof.* Letting $f_t(\overline{y}) = \exp(-\alpha_t \ell_t(\overline{y}))$ we have for any $\overline{y} \in \widehat{\mathcal{Y}}$:

$$\begin{aligned}
f_t'(\overline{y}) &= \left[\exp\left(-\frac{\alpha_t}{2}(y_t - \overline{y})^2\right)\right]' = \exp\left(-\frac{\alpha_t}{2}(y_t - \overline{y})^2\right)\alpha_t(y_t - \overline{y}) \\
f_t''(\overline{y}) &= \exp\left(-\frac{\alpha_t}{2}(y_t - \overline{y})^2\right)\left[\alpha_t^2(y_t - \overline{y})^2 - \alpha_t\right] \\
&= \exp\left(-\frac{\alpha_t}{2}(y_t - \overline{y})^2\right)\left[2\alpha_t^2\ell_t(\overline{y}) - \alpha_t\right]
\end{aligned}$$

Hence for $\alpha_t \le \frac{1}{2\max_i \ell_t(y^{(i)})}$ we have

$$f_t''(\overline{y}) \le \exp\left(-\frac{\alpha_t}{2}(y_t - \overline{y})^2\right)\alpha_t\left[2\alpha_t\ell_t(\overline{y}) - 1\right] \le 0$$

so $f_t(\overline{y}) = \exp(-\alpha_t\ell_t(\overline{y}))$ is concave and $\ell_t$ is $\alpha_t$-Exp-Concave over $\widehat{\mathcal{Y}}$ for $\alpha_t \le \frac{1}{2\max_i \ell_t(y^{(i)})}$. □

### Regret of the Range-Clipped Meta-Algorithm

In this section we prove a simple result showing that the range-clipping reduction described by Algorithm 13 incurs only an constant additional penalty. This lemma will be used to do most of the heavy-lifting in proving Theorem 10.3.1, which simply applies the following lemma and then chooses a specific meta-algorithm for $\mathcal{A}_{\text{Meta}}$.

**Lemma C.2.9.** *For any $[a,b] \subseteq [1,T]$, sequence $\boldsymbol{u} = (u_a, \dots, u_b)$ in $\mathbb{R}$, and $j \in [N]$, Algorithm 13 guarantees*

$$R_{[a,b]}(\boldsymbol{u}) \le \frac{1}{2}\max_t (y_t - y_t^{Ref})^2 + R_{[a,b]}^{\mathcal{A}_j}(\boldsymbol{u}) + R_{[a,b]}^{Meta}(e_j),$$

*where $R_{[a,b]}^{\mathcal{A}_j}(\boldsymbol{u}) = \sum_{t=a}^b \ell_t(w_t^{(j)}) - \ell_t(u_t)$ is the dynamic regret $\mathcal{A}_j$ and $R_{[a,b]}^{Meta}(e_j) = \sum_{t=a}^b \ell_t(\overline{y}_t) - \ell_t(\overline{y}_t^{(j)})$.*

*Proof.* For ease of notation we let $y_t^{(i)} = \left\langle x_t, w_t^{(i)}\right\rangle$, where $w_t^{(i)}$ is the output of algorithm $\mathcal{A}_i$, and slightly abuse notation by writing $\ell_t(y) = \frac{1}{2}(y_t - y)^2$ for $y \in \mathbb{R}$. Hence, we may write $\ell_t(w_t) \equiv \ell_t(y_t^{(i)})$

interchangeably. Note that this equivalence is valid in the improper online learning setting since the features are observed *before* the learner makes a prediction, as discussed in the introduction.

Now, for for any $j \in [N]$ we have

$$\begin{aligned}
R_{[a,b]}(\boldsymbol{u}) &= \sum_{t=a}^{b} \ell_t(\overline{y}_t) - \ell_t(u_t) \\
&= \sum_{t=a}^{b} \ell_t(w_t^{(j)}) - \ell_t(u_t) + \sum_{t=a}^{b} \ell_t(\overline{y}_t) - \ell_t(w_t^{(j)}) \\
&= R_{[a,b]}^{\mathcal{A}_j}(\boldsymbol{u}) + \sum_{t=a}^{b} \ell_t(\overline{y}_t) - \ell_t\left(y_t^{(j)}\right),
\end{aligned}$$

where we have observed $y_t^{(j)} = \left\langle x_t, w_t^{(j)} \right\rangle$. Observe that by Lemma C.2.7 we have

$$\begin{aligned}
\ell_t(y_t^{(j)}) &\geq \frac{1}{2}M_t^2 - \frac{1}{2}M_{t+1}^2 + \frac{1}{2}(y_t - \overline{y}_t^{(j)})^2 \\
&= \frac{1}{2}M_t^2 - \frac{1}{2}M_{t+1}^2 + \ell_t\left(\overline{y}_t^{(j)}\right),
\end{aligned}$$

where $M_t = \max_{s<t} \left|y_s - y_s^{\text{Ref}}\right|$. Hence,

$$\begin{aligned}
R_{[a,b]}(\boldsymbol{u}) &\leq R_{[a,b]}^{\mathcal{A}_j}(\boldsymbol{u}) + \sum_{t=a}^{b} \ell_t(\overline{y}_t) - \ell_t\left(y_t^{(j)}\right) \\
&\leq R_{[a,b]}^{\mathcal{A}_j}(\boldsymbol{u}) + \sum_{t=a}^{b} \ell_t(\overline{y}_t) - \ell_t\left(\overline{y}_t^{(j)}\right) \\
&\quad + \sum_{t=a}^{b} \frac{1}{2}M_{t+1}^2 - \frac{1}{2}M_t^2 \\
&\leq \frac{1}{2}M_{b+1}^2 + R_{[a,b]}^{\mathcal{A}_j}(\boldsymbol{u}) + \underbrace{\sum_{t=a}^{b} \ell_t(\overline{y}_t) - \ell_t\left(\overline{y}_t^{(j)}\right)}_{=:R_{[a,b]}^{\text{Meta}}(e_j)}
\end{aligned}$$

□

### Proof of Theorem 10.3.1

**Theorem 10.3.1.** *Let $\mathcal{A}_{Meta}$ be an instance of Algorithm 15 with $\alpha_t = \frac{1}{2\max_{t,i}\ell_t(\overline{y}_t^{(i)})}$, $\beta_{t+1} = \frac{1}{(e+t)\log^2(e+t)+1}$ and $p_1 = \mathbf{1}_N/N$. Then for any sequence $\boldsymbol{u} = (u_1,\ldots,u_T)$ in $\mathbb{R}$ and any $j \in [N]$, Algorithm 13 guarantees*

$$R_{[a,b]}(\boldsymbol{u}) \le O\left(R_{[a,b]}^{\mathcal{A}_j}(\boldsymbol{u}) + \max_t (y_t - \widetilde{y}_t)^2 \log\left(Nb\log^2(b)\right)\right),$$

*where $R_{[a,b]}$ denotes regret over the sub-interval $[a,b]$.*

*Proof.* The proof follows almost immediately using the regret guarantee of the range-clipped meta-algorithm (Lemma C.2.9), from which we have

$$R_{[a,b]}(\boldsymbol{u}) \le \frac{1}{2}\max_t (y_t - \widetilde{y}_t)^2 + R_{[a,b]}^{\mathcal{A}_j}(\boldsymbol{u}) + R_{[a,b]}^{\text{Meta}}(e_j).$$

Now applying the guarantee of an appropriate instance of the fixed-share algorithm (Theorem C.2.12 with $\alpha_t = \frac{1}{2\max_{t,i}\ell_t(\overline{y}_t^{(i)})}$, $\beta_t = \frac{1}{(e+t)\log^2(e+t)+1}$, and $p_1 = \mathbf{1}_N/N$), we have

$$\begin{aligned}
R_{[a,b]}^{\text{Meta}}(e_j) &\le \frac{1}{\alpha_{b+1}}\left[2\log\left(\frac{1}{\beta_{b+1}p_{1j}}\right) + 1\right] \\
&\le \max_{t,i}\ell_t(\overline{y}_t^{(i)})\left[2\log\left(((e+b)\log^2(e+b)+1)N\right)+1\right] \\
&\le O\left(\max_t (y_t - \widetilde{y}_t)^2 \log\left(b\log^2(b)N\right)\right),
\end{aligned}$$

where the last line applies Lemma C.2.7 and hides constants. All together, we have

$$R_{[a,b]}(\boldsymbol{u}) \le O\left(R_{[a,b]}^{\mathcal{A}_j}(\boldsymbol{u}) + \max_t (y_t - \widetilde{y}_t)^2 \log\left(Nb\log^2(b)\right)\right).$$

□

### Proof of Theorem 10.3.2

The proof of Theorem 10.3.2 follows by applying Theorem 10.3.1, and then showing that there exists a $\mathcal{A}_\gamma$ which attains the desired bound. We first provide proof of the latter claim in Lemma C.2.10 for the sake of modularity. In particular, we will also re-use this result to argue strongly-adaptive guarantees in Section 10.4. Proof of Theorem 10.3.2 is then easily proven at the end of this section.

**Lemma C.2.10.** *Let $b > 1$, $\eta_{\min} = 2d$, $\eta_{\max} = dT$, and define $\mathcal{S}_\eta = \left\{\eta_i = \eta_{\min} b^i \wedge \eta_{\max} : i = 0, 1, \dots\right\}$ and $\mathcal{S}_\gamma = \left\{\gamma_i = \frac{\eta_i}{1+\eta_i} : i = 0, 1, \dots\right\} \cup \{0\}$. For any $\gamma$ in $\mathcal{S}_\gamma$, let $\mathcal{A}_\gamma$ denote an instance of Algorithm 12 with discount $\gamma$. Then for any $\boldsymbol{u} = (u_1, \dots, u_T)$ in $\mathbb{R}^d$, there is a $\gamma^* \in [0, 1]$ satisfying $\gamma^* = \frac{\sqrt{d \sum_{t=1}^T \frac{1}{2}(y_t - \widetilde{y}_t)^2}}{\sqrt{d \sum_{t=1}^T \frac{1}{2}(y_t - \widetilde{y}_t)^2} + \sqrt{P_T^{\gamma^*}(\boldsymbol{u})}}$ and a $\gamma \in \mathcal{S}_\gamma$ such that*

$$R_T^{\mathcal{A}_\gamma}(\boldsymbol{u}) \le O\left(d \max_t (y_t - \widetilde{y}_t)^2 \log(T) + b\sqrt{d P_T^{\gamma^*}(\boldsymbol{u}) \sum_{t=1}^T (y_t - \widetilde{y}_t)^2}\right).$$

*Proof.* Denote $V_T = \frac{d}{2} \sum_{t=1}^T (y_t - \widetilde{y}_t)^2$. By Lemma C.2.5, there exists a $\gamma^* \in [0, 1]$ such that

$$\gamma^* = \frac{\sqrt{V_T}}{\sqrt{V_T} + \sqrt{P_T^{\gamma^*}(\boldsymbol{u})}}.$$

Throughout the proof it will be convenient to work in terms of the related quantity $\eta^* = \frac{\gamma^*}{1-\gamma^*} = \sqrt{\frac{V_T}{P_T^{\gamma}(\boldsymbol{u})}}$. Let us first suppose that $0 \le \eta^* \le \eta_{\min}$. In this case, we have

$$\eta^* = \sqrt{\frac{V_T}{P_T^{\gamma^*}(\boldsymbol{u})}} \le \eta_{\min} \implies \sqrt{\frac{1}{2} \sum_{t=1}^T (y_t - \widetilde{y}_t)^2} \le \eta_{\min} \sqrt{\frac{1}{d} P_T^{\gamma}(\boldsymbol{u})}.$$

Consider the algorithm $\mathcal{A}_0$ with $\gamma = 0$: in this case we have $w_t = \arg\min_{w \in \mathbb{R}^d} h_t(w)$, so $\langle x_t, w_t \rangle = \widetilde{y}_t$ and the regret is trivially

$$\begin{aligned}
\sum_{t=1}^T \ell_t(w_t^{\mathcal{A}_0}) - \ell_t(u_t) &\le \sum_{t=1}^T \frac{1}{2}(y_t - \widetilde{y}_t)^2 \\
&= \sqrt{\sum_{t=1}^T \frac{1}{2}(y_t - \widetilde{y}_t)^2 \sum_{t=1}^T \frac{1}{2}(y_t - \widetilde{y}_t)^2} \\
&\le \frac{\eta_{\min}}{\sqrt{d}} \sqrt{P_T^{\gamma^*}(\boldsymbol{u}) \sum_{t=1}^T \frac{1}{2}(y_t - \widetilde{y}_t)^2} \\
&= 2\sqrt{V_T P_T^{\gamma^*}(\boldsymbol{u})} \qquad \text{(C.24)}
\end{aligned}$$

for $\eta_{\min} = 2d$.

Otherwise, for $\eta^* \geq \eta_{\min}$, using Theorem 10.2.1 we have that for any $\gamma \in \mathcal{S}_\gamma$,

$$\begin{aligned}
R_T^{\mathcal{A}_\gamma}(\boldsymbol{u}) &\leq \frac{\gamma\lambda}{2}\left\|u_1\right\|_2^2 + \frac{d}{2}\max_t(y_t - \widetilde{y}_t)^2 \log\left(1 + \frac{\sum_{t=1}^T \gamma^{T-t}\left\|x_t\right\|_2^2}{\lambda d}\right) \\
&\quad + \gamma\sum_{t=1}^{T-1}\left[F_t^\gamma(u_{t+1}) - F_t^\gamma(u_t)\right] + \log\left(1/\gamma\right)V_T \\
&\overset{(*)}{\leq} \frac{\gamma\lambda}{2}\left\|u_1\right\|_2^2 + \frac{d}{2}\max_t(y_t - \widetilde{y}_t)^2 \log\left(1 + \frac{\sum_{t=1}^T \gamma^{T-t}\left\|x_t\right\|_2^2}{\lambda d}\right) \\
&\quad + \eta^* P_T^{\gamma^*}(\boldsymbol{u}) + \frac{V_T}{\eta}
\end{aligned}$$

where $(*)$ observes that $\eta_{\min} = \frac{\gamma_{\min}}{1-\gamma_{\min}} \leq \eta^* = \frac{\gamma^*}{1-\gamma^*} \implies \gamma_{\min} \leq \gamma^*$ and applies Lemma 10.2.2 (with $\beta = \gamma^*$) and substitutes $\eta = \frac{\gamma}{1-\gamma}$. If $\eta^* \geq \eta_{\max}$ then choosing $\eta = \eta_{\max} = dT$ yields

$$\frac{V_T}{\eta} = \frac{d}{2dT}\sum_{t=1}^T(y_t - \widetilde{y}_t)^2 \leq \frac{1}{2}\max_t(y_t - \widetilde{y}_t)^2,$$

and otherwise, there is an $\eta_k$ in $\mathcal{S}_\eta$ such that $\eta_k \leq \eta^* \leq b\eta_k$, so choosing $\eta = \eta_k$ yields

$$\frac{V_T}{\eta_k} \leq b\frac{V_T}{\eta^*} = b\sqrt{P_T^{\gamma^*}(\boldsymbol{u})V_T}$$

Hence, overall we have that there is a $\gamma \in \mathcal{S}_\gamma$ such that

$$\begin{aligned}
R_T^{\mathcal{A}_\gamma}(\boldsymbol{u}) &\leq \frac{\gamma\lambda}{2}\left\|u_1\right\|_2^2 + \frac{1}{2}\max_t(y_t - \widetilde{y}_t)^2\left[d\log\left(1 + \frac{\sum_{t=1}^T \gamma^{T-t}\left\|x_t\right\|_2^2}{\lambda d}\right) \vee 1\right] + \eta^* P_T^{\gamma^*}(\boldsymbol{u}) + b\frac{V_T}{\eta^*} \\
&= \frac{\gamma\lambda}{2}\left\|u_1\right\|_2^2 + \frac{1}{2}\max_t(y_t - \widetilde{y}_t)^2\left[d\log\left(1 + \frac{\sum_{t=1}^T \gamma^{T-t}\left\|x_t\right\|_2^2}{\lambda d}\right) \vee 1\right] + (b+1)\sqrt{V_T P_T^{\gamma^*}(\boldsymbol{u})} \\
&\leq O\left(d\max_t(y_t - \widetilde{y}_t)^2\log\left(T\right) \vee b\sqrt{dP_T^{\gamma^*}(\boldsymbol{u})\sum_{t=1}^T(y_t - \widetilde{y}_t)^2}\right).
\end{aligned}$$

□

With the previous lemma in hand, the proof of Theorem 10.3.2 follows easily. The theorem is re-stated for convenience.

**Theorem 10.3.2.** *Let $b > 1$, $\eta_{\min} = 2d$, $\eta_{\max} = dT$, and for all $i \in \mathbb{N}$ let $\eta_i = \eta_{\min} b^i \wedge \eta_{\max}$, and construct the set of discount factors $\mathcal{S}_\gamma = \left\{\gamma_i = \frac{\eta_i}{1+\eta_i} : i \in \mathbb{N}\right\} \cup \{0\}$. For any $\gamma$ in $\mathcal{S}_\gamma$, let $\mathcal{A}_\gamma$ denote an instance of Algorithm 12 with discount $\gamma$.[1] Let $\mathcal{A}_{Meta}$ be an instance of the algorithm characterized in Theorem 10.3.1, and suppose we set $y_t^{Ref} = \widetilde{y}_t$ for all $t$. Then for any $\boldsymbol{u} = (u_1, \ldots, u_T)$ in $\mathbb{R}^d$, Algorithm 13 guarantees*

$$R_T(\boldsymbol{u}) \le O\left(d \max_t (y_t - y_t^{Ref})^2 \log(T) + b\sqrt{dP_T^{\gamma^*}(\boldsymbol{u}) \sum_{t=1}^T (y_t - \widetilde{y}_t)^2}\right)$$

*where $\gamma^* \in [0,1]$ satisfies Equation (10.2).*

*Proof.* Applying Theorem 10.3.1, for any sequence $\boldsymbol{u} = (u_1, \ldots, u_T)$ in $\mathbb{R}^d$ and any $\gamma \in \mathcal{S}_\gamma$ we have

$$\begin{aligned} R_T(\boldsymbol{u}) &\le \widehat{O}\left(R_T^{\mathcal{A}_\gamma}(\boldsymbol{u}) + \max_t (y_t - y_t^{\text{Ref}})^2 \log(NT)\right) \\ &\le \widehat{O}\left(R_T^{\mathcal{A}_\gamma}(\boldsymbol{u}) + \max_t (y_t - y_t^{\text{Ref}})^2 \log(T)\right), \end{aligned} \tag{C.25}$$

where the last line uses $N = |\mathcal{S}_\gamma| = \log_b(\eta_{\max}/\eta_{\min}) \le O(\log_b(T))$, then hides log(log) factors. Finally, by Lemma C.2.10, there is indeed a $\gamma^* \in [0,1]$ satisfying $\gamma^* = \frac{\sqrt{d\sum_{t=1}^T \frac{1}{2}(y_t - \widetilde{y}_t)^2}}{\sqrt{d\sum_{t=1}^T \frac{1}{2}(y_t - \widetilde{y}_t)^2} + \sqrt{P_T^{\gamma^*}(\boldsymbol{u})}}$ and a $\gamma \in \mathcal{S}_\gamma$ such that

$$R_T^{\mathcal{A}_\gamma}(\boldsymbol{u}) \le O\left(d \max_t (y_t - \widetilde{y}_t)^2 \log(T) + b\sqrt{dP_T^{\gamma^*}(\boldsymbol{u}) \sum_{t=1}^T (y_t - \widetilde{y}_t)^2}\right).$$

Plugging this back into Equation (C.25) and choosing $y_t^{\text{Ref}} = \widetilde{y}_t$ proves the result. □

### Proof of Theorem 10.3.3

As in Appendix C.2.4, the proof of Theorem 10.3.3 follows by applying Theorem 10.3.1 and then showing that there is a $\mathcal{A}_\gamma$ attaining the desired regret bound. We first provide proof of the latter claim in Lemma C.2.11 for the sake of modularity, so that we can use it when arguing strongly-adaptive guarantees in Section 10.4. Proof of Theorem 10.3.3 is proven at the end of this section.

[1] For brevity, here we refer to an algorithm that directly predicts $\widetilde{y}_t$ on every round as being an instance of the discounted VAW forecaster with $\gamma = 0$. This terminology can be justified by Remark C.2.2, but for our purposes here it's sufficient to consider it convenient alias.

**Lemma C.2.11.** *Under the same conditions as Lemma C.2.10, suppose each $\mathcal{A}_\gamma$ sets hints $\widetilde{y}_t = \overline{y}_t^\gamma = \mathrm{Clip}_{\mathcal{B}_t}(\langle x_t, w_t^\gamma\rangle)$, where $\mathcal{B}_t = [y_t^{Ref} - M_t, y_t^{Ref} + M_t]$ and $M_t = \max_{s<t}\left|y_s - y_s^{Ref}\right|$. Then for any $\boldsymbol{u} = (u_1, \ldots, u_T)$ in $W$, there is a $\gamma^\circ \in [0,1]$ satisfying $\gamma^\circ = \frac{\sqrt{d\sum_{t=1}^T \ell_t(u_t)}}{\sqrt{d\sum_{t=1}^T \ell_t(u_t)} + \sqrt{dP_T^{\gamma^\circ}(\boldsymbol{u})}}$ and a $\gamma \in \mathcal{S}_\gamma$ such that*

$$R_T^{\mathcal{A}_\gamma}(\boldsymbol{u}) \le O\Bigg(dP_T^{\gamma_{\min}}(\boldsymbol{u}) + d\max_t\left(y_t - y_t^{Ref}\right)^2\log(T) \\ + b\sqrt{dP_T^{\gamma^\circ}(\boldsymbol{u})\sum_{t=1}^T \ell_t(u_t)}\Bigg),$$

*where $\gamma^{\min} = \min\{\gamma \in \mathcal{S}_\gamma\} = \frac{2d}{2d+1}$.*

*Proof.* Using Lemma C.2.6, for any $\boldsymbol{u} = (u_1, \ldots, u_T)$, $\gamma \in (0,1)$, and $\beta \ge \gamma \ge \gamma_{\min} = \frac{2d}{2d+1}$, we have

$$R_T(\boldsymbol{u}) \le \gamma\lambda\|u_1\|_2^2 + 4d\max_t(y_t - y_t^{\mathrm{Ref}})^2\log\left(1 + \frac{\sum_{t=1}^T \gamma^{T-t}\|x_t\|_2^2}{\lambda d}\right) \\ + 2\frac{\beta}{1-\beta}P_T^\beta(\boldsymbol{u}) + \frac{1-\gamma}{\gamma}2d\sum_{t=1}^T \ell_t(u_t),$$

We will proceed by showing that there is a $\beta$ and $\gamma$ that suitably balances the summations in the last line. To this end, recall that by Lemma C.2.5, there is a $\gamma^\circ$ satisfying

$$\gamma^\circ = \frac{\sqrt{d\sum_{t=1}^T \ell_t(u_t)}}{\sqrt{d\sum_{t=1}^T \ell_t(u_t)} + \sqrt{P_T^{\gamma^\circ}(\boldsymbol{u})}}$$

Denote $\eta = \frac{\gamma}{1-\gamma}$ and $\eta^\circ = \frac{\gamma^\circ}{1-\gamma^\circ} = \sqrt{\frac{d\sum_{t=1}^T \ell_t(u_t)}{P_T^{\gamma^\circ}(\boldsymbol{u})}}$. If $\eta^\circ \ge \eta_{\max} = \frac{\gamma_{\max}}{1-\gamma_{\max}}$, then we can take $\beta = \gamma^\circ$ and $\gamma = \gamma_{\max}$ to get

$$\begin{aligned}\frac{\beta}{1-\beta}P_T^\beta(\boldsymbol{u}) + \frac{\gamma}{1-\gamma}d\sum_{t=1}^T \ell_t(u_t) &= \eta^\circ P_T^{\gamma^\circ}(\boldsymbol{u}) + \frac{d\sum_{t=1}^T \ell_t(u_t)}{\eta_{\max}} \\ &= \sqrt{dP_T^{\gamma^\circ}(\boldsymbol{u})\sum_{t=1}^T \ell_t(u_t)} + \frac{d\sum_{t=1}^T \ell_t(u_t)}{\eta_{\max}} \\ &\le \sqrt{dP_T^{\gamma^\circ}(\boldsymbol{u})\sum_{t=1}^T \ell_t(u_t)} + \max_t \ell_t(u_t),\end{aligned}$$

where the last line recalls $\eta_{\max} = dT$. Otherwise, if $\eta^\circ \le \eta_{\min} = \frac{\gamma_{\min}}{1-\gamma_{\min}} = 2d$, then taking $\beta = \gamma = \gamma_{\min}$

yields

$$\eta_{\min} P_T^{\gamma_{\min}}(\boldsymbol{u}) + \frac{d\sum_{t=1}^T \ell_t(u_t)}{\eta_{\min}} \leq \eta_{\min} P_T^{\gamma_{\min}}(\boldsymbol{u}) + \frac{d\sum_{t=1}^T \ell_t(u_t)}{\eta^\circ}$$

$$= 2dP_T^{\gamma_{\min}}(\boldsymbol{u}) + \sqrt{dP_T^{\gamma^\circ}(\boldsymbol{u})\sum_{t=1}^T \ell_t(u_t)}.$$

Lastly, if $\eta_{\min} \leq \eta^\circ \leq \eta_{\max}$, there is a $\eta_k = \frac{\gamma_k}{1-\gamma_k} \in \mathcal{S}_\eta$ such that $\eta_k \leq \eta^\circ \leq b\eta_k$, so choosing $\beta = \gamma^\circ$ and $\gamma = \gamma_k$ yields

$$\eta^\circ P_T^{\gamma^\circ}(\boldsymbol{u}) + \frac{d\sum_{t=1}^T \ell_t(u_t)}{\eta_k} \leq \eta^\circ P_T^{\gamma^\circ}(\boldsymbol{u}) + b\frac{d\sum_{t=1}^T \ell_t(u_t)}{\eta^\circ}$$

$$= (b+1)\sqrt{dP_T^{\gamma^\circ}(\boldsymbol{u})\sum_{t=1}^T \ell_t(u_t)}$$

Combining the three cases, we have

$$2\frac{\beta}{1-\beta}P_T^{\beta}(\boldsymbol{u}) + \frac{1-\gamma}{\gamma}2d\sum_{t=1}^T \ell_t(u_t) \leq 4dP_T^{\gamma_{\min}}(\boldsymbol{u}) + 2\max_t \ell_t(u_t) + 2(b+1)\sqrt{dP_T^{\gamma^\circ}(\boldsymbol{u})\sum_{t=1}^T \ell_t(u_t)}$$

Hence, overall the regret can be bound as

$$R_T^{\mathcal{A}_\gamma}(\boldsymbol{u}) \leq \gamma\lambda\left\|u_1\right\|_2^2 + d\max_t(y_t - \overline{y}_t^\gamma)^2\log\left(1 + \frac{\sum_{t=1}^T \gamma^{T-t}\left\|x_t\right\|_2^2}{\lambda d}\right)$$

$$+ 4dP_T^{\gamma_{\min}}(\boldsymbol{u}) + 2\max_t \ell_t(u_t) + 2(b+1)\sqrt{dP_T^{\gamma^\circ}(\boldsymbol{u})\sum_{t=1}^T \ell_t(u_t)}$$

$$\leq O\left(dP_T^{\gamma_{\min}}(\boldsymbol{u}) + d\max_t(y_t - y_t^{\mathrm{Ref}})^2\log(T) + b\sqrt{dP_T^{\gamma^\circ}(\boldsymbol{u})\sum_{t=1}^T \ell_t(u_t)}\right),$$

where we've applied Lemma C.2.7 to bound $\max_t(y_t - \overline{y}_t^\gamma)^2 \leq 4M_{T+1}^2 = 4\max_t(y_t - y_t^{\mathrm{Ref}})^2$. Plugging this back into Equation (C.26) proves the stated bound. □

Now the proof of Theorem 10.3.3 follows by composing Theorem 10.3.1 and Lemma C.2.11. The theorem is restated below for convenience.

**Theorem 10.3.3.** *Under the same conditions as Theorem 10.3.2, suppose each $\mathcal{A}_\gamma$ sets hints $\widetilde{y}_t = \overline{y}_t^\gamma = \mathrm{Clip}_{\mathcal{B}_t}(\langle, x_t, w_t^\gamma\rangle)$, where $\mathcal{B}_t = [y_t^{Ref} - M_t, y_t^{Ref} + M_t]$ and $M_t = \max_{s<t} \left| y_s - y_s^{Ref} \right|$. Then for any $\boldsymbol{u} = (u_1, \ldots, u_T)$ in $\mathbb{R}^d$, Algorithm 13 guarantees*

$$R_T(\boldsymbol{u}) \le O\left( dP_T^{\gamma_{\min}}(\boldsymbol{u}) + d\max_t \left( y_t - y_t^{Ref} \right)^2 \log(T) + b\sqrt{dP_T^{\gamma^\circ}(\boldsymbol{u}) \sum_{t=1}^T \ell_t(u_t)} \right)$$

*where $\gamma_{\min} = \frac{2d}{2d+1}$ and $\gamma^\circ \in [0,1]$ satisfies Equation (10.3).*

*Proof.* As in the proof of Theorem 10.3.2, we apply Theorem 10.3.1, from which it follows that for any $\boldsymbol{u} = (u_1, \ldots, u_T)$ in $\mathbb{R}^d$ and any $\gamma \in \mathcal{S}_\gamma$, the dynamic regret is bounded as

$$\begin{aligned} R_T(\boldsymbol{u}) &\le \widehat{O}\left( R_T^{\mathcal{A}_\gamma}(\boldsymbol{u}) + \max_t (y_t - y_t^{\mathrm{Ref}})^2 \log(NT) \right) \\ &\le \widehat{O}\left( R_T^{\mathcal{A}_\gamma}(\boldsymbol{u}) + \max_t (y_t - y_t^{\mathrm{Ref}})^2 \log(T) \right), \end{aligned} \tag{C.26}$$

where the last line uses $N = |\mathcal{S}_\gamma| = \log_b(\eta_{\max}/\eta_{\min}) \le O(\log_b(T))$, then hides log(log) factors. And using Lemma C.2.11, for any $\boldsymbol{u} = (u_1, \ldots, u_T)$ there is a $\gamma^\circ \in [0,1]$ satisfying $\gamma^\circ = \frac{\sqrt{d\sum_{t=1}^T \ell_t(u_t)}}{\sqrt{d\sum_{t=1}^T \ell_t(u_t)} + \sqrt{P_T^\circ(\boldsymbol{u})}}$ and a $\gamma \in \mathcal{S}_\gamma$ such that

$$R_T^{\mathcal{A}_\gamma}(\boldsymbol{u}) \le O\left( dP_T^{\gamma_{\min}}(\boldsymbol{u}) + d\max_t (y_t - y_t^{\mathrm{Ref}})^2 \log(T) + b\sqrt{dP_T^{\gamma^\circ}(\boldsymbol{u}) \sum_{t=1}^T \ell_t(u_t)} \right),$$

Plugging this back into Equation (C.26) completes the proof. □

### C.2.5 Proofs for Section 3.2 (Strongly-Adaptive Guarantees)

In this section we provide a formal statement of the result sketched in Section 10.4. The result follows easily from the results in Section 10.3, after borrowing the geometric covering intervals from Daniely, Gonen, and Shalev-Shwartz (2015).

**Theorem 10.4.1.** *Let $\mathcal{S}_\gamma$ be the set of discount factors defined in Theorem 10.3.2, let $S$ denote a set of geometric covering intervals over $[1,T]$, and for each $\gamma \in \mathcal{S}_\gamma$ and $I \in S$, let $\mathcal{A}_{\gamma,I}$ be an instance of Algorithm 12 using discount $\gamma$ and applied during interval $I$. Let $\mathcal{A}_{Meta}$ be an instance of the meta-algorithm characterized in Theorem 10.3.1. Then for any $[s,\tau] \subseteq [1,T]$, there is a set of disjoint intervals $I_1,\ldots,I_K$ in $S$ such that $\cup_{i=1}^K I_i = [s,\tau]$, and moreover, for any $\boldsymbol{u} = (u_s,\ldots,u_\tau)$ Algorithm 13 with $y_t^{Ref} = \widetilde{y}_t$ guarantees*

$$R_{[s,\tau]}(\boldsymbol{u}) \le \widehat{O}\left(d\max_t(y_t - y_t^{Ref})^2\log^2(T) + b\sqrt{dP_{[s,\tau]}^{\gamma^*}(\boldsymbol{u})\sum_{t\in[s,\tau]}(y_t-\widetilde{y}_t)^2}\right)$$

*where $P_{[s,\tau]}^{\gamma^*}(\boldsymbol{u}) = \sum_{i=1}^K P_{I_i}^{\gamma_i^*}(\boldsymbol{u})$ and each $\gamma_i^* \in [0,1]$ satisfies $\gamma_i^* = \frac{\sqrt{\frac{d}{2}\sum_{t\in I_i}(y_t-\widetilde{y}_t)^2}}{\sqrt{\frac{d}{2}\sum_{t\in I_i}(y_t-\widetilde{y}_t)^2}+\sqrt{P_{I_i}^{\gamma_i^*}(\boldsymbol{u})}}$.*

*If we instead suppose each $\mathcal{A}_{\gamma,I}$ sets hints as in Theorem 10.3.3, then for any $\boldsymbol{u} = (u_s,\ldots,u_\tau)$ Algorithm 13 guarantees*

$$R_{[s,\tau]}(\boldsymbol{u}) \le \widehat{O}\left(dP_{[s,\tau]}^{\gamma_{\min}}(\boldsymbol{u}) + d\max_t(y_t - y_t^{Ref})^2\log^2(T) + b\sqrt{dP_{[s,\tau]}^{\gamma^\circ}(\boldsymbol{u})\sum_{t\in[s,\tau]}\ell_t(u_t)}\right)$$

*where $P_{[s,\tau]}^{\gamma^\circ}(\boldsymbol{u}) = \sum_{i=1}^K P_{I_i}^{\gamma_i^\circ}(\boldsymbol{u})$ and each $\gamma_i^\circ \in [0,1]$ satisfies $\gamma_i^\circ = \frac{\sqrt{d\sum_{t\in I_i}\ell_t(u_t)}}{\sqrt{d\sum_{t\in I_i}\ell_t(u_t)}+\sqrt{P_{I_i}^{\gamma_i^\circ}(\boldsymbol{u})}}$.*

*Proof.* For any $[s,\tau] \subseteq [1,T]$, Daniely, Gonen, and Shalev-Shwartz (2015, Lemma 1.2) shows that there exists a disjoint set of intervals $I_1,\ldots,I_K$ in $S$ such that $\cup_{i=1}^K I_i = [s,\tau]$ and $K \le O(\log(\tau - s))$. Hence, we can decompose $\sum_{i=1}^K R_{I_i}(\boldsymbol{u})$, so applying Theorem 10.3.1 to each of these sub-intervals, for any $\gamma_1,\ldots\gamma_k \in \mathcal{S}_\gamma$ we have:

$$\begin{aligned} R_{[s,\tau]}(\boldsymbol{u}) = \sum_{i=1}^K R_{I_i}(\boldsymbol{u}) &\le \sum_{i=1}^K \widehat{O}\left(R_{I_i}^{\mathcal{A}_{\gamma_i,I_i}}(\boldsymbol{u}) + \max_t(y_t-\widetilde{y}_t)^2\log(N|I_i|)\right) \\ &\le \widehat{O}\left(\sum_{i=1}^K R_{I_i}^{\mathcal{A}_{\gamma_i,I_i}}(\boldsymbol{u}) + K\max_t(y_t-\widetilde{y}_t)^2\log(N(\tau-s))\right) \\ &\le \widehat{O}\left(\sum_{i=1}^K R_{I_i}^{\mathcal{A}_{\gamma_i,I_i}}(\boldsymbol{u}) + \max_t(y_t-\widetilde{y}_t)^2\log^2(T)\right), \end{aligned} \tag{C.27}$$

where $\widehat{O}(\cdot)$ hides log(log) factors and the last line bounds $K \le O(\log(\tau - s)) \le O(\log(T))$ and

$N \le O(T\log(T))$. The bound on $N$ can be seen from the fact that $|\mathcal{S}_\gamma| \le O(\log(T))$, and from the fact that $S$ is constructed as $S = \cup_{i=1}^{\lfloor \log(T)\rfloor} S_i$ where $S_i = \left\{[k2^i, (k+1)2^i - 1] : k = 0, 1, \dots\right\}$, from which it is easily seen that $|S| \le O(T)$ by observing that each $S_i$ has at most $T/2^i$ intervals, hence summing them all up yields $|S| = \sum_{i=1}^{\lfloor \log(T)\rfloor} |S_i| \le O(T)$.

Now for any interval $I_i$, Lemma C.2.10 shows that there is a $\gamma_i^* \in [0,1]$ satisfying

$$\gamma_i^* = \frac{\sqrt{d\sum_{t\in I_i}\frac{1}{2}(y_t - \widetilde{y}_t)^2}}{\sqrt{d\sum_{t\in I_i}\frac{1}{2}(y_t - \widetilde{y}_t)^2} + \sqrt{P_{I_i}^{\gamma_i^*}(\boldsymbol{u})}}$$

and a $\gamma_i \in \mathcal{S}_\gamma$ such that

$$R_{I_i}^{\mathcal{A}_{\gamma_i, I_i}}(\boldsymbol{u}) \le O\left(d\max_t (y_t - \widetilde{y}_t)^2 \log(|I_i|) + b\sqrt{dP_{I_i}^{\gamma_i^*}(\boldsymbol{u})\sum_{t\in I_i}(y_t - \widetilde{y}_t)^2}\right)$$

so summing these up and applying Cauchy-Schwarz inequlity leads to

$$\begin{aligned}\sum_{i=1}^K R_{I_i}^{\mathcal{A}_{\gamma, I_i}}(\boldsymbol{u}) &\le O\left(Kd\max_t (y_t - \widetilde{y}_t)^2 \log(|I_i|) + \sum_{i=1}^K b\sqrt{dP_{I_i}^{\gamma_i^*}(\boldsymbol{u})\sum_{t\in I_i}(y_t - \widetilde{y}_t)^2}\right)\\ &\le O\left(d\max_t (y_t - \widetilde{y}_t)^2 \log^2(\tau - s) + b\sqrt{dP_{[s,\tau]}^{\gamma^*}(\boldsymbol{u})\sum_{t\in[s,\tau]}(y_t - \widetilde{y}_t)^2}\right)\end{aligned}$$

where we've defined $P_{[s,\tau]}^{\gamma^*}(\boldsymbol{u}) = \sum_{i=1}^K P_{I_i}^{\gamma_i^*}(\boldsymbol{u})$. Plugging this back into Equation (C.27), overall we may bound:

$$R_{[s,\tau]}(\boldsymbol{u}) \le \widehat{O}\left(d\max_t (y_t - y_t^{\text{Ref}})^2 \log^2(T) + b\sqrt{dP_{[s,\tau]}^{\gamma^*}(\boldsymbol{u})\sum_{t\in[s,\tau]}(y_t - \widetilde{y}_t)^2}\right)$$

where we've chosen $\widetilde{y}_t = y_t^{\text{Ref}}$ for simplicity.

An identical argument holds for the second statement: for any interval $I_i$, Lemma C.2.11 shows that there is a $\gamma_i^\circ \in [0,1]$ satisfying $\gamma_i^\circ = \frac{\sqrt{d\sum_{t\in I_i}\ell_t(u_t)}}{\sqrt{d\sum_{t\in I_i}\ell_t(u_t)} + \sqrt{P_{I_i}^{\gamma_i^\circ}(\boldsymbol{u})}}$ and a $\gamma_i \in \mathcal{S}_\gamma$ such that

$$R_{I_i}^{\mathcal{A}_{\gamma_i, I_i}}(\boldsymbol{u}) \le O\left(dP_{I_i}^{\gamma_{\min}}(\boldsymbol{u}) + d\max_t (y_t - y_t^{\text{Ref}})^2 \log(|I_i|) + b\sqrt{dP_{I_i}^{\gamma_i^\circ}(\boldsymbol{u})\sum_{t\in I_i}\ell_t(u_t)}\right)$$

so summing these up and applying Cauchy-Schwarz inequality again leads to

$$\begin{aligned}\sum_{i=1}^{K} R_{I_i}^{\mathcal{A}_{\gamma,I_i}}(\boldsymbol{u}) &\le O\left(dP_{[s,\tau]}^{\gamma_{\min}}(\boldsymbol{u}) + Kd\max_t(y_t-\widetilde{y}_t)^2\log\left(|I_i|\right) + \sum_{i=1}^{K} b\sqrt{dP_{I_i}^{\gamma^\circ}(\boldsymbol{u})\sum_{t\in I_i}\ell_t(u_t)}\right)\\ &\le O\left(dP_{[s,\tau]}^{\gamma_{\min}}(\boldsymbol{u}) + d\max_t(y_t-\widetilde{y}_t)^2\log^2(\tau-s) + b\sqrt{dP_{[s,\tau]}^{\gamma^\circ}(\boldsymbol{u})\sum_{t\in[s,\tau]}\ell_t(u_t)}\right)\end{aligned}$$

where we've defined $P_{[s,\tau]}^{\gamma^*}(\boldsymbol{u}) = \sum_{i=1}^{K} P_{I_i}^{\gamma_i^*}(\boldsymbol{u})$, so plugging this back into Equation (C.27), overall we may bound:

$$R_{[s,\tau]}(\boldsymbol{u}) \le \widehat{O}\left(dP_{[s,\tau]}^{\gamma_{\min}}(\boldsymbol{u}) + d\max_t(y_t-y_t^{\text{Ref}})^2\log^2(T) + b\sqrt{dP_{[s,\tau]}^{\gamma^\circ}(\boldsymbol{u})\sum_{t\in[s,\tau]}\ell_t(u_t)}\right),$$

where we've defined $P_{[s,\tau]}^{\gamma^\circ} = \sum_{i=1}^{K} P_{I_i}^{\gamma_i^\circ}(\boldsymbol{u})$.

□

### Matching the Exp-concave Guarantee in Unbounded Domains

Recall from Section 10.2.2 that in the Exp-concave setting, the algorithm of Baby and Y.-X. Wang (2021) achieves a dynamic regret bound of the form $R_T(\boldsymbol{u}) \le \widetilde{O}\left(T^{1/3}C_T^{2/3}\right)$ for $C_T = \sum_{t=1}^{T-1}\|u_t - u_{t-1}\|_1$. Our strongly-adaptive guarantees in Theorem 10.4.1 show that a bound of this form can be achieved even in the unbounded domain setting. To see why, note that the essential intuition of Baby and Y.-X. Wang (2021) is that if we have access to a *strongly-adaptive* algorithm guaranteeing $R_{[a,b]}(u) \le O(\log(b-a))$ *static* regret on all intervals $[a,b] \subseteq [1,T]$, then to attain the desired bound up to log terms it suffices to show that there *exists* a set of intervals $\{I_1,\ldots,I_N\}$ partitioning $[1,T]$ such that $N \le T^{1/3}C_T^{2/3}$ and that the dynamic regret is bounded by the static regrets over the partition, leading to regret matching $O(T^{1/3}C_T^{2/3})$ up to logarithmic terms.

Our strongly-adaptive guarantee in Theorem 10.4.1 actually achieves a stronger guarantee than is necessary to invoke the above argument, by guaranteeing $O\big(\log(b-a)\vee\sqrt{dP_{[a,b]}^{\gamma}(\boldsymbol{u})|b-a|}\big)$ *dynamic* regret on every interval $[a,b]$, and hence as a special case we have $O(\log(b-a))$ static regret on each interval as well. A similar partitioning argument then provides an analogous $T^{1/3}C_T^{2/3}$ bound, even in unbounded domains. If this is surprising, note that the exp-concave (and hence bounded domain) restriction is only really used to provide an algorithm which achieves logarithmic static regret, not to construct the essential partition. In the online linear regression setting, we do not need exp-concavity to guarantee logarithmic static regret — the VAW forecaster can provide the necessary guarantee even in an unbounded domain.

### C.2.6 Adaptive Fixed-share

**Algorithm 15:** Adaptive Fixed-Share

```
**Input**: Experts $\mathcal{A}_1, \dots, \mathcal{A}_N$, $p_1 \in \Delta_N$
**for** $t = 1 : T$ **do**
    Get $y_t^{(i)}$ from $\mathcal{A}_i$ for all $i$
    Play $\overline{y}_t = \sum_{i=1}^N p_{ti} y_t^{(i)}$
    Observe loss $\ell_t(y) = \frac{1}{2}(y_t - y)^2$ and let $\ell_{ti} = \ell_t(y_t^{(i)})$ for all $i$
    Let $q_{t+1,i} = \frac{p_{ti}\exp(-\alpha_t \ell_{ti})}{\sum_{j=1}^N p_{tj}\exp(-\alpha_t \ell_{tj})}$ for all $i$
    Choose $\beta_{t+1}$ and set $p_{t+1} = (1 - \beta_{t+1}) q_{t+1} + \beta_{t+1} p_1$
**end**
```

In this section, we provide for completeness analysis related to the fixed-share algorithm Nicolo Cesa-Bianchi, Gaillard, et al. 2012 with time-varying modulus. The following is a modest generalization of the analysis of Hazan (2019, Theorem 10.3). Throughout this section we assume that the losses $\ell_t : \widehat{\mathbb{Y}} \to \mathbb{R}$ are exp-concave in their domain.

**Theorem C.2.12.** *For all $t$ let $\ell_t$ be an $\alpha_t$-Exp-Concave function and assume that $\alpha_t \geq \alpha_{t+1}$ for all $t$. For all $t$, set $\beta_t \leq \frac{1}{(e+t)\log^2(e+t)+1}$. Then for any $j \in [N]$ and any $[a, b] \subseteq [1, T]$, Algorithm 15 guarantees*

$$\sum_{t=a}^{b} \ell_t(\overline{y}_t) - \ell_t(y_t^{(j)}) \leq \frac{1}{\alpha_{b+1}}\left[2\log\left(\frac{1}{\beta_{b+1} p_{1j}}\right) + 1\right]$$

*Proof.* The heavy lifting is done mostly using Lemma C.2.13, after which the proof follows by choosing the sequence of mixing parameters $\beta_t$. Applying Lemma C.2.13 and observing the telescoping sum, we have

$$\begin{aligned}\sum_{t=a}^{b} \ell_t(\overline{y}_t) - \ell_t\left(y_t^{(j)}\right) &\leq \sum_{t=a}^{b} \frac{1}{\alpha_t} \log\left(\frac{1}{p_{tj}}\right) - \frac{1}{\alpha_{t+1}} \log\left(\frac{1}{p_{t+1,j}}\right) \\ &\quad + \sum_{t=a}^{b} \frac{1}{\alpha_t} \log\left(\frac{1}{1-\beta_{t+1}}\right) \\ &\quad + \sum_{t=a}^{b} \left|\frac{1}{\alpha_{t+1}} - \frac{1}{\alpha_t}\right| \log\left(\frac{1}{\beta_{t+1} p_{1j}}\right) \\ &= \frac{1}{\alpha_a} \log\left(\frac{1}{p_{aj}}\right) - \frac{1}{\alpha_{b+1}} \log\left(\frac{1}{p_{b+1,j}}\right) \\ &\quad + \sum_{t=a}^{b} \frac{1}{\alpha_t} \log\left(\frac{1}{1-\beta_{t+1}}\right) \\ &\quad + \sum_{t=a}^{b} \left|\frac{1}{\alpha_{t+1}} - \frac{1}{\alpha_t}\right| \log\left(\frac{1}{\beta_{t+1} p_{1j}}\right).\end{aligned}$$

Now observe that with $\beta_{t+1} \leq \frac{1}{(e+t)\log^2(e+t)+1}$, using the elementary inequality $\log(1+y) \leq y$ we have

$$\log\left(\frac{1}{1-\beta_{t+1}}\right) = \log\left(1 + \frac{\beta_{t+1}}{1-\beta_{t+1}}\right) \leq \frac{\beta_{t+1}}{1-\beta_{t+1}} = \frac{1}{(e+t)\log^2(e+t)}$$

so for non-increasing $\alpha_t$ we have

$$\begin{aligned}\sum_{t=a}^{b} \frac{1}{\alpha_t} \log\left(\frac{1}{1-\beta_{t+1}}\right) &\leq \sum_{t=a}^{b} \frac{1}{\alpha_t} \frac{1}{(e+t)\log^2(e+t)} \\ &\leq \frac{1}{\alpha_b} \sum_{t=a}^{b} \frac{1}{(e+t)\log^2(e+t)} \\ &\leq \frac{1}{\alpha_b} \int_e^{e+b} \frac{1}{y\log^2 y} dy \\ &= \frac{1}{\alpha_b} \frac{-1}{\log(y)}\bigg|_{y=e}^{e+b} \leq \frac{1}{\alpha_b}\end{aligned}$$

and similarly,

$$\begin{aligned}\sum_{t=a}^{b} \left|\frac{1}{\alpha_{t+1}} - \frac{1}{\alpha_t}\right| \log\left(\frac{1}{\beta_{t+1} p_{1j}}\right) &\leq \log\left(\frac{1}{\beta_{b+1} p_{1j}}\right) \sum_{t=a}^{b} \frac{1}{\alpha_{t+1}} - \frac{1}{\alpha_t} \\ &\leq \frac{1}{\alpha_{b+1}} \log\left(\frac{1}{\beta_{b+1} p_{1j}}\right),\end{aligned}$$

so overall we have

$$\begin{aligned}
\sum_{t=a}^{b} \ell_t(\overline{y}_t) - \ell_t\left(y_t^{(j)}\right) &\leq \frac{1}{\alpha_a}\log\left(\frac{1}{p_{aj}}\right) - \frac{1}{\alpha_{b+1}}\log\left(\frac{1}{p_{b+1,j}}\right) + \frac{\log\left(\frac{1}{\beta_{b+1}p_{1j}}\right)+1}{\alpha_{b+1}} \\
&= \frac{1}{\alpha_a}\log\left(\frac{1}{p_{aj}}\right) + \frac{\log\left(\frac{p_{b+1,j}}{\beta_{b+1}p_{1j}}\right)+1}{\alpha_{b+1}} \\
&\leq \frac{1}{\alpha_{b+1}}\log\left(\frac{1}{(1-\beta_a)q_{aj}+\beta_a p_{1j}}\right) + \frac{\log\left(\frac{p_{b+1,j}}{\beta_{b+1}p_{1j}}\right)+1}{\alpha_{b+1}} \\
&\leq \frac{1}{\alpha_{b+1}}\left[2\log\left(\frac{1}{\beta_{b+1}p_{1j}}\right)+1\right] \qquad \leq
\end{aligned}$$

□

### Proof of Lemma C.2.13

The following provides an initial one-step bound to work from, which we use in the proof of Theorem C.2.12.

**Lemma C.2.13.** *For all t let $\ell_t$ be an $\alpha_t$-Exp-Concave function. Then for any $j \in [N]$, Algorithm 15 guarantees*

$$\begin{aligned}
\ell_t(\overline{y}_t) - \ell_t(y_t^{(j)}) \leq \frac{1}{\alpha_t}\log\left(\frac{1}{p_{tj}}\right) &- \frac{1}{\alpha_{t+1}}\log\left(\frac{1}{p_{t+1,j}}\right) \\
&+ \frac{1}{\alpha_t}\log\left(\frac{1}{1-\beta_{t+1}}\right) \\
&+ \left|\frac{1}{\alpha_{t+1}} - \frac{1}{\alpha_t}\right|\log\left(\frac{1}{\beta_{t+1}p_{1j}}\right)
\end{aligned}$$

*Proof.* By $\alpha_t$-Exp-Concavity of $\ell_t$, we have that $y \mapsto \exp\left(-\alpha_t \ell_t(y)\right)$ is concave. Hence, applying Jensen's inequality:

$$\exp\left(-\alpha_t \ell_t(\overline{y}_t)\right) \geq \sum_{i=1}^{N} p_{ti}\exp\left(-\alpha_t \ell_t\left(y_t^{(i)}\right)\right) = \sum_{i=1}^{N} p_{ti}\exp\left(-\alpha_t \ell_{ti}\right)$$

and taking the natural logarithm of both sides we have

$$\begin{aligned}
-\alpha_t \ell_t(\overline{y}_t) &\geq \log\left(\sum_{i=1}^{N} p_{ti}\exp\left(-\alpha_t \ell_{ti}\right)\right) \\
\ell_t(\overline{y}_t) &\leq -\frac{1}{\alpha_t}\log\left(\sum_{i=1}^{N} p_{ti}\exp\left(-\alpha_t \ell_{ti}\right)\right).
\end{aligned}$$

Hence, for any $j \in [N]$ we have

$$\begin{aligned}
\ell_t(\overline{y}_t) - \ell_t\left(y_t^{(j)}\right) &\leq -\frac{1}{\alpha_t}\log\left(\sum_{i=1}^{N} p_{ti}\exp\left(-\alpha_t \ell_{ti}\right)\right) - \ell_{tj} \\
&= -\frac{1}{\alpha_t}\log\left(\sum_{i=1}^{N} p_{ti}\exp\left(-\alpha_t \ell_{ti}\right)\right) + \frac{1}{\alpha_t}\log\left(\exp\left(-\alpha_t \ell_{tj}\right)\right) \\
&= \frac{1}{\alpha_t}\log\left(\frac{\exp\left(-\alpha_t \ell_{tj}\right)}{\sum_{i=1}^{N} p_{ti}\exp\left(-\alpha_t \ell_{ti}\right)}\right) \\
&= \frac{1}{\alpha_t}\log\left(\frac{p_{tj}\exp\left(-\alpha_t \ell_{tj}\right)}{p_{tj}\sum_{i=1}^{N} p_{ti}\exp\left(-\alpha_t \ell_{ti}\right)}\right) \\
&= \frac{1}{\alpha_t}\left[\log\left(\frac{q_{t+1,j}}{p_{tj}}\right)\right] \\
&= \frac{1}{\alpha_t}\left[\log\left(\frac{1}{p_{tj}}\right) - \log\left(\frac{1}{q_{t+1,j}}\right)\right].
\end{aligned}$$

Adding and subtracting $\frac{1}{\alpha_{t+1}}\log\left(\frac{1}{p_{t+1,j}}\right)$,

$$\begin{aligned}
\ell_t(\overline{y}_t) - \ell_t\left(y_t^{(j)}\right) &\leq \frac{1}{\alpha_t}\log\left(\frac{1}{p_{tj}}\right) - \frac{1}{\alpha_{t+1}}\log\left(\frac{1}{p_{t+1,j}}\right) \\
&\qquad + \frac{1}{\alpha_{t+1}}\log\left(\frac{1}{p_{t+1,j}}\right) - \frac{1}{\alpha_t}\log\left(\frac{1}{q_{t+1,j}}\right) \\
&= \frac{1}{\alpha_t}\log\left(\frac{1}{p_{tj}}\right) - \frac{1}{\alpha_{t+1}}\log\left(\frac{1}{p_{t+1,j}}\right) \\
&\qquad + \underbrace{\frac{1}{\alpha_t}\log\left(\frac{1}{p_{t+1,j}}\right) - \frac{1}{\alpha_t}\log\left(\frac{1}{q_{t+1,j}}\right)}_{\log(q_{t+1,j}/p_{t+1,j})/\alpha_t} \\
&\qquad + \left[\frac{1}{\alpha_{t+1}} - \frac{1}{\alpha_t}\right]\log\left(\frac{1}{p_{t+1,j}}\right)
\end{aligned}$$

recalling $p_{t+1,j} = (1 - \beta_{t+1})q_{t+1,j} + \beta_{t+1}p_{1j}$,

$$\begin{aligned}
&= \frac{1}{\alpha_t}\log\left(\frac{1}{p_{tj}}\right) - \frac{1}{\alpha_{t+1}}\log\left(\frac{1}{p_{t+1,j}}\right) \\
&\quad + \frac{1}{\alpha_t}\log\left(\frac{q_{t+1,j}}{(1-\beta_{t+1})q_{t+1,j} + \beta_{t+1}p_{1j}}\right) \\
&\quad + \left[\frac{1}{\alpha_{t+1}} - \frac{1}{\alpha_t}\right]\log\left(\frac{1}{(1-\beta_{t+1})q_{t+1,j} + \beta_{t+1}p_{1j}}\right) \\
&\le \frac{1}{\alpha_t}\log\left(\frac{1}{p_{tj}}\right) - \frac{1}{\alpha_{t+1}}\log\left(\frac{1}{p_{t+1,j}}\right) \\
&\quad + \frac{1}{\alpha_t}\log\left(\frac{1}{1-\beta_{t+1}}\right) \\
&\quad + \left|\frac{1}{\alpha_{t+1}} - \frac{1}{\alpha_t}\right|\log\left(\frac{1}{\beta_{t+1}p_{1j}}\right)
\end{aligned}$$

□

### C.2.7 Supporting Lemmas

The following provides a useful relation between the squared loss and its Bregman divergence.

**Lemma C.2.14.** *Let $\ell_t(w) = \frac{1}{2}(y_t - \langle x_t, w_t\rangle)^2$. Then for any $u, w \in W$,*

$$D_{\ell_t}(u|w) = \frac{1}{2}\langle x_t, u - w\rangle^2$$

*Proof.* By definition of Bregman divergence, we have:

$$D_{\ell_t}(u|w) = \ell_t(u) - \ell_t(w) - \langle \nabla \ell_t(w), u - w\rangle .$$

Expanding the definition of $\ell_t$, we have

$$\begin{aligned}
\ell_t(u) - \ell_t(w) &= \frac{1}{2}(y_t - \langle x_t, u\rangle)^2 - \frac{1}{2}(y_t - \langle x_t, w\rangle)^2 \\
&= \frac{1}{2}y_t^2 + \frac{1}{2}\langle x_t, u\rangle^2 - y_t\langle x_t, u\rangle - \frac{1}{2}y_t^2 - \frac{1}{2}\langle x_t, w\rangle^2 + y_t\langle x_t, w\rangle \\
&= \frac{1}{2}\langle x_t, u\rangle^2 - \frac{1}{2}\langle x_t, w\rangle^2 + y_t\langle x_t, w - u\rangle .
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
-\langle \nabla \ell_t(w), u - w\rangle &= \langle (y_t - \langle x_t, w\rangle)x_t, u - w\rangle \\
&= -y_t\langle x_t, w - u\rangle + \langle x_t, w\rangle^2 - \langle x_t, w\rangle\langle x_t, u\rangle ,
\end{aligned}$$

so combining with the previous display we have

$$\begin{aligned}
\ell_t(u) - \ell_t(w) - \langle \nabla \ell_t(w), u - w\rangle &= \frac{1}{2}\langle x_t, u\rangle^2 - \frac{1}{2}\langle x_t, w\rangle^2 + y_t\langle x_t, w - u\rangle \\
&\quad - y_t\langle x_t, w - u\rangle + \langle x_t, w\rangle^2 - \langle x_t, w\rangle\langle x_t, u\rangle \\
&= \frac{1}{2}\langle x_t, u\rangle^2 + \frac{1}{2}\langle x_t, w\rangle^2 - \langle x_t, w\rangle\langle x_t, u\rangle \\
&= \frac{1}{2}\left(\langle x_t, u\rangle - \langle x_t, w\rangle\right)^2 \\
&= \frac{1}{2}\langle x_t, u - w\rangle^2 .
\end{aligned}$$

□

The following provides a discounted version of the log-determinant lemma.

**Lemma C.2.15.** *Let $\gamma \in (0,1]$, $\lambda > 0$, $x_t \in \mathbb{R}^d$, and define $M_0 = \lambda I$ and $M_t = x_t x_t^\top + \gamma M_{t-1}$ for each $t > 0$. Then for any sequence $\Delta_1, \Delta_2, \ldots$ in $\mathbb{R}$,*

$$\sum_{t=1}^T \Delta_t^2 \|x_t\|_{M_t^{-1}}^2 \le d\log(1/\gamma)\,\Delta_{1:T}^2 + \max_t \Delta_t^2 d\log\left(1 + \frac{\sum_{t=1}^T \gamma^{T-t}\|x_t\|_2^2}{\lambda d}\right)$$

*Proof.* By definition we have $M_t = x_t x_t^\top + \gamma M_{t-1}$, so re-arranging and taking the determinant of both sides we have

$$\begin{aligned}
\mathrm{Det}(\gamma M_{t-1}) = \mathrm{Det}\left(M_t - x_t x_t^\top\right) &= \mathrm{Det}(M_t)\,\mathrm{Det}\left(I - M_t^{-\frac{1}{2}} x_t x_t^\top M_t^{-\frac{1}{2}}\right) \\
&= \mathrm{Det}(M_t)\,(1 - \|x_t\|_{M_t^{-1}}^2)
\end{aligned}$$

where the last line uses the fact that $\mathrm{Det}(I - yy^\top) = 1 - \|y\|_2^2$. Re-arranging, using $\mathrm{Det}(\gamma M_{t-1}) = \gamma^d \mathrm{Det}(M_{t-1})$, and using the fact that $1 - x \le -\log(x)$ we have

$$\begin{aligned}
\sum_{t=1}^T \Delta_t^2 \|x_t\|_{M_t^{-1}}^2 &= \sum_{t=1}^T \Delta_t^2 \left[1 - \frac{\gamma^d \mathrm{Det}(M_{t-1})}{\mathrm{Det}(M_t)}\right] \\
&\le \sum_{t=1}^T \Delta_t^2 \log\left(\frac{\mathrm{Det}(M_t)}{\gamma^d \mathrm{Det}(M_{t-1})}\right) \\
&= \sum_{t=1}^T \Delta_t^2 d\log(1/\gamma) + \sum_{t=1}^T \Delta_t^2 \log\left(\frac{\mathrm{Det}(M_t)}{\mathrm{Det}(M_{t-1})}\right) \\
&\le d\log(1/\gamma)\,\Delta_{1:T}^2 + \max_t \Delta_t^2 \log\left(\prod_{t=1}^T \frac{\mathrm{Det}(M_t)}{\mathrm{Det}(M_{t-1})}\right) \\
&= d\log(1/\gamma)\,\Delta_{1:T}^2 + \max_t \Delta_t^2 \log\left(\frac{\mathrm{Det}(M_T)}{\mathrm{Det}(M_0)}\right).
\end{aligned}$$

Observe that $\operatorname{Det}(M_0) = \operatorname{Det}(\lambda I) = \lambda^d$, and using AM-GM inequality we have

$$\begin{aligned}\operatorname{Det}(M_T) &\leq \left(\frac{\operatorname{Tr}(M_t)}{d}\right)^d = \left(\frac{\operatorname{Tr}\left(\lambda\gamma^T I + \sum_{t=1}^T \gamma^{T-t} x_t x_t^\top\right)}{d}\right)^d \\ &= \left(\frac{d\lambda\gamma^T + \sum_{t=1}^T \gamma^{T-t} \|x_t\|_2^2}{d}\right)^d,\end{aligned}$$

Hence $\frac{\operatorname{Det}(M_T)}{\operatorname{Det}(M_0)} \leq \left(\frac{d\lambda\gamma^T + \sum_{t=1}^T \gamma^{T-t}\|x_t\|_2^2}{d\lambda}\right)^d$, so overall we have

$$\begin{aligned}\sum_{t=1}^T \Delta_t^2 \|x_t\|_{M_t^{-1}}^2 &\leq d\log(1/\gamma)\,\Delta_{1:T}^2 + \max_t \Delta_t^2 \log\left(\left(\frac{d\lambda\gamma^T + \sum_{t=1}^T \gamma^{T-t}\|x_t\|_2^2}{\lambda d}\right)^d\right) \\ &= d\log(1/\gamma)\,\Delta_{1:T}^2 + \max_t \Delta_t^2 d\log\left(\frac{d\lambda\gamma^T + \sum_{t=1}^T \gamma^{T-t}\|x_t\|_2^2}{\lambda d}\right) \\ &\leq d\log(1/\gamma)\,\Delta_{1:T}^2 + \max_t \Delta_t^2 d\log\left(1 + \frac{\sum_{t=1}^T \gamma^{T-t}\|x_t\|_2^2}{\lambda d}\right)\end{aligned}$$

□

Note that the Lemma C.2.15 also immediately gives us the usual log determinant lemma as a special case where $\gamma = 1$:

**Lemma C.2.16.** *Let $\lambda > 0$, $x_t \in \mathbb{R}^d$, and define Let $M_0 = \lambda I$ and $M_t = x_t x_t^\top + M_{t-1}$ for each $t > 0$. Then for any sequence $\Delta_1, \Delta_2, \ldots$ in $\mathbb{R}$,*

$$\sum_{t=1}^T \Delta_t^2 \|x_t\|_{M_t^{-1}}^2 \leq d\max_t \Delta_t^2 \log\left(1 + \frac{\sum_{t=1}^T \|x_t\|_2^2}{\lambda d}\right)$$