Unconstrained Online Learning with Unbounded Losses

Andrew Jacobsen [1 2] **Ashok Cutkosky** [3]

Abstract

Algorithms for online learning typically require one or more boundedness assumptions: that the domain is bounded, that the losses are Lipschitz, or both. In this paper, we develop a new setting for online learning with unbounded domains and non-Lipschitz losses. For this setting we provide an algorithm which guarantees $R_T(u) \le \widetilde{O}(G\|u\|\sqrt{T}+L\|u\|^2\sqrt{T})$ regret on any problem where the subgradients satisfy $\|g_t\| \le G+L\|w_t\|$, and show that this bound is unimprovable without further assumptions. We leverage this algorithm to develop new saddle-point optimization algorithms that converge in duality gap in unbounded domains, even in the absence of meaningful curvature. Finally, we provide the first algorithm achieving non-trivial dynamic regret in an unbounded domain for non-Lipschitz losses, as well as a matching lower bound. The regret of our dynamic regret algorithm automatically improves to a novel L^* bound when the losses are smooth.

1. Online Learning

This paper introduces new techniques for online convex optimization (OCO), a standard framework used to model learning from a stream of data (Cesa-Bianchi & Lugosi, 2006; Shalev-Shwartz, 2011; Hazan, 2016; Orabona, 2019). Formally, consider T rounds of interaction between an algorithm and an environment. In each round, the algorithm chooses a w_t in some convex subset W of a Hilbert space, after which the environment chooses a convex loss function $\ell_t : W \to \mathbb{R}$. The standard performance metric in this setting is *regret* $R_T(u)$, the cumulative loss relative to an unknown benchmark point $u \in W$:

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

In many applications of interest the appropriate baseline is not any *fixed* comparator, but rather a *trajectory* of points. This is often the case in true streaming settings, wherein the losses are generated from a data distribution that may be slowly shifting over time. To better model settings such as these, *dynamic* regret measures the total loss relative to that of a benchmark *sequence* of points $\boldsymbol{u} = (u_1, \dots, u_T)$ in W:

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

Our goal in this work is to develop algorithms that achieve favorable regret and dynamic regret guarantees when *both* the domain W and range of ℓ_t may be unbounded.

To illustrate the difficulty of our goal, let us consider the special case where the loss functions are linear functions, $\ell_t(w) = \langle g_t, w\rangle$. Clearly, if both $\|g_t\|$ and $\|w\|$ are allowed to be arbitrarily large then the adversary can always ensure that the learner takes an arbitrarily large loss on each round. To alleviate this difficulty, prior works assume that one has access to a bound $D \ge \|u\|$ (usually by assuming that the domain is bounded with $D \ge \sup_{x,y\in W}\|x-y\|$), that the subgradients are bounded $\mathfrak{G}_T \ge \max_t \|g_t\|$, that the losses map to a bounded range $\ell_t : W \to [a, b]$, or some combination thereof.

In the simplest case, when one has access to both a bound $D \ge \|u\|$ and a bound on the subgradients $\mathfrak{G}_T \ge \|g_t\|$ for all t, classic methods based on Mirror Descent and Follow the Regularized Leader achieve minimax optimal regret of $R_T(u) \le O\big(D\mathfrak{G}_T\sqrt{T}\big)$ using a strongly convex regularizer (Hazan et al., 2007; Duchi et al., 2010; McMahan & Streeter, 2010). When D is available but not the Lipschitz bound $\mathfrak{G}_T$, it is still possible to match this guarantee up to constant factors, in which case the algorithm is said to be *Lipschitz adaptive* (Orabona & Pál, 2018; Mayo et al., 2022; Cutkosky, 2019). 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 — though prior works still

[1] Department of Computing Science, University of Alberta, Edmonton, Canada [2] Alberta Machine Intelligence Institute (Amii), Edmonton, Canada [3] Department of Electrical and Computer Engineering, Boston University, Boston, Massachussetts. Correspondence to: Andrew Jacobsen <ajjacobs@ualberta.ca>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

require one or more of the following assumptions: prior knowledge of $\mathfrak{G}_T$, that ℓ_t has bounded range (known in advance), prior knowledge of a lower bound $\ell_t^* \leq \ell_t(w)$ 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 (Cesa-Bianchi et al., 1996; Kivinen & Warmuth, 1997; Srebro et al., 2010; Orabona et al., 2012).

If a bound $\mathfrak{G}_T \geq \max_t \|g_t\|$ is known but not the bound $D \geq \|u\|$, the situation gets significantly trickier. The essential difficulty is that without prior knowledge of how large the comparator might be, the predictions w_t could at any point be arbitrarily far away from the benchmark, 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 the unknown comparator norm $\|u\|$. A standard result in this setting is

$$R_T(u) \leq O\left(\|u\|\mathfrak{G}_T\sqrt{T\log(T\|u\|+1)}\right), \qquad (1)$$

which holds for all $u \in W$ and is known to be optimal up to constants (Orabona, 2013). Bounds of this form are commonly referred to as "comparator adaptive" or "parameter-free" (Foster et al., 2015; Orabona & Pál, 2016; van der Hoeven, 2019; Cutkosky & Orabona, 2018; Mhammedi & Koolen, 2020; Jacobsen & Cutkosky, 2022).

The first results to avoid both the bounded domain and bounded gradient assumptions have only been achieved in recent years. Cutkosky (2019) develops an algorithm which achieves $R_T(u) \leq O\left(\|u\|\,\mathfrak{G}_T\sqrt{T\log\left(\|u\|\,T+1\right)} + \mathfrak{G}_T\,\|u\|^3\right)$, and Mhammedi & Koolen (2020) shows that the additional cubic penalty is unavoidable while maintaining the $\widetilde{O}\left(\|u\|\,\mathfrak{G}_T\sqrt{T}\right)$ dependence. Alternatively, Orabona & Pál (2018) show that $R_T(u) \leq O(\|u\|^2\,\mathfrak{G}_T\sqrt{T})$ 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 et al. (2022) and Kempka et al. (2019) show that Equation (1) can be achieved with essentially no extra penalty in certain special cases such as regression-type losses.

When it comes to dynamic regret, much less progress has been made in alleviating boundedness assumptions, with nearly all existing results assuming both a bounded domain and Lipschitz losses. Under both boundedness assumptions, prior works have achieved minimax optimal dynamic regret of $R_T(\boldsymbol{u}) \leq O\left(\mathfrak{G}_T\sqrt{(D^2+DP_T)T}\right)$, where $P_T = \sum_{t=2}^T \|u_t - u_{t-1}\|$ is the *path-length* of the comparator sequence (Zhang et al., 2018; Jadbabaie et al., 2015; Cutkosky, 2020). In an unbounded domain with Lipschitz losses, recent works have achieved an analogous guarantee of $R_T(\boldsymbol{u}) \leq \widetilde{O}(\mathfrak{G}_T\sqrt{(M^2+MP_T)T})$, where $M = \max_t \|u_t\|$ (Jacobsen & Cutkosky, 2022; Luo et al., 2022). We are unaware of any existing works that explicitly investigate Lipschitz-adaptive dynamic regret, though existing results can likely be made Lipschitz-adaptive in a black-box manner using the gradient clipping approach of Cutkosky (2019) in exchange for an appropriate $\mathfrak{G}_T \max_t \|u_t\|^3$ penalty.

Importantly, note that *in all of these prior works* there is an implicit assumption that *there exists* a uniform bound such that $G \geq \|\nabla\ell_t(w)\|$ for any $w \in W$ and $\nabla\ell_t(w) \in \partial\ell_t(w)$ — even if it is not known in advance. Otherwise, the terms $\mathfrak{G}_T = \max_t \|g_t\|$ can easily make any of the aforementioned regret guarantees vacuous.

In this work, we study unconstrained online convex optimization under a more general boundedness assumption on the gradients, allowing the gradient norms to grow arbitrarily large away from a given "reference point" $w_0 \in W$. In Section 2 we provide an algorithm for this more general problem setting which achieves a strict generalization of the usual comparator-adaptive bound in Equation (1), as well as a lower bound showing that our result is unimprovable in general. In Section 3 we leverage this algorithm to develop a new saddle-point optimization algorithm which converges in duality gap in an unbounded domain without requiring additional curvature assumptions such as strong-convexity/concavity. In Section 4, we turn to the problem of dynamic regret minimization and develop an algorithm which achieves dynamic regret $R_T(\boldsymbol{u}) \leq \widetilde{O}\left(M^{3/2}\sqrt{(M+P_T)T}\right)$ and provide a matching lower bound. This is the first algorithm to significantly alleviate both the bounded domain and bounded subgradient assumptions for dynamic regret. Moreover, when the losses are L_t-smooth, the same algorithm automatically improves to an L^* bound of the form $R_T(\boldsymbol{u}) \leq \widetilde{O}\left(\sqrt{(M^2+MP_T)\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t^*\right]}\right)$. To the best of our knowledge, this is in fact the first L^* bound to be achieved for general smooth losses without making either a uniformly-bounded subgradient or bounded range assumption in an unbounded domain.

Notations. For brevity, we occasionally abuse notation by letting $\nabla f(x)$ denote an element of $\partial f(x)$. The Bregman divergence *w.r.t.* a differentiable function ψ is $D_\psi(x|y) = \psi(x) - \psi(y) - \langle\nabla\psi(y), x-y\rangle$. We use the compressed sum notation $g_{i:j} = \sum_{t=i}^j g_t$ and $\|g\|_{a:b}^2 = \sum_{t=a}^b \|g_t\|^2$. We denote $a \vee b = \max\{a,b\}$ and $a \wedge b = \min\{a,b\}$. Δ_N denotes the N-dimensional simplex. The notation $O(\cdot)$ hides constants, $\widehat{O}(\cdot)$ hides constants and $\log(\log)$ terms, and $\widetilde{O}(\cdot)$ hides up to and including $\log$ factors.

2. 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 2.1. 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

$$\|\nabla\ell(w)\| \leq G + L\|w - w_0\|. \tag{2}$$

Note that Definition 2.1 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. [1] For the remainder of the paper we assume without loss of generality that $w_0 = \mathbf{0}$ and $\|\cdot\|$ is the Euclidean norm.

This 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 $\|g_t\| \leq G_t + L_t\|w_t\|$. Then using the standard reduction from OCO to OLO (see *e.g.* Zinkevich (2003)), for any sequence of (G_t, L_t)-quadratically bounded convex functions we have the following regret upper bound:

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

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 $\|w_t\| \propto \|u\|$. For brevity we refer to this harder problem setting as the QB-OLO setting, and QB-OCO for the setting where adversary must play ℓ_t satisfying Definition 2.1.

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.

Theorem 2.2. *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 $\Psi_t(w) = 3\int_0^{\|w\|} \min_{\eta \leq \frac{1}{G_{\max}}} \left[\frac{\log(x/\alpha_t + 1)}{\eta} + \eta V_t\right] dx$ 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 1 guarantees

$$R_T(u) \leq O\left(\epsilon + \|u\|\sqrt{G_{1:T}^2 F_T(\|u\|)} + \|u\|^2\sqrt{L_{1:T}^2}\right)$$

where $F_T(\|u\|) \leq \log\left(\frac{\|u\|\sqrt{T}\log^2(T)}{\epsilon} + 1\right)$.

Let us briefly develop some intuition for how the above result is constructed. Algorithm 1 can be interpreted as an instance of the Centered Mirror Descent algorithm recently developed by Jacobsen & Cutkosky (2022), which admits a generic regret guarantee of the form $R_T(u) \leq \psi_T(u) + \sum_{t=1}^{T}\varphi_t(u) + \sum_{t=1}^{T}\delta_t$, where the δ_t are similar to the "stability" terms encountered in vanilla Mirror Descent, but

[1] 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.

Algorithm 1 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 for**
```

with certain additional negative terms Δ_t and φ_t:

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

It's easily verified that that $\psi_{T+1}(u) + \sum_{t=1}^T \varphi_t(u)$ match the terms in the upper bound, so the main difficulty is making sure that the stability terms $\sum_{t=1}^T \delta_t$ 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 Jacobsen & Cutkosky (2022), by setting $\Psi_t(w) = O\Big(G_{\max}\|w\|\sqrt{T\log\left(\|w\|\sqrt{T}/\epsilon\right)}\Big)$ 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 Ψ_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

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

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 comparator-adaptive rate of $R_T(u) \leq \widehat{O}\left(\|u\|\sqrt{\sum_{t=1}^T\|g_t\|^2\log\left(\frac{\|u\|\sqrt{T}}{\epsilon}+1\right)}\right)$, which is known to be optimal up to constant and $\log(\log)$ terms (Mcmahan & 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, *i.e.*, 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 2.2 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

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

so $\{\nabla\ell_t(w_t)\}$ is a (G_t, L_t)-quadratically bounded sequence w.r.t $\{w_t\}$, and Theorem 2.2 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 $R_T(u) \leq \widetilde{O}\left(\|u\||y_{\max}|\|x_{\max}\|\sqrt{T} + \|u\|^2\|x_{\max}\|^2\sqrt{T}\right)$.

Finally, notice that for $L_{\max} > 0$ Algorithm 1 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 our problem setting. Proof can be found in Appendix B.2.

Theorem 2.3. *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 $\|g_t\| \leq G + L\|w_t\|$ and a comparator $u \in \mathbb{R}^2$ such that*

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

Remark 2.4. An alternative way to approach online learning in our problem setting would be to apply an algo-

Algorithm 2 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 $g^x_t \in \partial_x \mathcal{L}(x_t, y_t)$ and $g^y_t \in \partial_y[-\mathcal{L}(x_t, y_t)]$
      Send $g_t = (g^x_t, g^y_t)$ to $\mathcal{A}$ as the $t^{\text{th}}$ subgradient
  end for
  Return $\overline{w}_T = \left(\frac{\sum_{t=1}^T x_t}{T}, \frac{\sum_{t=1}^T y_t}{T}\right)$
```

rithm 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 2.3 demonstrates that this approach would be sub-optimal in our setting. Indeed, Mhammedi & 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.

3. 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:

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

where $\mathcal{X}$ and $\mathcal{Y}$ are unbounded compact convex sets, $x \mapsto \mathcal{L}(x, y)$ is convex for all $y \in \mathcal{Y}$, and $y \mapsto \mathcal{L}(x, y)$ is concave for all $x \in \mathcal{X}$. The saddle-point (x^*, y^*) of $\mathcal{L}$ is the point $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ satisfying $\mathcal{L}(x^*, y^*) = \min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \mathcal{L}(x, y) = \max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} \mathcal{L}(x, y)$, where the last equality follows from Sion's minimax theorem. Then for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we have

$$\mathcal{L}(x, y) - \mathcal{L}(x^*, y^*) \le \underbrace{\mathcal{L}(x, y^*) - \mathcal{L}(x^*, y)}_{=:G(x,y)}.$$

Hence, the sub-optimality of a point $(x, y) \in \mathcal{X} \times \mathcal{Y}$ can be controlled so long as we can control the quantity $G(x, y)$, which we refer to as the *duality gap*. Fortunately, the duality gap is easily controlled using an online learning algorithm via the well-known reduction to Online Linear Optimization (OLO) shown in Algorithm 2.

Lemma 3.1. *For any $\mathring{w} = (\mathring{x}, \mathring{y}) \in \mathcal{X} \times \mathcal{Y}$, Algorithm 2 guarantees*

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

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

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

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

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

□

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.

To the best of our knowledge, the only existing work to achieve a comparator-adaptive convergence guarantee for the duality gap in general saddle-point problems is Liu & Orabona (2022). Their approach does indeed guarantee a rate of the form $G(\overline{x}_T, \overline{y}_T) \le \frac{R^{\mathcal{A}}_T(w^*)}{T} \le \widetilde{O}\left(\frac{G\|w^*\|}{\sqrt{T}}\right)$ 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 $R^{\mathcal{A}}_T(w^*) \le \widetilde{O}\left(\|w^*\| G\sqrt{T}\right)$ 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 3.1 for a more explicit example of this issue.

On the other hand, for many interesting problems $\mathcal{L}(\cdot, \cdot)$ is quadratically bounded in both arguments, which will enable us to immediately apply Algorithm 1 to the linear losses $g_t = (g^x_t, g^y_t)$ as described above. In particular, we have the following:

Proposition 3.2. *Suppose that for all $\widetilde{y} \in \mathcal{Y}$, the function $x \mapsto \mathcal{L}(x, \widetilde{y})$ is $(G_x + L_{xy}\left\|\widetilde{y}\right\|, L_{xx})$-quadratically bounded, and for all $\widetilde{x} \in \mathcal{X}$ the function $y \mapsto -\mathcal{L}(\widetilde{x}, y)$ is $(G_y + L_{yx}\left\|\widetilde{x}\right\|, L_{yy})$-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

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

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 1 to guarantee that for any $\mathring{w} = (\mathring{x}, \mathring{y}) \in W$,

$$
\begin{aligned}
&\mathcal{L}(\overline{x}_T, \mathring{y}) - \mathcal{L}(\mathring{x}, \overline{y}_T) \\
&\quad \overset{Lemma\ 3.1}{\le} \frac{R_T^{\mathcal{A}}(\mathring{w})}{T} \\
&\quad \overset{Theorem\ 2.2}{\le} \widetilde{O}\left(\frac{G_w\|\mathring{w}\| + L_w\|\mathring{w}\|^2}{\sqrt{T}}\right),
\end{aligned}
$$

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 ℓ_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 ℓ_t^x and ℓ_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 ℓ_t^x and ℓ_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.

3.1. 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

$$
\mathcal{L}(x, y) = F_x(x) + H(x, y) - F_y(y) \tag{3}
$$

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 3.2.

Proposition 3.3. *Equation* (3) *satisfies the assumptions of Proposition 3.2 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 $L_{yx} = \left\|B^\top\right\|_{op}$.*

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 3.2 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 $L_{yx} = \left\|B^\top\right\|_{\text{op}}$. □

We note that this specific example is mainly for illustrative purposes — in many instances of Equation (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 & 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.

4. Dynamic Regret

Next, we turn our attention to *dynamic* regret. In the static regret setting, we saw in Section 2 that to control the stability of the algorithm it was necessary to add an additional term $\Phi_t(w) = O\left(L_{\max}\sqrt{T}\left\|w\right\|^2\right)$ 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 difficulties. To see why, consider the dynamic regret of gradient descent with a fixed step-size η. Using existing analyses one can arrive at

$$R_T(\boldsymbol{u}) \le O\left(\frac{\left\|u_T\right\|^2 + \max_t \left\|w_t\right\| P_T}{2\eta} + \frac{\eta}{2}\sum_{t=1}^T \left\|g_t\right\|^2\right), \tag{4}$$

where $g_t \in \partial\ell_t(w_t)$ and $P_T = \sum_{t=2}^T \left\|u_t - u_{t-1}\right\|$. In a bounded domain of diameter D, we can bound $\left\|u_T\right\|^2 \le D^2$ and $\max_t \left\|w_t\right\| \le D$, and then by optimally tuning η we get $R_T(\boldsymbol{u}) \le O\left(\sqrt{(D^2 + DP_T)\sum_{t=1}^T \left\|g_t\right\|^2}\right)$ which is optimal in the Lipschitz loss setting (Zhang et al., 2018). More generally, using Mirror Descent with regularizer $\psi(w)$, one can derive an analogous bound:

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

where δ_t are the stability terms discussed in Section 2. In an unbounded domain, Jacobsen & Cutkosky (2022) use a regularizer of the form $\psi(w) = O\left(\left\|w\right\| \log\left(\left\|w\right\| T\right)/\eta\right)$, which enables them to bound $\sum_{t=1}^T \delta_t \le O(\eta)$, and moreover, they show that $\max_t \left\|\nabla\psi_t(w_t)\right\|$ can be bound from above by $O\left(\log\left(MT/\epsilon\right)/\eta\right)$ after adding a composite penalty $\varphi_t(w) = \eta\left\|g_t\right\|^2 \left\|w\right\|$ to the update. Then optimally tuning η leads to regret scaling as

$$R_T(\boldsymbol{u}) \le \widetilde{O}\left(\sqrt{(M^2 + MP_T)\sum_{t=1}^T \left\|g_t\right\|^2}\right), \tag{5}$$

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

Algorithm 3 Dynamic Regret Algorithm

```
Input: $G_{\max}$, $L_{\max}$, weights $\beta_1, \ldots, \beta_T$ 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
for $t = 1 : T$ do
    Play $w_t = \sum_{\tau\in\mathcal{S}} p_t(\tau) w_t^{(\tau)}$
    Query $\ell_t(\widetilde{w}_t)$ for any $\widetilde{w}_t \in W$ 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 for
    Set $q_{t+1} = \operatorname{argmin}_{q\in\Delta_{|\mathcal{S}|}} \sum_{\tau\in\mathcal{S}} (\widetilde{\ell}_{t\tau} + \mu_\tau \widetilde{\ell}_{t\tau}^2) q_\tau + D_{\psi_\tau}(q_\tau | p_{t\tau})$
    Set $p_{t+1} = (1 - \beta_t) q_{t+1} + \beta_t p_1$.
end for
```

In our setting, the situation gets significantly more challenging. As in Section 2, we will need to include an $O(\left\|w\right\|^2/\eta)$ term in the regularizer ψ_t in order to control the "non-Lipschitz" part of the loss function. However, as above this leads to coupling $\max_t \left\|\nabla\psi_t(w_t)\right\| P_t = \max_t \left\|w_t\right\| P_T/\eta$ in the dynamic regret, and the term $\max_t \left\|w_t\right\|$ is generally too large to cancel out with additional regularization as done by Jacobsen & Cutkosky (2022). Even more troubling is that our lower bound in Theorem 4.2 suggests that the ideal dependence would be $O(MP_T/\eta)$, which we can only hope to achieve by constraining $\left\|w_t\right\|$ to a ball of diameter proportional to $M = \max_t \left\|u_t\right\|$. Yet M is unknown to the learner!

Luckily, hope is not all lost. Taking inspiration from Luo et al. (2022), we can still attain a bound similar to Equation (5) by tuning the diameter of an artificial domain constraint. The approach is as follows: for each (η, D) in some set $\mathcal{S}$, we run an instance of gradient descent $\mathcal{A}(\eta, D)$ which uses step-size η and projects to the set $W_D = \{w \in W : \left\|w\right\| \le D\}$. Then, using a carefully designed experts algorithm, it is possible to ensure that the overall regret of the algorithm scales roughly as $R_T(\boldsymbol{u}) \leq \widetilde{O}\left(R_T^{\mathcal{A}(\eta,D)}(\boldsymbol{u})\right)$ 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 η 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, characterizes an algorithm which achieves dynamic regret analogous to the above bounds, and in Theorem 4.2 we show that this is indeed unimprovable. Notably, our result also *automatically* improves to a novel L^* bound when the losses are smooth.

Theorem 4.1. *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 $\epsilon > 0$, $\beta_t \leq 1 - \exp(-1/T)$ for all t, and for any $i, j \geq 0$ let $D_j = \frac{\epsilon}{T}\left[2^j \wedge 2^T\right]$ and $\eta_i = \left[\frac{\epsilon 2^i}{8(G_{\max}+\epsilon L_{\max})T} \wedge \frac{1}{8L_{\max}}\right]$, 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 3 guarantees*

$$
\begin{aligned}
R_T(\boldsymbol{u}) \leq O\Bigg(& G_{\max}(M+\epsilon)\Lambda_T^* + L_{\max}(M+\epsilon)^2\Lambda_T^* \\
& + G_{\max}P_T + L_{\max}(M+\epsilon)P_T \\
& + \sqrt{(M^2\Lambda_T^* + MP_T)\Omega_T}, \Bigg).
\end{aligned}
$$

where $\Lambda_T^* \leq O\left(\log\left(\frac{MT\log(T)}{\epsilon}\right) + \log\left(\log\left(\frac{G_{\max}}{\epsilon L_{\max}}\right)\right)\right)$, $\Omega_T \leq \sum_{t=1}^T \left[G_t^2 + L_t^2 M^2\right]$, $P_T = \sum_{t=2}^T \|u_t - u_{t-1}\|$, and $M = \max_t \|u_t\|$. Moreover, when the losses are L_t-smooth, Ω_T automatically improves to $\Omega_T \leq \min\left\{\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t^*\right], \sum_{t=1}^T G_t^2 + L_t^2 M^2\right\}$.

Hiding constants and logarithmic terms, the bound is effectively

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

Notice that our result again generalizes the bounds established in prior works. Unfortunately, the result is not a strict generalization as Theorem 4.1 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 of Jacobsen & Cutkosky (2022) when $L_{\max} = 0$ and Algorithm 3 otherwise; this is possible because $L_{\max}$ must be provided as input to the algorithm. Notably, the algorithm of Jacobsen & Cutkosky (2022) 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 2, 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.3).

Theorem 4.2. *For any $M > 0$ there is a sequence of (G, L)-quadratically bounded functions with $\frac{G}{L} \leq M$ such that for any $\gamma \in [0, \frac{1}{2}]$,*

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

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

Notice that with $\gamma = \frac{1}{2}$, we have $R_T(\boldsymbol{u}) \geq \Omega\left(G\sqrt{MP_T T} + LM^{3/2}\sqrt{P_T T}\right)$, matching our upper bound in Theorem 4.1 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

$$
\begin{aligned}
R_T(\boldsymbol{u}) &\geq \Omega\left((G + LM)\sqrt{MP_T T}\right) \\
&= \Omega\left(G\sqrt{MP_T T} + LM^{3/2}\sqrt{P_T T}\right).
\end{aligned}
$$

Interestingly, when the losses are smooth, the bound in Theorem 4.1 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 et al., 2010; Orabona et al., 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 4.1 requires a collection of $O(T\log_2(\sqrt{T}) + T\log_2(G_{\max}/L_{\max}\epsilon))$ experts, so in practice we can only tolerate $G_{\max}/L_{\max}\epsilon \leq \mathrm{poly}(T)$ without increasing the (already quite high!) order of computation. We note that any algorithm that guarantees $R_T(\mathbf{0}) \leq G_{\max}\epsilon$ can't 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.2.

Theorem 4.3. *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 3 guarantees*

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

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

5. Conclusion

In this paper, we developed new algorithms for online learning in unbounded domains with potentially unbounded losses. We achieve several regret guarantees that have previously only been attained by assuming Lipschitz losses, losses with bounded range, a bounded domain, or a combination thereof. We provide algorithms for both static and dynamic regret, as well as an application in saddle-point optimization leading to new results for unbounded decision sets. Our lower bounds show that our results are optimal, and moreover, our algorithms achieve these results without appealing to any instance-specific hyperparameter tuning.

There are a few natural directions for future work. It is still unclear whether the dynamic regret achieved in Theorem 4.1 can be achieved in the more general QB-OLO setting. Moreover, while our dynamic regret algorithm attains the optimal bound, it requires $O(dT\log(\sqrt{T}))$ computation per round, whereas the optimal bounds in the Lipschitz loss setting are attained using $O(d\log(T))$ per-round computation. Yet it is unclear how to achieve the lower bound in Theorem 4.2 without the artificial domain trick discussed in Section 4. We leave these questions as exciting directions for future work.

Acknowledgements

AJ is supported by an NSERC CGS-D Scholarship. AC is supported by NSF grant CCF-2211718, a Google gift and an Amazon Research Award.

References

Azizian, W., Scieur, D., Mitliagkas, I., Lacoste-Julien, S., and Gidel, G. Accelerating smooth games by manipulating spectral shapes. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1705–1715. PMLR, 26–28 Aug 2020.

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

Cesa-Bianchi, N., Long, P. M., and Warmuth, M. K. Worst-case quadratic loss bounds for prediction using linear functions and gradient descent. *IEEE Transactions on Neural Networks*, 7(3):604–619, 1996.

Cesa-Bianchi, N., Gaillard, P., Lugosi, G., and Stoltz, G. Mirror descent meets fixed share (and feels no regret). *Advances in Neural Information Processing Systems*, 25, 2012.

Chen, L., Luo, H., and Wei, C.-Y. Impossible tuning made possible: A new expert algorithm and its applications. In Belkin, M. and Kpotufe, S. (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 1216–1259. PMLR, 15–19 Aug 2021.

Cutkosky, A. Artificial constraints and hints for unbounded online learning. In Beygelzimer, A. and Hsu, D. (eds.), *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 874–894, Phoenix, USA, 25–28 Jun 2019. PMLR.

Cutkosky, A. Parameter-free, dynamic, and strongly-adaptive online learning. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2250–2259, Virtual, 13–18 Jul 2020. PMLR.

Cutkosky, A. and Orabona, F. Black-box reductions for parameter-free online learning in banach spaces. In Bubeck, S., Perchet, V., and Rigollet, P. (eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1493–1529. PMLR, 06–09 Jul 2018.

Du, S. S., Gidel, G., Jordan, M. I., and Li, C. J. Optimal extragradient-based bilinearly-coupled saddle-point optimization, 2022.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. In *Conference on Learning Theory*, 2010.

Foster, D. J., Rakhlin, A., and Sridharan, K. Adaptive online learning. In *Advances in Neural Information Processing Systems 28*. 2015.

Gyorgy, A. and Szepesvari, C. Shifting regret, mirror descent, and matrices. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2943–2951, New York, New York, USA, 20–22 Jun 2016. PMLR.

Hall, E. C. and Willett, R. M. Online optimization in dynamic environments, 2016.

Hazan, E. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016. ISSN 2167-3888. doi: 10.1561/2400000013.

Hazan, E., Rakhlin, A., and Bartlett, P. Adaptive online gradient descent. In Platt, J., Koller, D., Singer, Y., and Roweis, S. (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

Ibrahim, A., Azizian, W., Gidel, G., and Mitliagkas, I. Linear lower bounds and conditioning of differentiable games. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4583–4593. PMLR, 13–18 Jul 2020.

Jacobsen, A. and Cutkosky, A. Parameter-free mirror descent, 2022.

Jadbabaie, A., Rakhlin, A., Shahrampour, S., and Sridharan, K. Online Optimization : Competing with Dynamic Comparators. In Lebanon, G. and Vishwanathan, S. V. N. (eds.), *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pp. 398–406, San Diego, California, USA, 09–12 May 2015. PMLR.

Kempka, M., Kotlowski, W., and Warmuth, M. K. Adaptive scale-invariant online algorithms for learning linear models. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3321–3330. PMLR, 09–15 Jun 2019.

Kivinen, J. and Warmuth, M. K. Exponentiated gradient versus gradient descent for linear predictors. *information and computation*, 132(1):1–63, 1997.

Liu, M. and Orabona, F. On the initialization for convex-concave min-max problems. In Dasgupta, S. and Haghtalab, N. (eds.), *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, pp. 743–767. PMLR, 29 Mar–01 Apr 2022.

Luo, H., Zhang, M., Zhao, P., and Zhou, Z.-H. Corralling a larger band of bandits: A case study on switching regret for linear bandits, 2022.

Mayo, J. J., Hadiji, H., and van Erven, T. Scale-free unconstrained online learning for curved losses, 2022.

Mcmahan, B. and Streeter, M. No-regret algorithms for unconstrained online convex optimization. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

McMahan, H. B. and Streeter, M. Adaptive bound optimization for online convex optimization. In *Conference on Learning Theory*, 2010.

Mhammedi, Z. and Koolen, W. M. Lipschitz and comparator-norm adaptivity in online learning. In Abernethy, J. and Agarwal, S. (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2858–2887. PMLR, 09–12 Jul 2020.

Orabona, F. Dimension-free exponentiated gradient. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Orabona, F. A modern introduction to online learning. *CoRR*, abs/1912.13213, 2019.

Orabona, F. and Pál, D. Coin betting and parameter-free online learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 577–585, Red Hook, NY, USA, 2016. Curran Associates Inc.

Orabona, F. and Pál, D. Scale-free online learning. *Theoretical Computer Science*, 716:50 – 69, 2018. ISSN 0304-3975. doi: https://doi.org/10.1016/j.tcs.2017.11.021. Special Issue on ALT 2015.

Orabona, F., Cesa-Bianchi, N., and Gentile, C. Beyond logarithmic bounds in online learning. In Lawrence, N. D. and Girolami, M. (eds.), *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pp. 823–831, La Palma, Canary Islands, 21–23 Apr 2012. PMLR.

Shalev-Shwartz, S. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2), 2011.

Srebro, N., Sridharan, K., and Tewari, A. Smoothness, low noise and fast rates. *Advances in neural information processing systems*, 23, 2010.

Telgarsky, M. Stochastic linear optimization never overfits with quadratically-bounded losses on general data, 2022.

van der Hoeven, D. User-specified local differential privacy in unconstrained adaptive online learning. In *NeurIPS*, pp. 14080–14089, 2019.

Zhang, L., Lu, S., and Zhou, Z.-H. Adaptive online learning in dynamic environments. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 1330–1340, 2018.

Zhao, P., Zhang, Y.-J., Zhang, L., and Zhou, Z.-H. Dynamic regret of convex and smooth functions. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12510–12520. Curran Associates, Inc., 2020.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *International Conference on Machine Learning*, 2003.

A. Centered Mirror Descent with Adjustment

Algorithm 4 Centered Mirror Descent with Adjustment

```
Input: $w_1 \in W$, $\psi_1 : W \to \mathbb{R}_{\geq 0}$
for $t = 1 : T$ do
    Play $w_t \in W$, observe $g_t$
    Choose functions $\psi_{t+1}$, $\phi_t$
    Define $\Delta_t(w) = D_{\psi_{t+1}}(w|w_1) - D_{\psi_t}(w|w_1)$
    Update $\widetilde{w}_{t+1} = \arg\min_{\widetilde{w} \in W} \langle g_t, \widetilde{w} \rangle + D_{\psi_t}(\widetilde{w}|w_t) + \Delta_t(\widetilde{w}) + \phi_t(\widetilde{w})$
    Choose mapping $\mathcal{M}_{t+1} : W \to W$
    Update $w_{t+1} = \mathcal{M}_{t+1}(\widetilde{w}_{t+1})$
end for
```

The key tool we'll use to build our algorithms is a slight generalization of the Centered Mirror Descent algorithm of Jacobsen & Cutkosky (2022), which accounts for an additional post-hoc "adjustment" of w_t 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 & Szepesvari (2016); Hall & Willett (2016), wherein $\mathcal{M}_t$ is interpreted as a dynamical model. In contrast, we will use $\mathcal{M}_t$ as a convenient way to formulate a multi-scale version of the fixed-share update, similar to the generalized share algorithm of Cesa-Bianchi et al. (2012). Note that when $\mathcal{M}_t$ is the identity mapping, Algorithm 4 is identical to the Centered Mirror Descent algorithm of Jacobsen & Cutkosky (2022).

The following lemma provides a regret template for Algorithm 4. Observe that several of the key terms related to the algorithm's stability replace the mirror descent iterates $\widetilde{w}_t$ with the adjusted iterates $w_t = \mathcal{M}_t(\widetilde{w}_t)$. The trade-off is that we must be careful to ensure that the new penalty terms $\xi_t = D_{\psi_{t+1}}(u_t|w_{t+1}) - D_{\psi_{t+1}}(u_t|\widetilde{w}_{t+1})$ are not too large, which places an implicit restriction on how much we can adjust the iterates via $\mathcal{M}_t$.

Lemma A.1. *For all t let $\psi_t : W \to \mathbb{R}_{\geq 0}$ be differentiable convex functions, $\phi_t : W \to \mathbb{R}_{\geq 0}$ 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 4 guarantees*

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

Proof. Letting $g_t \in \partial \ell_t(w_t)$, we have

$$R_T(\boldsymbol{u}) \leq \sum_{t=1}^{T} \langle g_t, w_t - u_t \rangle = \sum_{t=1}^{T} \langle g_t, \widetilde{w}_{t+1} - u_t \rangle + \sum_{t=1}^{T} \langle g_t, w_t - \widetilde{w}_{t+1} \rangle .$$

From the first-order optimality condition $\widetilde{w}_{t+1} = \arg\min_{\widetilde{w} \in W} \langle g_t, \widetilde{w} \rangle + D_{\psi_t}(\widetilde{w}|w_t) + \Delta_t(\widetilde{w}) + \phi_t(\widetilde{w})$, for any $u_t \in W$ we have

$$0 \geq \langle g_t + \nabla\psi_t(\widetilde{w}_{t+1}) - \nabla\psi_t(w_t), \widetilde{w}_{t+1} - u_t \rangle + \langle \nabla\Delta_t(\widetilde{w}_{t+1}) + \nabla\phi_t(\widetilde{w}_{t+1}), \widetilde{w}_{t+1} - u_t \rangle$$

so re-arranging,

$$\begin{aligned}
\langle g_t, \widetilde{w}_{t+1} - u_t \rangle &\leq \langle \nabla\psi_t(w_t) - \nabla\psi_t(\widetilde{w}_{t+1}), \widetilde{w}_{t+1} - u_t \rangle \\
&\quad - \langle \nabla\Delta_t(\widetilde{w}_{t+1}), \widetilde{w}_{t+1} - u_t \rangle - \langle \nabla\phi_t(\widetilde{w}_{t+1}), \widetilde{w}_{t+1} - u_t \rangle \\
&\overset{(*)}{\leq} \langle \nabla\psi_t(w_t) - \nabla\psi_t(\widetilde{w}_{t+1}), \widetilde{w}_{t+1} - u_t \rangle \\
&\quad + \phi_t(u_t) - \phi_t(\widetilde{w}_{t+1}) \\
&\quad + \Delta_t(u_t) - \Delta_t(\widetilde{w}_{t+1}) - D_{\Delta_t}(u_t|\widetilde{w}_{t+1})
\end{aligned}$$

where $(*)$ uses the definition of Bregman divergence to write $\langle -\nabla f(x), x - y \rangle = f(y) - f(x) - D_f(y|x)$ and bounds $\langle \nabla\phi_t(\widetilde{w}_{t+1}), u_t - \widetilde{w}_{t+1} \rangle \leq \phi_t(u_t) - \phi_t(\widetilde{w}_{t+1})$ by convexity of ϕ_t. Plugging this back into the previous display, we have

$$\begin{aligned}
R_T(\boldsymbol{u}) \leq \sum_{t=1}^T &\langle \nabla\psi_t(w_t) - \nabla\psi_t(\widetilde{w}_{t+1}), \widetilde{w}_{t+1} - u_t \rangle \\
&+ \sum_{t=1}^T [\Delta_t(u_t) + \phi_t(u_t)] + \sum_{t=1}^T -D_{\Delta_t}(u_t|\widetilde{w}_{t+1}) \\
&+ \sum_{t=1}^T \langle g_t, w_t - \widetilde{w}_{t+1} \rangle - \Delta_t(\widetilde{w}_{t+1}) - \phi_t(\widetilde{w}_{t+1}),
\end{aligned}$$

and using the three-point relation for Bregman divergences $\langle \nabla f(y) - \nabla f(x), x - z \rangle = D_f(z|y) - D_f(z|x) - D_f(x|y)$:

$$\begin{aligned}
&= \sum_{t=1}^T D_{\psi_t}(u_t|w_t) - D_{\psi_t}(u_t|\widetilde{w}_{t+1}) + \sum_{t=1}^T [\Delta_t(u_t) + \phi_t(u_t)] + \sum_{t=1}^T -D_{\Delta_t}(u_t|\widetilde{w}_{t+1}) \\
&\quad + \sum_{t=1}^T \underbrace{\langle g_t, w_t - \widetilde{w}_{t+1} \rangle - D_{\psi_t}(\widetilde{w}_{t+1}|w_t) - \Delta_t(\widetilde{w}_{t+1}) - \phi_t(\widetilde{w}_{t+1})}_{=:\delta_t} \\
&\overset{(a)}{=} \sum_{t=1}^T D_{\psi_t}(u_t|w_t) - D_{\psi_{t+1}}(u_t|\widetilde{w}_{t+1}) + \sum_{t=1}^T [\Delta_t(u_t) + \phi_t(u_t)] + \delta_{1:T} \\
&= \sum_{t=1}^T D_{\psi_t}(u_t|w_t) - D_{\psi_{t+1}}(u_t|w_{t+1}) + \sum_{t=1}^T \underbrace{D_{\psi_{t+1}}(u_t|w_{t+1}) - D_{\psi_{t+1}}(u_t|\widetilde{w}_{t+1})}_{=:\xi_t} + \sum_{t=1}^T [\Delta_t(u_t) + \phi_t(u_t)] + \delta_{1:T} \\
&= D_{\psi_1}(u_1|w_1) - D_{\psi_{T+1}}(u_T|w_{T+1}) + \underbrace{\sum_{t=2}^T [D_{\psi_t}(u_t|w_t) - D_{\psi_t}(u_{t-1}|w_t)] + \sum_{t=1}^T \Delta_t(u_t)}_{=:(\star)} \\
&\quad + \sum_{t=1}^T \phi_t(u_t) + \xi_{1:T} + \delta_{1:T},
\end{aligned}$$

where (a) observes $D_{\Delta_t}(u_t|\widetilde{w}_{t+1}) = D_{\psi_{t+1}}(u_t|\widetilde{w}_{t+1}) - D_{\psi_t}(u_t|\widetilde{w}_{t+1})$. Observe that the term $(\star)$ simplifies as

$$\begin{aligned}
(\star) &= \sum_{t=1}^{T}\Delta_t(u_t) + \sum_{t=2}^{T} D_{\psi_t}(u_t|w_t) - D_{\psi_t}(u_{t-1}|w_t)\\
&= \sum_{t=1}^{T} D_{\psi_{t+1}}(u_t|w_1) - D_{\psi_t}(u_t|w_1) + \sum_{t=2}^{T} D_{\psi_t}(u_t|w_t) - D_{\psi_t}(u_{t-1}|w_t)\\
&= D_{\psi_{T+1}}(u_T|w_1) - D_{\psi_1}(u_1|w_1)\\
&\quad + \sum_{t=1}^{T-1} D_{\psi_{t+1}}(u_t|w_1) - D_{\psi_{t+1}}(u_{t+1}|w_1)\\
&\quad + \sum_{t=2}^{T} D_{\psi_t}(u_t|w_t) - D_{\psi_t}(u_{t-1}|w_t)\\
&= D_{\psi_{T+1}}(u_T|w_1) - D_{\psi_1}(u_1|w_1)\\
&\quad + \sum_{t=2}^{T} D_{\psi_t}(u_{t-1}|w_1) - D_{\psi_t}(u_t|w_1)\\
&\quad + \sum_{t=2}^{T} D_{\psi_t}(u_t|w_t) - D_{\psi_t}(u_{t-1}|w_t)\\
&= D_{\psi_{T+1}}(u_T|w_1) - D_{\psi_1}(u_1|w_1)\\
&\quad + \sum_{t=2}^{T} \psi_t(u_{t-1}) - \psi_t(u_t) - \langle\nabla\psi_t(w_1), u_{t-1} - u_t\rangle\\
&\quad + \sum_{t=2}^{T} \psi_t(u_t) - \psi_t(u_{t-1}) - \langle\nabla\psi_t(w_t), u_t - u_{t-1}\rangle\\
&= D_{\psi_{T+1}}(u_T|w_1) - D_{\psi_1}(u_1|w_1) + \sum_{t=2}^{T}\underbrace{\langle\nabla\psi_t(w_t) - \nabla\psi_t(w_1), u_{t-1} - u_t\rangle}_{=:\mathcal{P}_t}
\end{aligned}$$

so plugging this back in above yields

$$\begin{aligned}
R_T(\boldsymbol{u}) &\le D_{\psi_1}(u_1|w_1) - D_{\psi_{T+1}}(u_T|w_{T+1}) + (\star) + \sum_{t=1}^{T}\phi_t(u_t) + \xi_{1:T} + \delta_{1:T}\\
&\le D_{\psi_1}(u_1|w_1) - D_{\psi_{T+1}}(u_T|w_{T+1}) + D_{\psi_{T+1}}(u_T|w_1) - D_{\psi_1}(u_1|w_1) + \mathcal{P}_{2:T}\\
&\quad + \sum_{t=1}^{T}\phi_t(u_t) + \xi_{1:T} + \delta_{1:T}\\
&= D_{\psi_{T+1}}(u_T|w_1) - D_{\psi_{T+1}}(u_T|w_{T+1}) + \sum_{t=1}^{T}\phi_t(u_t) + \mathcal{P}_{2:T} + \xi_{1:T} + \delta_{1:T}.
\end{aligned}$$

□

In this paper, we will frequently use composite penalties ϕ_t which are a linearization of some other function φ_t. The next lemma shows how this changes the bound.

Lemma A.2. *Under the same conditions as Lemma A.1, let $\varphi_t : W \to \mathbb{R}_+$ be subdifferentiable convex functions and suppose we set $\phi_t(w) = \langle \nabla\varphi_t(w_t), w\rangle$ for some $\nabla\varphi_t(w_t) \in \partial\varphi_t(w_t)$. Then Algorithm 4 guarantees*

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

where $\widetilde{g}_t = g_t + \nabla\varphi_t(w_t)$.

Proof. The proof is immediate by convexity of φ_t. From Lemma A.1 we have

$$R_T(\boldsymbol{u}) \le D_{\psi_{T+1}}(u_T|w_1) - D_{\psi_{T+1}}(u_T|w_{T+1}) + \sum_{t=1}^T \phi_t(u_t) + \mathcal{P}_{2:T} + \xi_{1:T}$$
$$+ \sum_{t=1}^T \langle g_t, w_t - \widetilde{w}_{t+1}\rangle - D_{\psi_t}(\widetilde{w}_{t+1}|w_t) - \Delta_t(\widetilde{w}_{t+1}) - \phi_t(\widetilde{w}_{t+1}).$$

Observe that we can write

$$\begin{aligned}\sum_{t=1}^T \phi_t(u_t) - \phi_t(\widetilde{w}_{t+1}) &= \sum_{t=1}^T \langle \nabla\varphi_t(w_t), u_t - \widetilde{w}_{t+1}\rangle \\ &= \sum_{t=1}^T \langle \nabla\varphi_t(w_t), u_t - w_t\rangle + \langle \nabla\varphi_t(w_t), w_t - \widetilde{w}_{t+1}\rangle \\ &\le \sum_{t=1}^T \varphi_t(u_t) - \varphi_t(w_t) + \langle \nabla\varphi_t(w_t), w_t - \widetilde{w}_{t+1}\rangle,\end{aligned}$$

so plugging this back in above gives the stated result:

$$R_T(\boldsymbol{u}) \le D_{\psi_{T+1}}(u_T|w_1) - D_{\psi_{T+1}}(u_T|w_{T+1}) + \sum_{t=1}^T \varphi_t(u_t) + \mathcal{P}_{2:T} + \xi_{1:T}$$
$$+ \sum_{t=1}^T \langle \widetilde{g}_t, w_t - \widetilde{w}_{t+1}\rangle - D_{\psi_t}(\widetilde{w}_{t+1}|w_t) - \Delta_t(\widetilde{w}_{t+1}) - \varphi_t(w_t),$$

where $\widetilde{g}_t = g_t + \nabla\varphi_t(w_t)$. □

A.1. Multi-scale Experts Algorithm

Algorithm 5 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 for
```

For completeness, in this section we provide a multi-scale experts algorithm which achieves the bound required for our dynamic regret algorithm in Section 4. Our approach is inspired by the Multi-scale Multiplicative-weight with Correction

(MsMwC) algorithm of Chen et al. (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, which requires several rather delicate conditions to come together in the right way.

Theorem A.3. *Let $k \geq \frac{9}{2}$ and assume $\mu_1, \ldots, \mu_N$ satisfy $\mu_i \widetilde{\ell}_{ti} \leq 1$ for all $t \in [T]$ and $i \in [N]$. Then for any $u \in \Delta_N$, Algorithm 5 guarantees*

$$\sum_{t=1}^T \left\langle \widetilde{\ell}_t, p_t - u \right\rangle \leq \sum_{i=1}^N u_i \left[\frac{k \left[\log\left(u_i / p_{1i}\right) + \sum_{t=1}^T \log\left(\frac{1}{1-\beta_t}\right)\right]}{\mu_i} + \mu_i \sum_{t=1}^T \widetilde{\ell}_{ti}^2 \right] + k(1 + \beta_{1:T}) \sum_{i=1}^N \frac{p_{1i}}{\mu_i}.$$

Moreover, for $\beta_t \leq 1 - \exp\left(-\frac{1}{T}\right)$,

$$\sum_{t=1}^T \left\langle \widetilde{\ell}_t, p_t - u \right\rangle \leq \sum_{i=1}^N u_i \left[\frac{k \left[\log\left(u_i/p_{1i}\right) + 1\right]}{\mu_i} + \mu_i \sum_{t=1}^T \widetilde{\ell}_{ti}^2 \right] + 2k \sum_{i=1}^N \frac{p_{1i}}{\mu_i}$$

Proof. The described algorithm is an instance of Algorithm 4 applied to the simplex Δ_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 A.2:

$$\sum_{t=1}^T \left\langle \widetilde{\ell}_t, p_t - u \right\rangle \leq D_\psi(u|p_1) - D_\psi(u|p_{T+1}) + \varphi_{1:T}(u) + \xi_{1:T} + \delta_{1:T},$$

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 Δ_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

$$\begin{aligned}
\sum_{t=1}^T \xi_t &= \sum_{t=1}^T D_\psi(u|p_{t+1}) - D_\psi(u|q_{t+1}) \\
&= k\sum_{t=1}^T\sum_{i=1}^N \frac{u_i \log\left(q_{t+1,i}/p_{t+1,i}\right) + p_{t+1,i} - q_{t+1,i}}{\mu_i} \\
&= k\sum_{t=1}^T\sum_{i=1}^N \frac{u_i \log\left(\frac{q_{t+1,i}}{(1-\beta_t)q_{t+1,i}+\beta_t q_{1,i}}\right)}{\mu_i} \\
&\quad + k\sum_{t=1}^T\sum_{i=1}^N \frac{(1-\beta_t)q_{t+1,i}+\beta_t q_{1,i} - q_{t+1,i}}{\mu_i} \\
&= k\sum_{t=1}^T\sum_{i=1}^N \frac{u_i}{\mu_i}\log\left(\frac{q_{t+1,i}}{(1-\beta_t)q_{t+1,i}+\beta_t q_{1,i}}\right) \\
&\quad + k\sum_{t=1}^T\sum_{i=1}^N \frac{\beta_t(q_{1,i}-q_{t+1,i})}{\mu_i} \\
&\le k\sum_{t=1}^T\sum_{i=1}^N \frac{u_i}{\mu_i}\log\left(\frac{1}{1-\beta_t}\right) + \frac{\beta_t q_{1i}}{\mu_i} \\
&= k\sum_{i=1}^N \frac{u_i}{\mu_i}\sum_{t=1}^T \log\left(\frac{1}{1-\beta_t}\right) + k\beta_{1:T}\sum_{i=1}^N \frac{p_{1i}}{\mu_i},
\end{aligned}$$

where the last line recalls $p_1 = q_1$. Plugging these bounds back into the above regret bound yields

$$\begin{aligned}
\sum_{t=1}^T \left\langle \widetilde{\ell}_t, p_t - u\right\rangle &\le k\sum_{i=1}^N \left[\frac{u_i\log\left(u_i/p_{1i}\right)}{\mu_i} + (1+\beta_{1:T})\frac{p_{1i}}{\mu_i} + \frac{u_i}{\mu_i}\sum_{t=1}^T \log\left(\frac{1}{1-\beta_t}\right)\right] + \varphi_{1:T}(u) + \delta_{1:T} \\
&= \sum_{i=1}^N u_i\left[\frac{k\left[\log\left(u_i/p_{1i}\right) + \sum_{t=1}^T \log\left(\frac{1}{1-\beta_t}\right)\right]}{\mu_i} + \mu_i\sum_{t=1}^T \widetilde{\ell}_{ti}^2\right] + k(1+\beta_{1:T})\sum_{i=1}^N \frac{p_{1i}}{\mu_i} \\
&\quad + \sum_{i=1}^N\sum_{t=1}^T \underbrace{(\widetilde{\ell}_{ti} + \mu_i\widetilde{\ell}_{ti}^2)(p_{ti} - q_{t+1,i}) - D_{\psi_i}(q_{t+1,i}|p_{t,i}) - \mu_i\widetilde{\ell}_{ti}^2 p_{ti}}_{=:\delta_{ti}}, \qquad (6)
\end{aligned}$$

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, δ_{ti}.

Note that by construction, we have $p_{ti} = (1-\beta_t)q_{ti} + \beta_t q_{1i} \ge \beta_t q_{1i} > 0$ 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

$$D_{\psi_i}(q_{t+1,i}|p_{ti}) \ge \frac{1}{2}(p_{ti} - q_{t+1,i})^2\psi_i''(\widetilde{p}_i) = \frac{1}{2}\frac{(p_{ti} - q_{t+1,i})^2 k}{\mu_i\widetilde{p}_i}$$

so using this with the assumption that $\mu_i \left|\widetilde{\ell}_{ti}\right| \leq 1$, we have

$$\begin{aligned}
\delta_{ti} &\leq \left|\widetilde{\ell}_{ti} + \mu_i \widetilde{\ell}_{ti}^2\right| |p_{ti} - q_{t+1,i}| - \frac{1}{2}\frac{(p_{ti} - q_{t+1,i})^2 k}{\mu_i \widetilde{p}_i} - \mu_i \widetilde{\ell}_{ti}^2 p_{ti} \\
&\leq 2\left|\widetilde{\ell}_{ti}\right| |p_{ti} - q_{t+1,i}| - \frac{1}{2}\frac{(p_{ti} - q_{t+1,i})^2 k}{\mu_i \widetilde{p}_i} - \mu_i \widetilde{\ell}_{ti}^2 \widetilde{p}_i + \mu_i \widetilde{\ell}_{ti}^2 |p_{ti} - \widetilde{p}_i| \\
&\overset{(a)}{\leq} 3\left|\widetilde{\ell}_{ti}\right| |p_{ti} - q_{t+1,i}| - \frac{1}{2}\frac{(p_{ti} - q_{t+1,i})^2 k}{\mu_i \widetilde{p}_i} - \mu_i \widetilde{\ell}_{ti}^2 \widetilde{p}_i \\
&\leq \frac{9}{2k}\mu_i \left|\widetilde{\ell}_{ti}\right|^2 \widetilde{p}_i - \mu_i \widetilde{\ell}_{ti}^2 \widetilde{p}_i \\
&\overset{(b)}{\leq} 0,
\end{aligned}$$

where (a) uses $|\widetilde{p}_i - p_{ti}| \leq |q_{t+1,i} - p_{ti}|$ 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 (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

$$\sum_{t=1}^T \left\langle \widetilde{\ell}_t, p_t - u \right\rangle \leq \sum_{i=1}^N u_i \left[\frac{k\left[\log(u_i/p_{1i}) + 1\right]}{\mu_i} + \mu_i \sum_{t=1}^T \widetilde{\ell}_{ti}^2\right] + 2k \sum_{i=1}^N \frac{p_{1i}}{\mu_i}$$

□

B. Proofs for Section 2 (Online Learning with Quadratically Bounded Losses)

B.1. Proof of Theorem 2.2

Theorem 2.2. *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 $\epsilon > 0$, $k \geq 3$, $\kappa \geq 4$, $c \geq 4$, $V_{t+1} = cG_{\max}^2 + G_{1:t}^2$, $\rho_{t+1} = \frac{1}{\sqrt{L_{\max}^2 + L_{1:t}^2}}$, $\alpha_{t+1} = \frac{\sqrt{V_{t+1}}\log^2\left(V_{t+1}/G_{\max}^2\right)}{\epsilon G_{\max}}$, and set*

$$\psi_t(w) = k\int_0^{\|w\|} \min_{\eta \leq \frac{1}{G_{\max}}} \left[\frac{\log(x/\alpha_t + 1)}{\eta} + \eta V_t\right] dx + \frac{\kappa \|w\|^2}{2\rho_t} \qquad \text{and} \qquad \varphi_t(w) = \frac{L_t^2}{2\sqrt{L_{1:t}^2}} \|w\|^2.$$

Then for any $u \in W$, Algorithm 1 guarantees

$$R_T(u) \leq 2\epsilon G_{\max} + \kappa \|u\|^2 \sqrt{L_{\max}^2 + L_{1:T}^2} + 2k\|u\| \max\left\{\sqrt{V_{T+1} F_{T+1}(\|u\|)}, G_{\max} F_{T+1}(\|u\|)\right\}$$

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 A.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,

$$\begin{aligned}\Psi'_t(\|u\|) &= k\min_{\eta\le 1/G_{\max}}\left[\frac{F_t(\|u\|)}{\eta} + \eta V_t\right]\\ &= \begin{cases}2k\sqrt{V_t F_t(\|u\|)} & \text{if } G_{\max}\sqrt{F_t(\|u\|)} \le \sqrt{V_t}\\ kG_{\max}F_t(\|u\|) + k\frac{V_t}{G_{\max}} & \text{otherwise}\end{cases}\\ &\overset{(*)}{\le} \begin{cases}2k\sqrt{V_t F_t(\|u\|)} & \text{if } G_{\max}\sqrt{F_t(\|u\|)} \le \sqrt{V_t}\\ 2kG_{\max}F_t(\|u\|) & \text{otherwise}\end{cases}\\ &= 2k\max\left\{\sqrt{V_t F_t(\|u\|)}, G_{\max}F_t(\|u\|)\right\}.\end{aligned}$$

where $(*)$ observes that $V_t/G_{\max} \le G_{\max}F_t(x)$ whenever $\Psi'_t(x) = kG_{\max}F_t(x) + kV_t/G_{\max}$. Next, using Lemma D.1 we have

$$\varphi_{1:T}(u) = \frac{1}{2}\|u\|^2\sum_{t=1}^T\frac{L_t^2}{\sqrt{L_{1:t}^2}} \le \|u\|^2\sqrt{L_{1:T}^2},$$

so overall we have

$$\sum_{t=1}^T\langle g_t, w_t - u\rangle \le 2k\|u\|\max\left\{\sqrt{V_{T+1}F_{T+1}(\|u\|)}, G_{\max}F_{T+1}(\|u\|)\right\} + \frac{\kappa}{2}\|u\|^2\sqrt{L_{\max}^2 + L_{1:T}^2} + \|u\|^2\sqrt{L_{1:T}^2} + \delta_{1:T} \tag{7}$$

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:

$$\begin{aligned}\delta_t &\le (\|g_t\| + \|\nabla\varphi_t(w_t)\|)\|w_t - w_{t+1}\|\\ &\quad - D_{\psi_t}(w_{t+1}|w_t) - \Delta_t(w_{t+1}) - \varphi_t(w_t)\\ &\le G_t\|w_t - w_{t+1}\| - D_{\Psi_t}(w_{t+1}|w_t) - \Delta_t(w_{t+1})\\ &\quad + 2L_t\|w_t\|\|w_t - w_{t+1}\| - D_{\Phi_t}(w_{t+1}|w_t) - \varphi_t(w_t),\end{aligned}$$

where we slightly abuse notations D_{Ψ_t} and D_{Φ_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

$$\begin{aligned}
2L_t\|w_t\|\|w_t - w_{t+1}\| - D_{\Phi_t}(w_{t+1}|w_t) - \varphi_t(w_t) &\leq 2L_t\|w_t\|\|w_t - w_{t+1}\| - \frac{\kappa}{2\rho_t}\|w_{t+1} - w_t\|^2 - \varphi_t(w_t)\\
&\leq \frac{4\rho_t L_t^2\|w_t\|^2}{2\kappa} - \varphi_t(w_t)\\
&= \frac{2L_t^2\|w_t\|^2}{\kappa\sqrt{L_{\max} + L_{1:t-1}^2}} - \frac{L_t^2}{2\sqrt{L_{1:t}^2}}\|w_t\|^2\\
&\leq \frac{2L_t^2\|w_t\|^2}{\kappa\sqrt{L_{1:t}^2}} - \frac{L_t^2}{2\sqrt{L_{1:t}^2}}\|w_t\|^2\\
&\leq 0
\end{aligned}$$

for $\kappa \geq 4$. Hence,

$$\delta_t \leq G_t\|w_t - w_{t+1}\| - D_{\Psi_t}(w_{t+1}|w_t) - \Delta_t(w_{t+1}),$$

which we will bound by showing that $\Delta_t(w) \geq \eta_t(w)G_t^2$ for some suitable G_t-Lipschitz convex function η_t and then invoking Lemma D.3. To this end, observe that

$$\begin{aligned}
\Delta_t(w) &= \psi_{t+1}(w) - \psi_t(w)\\
&= \underbrace{\Psi_{t+1}(\|w\|) - \Psi_t(\|w\|)}_{=:\Delta_t^\Psi(w)} + \underbrace{\Phi_{t+1}(\|w\|) - \Phi_t(\|w\|)}_{=:\Delta_t^\Phi(w)}\\
&\geq \Delta_t^\Psi(w).
\end{aligned}$$

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

$$\begin{aligned}
\Psi'_{t+1}(x) - \Psi'_t(x) &= k\min_{\eta \leq 1/G_{\max}}\left[\frac{F_{t+1}(x)}{\eta} + \eta V_{t+1}\right] - k\min_{\eta \leq 1/G_{\max}}\left[\frac{F_t(x)}{\eta} + \eta V_t\right]\\
&\geq k\min_{\eta \leq 1/G_{\max}}\left[\frac{F_t(x)}{\eta} + \eta V_{t+1}\right] - k\min_{\eta \leq 1/G_{\max}}\left[\frac{F_t(x)}{\eta} + \eta V_t\right]
\end{aligned}$$

and using the fact that for any $\eta \leq 1/G_{\max}$, we can bound $\frac{F_t(x)}{\eta} + \eta V_t + \eta G_t^2 \geq \min_{\eta^* \leq 1/G}\left[\frac{F_t(x)}{\eta^*} + \eta^* V_t\right] + \eta G_t^2$, we have

$$\begin{aligned}
&\geq k\min_{\eta \leq 1/G_{\max}}\left[\frac{F_t(x)}{\eta} + \eta V_t\right] - k\min_{\eta \leq 1/G_{\max}}\left[\frac{F_t(x)}{\eta} + \eta V_t\right] + kG_t^2\min\left\{\sqrt{\frac{F_t(x)}{V_{t+1}}}, \frac{1}{G_{\max}}\right\}\\
&\geq kG_t^2\min\left\{\sqrt{\frac{F_t(x)}{2V_t}}, \frac{1}{G_{\max}}\right\} \geq G_t^2\min\left\{\sqrt{\frac{F_t(x)}{V_t}}, \frac{1}{G_{\max}}\right\},
\end{aligned}$$

where the last line observes that $\frac{1}{V_t} = \frac{1}{V_{t+1}}\frac{V_{t+1}}{V_t} = \frac{1}{V_{t+1}}\left(1 + \frac{G_t^2}{V_t}\right) \leq \frac{2}{V_{t+1}}$ for $V_t \geq G_t^2$ and recalls $k \geq 3$. 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:

$$\Delta_t^\Psi(\|w\|) \geq G_t^2\int_0^{\|w\|}\min\left\{\sqrt{\frac{F_t(x)}{V_t}}, \frac{1}{G_{\max}}\right\}dx = \eta_t(\|w\|)G_t^2.$$

Hence:

$$
\begin{aligned}
\delta_t &\le G_t \left\|w_t - w_{t+1}\right\| - D_{\psi_t}(w_{t+1}|w_t) - \Delta_t(w_{t+1}) \\
&\le G_t \left\|w_t - w_{t+1}\right\| - D_{\psi_t}(w_t|w_{t+1}) - \eta_t(\|w_{t+1}\|)G_t^2
\end{aligned}
\tag{8}
$$

Finally, we conclude by showing that ψ_t satisfies the assumptions of Lemma D.3 w.r.t this function η_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

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

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

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

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, Ψ_t satisfies the conditions of Lemma D.3 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 (8) over all t, we have

$$\begin{aligned}
\sum_{t=1}^T \delta_t &\le \sum_{t=1}^T G_t \left\|w_t - w_{t+1}\right\| - D_{\Psi_t}(w_{t+1}|w_t) - \eta_t(\|w_{t+1}\|)G_t^2 \\
&\le \sum_{t=1}^T \frac{2G_t^2}{\Psi_t''(\mathring{x}_t)} = \sum_{t=1}^T \frac{2G_t^2}{k\sqrt{V_t}}(\|\mathring{x}_t\| + \alpha_t) \\
&\le \sum_{t=1}^T \frac{2e\alpha_t G_t^2}{k\sqrt{V_t}} \le \sum_{t=1}^T 2\frac{\alpha_t G_t^2}{\sqrt{V_t}}
\end{aligned}$$

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 $c \ge 4$. Finally, plugging this back into Equation (7) yields

$$\begin{aligned}
\sum_{t=1}^T \langle g_t, w_t - u\rangle &\le 2k\left\|u\right\| \max\left\{\sqrt{V_{T+1}F_{T+1}(\|u\|)}, G_{\max}F_{T+1}(\|u\|)\right\} \\
&\quad + \frac{\kappa}{2}\left\|u\right\|^2 \sqrt{L_{\max}^2 + L_{1:T}^2} + \left\|u\right\|^2 \sqrt{L_{1:T}^2} + \delta_{1:T} \\
&\le 2k\left\|u\right\| \max\left\{\sqrt{V_{T+1}F_{T+1}(\|u\|)}, G_{\max}F_{T+1}(\|u\|)\right\} \\
&\quad + \frac{\kappa}{2}\left\|u\right\|^2 \sqrt{L_{\max}^2 + L_{1:T}^2} + \left\|u\right\|^2 \sqrt{L_{1:T}^2} + 2\epsilon G_{\max} \\
&\le 2\epsilon G_{\max} + \kappa\left\|u\right\|^2 \sqrt{L_{\max}^2 + L_{1:T}^2} \\
&\quad 2k\left\|u\right\| \max\left\{\sqrt{V_{T+1}F_{T+1}(\|u\|)}, G_{\max}F_{T+1}(\|u\|)\right\}
\end{aligned}$$

□

B.2. Proof of Theorem 2.3

Theorem 2.3. *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 $g_1, \dots, g_T$ satisfying $\|g_t\| \le G + L\left\|w_t\right\|$ and a comparator $u \in \mathbb{R}^2$ such that*

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

Proof. Let $w_t \in \mathbb{R}^2$ be the output of algorithm $\mathcal{A}$ at time t. Consider sequences $g_1, \dots, g_T$ 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 ε_t are independent random signs. Consider the worst-case regret against a comparator constrained to an ℓ_∞ ball of radius U:

$$
\begin{aligned}
\sup_{g_1,\dots,g_T} R_T &= \sup_{g_1,\dots,g_T} \sum_{t=1}^T \langle g_t, w_t \rangle - \min_{u:\|u\|_\infty \le U} \sum_{t=1}^T \langle g_t, u \rangle \\
&\ge \mathbb{E}_{\varepsilon_1,\dots,\varepsilon_T}\left[\sum_{t=1}^T \langle \widetilde{g}_t, w_t \rangle - \min_{u:\|u\|_\infty \le U} \sum_{t=1}^T \langle \widetilde{g}_t, u \rangle \right] \\
&\ge \mathbb{E}_{\varepsilon_1,\dots,\varepsilon_T}\left[-\sum_{t=1}^T G\left\|w_t\right\| - \min_{u:\|u\|_\infty \le U} \sum_{t=1}^T -Gu_1 - u_2\varepsilon_t L\left\|w_t\right\| \right] \\
&= \mathbb{E}_{\varepsilon_1,\dots,\varepsilon_T}\left[-G\sum_{t=1}^T \left\|w_t\right\| + GTU + \max_{|u_2|\le U} u_2 L \sum_{t=1}^T \varepsilon_t \left\|w_t\right\| \right] \\
&= \mathbb{E}_{\varepsilon_1,\dots,\varepsilon_T}\left[GTU + UL\left| \sum_{t=1}^T \varepsilon_t \left\|w_t\right\| \right| - G\sum_{t=1}^T \left\|w_t\right\| \right] \\
&\overset{(a)}{\ge} \mathbb{E}_{\varepsilon_1,\dots,\varepsilon_T}\left[GTU + \frac{UL}{\sqrt{2}}\sqrt{\sum_{t=1}^T \left\|w_t\right\|^2} - G\sum_{t=1}^T \left\|w_t\right\| \right] \\
&\overset{(b)}{\ge} \mathbb{E}_{\varepsilon_1,\dots,\varepsilon_T}\left[GTU + \frac{UL}{\sqrt{2}}\sqrt{\sum_{t=1}^T \left\|w_t\right\|^2} - G\sqrt{T\sum_{t=1}^T \left\|w_t\right\|^2} \right]
\end{aligned}
$$

where (a) applies Khintchine inequality, (b) applies Cauchy-Schwarz inequality, and choosing $U = \frac{G}{L}\sqrt{2T}$ we have

$$
= GTU = \frac{L}{\sqrt{2}} U^2 \sqrt{T} = \frac{L\left\|u\right\|^2\sqrt{T}}{2\sqrt{2}},
$$

where the final equality bounds $\left\|u\right\|^2 = u_1^2 + u_2^2 \le 2U^2$. Hence, there exists a sequence of g_t which incurs at least $\Omega(L\left\|u\right\|^2\sqrt{T})$ regret. Moreover, for any algorithm which guarantees $R_T(\mathbf{0}) \le \epsilon$, there exists a sequence $g_1, \dots, g_T$ with $\left\|g_t\right\| \le G$ for all t such that for any T and u, $R_T(u) \ge \frac{G}{3\sqrt{2}}\left\|u\right\|\sqrt{T\log\left(\left\|u\right\|\sqrt{T}/\sqrt{2}\epsilon\right)}$ (Mcmahan & Streeter, 2012, Theorem 8). Thus, taking the worst of these two sequences yields

$$
\sup_{g_1,\dots,g_T} R_T \ge \max\left\{ \frac{G}{3\sqrt{2}}\left\|u\right\|\sqrt{T\log\left(\left\|u\right\|\sqrt{T}/\sqrt{2}\epsilon\right)}, \frac{L\left\|u\right\|^2\sqrt{T}}{2\sqrt{2}} \right\}
$$

□

C. Proofs for Section 4 (Dynamic Regret)

The main objective of this section is to prove Theorems 4.1 and 4.3. 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 η, 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), the regret of Algorithm 3 in terms of the regret of any of the base algorithms (Lemma C.2), as well as some utility lemmas (Lemmas C.3 to C.6) to help tame some unwieldy algebraic expressions and case work. We then prove the main results Theorems 4.1 and 4.3 in Appendices C.1 and C.2 respectively. Finally, we prove our lowerbound Theorem 4.2 in Appendix C.3.

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. *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 4 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 A.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 ψ, (b) is Fenchel-Young inequality, and (c) uses $K\eta L_t \leq 1$.

□

The following lemma provides a generic regret bound for Algorithm 3. 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.2. *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 3 guarantees*

$$\begin{aligned} R_T(\boldsymbol{u}) &\le 2kC_{\mathcal{S}} + 2kDG_{\max}\Lambda_T(\tau) + \frac{\|u_T\|^2 + 2DP_T + 4kD^2\Lambda_T(\tau)}{2\eta} \\ &\quad + K\eta\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t(w_t^{(\tau)})\right] + 4\eta\sum_{t=1}^T \left\|g_t^{(\tau)}\right\|^2 \end{aligned}$$

where $k \ge 9/2$ 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 playing $w_{t+1}^{(\tau)} = \Pi_{w\in W:\|w\|\le D}\left(w_t^{(\tau)} - \eta(1 + K\eta L_t)g_t^{(\tau)}\right)$ for $g_t^{(\tau)} \in \partial\ell_t(w_t^{(\tau)})$. Algorithm 3 is constructed as a collection of algorithms $\mathcal{A}_\tau$, with an multi-scale experts algorithm (Algorithm 5) 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 ℓ_t and Jensen's inequality:

$$\begin{aligned} &\le R_T^{\mathcal{A}_\tau}(\boldsymbol{u}) + \sum_{t=1}^T \left[\sum_{\widetilde{\tau}\in\mathcal{S}} p_t(\widetilde{\tau})\ell_t\left(w_t^{(\widetilde{\tau})}\right)\right] - \ell_t\left(w_t^{(\tau)}\right) \\ &= R_T^{\mathcal{A}_\tau}(\boldsymbol{u}) + \sum_{t=1}^T \sum_{\widetilde{\tau}\in\mathcal{S}} \ell_t\left(w_t^{(\widetilde{\tau})}\right)\left[p_t(\widetilde{\tau}) - \mathbf{1}\left\{\tau = \widetilde{\tau}\right\}\right] \\ &\overset{(a)}{=} R_T^{\mathcal{A}_\tau}(\boldsymbol{u}) \\ &\qquad + \sum_{t=1}^T \sum_{\widetilde{\tau}\in\mathcal{S}} \left[\ell_t\left(w_t^{(\widetilde{\tau})}\right) - \ell_t(\widetilde{w}_t)\right]\left[p_t(\widetilde{\tau}) - p_\tau^*(\widetilde{\tau})\right] \\ &\qquad + \sum_{t=1}^T \sum_{\widetilde{\tau}\in\mathcal{S}} \ell_t(\widetilde{w}_t)(p_t(\widetilde{\tau}) - p_\tau^*(\widetilde{\tau})) \\ &\overset{(b)}{=} R_T^{\mathcal{A}_\tau}(\boldsymbol{u}) + \sum_{t=1}^T \sum_{\widetilde{\tau}\in\mathcal{S}} \left[\ell_t\left(w_t^{(\widetilde{\tau})}\right) - \ell_t(\widetilde{w}_t)\right]\left[p_t(\widetilde{\tau}) - p_\tau^*(\widetilde{\tau})\right] \\ &\overset{(c)}{=} R_T^{\mathcal{A}_\tau}(\boldsymbol{u}) + \underbrace{\sum_{t=1}^T \left\langle \widetilde{\ell}_t, p_t - p_\tau^*\right\rangle}_{=:R_T^{\text{Meta}}(p_\tau^*)} \\ &= R_T^{\mathcal{A}_\tau}(\boldsymbol{u}) + R_T^{\text{Meta}}(p_\tau^*), \end{aligned} \tag{9}$$

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 that

$$R_T^{\mathcal{A}_\tau}(u) \leq \frac{\|u_T\|^2 + 2DP_T}{2\eta} + K\eta\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t(w_t^{(\tau)})\right] + 2\eta\sum_{t=1}^T \left\|g_t^{(\tau)}\right\|^2. \tag{10}$$

To bound $R_T^{\text{Meta}}(p^*_\tau)$, 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 $\mu_{\widetilde{\tau}} = \frac{1}{2\widetilde{D}\left(G_{\max}+\widetilde{D}/\widetilde{\eta}\right)}$ 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 A.3 and we have:

$$R_T^{\text{Meta}}(p^*_\tau) \leq \sum_{\widetilde{\tau}\in\mathcal{S}} p^*_\tau(\widetilde{\tau})\left[\frac{k\left[\log\left(p^*_\tau(\widetilde{\tau})/p_{1\widetilde{\tau}}\right) + 1\right]}{\mu_{\widetilde{\tau}}} + \mu_{\widetilde{\tau}}\sum_{t=1}^T \widetilde{\ell}^2_{t\widetilde{\tau}}\right] + 2k\sum_{\widetilde{\tau}\in\mathcal{S}}\frac{p_{1\widetilde{\tau}}}{\mu_{\widetilde{\tau}}}$$

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

$$\begin{aligned}\frac{k\left[\log\left(p^*_\tau(\tau)/p_{1\tau}\right) + 1\right]}{\mu_\tau} + \mu_\tau\sum_{t=1}^T \widetilde{\ell}^2_{t\tau} &= 2kD\left(G_{\max} + \frac{D}{\eta}\right)\left[\log\left(1/p_{1\tau}\right) + 1\right] + \frac{\eta}{2D\left(G_{\max}\eta + D\right)}\sum_{t=1}^T \widetilde{\ell}^2_{t,\tau} \\ &\leq 2kD\left(G_{\max} + \frac{D}{\eta}\right)\left[\log\left(1/p_{1\tau}\right) + 1\right] + \frac{\eta}{2D^2}\sum_{t=1}^T \left\|\nabla\ell_t(w_t^{(\tau)})\right\|^2 4D^2 \\ &= 2kD\left(G_{\max} + \frac{D}{\eta}\right)\left[\log\left(1/p_{1\tau}\right) + 1\right] + 2\eta\sum_{t=1}^T \left\|g_t^{(\tau)}\right\|^2,\end{aligned}$$

and so with $p_{1,\tau} = \frac{\mu_\tau^2}{\sum_{\widetilde{\tau}\in\mathcal{S}}\mu^2_{\widetilde{\tau}}}$, we have

$$R_T^{\text{Meta}}(p^*_\tau) \leq 2kD\left(G_{\max} + \frac{D}{\eta}\right)\left[\log\left(\frac{\sum_{\widetilde{\tau}\in\mathcal{S}}\mu^2_{\widetilde{\tau}}}{\mu_\tau^2}\right) + 1\right] + 2\eta\sum_{t=1}^T \left\|g_t^{(\tau)}\right\|^2 + 2k\sum_{\widetilde{\tau}\in\mathcal{S}}\frac{\mu_{\widetilde{\tau}}}{\sum_{\widetilde{\tau}\in\mathcal{S}}\mu^2_{\widetilde{\tau}}}.$$

Combining this with Equations (9) and (10) yields the stated result:

$$
\begin{aligned}
R_T(\boldsymbol{u}) &\leq \frac{\|u_T\|^2 + 2DP_T}{2\eta} + K\eta \sum_{t=1}^{T} L_t \left[\ell_t(u_t) - \ell_t(w_t^{(\tau)})\right] + 4\eta \sum_{t=1}^{T} \left\|g_t^{(\tau)}\right\|^2 \\
&\quad + 2kD\left(G_{\max} + \frac{D}{\eta}\right)\left[\log\left(\frac{\sum_{\widetilde{\tau}\in\mathcal{S}} \mu_{\widetilde{\tau}}^2}{\mu_\tau}\right) + 1\right] + 2k\frac{\sum_{\widetilde{\tau}\in\mathcal{S}} \mu_{\widetilde{\tau}}}{\sum_{\widetilde{\tau}\in\mathcal{S}} \mu_{\widetilde{\tau}}^2} \\
&= 2kC_{\mathcal{S}} + 2kDG_{\max}\Lambda_T(\tau) + \frac{\|u_T\|^2 + 2DP_T + 4kD^2\Lambda_T(\tau)}{2\eta} \\
&\quad + K\eta \sum_{t=1}^{T} L_t \left[\ell_t(u_t) - \ell_t(w_t^{(\tau)})\right] + 4\eta \sum_{t=1}^{T} \left\|g_t^{(\tau)}\right\|^2
\end{aligned}
$$

where 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 Λ_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.3. *Let $0 < \eta_{\min} \leq \eta_{\max}$, $0 < D_{\min} \leq D_{\max}$, and define the hyperparameter set $\mathcal{S} = \mathcal{S}_\eta \times \mathcal{S}_D$ for $\mathcal{S}_\eta = \left\{\eta_i = \left[\eta_{\min}2^i \wedge \eta_{\max}\right] : i \geq 0\right\}$ and $\mathcal{S}_D = \left\{D_j = \left[D_{\min}2^j \wedge D_{\max}\right] : j \geq 0\right\}$. For each $\tau = (\eta, D) \in \mathcal{S}$, let $\mu_\tau = \frac{1}{2D(G_{\max}+D/\eta)}$. Then*

$$
C_{\mathcal{S}} \overset{\text{def}}{=} \frac{\sum_{\tau\in\mathcal{S}} \mu_\tau}{\sum_{\tau\in\mathcal{S}} \mu_\tau^2} \leq 2\sqrt{T}D_{\min}\left(G_{\max} + \frac{D_{\min}}{\eta_{\max}}\right)
$$

and for any $\tau \in \mathcal{S}$,

$$
\Lambda_T(\tau) \overset{\text{def}}{=} \log\left(\frac{\sum_{\widetilde{\tau}\in\mathcal{S}} \mu_{\widetilde{\tau}}^2}{\mu_\tau^2}\right) + 1 \leq \log\left(\frac{24\eta_{\max}^2 D^4}{\eta_{\min}^2 D_{\min}^4} \wedge \frac{6\left|\mathcal{S}_\eta\right| D^2}{D_{\min}^2}\right) + 1
$$

Proof. For the first statement, we have

$$
\begin{aligned}
C_{\mathcal{S}} = \frac{\sum_{\widetilde{\tau}\in\mathcal{S}} \mu_{\widetilde{\tau}}}{\sum_{\widetilde{\tau}\in\mathcal{S}} \mu_{\widetilde{\tau}}^2} &\leq \sqrt{\frac{T}{\sum_{\widetilde{\tau}\in\mathcal{S}} \mu_{\widetilde{\tau}}^2}} \leq \sqrt{\frac{T}{\mu_{(\eta_{\max}, D_{\min})}^2}} \\
&= \sqrt{T(2D_{\min})^2\left(G_{\max} + D_{\min}/\eta_{\max}\right)^2} \\
&= 2\sqrt{T}D_{\min}\left(G_{\max} + \frac{D_{\min}}{\eta_{\max}}\right)
\end{aligned}
$$

where the first inequality applies Cauchy-Schwarz inequality. Moreover, for any $\tau = (\eta, D) \in \mathcal{S}$ we have

$$\begin{aligned}
\frac{\sum_{\widetilde{\tau}\in\mathcal{S}} \mu_{\widetilde{\tau}}^2}{\mu_\tau^2} &= \frac{1}{\mu_{(\eta,D)}^2}\left[\sum_{(\eta_i,D_j)\in\mathcal{S}} \frac{1}{(2D_j)^2\left[G_{\max} + D_j/\eta_i\right]^2}\right] \\
&\le \frac{1}{4\mu_{(\eta,D)}^2}\sum_{(\eta_i,D_j)\in\mathcal{S}} \frac{\eta_i^2}{D_j^4} \\
&= \frac{1}{4\mu_{(\eta,D)}^2}\sum_{(\eta_i,D_j)\in\mathcal{S}} \frac{2^{2i}\eta_{\min}^2}{D_{\min}^4 2^{4j}} \\
&= \frac{\eta_{\min}^2}{4\mu_{(\eta,D)}^2 D_{\min}^4}\sum_{i=0}^{\lceil\log_2(\eta_{\max}/\eta_{\min})\rceil}\sum_{j=0}^{\lceil\log_2(D_{\max}/D_{\min})\rceil} \frac{2^{2i}}{2^{4j}} \\
&\le \frac{\eta_{\min}^2}{4\mu_{(\eta,D)}^2 D_{\min}^4}\frac{2^{2\lceil\log_2(\eta_{\max}/\eta_{\min})\rceil+2}-1}{3}\frac{1}{1-\frac{1}{16}} \\
&\le \frac{\eta_{\min}^2}{4\mu_{(\eta,D)}^2 D_{\min}^4}\frac{2^{\log_2(\eta_{\max}^2/\eta_{\min}^2)+4}-1}{3}\frac{16}{15} \\
&\le \frac{\eta_{\min}^2}{4\mu_{(\eta,D)}^2 D_{\min}^4}16\frac{\eta_{\max}^2}{\eta_{\min}^2}\frac{16}{45} \\
&\le \frac{6\eta_{\max}^2}{4\mu_{(\eta,D)}^2 D_{\min}^4} = \frac{3\eta_{\max}^2}{2\mu_{(\eta,D)}^2 D_{\min}^4}
\end{aligned}$$

At the same time, we can also bound this term as

$$\begin{aligned}
\frac{\sum_{\widetilde{\tau}\in\mathcal{S}} \mu_{\widetilde{\tau}}^2}{\mu_\tau^2} &= \frac{1}{\mu_{(\eta,D)}^2}\left[\sum_{(\eta_i,D_j)\in\mathcal{S}} \frac{1}{(2D_j)^2\left[G_{\max} + D_j/\eta_i\right]^2}\right] \\
&\le \frac{1}{4\mu_{(\eta,D)}^2}\sum_{(\eta_i,D_j)\in\mathcal{S}} \frac{1}{D_j^2 G_{\max}^2} \\
&\le \frac{|\mathcal{S}_\eta|}{4\mu_{(\eta,D)}^2 G_{\max}^2}\sum_{j=0}^{\lceil\log_2(D_{\max}/D_{\min})\rceil} \frac{1}{D_{\min}^2 2^{2j}} \\
&\le \frac{|\mathcal{S}_\eta|}{4\mu_{(\eta,D)}^2 G_{\max}^2 D_{\min}^2}\frac{1}{1-\frac{1}{4}} \\
&\le \frac{4\,|\mathcal{S}_\eta|}{4\cdot 3\mu_{(\eta,D)}^2 G_{\max}^2 D_{\min}^2} = \frac{|\mathcal{S}_\eta|}{3\mu_{(\eta,D)}^2 G_{\max}^2 D_{\min}^2}
\end{aligned}$$

Hence,

$$\begin{aligned}
\Lambda_T(\eta, D) &= \log\left(\frac{\sum_{\widetilde{\tau}} \mu_{\widetilde{\tau}}}{\mu_\tau^2}\right) + 1 \\
&\le \log\left(\left[\frac{3\eta_{\max}^2}{2D_{\min}^2} \wedge \frac{|\mathcal{S}_\eta|}{3G_{\max}^2}\right]\frac{1}{\mu_{(\eta,D)}^2 D_{\min}^2}\right) + 1 \\
&= \log\left(\left[\frac{3\eta_{\max}^2}{2D_{\min}^2} \wedge \frac{|\mathcal{S}_\eta|}{3G_{\max}^2}\right]\frac{(2D)^2\left[G_{\max} + D/\eta\right]^2}{D_{\min}^2}\right) + 1.
\end{aligned}$$

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.4 provides a simple but tedius calculation which we will use a few times in the proof of Theorem 4.1.

Lemma C.4. *Let ℓ_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 ℓ_t is (G_t, L_t)-quadratically bounded, and $g_t \in \partial\ell_t(w)$ where $\|w\| \leq D$ we have

$$\left\|g_t\right\|^2 \leq (G_t + L_t\|w\|)^2 \leq 2G_t^2 + 2L_t^2\|w\|^2 \leq 2G_t^2 + 2L_t^2 D^2.$$

Moreover, letting $\nabla\ell_t(u) \in \partial\ell_t(u)$ and $\|u\| \leq D$ we have

$$\begin{aligned}
L_t\left(\ell_t(u) - \ell_t(w)\right) &\leq L_t\left\|\nabla\ell_t(u)\right\|\|u - w\| \\
&\leq 2DL_t\left\|\nabla\ell_t(u)\right\| \\
&\leq 2DL_t\left(G_t + L_t D\right) \\
&= 2DL_t G_t + 2L_t^2 D^2 \\
&\leq G_t^2 + L_t^2 D^2 + 2L_t^2 D^2 \\
&= G_t^2 + 3L_t^2 D^2.
\end{aligned}$$

Thus,

$$\begin{aligned}
c_1 L_t\left(\ell_t(u) - \ell_t(w)\right) + c_2\left\|g_t\right\|^2 &\leq (c_1 + 2c_2)G_t^2 + (3c_1 + 2c_2)L_t^2 D^2 \\
&\leq 3(c_1 + c_2)\left(G_t^2 + L_t^2 D^2\right)
\end{aligned}$$

□

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.5) or "too small" (Lemma C.6).

Lemma C.5. *For all t let ℓ_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.6. *For all t let ℓ_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.2 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.2 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 $M = \max_t \|u_t\| \leq D_{\min}$ and $\frac{D_{\min}}{\eta_{\min}} = K(G_{\max} + \epsilon L_{\max})$, we have

$$\begin{aligned}
\frac{\|u_T\|^2 + 2D_{\min}P_T + 4kD_{\min}^2\Lambda_T(\tau)}{2\eta_{\min}} &\leq \frac{D_{\min}}{\eta_{\min}}\frac{1}{2}\left(\|u_T\| + 2P_T + 4kD_{\min}\Lambda_T(\tau)\right) \\
&= \frac{1}{2}K(G_{\max} + \epsilon L_{\max})\left(\|u_T\| + 2P_T + 4k\frac{\epsilon\Lambda_T(\tau)}{T}\right).
\end{aligned}$$

Moreover, by Lemma C.4 we have

$$\begin{aligned}\sum_{t=1}^{T} K\eta_{\min}L_t\left[\ell_t(u_t)-\ell_t(w_t^{(\tau)})\right]+4\eta_{\min}\left\|g_t^{(\tau)}\right\|^2 &\le \eta_{\min}\sum_{t=1}^{T}3(K+4)\left(G_t^2+L_t^2D_{\min}^2\right)\\ &\le 3(K+4)\frac{\epsilon\left(G_{\max}^2+L_{\max}^2D_{\min}^2\right)}{K(G_{\max}+\epsilon L_{\max})}\\ &\le \frac{3(K+4)}{K}\left(\epsilon G_{\max}+\frac{\epsilon^2L_{\max}}{T^2}\right)\end{aligned}$$

Plugging in the previous two displays back into the full regret bound yields

$$\begin{aligned}R_T(\boldsymbol{u}) &\le 2kC_{\mathcal{S}}+2k\epsilon G_{\max}\frac{\Lambda_T(\tau)}{T}+\frac{1}{2}K(G_{\max}+\epsilon L_{\max})\left(\|u_T\|+2P_T+4k\frac{\epsilon\Lambda_T(\tau)}{T}\right)\\ &\quad+\frac{3(K+4)}{K}\left[\epsilon G_{\max}+\frac{L_{\max}\epsilon^2}{T^2}\right]\\ &\le 2kC_{\mathcal{S}}+\epsilon G_{\max}\left[\frac{3(K+4)}{K}+\frac{(K+1)2k\Lambda_T(\tau)}{T}\right]+\epsilon^2L_{\max}\left[\frac{3(K+4)}{KT^2}+\frac{2kK\Lambda_T(\tau)}{T}\right]\\ &\quad+K(G_{\max}+\epsilon L_{\max})\left[M+P_T\right].\end{aligned}$$

Finally, Lemma C.3 bounds

$$\begin{aligned}2kC_{\mathcal{S}} &\le 2k\cdot 2\sqrt{T}D_{\min}\left(G_{\max}+\frac{D_{\min}}{\eta_{\max}}\right)\\ &\le 4k\sqrt{T}\frac{\epsilon}{T}\left[G_{\max}+\frac{K\epsilon L_{\max}}{T}\right]\\ &\le \frac{4k\left(\epsilon G_{\max}+K\epsilon^2L_{\max}/T\right)}{\sqrt{T}}\end{aligned}$$

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:

$$\begin{aligned}R_T(\boldsymbol{u}) &\le \epsilon G_{\max}(K+4)\left[\frac{3}{K}+\frac{k}{\sqrt{T}}+\frac{2k\Lambda_T(\eta_{\min},D_{\min})}{T}\right]\\ &\quad+\epsilon^2L_{\max}(K+4)\left[\frac{3}{KT^2}+\frac{4k}{T^{3/2}}+\frac{2k\Lambda_T(\eta_{\min},D_{\min})}{T}\right]\\ &\quad+K(G_{\max}+\epsilon L_{\max})\left[M+P_T\right]\\ &\le \mathcal{C}_T\left(\epsilon G_{\max}+\epsilon^2L_{\max}\right)+K(G_{\max}+\epsilon L_{\max})\left[M+P_T\right]\\ &=(G_{\max}+\epsilon L_{\max})\left[K(M+P_T)+\epsilon\mathcal{C}_T\right]\end{aligned}$$

where

$$\begin{aligned}\mathcal{C}_T &\le (K+4)\left(\frac{3}{K}+\frac{4k}{\sqrt{T}}+\frac{2k\left(\log\left(\log_2\left(\frac{TG_{\max}}{\epsilon L_{\max}}\right)+2\right)+3\right)}{T}\right)\\ &\le O\left(\frac{\log\left(\log\left(\frac{G_{\max}}{\epsilon L_{\max}}\right)\right)}{T}\right)\end{aligned}$$

□

C.1. Proof of Theorem 4.1

Theorem 4.1. *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 $\epsilon > 0$, $K \ge 8$, $\beta_t = 1 - \exp(-1/T)$ for all t, and for any $i, j \ge 0$ let $D_j = \frac{\epsilon}{T}\left[2^j \wedge 2^T\right]$ and $\eta_i = \left[\frac{\epsilon 2^i}{K(G_{\max}+\epsilon L_{\max})T} \wedge \frac{1}{KL_{\max}}\right]$, 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 3 guarantees*

$$R_T(\boldsymbol{u}) \le O\Bigg(\Big[G_{\max} + (M+\epsilon)L_{\max}\Big]\Big[(M+\epsilon)\Lambda_T^* + P_T\Big] + \sqrt{(M^2\Lambda_T^* + MP_T)\sum_{t=1}^T G_t^2 + L_t^2 M^2,}\Bigg).$$

where $P_T = \sum_{t=2}^T \|u_t - u_{t-1}\|$, $M = \max_t \|u_t\|$, and $\Lambda_T^* \le O\left(\log\left(\frac{MT\log(T)}{\epsilon}\right) + \log\left(\log\left(\frac{G_{\max}}{\epsilon L_{\max}}\right)\right)\right)$. Moreover, when the losses are L_t-smooth, the bound automatically improves to

$$\begin{aligned}R_T(\boldsymbol{u}) \le O\Bigg(&\Big[G_{\max} + (M+\epsilon)L_{\max}\Big]\Big[(M+\epsilon)\Lambda_T^* + P_T\Big]\\ &+ \sqrt{(M^2\Lambda_T^* + MP_T)\left[\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t^*\right] \wedge \sum_{t=1}^T G_t^2 + L_t^2 M^2\right]}\Bigg).\end{aligned}$$

Proof. First observe that we can assume that there is a $\tau = (\eta, D) \in \mathcal{S}$ for which $D \ge \max_t \|u_t\| = M$, since otherwise using Lemma C.5 with $\varepsilon = \frac{\epsilon}{T}$ the regret is bounded as

$$R_T(\boldsymbol{u}) \le 2M\left(G_{\max} + ML_{\max}\right)\log\left(\frac{MT}{\epsilon}\right). \tag{11}$$

Likewise, if $M \le D_{\min}$ then by Lemma C.6 we have

$$R_T(\boldsymbol{u}) \le (G_{\max} + L_{\max}\epsilon)\left[K(M + P_T) + \epsilon\mathcal{C}_T\right], \tag{12}$$

where $\mathcal{C}_T \le O\left(\frac{\log\left(\log\left(\frac{G_{\max}}{\epsilon L_{\max}}\right)\right)}{T}\right)$. Otherwise, we have $M \in [D_{\min}, D_{\max}]$, in which case there is a $D_j = \frac{\epsilon 2^j}{T}$ for which $D_j \ge M \ge D_{j-1} = \frac{1}{2}D_j$, so for any $\tau = (\eta, D_j) \in \mathcal{S}$ we can apply Lemma C.2 to get

$$\begin{aligned}R_T(\boldsymbol{u}) \le 2kC_{\mathcal{S}} + 2kD_j G_{\max}\Lambda_T(\tau) + \frac{\|u_T\|^2 + 2D_jP_T + 4kD_j^2\Lambda_T(\tau)}{2\eta}\\ + K\eta\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t(w_t^{(\tau)})\right] + 4\eta\sum_{t=1}^T \left\|g_t^{(\tau)}\right\|^2,\end{aligned}$$

where $g_t^{(\tau)} \in \partial \ell_t(w_t^{(\tau)})$, $P_T = \sum_{t=2}^T \|u_t - u_{t-1}\|$, and

$$
\begin{aligned}
C_{\mathcal{S}} &= \frac{\sum_{\widetilde{\tau} \in \mathcal{S}} \mu_{\widetilde{\tau}}}{\sum_{\widetilde{\tau} \in \mathcal{S}} \mu_{\widetilde{\tau}}} \\
\Lambda_T(\eta, D_j) &= \log\left(\frac{\sum_{\widetilde{\tau} \in \mathcal{S}} \mu_{\widetilde{\tau}}^2}{\mu_{(\eta, D_j)}^2}\right) + 1 \\
&= \log\left(D_j^2 \left[G_{\max} + \frac{D_j}{\eta}\right]^2 \sum_{\widetilde{\tau} \in \mathcal{S}} \mu_{\widetilde{\tau}}^2\right) + 1 \\
&\le \log\left((2M)^2 \left[G_{\max} + \frac{2M}{\eta_{\min}}\right]^2 \sum_{\widetilde{\tau} \in \mathcal{S}} \mu_{\widetilde{\tau}}^2\right) + 1 \\
&= \Lambda_T(\eta_{\min}, 2M).
\end{aligned}
$$

Thus, bounding $D_j \le 2M$ 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:

$$
\begin{aligned}
R_T(\boldsymbol{u}) \le 2kC_{\mathcal{S}} + 4kMG_{\max}\Lambda_T(\eta_{\min}, 2M) + \frac{M^2\left(1 + 16k\Lambda_T(\eta_{\min}, 2M)\right) + 4MP_T}{2\eta} \\
\eta \underbrace{\sum_{t=1}^T \left[KL_t\left[\ell_t(u_t) - \ell_t(w_t^{(\tau)})\right] + 4\left\|g_t^{(\tau)}\right\|^2\right]}_{=:\Omega_T}.
\end{aligned}
\tag{13}
$$

Next, we show that there is an η for which the above expression is well-controlled.

Observe that choosing η optimally in Equation (13) would yield

$$
\eta^* = \sqrt{\frac{M^2(1 + 16k\Lambda_T(\eta_{\min}, 2M)) + 4MP_T}{2\Omega_T}}.
$$

If $\eta^* \ge \eta_{\max}$, then choosing $\eta = \eta_{\max}$ yields

$$
\begin{aligned}
R_T(\boldsymbol{u}) &\le 2kC_{\mathcal{S}} + 4kMG_{\max}\Lambda_T(\eta_{\min}, 2M) + \frac{M^2\left(1 + 16k\Lambda_T(\eta_{\min}, 2M)\right) + 4MP_T}{2\eta_{\max}} + \eta^*\Omega_T \\
&= 2kC_{\mathcal{S}} + 4kMG_{\max}\Lambda_T(\eta_{\min}, 2M) + \frac{KL_{\max}}{2}\left[M^2(1 + 16k\Lambda_T(\eta_{\min}, 2M)) + 4MP_T\right] \\
&\quad + \sqrt{\frac{1}{2}\left[M^2\left(1 + 16k\Lambda_T(\eta_{\min}, 2M)\right) + 4MP_T\right]\Omega_T}.
\end{aligned}
\tag{14}
$$

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.4, we have

$$
\begin{aligned}
\Omega_T &= \sum_{t=1}^T KL_t\left[\ell_t(u_t) - \ell_t(w_t^{(\tau)})\right] + 4\left\|g_t^{(\tau)}\right\|^2 \\
&\le \sum_{t=1}^T 3(K+4)\left(G_{\max}^2 + L_{\max}^2 D_j^2\right) \\
&\le 3(K+4)\left(G_{\max}^2 + 4M^2L_{\max}^2\right)T.
\end{aligned}
$$

Thus

$$\begin{aligned}\frac{\epsilon\Omega_T}{K\left(G_{\max}+\epsilon L_{\max}\right)T} &\le \frac{\epsilon\cdot 3(K+4)\left(G_{\max}^2+4M^2L_{\max}^2\right)T}{K\left(G_{\max}+\epsilon L_{\max}\right)T}\\ &\le \frac{3(K+4)}{K}\left(\epsilon G_{\max}+4M^2L_{\max}\right)\\ &\le (K+4)\left(\epsilon G_{\max}+4M^2L_{\max}\right)\end{aligned}$$

for $K\ge 3$. so overall when $\eta^*\le\eta_{\min}$ the regret can be bounded as

$$\begin{aligned}R_T(\boldsymbol{u}) \le{}& 2kC_{\mathcal{S}}+4kMG_{\max}\Lambda_T(\eta_{\min},2M)+\sqrt{\frac{1}{2}\left[M^2\left(1+16k\Lambda_T(\eta_{\min},2M)\right)+4MP_T\right]\Omega_T}\\ &+(K+4)\epsilon G_{\max}+4(K+4)M^2L_{\max}. \end{aligned}\tag{15}$$

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 (13) is bounded by

$$\begin{aligned}R_T(\boldsymbol{u}) &\le 2kC_{\mathcal{S}}+4kMG_{\max}\Lambda_T(\eta_{\min},2M)+\frac{M^2\left(1+16k\Lambda_T(\eta_{\min},2M)\right)+4MP_T}{\eta^*}+\eta^*\Omega_T\\ &\le 2kC_{\mathcal{S}}+4kMG_{\max}\Lambda_T(\eta_{\min},2M)+3\sqrt{\frac{1}{2}\left[M^2\left(1+16k\Lambda_T(\eta_{\min},2M)\right)+4MP_T\right]\Omega_T}. \end{aligned}\tag{16}$$

Now combining Equations (11), (12) and (14) to (16), we have

$$\begin{aligned}R_T(\boldsymbol{u}) \le{}& 2M\left(G_{\max}+ML_{\max}\right)\log\left(\frac{MT}{\epsilon}\right)\\ &+\left(G_{\max}+L_{\max}\epsilon\right)\left[K(M+P_T)+\epsilon\mathcal{C}_T\right]\\ &+2kC_{\mathcal{S}}+4kMG_{\max}\Lambda_T(\eta_{\min},2M)\\ &+3\sqrt{\frac{1}{2}\left[M^2\left(1+16k\Lambda_T(\eta_{\min},2M)\right)+4MP_T\right]\Omega_T}\\ &+(K+4)\epsilon G_{\max}+4(K+4)M^2L_{\max}\\ &+\frac{KL_{\max}}{2}\left[M^2(1+16k\Lambda_T(\eta_{\min},2M))+4MP_T\right].\end{aligned}$$

From Lemma C.3 we have

$$\begin{aligned}C_{\mathcal{S}} &\le 2\sqrt{T}D_{\min}\left(G_{\max}+\frac{D_{\min}}{\eta_{\max}}\right)\\ &\le \frac{2K\left(\epsilon G_{\max}+\epsilon^2L_{\max}\right)}{\sqrt{T}}\\ \Lambda_T&(\eta_{\min},2M)\\ &\le \log\left(\frac{6\left|\mathcal{S}\right|(2M)^2}{D_{\min}^2}\right)+1\\ &\le \log\left(\frac{24M^2T^2\left(\left\lceil\log_2\left(\frac{TG_{\max}}{\epsilon L_{\max}}\right)\right\rceil+1\right)}{\epsilon^2}\right)+1\\ &\le 2\log\left(\frac{5MT}{\epsilon}\right)+\log\left(\log_2\left(\frac{TG_{\max}}{\epsilon L_{\max}}\right)+2\right)+1\end{aligned}$$

Hence, hiding constants we may write

$$R_T(\boldsymbol{u})\le O\left(G_{\max}((M+\epsilon)\Lambda_T^*+P_T)+L_{\max}\left[(M+\epsilon)^2\Lambda_T^*+(M+\epsilon)P_T\right]+\sqrt{(M^2\Lambda_T^*+MP_T)\Omega_T,}\right).$$

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 ℓ_t are L_t-smooth, then using the self-bounding property we have $\left\|g_t^{(\tau)}\right\|^2 \leq 2L_t\left(\ell_t(w_t^{(\tau)}) - \ell_t^*\right)$ for $\ell_t^* = \min_{w\in W}\ell_t(w)$, and thus

$$\begin{aligned}\Omega_T &= \sum_{t=1}^T KL_t\left[\ell_t(u_t) - \ell_t(w_t^{(\tau)})\right] + 4\sum_{t=1}^T \left\|g_t^{(\tau)}\right\|^2 \\ &\leq \sum_{t=1}^T KL_t\left[\ell_t(u_t) - \ell_t(w_t^{(\tau)})\right] + 8\sum_{t=1}^T L_t\left[\ell_t(w_t^{(\tau)}) - \ell_t^*\right] \\ &\leq \sum_{t=1}^T KL_t\left[\ell_t(u_t) - \ell_t^*\right]\end{aligned}$$

where the second-to-last line chooses $K \geq 8$, and simultaneously we have using Lemma C.4 that

$$\begin{aligned}\Omega_T &\leq 3(K+4)\sum_{t=1}^T\left[G_t^2 + L_t^2 D_j^2\right] \\ &\leq 3(K+4)\sum_{t=1}^T\left[G_t^2 + 4L_t^2M^2\right],\end{aligned}$$

and so we have $\Omega_T \leq O\left(\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t^*\right] \wedge \sum_{t=1}^T G_t^2 + L_t^2M^2\right)$. □

C.2. Proof of Theorem 4.3

Theorem 4.3. *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 $\epsilon > 0$, $K \geq 8$, and for any $i, j \geq 0$ let $D_j = \frac{\epsilon}{\sqrt{T}}\left[2^j \wedge 2^T\right]$ and $\eta_i = \frac{1}{KL_{\max}\sqrt{T}}\left[2^i \wedge \sqrt{T}\right]$, 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 3 guarantees*

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

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

Proof. By Lemma C.5, we can assume that there is a $\tau = (\eta, D) \in \mathcal{S}$ for which $D \geq \max_t\|u_t\| = M$, since otherwise the regret is bounded as

$$R_T(\boldsymbol{u}) \leq 2M\left(G_{\max} + ML_{\max}\right)\log\left(\frac{M\sqrt{T}}{\epsilon}\right). \tag{17}$$

Hence, we can assume there is a $(\eta, D) \in \mathcal{S}$ which has $M \leq D$. For any such $(\eta, D) \in \mathcal{S}$, we can apply Lemma C.2 to get

$$\begin{aligned}R_T(\boldsymbol{u}) \leq\ & 2kC_{\mathcal{S}} + 2kDG_{\max}\Lambda_T(\tau) + \frac{\|u_T\|^2 + 2DP_T + 4kD^2\Lambda_T(\tau)}{2\eta} \\ &+ K\eta\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t(w_t^{(\tau)})\right] + 4\eta\sum_{t=1}^T\left\|g_t^{(\tau)}\right\|^2,\end{aligned}$$

where $g_t^{(\tau)} \in \partial\ell_t(w_t^{(\tau)})$, $P_T = \sum_{t=2}^T \|u_t - u_{t-1}\|$ and

$$C_{\mathcal{S}} \overset{\text{def}}{=} \frac{\sum_{\widetilde{\tau}\in\mathcal{S}} \mu_{\widetilde{\tau}}}{\sum_{\widetilde{\tau}\in\mathcal{S}} \mu_{\widetilde{\tau}}^2}, \qquad \Lambda_T(\tau) \overset{\text{def}}{=} \log\left(\frac{\sum_{\widetilde{\tau}\in\mathcal{S}}\mu^2_{\widetilde{\tau}}}{\mu_\tau^2}\right)+1,$$

where for any $\widetilde{\tau} = (\widetilde{D},\widetilde{\eta}) \in \mathcal{S}$ we define $\mu_{(\widetilde{\eta},\widetilde{D})} = \frac{1}{2\widetilde{D}\left(G_{\max}+\widetilde{D}/\widetilde{\eta}\right)}$. Using the self-bounding property of smooth functions, for any $g_t^{(\tau)} \in \partial\ell_t(w_t^{(\tau)})$ we have $\left\|g_t^{(\tau)}\right\|^2 \leq 2L_t\left[\ell_t(w_t^{(\tau)}) - \ell_t^*\right]$ for $\ell_t^* = \arg\min_{w\in W}\ell_t(w)$, so the last line is bound as

$$K\eta\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t(w_t^{(\tau)})\right] + 8\eta\sum_{t=1}^T L_t\left[\ell_t(w_t^{(\tau)}) - \ell_t^*\right] \leq K\eta\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t^*\right]$$

for $K \geq 8$. Hence,

$$\begin{aligned} R_T(\boldsymbol{u}) &\leq 2kC_{\mathcal{S}} + 2kDG_{\max}\Lambda_T(\tau) + \frac{\|u_T\|^2 + 2DP_T + 4kD^2\Lambda_T(\tau)}{2\eta} \\ &\quad + K\eta\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t^*\right] \end{aligned} \tag{18}$$

Now suppose that $M \leq D_{\min}$, then choosing $\tau = \tau_{\min} = (\eta_{\min}, D_{\min})$ we would have

$$\begin{aligned} R_T(\boldsymbol{u}) &\leq 2kC_{\mathcal{S}} + 2kD_{\min}G_{\max}\Lambda_T(\tau_{\min}) + \frac{\|u_T\|^2 + 2D_{\min}P_T + 4kD_{\min}^2\Lambda_T(\tau_{\min})}{2\eta_{\min}} \\ &\quad + K\eta_{\min}\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t^*\right] \\ &\leq 2kC_{\mathcal{S}} + 2kD_{\min}G_{\max}\Lambda_T(\tau_{\min}) + \frac{D_{\min}}{\eta_{\min}}\frac{1}{2}\left(M + 2P_T + 4kD_{\min}\Lambda_T(\tau_{\min})\right) \\ &\quad + \frac{1}{\sqrt{T}L_{\max}}\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t^*\right] \\ &\leq 2kC_{\mathcal{S}} + 2kG_{\max}\frac{\epsilon\Lambda_T(\tau_{\min})}{\sqrt{T}} + K\epsilon L_{\max}\left(M + P_T + 2k\frac{\epsilon\Lambda_T(\tau_{\min})}{\sqrt{T}}\right) \\ &\quad + \sqrt{\sum_{t=1}^T\left[\ell_t(u_t) - \ell_t^*\right]^2} \end{aligned} \tag{19}$$

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 (18) yields

$$\begin{aligned} R_T(\boldsymbol{u}) &\leq 2kC_{\mathcal{S}} + 4kMG_{\max}\Lambda_T(\eta_{\min}, 2M) + \frac{M^2 + 4MP_T + 16kM^2\Lambda_T(\eta_{\min}, 2M)}{2\eta} \\ &\quad + K\eta\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t^*\right] \end{aligned} \tag{20}$$

where we've observed that

$$\begin{aligned}\Lambda(\eta, D_j) &= \log\left(\frac{\sum_{\widetilde{\tau}\in\mathcal{S}}\mu_{\widetilde{\tau}}^2}{\mu_{(\eta,D_j)}}\right) + 1\\ &= \log\left(\sum_{\widetilde{\tau}\in\mathcal{S}}\mu_{\widetilde{\tau}}^2 D_j^2\left[G_{\max} + D_j/\eta\right]^2\right) + 1\\ &\le \log\left(\sum_{\widetilde{\tau}\in\mathcal{S}}\mu_{\widetilde{\tau}}^2 (2M)^2\left[G_{\max} + 2M/\eta\right]^2\right) + 1\\ &= \Lambda_T(\eta, 2M)\end{aligned}$$

so it remains to show that there is an η that favorably balances the last two terms of Equation (20).

Observe that the optimal choice for η 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 $\eta^* \le \eta_{\min}$ then choosing $\eta = \eta_{\min}$ we have

$$\begin{aligned}R_T(\boldsymbol{u}) &\le 2kC_{\mathcal{S}} + 4kMG_{\max}\Lambda_T(\eta_{\min}, 2M) + \frac{M^2(1 + 16k\Lambda_T(\eta_{\min}, 2M) + 4MP_T}{2\eta^*} + K\eta_{\min}\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t^*\right]\\ &\le 2kC_{\mathcal{S}} + 4kMG_{\max}\Lambda_T(\eta_{\min}, 2M) + \sqrt{\frac{K}{2}\left(M^2(1 + 16k\Lambda_T(\eta_{\min}, 2M)) + 4MP_T\right)\Omega_T}\\ &\quad + \sqrt{\sum_{t=1}^T\left[\ell_t(u_t) - \ell_t^*\right]^2},\end{aligned}\tag{21}$$

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 $K\eta_{\min}\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t^*\right] \le \sqrt{\sum_{t=1}^T\left[\ell_t(u_t) - \ell_t^*\right]^2}$. Likewise, if $\eta^* \ge \eta_{\max}$ then by choosing $\eta = \eta_{\max}$ we have via Equation (20) that

$$\begin{aligned}R_T(\boldsymbol{u}) &\le 2kC_{\mathcal{S}} + 4kMG_{\max}\Lambda_T(\eta_{\min}, 2M) + \frac{M^2 + 4MP_T + 16kM^2\Lambda_T(\eta_{\min}, 2M)}{2\eta_{\max}}\\ &\quad + K\eta^*\sum_{t=1}^T L_t\left[\ell_t(u_t) - \ell_t^*\right]\\ &= 2kC_{\mathcal{S}} + 4kMG_{\max}\Lambda_T(\eta_{\min}, 2M) + KL_{\max}\left(M^2(1 + 8k\Lambda_T(\eta_{\min}, 2M)) + 2MP_T\right)\\ &\quad + \sqrt{\frac{K}{2}\left(M^2(1 + 16k\Lambda_T(\eta_{\min}, 2M)) + 4MP_T\right)\Omega_T}\end{aligned}\tag{22}$$

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 (20) is gives us

$$R_T(\boldsymbol{u}) \le 2kC_{\mathcal{S}} + 4kMG_{\max}\Lambda_T(\eta_{\min}, 2M) + 3\sqrt{\frac{K}{2}\left(M^2(1 + 16k\Lambda_T(\eta_{\min}, 2M)) + 4MP_T\right)\Omega_T}\tag{23}$$

Finally, combining Equations (17), (19) and (21) to (23), we have

$$\begin{aligned}
R_T(\boldsymbol{u}) \leq 2kC_{\mathcal{S}} + 4kG_{\max} &\left[M\Lambda_T(\eta_{\min}, 2M) + \frac{\epsilon\Lambda_T(\tau_{\min})}{\sqrt{T}}\right] \\
&+ 2M\left(G_{\max} + ML_{\max}\right)\log\left(\frac{M\sqrt{T}}{\epsilon}\right) \\
&+ K\epsilon L_{\max}\left(M + P_T + 2k\frac{\epsilon\Lambda_T(\tau_{\min})}{\sqrt{T}}\right) \\
&+ KL_{\max}\left(M^2(1 + 8k\Lambda_T(\eta_{\min}, 2M)) + 2MP_T\right) \\
&+ \sqrt{\sum_{t=1}^{T}\left[\ell_t(u_t) - \ell_t^*\right]^2} \\
&+ 3\sqrt{\frac{K}{2}\left(M^2(1 + 16k\Lambda_T(\eta_{\min}, 2M)) + 4MP_T\right)\Omega_T}
\end{aligned}$$

Lastly, note that by Lemma C.3, 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 $\Lambda_T(\eta_{\min}, D_{\min}) \leq \log\left(6\log_2\left(\lceil\log_2(\sqrt{T})\rceil + 1\right)\right) \leq O\left(\log(\log(\sqrt{T}))\right)$ and $\Lambda_T(\eta_{\min}, 2M) \leq \log\left(\frac{24TM^2\log_2\left(\lceil\log_2(\sqrt{T})\rceil+1\right)}{\epsilon^2}\right) \leq O\left(\log\left(M\sqrt{T}\log\left(\sqrt{T}\right)/\epsilon\right)\right)$. Overall, we have

$$\begin{aligned}
R_T(\boldsymbol{u}) \leq O\Bigg(&G_{\max}(M + \epsilon)\Lambda_T^* \\
&+ L_{\max}(M + \epsilon)^2\Lambda_T^* + L_{\max}(M + \epsilon)P_T \\
&+ \sqrt{\sum_{t=1}^{T}\left[\ell_t(u_t) - \ell_t^*\right]^2} \\
&+ \sqrt{\left(M^2\Lambda_T^* + MP_T\right)\Omega_T}\Bigg)
\end{aligned}$$

where $\Lambda_T^* \leq O\left(\log\left(\frac{M\sqrt{T}\log\left(\sqrt{T}\right)}{\epsilon}\right) + \log\left(\log\left(\sqrt{T}\right)\right)\right) \leq O\left(\log\left(\frac{M\sqrt{T}\log\left(\sqrt{T}\right)}{\epsilon}\right)\right)$. □

C.3. Proof of Theorem 4.2

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 ℓ_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 $P_T = \sum_{t=2}^{T}\|u_t - u_{t-1}\|$ (Zhang et al., 2018).

Theorem 4.2. *For any $M > 0$ there is a sequence of (G, L)-quadratically bounded functions with $\frac{G}{L} \leq M$ such that for any $\gamma \in [0, \frac{1}{2}]$,*

$$R_T(\boldsymbol{u}) \geq \frac{G}{4} M^{1-\gamma} \left[P_T T\right]^{\gamma} + \frac{L}{8} M^{2-\gamma} \left[P_T T\right]^{\gamma}.$$

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

Proof. On each round t, we can always find a u_t such that $u_t \perp w_t$. Let $\|u_t\| := \sigma \leq M$ for some σ to be decided. Let $G > 0$, $L \geq 0$ such that $G/L \leq \sigma$, let $\xi_t = \frac{u_t}{\|u_t\|}$, and on each round set

$$\ell_t(w) = -\frac{1}{2} G \left\langle \xi_t, w \right\rangle + \frac{L}{4} (\sigma - \left\langle \xi_t, w \right\rangle)^2.$$

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 $\widetilde{G}/\widetilde{L} \leq \sigma \leq M$ as required. Since $w_t \perp \xi_t$ and $\langle \xi_t, u_t \rangle = \|u_t\| = \sigma$, we have

$$R_T(\boldsymbol{u}) = \sum_{t=1}^{T} \ell_t(w_t) - \ell_t(u_t) \geq \frac{1}{2} G \sigma T + \frac{L}{4} T \sigma^2.$$

Note also that the path-length of this comparator sequence is bounded as

$$P_T = \sum_{t=2}^{T} \|u_t - u_{t-1}\| \leq 2 \sigma T.$$

Now for $\mu \in [0, 1/2]$ set $\sigma = M T^{-\mu}$, then the path-length is bounded as

$$P_T \leq 2 M T^{1-\mu}$$

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:

$$\begin{aligned}
\frac{L}{4} T^{1-2\mu} M^2 &= \frac{L}{4} (M T^{1-\mu})^{\gamma} (M T^{1-\mu})^{1-\gamma} T^{-\mu} M \\
&\geq \frac{L}{4 \cdot 2^{\gamma}} (P_T)^{\gamma} (M T^{1-\mu})^{1-\gamma} T^{-\mu} M \\
&= \frac{L}{8} M^{2-\gamma} P_T^{\gamma} T^{(1-\mu)(1-\gamma)-\mu} \\
&= \frac{L}{8} M^{2-\gamma} \left[P_T T\right]^{\gamma}
\end{aligned}$$

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,

$$\begin{aligned}
\frac{1}{2} G M T^{1-\mu} &= \frac{1}{2} G (M T^{1-\mu})^{\gamma} (M T^{1-\mu})^{1-\gamma} \geq \frac{1}{2 \cdot 2^{\gamma}} G M^{1-\gamma} (P_T)^{\gamma} T^{(1-\mu)(1-\gamma)} \\
&\geq \frac{1}{4} G M^{1-\gamma} (P_T T)^{\gamma}.
\end{aligned}$$

so

$$R_T(\boldsymbol{u}) \geq \frac{G}{4} M^{1-\gamma} \left[P_T T\right]^{\gamma} + \frac{L}{8} M^{2-\gamma} \left[P_T T\right]^{\gamma}.$$

□

D. Supporting Lemmas

We collect here various miscellaneous supporting lemmas that we use throughout paper. The following lemma is standard but shown here for completeness.

Lemma D.1. *Let $a_1, \ldots, a_T$ be arbitrary non-negative numbers in $\mathbb{R}$. Then*

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

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

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

so summing over t and observing the resulting telescoping sum yields

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

For the lower bound, observe that

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

□

We borrow the following lemma from Jacobsen & Cutkosky (2022).

Lemma D.2. *(Jacobsen & Cutkosky, 2022, Lemma 7) Let $f : \mathbb{R} \to \mathbb{R}$ and let $g : W \to \mathbb{R}$ be defined as $g(x) = f(\|x\|)$. Suppose that $f'(x)$ is concave and non-negative. If f is twice-differentiable at $\|x\|$ and $\|x\| > 0$, then*

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

The following is a simple modification of the stability lemma used in Jacobsen & Cutkosky (2022), reported here with slight modification to handle a leading constant.

Lemma D.3. *For all t, set $\psi_{t+1}(w) = \Psi_{t+1}(\|w\|)$ 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 $c > 0$ and assume that there exists an $\mathring{x}_t > 0$ and G_t-Lipschitz convex function $\eta_t : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ 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 $w_{t+1}, w_t \in W$,*

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

Proof. The proof follows using similar arguments to Jacobsen & 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 ψ is twice differentiable everywhere except the origin, if we define $\widehat{\delta}_t^n = G_t \left\|\widehat{w}_t^n - \widehat{w}_{t+1}^n\right\| - D_{\psi_t}(\widehat{w}_{t+1}^n|\widehat{w}_t^n) - \eta_t(\|\widehat{w}_{t+1}^n\|)G_t^2$, 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}, ψ_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

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

where the last line observes $\psi_t(w) = \Psi_t(\|w\|)$ and uses the regularity assumptions $\Psi_t'''(x) \le 0$, and $\Psi_t'(x) \ge 0$ for $x \ge 0$ to apply Lemma D.2. Hence,

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

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

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

□