Parameter-free Mirror Descent

Andrew Jacobsen AJJACOBS@UALBERTA.CA
University of Alberta

Ashok Cutkosky ASHOK@CUTKOSKY.COM
Boston University

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We develop a modified online mirror descent framework that is suitable for building adaptive and parameter-free algorithms in unbounded domains. We leverage this technique to develop the first unconstrained online linear optimization algorithm achieving an optimal dynamic regret bound, and we further demonstrate that natural strategies based on Follow-the-Regularized-Leader are unable to achieve similar results. We also apply our mirror descent framework to build new parameter-free implicit updates, as well as a simplified and improved unconstrained scale-free algorithm.
Keywords: Online Learning, Dynamic Regret, Parameter-free, Mirror Descent

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 and 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 $\ell_t : W \to \mathbb{R}$. Performance is measured by the *regret* $R_T(u)$, the total loss relative to a benchmark $u \in W$:

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

Almost all of our development is focused on *online linear optimization* (OLO), in which the ℓ_t are assumed to be linear functions, $\ell_t(w) = \langle g_t, w\rangle$. This focus is justified by a well-known reduction (*e.g.* Zinkevich (2003)) which employs the identity $\ell_t(w_t) - \ell_t(u) \le \langle g_t, w_t - u\rangle$ for any $g_t \in \partial\ell_t(w_t)$ to show that OLO algorithms can be used to solve OCO problems. The classical algorithm for this setting is *online gradient descent*, which achieves the minimax optimal regret $R_T(u) \le \|u\|\sqrt{\sum_{t=1}^{T} \|g_t\|^2}$ when the learning rate η is set as $\eta = \frac{\|u\|}{\sqrt{\sum_{t=1}^{T} \|g_t\|^2}}$.

This optimal η is of course unknown *a priori*, and so there has been a concerted push to develop algorithms that achieve similar bounds *without* requiring such oracle tuning (Duchi et al., 2010a; McMahan and Streeter, 2010; Foster et al., 2015; Mhammedi and Koolen, 2020; van der Hoeven, 2019; Cutkosky and Orabona, 2018). A standard result in this setting is

$$R_T(u) \le O\left(\epsilon + \|u\|\sqrt{\sum_{t=1}^{T} \|g_t\|^2 \log(T\|u\|/\epsilon)}\right), \tag{1}$$

which holds for any user-specified ϵ for all u. This is known to be optimal up to constants (Orabona, 2013). Bounds of this form have many names in the literature, such as "comparator adaptive" or "parameter-free". We will use "parameter-free" in the following.

We will develop a new framework for parameter-free regret bounds that is based on *online mirror descent* (OMD). While OMD is already a standard technique in online learning, it has proven difficult to apply it to the unconstrained setting and achieve parameter-free regret. As a consequence of our development, we are able to produce several new kinds of algorithms. First, we consider the *dynamic regret*. In this setting, the benchmark point u is not a fixed value. Instead, we define the regret with respect to an arbitrary *sequence* of benchmarks $\boldsymbol{u} = (u_1, \ldots, u_T)$:

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

Dynamic regret is clearly more demanding than the previous definition (called *static* regret). It is also more appropriate for true streaming settings in which data might drift over time, so that a fixed benchmark u is too weak. Using our approach, for any ϵ we obtain a dynamic regret bound of:

$$R_T(\boldsymbol{u}) \leq \tilde{O}\left(\epsilon + \sqrt{\sum_{t=1}^{T} \|g_t\|^2 \|u_t\| \sum_{i=1}^{T-1} \|u_{i+1} - u_i\| \log(T \max_t \|u_t\|/\epsilon)}\right)$$

This bound holds in both unconstrained and constrained settings. In the unconstrained setting, this is to our knowledge the *first* non-trivial dynamic regret bound of any form. In the constrained setting, this bound still improves upon prior work (*e.g.* Zhang et al. (2018); Jadbabaie et al. (2015); Zhao et al. (2020)) by virtue of increased adaptivity to the individual $\|u_t\|$ values. Moreover, we show that our OMD-based analysis appears to be crucial to this result: essentially no "reasonable" variant of prior methods for unconstrained online learning is capable of achieving a similar result.

We are also able to apply our OMD analysis in two other ways: first, we can essentially immediately produce a general framework for parameter-free regret using *implicit* updates and so move beyond pure OLO. Finally, we show how to use OMD to improve upon the "scale-free" bounds presented by Mhammedi and Koolen (2020) by removing impractical doubling-like restart strategies as well as reducing logarithmic factors in the regret.

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 $a \wedge b = \min\{a, b\}$. The indicator function $\mathcal{I}_W(\cdot)$ is the function such that $\mathcal{I}_W(w) = 0$ if $w \in W$ and $\mathcal{I}_W = \infty$ otherwise. 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. Centered Mirror Descent

In this section we introduce our framework and key technical tools. Our algorithms are constructed from an instance of composite mirror descent (Duchi et al., 2010b) depicted in Algorithm 1. Composite mirror descent can be interpreted as a mirror descent update which adds an auxiliary penalty $\phi_t(w)$ to the loss function $\ell_t(w)$. Typically, $\phi_t(w)$ is a composite loss function which enforces some additional desirable properties of the solution, such as sparsity. In contrast,

Algorithm 1: Centered Mirror Descent

Input *Initial regularizer* $\psi_1 : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$
Initialize $w_1 = \arg\min_w \psi_1(w)$
for $t = 1 : T$ **do**
  Play w_t, observe loss function $\ell_t(\cdot)$
  Choose regularizer ψ_{t+1}, composite penalty φ_t
  Define $\Delta_t(w) = D_{\psi_{t+1}}(w|w_1) - D_{\psi_t}(w|w_1)$ and $\phi_t(w) = \Delta_t(w) + \varphi_t(w)$
  Update $w_{t+1} = \arg\min_w \ell_t(w) + D_{\psi_t}(w|w_t) + \phi_t(w)$
end

we will use these terms $\phi_t(w)$ as a crucial part of our algorithms. This composite term is composed of two parts, $\Delta_t(w)$ and $\varphi_t(w)$, with the distinguishing feature of our approach being the $\Delta_t(w) = D_{\psi_{t+1}}(w|w_1) - D_{\psi_t}(w|w_1)$.

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

In the unconstrained setting, the function $\Delta_t(w)$ is a critical feature of the update. Indeed, Orabona and Pál (2018) showed that adaptive mirror descent algorithms can incur *linear* regret in settings where the divergence $D_{\psi_t}(\cdot|\cdot)$ may be unbounded. The issue is that vanilla mirror descent doesn't properly account for changes in the regularizer ψ_t, allowing the iterates w_t to travel away from their initial position w_1 too quickly; Algorithm 1 fixes this by adding a corrective penalty $\Delta_t(w)$ related to how much ψ_t has changed between rounds. Since this penalty acts to bias the iterates back towards some central reference point w_1, we refer to Algorithm 1 as *Centered Mirror Descent*.

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

Lemma 1 ***(Centered Mirror Descent Lemma)*** *Let $\psi_t(\cdot)$ be an arbitrary sequence of differentiable non-negative convex functions, and assume that $w_1 \in \arg\min_{w\in\mathbb{R}^d} \psi_t(w)$ for all t. Let $\varphi_t(\cdot)$ be an arbitrary sequence of sub-differentiable non-negative convex functions. Then for any $u_1, \ldots, u_T$, Algorithm 1 guarantees*

$$R_T(\boldsymbol{u}) \le \psi_{T+1}(u_T) + \sum_{t=1}^{T} \varphi_t(u_t) + \sum_{t=1}^{T-1} \underbrace{\langle \nabla\psi_{t+1}(w_{t+1}), u_{t+1} - u_t\rangle}_{=:\rho_t}$$
$$+ \sum_{t=1}^{T} \underbrace{\langle g_t, w_t - w_{t+1}\rangle - D_{\psi_t}(w_{t+1}|w_t) - \phi_t(w_{t+1})}_{=:\delta_t}, \quad (2)$$

where $g_t \in \partial\ell_t(w_t)$ and $\phi_t(w) = \Delta_t(w) + \varphi_t(w)$.

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

Lemma 2 ***(Stability Lemma)*** *Let $\Psi_t : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ be a twice differentiable, three-times subdifferentiable function such that $\Psi_t'(x) \ge 0$, $\Psi_t''(x) \ge 0$, and $\Psi_t'''(x) \le 0$ for all $x > 0$. Let $G_t \ge \|g_t\|$ and $\eta_t : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ be a $1/G_t$ Lipschitz convex function, and assume there is an $x_0 \ge 0$ such that $|\Psi_t'''(x)| \le \frac{\eta_t'(x)}{2}\Psi_t''(x)^2$ for all $x > x_0$. Then with $\psi_t(w) = \Psi_t(\|w\|)$, for all w_t, w_{t+1}:*

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

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

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

This example demonstrates the purpose of φ_t in the update. When $\Delta_t(w_{t+1}) \ge \eta_t(\|w_{t+1}\|)\|g_t\|^2$, we already obtain $\delta_t \le \widehat{\delta}_t$. However, this identity may be false (as in the previous example) or difficult to prove.[1] In such cases, we include a small additional φ_t term to easily ensure the desired

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

bounds. In fact, this strategy can be viewed as generalizing a certain "correction" term which appears in the experts literature (e.g. Steinhardt and Liang (2014); Chen et al. (2021)), but to our knowledge is not typically employed in the general online linear optimization setting.

3. Parameter-free Learning

As a warm-up, we first use our framework to construct a parameter-free algorithm which achieves the optimal static regret (Equation (1)).

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

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

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

The full proof can be found in Appendix C, along with an efficient closed-form update formula. It follows the intuition developed in the previous section: Lemma 1 implies $R_T(u) \le \psi_{T+1}(u) + \sum_{t=1}^T \delta_t$. Then, we show that ψ_t satisfies the conditions of Lemma 2 while the growth rate $\Delta_t(w)$ ensures that $\delta_t \le \widehat{\delta}_t$, so that $\delta_t \le \widehat{\delta}_t \le O\left(\frac{2\|g_t\|^2}{\Psi''_t(\alpha_t)}\right) \le O(\frac{\alpha_t \|g_t\|^2}{\sqrt{V_t}})$. Finally, we choose α_t small enough to ensure $\sum_{t=1}^T \delta_t \le O(1)$.

Treating $\log(\log)$ terms as effectively constant, the bound in Theorem 1 achieves the "ideal" dependence on V_T in the logarithmic factors. Indeed, given oracle access to $\|u\|$ and V_T, we could set $\epsilon = O\left(\frac{\|u\|\sqrt{V_T}}{G}\right)$, causing all the log terms to disappear from the bound and leaving only $R_T(u) \le \widehat{O}\big(\|u\|\sqrt{V_{T+1}}\big)$, which matches the optimal rate vanilla gradient descent would achieve with oracle tuning up to a $\log(\log)$ factor.

4. Dynamic Regret

Now that we've had a taste of how to use our techniques, we turn to the challenging problem of competing with an arbitrary *sequence* of comparators u_t. To appreciate why this is difficult using existing techniques, let us recall the *reward-regret duality*, a key player in most analyses of parameter-free algorithms. Suppose that we wish to guarantee static regret of $R_T(u) = \sum_{t=1}^T \langle g_t, w_t - u\rangle \le B_T(u)$ for all $u \in \mathbb{R}^d$ for some function B_T. Because this holds for *all* $u \in \mathbb{R}^d$, we must have:

$$\sum_{t=1}^T \langle g_t, w_t\rangle + \sup_{u\in\mathbb{R}^d} \langle -g_{1:T}, u\rangle - B_T(u) = \sum_{t=1}^T \langle g_t, w_t\rangle + B_T^*(-g_{1:t}) \le 0,$$

where $B_T^*(\cdot)$ denotes the Fenchel conjugate of $B_T(\cdot)$. Rearranging, we find that guaranteeing $R_T(u) \le B_T(u)$ for all $u \in \mathbb{R}^d$ is equivalent to guaranteeing $\sum_{t=1}^T \langle -g_t, w_t\rangle \ge B_T^*(-g_{1:T})$. This latter expression has no dependence on the *unknown* comparator u, making it appealing from an algorithm design perspective. Hence, many prior works revolve around designing algorithms which ensure that $\sum_{t=1}^{T}\langle -g_t, w_t\rangle \geq B_T^*(-g_{1:T})$ for some B_T^* of interest.

However, notice that the assumption of a *fixed* comparator $u \in \mathbb{R}^d$ was vital for the argument above to work. It is unclear what the analogue of this argument should be for dynamic regret, where we instead have a *sequence* of comparators. In fact, in the next section, we show that most algorithms designed using this duality cannot guarantee sublinear dynamic regret. Then, in Section 4.2, we will remedy this issue via our mirror-descent framework.

4.1. Lower Bounds

In this section, we show that common algorithm design strategies for the unconstrained setting cannot achieve optimal dynamic regret. Specifically, we consider 1-D OLO algorithms such that:

1. The algorithm sets $w_{t+1} = F_t(g_{1:t}, \|g_1\|, \ldots, \|g_t\|)$ for some functions $F_1, \ldots, F_t$. We will frequently elide the dependence on $\|g_t\|$ to write $F_t(g_{1:t})$.
2. F_t satisfies $\text{sign}(F_t(g_{1:t})) = -\text{sign}(g_{1:t})$.
3. F_t is odd: $F_t(g_{1:t}) = -F_t(-g_{1:t})$.
4. $F_t(-x)$ is non-decreasing for positive x.

Notice that the vast majority of parameter-free FTRL algorithms satisfy these properties with $F_t = \nabla\psi_t^\star$. First, we consider the *constrained setting*. Here, it is actually relatively easy to show that no algorithm satisfying these conditions can obtain low dynamic regret:

Theorem 2 *Suppose an algorithm $\mathcal{A}$ satisfies the conditions [1, 2, 3, 4] for a 1-d OLO game with domain $[-1, 1]$. Then for all even T there exists a sequence of costs $g_1, \ldots, g_T$ with $\|g_t\| = 1$ for all t and a comparator sequence $u_1, \ldots, u_T$ such that the path length $P = \sum_{t=1}^{T-1}\|u_t - u_{t+1}\| \leq 2$, but the regret is at least $T/2$.*

Proof Set $g_t = 1$ for $t \leq T/2$ and $g_t = -1$ otherwise. Notice that $g_{1:t} \geq 0$ for all t, so that $-1 \leq w_t \leq 0$ for all t. Thus $\sum_{t=1}^{T} w_t g_t \geq \sum_{t=1}^{T/2} w_t g_t \geq -T/2$. Let $u_t = -1$ for $t \leq T/2$ and $u_t = 1$ otherwise. Then clearly $P = 2$, and $\sum_{t=1}^{T} g_t u_t = -T$ for a total regret of at least $T/2$. ■

The essential idea behind this result is that when presented with a sequence of alternating runs of $+1$ and -1, we have $g_{1:t} \geq 0$ for all t so that $w_t \leq 0$ for all t. Intuitively, this means that we cannot compete with the competitor when $g_t = -1$ and $u_t \geq 0$.

To formally establish bounds in the unconstrained setting is a bit more complicated. In this case, we must guard against the possibility that the algorithm experiences some significant *negative regret* during the periods in which g_t has the opposite sign of w_t. This is not an issue in the constrained setting because the algorithm clearly cannot have less than -1 loss on any given round. In the unconstrained setting however, the loss is in principle unbounded. Thus, we consider the maximum possible value of $\text{Regret}_T(0)$ as a measure of "complexity" of the algorithm. The following Lemma 3 provides an constraint on the algorithm in terms of this complexity. Then, by carefully tuning our adversarial sequence to this complexity measure, we are able to guarantee poor dynamic regret. Proofs of the following two results can be found in Appendix D.

Lemma 3 *Suppose an algorithm $\mathcal{A}$ satisfying the conditions [1, 2, 3, 4] also guarantees that $\sum_{i=1}^{t} g_i w_i \le \epsilon$ for all t for some ϵ if $\|g_i\| = 1$ for all i. Then there is a universal constant C (not depending on $\mathcal{A}$) such that for any sequence $g_1, \ldots, g_t$ satisfying $\|g_i\| = 1$ for all $i \le t$ and $\|g_{1:t}\| \le C\sqrt{t}/2$, we have $|w_{t+1}| \le \frac{2\epsilon}{C^2}$.*

Using this Lemma, we can show our lower bound:

Theorem 3 *Let C be the universal constant from Lemma 3. Suppose an algorithm $\mathcal{A}$ satisfying the conditions [1, 2, 3, 4] also guarantees $\sum_{i=1}^{t} g_i w_i \le \epsilon$ for all t for some ϵ if $\|g_i\| = 1$ for all i. Then there is a universal $\widetilde{T}$ such that for all $T \ge \widetilde{T}$, there exists a sequence $\{g_t\}$ and a comparator sequence $\{u_t\}$ that does not depend on $\mathcal{A}$ such that:*

$$P_T + \max_t |u_t| \in \left[\frac{\epsilon\sqrt{2T}}{C^3}, \frac{8\epsilon\sqrt{T}}{C^3}\right]$$

$$\sum_{t=1}^{T} g_t(w_t - u_t) \ge \frac{C}{16}(P_T + \max_t |u_t|)\sqrt{T}$$

Theorem 3 essentially shows that all FTRL-based algorithms we are aware of cannot achieve a dynamic regret better than $O(P_T\sqrt{T})$. It should be noted that there are a few algorithms (e.g. the ONS-based betting algorithm of Cutkosky and Orabona (2018)) that are not captured by the conditions imposed on the Algorithm in Theorem 3. However, we believe all such previously known exceptions satisfy the constraint that $w_t \le 0$ for all t when presented with alternating signs of gradients as in our lower bound constructions. We therefore hypothesize that they also fail to achieve optimal dynamic regret, but we leave open establishing formal bounds.

Finally, we stress that the adversarial sequence of Theorem 3 is *algorithm independent*. Thus, one likely cannot hope to "combine" several such suboptimal algorithms into an optimal algorithm as is done in the constrained setting: all the suboptimal algorithms could be *simultaneously bad*.

4.2. Dynamic Regret Algorithm

The results in the previous section suggest that any algorithm which updates using an FTRL-like update of the form $w_{t+1} = F_t(g_{1:t})$ will be unable to guarantee dynamic regret with a sublinear dependence on the path-length P_T due to an excessive resistance to changing direction. In what follows, we'll show that our additional φ_t term can mitigate this issue by biasing the iterates ever-so-slightly back towards the origin. This facilitates a more rapid change of sign when the losses change direction, and enables us to avoid the pathogenic behavior observed in the previous section.

Our approach is as follows. Using the tools developed in Section 2, we'll first derive an algorithm which, for any $\eta \le \frac{1}{G}$, guarantees $R_T(u) \le \widetilde{O}\left(\frac{P_T + \max_t \|u_t\|}{\eta} + \eta \sum_{t=1}^{T} \|g_t\|^2 \|u_t\|\right)$. Note that such a bound is out-of-reach for algorithms covered by Theorem 3. Further, for all η we have $R_T(\mathbf{0}) = O(1)$. Hence, following Cutkosky (2019b), we run an instance of this algorithm $\mathcal{A}_\eta$ for each η in some set $\mathcal{S} = \left\{\eta \in \mathbb{R} : 0 < \eta \le \frac{1}{G}\right\}$, and on each round we play $w_t = \sum_{\eta \in \mathcal{S}} w_t^\eta$ where w_t^η is the output of $\mathcal{A}_\eta$. Then for any arbitrary $\widetilde{\eta} \in \mathcal{S}$, we can write $\langle g_t, w_t - u_t \rangle = \left\langle g_t, w_t^{\widetilde{\eta}} - u_t \right\rangle + \sum_{\eta \ne \widetilde{\eta}} \langle g_t, w_t^\eta \rangle$, so the regret is bounded as

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

Algorithm 2: Dynamic Regret Algorithm

Input *Lipschitz bound* G*, value* $\varepsilon > 0$*, step-sizes* $\mathcal{S} = \left\{ \frac{2^k}{G\sqrt{T}} \wedge \frac{1}{G} : 1 \leq k \leq \lceil \log_2 \sqrt{T} \rceil \right\}$

Initialize *Initialize* $\epsilon = \frac{\varepsilon}{|\mathcal{S}|} = \frac{\varepsilon}{\lceil \log_2(\sqrt{T}) \rceil}$, $V_1 = 4G^2$, $w_1^\eta = \mathbf{0}$, $\widetilde{\theta}_t^\eta = \mathbf{0}$ *for each* $\eta \in \mathcal{S}$

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

  Play $w_t = \sum_{\eta \in \mathcal{S}} w_t^\eta$, receive subgradient g_t

  Update $V_{t+1} = V_t + \|g_t\|^2$ and $\alpha_{t+1} = \frac{\epsilon G^2}{V_{t+1} \log^2(V_{t+1}/G^2)}$

  for $\eta \in \mathcal{S}$ **do**

    Set $\theta_t^\eta = \frac{2 w_t^\eta \log(\|w_t^\eta\|/\alpha_t + 1)}{\eta \|w_t^\eta\|} - g_t$ (with $\theta_t^\eta = -g_t$ if $w_t^\eta = \mathbf{0}$)

    Update $w_{t+1}^\eta = \frac{\alpha_{t+1} \theta_t^\eta}{\|\theta_t^\eta\|} \left[\exp\left[\frac{\eta}{2} \max(\|\theta_t^\eta\| - 2\eta \|g_t\|^2, 0) \right] - 1 \right]$

  end

end

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

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

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

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

The proof can be found in Appendix D.3, and again follows the intuition in Section 2: we first apply Lemma 1 to get $R_T(\boldsymbol{u}) \leq \psi_{T+1}(u_T) + \sum_{t=1}^{T-1} \rho_t + \sum_{t=1}^T \varphi_t(u_t) + \sum_{t=1}^T \delta_t$. Unlike in Section 3, $\Delta_t(w)$ is generally not large enough to ensure that $\delta_t \leq \widehat{\delta}_t$. Instead, we include an additional composite regularizer φ_t in the update and show that this now ensures $\delta_t \leq \widehat{\delta}_t$, so that by Lemma 2 we have $\delta_t \leq \widehat{\delta}_t \leq \frac{2\|g_t\|^2}{\Psi_t''(0)} \leq 2\eta\alpha_t \|g_t\|^2$. Then we choose α_t to be small enough to ensure that $\sum_{t=1}^T \delta_t \leq O(1)$. We also now need to control the additional terms associated with the time-varying comparator, $\rho_t = \langle \nabla \psi_{t+1}(w_{t+1}), u_t - u_{t+1} \rangle$. To handle these, we again exploit φ_t: by increasing it slightly more, we can decrease δ_t enough to cancel out the additional ρ_t.

With this result in hand, we proceed to "tune" the optimal step-size by simply adding the iterates of a collection of these simple learners $\mathcal{A}_\eta$, as discussed above. The full algorithm is given in Algorithm 2, and the overall regret guarantee is given in Theorem 4 (with proof in Appendix D.4).

Theorem 4 *For any $\varepsilon > 0$ and $u_1, \ldots, u_T$ in $\mathbb{R}^d$, Algorithm 2 guarantees*

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

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

The bound achieved by Algorithm 2 is the first non-vacuous dynamic regret guarantee of any kind that we are aware of in unbounded domains. Further, Theorem 4 exhibits a stronger *per-comparator adaptivity* than previously obtained by depending on the individual comparators $\|u_t\|$, in contrast to the $R_T(\boldsymbol{u}) \leq \widetilde{O}\Big(\sqrt{(M^2 + MP_T)\sum_{t=1}^{T} \|g_t\|^2}\Big)$ rate attained by prior works in bounded domains (Zhang et al., 2018; Jadbabaie et al., 2015).

To see why this per-comparator adaptivity is interesting, let us consider a learning scenario in which there is a nominal "default" decision $\overline{u}$ which we expect to perform well *most* of the time, but may perform poorly during certain rare/unpredictable events. One example of such a situation is when one has access to an batch of data collected *offline*, which we can leverage to fit a parameterized model $\mathcal{M}(\overline{u})$ to the data to use as a baseline predictor. Deploying such a model online can be dangerous in practice because there may be certain events that are poorly covered by our dataset, leading to unpredictable behavior from the model. In this context, we can think of $\overline{u}$ as the learned model parameters, and without loss of generality we can assume $\overline{u} = \mathbf{0}$ (since otherwise we could just translate the decision space to be centered at $\overline{u}$). In this context, Theorem 4 tells us that Algorithm 2 will accumulate *no regret* over any intervals where we would want to compare performance against the baseline model, and over any intervals $[a, b]$ where the model is a poor comparison we are still guaranteed to accumulate no more than a $\widetilde{O}\Big(\sqrt{(M^2 + MP_{[a,b]})\|g\|_{a:b}^2}\Big)$ penalty, where $P_{[a,b]} = \sum_{t=a+1}^{b} \|u_t - u_{t-1}\|$ is the path-length of any other arbitrary sequence of comparators over the interval $[a, b]$.

The property in the preceeding discussion is similar to the notion of *strong adaptivity* in the *constrained* setting, in which an algorithm guarantees the optimal static regret over all sub-intervals of $[1, T]$ *simultaneously* (Jun et al., 2017; Daniely et al., 2015). One might wonder if instead we should hope for the natural analog in the unconstrained setting: $R_{[a,b]}(u) = \sum_{t=a}^{b}\langle g_t, w_t - u\rangle \leq \tilde{O}(\|u\|\sqrt{b-a})$ for all $[a, b]$. Unfortunately, this natural analog is likely unattainable. To see why, notice that for all intervals $[a, b]$ of some fixed length $\tau = b - a$, we would require $R_{[a,b]}(\mathbf{0}) = \sum_{t=a}^{b} \langle g_t, w_t\rangle \leq O(1)$, suggesting that no w_t can be larger than some fixed constant (dependent on τ). Yet clearly for large enough T this can't be guaranteed while simultaneously guaranteeing $R_T(u) \leq O\big(\|u\| G\sqrt{T \log(\|u\| GT)}\big)$ for all $u \in \mathbb{R}^d$, since via reward-regret duality this entails competing against a fixed comparator $u \in \mathbb{R}^d$ with $\|u\| = O\big(\exp(T)/T\big)$ in the worst-case, which can get arbitrarily large as T increases. For this reason, we consider Theorem 4 to be a suitable relaxation of the strongly adaptive guarantee for unbounded domains.

Interestingly, if one is willing to forego adaptivity to the individual $\|u_t\|$, we show in Appendix I that achieving the weaker $R_T(\boldsymbol{u}) \leq \widetilde{O}\Big(\sqrt{(M^2 + MP_T)\|g\|_{1:T}^2}\Big)$ can be attained in unbounded domains using the one-dimensional reduction of Cutkosky and Orabona (2018, Algorithm 2).

Algorithm 3: Implicit-Optimistic Centered Mirror Descent

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

5. Adapting to Gradient Variability

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

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

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

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

Our algorithm is derived as a special case of the algorithm shown in Algorithm 3, which can be understood as an instance of centered mirror descent with an additional *optimistic* step on each round. The optimistic step leverages an arbitrary guess $\widehat{\ell}_{t+1}(\cdot)$ about what the next loss function will be. Intuitively, if the learner could deduce the trajectory of the loss functions, they'd be able to "think ahead" and play a point w_{t+1} for which the next loss $\ell_{t+1}(\cdot)$ is minimized. The following theorem provides an algorithm which guarantees $R_T(u) \leq \widetilde{O}\Big(\|u\| \sqrt{\sum_{t=1}^T \|\nabla \ell_t(w_t) - \nabla \widehat{\ell}_t(w_t)\|^2}\Big)$ using an arbitrary sequence of optimistic guesses $\widehat{\ell}_t(\cdot)$.

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

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

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

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

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

6. Lipschitz Adaptivity and Scale-free Learning

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

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

While this hint-based strategy can be used to mitigate the problem of unknown Lipschitz constant G, a truly ideal algorithm would be *scale-free*. That is, the algorithm's outputs w_t are invariant to any constant rescaling of the gradients $g_t \mapsto cg_t$ for all t. Scale-free regret bounds scale with the maximal subgradient *encountered* $G_T = \max_{t\le T}\|g_t\|$, while non-scale free bounds typically depend on some user-specified estimate of G_T and may perform much worse if this estimate is very poor[2]. Mhammedi and Koolen (2020) used the approach proposed by Cutkosky (2019a) to develop FreeGrad, the first parameter-free and scale-free algorithm.

Unfortunately, FreeGrad suffers from an analytical difficulty called the *range-ratio* problem. Briefly, the range-ratio problem occurs when h_T/h_1 (called the range-ratio) is very large: in principle if we set $h_1 = \|g_1\|$, then this quantity could grow arbitrarily large, and so even logarithmic dependencies can make the regret bound vacuous. In order to circumvent this difficulty, Mhammedi and Koolen (2020) utilize a doubling-based scheme, restarting FreeGrad whenever a particular technical condition was met. While such restart strategies only lose a constant factor in the regret in theory, they are unsatisfying: scale-free updates are motivated by potential practical performance benefits, yet any algorithm which is forced to restart from scratch several times during deployment is unlikely to achieve high performance in practice. The following theorem, proven in Appendix G, characterizes a new base algorithm that, when combined with the reduction of Cutkosky (2019a), generates a scale-free algorithm which avoids the range-ratio problem without resorting to restarts. Our approach employs a simple analysis which follows easily using the tools developed in Section 2, enabling us to achieve tighter logarithmic factors than FreeGrad.

Theorem 6 *Let $\ell_1,\ldots,\ell_T$ be G-Lipschitz convex functions and $g_t \in \partial\ell_t(w_t)$ for all t. Let $h_1 \le \ldots \le h_T$ be a sequence of hints such that $h_t \ge \|g_t\|$, and assume that h_t is provided at the start of each round t. Set $\psi_t(w) = 3\int_0^{\|w\|}\min_{\eta\le\frac{1}{h_t}}\left[\frac{\log(x/\alpha_t+1)}{\eta}+\eta V_t\right]dx$ where $V_t = 4h_t^2 + \|g\|_{1:t-1}^2$, $\alpha_t = \frac{\epsilon}{\sqrt{B_t}\log^2(B_t)}$, $B_t = 4\sum_{s=1}^t\left(4+\sum_{s'=1}^{s-1}\frac{\|g_{s'}\|^2}{h_{s'}^2}\right)$, and $\epsilon > 0$. Then for all $u\in\mathbb{R}^d$, Algorithm 1 guarantees*

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

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

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

The terms B_t are carefully chosen to address the range-ratio problem in an *online* fashion: we show that $\sqrt{B_t}$ upper bounds the quantity h_t/h_{τ_t}, where starting from $\tau_1 = 1$, the variable τ_t roughly tracks the most-recent round t where the ratio $h_t/h_{\tau_{t-1}}$ exceeds a threshold analogous to the one used by FreeGrad to trigger restarts. That is, B_t enacts a kind of "soft restarting" by shrinking w_t according to the restarting threshold, just as setting a learning rate of $1/\sqrt{t}$ in online gradient

2. Typical parameter-free algorithms rely on an *upper bound* for G_T, while Cutkosky (2019a) implicitly relies on a *lower bound*.

descent can be viewed as a "soft restart" in contrast to the standard doubling trick. It is quite possible that FreeGrad could be similarly modified to avoid restarts by incorporating B_t directly, but this is difficult to verify due to highly non-trivial polynomial expressions appearing in the analysis.

We include the pseudocode for the complete scale-free algorithm — consisting of the gradient clipping and artificial constraints reductions of Cutkosky (2019a) applied with the algorithm characterized in Theorem 6 — and corresponding regret guarantee in Appendix F.

7. Conclusion

In this work, we developed a specialization of the standard mirror descent framework that is particularly suitable for building parameter-free algorithms. Although we focus our discussion on the unconstrained setting, we emphasize that our techniques apply equally well in the constrained setting, either by adding an indicator function to our regularizers ψ or φ, or via the unconstrained-to-constrained conversion proposed by Cutkosky and Orabona (2018). The mirror descent formulation allows us to obtain optimal dynamic regret bounds, implicit updates, and streamlined scale-free algorithms. We nevertheless leave important open questions. For example, observe that our dynamic regret algorithm requires $O(d\log(T))$ time and space: can this be improved? As some partial progress, in Appendix J we show that one can maintain the $O\left(\sqrt{(M^2 + MP_T)\,\|g\|_{1:T}^2}\right)$ bound up to poly-logarithmic factors using only $O(d)$ *amortized* computation. Further, we focused on the use of φ_t to achieve novel bounds, and did not explore its more traditional use in incorporating a user-provided composite objective. We look forward to exciting developments in this area.

Acknowledgments

This work was completed in part while AJ was visiting Boston University. AJ gratefully acknowledges support from the NSERC Alexander Graham Bell Scholarship and Alberta Innovates Data-enabled Innovation Scholarhip.

References

Hilal Asi and John C. Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, (3), 2019. ISSN 1095-7189. doi: 10.1137/18m1230323.

Nicolò Campolongo and Francesco Orabona. Temporal variability in implicit online learning. In *Advances in Neural Information Processing Systems*, 2020.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Keyi Chen, Ashok Cutkosky, and Francesco Orabona. Implicit parameter-free online learning with truncated linear models. In *Algorithmic Learning Theory*, 2022.

Liyu Chen, Haipeng Luo, and Chen-Yu Wei. Impossible tuning made possible: A new expert algorithm and its applications. In *Conference on Learning Theory*, 2021.

Ashok Cutkosky. Artificial constraints and hints for unbounded online learning. In *Conference on Learning Theory*, 2019a.

Ashok Cutkosky. Combining online learning guarantees. In *Conference on Learning Theory*, 2019b.

Ashok Cutkosky and Kwabena Boahen. Online learning without prior information. In *Conference on Learning Theory*, pages 643–677. PMLR, 2017.

Ashok Cutkosky and Francesco Orabona. Black-box reductions for parameter-free online learning in banach spaces. In *Conference On Learning Theory*, 2018.

Amit Daniely, Alon Gonen, and Shai Shalev-Shwartz. Strongly adaptive online learning. In *International Conference on Machine Learning*, 2015.

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

John C Duchi, Shai Shalev-Shwartz, Yoram Singer, and Ambuj Tewari. Composite objective mirror descent. In *Conference on Learning Theory*. Citeseer, 2010b.

Huang Fang, Nick Harvey, Victor Portella, and Michael Friedlander. Online mirror descent and dual averaging: keeping pace in the dynamic case. In *International Conference on Machine Learning*, 2020.

Dylan J Foster, Alexander Rakhlin, and Karthik Sridharan. Adaptive online learning. In *Advances in Neural Information Processing Systems 28*. 2015.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2016. ISSN 2167-3888. doi: 10.1561/2400000013.

Ali Jadbabaie, Alexander Rakhlin, Shahin Shahrampour, and Karthik Sridharan. Online Optimization : Competing with Dynamic Comparators. In *International Conference on Artificial Intelligence and Statistics*, 2015.

Kwang-Sung Jun, Francesco Orabona, Stephen Wright, and Rebecca Willett. Improved Strongly Adaptive Online Learning using Coin Betting. In *Conference on Artificial Intelligence and Statistics*, 2017.

H. Brendan McMahan. A survey of algorithms and analysis for adaptive online learning. *Journal of Machine Learning Research*, 18(90):1–50, 2017.

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

Zakaria Mhammedi and Wouter M. Koolen. Lipschitz and comparator-norm adaptivity in online learning. In *Conference on Learning Theory*, 2020.

Francesco Orabona. Dimension-free exponentiated gradient. In *Advances in Neural Information Processing Systems*, 2013.

Francesco Orabona. A modern introduction to online learning. *CoRR*, abs/1912.13213, 2019. URL http://arxiv.org/abs/1912.13213.

Francesco Orabona and Dávid Pál. Parameter-free stochastic optimization of variationally coherent functions. 2021.

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

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

J. Steinhardt and P. Liang. Adaptivity and optimism: An improved exponentiated gradient algorithm. In *International Conference on Machine Learning*, 2014.

Dirk van der Hoeven. User-specified local differential privacy in unconstrained adaptive online learning. In *Advances in Neural Information Processing Systems*, 2019.

Lijun Zhang, Shiyin Lu, and Zhi-Hua Zhou. Adaptive online learning in dynamic environments. In *Neural Information Processing Systems*, 2018.

Peng Zhao, Yu-Jie Zhang, Lijun Zhang, and Zhi-Hua Zhou. Dynamic regret of convex and smooth functions. In *Advances in Neural Information Processing Systems*, 2020.

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

Appendix A. A Strong Mirror Descent Lemma

In this section we derive a regret template for Centered Mirror Descent which holds for arbitrary sequences of loss functions and choices of ψ_t and φ_t. The result is analogous to the Strong FTRL Lemma of McMahan (2017), but applies to a sequence of comparators and is tailored to mirror descent-style analysis.

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

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

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

Lemma 4 ***(Strong Centered Mirror Descent Lemma)*** *For all t, let $\ell_t(\cdot)$ be a subdifferentiable function, $\varphi_t(\cdot)$ be a subdifferentiable non-negative function, and $\psi_t(\cdot)$ be a differentiable non-negative function. Define $\Delta_t(w) = D_{\psi_{t+1}}(w|w_1) - D_{\psi_t}(w|w_1)$, $\phi_t(w) = \Delta_t(w) + \varphi_t(w)$, and set $w_{t+1} = \arg\min_{w \in \mathbb{R}^d} \ell_t(w) + D_{\psi_t}(w|w_t) + \phi_t(w)$.*

Then, for all t there is some $\nabla \ell_t(w_{t+1}) \in \partial \ell_t(w_{t+1})$ and $\nabla \varphi_t(w_{t+1}) \in \partial \varphi_t(w_{t+1})$ such that $\nabla \ell_t(w_{t+1}) + \nabla \psi_{t+1}(w_{t+1}) - \nabla \psi_t(w_t) + \nabla \varphi_t(w_{t+1}) = \mathbf{0}$, and for any $u_1, \ldots, u_T$ in $\mathbb{R}^d$,

$$\begin{aligned}
\sum_{t=1}^T \ell_t(w_t) - \ell_t(u_t) = D_{\psi_{T+1}}(u_T|w_1) - D_{\psi_{T+1}}(u_T|w_{T+1}) + \sum_{t=1}^T \varphi_t(u_t) \\
+ \sum_{t=2}^T \underbrace{\langle \nabla \psi_t(w_t) - \nabla \psi_t(w_1), u_{t-1} - u_t \rangle}_{=:\mathcal{P}_t} \\
+ \sum_{t=1}^T \underbrace{\langle g_t, w_t - w_{t+1} \rangle - D_{\psi_t}(w_{t+1}|w_t) - \phi_t(w_{t+1})}_{=:\delta_t} \\
+ \sum_{t=1}^T \underbrace{-\widehat{D}_{\ell_t + \varphi_t}\Big(u_t, w_{t+1}, \nabla \ell_t(w_{t+1}) + \nabla \varphi_t(w_{t+1})\Big) - \widehat{D}_{\ell_t}\Big(w_{t+1}, w_t, g_t\Big)}_{=:\mathcal{L}_t} .
\end{aligned}$$

where g_t is an arbitrary element of $\partial \ell_t(w_t)$.

The bound includes three terms not included in Lemma 1: $-D_{\psi_T}(u_T|w_{T+1})$, $-\widehat{D}_{\ell_t}(w_{t+1}, w_t, g_t)$, and $-\widehat{D}_{\ell_t + \varphi_t}(u_t, w_{t+1}, \nabla \ell_t(w_{t+1}) + \nabla \varphi_t(w_{t+1}))$. Observe that the lemma holds even for non-convex losses; in this case we'll need to account for the fact that the terms $-\widehat{D}_{\ell_t}(u_t, w_{t+1}, \nabla \ell_t(w_{t+1}))$ and $-\widehat{D}_{\ell_t}(w_{t+1}, w_t, g_t)$ may be positive and may require additional effort to control. When the losses are convex the terms $-\widehat{D}_{\ell_t}(u_t, w_{t+1}, \nabla \ell_t(w_{t+1}))$ and $-\widehat{D}_{\ell_t}(w_{t+1}, w_t, g_t)$ can often be leveraged in useful ways, particularly when the ℓ_t have nice properties such as strong convexity. In this work we only assume convexity of ℓ_t and drop these terms. Similarly, for simplicity we assume that φ_t is convex so that we can bound $-\widehat{D}_{\varphi_t}(u_t, w_{t+1}, \nabla \varphi_t(w_{t+1})) \leq 0$. It's possible that this term could also be leveraged in some useful way, but we do not investigate this in the current work.

Proof of Lemma 4

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

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

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

Now, we define $\widetilde{\ell}_t(w) = \ell_t(w) + \phi_t(w)$, and begin by writing

$$\begin{aligned}\sum_{t=1}^T \ell_t(w_t) - \ell_t(u_t) &= \sum_{t=1}^T \ell_t(w_{t+1}) - \ell_t(u_t) + \sum_{t=1}^T \ell_t(w_t) - \ell_t(w_{t+1}) \\ &= \sum_{t=1}^T \widetilde{\ell}_t(w_{t+1}) - \widetilde{\ell}_t(u) + \phi_t(u_t) + \sum_{t=1}^T \ell_t(w_t) - \ell_t(w_{t+1}) - \phi_t(w_{t+1}), \quad (3)\end{aligned}$$

where the last line adds and subtracts $\phi_t(w_{t+1})$ and $\phi_t(u_t)$. For the rest of this proof, we define $\tilde{g}_t = \nabla \ell_t(w_{t+1}) + \nabla \varphi_t(w_{t+1}) + \nabla \Delta_t(w_{t+1})$. Thus, $\widetilde{g}_t + \nabla \psi_t(w_{t+1}) - \nabla \psi_t(w_t) = \mathbf{0}$. Further, with $\widetilde{\ell}_t(w) = \ell_t(w) + \phi_t(w)$, we have $\tilde{g}_t \in \partial \widetilde{\ell}_t(w_{t+1})$. Now observe that we can write

$$\begin{aligned}\sum_{t=1}^T \widetilde{\ell}_t(w_{t+1}) - \widetilde{\ell}_t(u_t) &= \sum_{t=1}^T \langle \widetilde{g}_t, w_{t+1} - u_t \rangle + \langle \widetilde{g}_t, u_t - w_{t+1} \rangle + \widetilde{\ell}_t(w_{t+1}) - \widetilde{\ell}_t(u_t) \\ &= \sum_{t=1}^T \langle \widetilde{g}_t, w_{t+1} - u_t \rangle - \widehat{D}_{\widetilde{\ell}_t}(u_t, w_{t+1}, \widetilde{g}_t) \\ &= \sum_{t=1}^T \langle \nabla \psi_t(w_t) - \nabla \psi_t(w_{t+1}), w_{t+1} - u_t \rangle - \widehat{D}_{\widetilde{\ell}_t}(u_t, w_{t+1}, \widetilde{g}_t) \\ &= \sum_{t=1}^T D_{\psi_t}(u_t|w_t) - D_{\psi_t}(u_t|w_{t+1}) - D_{\psi_t}(w_{t+1}|w_t) - \widehat{D}_{\widetilde{\ell}_t}(u_t, w_{t+1}, \widetilde{g}_t)\end{aligned}$$

where the last line uses the well-known three-point relation $\langle \nabla f(a) - \nabla f(b), b - c \rangle = D_f(c|a) - D_f(c|b) - D_f(b|a)$. Hence, recalling that $\phi_t(w_t) = \Delta_t(w_t) + \varphi_t(w)$ and observing that $D_{\Delta_t}(u_t|w_{t+1}) =$

$D_{\psi_{t+1}-\psi_t}(u_t|w_{t+1}) = D_{\psi_{t+1}}(u_t|w_{t+1}) - D_{\psi_t}(u_t|w_{t+1})$ leaves us with

$$\begin{aligned}
\sum_{t=1}^T \widetilde{\ell}_t(w_{t+1}) - \widetilde{\ell}_t(u_t) &= \sum_{t=1}^T D_{\psi_t}(u_t|w_t) - D_{\psi_t}(u_t|w_{t+1}) - D_{\psi_t}(w_{t+1}|w_t) \\
&\quad - \widehat{D}_{\ell_t+\varphi_t+\Delta_t}(u_t, w_{t+1}, \nabla\ell_t(w_{t+1}) + \nabla\varphi_t(w_{t+1}) + \nabla\Delta_t(w_{t+1})) \\
&\overset{(*)}{=} \sum_{t=1}^T D_{\psi_t}(u_t|w_t) - D_{\psi_t}(u_t|w_{t+1}) - D_{\psi_t}(w_{t+1}|w_t) \\
&\quad - D_{\Delta_t}(u_t|w_{t+1}) - \widehat{D}_{\ell_t+\varphi_t}(u_t, w_{t+1}, \nabla\ell_t(w_{t+1}) + \nabla\varphi_t(w_{t+1})) \\
&= \sum_{t=1}^T D_{\psi_t}(u_t|w_t) - D_{\psi_{t+1}}(u_t|w_{t+1}) - D_{\psi_t}(w_{t+1}|w_t) \\
&\quad - \widehat{D}_{\ell_t+\varphi_t}(u_t, w_{t+1}, \nabla\ell_t(w_{t+1}) + \nabla\varphi_t(w_{t+1})) \\
&= \underbrace{D_{\psi_1}(u_1|w_1) - D_{\psi_{T+1}}(u_T|w_{T+1}) + \sum_{t=2}^T D_{\psi_t}(u_t|w_t) - D_{\psi_t}(u_{t-1}|w_t)}_{\text{Ⓐ}} \\
&\quad + \sum_{t=1}^T -D_{\psi_t}(w_{t+1}|w_t) - \widehat{D}_{\ell_t+\varphi_t}(w_t, w_{t+1}, \nabla\ell_t(w_{t+1}) + \nabla\varphi_t(w_{t+1})),
\end{aligned}$$

where $(*)$ uses the fact that $\widehat{D}_{\Delta_t}(x, y, g_y) = D_{\Delta_t}(x|y)$ since $\Delta_t(x)$ is a differentiable function. Returning to the full regret bound we have

$$\begin{aligned}
\sum_{t=1}^T \ell_t(w_t) - \ell_t(u_t) &= \text{Ⓐ} + \sum_{t=1}^T \phi_t(u_t) + \sum_{t=1}^T \ell_t(w_t) - \ell_t(w_{t+1}) - D_{\psi_t}(w_{t+1}|w_t) - \phi_t(w_{t+1}) \\
&\quad + \sum_{t=1}^T -\widehat{D}_{\ell_t+\varphi_t}(u_t, w_{t+1}, \nabla\ell_t(w_{t+1}) + \nabla\varphi_t(w_{t+1})) \\
&= \text{Ⓐ} + \sum_{t=1}^T [\Delta_t(u_t) + \varphi_t(u_t)] + \sum_{t=1}^T \underbrace{\langle g_t, w_t - w_{t+1}\rangle - D_{\psi_t}(w_{t+1}|w_t) - \phi_t(w_{t+1})}_{=:\delta_t} \\
&\quad + \sum_{t=1}^T \underbrace{-\widehat{D}_{\ell_t+\varphi_t}(u_t, w_{t+1}, \nabla\ell_t(w_{t+1}) + \nabla\varphi_t(w_{t+1})) - \widehat{D}_{\ell_t}(w_{t+1}, w_t, g_t)}_{=:\mathcal{L}_t},
\end{aligned}$$

where the last line lets $g_t \in \partial\ell_t(w_t)$ and observes $\ell_t(w_t) - \ell_t(w_{t+1}) = \langle g_t, w_t - w_{t+1}\rangle - \langle g_t, w_t - w_{t+1}\rangle + \ell_t(w_t) - \ell_t(w_{t+1}) = \langle g_t, w_t - w_{t+1}\rangle - \widehat{D}_{\ell_t}(w_{t+1}, w_t, g_t)$. Finally, observe that $\sum_{t=1}^T \Delta_t(u_t) =$

$D_{\psi_{T+1}}(u_T|w_1) - D_{\psi_1}(w_1|u_1) + \sum_{t=2}^{T} D_{\psi_t}(u_{t-1}|w_1) - D_{\psi_t}(u_t|w_1)$, so combining with Ⓐ yields

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

and hence

$$\begin{aligned}
\sum_{t=1}^{T} \ell_t(w_t) - \ell_t(u_t) &= \text{Ⓐ} + \sum_{t=1}^{T} \Delta_t(u_t) + \sum_{t=1}^{T} \varphi_t(u_t) + \sum_{t=1}^{T} \delta_t + \sum_{t=1}^{T} \mathcal{L}_t \\
&= D_{\psi_{T+1}}(u_T|w_1) - D_{\psi_{T+1}}(u_T|w_{T+1}) + \sum_{t=2}^{T} \mathcal{P}_t + \sum_{t=1}^{T} \varphi_t(u_t) + \sum_{t=1}^{T} \delta_t + \sum_{t=1}^{T} \mathcal{L}_t.
\end{aligned}$$

■

Appendix B. Proofs for Section 2 (Centered Mirror Descent)

B.1. Proof of Lemma 1

Lemma 1 ***(Centered Mirror Descent Lemma)*** *Let $\psi_t(\cdot)$ be an arbitrary sequence of differentiable non-negative convex functions, and assume that $w_1 \in \arg\min_{w\in\mathbb{R}^d} \psi_t(w)$ for all t. Let $\varphi_t(\cdot)$ be an arbitrary sequence of sub-differentiable non-negative convex functions. Then for any $u_1, \ldots, u_T$, Algorithm 1 guarantees*

$$R_T(\boldsymbol{u}) \leq \psi_{T+1}(u_T) + \sum_{t=1}^{T} \varphi_t(u_t) + \sum_{t=1}^{T-1} \underbrace{\langle \nabla\psi_{t+1}(w_{t+1}), u_{t+1} - u_t\rangle}_{=:\rho_t}$$
$$+ \sum_{t=1}^{T} \underbrace{\langle g_t, w_t - w_{t+1}\rangle - D_{\psi_t}(w_{t+1}|w_t) - \phi_t(w_{t+1})}_{=:\delta_t}, \tag{2}$$

where $g_t \in \partial\ell_t(w_t)$ and $\phi_t(w) = \Delta_t(w) + \varphi_t(w)$.

Proof From Lemma 4 we have that

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

where

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

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

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

The stated bound then follows by re-indexing $\sum_{t=2}^T \langle \nabla\psi_t(w_t), u_{t-1} - u_t\rangle$. ■

B.2. Proof of Lemma 2

Lemma 2 ***(Stability Lemma)*** *Let $\Psi_t : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be a twice differentiable, three-times subdifferentiable function such that $\Psi_t'(x) \geq 0$, $\Psi_t''(x) \geq 0$, and $\Psi_t'''(x) \leq 0$ for all $x > 0$. Let $G_t \geq \|g_t\|$ and $\eta_t : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be a $1/G_t$ Lipschitz convex function, and assume there is an $x_0 \geq 0$ such that $|\Psi_t'''(x)| \leq \frac{\eta_t'(x)}{2}\Psi_t''(x)^2$ for all $x > x_0$. Then with $\psi_t(w) = \Psi_t(\|w\|)$, for all w_t, w_{t+1}:*

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

Proof First, consider the case that the origin is contained in the line segment connecting w_t and w_{t+1}. Then, there exists sequences $\widehat{w}_t^1, \widehat{w}_t^2 \dots$ and $\widehat{w}_{t+1}^1, \widehat{w}_{t+1}^2 \dots$ 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 $\widehat{\delta}_t^n = \langle g_t, \widehat{w}_t^n - \widehat{w}_{t+1}^n\rangle - D_{\psi_t}(\widehat{w}_{t+1}^n|\widehat{w}_t^n) - \eta_t(\|\widehat{w}_{t+1}^n\|)\|g_t\|^2$, 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 $\widehat{\delta}_t \overset{\text{def}}{=} \langle g_t, w_t - w_{t+1}\rangle - D_{\psi_t}(w_{t+1}|w_t) - \eta_t(\|w_{t+1}\|)\|g_t\|^2$. 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

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

where the last line observes $\psi_t(w) = \Psi_t(\|w\|)$ and uses the assumptions that $\Psi_t'(x) \geq 0$ and $\Psi_t'''(x) \leq 0$ (hence Ψ_t' is concave) to apply Lemma 7. Thus,

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

where (a) uses convexity of $\eta_t(x)$, (b) uses the Lipschitz assumption $\eta_t'(\|\widetilde{w}\|) \leq 1/G_t \leq \frac{1}{\|g_t\|}$ and triangle inequality, and (c) uses Cauchy-Schwarz inequality and the fact that $\|\widetilde{w} - w_{t+1}\| \leq \|w_t - w_{t+1}\|$ for any $\widetilde{w}$ on the line connecting w_{t+1} and w_t. Next, by assumption we know that there is an $x_0 \geq 0$ such that $|\Psi_t'''(x)| \leq \frac{\eta_t'(x)}{2}\Psi_t''(x)^2$. If it happens that $\|\widetilde{w}\| \leq x_0$, then

$$\widehat{\delta}_t \leq \frac{2\|g_t\|^2}{\Psi_t''(\|\widetilde{w}\|)} - \eta_t(\|\widetilde{w}\|)\|g_t\|^2 \leq \frac{2\|g_t\|^2}{\Psi_t''(x_0)},$$

which follows from the fact that $\Psi_t'''(x) \leq 0$ implies that $\Psi_t''(x)$ is non-increasing in x, and hence $\Psi_t''(\|\widetilde{w}\|) \geq \Psi_t''(x_0)$ whenever $\|\widetilde{w}\| \leq x_0$. Otherwise, when $\|\widetilde{w}\| > x_0$ we have by assumption that $\frac{|\Psi_t'''(x)|}{\Psi_t''(x)^2} = \frac{-\Psi_t'''(x)}{\Psi_t''(x)^2} = \frac{d}{dx}\frac{1}{\Psi_t''(x)} \leq \frac{\eta_t'(x)}{2}$ for any $x > x_0$, so integrating from x_0 to $\|\widetilde{w}\|$ we have

$$\begin{aligned}
&\frac{1}{\Psi_t''(\|\widetilde{w}\|)} - \frac{1}{\Psi_t''(x_0)} \leq \frac{1}{2}\int_{x_0}^{x} \eta_t'(x)dx \\
&\qquad \implies \frac{1}{\Psi_t''(\|\widetilde{w}\|)} \leq \frac{1}{\Psi_t''(x_0)} + \frac{1}{2}\int_{x_0}^{x} \eta_t'(x)dx \leq \frac{1}{\Psi_t''(x_0)} + \frac{1}{2}\int_{0}^{x} \eta_t'(x)dx = \frac{1}{\Psi_t''(x_0)} + \frac{\eta_t(x)}{2},
\end{aligned}$$

so

$$\widehat{\delta}_t \leq \frac{2\|g_t\|^2}{\Psi_t''(\|\widetilde{w}\|)} - \eta_t(\|\widetilde{w}\|)\|g_t\|^2 \leq 2\|g_t\|^2\left(\frac{1}{\Psi_t''(x_0)} + \frac{\eta_t(\|\widetilde{w}\|)}{2}\right) - \eta_t(\|\widetilde{w}\|)\|g_t\|^2 = \frac{2\|g_t\|^2}{\Psi_t''(x_0)}.$$

Thus, in either case we have $\widehat{\delta}_t \leq \frac{2\|g_t\|^2}{\Psi_t''(x_0)}$.

■

Appendix C. Proofs for Section 3 (Parameter-free Learning)

C.1. Proof of Theorem 1

The theorem is restated below. Pseudocode for the algorithm characterized by this theorem is given in Algorithm 4 for convenience.

Algorithm 4: Parameter-free Learning via Centered Mirror Descent

Input *Lipschitz bound G, Value $\epsilon > 0$*
Initialize $V_1 = 4G^2$, $w_1 = \mathbf{0}$, $\theta_1 = \mathbf{0}$
for $t = 1 : T$ **do**
  Play w_t, receive subgradient g_t
  Set $\theta_{t+1} = \theta_t - g_t$, $V_{t+1} = V_t + \|g_t\|^2$, $\alpha_{t+1} = \frac{\epsilon G}{\sqrt{V_{t+1}}\log^2(V_{t+1}/G^2)}$, and define

$$f_{t+1}(\theta) = \begin{cases} \frac{\|\theta\|^2}{36V_{t+1}} & \text{if } \|\theta\| \leq \frac{6V_{t+1}}{G} \\ \frac{\|\theta\|}{3G} - \frac{V_{t+1}}{G^2} & \text{otherwise} \end{cases}$$

  Update $w_{t+1} = \frac{\alpha_{t+1}\theta_{t+1}}{\|\theta_{t+1}\|}\left[\exp\left(f_{t+1}(\theta_{t+1})\right) - 1\right]$
end

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

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

Then for all $u \in \mathbb{R}^d$, Algorithm 4 guarantees

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

Proof First, let us derive the update formula, which can be seen in Algorithm 4. By first-order optimality conditions for $w_{t+1} = \arg\min_{w \in \mathbb{R}^d} \langle g_t, w\rangle + D_{\psi_t}(w|w_t) + \Delta_t(w)$ we have:

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

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

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

and unrolling the recursion we have

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

Inspecting the equation for ψ_{t+1} then yields:

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

where we define the function

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

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

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

Now we see that one of two cases occurs: either

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

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

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

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

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

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

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

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

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

Putting these cases together yields the update described in Algorithm 4 (with $k = 3$, which is important later in the regret analysis).

Now, we concentrate on proving the regret bound.

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

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

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

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

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

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

Thus, we need only bound the stability terms $\sum_{t=1}^T \delta_t$, which we will handle using the Stability Lemma (Lemma 2).

For any $x > 0$, we have

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

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

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

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

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

where the last line defines $\eta_t(x) = \int_0^x \min\left\{ \sqrt{\frac{F_t(v)}{V_t}}, \frac{1}{G} \right\} dv$ and uses $k \ge 3$. We also have $\eta_t'(x) = \min\left\{ \sqrt{\frac{F_t(x)}{V_t}}, \frac{1}{G} \right\} \le \frac{1}{G}$, and $\eta_t'(x)$ is monotonic, so $\eta_t(x)$ is convex and $1/G$ Lipschitz. Hence, by Lemma 2 we have

$$\widehat{\delta}_t = \langle g_t, w_t - w_{t+1} \rangle - D_{\psi_t}(w_{t+1}|w_t) - \eta_t(\|w_{t+1}\|) \|g_t\|^2 \le \frac{2\|g_t\|^2}{\Psi_t''(x_0)} \tag{4}$$

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

Next, we want to show that $\phi_t(w) = \Delta_t(w) \ge \eta_t(\|w\|) \|g_t\|^2$, so that $\delta_t \le \widehat{\delta}_t$. To this end, let $x > 0$ and observe that for $\alpha_{t+1} \le \alpha_t$, we have $F_{t+1}(x) = \log(x/\alpha_{t+1} + 1) \ge \log(x/\alpha_t + 1) =$

$F_t(x)$, so

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

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

$$\begin{aligned}&\geq k \|g_t\|^2 \min\left\{\sqrt{\frac{F_t(x)}{V_{t+1}}}, \frac{1}{G}\right\} + k \min_{\eta \leq \frac{1}{G}} \left[\frac{F_t(x)}{\eta} + \eta V_t\right] - k \min_{\eta \leq \frac{1}{G}} \left[\frac{F_t(x)}{\eta} + \eta V_t\right] \\ &= k \|g_t\|^2 \min\left\{\sqrt{\frac{F_t(x)}{V_{t+1}}}, \frac{1}{G}\right\} \geq \frac{k}{\sqrt{2}} \|g_t\|^2 \min\left\{\sqrt{\frac{F_t(x)}{V_t}}, \frac{1}{G}\right\} \geq \|g_t\|^2 \eta'_t(x),\end{aligned}$$

where the last line uses $k \geq 3$ and $\frac{1}{V_t} = \frac{1}{V_{t+1}} \frac{V_{t+1}}{V_t} = \frac{1}{V_{t+1}} \left(1 + \|g_t\|^2 / V_t\right) \leq \frac{2}{V_{t+1}}$ for $V_t \geq \|g_t\|^2$. From this, we immediately have

$$\Delta_t(w) = \int_0^{\|w\|} \Psi'_{t+1}(x) - \Psi'_t(x) dx \geq \|g_t\|^2 \int_0^{\|w\|} \eta'_t(x) dx = \eta_t(\|w\|) \|g_t\|^2,$$

and hence

$$\begin{aligned}\delta_t &= \langle g_t, w_t - w_{t+1}\rangle - D_{\psi_t}(w_{t+1}|w_t) - \Delta_t(w_{t+1}) \\ &\leq \langle g_t, w_t - w_{t+1}\rangle - D_{\psi_t}(w_{t+1}|w_t) - \eta_t(\|w_{t+1}\|) \|g_t\|^2 = \widehat{\delta}_t \leq \frac{2\|g_t\|^2}{\Psi''_t(x_0)}\end{aligned}$$

for $x_0 = \alpha_t(e-1)$ via Equation (4). Summing over t then yields

$$\begin{aligned}\sum_{t=1}^T \delta_t &\leq \sum_{t=1}^T \frac{2\|g_t\|^2}{\Psi''_t(\alpha_t(e-1))} \leq \sum_{t=1}^T \frac{2e\alpha_t}{k} \|g_t\|^2 \sqrt{\frac{F_t(\alpha_t(e-1))}{V_t}} \\ &\leq \sum_{t=1}^T \frac{2e\alpha_t}{k} \|g_t\|^2 \frac{1}{\sqrt{V_t}} \leq \sum_{t=1}^T \frac{6}{k} \frac{\alpha_t \|g_t\|^2}{\sqrt{V_t}} \\ &\overset{(a)}{\leq} 2G\epsilon \sum_{t=1}^T \frac{\|g_t\|^2}{V_t \log^2(V_t/G^2)} \\ &\overset{(b)}{\leq} 4G\epsilon\end{aligned}$$

where (a) chooses $\alpha_t = \frac{\epsilon G}{\sqrt{V_t}\log^2(V_t/G^2)}$ and recalls $k \geq 3$, and (b) recalls $V_t = 4G^2 + \|g\|_{1:t-1}^2$ and uses Lemma 9 to bound $\sum_{t=1}^T \frac{\|g_t\|^2}{V_t \log^2(V_t/G^2)} \leq 2$. Returning to our regret bound we have

$$\begin{aligned}
R_T(u) &\leq 2k\,\|u\| \max\left\{\sqrt{V_{T+1}\log\left(\|u\|/\alpha_{T+1}+1\right)}, G\log\left(\|u\|/\alpha_{T+1}+1\right)\right\} + \sum_{t=1}^T \delta_t \\
&\leq 4G\epsilon + 2k\,\|u\| \max\left\{\sqrt{V_{T+1}\log\left(\|u\|/\alpha_{T+1}+1\right)}, G\log\left(\|u\|/\alpha_{T+1}+1\right)\right\} \\
&\leq \widehat{O}\left(G\epsilon + \|u\|\left[\sqrt{\|g\|_{1:T}^2 \log\left(\frac{\|u\|\sqrt{\|g\|_{1:T}^2}}{\epsilon G}+1\right)} \vee G\log\left(\frac{\|u\|\sqrt{\|g\|_{1:T}^2}}{\epsilon G}+1\right)\right]\right)
\end{aligned}$$

■

Appendix D. Proofs for Section 4 (Dynamic Regret)

D.1. Proof of Lemma 3

Lemma 3 *Suppose an algorithm $\mathcal{A}$ satisfying the conditions [1, 2, 3, 4] also guarantees that $\sum_{i=1}^t g_i w_i \leq \epsilon$ for all t for some ϵ if $\|g_i\| = 1$ for all i. Then there is a universal constant C (not depending on $\mathcal{A}$) such that for any sequence $g_1, \dots, g_t$ satisfying $\|g_i\| = 1$ for all $i \leq t$ and $\|g_{1:t}\| \leq C\sqrt{t}/2$, we have $|w_{t+1}| \leq \frac{2\epsilon}{C^2}$.*

Proof Notice that if $\|g_t\| = 1$ for all t, since F_t is odd, we can write

$$F_t(g_{1:t-1}, \|g_1\|, \dots, \|g_t\|) = -\text{sign}(g_{1:t-1})F_t(-|g_{1:t-1}|, 1, 1, \dots, 1)$$

which justifies our simpler notation of dropping the dependence on $\|g_t\|$ from F_t in this setting. Further, since $F_t(-x)$ is non-decreasing for positive x, it suffices to show that $F_t(-C\sqrt{t}/2) \leq \frac{2\epsilon}{C^2}$.

Let $g_1, \dots, g_{t-1}$ all be independent random signs, and let $g_t = \text{sign}(F_t(g_{1:t-1}))$. Notice that since F_t is odd, we can write $F_t(g_{1:t-1}) = -\text{sign}(g_{1:t-1})F_t(-|g_{1:t-1}|)$. Further, notice that $|g_{1:t-1}|$ satisfies $\mathbb{E}[|g_{1:t-1}|] \geq C\sqrt{t}$ for some absolute constant C, and $\mathbb{E}[|g_{1:t-1}|^2] \leq t$. Thus, by Paley-Zygmund inequality, $P[|g_{1:t-1}| > \theta C\sqrt{t}] \geq (1-\theta)^2 C^2$ for all θ. In particular, setting $\theta = 1/2$ yields $P[|g_{1:t-1}| > C/2\sqrt{t}] \geq C^2/4$

Then we have:

$$\epsilon \geq \mathbb{E}\left[\sum_{i=1}^t w_i g_i\right]$$

using $\mathbb{E}[g_i] = 0$ for $i < t$:

$$= \mathbb{E}[w_t g_t]$$

Using $g_t w_t = |F_t(g_{1:t-1})| = F(-|g_{1:t-1}|)$:

$$\begin{aligned}
&= \mathbb{E}\left[F_t(-|g_{1:t-1}|)\right] \\
&\geq \sum_{S\in\mathbb{N}} F_t(-S)P(|g_{1:t-1}| = S)
\end{aligned}$$

using $F_t(-|x|) \geq 0$ for all x:

$$\geq 2 \sum_{S \geq C/2\sqrt{t}} F_t(-S)P(|g_{1:t-1}| = S)$$
$$\geq 2F_t(-C\sqrt{t}/2)P(|g_{1:t-1}| \geq C\sqrt{t}/2)$$
$$\geq \frac{C^2}{2}F_t(-C\sqrt{t}/2)$$

rearranging:

$$\frac{2\epsilon}{C^2} \geq F_t(-C\sqrt{t}/2)$$

■

D.2. Proof of Theorem 3

Theorem 3 *Let C be the universal constant from Lemma 3. Suppose an algorithm $\mathcal{A}$ satisfying the conditions [1, 2, 3, 4] also guarantees $\sum_{i=1}^{t} g_t w_t \leq \epsilon$ for all t for some ϵ if $\|g_t\| = 1$ for all t. Then, for all T large enough that:*

1. $C\sqrt{T/2} \geq 2$
2. $T - \lfloor C\sqrt{T/2}/2 \rfloor \geq T/2$
3. $2 + \frac{2}{C}\sqrt{2T} \leq \frac{4}{C}\sqrt{T}$

there exists a sequence $\{g_t\}$ and a comparator sequence $\{u_t\}$ that does not depend on $\mathcal{A}$ such that such that:

$$\frac{\epsilon\sqrt{2T}}{C^3} \leq P_T + \max_t |u_t| \leq \frac{8\epsilon\sqrt{T}}{C^3}$$
$$\sum_{t=1}^{T} g_t(w_t - u_t) \geq \frac{T\epsilon}{2C^2}$$
$$\geq \frac{C}{16}(P_T + \max_t |u_t|)\sqrt{T}$$

Proof Consider the sequence $g_t = (-1)^{t+1}$ for $t \in [1, \lceil T/2 \rceil]$, and afterwards for all natural numbers k and j with $j < \lfloor C\sqrt{T/2}/2 \rfloor$: $g_{\lceil T/2 \rceil + k\lfloor C\sqrt{T/2}/2 \rfloor + j} = (-1)^k$. Notice that $|g_t| = 1$ for all t, and also, for $t \leq \lceil T/2 \rceil - 1$, we have $g_{1:t-1} = 0$ for all odd $t \leq T/2 - 1$, so that $w_t = 0$ for all odd t. Next, for even $t \leq \lceil T/2 \rceil - 1$, $g_{1:t-1} = 1$ so that $w_t \leq 0$ and so $w_t g_t = -w_t \geq 0$. Therefore $\sum_{t=1}^{\lceil T/2 \rceil - 1} w_t g_t \geq 0$.

Now, let's consider the indices $t \geq \lceil T/2 \rceil$. Observe that our sequence satisfies $g_{1:t} \geq 0$ for all t so that $w_t \leq 0$ for all t. Further, $|g_{1:t}| \leq \lfloor C\sqrt{T/2}/2 \rfloor \leq C\sqrt{t}/2$ for all $t \geq \lceil T/2 \rceil$. Thus, by Lemma 3, we have that $-\frac{2\epsilon}{C^2} \leq w_t$ for all t. Now, let S^+ be the indices t such that $t = \lceil T/2 \rceil + k\lfloor C\sqrt{T/2}/2 \rfloor + j$ for even k and $j \in [0, \lfloor C\sqrt{T/2}/2 \rfloor]$, and let S^- be the indices with odd k. Then for $t \in S^+$, $g_t = 1$ and for $t \in S^-$, $g_t = -1$. Further, $|S^+| \le |S^-| + \lfloor C\sqrt{T/2}/2 \rfloor$. Thus, since $|S^+| + |S^-| = T - \lceil T/2 \rceil + 1 \ge T/2$, we have $|S^+| \ge \frac{T - \lfloor C\sqrt{T/2}/2 \rfloor}{2}$.

Putting all these observations together:

$$\sum_{t=\lceil T/2 \rceil}^{T} g_t w_t \ge \sum_{t\in S^+} -\frac{2\epsilon}{C^2} + \sum_{t \in S^-} -w_t$$

Using the fact that $w_t \in [-2\epsilon/C^2, 0]$:

$$\begin{aligned}
&\ge -\frac{2\epsilon}{C^2}|S^+| \\
&\ge -\frac{2\epsilon}{C^2} \cdot \frac{T - \lfloor C\sqrt{T/2}/2 \rfloor}{2} \\
&\ge -\frac{T\epsilon}{2C^2}
\end{aligned}$$

Next, consider the comparator sequence $u_t = 0$ for $t \le \lceil T/2 \rceil - 1$, and $u_t = -2\epsilon/C^2$ for $t \in S^+$ and $u_t = 2\epsilon/C^2$ for $t \in S^-$. Notice that the path length is

$$\begin{aligned}
P_T + \max_t |u_t| &= \frac{2\epsilon}{C^2} + \sum_{t=1}^{T-1} |u_t - u_{t+1}| \\
&\le \frac{2\epsilon}{C^2}\left(1 + \left\lceil \frac{\lceil T/2 \rceil + 1}{2\lfloor C\sqrt{T/2}/2 \rfloor} \right\rceil\right) \\
&\le \frac{2\epsilon}{C^2}\left(2 + \frac{T/2}{\lfloor C\sqrt{T/2}/2 \rfloor}\right) \\
&\le \frac{2\epsilon}{C^2}\left(2 + \frac{T}{C\sqrt{T/2} - 1}\right) \\
&\le \frac{2\epsilon}{C^2}\left(2 + \frac{2T}{C\sqrt{T/2}}\right) \\
&\le \frac{2\epsilon}{C^2}\left(2 + \frac{2}{C}\sqrt{2T}\right) \\
&\le \frac{8\epsilon\sqrt{T}}{C^3}
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
P_T + \max_t |u_t| &\ge \frac{2\epsilon}{C^2}\left(1 + \left\lfloor \frac{T/2}{2\lfloor C\sqrt{T/2}/2 \rfloor} \right\rfloor\right) \\
&\ge \frac{2\epsilon}{C^2}\left(\frac{T}{2C\sqrt{T/2}}\right) \\
&\ge \frac{\epsilon\sqrt{2T}}{C^3}
\end{aligned}$$

Further,

$$\sum_{t=1}^{T} g_t u_t = -\frac{2(T - \lceil T/2 \rceil + 1)\epsilon}{C^2} \leq -\frac{T\epsilon}{C^2}$$

Thus, overall we obtain dynamic regret:

$$\sum_{t=1}^{T} g_t(w_t - u_t) \geq \frac{T\epsilon}{2C^2}$$

Substituting the bound on $P_T + \max_t |u_t|$ completes the argument. ■

D.3. Proof of Proposition 1

We break the proof of Proposition 1 into parts; we first derive a partial result in Proposition 2, and then make particular choices for the unspecified parameters α_t and b_t.

Proposition 2 *$(\alpha_t)_{t=1}^{T}$ be a non-increasing sequence and consider Algorithm 1 with*

$$\psi_t(w) = 2\int_0^{\|w\|} \frac{\log(x/\alpha_t + 1)}{\eta} dx$$
$$\varphi_t(w) = \left(\eta \|g_t\|^2 + b_t\right) \|w\|,$$

where $b_t \geq 0$ and $\eta \leq \frac{1}{G}$. Then for all $u_1, \ldots, u_T$ in $\mathbb{R}^d$, Algorithm 1 guarantees

$$R_T(\boldsymbol{u}) \leq \frac{2M \log(M/\alpha_{T+1} + 1)}{\eta} + \sum_{t=1}^{T-1} \left[\frac{2\|u_{t+1} - u_t\| \log(\|w_{t+1}\|/\alpha_{t+1} + 1)}{\eta} - b_t \|w_{t+1}\|\right]$$
$$+ \sum_{t=1}^{T} \left(\eta \|g_t\|^2 + b_t\right) \|u_t\| + \eta \sum_{t=1}^{T} \alpha_t \|g_t\|^2,$$

where $M = \max_t \|u_t\|$.

Proof Using Lemma 1 we have

$$\begin{aligned}
R_T(\boldsymbol{u}) &\leq \psi_{T+1}(u_T) + \sum_{t=1}^{T-1} \rho_t + \sum_{t=1}^{T} \varphi_t(u_t) + \sum_{t=1}^{T} \delta_t \\
&\leq \frac{2\|u_T\| \log(\|u_T\|/\alpha_{T+1} + 1)}{\eta} + \sum_{t=1}^{T-1} \rho_t + \sum_{t=1}^{T} \varphi_t(u_t) + \sum_{t=1}^{T} \delta_t \\
&\leq \frac{2M \log(M/\alpha_{T+1} + 1)}{\eta} + \sum_{t=1}^{T-1} \rho_t + \sum_{t=1}^{T} \varphi_t(u_t) + \sum_{t=1}^{T} \delta_t
\end{aligned}$$

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

$$\begin{aligned}
\sum_{t=1}^{T-1} \rho_t &= \sum_{t=1}^{T-1} \langle \nabla \psi_{t+1}(w_{t+1}), u_t - u_{t+1} \rangle \le \sum_{t=1}^{T-1} \|\nabla \psi_{t+1}(w_{t+1})\| \|u_t - u_{t+1}\| \\
&= 2 \sum_{t=1}^{T-1} \frac{\log\left(\|w_{t+1}\| / \alpha_{t+1} + 1\right)}{\eta} \|u_t - u_{t+1}\| \\
\sum_{t=1}^{T} \delta_t &= \sum_{t=1}^{T} \langle g_t, w_t - w_{t+1} \rangle - D_{\psi_t}(w_{t+1}|w_t) - \phi_t(w_{t+1}) \\
&= \sum_{t=1}^{T} \langle g_t, w_t - w_{t+1} \rangle - D_{\psi_t}(w_{t+1}|w_t) - \Delta_t(w_{t+1}) - \varphi_t(w_{t+1})
\end{aligned}$$

First consider the terms $\sum_{t=1}^{T} \delta_t$. Since $(\alpha_t)_{t=1}^{T}$ is a non-increasing sequence, we have $\Delta_t(w_{t+1}) = \psi_{t+1}(w_{t+1}) - \psi_t(w_{t+1}) \ge 0$ and

$$\begin{aligned}
\delta_t &= \langle g_t, w_t - w_{t+1} \rangle - D_{\psi_t}(w_{t+1}|w_t) - \Delta_t(w_{t+1}) - \varphi_t(w_{t+1}) \\
&\le \langle g_t, w_t - w_{t+1} \rangle - D_{\psi_t}(w_{t+1}|w_t) - \varphi_t(w_{t+1}) \\
&= \langle g_t, w_t - w_{t+1} \rangle - D_{\psi_t}(w_{t+1}|w_t) - \eta \|g_t\|^2 \|w_{t+1}\| - b_t \|w_{t+1}\| .
\end{aligned}$$

We proceed by showing that the regularizers $\psi_t(\cdot)$ satisfy the conditions of Lemma 2. we have $\psi_t(w) = \Psi_t(\|w\|) = 2\int_0^{\|w\|} \frac{\log(x/\alpha_t+1)}{\eta} dx$ and

$$\Psi_t'(x) = 2\frac{\log\left(x/\alpha_t + 1\right)}{\eta}, \qquad \Psi_t''(x) = \frac{2}{\eta\left(x + \alpha_t\right)}, \qquad \Psi_t'''(x) = \frac{-2}{\eta\left(x + \alpha_t\right)^2},$$

so $\Psi_t(x) \ge 0$, $\Psi_t'(x) \ge 0$, $\Psi_t''(x) \ge 0$, and $\Psi_t'''(x) \le 0$ for all $x > 0$. Moreover,

$$\frac{-\Psi_t'''(x)}{\Psi_t''(x)^2} = \frac{2}{\eta(x+\alpha_t)^2} \frac{\eta^2(x+\alpha_t)^2}{2^2} = \frac{\eta}{2},$$

so assuming $\eta \le \frac{1}{G}$ and letting $\eta_t(\|w\|) = \eta \|w\|$, we have $|\Psi_t'''(x)| \le \frac{\eta_t'(x)}{2} \Psi_t''(x)^2$ for all $x > 0$, and $\eta_t(x)$ is a $1/G$ Lipschitz convex function. Hence, using Lemma 2 we have

$$\begin{aligned}
\delta_t &\le \langle g_t, w_t - w_{t+1} \rangle - D_{\psi_t}(w_{t+1}|w_t) - \eta \|g_t\|^2 \|w_{t+1}\| - b_t \|w_{t+1}\| \\
&= \langle g_t, w_t - w_{t+1} \rangle - D_{\psi_t}(w_{t+1}|w_t) - \eta_t(\|w_{t+1}\|) \|g_t\|^2 - b_t \|w_{t+1}\| \\
&\le \frac{2\|g_t\|^2}{\Psi_t''(0)} - b_t \|w_{t+1}\| = \eta\alpha_t \|g_t\|^2 - b_t \|w_{t+1}\| .
\end{aligned}$$

Plugging this back into the full regret bound we have

$$
\begin{aligned}
R_T(\boldsymbol{u}) \le& \frac{2M\log\left(M/\alpha_{T+1}+1\right)}{\eta} + 2\sum_{t=1}^{T-1}\frac{\|u_t-u_{t+1}\|\log\left(\|w_{t+1}\|/\alpha_{t+1}+1\right)}{\eta} + \sum_{t=1}^{T}\varphi_t(u_t) \\
&+ \sum_{t=1}^{T}\eta\alpha_t\|g_t\|^2 - b_t\|w_{t+1}\| \\
=& \frac{2M\log\left(M/\alpha_{T+1}+1\right)}{\eta} + \sum_{t=1}^{T-1}\left[\frac{2\|u_{t+1}-u_t\|\log\left(\|w_{t+1}\|/\alpha_{t+1}+1\right)}{\eta} - b_t\|w_{t+1}\|\right] \\
&+ \sum_{t=1}^{T}\left(\eta\|g_t\|^2+b_t\right)\|u_t\| + \eta\sum_{t=1}^{T}\alpha_t\|g_t\|^2.
\end{aligned}
$$

■

With this result in hand, we prove Proposition 1 by choosing values $\alpha_t = \frac{\epsilon G^2}{V_t\log^2(V_t/G^2)}$ and $b_t = \eta\|g_t\|^2$. The full version of the result is given below.

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

$$
\begin{aligned}
\theta_t &= \nabla\psi_t(w_t) - g_t \\
w_{t+1} &= \frac{\alpha_{t+1}\theta_t}{\|\theta_t\|}\left[\exp\left[\frac{\eta}{2}\max\left(\|\theta_t\|-2\eta\|g_t\|^2,0\right)\right]-1\right]
\end{aligned}
$$

where we we define $C\frac{x}{\|x\|} = \mathbf{0}$ for all C when $x=\mathbf{0}$. Moreover, for any $u_1,\dots,u_T$ in $\mathbb{R}^d$, Algorithm 1 guarantees

$$
R_T(\boldsymbol{u}) \le 2\epsilon G + \frac{4\left(M+P_T\right)\left[\log\left(\frac{9MT^2}{4\alpha_{T+1}}+1\right)\vee 1\right]}{\eta} + 2\eta\sum_{t=1}^{T}\|g_t\|^2\|u_t\|.
$$

where $M = \max_t\|u_t\|$.

Proof First, we will verify the update equation, and then show the regret bound. To compute the update, observe that from the first-order optimality conditions, there is some $\nabla\phi_t(w_{t+1})\in\partial\phi_t(w_{t+1})$ such that

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

Now, notice that we can write $\nabla\phi_t(w_{t+1}) = \nabla\psi_{t+1}(w_{t+1}) - \nabla\psi_t(w_{t+1}) + \nabla\varphi_t(w_{t+1})$ for some $\nabla\varphi_t(w_{t+1})\in\partial\varphi_t(w_{t+1})$. Thus, we have:

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

Moreover, *any* value for w_{t+1} such that there is a $\varphi_t(w_{t+1}) \in \partial\varphi_t(w_{t+1})$ satisfying the above condition is valid solution to the mirror descent update. We justify our update equation in two cases.

First, consider the case $\max(\|\theta_t\| - 2\eta\|g_t\|, 0) = 0$, In this case, the update equation suggests $w_{t+1} = \mathbf{0}$. To justify this, notice that $\partial\varphi_t(\mathbf{0})$ consists of all vectors of norm at most $2\eta \|g_t\|^2$. Further, $\nabla\psi_{t+1}(\mathbf{0}) = \mathbf{0}$. Thus, whenever $\max(\|\theta_t\| - 2\eta\|g_t\|, 0) = 0$, we can set $w_{t+1} = \mathbf{0}$ as described by our update.

Now, let us suppose $\max(\|\theta_t\| - 2\eta\|g_t\|, 0) = \|\theta_t\| - 2\eta\|g_t\| > 0$. Note that this implies $\theta_t \neq \mathbf{0}$, and the update equation sets $w_{t+1} \neq \mathbf{0}$. In the case $w_{t+1} \neq \mathbf{0}$, $\varphi_t(w_{t+1})$ is differentiable so that $\varphi_t(w_{t+1}) = 2\eta \|g_t\|^2 \frac{w_{t+1}}{\|w_{t+1}\|}$. Thus, we need to establish that indeed a non-zero w_{t+1} given by the update equation is a solution to the optimality condition:

$$g_t + \nabla\psi_{t+1}(w_{t+1}) - \nabla\psi_t(w_t) + 2\eta \|g_t\|^2 \frac{w_{t+1}}{\|w_{t+1}\|} = \mathbf{0}.$$

Writing $\psi_t(w) = \Psi_t(\|w\|) = \int_0^{\|w\|} \Psi_t'(x)dx$, we have $\nabla\psi_{t+1}(w_{t+1}) = \frac{w_{t+1}}{\|w_{t+1}\|}\Psi_{t+1}'(\|w_{t+1}\|)$ (where we define $\frac{w_{t+1}}{\|w_{t+1}\|} \cdot 0 = \mathbf{0}$) and hence the optimality condition can be re-written:

$$\frac{w_{t+1}}{\|w_{t+1}\|}\left[\Psi_{t+1}'(\|w_{t+1}\|) + 2\eta \|g_t\|^2\right] = \nabla\psi_t(w_t) - g_t = \theta_t$$

Now we need only verify that our expression $w_{t+1} = \frac{\alpha_{t+1}\theta_t}{\|\theta_t\|}\left[\exp\left[\frac{\eta}{2}(\|\theta_t\| - 2\eta\|g_t\|)\right] - 1\right]$ satisfies this condition. Fortunately, this is easily checked by observing the stated update satisfies:

$$\Psi_{t+1}'(\|w_{t+1}\|) = \frac{2}{\eta}\log(\|w_{t+1}\|/\alpha_{t+1} + 1) = \|\theta_t\| - 2\eta\|g_t\|^2.$$

Turning now to the regret, we begin by replacing the comparator sequence with an auxiliary sequence $\widehat{u}_1, \dots, \widehat{u}_T$ to be determined later. This alternative sequence will eventually be designed to have some useful stability properties while still being "close" to the real sequence $u_1, \dots, u_T$:

$$\begin{aligned} R_T(\boldsymbol{u}) &= \sum_{t=1}^T \langle g_t, w_t - u_t\rangle = \sum_{t=1}^T \langle g_t, w_t - \widehat{u}_t\rangle + \sum_{t=1}^T \langle g_t, \widehat{u}_t - u_t\rangle \\ &\le R_T(\widehat{\boldsymbol{u}}) + \sum_{t=1}^T \|g_t\| \|\widehat{u}_t - u_t\| \end{aligned}$$

The first term is bounded via Proposition 2 as

$$\begin{aligned} R_T(\widehat{\boldsymbol{u}}) \le \frac{2\widehat{M}\log\left(\widehat{M}/\alpha_{T+1} + 1\right)}{\eta} + 2\eta\sum_{t=1}^T \|g_t\|^2 \|\widehat{u}_t\| + \eta\sum_{t=1}^T \alpha_t \|g_t\|^2 \\ + \sum_{t=1}^{T-1}\left[\frac{2\|\widehat{u}_t - \widehat{u}_{t+1}\|\log\left(\|w_{t+1}\|/\alpha_{t+1} + 1\right)}{\eta} - \eta\|g_t\|^2\|w_{t+1}\|\right] \end{aligned}$$

where $\widehat{M} = \max_{t\le T}\|\widehat{u}_t\|$. We focus first on bounding the sum in the second line. To do so, we first provide the definition of $\widehat{u}_t$:

$$\text{Let } \mathcal{T} > 0 \text{ and set } \widehat{u}_T = u_T \text{ and } \widehat{u}_t = \begin{cases} u_t & \text{if } \|g_t\| \geq \mathcal{T} \\ \widehat{u}_{t+1} & \text{otherwise} \end{cases} \text{ for } t < T.$$

Hence, by definition we have $\|\widehat{u}_t - \widehat{u}_{t+1}\| = 0$ whenever $\|g_t\| \leq \mathcal{T}$, so

$$\begin{aligned}
&\sum_{t=1}^{T-1} \left[\frac{2\left\|\widehat{u}_t - \widehat{u}_{t+1}\right\| \log\left(\left\|w_{t+1}\right\|/\alpha_{t+1} + 1\right)}{\eta} - \eta\left\|g_t\right\|^2\left\|w_{t+1}\right\|\right] \\
&\quad\leq \sum_{t:\|g_t\|\geq\mathcal{T}} \left[\frac{2\left\|\widehat{u}_t - \widehat{u}_{t+1}\right\| \log\left(\left\|w_{t+1}\right\|/\alpha_{t+1} + 1\right)}{\eta} - \eta\mathcal{T}^2\left\|w_{t+1}\right\|\right] \\
&\quad\leq \sum_{t:\|g_t\|\geq\mathcal{T}} \left[\sup_{X\geq 0} \frac{2\left\|\widehat{u}_t - \widehat{u}_{t+1}\right\| \log\left(X/\alpha_{t+1} + 1\right)}{\eta} - \eta\mathcal{T}^2 X\right] \\
&\quad\overset{(*)}{\leq} \sum_{t:\|g_t\|\geq\mathcal{T}} \frac{2\left\|\widehat{u}_t - \widehat{u}_{t+1}\right\| \log\left(\frac{2\|\widehat{u}_t - \widehat{u}_{t+1}\|}{\alpha_{t+1}\eta^2\mathcal{T}^2}\right)}{\eta}
\end{aligned}$$

where $(*)$ observes that either the max is obtained at $X = 0$, for which $\sup_{X\geq 0} \frac{2\|\widehat{u}_t - \widehat{u}_{t+1}\| \log(X/\alpha_{t+1}+1)}{\eta} - \eta\mathcal{T}^2 X = 0$, and otherwise the max is obtained at $X = \frac{2\|\widehat{u}_{t+1} - u_t\|}{\eta^2\mathcal{T}^2} - \alpha_{t+1} > 0$, which leads to an upperbound of

$$\sup_{X\geq 0} \frac{2\left\|\widehat{u}_t - \widehat{u}_{t+1}\right\| \log\left(X/\alpha_{t+1} + 1\right)}{\eta} - \eta\mathcal{T}^2 X \leq \frac{2\left\|\widehat{u}_t - \widehat{u}_{t+1}\right\| \log\left(\frac{2\|\widehat{u}_{t+1} - \widehat{u}_t\|}{\alpha_{t+1}\eta^2\mathcal{T}^2}\right)}{\eta}$$

in both cases. Moreover, for any t such that $\|g_t\| \geq \mathcal{T}$ let t' denote the smallest index greater than t for which $\|g_{t'}\| \geq \mathcal{T}$; then by triangle inequality we have $\|\widehat{u}_t - \widehat{u}_{t+1}\| = \|u_t - u_{t'}\| \leq \sum_{s=t}^{t'} \|u_s - u_{s+1}\|$ and

$$\begin{aligned}
\sum_{t:\|g_t\|\geq\mathcal{T}} \frac{2\left\|\widehat{u}_t - \widehat{u}_{t+1}\right\| \log\left(\frac{2\|\widehat{u}_t - \widehat{u}_{t+1}\|}{\alpha_{t+1}\eta^2\mathcal{T}^2}\right)}{\eta} &\leq \sum_{t:\|g_t\|\geq\mathcal{T}} \frac{\sum_{s=t}^{t'} 2\left\|u_s - u_{s+1}\right\| \log\left(\frac{4\widehat{M}}{\alpha_{T+1}\eta^2\mathcal{T}^2}\right)}{\eta} \\
&= \frac{2P_T \log\left(\frac{4\widehat{M}}{\alpha_{T+1}\eta^2\mathcal{T}^2}\right)}{\eta}.
\end{aligned}$$

Returning to the regret against the auxiliary comparator sequence we have

$$
\begin{aligned}
R_T(\widehat{\boldsymbol{u}}) &\leq \frac{2\widehat{M}\log\left(\widehat{M}/\alpha_{T+1}+1\right)}{\eta} + 2\eta\sum_{t=1}^T \|g_t\|^2\|\widehat{u}_t\| + \eta\sum_{t=1}^T \alpha_t\|g_t\|^2 + \frac{2P_T\log\left(\frac{4\widehat{M}}{\alpha_{t+1}\eta^2\mathcal{T}^2}\right)}{\eta}\\
&\overset{(a)}{\leq} \frac{2M\log\left(M/\alpha_{T+1}+1\right) + 2P_T\log\left(\frac{4M}{\alpha_{t+1}\eta^2\mathcal{T}^2}\right)}{\eta} + 2\eta\sum_{t=1}^T\|g_t\|^2\|\widehat{u}_t\| + \eta\sum_{t=1}^T\alpha_t\|g_t\|^2\\
&\overset{(b)}{\leq} \frac{2M\log\left(M/\alpha_{T+1}+1\right) + 2P_T\log\left(\frac{4M}{\alpha_{t+1}\eta^2\mathcal{T}^2}\right)}{\eta} + \eta\sum_{t=1}^T\alpha_t\|g_t\|^2\\
&\quad + 2\eta\sum_{t=1}^T\|g_t\|^2\|u_t\| + 2\sum_{t=1}^T\|g_t\|\|\widehat{u}_t-u_t\|\\
&\overset{(c)}{\leq} \frac{2M\log\left(M/\alpha_{T+1}+1\right) + 2P_T\log\left(\frac{4M}{\alpha_{t+1}\eta^2\mathcal{T}^2}\right)}{\eta} + 2\epsilon G\\
&\quad + 2\eta\sum_{t=1}^T\|g_t\|^2\|u_t\| + 2\sum_{t=1}^T\|g_t\|\|\widehat{u}_t-u_t\|,
\end{aligned}
$$

where (a) observes that $\widehat{M} = \max_{t\leq T}\|\widehat{u}_t\| \leq \max_{t\leq T}\|u_t\| = M$ and (b) recalls $\eta \leq \frac{1}{G}$ and uses $\eta\|g_t\|^2\|\widehat{u}_t\| \leq \eta\|g_t\|^2(\|u_t-\widehat{u}_t\|+\|u_t\|) \leq \eta\|g_t\|^2\|u_t\| + \|g_t\|\|u_t-\widehat{u}_t\|$, and (c) chooses $\alpha_t = \frac{\epsilon G^2}{V_t\log^2(V_t/G^2)}$ for $V_t = 4G^2 + \|g\|^2_{1:t-1}$ and applies Lemma 9 to bound

$$
\begin{aligned}
\eta\sum_{t=1}^T\alpha_t\|g_t\|^2 &= \eta\epsilon G^2\sum_{t=1}^T\frac{\|g_t\|^2}{V_t\log^2(V_t/G^2)}\\
&\leq 2\eta\epsilon G^2 \leq 2\epsilon G
\end{aligned}
$$

Returning now to the full regret bound and recalling $\widehat{u}_t = u_t$ whenever $\|g_t\| \geq \mathcal{T}$ and $\widehat{u}_t = \widehat{u}_{t+1}$ otherwise, we have

$$
\begin{aligned}
R_T(\boldsymbol{u}) &\leq R_T(\widehat{\boldsymbol{u}}) + \sum_{t=1}^{T} \|g_t\| \|\widehat{u}_t - u_t\| \\
&\leq 2\epsilon G + \frac{2M \log\left(M/\alpha_{T+1} + 1\right) + 2P_T \log\left(\frac{4M}{\alpha_{t+1}\eta^2\mathcal{T}^2}\right)}{\eta} \\
&\quad + 2\eta \sum_{t=1}^{T} \|g_t\|^2 \|u_t\| + 3\sum_{t=1}^{T} \|g_t\| \|\widehat{u}_t - u_t\| \\
&\leq 2\epsilon G + \frac{2M \log\left(M/\alpha_{T+1} + 1\right) + 2P_T \log\left(\frac{4M}{\alpha_{t+1}\eta^2\mathcal{T}^2}\right)}{\eta} \\
&\quad + 2\eta \sum_{t=1}^{T} \|g_t\|^2 \|u_t\| + 3\mathcal{T} \sum_{t:\|g_t\| \leq \mathcal{T}} \|\widehat{u}_{t+1} - u_t\| \\
&\overset{(a)}{\leq} 2\epsilon G + \frac{2M \log\left(M/\alpha_{T+1} + 1\right) + 2P_T \log\left(\frac{4M}{\alpha_{t+1}\eta^2\mathcal{T}^2}\right)}{\eta} \\
&\quad + 2\eta \sum_{t=1}^{T} \|g_t\|^2 \|u_t\| + 3\mathcal{T}TP_T.
\end{aligned}
$$

where (a) uses the fact that $\widehat{u}_{t+1} = u_{t'}$ for *some* $t' \geq t$, so that $\|\widehat{u}_{t+1} - u_t\| \leq \sum_{s=1}^{t'-1} \|u_{s+1} - u_s\| \leq P_T$. Since this bound holds for an arbitrary $\mathcal{T} > 0$ we are free to choose a $\mathcal{T}$ which tightens the upperbound, such as $\mathcal{T} = \frac{4}{3\eta T}$:

$$
\begin{aligned}
R_T(\boldsymbol{u}) &\leq \inf_{\mathcal{T}>0} 2\epsilon G + \frac{2M \log\left(M/\alpha_{T+1} + 1\right) + 2P_T \log\left(\frac{4M}{\alpha_{t+1}\eta^2\mathcal{T}^2}\right)}{\eta} \\
&\quad + 2\eta \sum_{t=1}^{T} \|g_t\|^2 \|u_t\| + \mathcal{T}3TP_T \\
&\leq \frac{2M \log\left(M/\alpha_{T+1} + 1\right) + 2P_T \left(\log\left(\frac{9MT^2}{4\alpha_{t+1}}\right) + 2\right)}{\eta} \\
&\quad + 2\epsilon G + 2\eta \sum_{t=1}^{T} \|g_t\|^2 \|u_t\| \\
&\leq 2\epsilon G + \frac{4\left(M + P_T\right)\left\{\log\left(\frac{9MT^2}{4\alpha_{T+1}} + 1\right) \vee 1\right\}}{\eta} + 2\eta \sum_{t=1}^{T} \|g_t\|^2 \|u_t\|.
\end{aligned}
$$

■

D.4. Proof of Theorem 4

The full statement of the theorem is given below.

Theorem 4 *For any* $\varepsilon > 0$ *and* $u_1, \dots, u_T$ *in* $\mathbb{R}^d$*, Algorithm 2 guarantees*

$$R_T(\boldsymbol{u}) \leq 2\varepsilon G + 6\sqrt{2(M+P_T)\left[\log\left(\frac{9M\Lambda_T}{4\varepsilon}+1\right)\vee 1\right]\sum_{t=1}^{T}\|g_t\|^2\|u_t\|}$$
$$+ 4G(M+P_T)\left[\log\left(\frac{9M\Lambda_T}{4\varepsilon}+1\right)\vee 1\right].$$

where $\Lambda_T = T^2\left(4+\frac{\|g\|_{1:T}^2}{G^2}\right)\log^2\left(4+\frac{\|g\|_{1:T}^2}{G^2}\right)\lceil\log_2(\sqrt{T})\rceil \leq O\left(T^3\log^3(T)\right)$ *and* $M = \max_t\|u_t\|$.

Proof Let $\mathcal{A}_\eta$ denote an instance of the algorithm in Proposition 1, w_t^η denote its iterates, and let $R_T^{\mathcal{A}_\eta}(\boldsymbol{u})$ denote the dynamic regret of $\mathcal{A}_\eta$. From Proposition 1, we have that for any $\eta \leq \frac{1}{G}$,

$$R_T^{\mathcal{A}_\eta}(\boldsymbol{u}) \leq 2\epsilon G + \frac{4\left(M+P_T\right)\left[\log\left(\frac{9MT^2}{4\alpha_{T+1}}+1\right)\vee 1\right]}{\eta} + 2\eta\sum_{t=1}^{T}\|g_t\|^2\|u_t\|,$$

where $\alpha_{T+1} = \frac{\epsilon G^2}{V_{T+1}\log^2(V_{T+1}/G^2)}$ and $V_{T+1} = 4G^2 + \|g\|_{1:T}^2$, $M = \max_{t\leq T}\|u_t\|$, $P_T = \sum_{t=2}^{T}\|u_t - u_{t-1}\|$, and $\epsilon > 0$. The stepsize which minimizes the right-hand side of the inequality is

$$\eta^* = \min\left\{\sqrt{\frac{2(M+P_T)\left[\log\left(\frac{9MT^2}{4\alpha_{T+1}}+1\right)\vee 1\right]}{\sum_{t=1}^{T}\|g_t\|^2\|u_t\|}}, \frac{1}{G}\right\},$$

for which we have

$$R_T^{\mathcal{A}_{\eta^*}}(\boldsymbol{u}) \leq 2\epsilon G + 4\sqrt{2(M+P_T)\left[\log\left(\frac{9MT^2}{4\alpha_{T+1}}+1\right)\vee 1\right]\sum_{t=1}^{T}\|g_t\|^2\|u_t\|}$$
$$+ 2G(M+P_T)\left[\log\left(\frac{9MT^2}{4\alpha_{T+1}}+1\right)\vee 1\right].$$

In what follows, we will match this bound up to constant factors using the iterate adding approach proposed by Cutkosky (2019b).

Suppose that we have a collection of step-sizes $\mathcal{S} = \left\{\eta \in \mathbb{R} : 0 < \eta \leq \frac{1}{G}\right\}$ and suppose that on each round we play $w_t = \sum_{\eta\in\mathcal{S}} w_t^\eta$ where w_t^η is the output of $\mathcal{A}_\eta$. Then for any $\widetilde{\eta} \in \mathcal{S}$ we can write

$$\begin{aligned}
R_T(\boldsymbol{u}) &= \sum_{t=1}^T \langle g_t, w_t - u_t\rangle = \sum_{t=1}^T \left\langle g_t, \sum_{\eta\in\mathcal{S}} w_t^\eta - u_t\right\rangle \\
&= \sum_{t=1}^T \left\langle g_t, w_t^{\widetilde{\eta}} - u_t\right\rangle + \sum_{\eta\neq\widetilde{\eta}\in\mathcal{S}} \sum_{t=1}^T \langle g_t, w_t^\eta - \mathbf{0}\rangle \\
&= R_T^{\mathcal{A}_{\widetilde{\eta}}}(\boldsymbol{u}) + \sum_{\eta\neq\widetilde{\eta}\in\mathcal{S}} R_T^{\mathcal{A}_\eta}(\mathbf{0}) \\
&\leq R_T^{\mathcal{A}_{\widetilde{\eta}}}(\boldsymbol{u}) + 2\epsilon G(|\mathcal{S}| - 1). \qquad (5)
\end{aligned}$$

Notice that since this holds for any $\widetilde{\eta} \in \mathcal{S}$, it holds for the one with the lowest dynamic regret, hence

$$R_T(\boldsymbol{u}) \leq 2\epsilon G(|\mathcal{S}| - 1) + \min_{\eta\in\mathcal{S}} R_T^{\mathcal{A}_\eta}(\boldsymbol{u}).$$

Thus, we need only ensure that there is *some* $\eta \in \mathcal{S}$ which is close to the optimal η^*. It is easy to see that

$$\eta^* = \min\left\{\sqrt{\frac{2(M + P_T)\left[\log\left(\frac{9MT^2}{4\alpha_{T+1}} + 1\right) \vee 1\right]}{\sum_{t=1}^T \|g_t\|^2 \|u_t\|}}, \frac{1}{G}\right\} \implies \frac{2}{G\sqrt{T}} \leq \eta^* \leq \frac{1}{G},$$

so if we let $\mathcal{S} = \left\{\frac{2^k}{G\sqrt{T}} \wedge \frac{1}{G} : 1 \leq k \leq \lceil\log_2\left(\sqrt{T}\right)\rceil\right\}$, we'll have

$$\eta_{\min} = \frac{2}{G\sqrt{T}} \leq \eta^* \leq \frac{1}{G} = \eta_{\max},$$

where $\eta_{\min}$ and $\eta_{\max}$ are the smallest and largest step-sizes in $\mathcal{S}$ respectively. Hence, there must be an $\eta_k \in \mathcal{S}$ such that $\eta_k \leq \eta^* \leq \eta_{k+1} \leq 2\eta_k$. Using $\widetilde{\eta} = \eta_k$ in Equation (5) yields

$$\begin{aligned}
R_T(\boldsymbol{u}) &\leq 2\epsilon G(|\mathcal{S}| - 1) + R_T^{\mathcal{A}_{\eta_k}}(\boldsymbol{u}) \\
&\leq 2\epsilon G\,|\mathcal{S}| + \frac{4\,(M + P_T)\left[\log\left(\frac{9MT^2}{4\alpha_{T+1}} + 1\right) \vee 1\right]}{\eta_k} + 2\eta_k \sum_{t=1}^T \|g_t\|^2 \|u_t\| \\
&\leq 2\epsilon G\,|\mathcal{S}| + \frac{8\,(M + P_T)\left[\log\left(\frac{9MT^2}{4\alpha_{T+1}} + 1\right) \vee 1\right]}{\eta^*} + 2\eta^* \sum_{t=1}^T \|g_t\|^2 \|u_t\| \\
&= 2\epsilon G\,|\mathcal{S}| + 6\sqrt{2(M + P_T)\left[\log\left(\frac{9MT^2}{4\alpha_{T+1}} + 1\right) \vee 1\right] \sum_{t=1}^T \|g_t\|^2 \|u_t\|} \\
&\quad + 4G(M + P_T)\left[\log\left(\frac{9MT^2}{4\alpha_{T+1}} + 1\right) \vee 1\right].
\end{aligned}$$

The result then follows by choosing $\epsilon = \frac{\varepsilon}{\lceil\log_2(\sqrt{T})\rceil} \leq \frac{\varepsilon}{|\mathcal{S}|}$. ■

Appendix E. Proofs for Section 5 (Adapting to Gradient Variability)

Theorem 5 *Let $\ell_1, \dots, \ell_T$ and $\widehat{\ell}_1, \dots, \widehat{\ell}_T$ be G-Lipschitz convex functions. Let $\epsilon > 0$, $k \ge 3$, and for all t set $\widehat{V}_t = 16G^2 + \sum_{s=1}^{t-1} \left\| \nabla \ell_s(w_s) - \nabla \widehat{\ell}_s(w_s) \right\|^2$, $\widehat{\alpha}_t = \frac{\epsilon G}{\sqrt{\widehat{V}_t} \log^2(\widehat{V}_t/G^2)}$, and*

$$\psi_t(w) = k \int_0^{\|w\|} \min_{\eta \le \frac{1}{2G}} \left[\frac{\log\left(x/\widehat{\alpha}_t + 1\right)}{\eta} + \eta \widehat{V}_t \right] dx.$$

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

$$R_T(u) \le 4\epsilon G + 2k \|u\| \max\left\{ \sqrt{\widehat{V}_t \log\left(\|u\| / \widehat{\alpha}_{T+1} + 1\right)}, 2G \log\left(\|u\| / \widehat{\alpha}_{T+1} + 1\right) \right\}$$

Proof The proof follows similar steps to Theorem 1. Let $g_t \in \ell_t(w_t)$ and let $h_t \in \partial \widehat{\ell}_t(w_t)$ be the subgradient of $\widehat{\ell}_t(w_t)$ for which the first-order optimality condition $h_t + \nabla \psi_t(w_t) - \nabla \psi_t(x_t) = \mathbf{0}$ holds. Then

$$\begin{aligned}
\sum_{t=1}^T \langle g_t, w_t - u \rangle &= \sum_{t=1}^T \langle g_t, x_{t+1} - u \rangle + \langle g_t, w_t - x_{t+1} \rangle \\
&= \sum_{t=1}^T \langle g_t, x_{t+1} - u \rangle + \langle h_t, w_t - x_{t+1} \rangle + \langle g_t - h_t, w_t - x_{t+1} \rangle .
\end{aligned}$$

Following the same steps as Lemma 1 we have

$$\begin{aligned}
\sum_{t=1}^T \langle g_t, x_{t+1} - u \rangle &\le D_{\psi_{T+1}}(u|x_1) - D_{\psi_{T+1}}(u|x_{T+1}) + \sum_{t=1}^T -D_{\psi_t}(x_{t+1}|x_t) - \phi_t(x_{t+1}) \\
&\le \psi_{T+1}(u) + \sum_{t=1}^T -D_{\psi_t}(x_{t+1}|x_t) - \phi_t(w_{t+1}),
\end{aligned}$$

where the last line observes $\arg\min_{x \in \mathbb{R}^d} \psi_{T+1}(x) = \psi_{T+1}(x_1) = 0$, so $D_{\psi_{T+1}}(u|x_1) = \psi_{T+1}(u)$ and $-D_{\psi_{T+1}}(u|x_{T+1}) \le 0$. Similarly, from the first-order optimality condition for w_t we have

$$\begin{aligned}
\sum_{t=1}^T \langle h_t, w_t - x_{t+1} \rangle &= \sum_{t=1}^T \langle \nabla \psi_t(w_t) - \nabla \psi_t(x_t), w_t - x_{t+1} \rangle \\
&= \sum_{t=1}^T D_{\psi_t}(x_{t+1}|x_t) - D_{\psi_t}(x_{t+1}|w_t) \underbrace{-D_{\psi_t}(w_t|x_t)}_{\le 0} \\
&\le \sum_{t=1}^T D_{\psi_t}(x_{t+1}|x_t) - D_{\psi_t}(x_{t+1}|w_t)
\end{aligned}$$

where the second line applies the three-point relation for Bregman divergences:

$$\langle \nabla f(y) - \nabla f(x), x - z \rangle = D_f(z|y) - D_f(z|x) - D_f(x|y).$$

Combining these two observations yields

$$R_T(u) \leq \psi_{T+1}(u) + \sum_{t=1}^{T} \underbrace{\langle g_t - h_t, w_t - x_{t+1}\rangle - D_{\psi_t}(x_{t+1}|w_t) - \phi_t(x_{t+1})}_{=:\delta_t}$$

To bound δ_t, define $\widehat{g}_t = \nabla \ell_t(w_t) - \nabla \widehat{\ell}_t(w_t)$, $\widehat{G} = 2G$, $\widehat{V}_t = 4\widehat{G}^2 + \sum_{s=1}^{t-1} \|\widehat{g}_s\|^2$, $\widehat{\alpha}_t = \frac{\epsilon G}{\sqrt{\widehat{V}_t}\log^2(\widehat{V}_t/G^2)}$, and observe that $\psi_t(w) = k\int_0^{\|w\|} \min_{\eta \leq 1/\widehat{G}} \left[\frac{\log(x/\widehat{\alpha}_t+1)}{\eta} + \eta\widehat{V}_t\right] dx$ is equivalent to the regularizer from Theorem 1. Hence, borrowing the arguments of Theorem 1, we can bound $\sum_{t=1}^{T} \delta_t \leq 4\epsilon G$. Returning to our regret bound, we have

$$\begin{aligned} R_T(u) &\leq \psi_{T+1}(u) + 4\epsilon G \overset{(a)}{\leq} 4\epsilon G + \|u\| \, \Psi'_{T+1}(\|u\|) \\ &\overset{(b)}{\leq} 4\epsilon G + 2k\|u\| \max\left\{\sqrt{\widehat{V}_t \log\left(\|u\|/\widehat{\alpha}_{T+1} + 1\right)}, 2G\log\left(\|u\|/\widehat{\alpha}_{T+1} + 1\right)\right\} \end{aligned}$$

where (a) defines

$$\begin{aligned} \Psi'_{T+1}(x) &= k \min_{\eta \leq 1/2G} \left[\frac{\log\left(x/\widehat{\alpha}_{T+1} + 1\right)}{\eta} + \eta\widehat{V}_{T+1}\right] \\ &= \begin{cases} 2k\sqrt{\widehat{V}_{T+1}\log\left(x/\widehat{\alpha}_{T+1}+1\right)} & \text{if } 2G\sqrt{\log\left(x/\widehat{\alpha}_{T+1}+1\right)} \leq \sqrt{\widehat{V}_{T+1}} \\ 2kG\log\left(x/\widehat{\alpha}_{T+1}+1\right) + \frac{k\widehat{V}_{T+1}}{2G} & \text{otherwise} \end{cases} \end{aligned}$$

and observes that $\psi_{T+1}(u) = \int_0^{\|u\|} \Psi'_t(x)dx \leq \|u\| \, \Psi'_t(\|u\|)$ since Ψ'_t is non-decreasing in its argument, and (b) observes that the case $\Psi'_t(x) = 2kG\log\left(x/\widehat{\alpha}_{T+1}+1\right) + \frac{k\widehat{V}_{T+1}}{2G}$, coincides with $\widehat{V}_{T+1}/2G \leq \sqrt{\widehat{V}_{T+1}\log\left(x/\widehat{\alpha}_{T+1}+1\right)} \leq 2G\log\left(x/\widehat{\alpha}_{T+1}+1\right)$, so

$$\begin{aligned} \Psi'_{T+1}(x) &\leq \begin{cases} 2k\sqrt{\widehat{V}_{T+1}\log\left(x/\widehat{\alpha}_{T+1}+1\right)} & \text{if } 2G\sqrt{\log\left(x/\widehat{\alpha}_{T+1}+1\right)} \leq \sqrt{\widehat{V}_{T+1}} \\ 4kG\log\left(x/\widehat{\alpha}_{T+1}+1\right) & \text{otherwise} \end{cases} \\ &= 2k\max\left\{\sqrt{\widehat{V}_{T+1}\log\left(x/\widehat{\alpha}_{T+1}+1\right)}, 2G\log\left(x/\widehat{\alpha}_{T+1}+1\right)\right\} \end{aligned}$$

■

Appendix F. A Lipschitz Adaptive, Scale-free Algorithm for Unbounded Domains

The full pseudocode for our Scale-free, Lipschitz adaptive algorithm for unbounded domains is given in Algorithm 5. The update equation is derived in a similar manner to the algorithm in Section 3.

The implementation can be understood as the Leashed meta-algorithm of Cutkosky (2019a) with an instance of the algorithm specified in Theorem 6 as the base algorithm. The corresponding regret guarantee is immediate using Theorem 6 along with the with the aforementioned reductions (Cutkosky, 2019a, Theorem 3).

Corollary 1 *For any $u \in \mathbb{R}^d$, Algorithm 5 guarantees*

$$R_T(u) \leq \widehat{O}\left(\epsilon G_T + \|u\| \left[\sqrt{V_{T+1} \log\left(\frac{\|u\| \sqrt{B_{T+1}}}{\epsilon} + 1 \right)} \vee G_T \log\left(\frac{\|u\| \sqrt{B_{T+1}}}{\epsilon} + 1 \right) \right] \right.$$
$$\left. + G_T \|u\|^3 + G_T \|u\| + G_T \sqrt{\sum_{t=1}^{T} \frac{\|g_t\|}{G_t}} \right),$$

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

Algorithm 5: Unbounded, Scale-Free, Lipschitz Adaptivity

Initialize $w_1 = \mathbf{0}$, $h_1 = 0$, $G_0 = 0$, $\widetilde{b}_1 = 4$, $\widetilde{B}_1 = 4\widetilde{b}_1$, $\widetilde{\theta}_1 = \mathbf{0}$
for $t = 1 : T$ **do**
  Define $D_t = \sum_{s=1}^{t-1} \frac{\|g_s\|}{G_s}$ and $W_t = \left\{ w \in \mathbb{R}^d : \|w\| \leq D_t \right\}$
  Play $\widehat{w_t} = \Pi_{W_t}(w_t) = w_t \min\left\{1, \frac{D_t}{\|w_t\|}\right\}$
  Receive subgradient g_t
  Set $\overline{g}_t = g_t \min\left\{1, \frac{h_t}{\|g_t\|}\right\}$, $G_t = \max\left\{\|g_t\|, G_{t-1}\right\}$, and $h_{t+1} = G_t$
  Set $\widetilde{\ell}_t(w) = \frac{1}{2}\langle \overline{g}_t, w\rangle + \frac{1}{2}\|\overline{g}_t\| \max\left\{0, \|w_t\| - D_t\right\}$ and compute $\widetilde{g}_t \in \partial \widetilde{\ell}_t(w_t)$
  Set $\widetilde{\theta}_{t+1} = \widetilde{\theta}_t - \widetilde{g}_t$, $\widetilde{V}_{t+1} = 4h_{t+1}^2 + \|\widetilde{g}\|_{1:t}^2$, $\widetilde{b}_{t+1} = \widetilde{b}_t + \frac{\|\widetilde{g}_t\|^2}{h_t^2}$, $\widetilde{B}_{t+1} = \widetilde{B}_t + 4\widetilde{b}_t$, and
    $\widetilde{\alpha}_{t+1} = \frac{\epsilon}{\sqrt{\widetilde{B}_{t+1}} \log^2(\widetilde{B}_{t+1})}$
  Define $f_{t+1}(\theta) = \begin{cases} \frac{\|\theta\|^2}{36\widetilde{V}_{t+1}} & \text{if } \|\theta\| \leq \frac{6\widetilde{V}_{t+1}}{h_{t+1}} \\ \frac{\|\theta\|}{3h_{t+1}} - \frac{\widetilde{V}_{t+1}}{h_{t+1}^2} & \text{otherwise} \end{cases}$
  Update $w_{t+1} = \frac{\widetilde{\alpha}_{t+1}\widetilde{\theta}_{t+1}}{\|\widetilde{\theta}_{t+1}\|}\left[\exp\left(f_{t+1}\left(\widetilde{\theta}_{t+1}\right)\right) - 1\right]$
end

Appendix G. Proofs for Section 6 (Lipschitz Adaptivity and Scale-free Learning)

G.1. Proof of Theorem 6

The complete theorem is stated below.

Theorem 6 *Let $\ell_1, \ldots, \ell_T$ be G-Lipschitz convex functions and $g_t \in \partial \ell_t(w_t)$ for all t. Let $h_1 \leq \ldots \leq h_T$ be a sequence of hints such that $h_t \geq \|g_t\|$, and assume that h_t is provided at the start of each round t. Let $\epsilon > 0$, $k \geq 3$, $V_t = 4h_t^2 + \|g\|_{1:t-1}^2$, $B_t = 4\sum_{s=1}^t \left(4 + \sum_{s'=1}^{s-1} \frac{\|g_{s'}\|^2}{h_{s'}^2}\right)$, $\alpha_t = \frac{\epsilon}{\sqrt{B_t}\log^2(B_t)}$, and set*

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

Then for all $u \in \mathbb{R}^d$, Algorithm 1 guarantees

$$R_T(u) \leq 4\epsilon h_T + 2k\|u\| \max\left\{\sqrt{V_{T+1}\log\left(\frac{\|u\|\sqrt{B_{T+1}}\log^2(B_{T+1})}{\epsilon} + 1\right)},\right.$$
$$\left. h_T \log\left(\frac{\|u\|\sqrt{B_{T+1}}\log^2(B_{T+1})}{\epsilon} + 1\right)\right\}$$

Proof The proof follows similar steps to Theorem 1. We have via Lemma 1 that

$$R_T(u) \leq \psi_{T+1}(u) + \sum_{t=1}^T \underbrace{\langle g_t, w_t - w_{t+1}\rangle - D_{\psi_t}(w_{t+1}|w_t) - \Delta_t(w_{t+1})}_{=:\delta_t},$$

so the main challenge is to bound the stability terms $\sum_{t=1}^T \delta_t$, which we focus on first.

Let $F_t(w) = \log(x/\alpha_t + 1)$ and define

$$\Psi_t(x) = k \int_0^x \min_{\eta \leq 1/h_t} \left[\frac{F_t(x)}{\eta} + \eta V_t\right] dx,$$

so that $\psi_t(w) = \Psi_t(\|w\|)$, and observe that

$$\begin{aligned}
\Psi_t'(x) &= k \min_{1/h_t} \left[\frac{F_t(x)}{\eta} + \eta V_t\right] \\
&= \begin{cases} 2k\sqrt{V_t F_t(x)} & \text{if } h_t\sqrt{F_t(x)} \leq \sqrt{V_t} \\ kh_t F_t(x) + \frac{kV_t}{h_t} & \text{otherwise} \end{cases} \\
\Psi_t''(x) &= \begin{cases} \frac{k}{x+\alpha_t}\sqrt{\frac{F_t(x)}{V_t}} & \text{if } h_t\sqrt{F_t(x)} \leq \sqrt{V_t} \\ \frac{kh_t}{x+\alpha_t} & \text{otherwise} \end{cases} \\
\Psi_t'''(x) &= \begin{cases} \frac{-k\sqrt{V_t}(1+2F_t(x))}{2(x+\alpha_t)^2 F_t(x)^{3/2}} & \text{if } h_t\sqrt{F_t(x)} \leq \sqrt{V_t} \\ \frac{-kh_t}{(x+\alpha_t)^2} & \text{otherwise.} \end{cases}
\end{aligned}$$

Hence, $\Psi_t(x) \geq 0$, $\Psi'_t(x) \geq 0$, $\Psi''_t(x) \geq 0$, and $\Psi'''_t(x) \leq 0$ for all $x > 0$. Moreover, for any $x > \alpha_t(e-1) \stackrel{\text{def}}{=} x_0$ we have

$$-\frac{\Psi'''_t(x)}{\Psi''_t(x)^2} = \begin{cases} \frac{k\sqrt{V_t}(1+2F_t(x))}{2(x+\alpha_t)F_t(x)^{3/2}} \frac{(x+\alpha_t)^2 F_t(x)}{k^2 V_t} & \text{if } h_t\sqrt{F_t(x)} \leq \sqrt{V_t} \\ \frac{kh_t}{(x+\alpha_t)^2} \frac{(x+\alpha_t)^2}{k^2 h_t^2} & \text{otherwise} \end{cases}$$
$$\leq \begin{cases} \frac{1}{2k\sqrt{V_t}} \left(\frac{1}{\sqrt{F_t(x)}} + 2\sqrt{F_t(x)} \right) & \text{if } h_t\sqrt{F_t(x)} \leq \sqrt{V_t} \\ \frac{1}{kh_t} & \text{otherwise,} \end{cases}$$

and since $x > x_0$, we have $\sqrt{F_t(x)} > 1$ and $\frac{1}{\sqrt{F_t(x)}} \leq \sqrt{F_t(x)}$, hence

$$\leq \begin{cases} \frac{3}{2k}\sqrt{\frac{F_t(x)}{V_t}} & \text{if } h_t\sqrt{F_t(x)} \leq \sqrt{V_t} \\ \frac{1}{kh_t} & \text{otherwise} \end{cases}$$
$$\leq \frac{1}{2}\min\left\{\sqrt{\frac{F_t(x)}{V_t}}, \frac{1}{h_t}\right\}$$
$$= \frac{1}{2}\eta'_t(x),$$

for $k \geq 3$ and $\eta_t(x) = \int_0^{\|w\|} \min\left\{\sqrt{\frac{F_t(x)}{V_t}}, \frac{1}{h_t}\right\} dx$. Notice that η_t is convex and $1/h_t$ Lipschitz with $h_t \geq \|g_t\|$. Hence, Ψ_t satisfies the conditions of Lemma 2 with $x_0 = \alpha_t(e-1)$, so

$$\widehat{\delta}_t = \langle g_t, w_t - w_{t+1}\rangle - D_{\psi_t}(w_{t+1}|w_t) - \eta_t(\|w_{t+1}\|)\|g_t\|^2 \leq \frac{2\|g_t\|^2}{\Psi''_t(x_0)}. \tag{6}$$

Next, we want to show that $\delta_t \leq \widehat{\delta}_t$, which will follow if we can show that $\Delta_t(w) \geq \eta_t(\|w\|)\|g_t\|^2$ for any w. Observe that for any $x > 0$ we have

$$\Psi'_{t+1}(x) - \Psi'_t(x) = k \min_{\eta \leq \frac{1}{h_{t+1}}}\left[\frac{F_{t+1}(x)}{\eta} + \eta V_{t+1}\right] - k \min_{\eta \leq \frac{1}{h_t}}\left[\frac{F_t(x)}{\eta} + \eta V_t\right]$$
$$\geq k \min_{\eta \leq \frac{1}{h_t}}\left[\frac{F_{t+1}(x)}{\eta} + \eta V_{t+1}\right] - k \min_{\eta \leq \frac{1}{h_t}}\left[\frac{F_t(x)}{\eta} + \eta V_t\right].$$

Now observe that for any $\eta \leq 1/h_t$, it holds that $\frac{F_{t+1}(x)}{\eta} + \eta V_{t+1} = \frac{F_{t+1}(x)}{\eta} + \eta V_t + \eta \|g_t\|^2 \geq \min_{\eta^* \leq 1/h_t}\left[\frac{F_{t+1}(x)}{\eta^*} + \eta^* V_t\right] + \eta \|g_t\|^2$, which yields

$$
\begin{aligned}
&\geq k\|g_t\|^2 \min\left\{\sqrt{\frac{F_{t+1}(x)}{V_{t+1}}}, \frac{1}{h_t}\right\} + k \min_{\eta \leq \frac{1}{h_t}}\left[\frac{F_{t+1}(x)}{\eta} + \eta V_t\right] - k \min_{\eta \leq \frac{1}{h_t}}\left[\frac{F_t(x)}{\eta} + \eta V_t\right] \\
&\overset{(a)}{\geq} k\|g_t\|^2 \min\left\{\sqrt{\frac{F_t(x)}{V_{t+1}}}, \frac{1}{h_t}\right\} + k \min_{\eta \leq \frac{1}{h_t}}\left[\frac{F_t(x)}{\eta} + \eta V_t\right] - k \min_{\eta \leq \frac{1}{h_t}}\left[\frac{F_t(x)}{\eta} + \eta V_t\right] \\
&= k\|g_t\|^2 \min\left\{\sqrt{\frac{F_t(x)}{V_{t+1}}}, \frac{1}{h_t}\right\} \overset{(b)}{\geq} \frac{k}{\sqrt{2}}\|g_t\|^2 \min\left\{\sqrt{\frac{F_t(x)}{V_t}}, \frac{1}{h_t}\right\} \\
&\geq \|g_t\|^2 \min\left\{\sqrt{\frac{F_t(x)}{V_t}}, \frac{1}{h_t}\right\} = \eta_t'(x)\|g_t\|^2,
\end{aligned}
$$

where (a) uses that $\alpha_{t+1} \leq \alpha_t$ so $F_{t+1}(x) = \log(x/\alpha_{t+1} + 1) \geq \log(x/\alpha_t + 1) = F_t(x)$, and (b) uses $\frac{1}{V_t} = \frac{1}{V_{t+1}}\frac{V_{t+1}}{V_t} \leq \frac{2}{V_{t+1}}$. From this we immediately have

$$
\Delta_t(w) = \int_0^{\|w\|} \Psi_{t+1}'(x) - \Psi_t'(x)dx \geq \|g_t\|^2 \int_0^{\|w\|} \eta_t'(x)dx = \eta_t(\|w\|)\|g_t\|^2,
$$

so combining this with Equation (6), we have

$$
\begin{aligned}
\delta_t &= \langle g_t, w_t - w_{t+1}\rangle - D_{\psi_t}(w_{t+1}|w_t) - \Delta_t(w_{t+1}) \\
&\leq \langle g_t, w_t - w_{t+1}\rangle - D_{\psi_t}(w_{t+1}|w_t) - \eta_t(\|w_{t+1}\|)\|g_t\|^2 \\
&= \widehat{\delta}_t \leq \frac{2\|g_t\|^2}{\Psi_t''(x_0)} = \frac{2\alpha_t e\|g_t\|^2}{k\sqrt{V_t}} \leq \frac{2\alpha_t\|g_t\|^2}{\sqrt{V_t}}
\end{aligned}
$$

for $k \geq 3$. Returning to our regret bound, we have

$$
\begin{aligned}
R_T(u) &\leq \psi_{T+1}(u) + \sum_{t=1}^T \delta_t \leq \psi_{T+1}(u) + 2\sum_{t=1}^T \frac{\alpha_t\|g_t\|^2}{\sqrt{V_t}} \\
&\overset{(a)}{\leq} \|u\|\Psi_{T+1}'(\|u\|) + 2\sum_{t=1}^T \frac{\alpha_t\|g_t\|^2}{\sqrt{V_t}} \\
&\overset{(b)}{\leq} 2k\|u\| \max\left\{\sqrt{V_{T+1}\log(\|u\|/\alpha_{T+1} + 1)}, h_{T+1}\log(\|u\|/\alpha_{T+1} + 1)\right\} \\
&\quad + 2\sum_{t=1}^T \frac{\alpha_t\|g_t\|^2}{\sqrt{V_t}} \qquad (7)
\end{aligned}
$$

where (a) observes that $\Psi_t'(x)$ is increasing in x, so

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

and (b) observes that the case $\Psi'_t(x) = kh_tF_t(x) + \frac{kV_t}{h_t}$ coincides with $\frac{V_t}{h_t} \le h_tF_t(x)$, so

$$\begin{aligned}\Psi'_{T+1}(\|u\|) &= \begin{cases} 2k\sqrt{V_{T+1}F_{T+1}(\|u\|)} & \text{if } h_{T+1}\sqrt{F_{T+1}(\|u\|)} \le \sqrt{V_{T+1}} \\ kh_{T+1}F_{T+1}(\|u\|) + \frac{kV_{T+1}}{h_{T+1}} & \text{otherwise} \end{cases} \\ &\le \begin{cases} 2k\sqrt{V_{T+1}F_{T+1}(\|u\|)} & \text{if } h_{T+1}\sqrt{F_{T+1}(\|u\|)} \le \sqrt{V_{T+1}} \\ 2kh_{T+1}F_{T+1}(\|u\|) & \text{otherwise} \end{cases} \\ &= 2k\max\left\{\sqrt{V_{T+1}F_{T+1}(\|u\|)}, h_{T+1}F_{T+1}(\|u\|)\right\}.\end{aligned}$$

Note that the regret does not depend on g_{T+1}, so without loss of generality we can assume $g_{T+1} = g_T$ and hence $h_{T+1} = h_T$. Finally, Lemma 5 bounds $2\sum_{t=1}^T \frac{\alpha_t\|g_t\|^2}{\sqrt{V_t}} \le 4\epsilon h_T$, so plugging this into Equation (7) yields the stated result. Notice that Lemma 5 is responsible for removing the "range-ratio" problem addressed by Mhammedi and Koolen (2020) via a doubling-like restart scheme. ■

Lemma 5 *Let $c \ge 4$, $V_t = ch_t + \|g\|^2_{1:t-1}$, $B_t = c\sum_{s=1}^t \left(4 + \sum_{s'=1}^{s-1} \frac{\|g_{s'}\|^2}{h^2_{s'}}\right)$ and set $\alpha_t = \frac{\epsilon}{\sqrt{B_t}\log^2(B_t)}$. Then*

$$\sum_{t=1}^T \frac{\alpha_t\|g_t\|^2}{\sqrt{V_t}} \le 2\epsilon h_T.$$

Proof Define $\tau_1 = 1$ and $\tau_t = \max\left\{t' : t' \le t \text{ and } \sum_{s=1}^{t'-1} \frac{\|g_s\|^2}{h_s^2} + 4 < \frac{h^2_{t'}}{h^2_{\tau_{t'-1}}}\right\}$ for $t > 1$. Then, we partition $[1, T]$ into the disjoint intervals $[1, T] = \mathcal{I}_1 \cup \ldots \cup \mathcal{I}_N$ over which τ_t is fixed. Denote $\mathcal{I}_j = [\widetilde{\tau}_j, \widetilde{\tau}_{j+1} - 1]$ where $\widetilde{\tau}_1 = 1$, $\widetilde{\tau}_{N+1} = T + 1$, and $\widetilde{\tau}_j = \min\{t > \widetilde{\tau}_{j-1} : \tau_t > \tau_{t-1}\}$ for $j \in [2, N]$. Observe that by definition, $\tau_t = \widetilde{\tau}_j$ for all $t \in \mathcal{I}_j$. Further, for all j and $t \in \mathcal{I}_j$, we have either $t = \widetilde{\tau}_j$ or $\tau_{t-1} = \widetilde{\tau}_j < t$, so that:

$$\frac{h_t^2}{h^2_{\widetilde{\tau}_j}} \le 4 + \sum_{s=1}^{t-1} \frac{\|g_s\|^2}{h_s^2}$$

Now, we show that $V_{t+1}/h_{\tau^2_{t+1}} \le B_{t+1}$. Notice that if t is the last round of an interval $\mathcal{I}_k$, then $t+1$ would be the start of the next epoch so $h_{\tau_{t+1}} = h_{t+1}$ and $V_{t+1}/h^2_{\tau_{t+1}} = V_{t+1}/h^2_{t+1} \le B_{t+1}$ (since $c \ge 1$). Otherwise, $t+1$ occurs before the end of interval $\mathcal{I}_k$ so

$$\begin{aligned}\frac{V_{t+1}}{h^2_{\tau_{t+1}}} &= \frac{ch^2_{t+1} + \|g\|^2_{1:t}}{h^2_{\widetilde{\tau}_k}} \le c\frac{h^2_{t+1}}{h^2_{\widetilde{\tau}_k}} + \sum_{j=1}^k \sum_{\substack{s\in\mathcal{I}_j \\ s\le t}} \frac{\|g_s\|^2}{h^2_{\widetilde{\tau}_j}} \\ &\le c\frac{h^2_{t+1}}{h^2_{\widetilde{\tau}_k}} + \sum_{j=1}^k \sum_{\substack{s\in\mathcal{I}_j \\ s\le t}} \frac{h_s^2}{h^2_{\widetilde{\tau}_j}}\end{aligned}$$

Now, apply the definition of $\widetilde{\tau}_j$ to get:

$$\leq c\left(4+\sum_{s=1}^{t}\frac{\|g_s\|^2}{h_s^2}\right)+\sum_{j=1}^{k}\sum_{\substack{s\in\mathcal{I}_j\\ s\leq t}}\left(4+\sum_{s'=1}^{s-1}\frac{\|g_{s'}\|^2}{h_{s'}^2}\right)$$
$$\leq c\left(4+\sum_{s=1}^{t}\frac{\|g_s\|^2}{h_s^2}\right)+\sum_{s=1}^{t}\left(4+\sum_{s'=1}^{s-1}\frac{\|g_{s'}\|^2}{h_{s'}^2}\right)$$
$$\leq c\sum_{s=1}^{t+1}\left(4+\sum_{s'=1}^{s-1}\frac{\|g_{s'}\|^2}{h_{s'}^2}\right)=B_{t+1}.$$

Now, using this we have that $\alpha_t=\frac{\varepsilon}{\sqrt{B_t}\log^2(B_t)}\leq\frac{\varepsilon h_{\tau_t}}{\sqrt{V_t}\log^2\left(V_t/h_{\tau_t^2}\right)}$ and thus

$$\sum_{t=1}^{T}\frac{\alpha_t\|g_t\|^2}{\sqrt{V_t}}=\sum_{j=1}^{N}\sum_{t\in\mathcal{I}_j}\frac{\alpha_t\|g_t\|^2}{\sqrt{V_t}}=\epsilon\sum_{j=1}^{N}\sum_{t\in\mathcal{I}_j}\frac{\|g_t\|^2}{\sqrt{V_t}\sqrt{B_t}\log^2(B_t)}\leq\epsilon\sum_{j=1}^{N}\sum_{t\in\mathcal{I}_j}h_{\tau_t}\frac{\|g_t\|^2}{V_t\log^2\left(V_t/h_{\tau_t}^2\right)}$$
$$=\epsilon\sum_{j=1}^{N}h_{\widetilde{\tau}_j}\sum_{t\in\mathcal{I}_j}\frac{\|g_t\|^2}{\left(ch_t^2+\|g\|_{1:t-1}^2\right)\log^2\left(\frac{ch_t^2+\|g\|_{1:t-1}^2}{h_{\tau_t}^2}\right)}$$
$$\leq\epsilon\sum_{j=1}^{N}h_{\widetilde{\tau}_j}\sum_{t\in\mathcal{I}_j}\frac{\|g_t\|^2}{\left((c-1)h_{\widetilde{\tau}_j}^2+\|g\|_{1:t}^2\right)\log^2\left(\frac{(c-1)h_{\widetilde{\tau}_j}^2+\|g\|_{1:t}}{h_{\widetilde{\tau}_j}^2}\right)}$$
$$\leq\epsilon\sum_{j=1}^{N}h_{\widetilde{\tau}_j}\int_{(c-1)h_{\widetilde{\tau}_j}^2}^{(c-1)h_{\widetilde{\tau}_j}^2+\|g\|_{1:t}^2}\frac{1}{x\log^2(x/h_{\widetilde{\tau}_j}^2)}dx$$
$$=\epsilon\sum_{j=1}^{N}h_{\widetilde{\tau}_j}\frac{-1}{\log\left(x/h_{\widetilde{\tau}_j}^2\right)}\Bigg|_{x=(c-1)h_{\widetilde{\tau}_j}^2}^{(c-1)h_{\widetilde{\tau}_j}^2+\|g\|_{1:t}^2}\leq\frac{\epsilon}{\log(c-1)}\sum_{j=1}^{N}h_{\widetilde{\tau}_j}.$$

Notice that each interval begins when $\frac{h_{\widetilde{\tau}_j}^2}{h_{\widetilde{\tau}_{j-1}}^2}>\sum_{s=1}^{t-1}\frac{\|g_s\|^2}{h_s^2}+4>4$, so $h_{\widetilde{\tau}_j}>2h_{\widetilde{\tau}_{j-1}}$ and hence

$$\frac{\epsilon}{\log(c-1)}\sum_{j=1}^{N}h_{\widetilde{\tau}_j}\leq\frac{\epsilon}{\log(c-1)}\sum_{j=0}^{N-1}\frac{1}{2^i}h_{\widetilde{\tau}_N}\leq\frac{2\epsilon h_T}{\log(c-1)}\leq 2\epsilon h_T,$$

for $c>4$. ■

Appendix H. Supporting Lemmas

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

Lemma 6 *(Orabona and Pál, 2021, Lemma 23) Let $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}$ be defined as $g(x) = f(\|x\|)$. If f is twice differentiable at $\|x\|$ and $\|x\| > 0$ then*

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

Lemma 7 *Under the same assumptions as Lemma 6, further suppose that $f'(x)$ is concave and non-negative. If f is twice-differentiable at $\|x\|$ and $\|x\| > 0$, then*

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

Proof Apply Lemma 6,

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

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

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

■

The following integral bound is standard and included for completeness.

Lemma 8 *Let $V_t \geq \sum_{s=1}^{t} \|g_s\|^2$. Then*

$$\sum_{t=1}^{T} \frac{\|g_t\|^2}{\sqrt{V_t}} \leq 2\sqrt{\sum_{t=1}^{T} \|g_t\|^2}$$

Proof Using the well-known integral bound $\sum_{t=1}^{T} a_t f(\sum_{s=0}^{t} a_s) \leq \int_{a_0}^{\sum_{s=0}^{t} a_s} f(x)dx$ for non-increasing f (See *e.g.* Orabona (2019, Lemma 4.13)), we have

$$\sum_{t=1}^{T} \frac{\|g_t\|^2}{\sqrt{V_t}} \leq \sum_{t=1}^{T} \frac{\|g_t\|^2}{\sqrt{\|g\|_{1:t}^2}} \leq \int_0^{\|g\|_{1:T}^2} \frac{1}{\sqrt{x}} dx = 2\sqrt{\|g\|_{1:T}^2}.$$

■

Lemma 9 *Let $V_t \geq 4G^2 + \sum_{s=1}^{t-1} \|g_s\|^2$ where $G \geq \|g_t\|$ for all t. Then*

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

Proof Let $c \geq 4$ and $V_t = cG^2 + \|g\|^2_{1:t-1}$. As in Lemma 8, we apply the integral bound $\sum_{t=1}^T a_t f(\sum_{i=0}^t a_t) \leq \int_{a_0}^{\sum_{s=0}^t a_s} f(x)dx$ for non-increasing f to get

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

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

■

Appendix I. A Simple Reduction for Dynamic Regret in Unbounded Domains

Interestingly, a dynamic regret bound of $R_T(\boldsymbol{u}) \leq \widetilde{O}\left(\sqrt{(M^2 + MP_T)\,\|g\|_{1:T}^2}\right)$ can be achieved very simply using a generalization of the one-dimensional reduction of Cutkosky and Orabona (2018) to dynamic regret. Note however, that this approach fails to achieve the improved per-comparator adaptivity observed in Section 4. The following lemma shows that achieving the $R_T(\boldsymbol{u}) \leq \widetilde{O}\left(\sqrt{(M^2 + MP_T)\,\|g\|_{1:T}^2}\right)$ bound in an unconstrained domain is essentially no harder than achieving it in a bounded domain, so long as one has access to an algorithm guaranteeing parameter-free *static* regret.

Algorithm 6: One-dimensional Reduction (Cutkosky and Orabona, 2018)

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

Lemma 10 *Suppose that $\mathcal{A}_S$ guarantees dynamic regret $R_T^{\mathcal{A}_S}(\boldsymbol{u})$ for any sequence $u_1, \ldots, u_T$ in the unit-ball $S = \left\{w \in \mathbb{R}^d : \|w\| \leq 1\right\}$ and suppose $\mathcal{A}_{1D}$ obtains static regret $R_T^{\mathcal{A}_{1D}}(u)$ for any $u \in \mathbb{R}$. Then for any $u_1, \ldots, u_T$ in $\mathbb{R}^d$, Algorithm 6 guarantees*

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

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

Proof the proof follows the same reasoning as in the static regret case (Cutkosky and Orabona, 2018, Theorem 2):

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

Algorithm 7: Lazy Reduction for Amortized Computation

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

■

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

Appendix J. Amortized Computation for Dynamic Regret

All known algorithms which achieve the optimal $O(\sqrt{TP_T})$ dynamic regret follow a similar construction, in which several instances of a simple base algorithm $\mathcal{A}$ are run simultaneously and a meta-algorithm combines their outputs in a way that guarantees near-optimal performance. Assuming the base algorithm $\mathcal{A}$ uses $O(d)$ computation per round, the full algorithm then requires $O(d \log(T))$ computation per round. Ideally we'd like to avoid this $\log(T)$ overhead. A simple way to combat this difficulty is to only update the algorithm every $O(\log(T))$ rounds, so that the *amortized* computation per round is $O(d)$ on average. The following proposition shows that $R_T(\boldsymbol{u}) \leq O\left(\sqrt{TP_T}\right)$ can be maintained up to poly-logarithmic terms using only $O(d)$ per-round computation on average by updating only at the end of itervals I_k of length $O(\log(T))$.

Proposition 3 *Suppose $\mathcal{A}$ is an online learning algorithm which guarantees*

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

for all $u_1, \ldots, u_T$ in $\mathbb{R}^d$ with $\max_{t \leq T} \|u_t\| \leq M$. Then for all $u_1, \ldots, u_T$ in $\mathbb{R}^d$, Algorithm 7 guarantees

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

Proof First observe that for any interval $I = [a, b]$, we have

$$\sum_{t\in I} \langle g_t, w_t - u_t\rangle = \sum_{t\in I} \langle g_t, w_t - u_b\rangle + \sum_{t\in I} \langle g_t, u_b - u_t\rangle\,,$$

and bound the second sum as

$$\begin{aligned}
\sum_{t\in I} \left\langle g_t, \sum_{s=t+1}^{b} u_s - u_{s-1}\right\rangle &= \sum_{t=a}^{b}\sum_{s=t+1}^{b} \langle g_t, u_s - u_{s-1}\rangle \\
&= \sum_{s=a+1}^{b} \langle g_{a:s-1}, u_s - u_{s-1}\rangle \le \sqrt{\sum_{s=a+1}^{b} \|g_{a:s-1}\|^2 \sum_{t=a+1}^{b} \|u_t - u_{t-1}\|^2} \\
&\le \sqrt{\left(\sum_{t=a+1}^{b} \|g_t\|^2 + \sum_{t=a+1}^{b}\sum_{t'\neq t}^{b} \|g_t\|\,\|g_{t'}\|\right) S_I} \\
&\le \sqrt{\left(\sum_{t=a+1}^{b} \|g_t\|^2 + \max_{s\in[a,b]} \|g_s\|^2\,|I|^2\right) S_I} \\
&\le \sqrt{2\,\|g\|_{a+1:b}^2\,|I|^2 S_I} = \sqrt{2\,\|g\|_{a+1:b}^2\, S_I}|I|.
\end{aligned}$$

where $S_I = \sum_{t=a+1}^{b} \|u_t - u_{t-1}\|^2$. Thus, denoting $I_1 = [1, \tau_1]$, $I_2 = [\tau_1 + 1, \tau_2], \ldots, I_K = [\tau_{K-1} + 1, \tau_K]$, we can bound

$$\begin{aligned}
\sum_{t=1}^{T} \langle g_t, w_t - u_t\rangle &= \sum_{k=1}^{K}\sum_{t\in I_k} \langle g_t, w_t - u_t\rangle = \sum_{k=1}^{K}\sum_{t\in I_k} \langle g_t, w_t - u_{\tau_k}\rangle + \langle g_t, u_{\tau_k} - u_t\rangle \\
&\le \sum_{k=1}^{K}\sum_{t\in I_k} \langle g_t, w_t - u_{\tau_k}\rangle + \sum_{k=1}^{K} \sqrt{2S_{I_k}\,\|g\|_{t\in I_k}^2}|I_k| \\
&\le \sum_{k=1}^{K} \left\langle \sum_{t\in I_k} g_t, w_{\tau_k} - u_{\tau_k}\right\rangle + \sqrt{2S_T\,\|g\|_{1:T}^2}\max_{k\le K} |I_k|
\end{aligned}$$

where the last line observes that w_t is fixed within each interval. From the regret guarantee of algorithm $\mathcal{A}$ we have

$$\begin{aligned}
\sum_{k=1}^{K} \left\langle \sum_{t\in I_k} g_t, w_{\tau_k} - u_{\tau_k}\right\rangle &= \sum_{k=1}^{K} \langle \widetilde{g}_{\tau_k}, w_{\tau_k} - u_{\tau_k}\rangle \le \widetilde{O}\left(\sqrt{(M^2 + M\widehat{P}_K)\,\|\widetilde{g}\|_{1:K}^2}\right) \\
&\le \widetilde{O}\left(\max_{k\le K} |I_k|\sqrt{2(M^2 + MP_T)\,\|g\|_{1:T}^2}\right),
\end{aligned}$$

where the first line defines $\widehat{P}_K = \sum_{k=2}^{K} \left\|u_{\tau_k} - u_{\tau_{k-1}}\right\|$ and the last line observes $\widehat{P}_K \le P_T$. Hence,

$$\sum_{t=1}^{T} \langle g_t, w_t - u_t\rangle \le \widetilde{O}\left(\max_{k\le K} |I_k| \left(\sqrt{2(M^2 + MP_T)\,\|g\|_{1:T}^2} + \sqrt{2S_T\,\|g\|_{1:T}^2}\right)\right).$$

The stated bound follows by observing that $S_T \leq MP_T \leq M^2 + MP_T$ and hiding constants. ■