Theoretical Computer Science 716 (2018) 50–69

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Scale-free online learning ☆

Francesco Orabona [a,1], Dávid Pál [b,*]

[a] Stony Brook University, Stony Brook, NY 11794, USA
[b] Yahoo Research, 11th Floor, 229 West 43rd Street, New York, NY 10036, USA

ARTICLE INFO

Article history:
Available online 1 December 2017

Keywords:
Online algorithms
Optimization
Regret bounds
Online learning

ABSTRACT

We design and analyze algorithms for online linear optimization that have optimal regret and at the same time do not need to know any upper or lower bounds on the norm of the loss vectors. Our algorithms are instances of the Follow the Regularized Leader (FTRL) and Mirror Descent (MD) meta-algorithms. We achieve adaptiveness to the norms of the loss vectors by scale invariance, i.e., our algorithms make exactly the same decisions if the sequence of loss vectors is multiplied by any positive constant. The algorithm based on FTRL works for any decision set, bounded or unbounded. For unbounded decisions sets, this is the first adaptive algorithm for online linear optimization with a non-vacuous regret bound. In contrast, we show lower bounds on scale-free algorithms based on MD on unbounded domains.

1. Introduction

Online Linear Optimization (OLO) is a problem where an algorithm repeatedly chooses a point w_t from a convex decision set K, observes an arbitrary, or even adversarially chosen, loss vector ℓ_t and suffers the loss $\langle \ell_t, w_t \rangle$. The goal of the algorithm is to have a small cumulative loss. The performance of an algorithm is evaluated by the so-called regret, which is the difference of the cumulative losses of the algorithm and of the (hypothetical) strategy that would choose in every round the same best point in hindsight.

OLO is a fundamental problem in machine learning [2–4]. Many learning problems can be directly phrased as OLO, e.g., learning with expert advice [5–8] and online combinatorial optimization [9–11]. Other problems can be reduced to OLO, e.g., online convex optimization [12], [4, Chapter 2], online classification [13,14] and regression [15], [2, Chapters 11 and 12], multi-armed bandits problems [2, Chapter 6], [16,17], and batch and stochastic optimization of convex functions [18,19]. Hence, a result in OLO immediately implies other results in all these domains.

The adversarial choice of the loss vectors received by the algorithm is what makes the OLO problem challenging. In particular, if an OLO algorithm commits to an upper bound on the norm of future loss vectors, its regret can be made arbitrarily large through an adversarial strategy that produces loss vectors with norms that exceed the upper bound.

For this reason, most of the existing OLO algorithms receive as an input—or implicitly assume—an upper bound B on the norm of the loss vectors. The input B is often disguised as the learning rate, the regularization parameter, or the parameter of strong convexity of the regularizer. However, these algorithms have two obvious drawbacks.

☆ A preliminary version of this paper was presented at ALT 2015.

* Corresponding author.
E-mail addresses: francesco@orabona.com (F. Orabona), dpal@yahoo-inc.com (D. Pál).

[1] Work done while at Yahoo Research.

https://doi.org/10.1016/j.tcs.2017.11.021
0304-3975/© 2017 Published by Elsevier B.V.

Table 1
Selected results for OLO. Best results in each column are in bold.

Algorithm	Decisions set(s)	Regularizer(s)	Scale-free
HEDGE [7]	Probability Simplex	Negative Entropy	No
GIGA [20]	Any Bounded	$\frac{1}{2}\|w\|_2^2$	No
RDA [21]	**Any**	**Any Strongly Convex**	No
FTRL-PROXIMAL [22,23]	Any Bounded	$\frac{1}{2}\|w\|_2^2+$ any convex func. [a]	**Yes**
ADAGRAD MD [24]	Any Bounded	$\frac{1}{2}\|w\|_2^2+$ any convex func.	**Yes**
ADAGRAD FTRL [24]	**Any**	$\frac{1}{2}\|w\|_2^2+$ any convex func.	No
ADAHEDGE [25]	Probability Simplex	Negative Entropy	**Yes**
NAG [26]	$\{u : \max_t \langle \ell_t, u\rangle \le C\}$	$\frac{1}{2}\|w\|_2^2$	N/A[b]
SCALE INVARIANT ALGORITHMS [27]	**Any**	$\frac{1}{2}\|w\|_p^2+$ any convex func. $1 < p \le 2$	N/A[b]
SCALE-FREE MD **[this paper]**	$\sup_{u,v\in K} \mathcal{B}_f(u,v) < \infty$	**Any Strongly Convex**	**Yes**
SOLO FTRL **[this paper]**	**Any**	**Any Strongly Convex**	**Yes**

[a] Even if, in principle the FTRL-Proximal algorithm can be used with any proximal regularizer, to the best of our knowledge a general way to construct proximal regularizers is not known. The only proximal regularizer we are aware is based on the 2-norm.
[b] These algorithms attempt to produce an invariant sequence of predictions $\langle w_t, \ell_t\rangle$, rather than a sequence of invariant w_t.

First, they do not come with any regret guarantee for sequences of loss vectors with norms exceeding B. Second, on sequences of loss vectors with norms bounded by $b \ll B$, these algorithms fail to have an optimal regret guarantee that depends on b rather than on B.

There is a clear practical need to design algorithms that adapt automatically to the norms of the loss vectors. A natural, yet overlooked, design method to achieve this type of adaptivity is by insisting to have a **scale-free** algorithm. That is, with the same parameters, the sequence of decisions of the algorithm does not change if the sequence of loss vectors is multiplied by a positive constant. The most important property of scale-free algorithms is that both their loss and their regret scale linearly with the maximum norm of the loss vector appearing in the sequence.

1.1. Previous results

The majority of the existing algorithms for OLO are based on two generic algorithms: FOLLOW THE REGULARIZER LEADER (FTRL) and MIRROR DESCENT (MD). FTRL dates back to the potential-based forecaster in [2, Chapter 11] and its theory was developed in [28]. The name FOLLOW THE REGULARIZED LEADER comes from [16]. Independently, the same algorithm was proposed in [29] for convex optimization under the name DUAL AVERAGING and rediscovered in [21] for online convex optimization. Time-varying regularizers were analyzed in [24] and the analysis tightened in [27]. MD was originally proposed in [18] and later analyzed in [30] for convex optimization. In the online learning literature it makes its first appearance, with a different name, in [15].

Both FTRL and MD are parametrized by a function called a *regularizer*. Based on different regularizers different algorithms with different properties can be instantiated. A summary of algorithms for OLO is presented in Table 1. All of them are instances of FTRL or MD.

Scale-free versions of MD include ADAGRAD MD [24]. However, the ADAGRAD MD algorithm has a non-trivial regret bounds only when the Bregman divergence associated with the regularizer is bounded. In particular, since a bound on the Bregman divergence implies that the decision set is bounded, the regret bound for ADAGRAD MD is vacuous for unbounded sets. In fact, as we show in Section 4.1, ADAGRAD MD and similar algorithms based on MD incur $\Omega(T)$ regret, in the worst case, if the Bregman divergence is not bounded.

Only one scale-free algorithm based on FTRL was known. It is the ADAHEDGE [25] algorithm for learning with expert advice, where the decision set is bounded. An algorithm based on FTRL that is "almost" scale-free is ADAGRAD FTRL [24]. This algorithm fails to be scale-free due to "off-by-one" issue; see [23] and the discussion in Section 3. Instead, FTRL-PROXIMAL [22,23] solves the off-by-one issue, but it requires proximal regularizers. In general, proximal regularizers do not have a simple form and even the simple 2-norm case requires bounded domains to achieve non-vacuous regret.

For unbounded decision sets no scale-free algorithm with a non-trivial regret bound was known. Unbounded decision sets are practically important (see, e.g., [31]), since learning of large-scale linear models (e.g., logistic regression) is done by gradient methods that can be reduced to OLO with decision set $\mathbb{R}^d$.

1.2. Overview of the results

We design and analyze two scale-free algorithms: SOLO FTRL and SCALE-FREE MD. A third one, ADAFTRL, is presented in the Appendix. SOLO FTRL and ADAFTRL are based on FTRL. ADAFTRL is a generalization of ADAHEDGE [25] to arbitrary strongly convex regularizers. SOLO FTRL can be viewed as the "correct" scale-free version of the diagonal version of ADAGRAD

FTRL [24] generalized to arbitrary strongly convex regularizers. SCALE-FREE MD is based on MD. It is a generalization of ADAGRAD MD [24] to arbitrary strongly convex regularizers. The three algorithms are presented in Sections 3 and 4, and Appendix B, respectively.

We prove that the regret of SOLO FTRL and ADAFTRL on bounded domains after T rounds is bounded by $O(\sqrt{\sup_{v\in K} f(v) \sum_{t=1}^T \|\ell_t\|_*^2})$ where f is a non-negative regularizer that is 1-strongly convex with respect to a norm $\|\cdot\|$ and $\|\cdot\|_*$ is its dual norm. For SCALE-FREE MD, we prove $O(\sqrt{\sup_{u,v\in K} B_f(u,v) \sum_{t=1}^T \|\ell_t\|_*^2})$ where B_f is the Bregman divergence associated with a 1-strongly convex regularizer f. In Section 5, we show that the $\sqrt{\sum_{t=1}^T \|\ell_t\|_*^2}$ term in the bounds is necessary by proving a $\frac{D}{\sqrt{8}}\sqrt{\sum_{t=1}^T \|\ell_t\|_*^2}$ lower bound on the regret of any algorithm for OLO for any decision set with diameter D with respect to the primal norm $\|\cdot\|$.

For SOLO FTRL, we prove that the regret against a competitor $u \in K$ is at most $O(f(u)\sqrt{\sum_{t=1}^T \|\ell_t\|_*^2} + \max_{t=1,2,\dots,T} \|\ell_t\|_* \sqrt{T})$. As before, f is a non-negative 1-strongly convex regularizer. This bound is non-trivial for any decision set, bounded or unbounded. The result makes SOLO FTRL the **first adaptive algorithm for unbounded decision sets** with a non-trivial regret bound.

All three algorithms are **any-time**, i.e., they do not need to know the number of rounds, T, in advance and the regret bounds hold for all T simultaneously.

Our proof techniques rely on new homogeneous inequalities (Lemmas 3, 7) which might be of independent interest.

Finally, in Section 4.1, we show negative results for existing popular variants of MD. We show two examples of decision sets and sequences of loss vectors of unit norm on which these variants of MD have $\Omega(T)$ regret. These results indicate that FTRL is superior to MD in a worst-case sense.

Preliminary version of the results were presented at ALT 2015 [1].

2. Notation and preliminaries

Let V be a finite-dimensional[2] real vector space equipped with a norm $\|\cdot\|$. We denote by V^* its dual vector space. The bi-linear map associated with (V^*, V) is denoted by $\langle\cdot,\cdot\rangle : V^* \times V \to \mathbb{R}$. The dual norm of $\|\cdot\|$ is $\|\cdot\|_*$.

In OLO, in each round $t = 1, 2, \dots,$ the algorithm chooses a point w_t in the decision set $K \subseteq V$ and then the algorithm observes a loss vector $\ell_t \in V^*$. The instantaneous loss of the algorithm in round t is $\langle \ell_t, w_t\rangle$. The cumulative loss of the algorithm after T rounds is $\sum_{t=1}^T \langle \ell_t, w_t\rangle$. The regret of the algorithm with respect to a point $u \in K$ is

$$\mathrm{Regret}_T(u) = \sum_{t=1}^T \langle \ell_t, w_t\rangle - \sum_{t=1}^T \langle \ell_t, u\rangle,$$

and the regret with respect to the best point is $\mathrm{Regret}_T = \sup_{u\in K} \mathrm{Regret}_T(u)$. We assume that K is a non-empty closed convex subset of V. Sometimes we will assume that K is also bounded. We denote by D its diameter with respect to $\|\cdot\|$, i.e., $D = \sup_{u,v\in K} \|u - v\|$. If K is unbounded, $D = +\infty$.

2.1. Convex analysis

The *Bregman divergence* of a convex differentiable function f is defined as $\mathcal{B}_f(u, v) = f(u) - f(v) - \langle \nabla f(v), u - v\rangle$. Note that $\mathcal{B}_f(u, v) \ge 0$ for any u, v which follows directly from the definition of convexity of f.

The *Fenchel conjugate* of a function $f : K \to \mathbb{R}$ is the function $f^* : V^* \to \mathbb{R} \cup \{+\infty\}$ defined as $f^*(\ell) = \sup_{w\in K}(\langle \ell, w\rangle - f(w))$. The Fenchel conjugate of any function is convex (since it is a supremum of affine functions) and satisfies the *Fenchel–Young inequality*

$$\forall w \in K,\ \forall \ell \in V^* \qquad f(w) + f^*(\ell) \ge \langle \ell, w\rangle\ .$$

Monotonicity of Fenchel conjugates follows easily from the definition: If $f, g : K \to \mathbb{R}$ satisfy $f(w) \le g(w)$ for all $w \in K$ then $f^*(\ell) \ge g^*(\ell)$ for every $\ell \in V^*$.

Given $\lambda > 0$, a function $f : K \to \mathbb{R}$ is called λ*-strongly convex* with respect to a norm $\|\cdot\|$ if and only if, for all $x, y \in K$,

$$f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle + \frac{\lambda}{2}\|x - y\|^2\ ,$$

where $\nabla f(x)$ is any subgradient of f at the point x.

The following proposition relates the range of values of a strongly convex function to the diameter of its domain. The proof can be found in Appendix A.

[2] Many, but not all, of our results can be extended to more general normed vector spaces.

Proposition 1 *(Diameter vs. range). Let $K \subseteq V$ be a non-empty bounded closed convex set. Let $D = \sup_{u,v \in K} \|u - v\|$ be its diameter with respect to $\|\cdot\|$. Let $f : K \to \mathbb{R}$ be a non-negative lower semi-continuous function that is 1-strongly convex with respect to $\|\cdot\|$. Then, $D \le \sqrt{8 \sup_{v \in K} f(v)}$.*

Fenchel conjugates and strongly convex functions have certain nice properties, which we list in Proposition 2 below.

Proposition 2 *(Fenchel conjugates of strongly convex functions). Let $K \subseteq V$ be a non-empty closed convex set with diameter $D := \sup_{u,v \in K} \|u - v\|$. Let $\lambda > 0$, and let $f : K \to \mathbb{R}$ be a lower semi-continuous function that is λ-strongly convex with respect to $\|\cdot\|$. The Fenchel conjugate of f satisfies:*

1. *f^* is finite everywhere and differentiable everywhere.*
2. *For any $\ell \in V^*$, $\nabla f^*(\ell) = \operatorname{argmin}_{w \in K} (f(w) - \langle \ell, w \rangle)$.*
3. *For any $\ell \in V^*$, $f^*(\ell) + f(\nabla f^*(\ell)) = \langle \ell, \nabla f^*(\ell) \rangle$.*
4. *f^* is $\frac{1}{\lambda}$-strongly smooth, i.e., for any $x, y \in V^*$, $\mathcal{B}_{f^*}(x, y) \le \frac{1}{2\lambda} \|x - y\|_*^2$.*
5. *f^* has $\frac{1}{\lambda}$-Lipschitz continuous gradients, i.e., for any $x, y \in V^*$, $\|\nabla f^*(x) - \nabla f^*(y)\| \le \frac{1}{\lambda} \|x - y\|_*$.*
6. *$\mathcal{B}_{f^*}(x, y) \le D \|x - y\|_*$ for any $x, y \in V^*$.*
7. *$\|\nabla f^*(x) - \nabla f^*(y)\| \le D$ for any $x, y \in V^*$.*
8. *For any $c > 0$, $(cf(\cdot))^* = cf^*(\cdot/c)$.*

Except for properties 6 and 7, the proofs can be found in [28]. Property 6 is proven in Appendix A. Property 7 trivially follows from property 2.

Algorithm 1 FTRL WITH VARYING REGULARIZER

Require: Non-empty closed convex set $K \subseteq V$
Initialize $L_0 \leftarrow 0$
for $t = 1, 2, 3, \dots$ **do**
  Choose a regularizer $R_t : K \to \mathbb{R}$
  $w_t \leftarrow \operatorname{argmin}_{w \in K} (\langle L_{t-1}, w \rangle + R_t(w))$
  Predict w_t
  Observe $\ell_t \in V^*$
  $L_t \leftarrow L_{t-1} + \ell_t$
end for

2.2. Generic FTRL with varying regularizer

Two of our scale-free algorithms are instances of FTRL with *varying regularizers*, presented as Algorithm 1. The algorithm is parametrized by a sequence $\{R_t\}_{t=1}^{\infty}$ of functions $R_t : K \to \mathbb{R}$ called *regularizers*. Each regularizer R_t can depend on the past loss vectors $\ell_1, \ell_2, \dots, \ell_{t-1}$ in an arbitrary way. The following lemma bounds its regret.

Lemma 1 *(Regret of FTRL). If the regularizers $R_1, R_2, \dots$ chosen by Algorithm 1 are strongly convex and lower semi-continuous, the algorithm's regret is upper bounded as*

$$\mathrm{Regret}_T(u) \le R_{T+1}(u) + R_1^*(0) + \sum_{t=1}^{T} \mathcal{B}_{R_t^*}(-L_t, -L_{t-1}) - R_t^*(-L_t) + R_{t+1}^*(-L_t) \, .$$

The proof of the lemma can be found in [27]. For completeness, we include it in Appendix A.

2.3. Generic mirror descent with varying regularizer

Algorithm 2 MIRROR DESCENT WITH VARYING REGULARIZER

Require: Non-empty closed convex set $K \subseteq V$
Choose a regularizer $R_0 : K \to \mathbb{R}$
$w_1 \leftarrow \operatorname{argmin}_{w \in K} R_0(w)$
for $t = 1, 2, 3, \dots$ **do**
  Predict w_t
  Observe $\ell_t \in V^*$
  Choose a regularizer $R_t : K \to \mathbb{R}$
  $w_{t+1} \leftarrow \operatorname{argmin}_{w \in K} \left(\langle \ell_t, w \rangle + \mathcal{B}_{R_t}(w, w_t)\right)$
end for

Mirror Descent (MD) is a generic algorithm similar to FTRL but quite different in the details. The algorithm is stated as Algorithm 2. The algorithm is parametrized by a sequence $\{R_t\}_{t=0}^{\infty}$ of convex functions $R_t : K \to \mathbb{R}$ called *regularizers*. Each regularizer R_t can depend on past loss vectors $\ell_1, \ell_2, \dots, \ell_t$ in an arbitrary way. If R_t is not differentiable,[3] the Bregman divergence, $\mathcal{B}_{R_t}(u, v) = R_t(u) - R_t(v) - \langle \nabla R_t(v), u - v \rangle$ needs to be defined. This is done by choosing a subgradient map $\nabla R_t : K \to V$, i.e., a function such that $\nabla R_t(w)$ is a subgradient of R_t at any point w. If R_t is a restriction of a differentiable function R'_t, it is convenient to define $\nabla R_t(w) = \nabla R'_t(w)$ for all $w \in K$. The following lemma bounds the regret of MD.

Lemma 2 *(Regret of* MD*). Algorithm 2 satisfies, for any $u \in K$,*

$$\text{Regret}_T(u) \le \sum_{t=1}^{T} \langle \ell_t, w_t - w_{t+1} \rangle - \mathcal{B}_{R_t}(w_{t+1}, w_t) + \mathcal{B}_{R_t}(u, w_t) - \mathcal{B}_{R_t}(u, w_{t+1}) \,.$$

The proof of the lemma can be found in [3,32]. For completeness, we give a proof in Appendix E.

2.4. Per-coordinate learning

An interesting class of algorithms proposed in [22] and [24] are based on so-called per-coordinate learning rates. As shown in [33], any algorithm for OLO can be used with per-coordinate learning rates as well.

Abstractly, we assume that the decision set is a Cartesian product $K = K_1 \times K_2 \times \cdots \times K_d$ of a finite number of convex sets. On each factor K_j, $j = 1, 2, \dots, d$, we can run any OLO algorithm separately and we denote by $\text{Regret}_T^{(j)}(u_j)$ its regret with respect to $u_j \in K_j$. The overall regret with respect to any $u = (u_1, u_2, \dots, u_d) \in K$ can be written as

$$\text{Regret}_T(u) = \sum_{j=1}^{d} \text{Regret}_T^{(j)}(u_j) \,.$$

If the algorithm for each factor is scale-free, the overall algorithm is clearly scale-free as well. Hence, even if not explicitly mentioned in the text, any algorithm we present can be trivially transformed to a per-coordinate version.

3. SOLO FTRL

In this section, we introduce our first scale-free algorithm; it will be based on FTRL. The closest algorithm to a scale-free FTRL in the existing literature is the AdaGrad FTRL algorithm [24]. It uses a regularizer on each coordinate of the form

$$R_t(w) = R(w) \left(\delta + \sqrt{\sum_{i=1}^{t-1} \|\ell_i\|_*^2} \right) .$$

This kind of regularizer would yield a scale-free algorithm *only* for $\delta = 0$. In fact, with this choice of δ it is easy to see that the predictions w_t in line 4 of Algorithm 1 would be independent of the scaling of the ℓ_t. Unfortunately, the regret bound in [24] becomes vacuous for such setting in the unbounded case. In fact, it requires δ to be greater than $\|\ell_t\|_*$ for all time steps t, requiring knowledge of the future (see Theorem 5 in [24]). In other words, despite of its name, AdaGrad FTRL is not fully adaptive to the norm of the gradient vectors. Similar considerations hold for the FTRL-Proximal in [22,23]: The scale-free setting of the learning rate is valid only in the bounded case.

One simple approach would be to use a doubling trick on δ in order to estimate on the fly the maximum norm of the losses. Note that a naive strategy would still fail because the initial value of δ should be data-dependent in order to have a scale-free algorithm. Moreover, we would have to upper bound the regret in all the rounds where the norm of the current loss is bigger than the estimate. Finally, the algorithm would depend on an additional parameter, the "doubling" power. Hence, even in the case one would prove a regret bound, such strategy would give the feeling that FTRL needs to be "fixed" in order to obtain a scale-free algorithm.

In the following, we propose a much simpler and better approach. We propose to use Algorithm 1 with the regularizer

$$R_t(w) = R(w) \sqrt{\sum_{i=1}^{t-1} \|\ell_i\|_*^2} \,, \tag{1}$$

[3] Note that this can happen even when R_t is a restriction of a differentiable function defined on a superset of K. If K is bounded and closed, R_t fails to be differentiable at the boundary of K. If K is a subset of an affine subspace of a dimension smaller than the dimension of V, then R_t fails to be differentiable everywhere.

where $R : K \to \mathbb{R}$ is any strongly convex function. Through a refined analysis, we show that this regularizer suffices to obtain an optimal regret bound for any decision set, bounded or unbounded. We call this variant SCALE-FREE ONLINE LINEAR OPTIMIZATION FTRL algorithm (SOLO FTRL). Our main result is Theorem 1 below, which is proven in Section 3.1.

The regularizer (1) does not uniquely define the FTRL minimizer $w_t = \operatorname{argmin}_{w \in K} R_t(w)$ when $\sqrt{\sum_{i=1}^{t-1} \|\ell_i\|_*^2}$ is zero. This happens if $\ell_1, \ell_2, \dots, \ell_{t-1}$ are all zero (and in particular for $t = 1$). In that case, we define $w_t = \operatorname{argmin}_{w \in K} R(w)$ which is consistent with $w_t = \lim_{a \to 0^+} \operatorname{argmin}_{w \in K} aR(w)$.

Theorem 1 *(Regret of SOLO FTRL). Suppose $K \subseteq V$ is a non-empty closed convex set. Let $D = \sup_{u,v \in K} \|u - v\|$ be its diameter with respect to a norm $\|\cdot\|$. Suppose that the regularizer $R : K \to \mathbb{R}$ is a non-negative lower semi-continuous function that is λ-strongly convex with respect to $\|\cdot\|$. The regret of SOLO FTRL satisfies*

$$\operatorname{Regret}_T(u) \le \left(R(u) + \frac{2.75}{\lambda} \right) \sqrt{\sum_{t=1}^{T} \|\ell_t\|_*^2} + 3.5 \min \left\{ \frac{\sqrt{T-1}}{\lambda}, D \right\} \max_{t \le T} \|\ell_t\|_* .$$

When K is unbounded, we pay a penalty that scales as $\max_{t \le T} \|\ell_t\|_* \sqrt{T}$, that has the same magnitude of the first term in the bound. On the other hand, when K is bounded, the second term is a constant and we can choose the optimal multiple of the regularizer. We choose $R(w) = \lambda f(w)$ where f is a 1-strongly convex function and optimize λ. The result of the optimization is Corollary 1.

Corollary 1 *(Regret bound for bounded decision sets). Suppose $K \subseteq V$ is a non-empty bounded closed convex set. Suppose that $f : K \to \mathbb{R}$ is a non-negative lower semi-continuous function that is 1-strongly convex with respect to $\|\cdot\|$. SOLO FTRL with regularizer*

$$R(w) = \frac{f(w)\sqrt{2.75}}{\sqrt{\sup_{v \in K} f(v)}} \quad \textit{satisfies} \quad \operatorname{Regret}_T \le 13.3 \sqrt{\sup_{v \in K} f(v) \sum_{t=1}^{T} \|\ell_t\|_*^2} .$$

Proof. Let $S = \sup_{v \in K} f(v)$. Theorem 1 applied to the regularizer $R(w) = \frac{c}{\sqrt{S}} f(w)$, together with Proposition 1 and a crude bound $\max_{t=1,2,\dots,T} \|\ell_t\|_* \le \sqrt{\sum_{t=1}^{T} \|\ell_t\|_*^2}$, give

$$\operatorname{Regret}_T \le \left(c + \frac{2.75}{c} + 3.5\sqrt{8} \right) \sqrt{S \sum_{t=1}^{T} \|\ell_t\|_*^2} .$$

We choose c by minimizing $g(c) = c + \frac{2.75}{c} + 3.5\sqrt{8}$. Clearly, $g(c)$ has minimum at $c = \sqrt{2.75}$ and has minimal value $g(\sqrt{2.75}) = 2\sqrt{2.75} + 3.5\sqrt{8} \le 13.3$. $\square$

3.1. Proof of regret bound for SOLO FTRL

The proof of Theorem 1 relies on an inequality (Lemma 3). Related and weaker inequalities, like Lemma 4, were proved in [34] and [35]. The main property of this inequality is that on the right-hand side C does *not* multiply the $\sqrt{\sum_{t=1}^{T} a_t^2}$ term.

Lemma 3 *(Useful inequality). Let $C, a_1, a_2, \dots, a_T \ge 0$. Then,*

$$\sum_{t=1}^{T} \min \left\{ \frac{a_t^2}{\sqrt{\sum_{i=1}^{t-1} a_i^2}}, C a_t \right\} \le 3.5\, C \max_{t=1,2,\dots,T} a_t \, + \, 3.5 \sqrt{\sum_{t=1}^{T} a_t^2} .$$

Proof. Without loss of generality, we can assume that $a_t > 0$ for all t. Since otherwise we can remove all $a_t = 0$ without affecting either side of the inequality. Let $M_t = \max\{a_1, a_2, \dots, a_t\}$ and $M_0 = 0$. We prove that for any $\alpha > 1$

$$\min \left\{ \frac{a_t^2}{\sqrt{\sum_{i=1}^{t-1} a_i^2}}, C a_t \right\} \le 2\sqrt{1+\alpha^2} \left(\sqrt{\sum_{i=1}^{t} a_i^2} - \sqrt{\sum_{i=1}^{t-1} a_i^2} \right) + \frac{C\alpha(M_t - M_{t-1})}{\alpha - 1}$$

from which the inequality follows by summing over $t = 1, 2, \dots, T$ and choosing $\alpha = \sqrt{2}$. The inequality follows by case analysis. If $a_t^2 \le \alpha^2 \sum_{i=1}^{t-1} a_i^2$, we have

$$\min\left\{\frac{a_t^2}{\sqrt{\sum_{i=1}^{t-1} a_i^2}}, Ca_t\right\} \le \frac{a_t^2}{\sqrt{\sum_{i=1}^{t-1} a_i^2}} = \frac{a_t^2}{\sqrt{\frac{1}{1+\alpha^2}\left(\alpha^2 \sum_{i=1}^{t-1} a_i^2 + \sum_{i=1}^{t-1} a_i^2\right)}}$$
$$\le \frac{a_t^2\sqrt{1+\alpha^2}}{\sqrt{a_t^2 + \sum_{i=1}^{t-1} a_i^2}} = \frac{a_t^2\sqrt{1+\alpha^2}}{\sqrt{\sum_{i=1}^{t} a_i^2}} \le 2\sqrt{1+\alpha^2}\left(\sqrt{\sum_{i=1}^{t} a_i^2} - \sqrt{\sum_{i=1}^{t-1} a_i^2}\right)$$

where we have used $x^2/\sqrt{x^2+y^2} \le 2(\sqrt{x^2+y^2} - \sqrt{y^2})$ in the last step. On the other hand, if $a_t^2 > \alpha^2 \sum_{t=1}^{t-1} a_i^2$, we have

$$\min\left\{\frac{a_t^2}{\sqrt{\sum_{i=1}^{t-1} a_i^2}}, Ca_t\right\} \le Ca_t = C\frac{\alpha a_t - a_t}{\alpha - 1} \le \frac{C}{\alpha-1}\left(\alpha a_t - \alpha\sqrt{\sum_{i=1}^{t-1} a_i^2}\right)$$
$$= \frac{C\alpha}{\alpha-1}\left(a_t - \sqrt{\sum_{i=1}^{t-1} a_i^2}\right) \le \frac{C\alpha}{\alpha-1}(a_t - M_{t-1}) = \frac{C\alpha}{\alpha-1}(M_t - M_{t-1})$$

where we have used that $a_t = M_t$ and $\sqrt{\sum_{i=1}^{t-1} a_i^2} \ge M_{t-1}$. □

Lemma 4 *([34, Lemma 3.5]). Let $a_1, a_2, \ldots, a_T$ be non-negative real numbers. If $a_1 > 0$ then,*

$$\sum_{t=1}^{T} \frac{a_t}{\sqrt{\sum_{i=1}^{t} a_i}} \le 2\sqrt{\sum_{t=1}^{T} a_t}\,.$$

For completeness, a proof of Lemma 4 is in Appendix D.

Proof of Theorem 1. Let $\eta_t = \frac{1}{\sqrt{\sum_{i=1}^{t-1} \|\ell_i\|_*^2}}$, hence $R_t(w) = \frac{1}{\eta_t} R(w)$. We assume without loss of generality that $\|\ell_t\|_* > 0$ for all t, since otherwise we can remove all rounds t where $\ell_t = 0$ without affecting the regret and the predictions of the algorithm on the remaining rounds. By Lemma 1,

$$\text{Regret}_T(u) \le \frac{1}{\eta_{T+1}} R(u) + \sum_{t=1}^{T}\left(\mathcal{B}_{R_t^*}(-L_t, -L_{t-1}) - R_t^*(-L_t) + R_{t+1}^*(-L_t)\right).$$

We upper bound the terms of the sum in two different ways. First, by Proposition 2, we have

$$\mathcal{B}_{R_t^*}(-L_t, -L_{t-1}) - R_t^*(-L_t) + R_{t+1}^*(-L_t) \le \mathcal{B}_{R_t^*}(-L_t, -L_{t-1}) \le \frac{\eta_t \|\ell_t\|_*^2}{2\lambda}.$$

Second, we have

$$\begin{aligned}
&\mathcal{B}_{R_t^*}(-L_t, -L_{t-1}) - R_t^*(-L_t) + R_{t+1}^*(-L_t)\\
&= \mathcal{B}_{R_{t+1}^*}(-L_t, -L_{t-1}) + R_{t+1}^*(-L_{t-1}) - R_t^*(-L_{t-1}) + \langle \nabla R_t^*(-L_{t-1}) - \nabla R_{t+1}^*(-L_{t-1}), \ell_t\rangle\\
&\le \frac{\eta_{t+1}\|\ell_t\|_*^2}{2\lambda} + \|\nabla R_t^*(-L_{t-1}) - \nabla R_{t+1}^*(-L_{t-1})\| \cdot \|\ell_t\|_*\\
&= \frac{\eta_{t+1}\|\ell_t\|_*^2}{2\lambda} + \|\nabla R^*(-\eta_t L_{t-1}) - \nabla R^*(-\eta_{t+1} L_{t-1})\| \cdot \|\ell_t\|_*\\
&\le \frac{\eta_{t+1}\|\ell_t\|_*^2}{2\lambda} + \min\left\{\frac{1}{\lambda}\|L_{t-1}\|_*(\eta_t - \eta_{t+1}), D\right\}\|\ell_t\|_*,
\end{aligned}$$

where in the first inequality we have used the fact that $R_{t+1}^*(-L_{t-1}) \le R_t^*(-L_{t-1})$, Hölder's inequality, and Proposition 2. In the second inequality we have used properties 5 and 7 of Proposition 2. Using the definition of η_{t+1} we have

$$\frac{\|L_{t-1}\|_*(\eta_t - \eta_{t+1})}{\lambda} \le \frac{\|L_{t-1}\|_*}{\lambda\sqrt{\sum_{i=1}^{t-1}\|\ell_i\|_*^2}} \le \frac{\sum_{i=1}^{t-1}\|\ell_i\|_*}{\lambda\sqrt{\sum_{i=1}^{t-1}\|\ell_i\|_*^2}} \le \frac{\sqrt{t-1}}{\lambda} \le \frac{\sqrt{T-1}}{\lambda}.$$

Denoting by $H = \min\left\{\frac{\sqrt{T-1}}{\lambda}, D\right\}$ we have

$$\begin{aligned}
\text{Regret}_T(u) &\le \frac{1}{\eta_{T+1}} R(u) + \sum_{t=1}^{T} \min\left\{\frac{\eta_t \|\ell_t\|_*^2}{2\lambda},\ H\|\ell_t\|_* + \frac{\eta_{t+1}\|\ell_t\|_*^2}{2\lambda}\right\} \\
&\le \frac{1}{\eta_{T+1}} R(u) + \frac{1}{2\lambda}\sum_{t=1}^{T} \eta_{t+1}\|\ell_t\|_*^2 + \frac{1}{2\lambda}\sum_{t=1}^{T} \min\left\{\eta_t\|\ell_t\|_*^2,\ 2\lambda H\|\ell_t\|_*\right\} \\
&= \frac{1}{\eta_{T+1}} R(u) + \frac{1}{2\lambda}\sum_{t=1}^{T} \frac{\|\ell_t\|_*^2}{\sqrt{\sum_{i=1}^{t}\|\ell_i\|_*^2}} + \frac{1}{2\lambda}\sum_{t=1}^{T} \min\left\{\frac{\|\ell_t\|_*^2}{\sqrt{\sum_{i=1}^{t-1}\|\ell_i\|_*^2}},\ 2\lambda H\|\ell_t\|_*\right\}.
\end{aligned}$$

We bound each of the three terms separately. By definition of η_{T+1}, the first term is $\frac{1}{\eta_{T+1}} R(u) = R(u)\sqrt{\sum_{t=1}^{T}\|\ell_t\|_*^2}$. We upper bound the second term using Lemma 4 as

$$\frac{1}{2\lambda}\sum_{t=1}^{T} \frac{\|\ell_t\|_*^2}{\sqrt{\sum_{i=1}^{t}\|\ell_i\|_*^2}} \le \frac{1}{\lambda}\sqrt{\sum_{t=1}^{T}\|\ell_t\|_*^2}\,.$$

Finally, by Lemma 3 we upper bound the third term as

$$\frac{1}{2\lambda}\sum_{t=1}^{T} \min\left\{\frac{\|\ell_t\|_*^2}{\sqrt{\sum_{i=1}^{t-1}\|\ell_i\|_*^2}},\ 2\lambda\|\ell_t\|_* H\right\} \le 3.5H \max_{t\le T}\|\ell_t\|_* + \frac{1.75}{\lambda}\sqrt{\sum_{t=1}^{T}\|\ell_t\|_*^2}\,.$$

Putting everything together gives the stated bound. □

4. Scale-free mirror descent

In this section, we analyze scale-free version of Mirror Descent. Our algorithm uses the regularizer

$$R_t(w) = R(w)\sqrt{\sum_{i=1}^{t}\|\ell_i\|_*^2}\,, \tag{2}$$

where $R : K \to \mathbb{R}$ an arbitrary strongly convex function. As for SOLO FTRL, it is easy to see that such regularizer gives rise to predictions w_t that are scale-free. We call the resulting algorithm Scale-Free MD. Similar to SOLO FTRL, the regularizer (2) does not uniquely define the MD minimizer $w_{t+1} = \operatorname{argmin}_{w\in K}\left(\langle \ell_t, w\rangle + \mathcal{B}_{R_t}(w, w_t)\right)$ when $\sqrt{\sum_{i=1}^{t}\|\ell_i\|_*^2}$ is zero. This happens when the loss vectors $\ell_1, \ell_2, \dots, \ell_t$ are all zero. In this case, we define $w_{t+1} = \operatorname{argmin}_{w\in K} R(w)$ which agrees with $w_{t+1} = \lim_{a\to 0^+} \operatorname{argmin}_{w\in K} a\mathcal{B}_R(w, w_t)$. Similarly, $w_1 = \operatorname{argmin}_{w\in K} R(w)$.

Per-coordinate version of Scale-Free MD with regularizer $R(w) = \frac{1}{2}\|w\|_2^2$ is exactly the same algorithm as the diagonal version of AdaGrad MD [24].

The theorem below upper bounds the regret of Scale-Free MD (see also [24,32,36]). The proof is in Appendix E.

Theorem 2 *(Regret of scale-free mirror descent). Suppose $K \subseteq V$ is a non-empty closed convex set. Suppose that $R : K \to \mathbb{R}$ is a λ-strongly convex function with respect to a norm $\|\cdot\|$. Scale-Free MD with regularizer R satisfies for any $u \in K$,*

$$\text{Regret}_T(u) \le \left(\frac{1}{\lambda} + \sup_{v\in K} \mathcal{B}_R(u, v)\right)\sqrt{\sum_{t=1}^{T}\|\ell_t\|_*^2}\,.$$

We choose the regularizer $R(w) = \lambda f(w)$ where f is a 1-strongly convex function and optimize λ. The result is the following Corollary. Its proof is trivial.

Corollary 2 *(Regret of scale-free mirror descent). Suppose $K \subseteq V$ is a non-empty bounded closed convex set. Suppose that $f : K \to \mathbb{R}$ is a 1-strongly convex function with respect to a norm $\|\cdot\|$. Scale-Free MD with regularizer*

$$R(w) = \frac{f(w)}{\sqrt{\sup_{u,v\in K} \mathcal{B}_f(u, v)}} \quad \textit{satisfies} \quad \text{Regret}_T \le 2\sqrt{\sup_{u,v\in K} \mathcal{B}_f(u, v)\sum_{t=1}^{T}\|\ell_t\|_*^2}\,.$$

The regret bound for SCALE-FREE MD in the Corollary 2 depends on $\sup_{u,v \in K} \mathcal{B}_f(u,v)$. In contrast, the regret bound for SOLO FTRL in Corollary 1 depend on $\sup_{u \in K} f(u)$. Similarly, the regret bound in Theorem 2 for SCALE-FREE MD depends on $\sup_{v \in K} \mathcal{B}_R(u,v)$ and the regret bounds in Theorem 1 for SOLO FTRL depend on $R(u)$. It is not hard to show that

$$\forall u \in K \qquad R(u) \le \sup_{v \in K} \mathcal{B}_R(u,v)\,, \tag{3}$$

provided that at the minimizer $v^* = \operatorname{argmin}_{v \in K} R(v)$ both $R(v^*)$ and $\nabla R(v^*)$ are zero. Indeed, in that case, $R(u) = \mathcal{B}_R(u, v^*) \le \sup_{v \in K} \mathcal{B}_R(u,v)$.

The assumption $R(v^*) = 0$ and $\nabla R(v^*) = 0$ are easy to achieve by adding an affine function to the regularizer:

$$R'(u) = R(u) - \langle \nabla R(v^*), u - v^* \rangle - R(v^*)\,.$$

The regularizer R' has the same parameter of strong convexity as R, the associated Bregman divergences $\mathcal{B}_{R'}$ and $\mathcal{B}_R$ are equal, R' and R have the same minimizer v^*, and $R'(v^*)$ and $\nabla R'(v^*)$ are both zero.

Thus, inequality (3) implies that—ignoring constant factors—the regret bound for SCALE-FREE MD is inferior to the regret bound for SOLO FTRL. In fact, it is not hard to come up with examples where $R(u)$ is finite whereas $\sup_{v \in K} \mathcal{B}_R(u,v)$ is infinite. We mention two such examples. The first example is $R(w) = \frac{1}{2}\|w\|_2^2$ defined on the whole space V, where for any $u \in V$, $R(u)$ is a finite value but $\sup_{v \in K} \mathcal{B}_R(u,v) = \sup_{v \in V} \frac{1}{2}\|u - v\|_2^2 = +\infty$. The second example is the shifted negative entropy regularizer $R(w) = \ln(d) + \sum_{j=1}^d w_j \ln w_j$ defined on the d-dimensional probability simplex $K = \{w \in \mathbb{R}^d \;:\; w_j \ge 0, \sum_{j=1}^d w_j = 1\}$, where for any $u \in K$, $R(u)$ is finite and in fact lies in the interval $[0, \ln d]$ but $\sup_{v \in K} \mathcal{B}_R(u,v) = \sup_{v \in K} \sum_{j=1}^d u_j \ln(u_j / v_j) = +\infty$. We revisit these examples in the following subsection.

4.1. Lower bounds for scale-free mirror descent

The bounds in Theorem 2 and Corollary 2 are vacuous when $\mathcal{B}_R(u,v)$ is not bounded. One might wonder if the assumption that $\mathcal{B}_R(u,v)$ is bounded is necessary in order for SCALE-FREE MD to have a sublinear regret. We show necessity of this assumption on two counter-examples. In these counter-examples, we consider strongly convex regularizers R such that $\mathcal{B}_R(u,v)$ is not bounded and we construct sequences of loss vectors $\ell_1, \ell_2, \ldots, \ell_T$ such that $\|\ell_1\|_* = \|\ell_2\|_* = \cdots = \|\ell_T\|_* = 1$ and SCALE-FREE MD has regret $\Omega(T)$ or worse.

The first counter-example is stated as Theorem 3 below; our proof is in Appendix E. The decision set is the whole space $K = V$ and the regularizer is $R(w) = \frac{1}{2}\|w\|_2^2$. Note that $R(w)$ is 1-strongly convex with respect to $\|\cdot\|_2$ and the dual norm of $\|\cdot\|_2$ is $\|\cdot\|_2$. The corresponding Bregman divergence is $\mathcal{B}_R(u,v) = \frac{1}{2}\|u - v\|_2^2$. The counter-example constructs a sequence of unit-norm loss vectors in the one-dimensional subspace spanned by the first vector of the standard orthonormal basis. On such a sequence, both versions of ADAGRAD MD as well as SCALE-FREE MD are identical to gradient descent with step size $1/\sqrt{t}$, i.e., they are identical Zinkevich's GENERALIZED INFINITESIMAL GRADIENT ASCENT (GIGA) algorithm [20]. Hence the lower bound applies to all these algorithms.

Theorem 3 *(First counter-example). Suppose $K = V$. For any $T \ge 42$, there exists a sequence of loss vectors $\ell_1, \ell_2, \ldots, \ell_T \in V^*$ such that $\|\ell_1\|_2 = \|\ell_2\|_2 = \cdots = \|\ell_T\|_2 = 1$ and* SCALE-FREE MD *with regularizer $R(w) = \frac{1}{2}\|w\|_2^2$,* GIGA, *and both versions of* ADAGRAD MD *satisfy*

$$\text{Regret}_T(0) \ge \frac{T^{3/2}}{20}\,.$$

The second counter-example is stated as Theorem 4 below; our proof is in Appendix E. The decision set is the d-dimensional probability simplex $K = \{w \in \mathbb{R}^d \;:\; w_j \ge 0, \sum_{j=1}^d w_j = 1\}$ and the regularizer is the negative entropy $R(w) = \sum_{j=1}^d w_j \ln w_j$. Negative entropy is 1-strongly convex with respect to $\|\cdot\|_1$ and the dual norm of $\|\cdot\|_1$ is $\|\cdot\|_\infty$. The corresponding Bregman divergence is the Kullback–Leibler divergence $\mathcal{B}_R(u,v) = \sum_{j=1}^d u_j \ln(u_j/v_j)$. Note that despite that negative entropy is upper- and lower-bounded, Kullback–Leibler divergence can be arbitrarily large.

Theorem 4 *(Second counter-example). Let $d \ge 2$, let $V = \mathbb{R}^d$, and let $K = \{w \in V \;:\; w_j \ge 0, \sum_{j=1}^d w_j = 1\}$ be the d-dimensional probability simplex. For any $T \ge 120$, there exists a sequence of loss vectors $\ell_1, \ell_2, \ldots, \ell_T \in V^*$ such that $\|\ell_1\|_\infty = \|\ell_2\|_\infty = \cdots = \|\ell_T\|_\infty = 1$ and* SCALE-FREE MD *with regularizer $R(w) = \sum_{j=1}^d w_j \ln w_j$ satisfies*

$$\text{Regret}_T \ge \frac{T}{6}\,.$$

5. Lower bound

We show a lower bound on the worst-case regret of any algorithm for OLO. The proof, presented in Appendix F, is a standard probabilistic argument.

Theorem 5 *(Lower bound). Let $K \subseteq V$ be any non-empty bounded closed convex subset. Let $D = \sup_{u,v \in K} \|u - v\|$ be the diameter of K. Let A be any (possibly randomized) algorithm for OLO on K. Let T be any non-negative integer and let $a_1, a_2, \dots, a_T$ be any non-negative real numbers. There exists a sequence of vectors $\ell_1, \ell_2, \dots, \ell_T$ in the dual vector space V^* such that $\|\ell_1\|_* = a_1$, $\|\ell_2\|_* = a_2, \dots, \|\ell_T\|_* = a_T$ and the regret of algorithm A satisfies*

$$\text{Regret}_T \geq \frac{D}{\sqrt{8}} \sqrt{\sum_{t=1}^{T} \|\ell_t\|_*^2} \,. \tag{4}$$

The upper bounds on the regret, which we have proved for our algorithms, have the same dependency on the norms of the loss vectors. However, a gap remains between the lower bound and the upper bounds.

The upper bound on regret of SOLO FTRL is of the form $O(\sqrt{\sup_{v \in K} f(v) \sum_{t=1}^{T} \|\ell_t\|_*^2})$ where f is any 1-strongly convex function with respect to $\|\cdot\|$. The same upper bound is also achieved by FTRL with a constant learning rate when $\sum_{t=1}^{T} \|\ell_t\|_*^2$ is known upfront [4, Chapter 2]. The lower bound is $\Omega(D\sqrt{\sum_{t=1}^{T} \|\ell_t\|_*^2})$.

The gap between D and $\sqrt{\sup_{v \in K} f(v)}$ can be substantial. For example, if K is the probability simplex in $\mathbb{R}^d$ and $f(w) = \ln(d) + \sum_{j=1}^{d} w_j \ln w_j$ is the shifted negative entropy, the $\|\cdot\|_1$-diameter of K is 2, f is non-negative and 1-strongly convex with respect to $\|\cdot\|_1$, but $\sup_{v \in K} f(v) = \ln(d)$. On the other hand, if the norm $\|\cdot\|_2 = \sqrt{\langle \cdot, \cdot \rangle}$ arises from an inner product $\langle \cdot, \cdot \rangle$, the lower bound matches the upper bounds within a constant factor. The reason is that for any K with $\|\cdot\|_2$-diameter D, the function $f(w) = \frac{1}{2}\|w - w_0\|_2^2$, where w_0 is an arbitrary point in K, is 1-strongly convex with respect to $\|\cdot\|_2$ and satisfies that $\sqrt{\sup_{v \in K} f(v)} \leq D$. This leads to the following open problem (posed also in [37]):

> *Given a bounded convex set K and a norm $\|\cdot\|$, construct a non-negative function $f : K \to \mathbb{R}$ that is 1-strongly convex with respect to $\|\cdot\|$ and minimizes $\sup_{v \in K} f(v)$.*

As shown in [38], the existence of f with small $\sup_{v \in K} f(v)$ is equivalent to the existence of an algorithm for OLO with $\widetilde{O}(\sqrt{T \sup_{v \in K} f(v)})$ regret assuming $\|\ell_t\|_* \leq 1$. The $\widetilde{O}$ notation hides a polylogarithmic factor in T.

6. Conclusions

We have investigated scale-free algorithms for online linear optimization and we have shown that the scale-free property leads to algorithms which have optimal regret and do not need to know or assume **anything** about the sequence of loss vectors. In particular, the algorithms do not assume any upper or lower bounds on the norms of the loss vectors or the number of rounds.

We have designed a scale-free algorithm based on Follow The Regularizer Leader. Its regret with respect to any competitor u is

$$O\left(f(u) \sqrt{\sum_{t=1}^{T} \|\ell_t\|_*^2 + \min\{\sqrt{T}, D\} \max_{t=1,2,\dots,T} \|\ell_t\|_*} \right),$$

where f is any non-negative 1-strongly convex function defined on the decision set and D is the diameter of the decision set. The result makes sense even when the decision set is unbounded.

A similar, but weaker result holds for a scale-free algorithm based on Mirror Descent. However, we have also shown this algorithm to be strictly weaker than algorithms based on Follow The Regularizer Leader. Namely, we gave examples of regularizers for which the scale-free version of Mirror Descent has $\Omega(T)$ regret or worse.

We have proved an $\frac{D}{\sqrt{8}} \sqrt{\sum_{t=1}^{T} \|\ell\|_*^2}$ lower bound on the regret of any algorithm for any decision set with diameter D.

Notice that with the regularizer $f(u) = \frac{1}{2}\|u\|_2^2$ the regret of SOLO FTRL depends quadratically on the norm of the competitor $\|u\|_2$. There exist non-scale-free algorithms [39–44] that have only a $O(\|u\|_2 \sqrt{\log \|u\|_2})$ or $O(\|u\|_2 \log \|u\|_2)$ dependency. These algorithms assume an a priori bound on the norm of the loss vectors. Recently, an algorithm that adapts to norms of loss vectors and has a $O(\|u\|_2 \log \|u\|_2)$ dependency was proposed [45]. However, the trade-off between the dependency on $\|u\|_2$ and the adaptivity to the norms of the loss vectors still remains to be explored.

Acknowledgements

We thank an anonymous reviewer for suggesting a simpler proof of Lemma 7.

Appendix A. Proofs for preliminaries

Proof of Proposition 1. Let $S = \sup_{u \in K} f(u)$ and $v^* = \operatorname{argmin}_{v \in K} f(v)$. The minimizer v^* is guaranteed to exist by lower semi-continuity of f and compactness of K. The optimality condition for v^* and 1-strong convexity of f imply that for any $u \in K$,

$$S \geq f(u) - f(v^*) \geq f(u) - f(v^*) - \langle \nabla f(v^*), u - v^* \rangle \geq \frac{1}{2}\|u - v^*\|^2 .$$

In other words, $\|u - v^*\| \leq \sqrt{2S}$. By the triangle inequality,

$$D = \sup_{u,v \in K} \|u - v\| \leq \sup_{u,v \in K} \left(\|u - v^*\| + \|v^* - v\|\right) \leq 2\sqrt{2S} = \sqrt{8S} . \quad \square$$

Proof of Property 6 of Proposition 2. To bound $\mathcal{B}_{f^*}(x, y)$ we add a non-negative divergence term $\mathcal{B}_{f^*}(y, x)$.

$$\begin{aligned} \mathcal{B}_{f^*}(x, y) &\leq \mathcal{B}_{f^*}(x, y) + \mathcal{B}_{f^*}(y, x) = \langle x - y, \nabla f^*(x) - \nabla f^*(y) \rangle \\ &\leq \|x - y\|_* \cdot \|\nabla f^*(x) - \nabla f^*(y)\| \leq D\|x - y\|_* , \end{aligned}$$

where we have used Hölder's inequality and property 7 of the Proposition. $\square$

Proof of Lemma 1. By the Fenchel–Young inequality,

$$\begin{aligned} \sum_{t=1}^{T} \left(R^*_{t+1}(-L_t) - R^*_t(-L_{t-1})\right) &= R^*_{T+1}(-L_T) - R^*_1(0) \\ &\geq -\langle L_T, u \rangle - R_{T+1}(u) - R^*_1(0) \\ &= -R_{T+1}(u) - R^*_1(0) - \sum_{t=1}^{T} \langle \ell_t, u \rangle . \end{aligned}$$

We add $\sum_{t=1}^{T} \langle \ell_t, w_t \rangle$ to both sides and we obtain $\text{Regret}_T(u)$ on the right side. After rearrangement of the terms, we get an upper bound on the regret:

$$\begin{aligned} \text{Regret}_T(u) &= \sum_{t=1}^{T} \langle \ell_t, w_t \rangle - \sum_{t=1}^{T} \langle \ell_t, u \rangle \\ &\leq R_{T+1}(u) + R^*_1(0) + \sum_{t=1}^{T} \left(R^*_{t+1}(-L_t) - R^*_t(-L_{t-1}) + \langle \ell_t, w_t \rangle\right) . \end{aligned}$$

By Proposition 2, property 2, we have $w_t = \nabla R^*_t(-L_{t-1})$ and therefore we can rewrite the sum in last expression as

$$\begin{aligned} &\sum_{t=1}^{T} R^*_{t+1}(-L_t) - R^*_t(-L_{t-1}) + \langle \ell_t, w_t \rangle \\ &= \sum_{t=1}^{T} R^*_{t+1}(-L_t) - R^*_t(-L_{t-1}) + \langle \ell_t, \nabla R^*_t(-L_{t-1}) \rangle \\ &= \sum_{t=1}^{T} R^*_t(-L_t) - R^*_t(-L_{t-1}) + \langle \ell_t, \nabla R^*_t(-L_{t-1}) \rangle - R^*_t(-L_t) + R^*_{t+1}(-L_t) \\ &= \sum_{t=1}^{T} \mathcal{B}_{R^*_t}(-L_t, -L_{t-1}) - R^*_t(-L_t) + R^*_{t+1}(-L_t) . \end{aligned}$$

This finishes the proof. $\square$

Appendix B. AdaFTRL

In this section, we show that it is possible to derive a scale-free algorithm different from SOLO FTRL. We generalize the AdaHedge algorithm [25] to the OLO setting, showing that it retains its scale-free property. We call the resulting algorithm AdaFTRL. The analysis is very general and based on general properties of strongly convex functions, rather than specific properties of the entropic regularizer as in the original analysis of AdaHedge.

Assume that K is bounded and that $R : K \to \mathbb{R}$ is a strongly convex lower semi-continuous function bounded from above. We instantiate Algorithm 1 with the sequence of regularizers

$$R_t(w) = \Delta_{t-1} R(w) \quad \text{where} \quad \Delta_t = \sum_{i=1}^{t} \Delta_{i-1} \mathcal{B}_{R^*}\left(-\frac{L_i}{\Delta_{i-1}}, -\frac{L_{i-1}}{\Delta_{i-1}}\right) . \tag{B.1}$$

The sequence $\{\Delta_t\}_{t=0}^{\infty}$ is non-negative and non-decreasing. Also, Δ_t as a function of $\ell_1, \ell_2, \ldots, \ell_t$ is positive homogeneous of degree one, making the algorithm scale-free.

If $\Delta_{i-1} = 0$, we define $\Delta_{i-1}\mathcal{B}_{R^*}(\frac{-L_i}{\Delta_{i-1}}, \frac{-L_{i-1}}{\Delta_{i-1}})$ as $\lim_{a\to 0^+} a\mathcal{B}_{R^*}(\frac{-L_i}{a}, \frac{-L_{i-1}}{a})$ which always exists and is finite; see Lemma 9 in Appendix C. Similarly, when $\Delta_{t-1} = 0$, we define $w_t = \operatorname{argmin}_{w\in K} \langle L_{t-1}, w\rangle$ where ties among minimizers are broken by taking the one with the smallest value of $R(w)$, which is unique due to strong convexity. As we show in Lemma 8 in Appendix C, this is the same as $w_t = \lim_{a\to 0^+} \operatorname{argmin}_{w\in K}(\langle L_{t-1}, w\rangle + aR(w))$.

Our main result is an $O(\sqrt{\sum_{t=1}^{T} \|\ell_t\|_*^2})$ upper bound on the regret of the algorithm after T rounds, without the need to know beforehand an upper bound on $\|\ell_t\|_*$. We prove the theorem in B.1.

Theorem 6 *(Regret bound). Suppose $K \subseteq V$ is a non-empty bounded closed convex set. Let $D = \sup_{x,y\in K} \|x - y\|$ be its diameter with respect to a norm $\|\cdot\|$. Suppose that the regularizer $R : K \to \mathbb{R}$ is a non-negative lower semi-continuous function that is λ-strongly convex with respect to $\|\cdot\|$ and is bounded from above. The regret of* AdaFTRL *satisfies*

$$\text{Regret}_T(u) \le \sqrt{3} \max\left\{D, \frac{1}{\sqrt{2\lambda}}\right\} \sqrt{\sum_{t=1}^{T} \|\ell_t\|_*^2}\,(1 + R(u)) .$$

The regret bound can be optimized by choosing the optimal multiple of the regularizer. Namely, we choose regularizer of the form $\lambda f(w)$ where $f(w)$ is 1-strongly convex and optimize over λ. The result of the optimization is the following corollary.

Corollary 3 *(Regret bound). Suppose $K \subseteq V$ is a non-empty bounded closed convex set. Suppose $f : K \to \mathbb{R}$ is a non-negative lower semi-continuous function that is 1-strongly convex with respect to $\|\cdot\|$ and is bounded from above. The regret of* AdaFTRL *with regularizer*

$$R(w) = \frac{f(w)}{16 \cdot \sup_{v\in K} f(v)} \qquad \textit{satisfies} \qquad \text{Regret}_T \le 5.3 \sqrt{\sup_{v\in K} f(v) \sum_{t=1}^{T} \|\ell_t\|_*^2} .$$

Proof. Let $S = \sup_{v\in K} f(v)$. Theorem 6 applied to the regularizer $R(w) = \frac{c}{S} f(w)$ and Proposition 1 gives

$$\text{Regret}_T \le \sqrt{3}(1 + c) \max\left\{\sqrt{8}, \frac{1}{\sqrt{2c}}\right\} \sqrt{S \sum_{t=1}^{T} \|\ell_t\|_*^2} .$$

It remains to find the minimum of $g(c) = \sqrt{3}(1 + c)\max\{\sqrt{8}, 1/\sqrt{2c}\}$. The function g is strictly convex on $(0, \infty)$ and has minimum at $c = 1/16$ and $g(\frac{1}{16}) = \sqrt{3}(1 + \frac{1}{16})\sqrt{8} \le 5.3$. □

B.1. Proof of regret bound for AdaFTRL

Lemma 5 *(Initial regret bound).* AdaFTRL *satisfies, for any $u \in K$ and any $T \ge 0$,*

$$\text{Regret}_T(u) \le (1 + R(u))\, \Delta_T .$$

Proof. Recall from (B.1) that $R_t(w) = \Delta_{t-1}R(w)$. Since R is non-negative, $\{R_t\}_{t=1}^{\infty}$ is non-decreasing. Hence, $R_t^*(\ell) \ge R_{t+1}^*(\ell)$ for every $\ell \in V^*$ and thus $R_t^*(-L_t) - R_{t+1}^*(-L_t) \ge 0$. So, by Lemma 1,

$$\mathrm{Regret}_T(u) \le R_{T+1}(u) + R_1^*(0) + \sum_{t=1}^{T} \mathcal{B}_{R_t^*}(-L_t, -L_{t-1}) \,. \tag{B.2}$$

Technically, (B.2) is not justified since R_t might not be strongly convex. This happens when $\Delta_{t-1} = 0$. In order to justify (B.2), we consider a different algorithm that initializes $\Delta_0 = \epsilon$ where $\epsilon > 0$; that ensures that $\Delta_{t-1} > 0$ and R_t is strongly convex. Applying Lemma 1 and then taking limit $\epsilon \to 0$, yields (B.2).

Since, $\mathcal{B}_{R_t^*}(u, v) = \Delta_{t-1}\mathcal{B}_{R^*}(\frac{u}{\Delta_{t-1}}, \frac{v}{\Delta_{t-1}})$ by definition of Bregman divergence and property 8 of Proposition 2, we have $\sum_{t=1}^{T} \mathcal{B}_{R_t^*}(-L_t, -L_{t-1}) = \Delta_T$. □

Lemma 6 *(Recurrence). Let $D = \sup_{u,v \in K} \|u - v\|$ be the diameter of K. The sequence $\{\Delta_t\}_{t=1}^{\infty}$ generated by* AdaFTRL *satisfies for any $t \ge 1$,*

$$\Delta_t \le \Delta_{t-1} + \min\left\{ D\|\ell_t\|_*, \, \frac{\|\ell_t\|_*^2}{2\lambda\Delta_{t-1}} \right\} .$$

Proof. By definition, Δ_t satisfies the recurrence

$$\Delta_t = \Delta_{t-1} + \Delta_{t-1}\mathcal{B}_{R^*}\left(-\frac{L_t}{\Delta_{t-1}}, -\frac{L_{t-1}}{\Delta_{t-1}} \right) .$$

Using parts 4 and 6 of Proposition 2, we can upper bound $\mathcal{B}_{R^*}\left(-\frac{L_t}{\Delta_{t-1}}, -\frac{L_{t-1}}{\Delta_{t-1}} \right)$ with two different quantities. Taking the minimum of the two quantities finishes the proof. □

The recurrence of Lemma 6 can be simplified. Defining

$$a_t = \|\ell_t\|_* \max\left\{ D, \frac{1}{\sqrt{2\lambda}} \right\} ,$$

we get a recurrence

$$\Delta_t \le \Delta_{t-1} + \min\left\{ a_t, \frac{a_t^2}{\Delta_{t-1}} \right\} .$$

The next lemma solves this recurrence, by giving an explicit upper bound on Δ_T in terms of $a_1, a_2, \ldots, a_T$.

Lemma 7 *(Solution of the recurrence). Let $\{a_t\}_{t=1}^{\infty}$ be any sequence of non-negative real numbers. Suppose that $\{\Delta_t\}_{t=0}^{\infty}$ is a sequence of non-negative real numbers satisfying*

$$\Delta_0 = 0 \qquad \text{and} \qquad \Delta_t \le \Delta_{t-1} + \min\left\{ a_t, \frac{a_t^2}{\Delta_{t-1}} \right\} \quad \text{for any } t \ge 1 .$$

Then, for any $T \ge 0$,

$$\Delta_T \le \sqrt{3\sum_{t=1}^{T} a_t^2} \,.$$

Proof. Observe that

$$\Delta_T^2 = \sum_{t=1}^{T} \Delta_t^2 - \Delta_{t-1}^2 = \sum_{t=1}^{T} (\Delta_t - \Delta_{t-1})^2 + 2(\Delta_t - \Delta_{t-1})\Delta_{t-1} \,.$$

We bound each term in the sum separately. The left term of the minimum inequality in the definition of Δ_t gives

$$(\Delta_t - \Delta_{t-1})^2 \le a_t^2,$$

while the right term gives

$$2(\Delta_t - \Delta_{t-1})\Delta_{t-1} \le 2a_t^2 \,.$$

So, we conclude

$$\Delta_T^2 \le 3\sum_{t=1}^{T} a_t^2 \,. \quad \square$$

Theorem 6 follows from Lemmas 5, 6 and 7.

Appendix C. Limits

In this section, we show that prediction of AdaFTRL is correctly defined when the regularizer is multiplied by zero.

Lemma 8 *(Prediction for zero regularizer). Let K be non-empty bounded closed convex subset of a finite dimensional normed real vector space $(V, \|\cdot\|)$. Let $R : K \to \mathbb{R}$ be strictly convex and lower semi-continuous, and let $L \in V^*$. The limit*

$$\lim_{\eta \to +\infty} \operatorname*{argmin}_{w \in K} \left(\langle L, w \rangle + \frac{1}{\eta} R(w) \right) \tag{C.1}$$

exists and it is equal to the unique minimizer of $R(w)$ over the set (of minimizers)

$$\left\{ w \in K \ : \ \langle L, w \rangle = \inf_{v \in K} \langle L, v \rangle \right\} .$$

Before we give the proof, we illustrate the lemma on a simple example. Let $K = [-1, 1]^2$ be a closed square in $\mathbb{R}^2$ and let $R(w) = \|w\|_2^2$. Let $L = (1, 0)$. The minimizers are

$$\operatorname*{argmin}_{w \in K} \langle L, w \rangle = \{(-1, y) \ : \ y \in [-1, 1]\} .$$

The minimizer with the smallest value of $R(w)$ is $(-1, 0)$. Hence the lemma implies that

$$\lim_{\eta \to +\infty} \operatorname*{argmin}_{w \in K} \left(\langle L, w \rangle + \frac{1}{\eta} \|w\|_2^2 \right) = (-1, 0) .$$

Proof of Lemma 8. Without loss of generality, we can assume that $R(w)$ is non-negative for any $w \in K$. For otherwise, we can replace $R(w)$ with $R'(w) = R(w) - \inf_{v \in K} R(v)$.

Since K is a non-empty bounded closed convex subset of a finite dimensional normed vector space, it is compact and $r^* = \min_{w \in K} \langle L, w \rangle$ exists and is attained at some $w \in K$. Consider the hyperplane

$$H = \{w \in V \ : \ \langle L, w \rangle = r^*\} .$$

The intersection $H \cap K$ is a non-empty compact convex set. Let

$$v^* = \operatorname*{argmin}_{v \in K \cap H} R(v) .$$

The existence of v^* follows from compactness of $H \cap K$ and lower semi-continuity of $R(v)$. Uniqueness of v^* follows from strict convexity of $R(v)$. We show that the limit (C.1) equals v^*.

By the definition of H,

$$v^* \in \operatorname*{argmin}_{w \in K} \langle L, w \rangle . \tag{C.2}$$

Let $S = \{w \in K \ : \ R(w) \le R(v^*)\}$. Since $R(w)$ is lower semi-continuous S is closed. Since $R(w)$ is strictly convex, $S \cap H = \{v^*\}$.

For any $\eta > 0$, let

$$w(\eta) = \operatorname*{argmin}_{w \in K} \left(\langle L, w \rangle + \frac{1}{\eta} R(w)) \right) .$$

We prove that $w(\eta) \in S$. Indeed, by optimality of v^* and $w(\eta)$,

$$\frac{1}{\eta} R(w(\eta)) + \langle L, w(\eta) \rangle \le \frac{1}{\eta} R(v^*) + \langle L, v^* \rangle \le \frac{1}{\eta} R(v^*) + \langle L, w(\eta) \rangle$$

and hence $R(w(\eta)) \le R(v^*)$.

By non-negativity of R and optimality of $w(\eta)$ we have

$$\langle L, w(\eta) \rangle \le \langle L, w(\eta) \rangle + \frac{1}{\eta} R(w(\eta)) \le \langle L, v^* \rangle + \frac{1}{\eta} R(v^*) .$$

Taking the limit $\eta \to +\infty$, we see that

$$\lim_{\eta\to+\infty} \langle L, w(\eta)\rangle \le \lim_{\eta\to+\infty} \left(\langle L, v^*\rangle + \frac{1}{\eta} R(v^*)\right) = \langle L, v^*\rangle \,.$$

From (C.2) we have $\langle L, v^*\rangle \le \langle L, w\rangle$ for any w, and therefore

$$\lim_{\eta\to+\infty} \langle L, w(\eta)\rangle = \langle L, v^*\rangle \,. \tag{C.3}$$

Consider any sequence $\{\eta_t\}_{t=1}^\infty$ of positive numbers approaching $+\infty$. Since K is compact, $w(\eta_t)$ has a convergent subsequence. Thus $\{w(\eta_t)\}_{t=1}^\infty$ has at least one accumulation point; let w^* be any of them. We will show that $w^* = v^*$.

Consider a subsequence $\{\xi_t\}_{t=1}^\infty$ of $\{\eta_t\}_{t=1}^\infty$ such that $\lim_{t\to\infty} w(\xi_t) = w^*$. Since $w(\xi_t) \in S$ and S is closed, $w^* \in S$. From (C.3) we have $\langle L, w^*\rangle = \langle L, v^*\rangle$ and hence $w^* \in H$. Thus $w^* \in S \cap H$. Since v^* is the only point in $S \cap H$ we must have $w^* = v^*$. $\square$

Lemma 9 *(Limit of Bregman divergence). Let K be a non-empty bounded closed convex subset of a finite dimensional normed real vector space $(V, \|\cdot\|)$. Let $R : K \to \mathbb{R}$ be a strongly convex lower semi-continuous function bounded from above. Then, for any $x, y \in V^*$,*

$$\lim_{a\to 0^+} a\mathcal{B}_{R^*}(x/a, y/a) = \langle x, u - v\rangle$$

where

$$u = \lim_{a\to 0^+} \operatorname*{argmin}_{w\in K} (aR(w) - \langle x, w\rangle) \quad \textit{and} \quad v = \lim_{a\to 0^+} \operatorname*{argmin}_{w\in K} (aR(w) - \langle y, w\rangle) \,.$$

Proof. Using property 3 of Proposition 2 we can write the divergence

$$\begin{aligned} a\mathcal{B}_{R^*}(x/a, y/a) &= aR^*(x/a) - aR^*(y/a) - \langle x - y, \nabla R^*(y/a)\rangle \\ &= a\left[\langle x/a, \nabla R^*(x/a)\rangle - R(\nabla R^*(x/a))\right] \\ &\quad - a\left[\langle y/a, \nabla R^*(y/a)\rangle - R(\nabla R^*(y/a))\right] - \langle x - y, \nabla R^*(y/a)\rangle \\ &= \langle x, \nabla R^*(x/a) - \nabla R^*(y/a)\rangle - aR(\nabla R^*(x/a)) + aR(\nabla R^*(y/a)) \,. \end{aligned}$$

Property 2 of Proposition 2 implies that

$$\begin{aligned} u &= \lim_{a\to 0^+} \nabla R^*(x/a) = \lim_{a\to 0^+} \operatorname*{argmin}_{w\in K} (aR(w) - \langle x, w\rangle) \,, \\ v &= \lim_{a\to 0^+} \nabla R^*(y/a) = \lim_{a\to 0^+} \operatorname*{argmin}_{w\in K} (aR(w) - \langle y, w\rangle) \,. \end{aligned}$$

The limits on the right exist according to Lemma 8. They are simply the minimizers $u = \operatorname{argmin}_{w\in K} -\langle x, w\rangle$ and $v = \operatorname{argmin}_{w\in K} -\langle y, w\rangle$ where ties in argmin are broken according to smaller value of $R(w)$.

By assumption $R(w)$ is upper bounded. It is also lower bounded, since it is defined on a compact set and it is lower semi-continuous. Thus,

$$\begin{aligned} &\lim_{a\to 0^+} a\mathcal{B}_{R^*}(x/a, y/a) \\ &= \lim_{a\to 0^+} \langle x, \nabla R^*(x/a) - \nabla R^*(y/a)\rangle - aR(\nabla R^*(x/a)) + aR(\nabla R^*(y/a)) \\ &= \lim_{a\to 0^+} \langle x, \nabla R^*(x/a) - \nabla R^*(y/a)\rangle = \langle x, u - v\rangle \,. \quad \square \end{aligned}$$

Appendix D. Proofs for SOLO FTRL

Proof of Lemma 4. We use the inequality $x/\sqrt{x+y} \le 2(\sqrt{x+y} - \sqrt{y})$ which holds for non-negative x, y that are not both zero. Substituting $x = a_t$ and $y = \sum_{i=1}^{t-1} a_i$, we get that for any $t \ge 1$,

$$\frac{a_t}{\sqrt{\sum_{i=1}^t a_i}} \le 2\sqrt{\sum_{i=1}^t a_i} - 2\sqrt{\sum_{i=1}^{t-1} a_i} \,.$$

Summing the above inequality over all $t = 1, 2, \ldots, T$, the right side telescopes to $2\sqrt{\sum_{t=1}^T a_t}$. $\square$

Appendix E. Proofs for scale-free mirror descent

Proof of Lemma 2. Let

$$\begin{aligned}\Psi_{t+1}(w) &= \langle \ell_t, w\rangle + \mathcal{B}_{R_t}(w, w_t)\\ &= \langle \ell_t, w\rangle + R_t(w) - R_t(w_t) - \langle \nabla R_t(w_t), w - w_t\rangle\,.\end{aligned}$$

Then, $w_{t+1} = \operatorname{argmin}_{w\in K} \Psi_{t+1}(w)$. Note that $\nabla\Psi_{t+1}(w) = \ell_t + \nabla R_t(w) - \nabla R_t(w_t)$. The optimality condition for w_{t+1} states that $\langle \nabla\Psi_{t+1}(w_{t+1}), u - w_{t+1}\rangle \ge 0$ for all $u \in K$. Written explicitly,

$$\langle \ell_t + \nabla R_t(w_{t+1}) - \nabla R_t(w_t), u - w_{t+1}\rangle \ge 0\,.$$

Adding $\langle \ell_t, w_{t+1} - w_t\rangle$ to both sides and rearranging, we have

$$\begin{aligned}\langle \ell_t, w_t - u\rangle &\le \langle \nabla R_t(w_{t+1}) - \nabla R_t(w_t), u - w_{t+1}\rangle + \langle \ell_t, w_t - w_{t+1}\rangle\\ &= \langle \ell_t, w_t - w_{t+1}\rangle - \mathcal{B}_{R_t}(w_{t+1}, w_t) + \mathcal{B}_{R_t}(u, w_t) - \mathcal{B}_{R_t}(u, w_{t+1})\,.\end{aligned}$$

The last equality follows by from definition of Bregman divergence. Summation over all $t = 1, 2, \dots, T$ gives the final regret bound. □

Proof of Theorem 2. Let $\eta_t = \frac{1}{\sqrt{\sum_{i=1}^t \|\ell_i\|_*^2}}$. We define $\eta_0 = +\infty$ and $1/\eta_0 = 0$. Hence $R_t(w) = \frac{1}{\eta_t} R(w)$. Since R_t is $\frac{\lambda}{\eta_t}$-strongly convex, we have

$$\begin{aligned}\langle \ell_t, w_t - w_{t+1}\rangle - \mathcal{B}_{R_t}(w_{t+1}, w_t) &\le \|\ell_t\|_* \cdot \|w_t - w_{t+1}\| - \frac{\lambda}{2\eta_t}\|w_t - w_{t+1}\|^2\\ &\le \max_{z\in\mathbb{R}}\left(\|\ell_t\|_* z - \frac{\lambda}{2\eta_t} z^2\right)\\ &= \frac{\eta_t}{2\lambda}\|\ell_t\|_*^2\,.\end{aligned}$$

Combining the last inequality with Lemma 2, we have

$$\text{Regret}_T(u) \le \sum_{t=1}^T \frac{\eta_t}{2\lambda}\|\ell_t\|_*^2 + \sum_{t=1}^T \left[\mathcal{B}_{R_t}(u, w_t) - \mathcal{B}_{R_t}(u, w_{t+1})\right]\,.$$

Since $R_t(w) = \frac{1}{\eta_t} R(w)$, we have

$$\begin{aligned}\text{Regret}_T(u) &\le \frac{1}{2\lambda}\sum_{t=1}^T \eta_t\|\ell_t\|_*^2 + \sum_{t=1}^T \frac{1}{\eta_t}\left[\mathcal{B}_R(u, w_t) - \mathcal{B}_R(u, w_{t+1})\right]\\ &\le \frac{1}{2\lambda}\sum_{t=1}^T \eta_t\|\ell_t\|_*^2 + \sum_{t=1}^T \mathcal{B}_R(u, w_t)\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\\ &\le \frac{1}{2\lambda}\sum_{t=1}^T \frac{\|\ell_t\|_*^2}{\sqrt{\sum_{i=1}^t \|\ell_i\|_*^2}} + \sup_{v\in K}\mathcal{B}_R(u, v)\sum_{t=1}^T\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\\ &\le \frac{1}{\lambda}\sqrt{\sum_{t=1}^T \|\ell_t\|_*^2} + \sup_{v\in K}\mathcal{B}_R(u, v)\sqrt{\sum_{t=1}^T \|\ell_t\|_*^2} \qquad \text{(by Lemma 4)}\\ &= \left(\frac{1}{\lambda} + \sup_{v\in K}\mathcal{B}_R(u, v)\right)\sqrt{\sum_{t=1}^T \|\ell_t\|_*^2}\,. \quad \square\end{aligned}$$

Proof of Theorem 3. We assume $d = 1$. For $d \ge 2$, we simply embed the one-dimensional loss vectors into the first coordinate of $\mathbb{R}^d$. Consider the sequence

$$(\ell_1, \ell_2, \dots, \ell_T) = (\underbrace{-1, -1, \dots, -1}_{\lceil T/2\rceil}, \underbrace{+1, +1, \dots, +1}_{\lfloor T/2\rfloor})\,.$$

The first half consists of -1's, the second of $+1$'s. For $t \le \lceil T/2 \rceil$

$$w_{t+1} = w_t + \frac{1}{\sqrt{t}}\,.$$

Unrolling the recurrence and using $w_1 = 0$ we get

$$w_t = \sum_{i=1}^{t-1} \frac{1}{\sqrt{i}} \qquad (\text{for } t \le \lceil T/2 \rceil + 1)\,.$$

On the other hand, for $t \ge \lceil T/2 \rceil + 1$, we have

$$w_{t+1} = w_t - \frac{1}{\sqrt{t}}\,.$$

Unrolling the recurrence up to $w_{\lceil T/2 \rceil+1}$ we get

$$w_t = w_{\lceil T/2 \rceil+1} - \sum_{i=\lceil T/2 \rceil+1}^{t-1} \frac{1}{\sqrt{i}} = \sum_{i=1}^{\lceil T/2 \rceil} \frac{1}{\sqrt{i}} - \sum_{i=\lceil T/2 \rceil+1}^{t-1} \frac{1}{\sqrt{i}} \qquad (\text{for } t \ge \lceil T/2 \rceil + 1)\,.$$

We are ready to lower bound the regret.

$$\begin{aligned}
\text{Regret}_T(0) &= \sum_{t=1}^{T} \ell_t w_t \\
&= -\sum_{t=1}^{\lceil T/2 \rceil} w_t + \sum_{t=\lceil T/2 \rceil+1}^{T} w_t \\
&= -\sum_{t=1}^{\lceil T/2 \rceil} \sum_{i=1}^{t-1} \frac{1}{\sqrt{i}} + \sum_{t=\lceil T/2 \rceil+1}^{T} \left(\sum_{i=1}^{\lceil T/2 \rceil} \frac{1}{\sqrt{i}} - \sum_{i=\lceil T/2 \rceil+1}^{t-1} \frac{1}{\sqrt{i}} \right) \\
&= -\sum_{i=1}^{\lceil T/2 \rceil} \frac{\lceil T/2 \rceil - i}{\sqrt{i}} + \lfloor T/2 \rfloor \sum_{i=1}^{\lceil T/2 \rceil} \frac{1}{\sqrt{i}} - \sum_{i=\lceil T/2 \rceil+1}^{T} \frac{T-i}{\sqrt{i}} \\
&= -\sum_{i=1}^{\lceil T/2 \rceil} \frac{\lceil T/2 \rceil - \lfloor T/2 \rfloor}{\sqrt{i}} + \sum_{i=1}^{T} \sqrt{i} - T \sum_{i=\lceil T/2 \rceil+1}^{T} \frac{1}{\sqrt{i}} \\
&\ge -\sum_{i=1}^{\lceil T/2 \rceil} \frac{1}{\sqrt{i}} + \sum_{i=1}^{T} \sqrt{i} - T \sum_{i=\lceil T/2 \rceil+1}^{T} \frac{1}{\sqrt{i}} \\
&\ge -1 - \int_{i=1}^{\lceil T/2 \rceil} \frac{1}{\sqrt{x}} dx + \int_{0}^{T} \sqrt{x}\, dx - T \int_{\lceil T/2 \rceil}^{T} \frac{1}{\sqrt{x}} dx \\
&= -1 - 2\left(\sqrt{\lceil T/2 \rceil} - 1\right) + \frac{2}{3} T^{3/2} - 2T \left(\sqrt{T} - \sqrt{\lceil T/2 \rceil}\right) \\
&\ge 1 - 2\sqrt{\lceil T/2 \rceil} + \left(\frac{2}{3} - 2 + \sqrt{2}\right) T^{3/2}\,.
\end{aligned}$$

The last expression is $\Omega(T^{3/2})$ with dominant term $(\frac{2}{3} - 2 + \sqrt{2})T^{3/2} \approx 0.08 \cdot T^{3/2}$. For any $T \ge 42$, the expression is lower bounded by $\frac{1}{20} T^{3/2}$. □

Proof of Theorem 4. Let $e_1, e_2, \dots, e_d$ be the standard orthonormal basis of $\mathbb{R}^d$. Consider the sequence of loss vectors

$$(\ell_1, \ell_2, \dots, \ell_T) = (\underbrace{-e_1, -e_1, \dots, -e_1}_{\lceil T/3 \rceil}, \underbrace{-e_2, -e_2, \dots, -e_2}_{\lfloor 2T/3 \rfloor})\,.$$

First, for any $t \ge \lceil T/3 \rceil + 1$,

$$\begin{aligned}
\frac{w_{t,1}}{w_{t,2}} &= \frac{\exp(-\sum_{i=1}^{t-1} \ell_{i,1}/\sqrt{i})}{\exp(-\sum_{i=1}^{t-1} \ell_{i,2}/\sqrt{i})} \\
&= \frac{\exp(\sum_{i=1}^{\lceil T/3 \rceil} 1/\sqrt{i})}{\exp(\sum_{i=\lceil T/3 \rceil+1}^{t} 1/\sqrt{i})} \\
&\ge \frac{\exp(\sum_{i=1}^{\lceil T/3 \rceil} 1/\sqrt{i})}{\exp(\sum_{i=\lceil T/3 \rceil+1}^{T} 1/\sqrt{i})} \\
&= \exp\left(\sum_{i=1}^{\lceil T/3 \rceil} \frac{1}{\sqrt{i}} - \sum_{i=\lceil T/3 \rceil+1}^{T} 1/\sqrt{i} \right) \\
&\ge \exp\left(\int_{1}^{\lceil T/3 \rceil+1} \frac{dx}{\sqrt{x}} - \int_{\lceil T/3 \rceil}^{T} \frac{dx}{\sqrt{x}} \right) \\
&= \exp\left(2\sqrt{\lceil T/3 \rceil + 1} - 2 - (2\sqrt{T} - 2\sqrt{\lceil T/3 \rceil})\right) \\
&\ge \exp\left(\left(\frac{4}{\sqrt{3}} - 2\right)\sqrt{T} - 2\right) \\
&\ge 4 \,,
\end{aligned}$$

where the last inequality follows from the fact that $\exp\left(\left(\frac{4}{\sqrt{3}} - 2\right)\sqrt{T} - 2\right)$ is an increasing function of T and the inequality can be easily verified for $T = 120$. Since $w_{t,1} + w_{t,1} \le 1$ and $w_{t,1} \ge 0$ and $w_{t,2} \ge 0$, the inequality $w_{t,1}/w_{t,2} \ge 4$ implies that

$$w_{t,2} \le \frac{1}{5} \qquad \text{(for any } t \ge \lceil T/3 \rceil + 1) \,.$$

Now, we lower bound the regret. Since $T \ge 120$,

$$\begin{aligned}
\text{Regret}_T &\ge \text{Regret}_T(e_2) \\
&= \sum_{t=1}^{T} \langle \ell_t, w_t \rangle - \sum_{t=1}^{T} \langle \ell_t, e_2 \rangle \\
&= -\sum_{t=1}^{\lceil T/3 \rceil} w_{t,1} - \sum_{t=\lceil T/3 \rceil+1}^{T} w_{t,2} + \lfloor 2T/3 \rfloor \\
&\ge -\lceil T/3 \rceil - \frac{1}{5}\lfloor 2T/3 \rfloor + \lfloor 2T/3 \rfloor \\
&\ge -T/3 - 1 - 2T/15 + 2T/3 - 1 \\
&= T/5 - 2 \\
&\ge T/6 \,. \quad \square
\end{aligned}$$

Appendix F. Lower bound proof

Proof of Theorem 5. Pick $x, y \in K$ such that $\|x - y\| = D$. This is possible since K is compact. Since $\|x - y\| = \sup\{\langle \ell, x - y \rangle \;:\; \ell \in V^*, \|\ell\|_* = 1\}$ and the set $\{\ell \in V^* \;:\; \|\ell\|_* = 1\}$ is compact, there exists $\ell \in V^*$ such that

$$\|\ell\|_* = 1 \qquad \text{and} \qquad \langle \ell, x - y \rangle = \|x - y\| = D \,.$$

Let $Z_1, Z_2, \dots, Z_T$ be i.i.d. Rademacher variables, that is, $\Pr[Z_t = +1] = \Pr[Z_t = -1] = 1/2$. Let $\ell_t = Z_t a_t \ell$. Clearly, $\|\ell_t\|_* = a_t$. The lemma will be proved if we show that (4) holds with positive probability. We show a stronger statement that the inequality holds in expectation, i.e., $\mathbf{E}[\text{Regret}_T] \ge \frac{D}{\sqrt{8}}\sqrt{\sum_{t=1}^{T} a_t^2}$. Indeed,

$$
\begin{aligned}
\mathbf{E}\left[\text{Regret}_T\right] &\geq \mathbf{E}\left[\sum_{t=1}^{T}\langle \ell_t, w_t\rangle\right] - \mathbf{E}\left[\min_{u\in\{x,y\}}\sum_{t=1}^{T}\langle \ell_t, u\rangle\right] \\
&= \mathbf{E}\left[\sum_{t=1}^{T} Z_t a_t\langle \ell, w_t\rangle\right] + \mathbf{E}\left[\max_{u\in\{x,y\}}\sum_{t=1}^{T} -Z_t a_t\langle \ell, u\rangle\right] \\
&= \mathbf{E}\left[\max_{u\in\{x,y\}}\sum_{t=1}^{T} -Z_t a_t\langle \ell, u\rangle\right] \\
&= \mathbf{E}\left[\max_{u\in\{x,y\}}\sum_{t=1}^{T} Z_t a_t\langle \ell, u\rangle\right] \\
&= \frac{1}{2}\mathbf{E}\left[\sum_{t=1}^{T} Z_t a_t\langle \ell, x+y\rangle\right] + \frac{1}{2}\mathbf{E}\left[\left|\sum_{t=1}^{T} Z_t a_t\langle \ell, x-y\rangle\right|\right] \\
&= \frac{D}{2}\mathbf{E}\left[\left|\sum_{t=1}^{T} Z_t a_t\right|\right] \\
&\geq \frac{D}{\sqrt{8}}\sqrt{\sum_{t=1}^{T} a_t^2}\,,
\end{aligned}
$$

where we used that $\mathbf{E}[Z_t] = 0$, the fact that distributions of Z_t and $-Z_t$ are the same, the formula $\max\{a, b\} = (a+b)/2 + |a-b|/2$, and Khinchin's inequality in the last step (Lemma A.9 in [2]). □

References

[1] F. Orabona, D. Pál, Scale-free algorithms for online linear optimization, in: K. Chaudhuri, C. Gentile, S. Zilles (Eds.), Proceedings of 26th International Conference on Algorithmic Learning Theory, ALT 2015, Banff, AB, Canada, October 4–6, 2015, pp. 287–301.

[2] N. Cesa-Bianchi, G. Lugosi, Prediction, Learning, and Games, Cambridge University Press, Cambridge, 2006.

[3] A. Rakhlin, K. Sridharan, Lecture notes on online learning, available from, http://www-stat.wharton.upenn.edu/~rakhlin/courses/stat991/papers/lecture_notes.pdf, 2009.

[4] S. Shalev-Shwartz, Online learning and online convex optimization, Found. Trends Mach. Learn. 4 (2) (2011) 107–194.

[5] N. Littlestone, M.K. Warmuth, The weighted majority algorithm, Inform. and Comput. 108 (2) (1994) 212–261.

[6] V. Vovk, A game of prediction with expert advice, J. Comput. System Sci. 56 (1998) 153–173.

[7] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. System Sci. 55 (1) (1997) 119–139.

[8] N. Cesa-Bianchi, Y. Freund, D. Haussler, D.P. Helmbold, R.E. Schapire, M.K. Warmuth, How to use expert advice, J. ACM 44 (3) (1997) 427–485.

[9] A. Kalai, S. Vempala, Efficient algorithms for online decision problems, J. Comput. System Sci. 71 (3) (2005) 291–307.

[10] D.P. Helmbold, M.K. Warmuth, Learning permutations with exponential weights, J. Mach. Learn. Res. 10 (2009) 1705–1736.

[11] W.M. Koolen, M.K. Warmuth, J. Kivinen, Hedging structured concepts, in: Proceedings of the 23rd Annual Conference on Computational Learning Theory, COLT, Haifa, Israel, June 27–29, 2010, 2010, pp. 93–105.

[12] J. Abernethy, P.L. Bartlett, A. Rakhlin, A. Tewari, Optimal strategies and minimax lower bounds for online convex games, in: R. Servedio, T. Zhang (Eds.), Proceedings of the 21st Annual Conference on Learning Theory, COLT, Helsinki, Finland, July 9–12, 2008, pp. 415–423.

[13] F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain, Psychol. Rev. 65 (6) (1958) 386.

[14] Y. Freund, R.E. Schapire, Large margin classification using the perceptron algorithm, Mach. Learn. 37 (3) (1999) 277–296.

[15] J. Kivinen, M.K. Warmuth, Exponentiated gradient versus gradient descent for linear predictors, Inform. and Comput. 132 (1) (1997) 1–63.

[16] J. Abernethy, E. Hazan, A. Rakhlin, Competing in the dark: an efficient algorithm for bandit linear optimization, in: R. Servedio, T. Zhang (Eds.), Proceedings of the 21st Annual Conference on Learning Theory, COLT, Helsinki, Finland, July 9–12, 2008, pp. 263–273.

[17] S. Bubeck, N. Cesa-Bianchi, Regret analysis of stochastic and nonstochastic multi-armed bandit problems, Found. Trends Mach. Learn. 5 (1) (2012) 1–122.

[18] A. Nemirovski, D.B. Yudin, Problem Complexity and Method Efficiency in Optimization, Wiley, 1983.

[19] S. Bubeck, Convex optimization: algorithms and complexity, Found. Trends Mach. Learn. 8 (3–4) (2015) 231–357.

[20] M. Zinkevich, Online convex programming and generalized infinitesimal gradient ascent, in: T. Fawcett, N. Mishra (Eds.), Proceedings of 20th International Conference On Machine Learning, ICML 2003, Washington, DC, USA, August 21–24, AAAI Press, 2003, pp. 928–936.

[21] L. Xiao, Dual averaging methods for regularized stochastic learning and online optimization, J. Mach. Learn. Res. 11 (2010) 2543–2596.

[22] H.B. McMahan, J.M. Streeter, Adaptive bound optimization for online convex optimization, in: Proceedings of the 23rd Annual Conference on Computational Learning Theory, COLT, Haifa, Israel, June 27–29, 2010, 2010, pp. 244–256.

[23] H.B. McMahan, Analysis techniques for adaptive online learning, arXiv:1403.3465, 2014.

[24] J. Duchi, E. Hazan, Y. Singer, Adaptive subgradient methods for online learning and stochastic optimization, J. Mach. Learn. Res. 12 (2011) 2121–2159.

[25] S. de Rooij, T. van Erven, P.D. Grünwald, W.M. Koolen, Follow the leader if you can, hedge if you must, J. Mach. Learn. Res. 15 (2014) 1281–1316.

[26] S. Ross, P. Mineiro, J. Langford, Normalized online learning, in: Proceedings of The Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, UAI 2013, 2013, pp. 537–545.

[27] F. Orabona, K. Crammer, N. Cesa-Bianchi, A generalized online mirror descent with applications to classification and regression, Mach. Learn. 99 (2014) 411–435.

[28] S. Shalev-Shwartz, Online Learning: Theory, Algorithms, and Applications, PhD thesis, Hebrew University, Jerusalem, 2007.

[29] Y. Nesterov, Primal-dual subgradient methods for convex problems, Math. Program. 120 (1) (2009) 221–259, appeared early as CORE discussion paper 2005/67, Catholic University of Louvain, Center for Operations Research and Econometrics.
[30] A. Beck, M. Teboulle, Mirror descent and nonlinear projected subgradient methods for convex optimization, Oper. Res. Lett. 31 (3) (2003) 167–175.
[31] H.B. McMahan, G. Holt, D. Sculley, M. Young, D. Ebner, J. Grady, L. Nie, T. Phillips, E. Davydov, D. Golovin, S. Chikkerur, D. Liu, M. Wattenberg, A.M. Hrafnkelsson, T. Boulos, J. Kubica, Ad click prediction: a view from the trenches, in: Proceedings of the 19th International Conference on Knowledge Discovery and Data Mining, KDD 2013, August 11–14, ACM, Chicago, Illinois, USA, 2013, pp. 1222–1230.
[32] J. Duchi, S. Shalev-Shwartz, Y. Singer, A. Tewari, Composite objective mirror descent, in: A.T. Kalai, M. Mohri (Eds.), Proceedings of the 23rd Annual Conference on Computational Learning Theory, COLT, Haifa, Israel, June 27–29, 2010, 2010, pp. 14–26.
[33] M. Streeter, H.B. McMahan, Less regret via online conditioning, arXiv:1002.4862, 2010.
[34] P. Auer, N. Cesa-Bianchi, C. Gentile, Adaptive and self-confident on-line learning algorithms, J. Comput. System Sci. 64 (1) (2002) 48–75.
[35] T. Jaksch, R. Ortner, P. Auer, Near-optimal regret bounds for reinforcement learning, J. Mach. Learn. Res. 11 (2010) 1563–1600.
[36] A. Rakhlin, K. Sridharan, Optimization, learning, and games with predictable sequences, in: C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 26, NIPS 2013, 2013, pp. 3066–3074.
[37] J. Kwon, P. Mertikopoulos, A continuous-time approach to online optimization, arXiv:1401.6956, February 2014.
[38] N. Srebro, K. Sridharan, A. Tewari, On the universality of online mirror descent, in: J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 24, NIPS 2011, 2011, pp. 2645–2653.
[39] H.B. Mcmahan, M. Streeter, No-regret algorithms for unconstrained online convex optimization, in: F. Pereira, C.J.C. Burges, L. Bottou, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 25, NIPS 2012, 2012, pp. 2402–2410.
[40] H.B. McMahan, J. Abernethy, Minimax optimal algorithms for unconstrained linear optimization, in: C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 26, NIPS 2013, 2013, pp. 2724–2732.
[41] F. Orabona, Dimension-free exponentiated gradient, in: C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 26, NIPS 2013, 2013, pp. 1806–1814.
[42] H.B. McMahan, F. Orabona, Unconstrained online linear learning in Hilbert spaces: minimax algorithms and normal approximations, in: M.F. Balcan, C. Szepesvári (Eds.), Proceedings of The 27th Conference on Learning Theory, COLT 2014, vol. 35, 2014, pp. 1020–1039.
[43] F. Orabona, Simultaneous model selection and optimization through parameter-free stochastic learning, in: Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 27, NIPS 2014, 2014, pp. 1116–1124.
[44] F. Orabona, D. Pál, Coin betting and parameter-free online learning, in: D.D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems 29, NIPS 2016, Curran Associates, Inc., 2016, pp. 577–585.
[45] A. Cutkosky, K.A. Boahen, Online convex optimization with unconstrained domains and losses, in: D.D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems 29, NIPS 2016, Curran Associates, Inc., 2016, pp. 748–756.