CrossMark

A generalized online mirror descent with applications to classification and regression

Francesco Orabona · Koby Crammer · Nicolò Cesa-Bianchi

Abstract Online learning algorithms are fast, memory-efficient, easy to implement, and applicable to many prediction problems, including classification, regression, and ranking. Several online algorithms were proposed in the past few decades, some based on additive updates, like the Perceptron, and some on multiplicative updates, like Winnow. A unifying perspective on the design and the analysis of online algorithms is provided by online mirror descent, a general prediction strategy from which most first-order algorithms can be obtained as special cases. We generalize online mirror descent to time-varying regularizers with generic updates. Unlike standard mirror descent, our more general formulation also captures second order algorithms, algorithms for composite losses and algorithms for adaptive filtering. Moreover, we recover, and sometimes improve, known regret bounds as special cases of our analysis using specific regularizers. Finally, we show the power of our approach by deriving a new second order algorithm with a regret bound invariant with respect to arbitrary rescalings of individual features.

1 Introduction

Online learning provides a scalable and flexible approach to the solution of a wide range of prediction problems, including classification, regression, and ranking. Popular online

F. Orabona (✉)
Yahoo Labs, New York, NY 10036, USA
e-mail: francesco@orabona.com

K. Crammer
Department of Electrical Engineering, The Technion, 32000 Haifa, Israel
e-mail: koby@ee.technion.ac.il

N. Cesa-Bianchi
Department of Computer Science, Università degli Studi di Milano, 20135 Milan, Italy
e-mail: nicolo.cesa-bianchi@unimi.it

strategies for classification and regression include first-order gradient-based methods, such as the standard Perceptron algorithm and its many variants (e.g., p-norm Perceptron Gentile 2003 and passive–aggressive Crammer et al. 2006), the Winnow algorithm of Littlestone (1988), the Widrow–Hoff rule, the Exponentiated Gradient algorithm of Kivinen and Warmuth (1997), and many others. A more sophisticated adaptation to the data sequence is achieved by second-order methods, which use the inverse of the empirical feature correlation matrix (or, more generally, the inverse of the Hessian) as an dynamic conditioner for the gradient step. These methods include the Vovk–Azoury–Warmuth algorithm (Vovk 2001; Azoury and Warmuth 2001) (see also (Forster 1999)), the second-order Perceptron (Cesa-Bianchi et al. 2005), the CW/AROW algorithms (Dredze et al. 2008; Crammer et al. 2008, 2009, 2012), the adaptive gradient algorithms proposed by Duchi et al. (2011) and McMahan and Streeter (2010), and the Online Newton Step algorithm of Hazan et al. (2007) for exp-concave losses.

Recently, online convex optimization has been proposed as a common unifying framework for designing and analyzing online algorithms. In particular, online mirror descent (OMD) is a generalized online gradient descent algorithm in which the gradient step is mediated by a strongly convex regularization function. By appropriately choosing the regularizer, most first-order online learning algorithms are recovered as special cases of OMD. Moreover, performance guarantees can be derived simply by instantiating the general OMD bounds to the specific regularizer being used. Note that OMD is the online version of the Mirror Descent algorithm for standard convex optimization (over a fixed convex function) by Nemirovsky and Yudin (1983)—see also (Beck and Teboulle 2003) and (Cesa-Bianchi and Lugosi 2006, Chapt. 11). However, before this connection was made explicit, specific instances of OMD had been discovered and analyzed by Warmuth and Jagota (1997) and Kivinen and Warmuth (2001). These works also pioneered the use of Bregman divergences (ways of measuring distances in linear spaces through convex functions) in the analysis of online algorithms, an approach later extended by Gentile and Warmuth (1998) to non-differentiable loss functions.

A series of recent works (Shalev-Shwartz and Singer 2007; Shalev-Shwartz 2007; Shalev-Shwartz and Kakade 2009) investigates a different approach to online analysis based on primal–dual optimization methods. The work of Kakade et al. (2012) showed indeed that many instances of OMD can be easily analyzed using only a few basic convex duality properties—see the recent survey by Shalev-Shwartz (2012) for a lucid description of these advances. A related algorithm is Follow the Regularized Leader (FTRL), introduced in (Abernethy et al. 2008, 2012),[1] where at each step the prediction is computed as the minimizer of a regularization term plus the sum of losses on all past rounds. When losses are linear, FTRL and OMD are easily seen to be equivalent (Abernethy et al. 2008; Rakhlin and Tewari 2008; Hazan 2011).

In this paper, we extend the theoretical framework of Kakade et al. (2012) by allowing OMD to use time-varying regularizers with updates that do not necessarily use the subgradient of the loss function. Time-varying regularizers have been recently proved to be the key to obtain improved guarantees in unconstrained settings (Streeter and McMahan 2012; Orabona 2013; McMahan and Orabona 2014), and updates that do not use the subgradient of the loss are known to lead to second-order algorithms. Indeed, we show that the

---

[1] There is no clear agreement in the community on the names used to indicate the different algorithms. Hazan (2011) distinguishes between an active and a lazy version of OMD, with the former corresponding to online learning through Bregman divergences, and the latter being equivalent to FTRL for linear losses. Xiao (2010) rediscovered the lazy version of OMD with time-varying regularizers, and—following the optimization terminology—named it regularized dual averaging. McMahan and Streeter (2010) denote any algorithm minimizing a (linearized) sum of the past losses plus a regularization term as FTRL. Here we follow Shalev-Shwartz (2012), viewing OMD as FTRL applied to linearized losses.

Vovk–Azoury–Warmuth algorithm, the second-order Perceptron, and the AROW algorithm are recovered as special cases of our generalized OMD. Our unified analysis is simple, and in certain cases achieves slightly improved theoretical guarantees. Our generalized OMD also captures the efficient variants of second-order classification algorithms that only use the diagonal elements of the feature correlation matrix, a result which was not within reach of previous techniques. Moreover, we also show that a proper choice of the time-varying regularizer allows to cope with the composite setting of (Duchi and Singer 2009; Xiao 2010; Duchi et al. 2010) without using ad-hoc proof techniques.

Our framework improves on previous results even in the case of first-order algorithms. For example, although aggressive algorithms for binary classification often exhibit a better empirical performance than their conservative counterparts, a theoretical explanation of this behavior remained elusive until now. Using our refined analysis, we are able to prove the first bound for passive–aggressive (PA-I) that is never worse (and sometimes better) than the Perceptron bound.

Time-varying regularizers can also be used to perform other types of adaptation to the sequence of observed data. We give a concrete example by introducing new adaptive regularizers corresponding to weighted variants of the standard p-norm regularizer. The resulting updates, and their associated regret bounds, enjoy the property of being invariant with respect to arbitrary rescalings of individual features. Moreover, if the best linear predictor for the loss sequence is sparse, then our analysis delivers a better bound than that of OMD with 1-norm regularization, which is the standard regularizer for the sparse target assumption.

2 Online convex optimization

Let $\mathbb{X}$ be any finite-dimensional linear space equipped with inner product $\langle \cdot , \cdot \rangle$. For example, $\mathbb{X} = \mathbb{R}^d$ where $\langle \cdot , \cdot \rangle$ is the vector dot product, or $\mathbb{X} = \mathcal{M}_{m,n}$, the space of $m \times n$ real matrices with inner product $\langle A , B \rangle = \mathrm{tr}(A^\top B)$.

In the online convex optimization protocol, an algorithm sequentially chooses elements from a convex set $S \subseteq \mathbb{X}$, each time incurring a certain loss. At each step $t = 1, 2, \ldots$ the algorithm chooses $\boldsymbol{w}_t \in S$ and then observes a convex loss function $\ell_t : S \to \mathbb{R}$. The value $\ell_t(\boldsymbol{w}_t)$ is the loss of the learner at step t, and the goal is to control the regret,

$$R_T(\boldsymbol{u}) = \sum_{t=1}^{T} \ell_t(\boldsymbol{w}_t) - \sum_{t=1}^{T} \ell_t(\boldsymbol{u}) \tag{1}$$

for all $\boldsymbol{u} \in S$ and for any sequence of convex loss functions ℓ_t. Important application domains for this protocol are online linear regression and classification. In these settings there is a fixed and known loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and a fixed but unknown sequence $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots$ of examples $(\boldsymbol{x}_t, y_t) \in \mathbb{X} \times \mathbb{R}$. At each step $t = 1, 2, \ldots$ the learner observes $\boldsymbol{x}_t$ and picks $\boldsymbol{w}_t \in S \subseteq \mathbb{X}$. The loss suffered at step t is then defined as $\ell_t(\boldsymbol{w}_t) = \ell(\langle \boldsymbol{w}, \boldsymbol{x}_t \rangle, y_t)$. For example, in regression tasks we might use the square loss $\ell(\langle \boldsymbol{w}, \boldsymbol{x}_t \rangle, y_t) = (\langle \boldsymbol{w}, \boldsymbol{x}_t \rangle - y_t)^2$. In binary classification, where $y_t \in \{-1, +1\}$, a popular loss function is the hinge loss $[1 - y_t \langle \boldsymbol{w}, \boldsymbol{x}_t \rangle]_+$, where $[a]_+ = \max\{0, a\}$. This loss is a convex upper bound on the mistake indicator function $\mathbb{I}_{\{y_t \langle \boldsymbol{w}, \boldsymbol{x}_t \rangle \leq 0\}}$, which is the true quantity of interest.

Note that Mirror Descent, and its online version OMD, can be also defined and analyzed in spaces much more general than those considered here—see, e.g., (Sridharan and Tewari 2010b; Srebro et al. 2011).

2.1 Further notation and definitions

We now introduce the basic notions of convex analysis that are used in the paper—see, e.g., Bauschke and Combettes (2011). We consider functions $f : \mathbb{X} \to \mathbb{R}$ that are closed and convex. This is equivalent to saying that their epigraph $\{(\boldsymbol{x}, y) : f(\boldsymbol{x}) \le y\}$ is a convex and closed subset of $\mathbb{X} \times \mathbb{R}$. The (effective) domain of f, defined by $\{\boldsymbol{x} \in \mathbb{X} : f(\boldsymbol{x}) < \infty\}$, is a convex set whenever f is convex. We can always choose any $S \subseteq \mathbb{X}$ as the domain of f by letting $f(\boldsymbol{x}) = \infty$ for $\boldsymbol{x} \notin S$.

Given a closed and convex function f with domain $S \subseteq \mathbb{X}$, its Fenchel conjugate $f^* : \mathbb{X} \to \mathbb{R}$ is defined by $f^*(\boldsymbol{u}) = \sup_{\boldsymbol{v} \in S} (\langle \boldsymbol{v}, \boldsymbol{u} \rangle - f(\boldsymbol{v}))$. Note that the domain of f^* is always $\mathbb{X}$. Moreover, one can prove that $f^{**} = f$.

A generic norm of a vector $\boldsymbol{u} \in \mathbb{X}$ is denoted by $\|\boldsymbol{u}\|$. Its dual $\|\cdot\|_*$ is the norm defined by $\|\boldsymbol{v}\|_* = \sup_{\boldsymbol{u}} \{\langle \boldsymbol{u}, \boldsymbol{v} \rangle : \|\boldsymbol{u}\| \le 1\}$. The Fenchel-Young inequality states that $f(\boldsymbol{u}) + f^*(\boldsymbol{v}) \ge \langle \boldsymbol{u}, \boldsymbol{v} \rangle$ for all $\boldsymbol{v}, \boldsymbol{u}$.

A vector $\boldsymbol{x}$ is a subgradient of a convex function f at $\boldsymbol{v}$ if $f(\boldsymbol{u}) - f(\boldsymbol{v}) \ge \langle \boldsymbol{u} - \boldsymbol{v}, \boldsymbol{x} \rangle$ for any $\boldsymbol{u}$ in the domain of f. The differential set of f at $\boldsymbol{v}$, denoted by $\partial f(\boldsymbol{v})$, is the set of all the subgradients of f at $\boldsymbol{v}$. If f is also differentiable at $\boldsymbol{v}$, then $\partial f(\boldsymbol{v})$ contains a single vector, denoted by $\nabla f(\boldsymbol{v})$, which is the gradient of f at $\boldsymbol{v}$. A consequence of the Fenchel-Young inequality is the following: for all $\boldsymbol{x} \in \partial f(\boldsymbol{v})$ we have that $f(\boldsymbol{v}) + f^*(\boldsymbol{x}) = \langle \boldsymbol{v}, \boldsymbol{x} \rangle$.

A function f is β-strongly convex with respect to a norm $\|\cdot\|$ if for any $\boldsymbol{u}, \boldsymbol{v}$ in its domain, and any $\boldsymbol{x} \in \partial f(\boldsymbol{u})$,

$$ f(\boldsymbol{v}) \ge f(\boldsymbol{u}) + \langle \boldsymbol{x}, \boldsymbol{v} - \boldsymbol{u} \rangle + \frac{\beta}{2} \|\boldsymbol{u} - \boldsymbol{v}\|^2 . $$

The Fenchel conjugate f^* of a β-strongly convex function f is everywhere differentiable and $\frac{1}{\beta}$-strongly smooth. This means that, for all $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{X}$,

$$ f^*(\boldsymbol{v}) \le f^*(\boldsymbol{u}) + \langle \nabla f^*(\boldsymbol{u}), \boldsymbol{v} - \boldsymbol{u} \rangle + \frac{1}{2\beta} \|\boldsymbol{u} - \boldsymbol{v}\|_*^2 . $$

See also the paper of Kakade et al. (2012) and references therein. A further property of strongly convex functions $f : S \to \mathbb{R}$ is the following: for all $\boldsymbol{u} \in \mathbb{X}$,

$$ \nabla f^*(\boldsymbol{u}) = \operatorname*{argmax}_{\boldsymbol{v} \in S} \big(\langle \boldsymbol{v}, \boldsymbol{u} \rangle - f(\boldsymbol{v}) \big). \tag{2} $$

This implies the useful identity

$$ f\big(\nabla f^*(\boldsymbol{u})\big) + f^*(\boldsymbol{u}) = \big\langle \nabla f^*(\boldsymbol{u}), \boldsymbol{u} \big\rangle. \tag{3} $$

Strong convexity and strong smoothness are key properties in the design of online learning algorithms.

3 Online mirror descent

We now introduce our main algorithmic tool: a generalization of the OMD algorithm in which the regularizer may change over time. The standard OMD algorithm for online convex optimization—see, e.g., (Shalev-Shwartz 2012)—sets $\boldsymbol{w}_t = \nabla f^*(\boldsymbol{\theta}_t)$, where f^* is a strongly convex regularizer and $\boldsymbol{\theta}_t$ is updated using subgradient descent: $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \boldsymbol{\ell}'_t$ for $\eta > 0$ and $\boldsymbol{\ell}'_t \in \partial \ell_t(\boldsymbol{w}_t)$. For instance, if $f = \frac{1}{2} \|\cdot\|_2^2$, then $f^* = f$ and OMD specializes to standard online subgradient descent.

We generalize OMD in two ways: first, we allow f to change over time; second, we do not necessarily use the subgradient of the loss to update $\boldsymbol{\theta}_t$, but rather use an input sequence of generic elements $\boldsymbol{z}_t$. The resulting algorithm is summarized in Algorithm 1.

Algorithm 1 Online Mirror Descent

1: **Parameters:** A sequence of strongly convex functions $f_1, f_2, \ldots$ defined on a common convex domain $S \subseteq \mathbb{X}$.
2: **Initialize:** $\boldsymbol{\theta}_1 = \mathbf{0} \in \mathbb{X}$
3: **for** $t = 1, 2, \ldots$ **do**
4:     Choose $\boldsymbol{w}_t = \nabla f_t^*(\boldsymbol{\theta}_t)$
5:     Observe $\boldsymbol{z}_t \in \mathbb{X}$
6:     Update $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \boldsymbol{z}_t$
7: **end for**

Note the following remarkable property: while $\boldsymbol{\theta}_t$ follows an arbitrary trajectory in $\mathbb{X}$, as determined by the input sequence $\boldsymbol{z}_t$, because of (2) the property $\boldsymbol{w}_t \in S$ holds uniformly over t.

OMD with specific time-varying regularizers was implicitly analyzed by Vovk (2001) and independently by Azoury and Warmuth (2001), who introduced the proof ideas currently used here. Another early example of OMD with time-varying regularizers is due to Bartlett et al. (2007). A more explicit analysis is contained in the work of Sridharan and Tewari (2010a). The following lemma is a generalization of two corollaries of (Kakade et al. 2012, Corollary 4) and of (Duchi et al. 2011, Corollary 3).

Lemma 1 *Assume OMD is run with functions $f_1, f_2, \ldots, f_T$ defined on a common convex domain $S \subseteq \mathbb{X}$ and such that each f_t is β_t-strongly convex with respect to the norm $\|\cdot\|_t$. Let $\|\cdot\|_{t,*}$ be the dual norm of $\|\cdot\|_t$, for $t = 1, 2, \ldots, T$. Then, for any $\boldsymbol{u} \in S$,*

$$\sum_{t=1}^{T} \langle \boldsymbol{z}_t, \boldsymbol{u} - \boldsymbol{w}_t \rangle \leq f_T(\boldsymbol{u}) + \sum_{t=1}^{T} \left(\frac{\|\boldsymbol{z}_t\|_{t,*}^2}{2\beta_t} + f_t^*(\boldsymbol{\theta}_t) - f_{t-1}^*(\boldsymbol{\theta}_t) \right)$$

where we set $f_0^*(\mathbf{0}) = 0$. Moreover, for all $t \geq 1$, we have

$$f_t^*(\boldsymbol{\theta}_t) - f_{t-1}^*(\boldsymbol{\theta}_t) \leq f_{t-1}(\boldsymbol{w}_t) - f_t(\boldsymbol{w}_t). \tag{4}$$

Proof Let $\Delta_t = f_t^*(\boldsymbol{\theta}_{t+1}) - f_{t-1}^*(\boldsymbol{\theta}_t)$. Then

$$\sum_{t=1}^{T} \Delta_t = f_T^*(\boldsymbol{\theta}_{T+1}) - f_0^*(\boldsymbol{\theta}_1) = f_T^*(\boldsymbol{\theta}_{T+1}).$$

Since the functions f_t^* are $\frac{1}{\beta_t}$-strongly smooth with respect to $\|\cdot\|_{t,*}$, and recalling that $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \boldsymbol{z}_t$,

$$\Delta_t = f_t^*(\boldsymbol{\theta}_{t+1}) - f_t^*(\boldsymbol{\theta}_t) + f_t^*(\boldsymbol{\theta}_t) - f_{t-1}^*(\boldsymbol{\theta}_t)$$

$$\leq f_t^*(\boldsymbol{\theta}_t) - f_{t-1}^*(\boldsymbol{\theta}_t) + \langle \nabla f_t^*(\boldsymbol{\theta}_t), \boldsymbol{z}_t \rangle + \frac{1}{2\beta_t} \|\boldsymbol{z}_t\|_{t,*}^2$$

$$= f_t^*(\boldsymbol{\theta}_t) - f_{t-1}^*(\boldsymbol{\theta}_t) + \langle \boldsymbol{w}_t, \boldsymbol{z}_t \rangle + \frac{1}{2\beta_t} \|\boldsymbol{z}_t\|_{t,*}^2$$

where we used the definition of $\boldsymbol{w}_t$ in the last step. On the other hand, the Fenchel-Young inequality implies

$$\sum_{t=1}^{T} \Delta_t = f_T^*(\boldsymbol{\theta}_{T+1}) \geq \langle \boldsymbol{u}, \boldsymbol{\theta}_{T+1} \rangle - f_T(\boldsymbol{u}) = \sum_{t=1}^{T} \langle \boldsymbol{u}, \boldsymbol{z}_t \rangle - f_T(\boldsymbol{u}).$$

Combining the upper and lower bound on Δ_t and summing over t we get

$$\sum_{t=1}^{T} \langle \boldsymbol{u}, \boldsymbol{z}_t \rangle - f_T(\boldsymbol{u}) \leq \sum_{t=1}^{T} \Delta_t \leq \sum_{t=1}^{T} \left(f_t^*(\boldsymbol{\theta}_t) - f_{t-1}^*(\boldsymbol{\theta}_t) + \langle \boldsymbol{w}_t, \boldsymbol{z}_t \rangle + \frac{1}{2\beta_t} \|\boldsymbol{z}_t\|_{t,*}^2 \right).$$

We now prove the second statement. Recalling again the definition of $\boldsymbol{w}_t$, we have that (3) implies $f_t^*(\boldsymbol{\theta}_t) = \langle \boldsymbol{w}_t, \boldsymbol{\theta}_t \rangle - f_t(\boldsymbol{w}_t)$. On the other hand, the Fenchel-Young inequality implies that $-f_{t-1}^*(\boldsymbol{\theta}_t) \leq f_{t-1}(\boldsymbol{w}_t) - \langle \boldsymbol{w}_t, \boldsymbol{\theta}_t \rangle$. Combining the two we get $f_t^*(\boldsymbol{\theta}_t) - f_{t-1}^*(\boldsymbol{\theta}_t) \leq f_{t-1}(\boldsymbol{w}_t) - f_t(\boldsymbol{w}_t)$, as desired.   □

We are now ready to prove regret bounds for OMD applied to three different classes of time-varying regularizers.

Corollary 1 *Let S a convex set, $F : S \rightarrow \mathbb{R}$ be a convex function, and let $g_1, g_2, \ldots, g_T$ be a sequence of convex functions $g_t : S \rightarrow \mathbb{R}$ such that $g_t(\boldsymbol{u}) \leq g_{t+1}(\boldsymbol{u})$ for all $t = 1, 2, \ldots, T$ and all $\boldsymbol{u} \in S$. Fix $\eta > 0$ and assume $f_t = g_t + \eta t F$ are β_t-strongly convex w.r.t. $\|\cdot\|_{f_t}$. For each $t = 1, 2, \ldots, T$ let $\|\cdot\|_{t,*}$ be the dual norm of $\|\cdot\|_t$. If OMD is run on the input sequence $\boldsymbol{z}_t = -\eta \boldsymbol{\ell}_t'$ for some $\boldsymbol{\ell}_t' \in \partial \ell_t(\boldsymbol{w}_t)$, then*

$$\sum_{t=1}^{T} \left(\ell_t(\boldsymbol{w}_t) + F(\boldsymbol{u}) \right) - \sum_{t=1}^{T} \left(\ell_t(\boldsymbol{u}) + F(\boldsymbol{u}) \right) \leq \frac{g_T(\boldsymbol{u})}{\eta} + \eta \sum_{t=1}^{T} \frac{\|\boldsymbol{\ell}_t'\|_{t,*}^2}{2\beta_t} \tag{5}$$

for all $\boldsymbol{u} \in S$.
Moreover, if $f_t = g\sqrt{t} + \eta t F$ where $g : S \rightarrow \mathbb{R}$ is β-strongly convex w.r.t. $\|\cdot\|$, then

$$\sum_{t=1}^{T} \left(\ell_t(\boldsymbol{w}_t) + F(\boldsymbol{u}) \right) - \sum_{t=1}^{T} \left(\ell_t(\boldsymbol{u}) + F(\boldsymbol{u}) \right) \leq \sqrt{T} \left(\frac{g(\boldsymbol{u})}{\eta} + \frac{\eta}{\beta} \max_{t \leq T} \|\boldsymbol{\ell}_t'\|_*^2 \right) \tag{6}$$

for all $\boldsymbol{u} \in S$.
Finally, if $f_t = t F$, where F is β-strongly convex w.r.t. $\|\cdot\|$, then

$$\sum_{t=1}^{T} \left(\ell_t(\boldsymbol{w}_t) + F(\boldsymbol{u}) \right) - \sum_{t=1}^{T} \left(\ell_t(\boldsymbol{u}) + F(\boldsymbol{u}) \right) \leq \max_{t \leq T} \|\boldsymbol{\ell}_t'\|_*^2 \frac{(1 + \ln T)}{2\beta} \tag{7}$$

for all $\boldsymbol{u} \in S$.

Proof By convexity, $\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{u}) \leq \frac{1}{\eta} \langle \boldsymbol{z}_t, \boldsymbol{u} - \boldsymbol{w}_t \rangle$. Using Lemma 1 we have,

$$\sum_{t=1}^{T} \langle \boldsymbol{z}_t, \boldsymbol{u} - \boldsymbol{w}_t \rangle \leq g_T(\boldsymbol{u}) + \eta T F(\boldsymbol{u}) + \eta^2 \sum_{t=1}^{T} \frac{\|\boldsymbol{\ell}_t'\|_{t,*}^2}{2\beta_t} + \eta \sum_{t=1}^{T} \left((t-1) F(\boldsymbol{w}_t) - t F(\boldsymbol{w}_t) \right)$$

where we used the fact that the terms $g_{t-1}(\boldsymbol{w}_t) - g_t(\boldsymbol{w}_t)$ are nonpositive as per our assumption. Reordering terms we obtain (5). In order to obtain (6) it is sufficient to note that, by definition of strong convexity, $g\sqrt{t}$ is $\beta\sqrt{t}$-strongly convex because g is β-strongly convex, hence f_t is $\beta\sqrt{t}$-strongly convex too. The elementary inequality $\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T}$ concludes the proof

of (6). Finally, bound (7) is proven by observing that $f_t = t F$ is βt-strongly convex because F is β-strongly convex. The elementary inequality $\sum_{t=1}^{T} \frac{1}{t} \leq 1 + \ln T$ concludes the proof. $\square$

Note that the regret bounds obtained in Corollary 1 are for the composite setting, where the algorithm minimizes the sum $\ell_t(\cdot) + F(\cdot)$ of two functions, where the first one is typically a loss and the other is a regularization term. Here, the only hypothesis on F is convexity, hence F can be a nondifferentiable function, like $\|\cdot\|_1$ for inducing sparse solutions. While the composite setting is considered more difficult than the standard one, and requires specific ad-hoc algorithms (Duchi and Singer 2009; Xiao 2010; Duchi et al. 2010), here we show that this setting can be efficiently solved using OMD with a specific choice of the time-varying regularizer. Thus, we recover the results about minimization of strongly convex and composite loss functions, and adaptive learning rates, in a simple unified framework. Note that (4), which was missing in previous analyses of OMD, is the key result to obtain this level of generality.

A special case of OMD is the regularized dual averaging framework of Xiao (2010), where the prediction at each step is defined by

$$\boldsymbol{w}_t = \underset{\boldsymbol{w}}{\operatorname{argmin}} \; \frac{1}{t-1} \sum_{s=1}^{t-1} \boldsymbol{w}^\top \boldsymbol{\ell}_s' + \frac{\beta_{t-1}}{t-1} g(\boldsymbol{w}) + F(\boldsymbol{w}) \tag{8}$$

for some $\boldsymbol{\ell}_s' \in \partial \ell_s(\boldsymbol{w}_s), s = 1, \dots, t-1$. Using (2), it is easy to see that this update is equivalent[2] to $\boldsymbol{w}_t = \nabla f_t^* \left(\sum_{s=1}^{t-1} \boldsymbol{\ell}_s' \right)$, where $f_t(\boldsymbol{w}) = \beta_{t-1} g(\boldsymbol{w}) + (t-1) F(\boldsymbol{w})$. The framework of Xiao (2010) has been extended by Duchi et al. (2010) to allow the strongly convex part of the regularizer to increase over time. A bound similar to (5) has been also recently presented by Duchi et al. (2010). There, a more immediate trade-off between the current gradient and the Bregman divergence from the new solution to the previous one is used to update at each time step. However, in both cases their analysis is not flexible enough to include algorithms whose update does not use the sub-gradient of the loss function. Examples of such algorithms are the Vovk–Azoury–Warmuth algorithm of the next section and the online binary classification algorithms of Sect. 6.

4 Online regression with square loss

In this section we apply Lemma 1 to recover known regret bounds for online regression and adaptive filtering with the square loss. For simplicity, we set $\mathbb{X} = \mathbb{R}^d$ and let the inner product $\langle \boldsymbol{u}, \boldsymbol{x} \rangle$ be the standard dot product $\boldsymbol{u}^\top \boldsymbol{x}$. We also set $\ell_t(\boldsymbol{u}) = \frac{1}{2} \left(y_t - \boldsymbol{u}^\top \boldsymbol{x}_t \right)^2$ where $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots$ is some arbitrary sequence of examples $(\boldsymbol{x}_t, y_t) \in \mathbb{R}^d \times \mathbb{R}$.

First, note that it is possible to specialize OMD to the Vovk–Azoury–Warmuth algorithm for online regression. Remember that, at each time step t, the Vovk–Azoury–Warmuth algorithm predicts with

$$\boldsymbol{w}_t = \underset{\boldsymbol{w}}{\operatorname{argmin}} \; \frac{a}{2} \|\boldsymbol{w}\|^2 + \frac{1}{2} \sum_{s=1}^{t-1} \left(y_s - \boldsymbol{w}^\top \boldsymbol{x}_s \right)^2 + \frac{1}{2} \left(\boldsymbol{w}^\top \boldsymbol{x}_t \right)^2$$

---

[2] Although Xiao (2010) explicitly mentions that his results cannot be recovered with the primal–dual proofs, here we prove the contrary.

$$= \operatorname*{argmin}_{\boldsymbol{w}} \frac{a}{2}\|\boldsymbol{w}\|^2 + \frac{1}{2}\sum_{s=1}^{t}(\boldsymbol{w}^\top \boldsymbol{x}_s)^2 - \sum_{s=1}^{t-1} y_s \boldsymbol{w}^\top \boldsymbol{x}_s$$

$$= \operatorname*{argmin}_{\boldsymbol{w}} \frac{1}{2}\boldsymbol{w}^\top \left(aI + \sum_{i=1}^{t} \boldsymbol{x}_s \boldsymbol{x}_s^\top\right)\boldsymbol{w} - \sum_{s=1}^{t-1} y_s \boldsymbol{w}^\top \boldsymbol{x}_s$$

$$= \left(aI + \sum_{s=1}^{t} \boldsymbol{x}_s \boldsymbol{x}_s^\top\right)^{-1} \sum_{i=1}^{t-1} y_s \boldsymbol{x}_s.$$

Now, by letting $A_0 = a\,I_d$, $A_t = A_{t-1} + \boldsymbol{x}_t \boldsymbol{x}_t^\top$ for $t \geq 1$, and $\boldsymbol{z}_s = y_s \boldsymbol{x}_s$, we obtain the OMD update $\boldsymbol{w}_t = A_t^{-1}\boldsymbol{\theta}_t = \nabla f_t^*(\boldsymbol{\theta}_t)$, where $f_t(\boldsymbol{u}) = \frac{1}{2}\boldsymbol{u}^\top A_t \boldsymbol{u}$ and $f_t^*(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^\top A_t^{-1}\boldsymbol{\theta}$. Note that $\boldsymbol{z}_t$ is *not* equal to the negative gradient of the square loss. In fact, the special structure of the square loss allows us to move some of the terms inside the regularizer, and use proxies for the gradients of the losses. The regret bound of this algorithm—see, e.g., Cesa-Bianchi and Lugosi (2006, Theorem 11.8)—is recovered from Lemma 1 by noting that f_t is 1-strongly convex with respect to the norm $\|\boldsymbol{u}\|_t = \sqrt{\boldsymbol{u}^\top A_t \boldsymbol{u}}$. Hence, the regret (1) is controlled as follows

$$R_T(\boldsymbol{u}) = \frac{1}{2}\sum_{t=1}^{T}\left(y_t - \boldsymbol{w}_t^\top \boldsymbol{x}_t\right)^2 - \frac{1}{2}\sum_{t=1}^{T}\left(y_t - \boldsymbol{u}^\top \boldsymbol{x}_t\right)^2$$

$$= \sum_{t=1}^{T}\left(y_t \boldsymbol{u}^\top \boldsymbol{x}_t - y_t \boldsymbol{w}_t^\top \boldsymbol{x}_t\right) - f_T(\boldsymbol{u}) + \frac{a}{2}\|\boldsymbol{u}\|^2 + \frac{1}{2}\sum_{t=1}^{T}\left(\boldsymbol{w}_t^\top \boldsymbol{x}_t\right)^2$$

$$\leq f_T(\boldsymbol{u}) + \sum_{t=1}^{T}\left(\frac{y_t^2 \|\boldsymbol{x}_t\|_{t,*}^2}{2} + f_t^*(\boldsymbol{\theta}_t) - f_{t-1}^*(\boldsymbol{\theta}_t)\right) - f_T(\boldsymbol{u}) + \frac{a}{2}\|\boldsymbol{u}\|^2$$

$$\quad + \frac{1}{2}\sum_{t=1}^{T}\left(\boldsymbol{w}_t^\top \boldsymbol{x}_t\right)^2$$

$$\leq \frac{a}{2}\|\boldsymbol{u}\|^2 + \frac{Y^2}{2}\sum_{t=1}^{T}\boldsymbol{x}_t^\top A_t^{-1}\boldsymbol{x}_t$$

since $f_t^*(\boldsymbol{\theta}_t) - f_{t-1}^*(\boldsymbol{\theta}_t) \leq f_{t-1}(\boldsymbol{w}_t) - f_t(\boldsymbol{w}_t) = -\frac{1}{2}(\boldsymbol{w}_t^\top \boldsymbol{x}_t)^2$, and where $Y = \max_t |y_t|$.

A related setting is that of adaptive filtering—see, e.g., (Kivinen et al. 2006) and references therein. In this setting the output signals y_t are given by $y_t = \boldsymbol{u}^\top \boldsymbol{x}_t + \nu_t$, where $\boldsymbol{u}$ is unknown and ν_t is some arbitrary noise process. The goal is to recover the uncorrupted output $\boldsymbol{u}^\top \boldsymbol{x}_t$. This can be achieved by minimizing a suitable notion of regret, namely the adaptive filtering regret w.r.t. the square loss,

$$R_T^{\mathrm{AF}}(\boldsymbol{u}) := \sum_{t=1}^{T}\left(\boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{u}^\top \boldsymbol{x}_t\right)^2.$$

Kivinen et al. (2006) introduced the p-norm LMS algorithm, addressing adaptive filtering problems in a purely online nonstochastic setting. We now show that OMD can be easily applied to online adaptive filtering. First, note that

$$R_T(\boldsymbol{u}) + \frac{1}{2} R_T^{\mathrm{AF}}(\boldsymbol{u}) = \sum_{t=1}^{T} \left((y_t - \boldsymbol{w}_t^\top \boldsymbol{x}_t)^2 - (y_t - \boldsymbol{u}^\top \boldsymbol{x}_t)^2 + \frac{1}{2} (\boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{u}^\top \boldsymbol{x}_t)^2 \right)$$

$$= \sum_{t=1}^{T} \left((y_t - \boldsymbol{w}_t^\top \boldsymbol{x}_t) \boldsymbol{u}^\top \boldsymbol{x}_t - (y_t - \boldsymbol{w}_t^\top \boldsymbol{x}_t) \boldsymbol{w}_t^\top \boldsymbol{x}_t \right)$$

$$= \sum_{t=1}^{T} (\boldsymbol{u} - \boldsymbol{w}_t)^\top \boldsymbol{z}_t \qquad (9)$$

where we set $\boldsymbol{z}_t = -(y_t - \boldsymbol{w}_t^\top \boldsymbol{x}_t) \boldsymbol{x}_t$. Now, pick any function f which is 1-strongly convex with respect to some norm $\|\cdot\|$ and let $f_t(\boldsymbol{u}) = X_t^2 f(\boldsymbol{u})$, where $X_t = \max_{s \le t} \|\boldsymbol{x}_s\|_*$. Lemma 1 then immediately implies that

$$\sum_{t=1}^{T} (\boldsymbol{u} - \boldsymbol{w}_t)^\top \boldsymbol{z}_t \le f_T(\boldsymbol{u}) + \frac{1}{2} \sum_{t=1}^{T} (y_t - \boldsymbol{w}_t^\top \boldsymbol{x}_t)^2 \qquad (10)$$

where we used the X_t^2-strong convexity of f_t and the fact the $f_t \ge f_{t-1}$. Combining (9) with (10) and simplifying, we obtain the following adaptive filtering bound

$$R_T^{\mathrm{AF}}(\boldsymbol{u}) \le 2 X_T^2 f(\boldsymbol{u}) + \sum_{t=1}^{T} (y_t - \boldsymbol{u}^\top \boldsymbol{x}_t)^2.$$

This is a direct generalization to arbitrary regularizers f of the bound by Kivinen et al. (2006). However, their algorithm requires the prior knowledge of the maximum norm of $\boldsymbol{x}_t$ to set a critical parameter. Instead, our algorithm, through the use of an increasing regularizer, has the ability to adapt to the maximum norm of $\boldsymbol{x}_t$, without using any prior knowledge.

5 Scale-invariant algorithms

In this section we show the full power of our framework by introducing two new scale-invariant algorithms for online linear regression with an arbitrary convex and Lipschitz loss function.

Recall the online linear regression setting: given a convex loss $\ell : \mathbb{R}^2 \to \mathbb{R}$ and a fixed but unknown sequence $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots$ of examples, at each step $t = 1, 2, \dots$ the online algorithm observes $\boldsymbol{x}_t$ and picks $\boldsymbol{w}_t$. The associated loss is $\ell_t(\boldsymbol{w}_t) = \ell(\boldsymbol{w}_t^\top \boldsymbol{x}_t, y_t)$.

Let $\boldsymbol{u} \in \mathbb{R}^d$ be any fixed predictor with small total loss $\ell_1(\boldsymbol{u}) + \dots + \ell_T(\boldsymbol{u})$. Because of linearity, an arbitrary rescaling of any individual feature, $x_{t,i} \to c\, x_{t,i}$ for $t = 1, \dots, T$ (while $y_1, \dots, y_T$ are kept fixed) can be offset by a corresponding rescaling of u_i without affecting the total loss. We might ask a similar scale-invariance for the online predictor. In other words, we would like the online algorithm to be independent of the units in which each data coordinate is expressed.

We now introduce two new time-varying regularizers that achieve this goal. As in the previous section, let $\mathbb{X} = \mathbb{R}^d$ and let the inner product $\langle \boldsymbol{u}, \boldsymbol{x} \rangle$ be the standard dot product $\boldsymbol{u}^\top \boldsymbol{x}$.

The new regularizers are based on the following generalization of the squared q-norm. Given $(a_1, \dots, a_d) \in \mathbb{R}_+$ and $q \in (1, 2]$ define the weighted q-norm of $\boldsymbol{w} \in \mathbb{R}^d$ by

$$\left(\sum_{i=1}^{d} |w_i|^q a_i\right)^{1/q}.$$

Define the corresponding regularization function by

$$f(\boldsymbol{w}) = \frac{1}{2(q-1)} \left(\sum_{i=1}^{d} |w_i|^q a_i\right)^{2/q}.$$

This function has the following properties (proof in the "Appendix").

Lemma 2 *The Fenchel conjugate of f is*

$$f^*(\boldsymbol{\theta}) = \frac{1}{2(p-1)} \left(\sum_{i=1}^{d} |\theta_i|^p a_i^{1-p}\right)^{2/p} \quad \text{for} \quad p = \frac{q}{q-1}. \tag{11}$$

Moreover, the function $f(\boldsymbol{w})$ is 1-strictly convex with respect to the norm

$$\left(\sum_{i=1}^{d} |x_i|^q a_i\right)^{1/q}$$

whose dual norm is defined by

$$\left(\sum_{i=1}^{d} |\theta_i|^p a_i^{1-p}\right)^{1/p}.$$

Given an arbitrary sequence $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots$ of examples, we assume OMD is run with $z_t = -\eta \boldsymbol{\ell}'_t$ where, as usual, $\boldsymbol{\ell}'_t \in \partial \ell_t(\boldsymbol{w}_t)$. We also assume the loss ℓ is such that $\ell(\cdot, y)$ is L-Lipschitz for each $y \in \mathbb{R}$ and $\ell_t(\boldsymbol{w}) = \ell(\boldsymbol{w}^\top \boldsymbol{x}_t, y_t)$ is convex for all t. In the rest of this section, the following notation is used: $b_{t,i} = \max_{s=1,\ldots,t} |x_{s,i}|$, $m_t = \max_{s=1,\ldots,t} \|\boldsymbol{x}_s\|_0$, $p_t = 2 \ln m_t$, and

$$\beta_t = \sqrt{eL^2(p_t - 1) + \sum_{s=1}^{t-1}(p_s - 1)\left(\sum_{i=1}^{d}\left(\frac{|\ell'_{s,i}|}{b_{s,i}}\right)^{p_s}\right)^{2/p_s}}.$$

The time-varying regularizers we consider are defined as follows,

$$f_t(\boldsymbol{u}) = \frac{\beta_t}{2}\left(\sum_{i=1}^{d}(|u_i|b_{t,i})^{q_t}\right)^{2/q_t} \quad \text{for} \quad q_t = \frac{p_t}{p_t - 1} \tag{12}$$

$$f_t(\boldsymbol{u}) = \frac{\sqrt{d}}{2}\left(\sum_{i=1}^{d}(|u_i|b_{t,i})^2 \sqrt{L^2 + \sum_{s=1}^{t-1}\left(\frac{\ell'_{s,i}}{b_{s,i}}\right)^2}\right). \tag{13}$$

As we show next, regularizers of type (12) give regret bounds that exhibit a logarithmic dependency on the maximum number of non-zero components observed. Instead, regularizers of type (13) give bounds that depend on $\sqrt{d}$, and used a different learning rate for each coordinate. Roughly speaking, the first regularizer provides scale-invariance with a logarithmic dependency on the dimension d obtained through the p-norm regularization (Gentile 2003). The second regularizer, instead, corresponds to a scale invariant version of AdaGrad (Duchi et al. 2011).

In order to spell out the OMD update, we compute the derivative of the Fenchel dual of the regularization functions. Using the fact that if $g(\boldsymbol{w}) = af(\boldsymbol{w})$, then $g^*(\boldsymbol{\theta}) = af^*(\frac{\theta}{a})$, for regularizers of type (12) we have

$$\left(\nabla f_t^*(\boldsymbol{\theta})\right)_j = \frac{1}{\beta_t(p_t - 1)}\left(\sum_{i=1}^{d}\left(\frac{|\theta_i|}{b_{t,i}}\right)^{p_t}\right)^{2/p_t - 1}\frac{|\theta_j|^{p_t - 1}}{b_{t,j}^{p_t}}\operatorname{sign}(\theta_j). \tag{14}$$

For regularizers of type (13) we have

$$\left(\nabla f_t^*(\boldsymbol{\theta})\right)_j = \frac{\theta_j}{b_{t,j}^2\sqrt{d}\sqrt{L^2 + \sum_{s=1}^{t-1}\left(\frac{\ell_{s,j}'}{b_{s,j}}\right)^2}}. \tag{15}$$

These updates are such that $\boldsymbol{w}_t^\top \boldsymbol{x}_t$ is independent of arbitrary rescalings of individual features. To see this, recall that $\boldsymbol{\theta}_t = -\eta\left(\boldsymbol{\ell}_1' + \cdots + \boldsymbol{\ell}_{t-1}'\right)$ and

$$\boldsymbol{\ell}_s' = \left.\frac{\partial \ell(z, y_s)}{\partial z}\right|_{z = \boldsymbol{w}_s^\top \boldsymbol{x}_s}\boldsymbol{x}_s$$

where the partial derivative is at most L by Lipschitzness of ℓ. Hence, the ratios $|\ell_{t,j}'|/b_{t,j}$ and $|\theta_{t,j}|/b_{t,j}$ are invariant with respect to arbitrary rescalings of the jth feature. So, in both (14) and (15), $w_{t,j}$ scales as $1/b_{t,j}$, and we have that $\boldsymbol{w}_t^\top \boldsymbol{x}_t$ is invariant to the rescaling of individual coordinates.

The formula in the right-hand side of (15) also shows that, similar to the updates studied in (McMahan and Streeter 2010; Duchi et al. 2011), the second type of regularizer induces a different learning rate for each component of $\boldsymbol{w}_t$.

We now prove the following regret bounds.

Theorem 1 *If OMD is run using regularizers of type* (12)*, then for any* $\boldsymbol{u} \in \mathbb{R}^d$

$$R_T(\boldsymbol{u}) \leq L\sqrt{e(T+1)(2\ln m_T - 1)}\left(\frac{1}{2\eta}\left(\sum_{i=1}^{d}|u_i|b_{T,i}\right)^2 + \eta\right).$$

If OMD is run using regularizers of type (13)*, then for any* $\boldsymbol{u} \in \mathbb{R}^d$

$$R_T(\boldsymbol{u}) \leq L\sqrt{d(T+1)}\left(\frac{1}{2\eta}\sum_{i=1}^{d}\left(u_i b_{T,i}\right)^2 + \eta\right).$$

Proof For the first algorithm, note that $m_t^{2/p_t} = e$, and setting $q_t = \left(1 - \frac{1}{p_t}\right)^{-1}$, we have $q_t(1 - p_t) = -p_t$. Further note that $f_t^*(\boldsymbol{\theta}_t) - f_{t-1}^*(\boldsymbol{\theta}_t) \leq f_{t-1}(\boldsymbol{w}_t) - f_t(\boldsymbol{w}_t) \leq 0$, where $f_{t-1} \leq f_t$ because q_t is decreasing, $b_{t,i}$ is increasing, and β_t is also increasing. Hence, using the convexity of ℓ_t and Lemma 1, we may write

$$R_T(\boldsymbol{u}) \leq \sum_{t=1}^{T}(\boldsymbol{\ell}_t')^\top\left(\boldsymbol{u} - \boldsymbol{w}_t\right)$$

$$\leq \frac{\beta_T}{2\eta}\left(\sum_{i=1}^{d}\left(|u_i|b_{T,i}\right)^{q_T}\right)^{2/q_T} + \eta\sum_{t=1}^{T}\frac{1}{2\beta_t(q_t - 1)}\left(\sum_{i=1}^{d}\frac{|\ell_{t,i}'|^{p_t}}{b_{t,i}^{p_t}}\right)^{2/p_t}. \tag{16}$$

For the second term in (16), using the fact that $\frac{1}{q_t-1} = p_t - 1$ and the L-Lipschitzness, we have

$$\sum_{t=1}^{T} \frac{1}{2\beta_t(q_t-1)} \left(\sum_{i=1}^{d} \frac{|\ell'_{t,i}|^{p_t}}{b_{t,i}^{p_t}} \right)^{2/p_t}$$

$$= \frac{1}{2} \sum_{t=1}^{T} \frac{(p_t-1) \left(\sum_{i=1}^{d} \frac{|\ell'_{t,i}|^{p_t}}{b_{t,i}^{p_t}} \right)^{2/p_t}}{\sqrt{L^2 m_t^{2/p_t}(p_t-1) + \sum_{s=1}^{t-1}(p_s-1) \left(\sum_{i=1}^{d} \frac{|\ell'_{s,i}|^{p_s}}{b_{s,i}^{p_s}} \right)^{2/p_s}}}$$

$$\leq \frac{1}{2} \sum_{t=1}^{T} \frac{(p_t-1) \left(\sum_{i=1}^{d} \frac{|\ell'_{t,i}|^{p_t}}{b_{t,i}^{p_t}} \right)^{2/p_t}}{\sqrt{\sum_{s=1}^{t}(p_s-1) \left(\sum_{i=1}^{d} \frac{|\ell'_{s,i}|^{p_s}}{b_{s,i}^{p_s}} \right)^{2/p_s}}}$$

$$\leq \sqrt{\sum_{t=1}^{T}(p_t-1) \left(\sum_{i=1}^{d} \frac{|\ell'_{t,i}|^{p_t}}{b_{t,i}^{p_t}} \right)^{\frac{2}{p_t}}}$$

$$\leq \beta_{T+1}$$

where the second inequality uses the elementary inequality

$$\sum_{t=1}^{T} \frac{a_t}{\sqrt{\sum_{s=1}^{t} a_s}} \leq 2\sqrt{\sum_{t=1}^{T} a_t} \quad \text{for} \quad a_1, \ldots, a_T \geq 0 \tag{17}$$

(see, e.g., Auer et al. 2002, Lemma 3.5). Hence we have

$$R_T(\boldsymbol{u}) \leq \beta_{T+1} \left(\frac{\left(\sum_{i=1}^{d} |u_i| b_{T,i} \right)^2}{2\eta} + \eta \right).$$

Finally, note that

$$\beta_{T+1} = \sqrt{eL^2 p_T + \sum_{s=1}^{T}(p_s-1) \left(\sum_{i=1}^{d} \left(\frac{|\ell'_{s,i}|}{b_{s,i}} \right)^{p_s} \right)^{2/p_s}} \leq L\sqrt{e(T+1)(2\ln m_T - 1)}.$$

The proof of the second bound is similar. First note that $f_t^*(\boldsymbol{\theta}_t) - f_{t-1}^*(\boldsymbol{\theta}_t) \leq f_{t-1}(\boldsymbol{w}_t) - f_t(\boldsymbol{w}_t) \leq 0$, where $f_{t-1} \leq f_t$ is easily verified by inspection of (13). Using the convexity of ℓ_t and Lemma 1 we then obtain

$$R_T(\boldsymbol{u}) \leq \sum_{t=1}^{T} (\boldsymbol{\ell}_t')^\top (\boldsymbol{u} - \boldsymbol{w}_t)$$

$$\leq \frac{\sqrt{d}}{2\eta} \left(\sum_{i=1}^{d} (u_i b_{T,i})^2 \sqrt{L^2 + \sum_{s=1}^{T-1} \left(\frac{\ell_{s,i}'}{b_{s,i}} \right)^2} \right)$$

$$+ \frac{\eta}{2\sqrt{d}} \sum_{t=1}^{T} \left(\sum_{i=1}^{d} \left(\frac{\ell_{t,i}'}{b_{t,i}} \right)^2 \frac{1}{\sqrt{L^2 + \sum_{s=1}^{t-1} \left(\frac{\ell_{s,i}'}{b_{s,i}} \right)^2}} \right).$$

For the second term, we have

$$\frac{\eta}{2\sqrt{d}} \sum_{t=1}^{T} \left(\sum_{i=1}^{d} \left(\frac{\ell_{t,i}'}{b_{t,i}} \right)^2 \frac{1}{\sqrt{L^2 + \sum_{s=1}^{t-1} \left(\frac{\ell_{s,i}'}{b_{s,i}} \right)^2}} \right) \leq \frac{\eta}{2\sqrt{d}} \sum_{t=1}^{T} \sum_{i=1}^{d} \frac{(\ell_{t,i}'/b_{t,i})^2}{\sqrt{\sum_{s=1}^{t} \left(\frac{\ell_{s,i}'}{b_{s,i}} \right)^2}}$$

$$\leq \frac{\eta}{\sqrt{d}} \sum_{i=1}^{d} \sqrt{\sum_{t=1}^{T} \left(\frac{\ell_{t,i}'}{b_{t,i}} \right)^2}$$

where the last inequality uses (17). The proof is finished by noting that

$$\frac{\sqrt{d}}{2\eta} \left(\sum_{i=1}^{d} (u_i b_{T,i})^2 \sqrt{L^2 + \sum_{t=1}^{T-1} \left(\frac{\ell_{t,i}'}{b_{t,i}} \right)^2} \right) + \frac{\eta}{\sqrt{d}} \sum_{i=1}^{d} \sqrt{\sum_{t=1}^{T} \left(\frac{\ell_{t,i}'}{b_{t,i}} \right)^2}$$

$$\leq \sum_{i=1}^{d} \left(\sqrt{L^2 + \sum_{t=1}^{T} \left(\frac{\ell_{t,i}'}{b_{t,i}} \right)^2} \left(\frac{\sqrt{d}}{2\eta} (u_i b_{T,i})^2 + \frac{\eta}{\sqrt{d}} \right) \right)$$

$$\leq L\sqrt{d(T+1)} \left(\frac{1}{2\eta} \sum_{i=1}^{d} (u_i b_{T,i})^2 + \eta \right). \tag{18}$$

$\square$

Note that both bounds are invariant with respect to arbitrary scaling of individual coordinates of the data points $\boldsymbol{x}_t$ in the following sense: if the ith feature is rescaled $x_{t,i} \to c\, x_{t,i}$ for all t, then a corresponding rescaling $u_i \to u_i/c$, leaves the bounds unchanged.

This invariance property is not exhibited by standard OMD run with non-adaptive regularizers, whose regret bounds are of the form $\|\boldsymbol{u}\| \max_t \|\boldsymbol{x}_t\|_* \sqrt{T}$. In particular, by an appropriate tuning of η the regret in Corollary 1 for the regularizer type (12) is bounded by a quantity of the order of

$$\left(\sum_{i=1}^{d} |u_i| \max_t |x_{t,i}| \right) \sqrt{T \ln d}.$$

When the good $\boldsymbol{u}$ are sparse, implying that the norms $\|\boldsymbol{u}\|_1$ are small, this is always better than running standard OMD with a non-weighted q-norm regularizer. For $q \to 1$ (the best choice for the sparse $\boldsymbol{u}$ case), this gives bounds of the form

$$\left(\|\boldsymbol{u}\|_1 \max_t \|\boldsymbol{x}_t\|_\infty \right) \sqrt{T \ln d}.$$

Indeed, for regularizer (12), we have

$$\left(\sum_{i=1}^d |u_i| \max_t |x_{t,i}| \right) \leq \left(\sum_{i=1}^d |u_i| \max_t \max_j |x_{t,j}| \right) = \|\boldsymbol{u}\|_1 \max_t \|\boldsymbol{x}_t\|_\infty.$$

Similar regularization functions are studied by Grave et al. (2011) although in a different context.

Recently, a framework for studying scale-invariance in online algorithms has been proposed by Ross et al. (2013). In the variant of their setting closest to our model, the sequence of instances $\boldsymbol{x}_t \in \mathbb{R}^d$ is such that there exists an unknown diagonal matrix S for which $\left\| S^{1/2} \boldsymbol{x}_t \right\|_\infty \leq 1$ for all t. The algorithm they propose is a form of projected gradient descent with a diagonal update (see Sect. 6.3 for an explanation of diagonal updates), where adaptivity is achieved by means of a variable learning rate rather than a variable regularizer. Their algorithm achieves a regret bound of the form

$$R_T(\boldsymbol{u}) \leq \frac{C}{2\sqrt{2}} \sum_{i=1}^d \frac{1 + 6\Delta_i + \Delta_i^2}{b_{T,i}} \sqrt{\sum_{t=1}^T (\ell'_{t,i})^2} \tag{19}$$

for any $\boldsymbol{u} \in \mathbb{R}^d$ such that $\max_{t=1,\dots,T} |\boldsymbol{u}^\top \boldsymbol{x}_t| \leq C$, where C is a parameter used by the algorithm. The quantity Δ_i is of the form

$$\Delta_i = \frac{b_{T,i}}{|x_{t_i,i}|}$$

where t_i is the first time step where the ith feature has a nonzero value.

Clearly enough, introducing the parameter C in our setting might allow a dynamical tuning of η which we could not afford in our analysis. However, a rough comparison can be made by considering the intermediate bound (18) for the regularizer of type (13). Tuning $\eta = C\sqrt{d/2}$ leads to the regret bound

$$R_T(\boldsymbol{u}) \leq C\sqrt{2} \sum_{i=1}^d \sqrt{L^2 + \sum_{t=1}^T \left(\frac{\ell'_{t,i}}{b_{t,i}} \right)^2}$$

for any $\boldsymbol{u} \in \mathbb{R}^d$ such that $\max_{i=1,\dots,d} \max_{t=1,\dots,T} (u_i x_{t,i})^2 \leq C$. This last bound now bears some resemblance to (19), although further study is clearly necessary to bring out the connections between these scale-invariant updates.

6 Binary classification: aggressive and diagonal updates

In this section we show the first mistake bounds for passive–aggressive (Crammer et al. 2006) that improve on the standard Perceptron mistake bound, and also prove the first known bound for AROW with diagonal updates. Moreover, we recover—with some minor improvement—the known bounds for the second-order Perceptron (Cesa-Bianchi et al. 2005) and non-diagonalized AROW (Crammer et al. 2009).

We start by introducing binary classification as a special case of online convex optimization. Let $\mathbb{X}$ be any finite-dimensional inner product space. Given a fixed but unknown sequence $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots$ of examples $(\boldsymbol{x}_t, y_t) \in \mathbb{X} \times \{-1, +1\}$, let $\ell_t(\boldsymbol{w}) = \ell(\langle \boldsymbol{w}, \boldsymbol{x}_t \rangle, y_t)$ be

the hinge loss $\left[1 - y_t \langle \boldsymbol{w}, \boldsymbol{x}_t \rangle\right]_+$. It is easy to verify that the hinge loss satisfies the following condition:

$$\text{if } \ell_t(\boldsymbol{w}) > 0 \text{ then } \ell_t(\boldsymbol{u}) \geq 1 + \langle \boldsymbol{u}, \boldsymbol{\ell}_t' \rangle \text{ for all } \boldsymbol{u}, \boldsymbol{w} \in \mathbb{R}^d \text{ with } \boldsymbol{\ell}_t' \in \partial \ell_t(\boldsymbol{w}). \qquad (20)$$

Note that when $\ell_t(\boldsymbol{w}) > 0$ the subgradient notation is redundant, as $\partial \ell_t(\boldsymbol{w})$ is the singleton $\{\nabla \ell_t(\boldsymbol{w})\}$. In this secton, we apply the OMD algorithm to online binary classification by setting $z_t = -\eta_t \boldsymbol{\ell}_t'$ if $\ell_t(\boldsymbol{w}_t) > 0$, and $z_t = \boldsymbol{0}$ otherwise.

We prove bounds on the number of steps t in which the algorithm made a prediction mistake, defined by the condition $y_t \boldsymbol{w}_t^\top \boldsymbol{x}_t \leq 0$ or, equivalently, by $\ell_t(\boldsymbol{w}_t) \geq 1$. In the following, when the number T of prediction steps is understood from the context, we denote by $\mathcal{M}$ the subset of steps t such that $y_t \boldsymbol{w}_t^\top \boldsymbol{x}_t \leq 0$ and by M its cardinality. Similarly, we denote by $\mathcal{U}$ the set of margin error steps; that is, steps t where $y_t \boldsymbol{w}_t^\top \boldsymbol{x}_t > 0$ and $\ell_t(\boldsymbol{w}_t) > 0$. Also, we use U to denote the cardinality of $\mathcal{U}$. Following a standard terminology, we call *conservative* or *passive* an algorithm that updates its classifier only on mistake steps, and *aggressive* an algorithm that updates its classifier both on mistake and margin-error steps.

6.1 First-order algorithms

We start by showing how our framework allows us to generalize and improve previous analyses for binary classification algorithms that use first-order aggressive updates. Let

$$L(\boldsymbol{u}) = \sum_{t=1}^T \left[1 - y_t \langle \boldsymbol{u}, \boldsymbol{x}_t \rangle\right]_+$$

be the cumulative hinge loss of $\boldsymbol{u} \in \mathbb{X}$ with respect to some sequence of examples. The next result provides a general mistake bound for first-order algorithms.

Corollary 2 *Assume OMD is run with $f_t = f$, where f has domain $\mathbb{X}$, is β-strongly convex with respect to the norm $\|\cdot\|$, and satisfies $f(\lambda \boldsymbol{u}) \leq \lambda^2 f(\boldsymbol{u})$ for all $\lambda \in \mathbb{R}$ and all $\boldsymbol{u} \in \mathbb{X}$. Further assume the input sequence is $z_t = \eta_t y_t \boldsymbol{x}_t$ for some $0 \leq \eta_t \leq 1$ such that $\eta_t = 1$ whenever $y_t \langle \boldsymbol{w}_t, \boldsymbol{x}_t \rangle \leq 0$. Then, for all $T \geq 1$,*

$$M \leq \operatorname*{argmin}_{\boldsymbol{u} \in \mathbb{X}} L(\boldsymbol{u}) + D + \frac{2}{\beta} f(\boldsymbol{u}) X_T^2 + X_T \sqrt{\frac{2}{\beta} f(\boldsymbol{u}) L(\boldsymbol{u})},$$

where $M = |\mathcal{M}|$, $X_t = \max_{i=1,\dots,t} \|\boldsymbol{x}_i\|_*$ and

$$D = \sum_{t \in \mathcal{U}} \eta_t \left(\frac{\eta_t \|\boldsymbol{x}_t\|_*^2 + 2\beta \, y_t \langle \boldsymbol{w}_t, \boldsymbol{x}_t \rangle}{X_t^2} - 2 \right).$$

Proof Fix any $\boldsymbol{u} \in \mathbb{X}$. Using the second bound of Lemma 3 in the "Appendix", with the assumption $\eta_t = 1$ when $t \in \mathcal{M}$, we get

$$M \leq L(\boldsymbol{u}) + \sqrt{2f(\boldsymbol{u})} \sqrt{\sum_{t \in \mathcal{M}} \frac{\|\boldsymbol{x}_t\|_*^2}{\beta} + \sum_{t \in \mathcal{U}} \left(\frac{\eta_t^2}{\beta} \|\boldsymbol{x}_t\|_*^2 + 2\eta_t y_t \langle \boldsymbol{w}_t, \boldsymbol{x}_t \rangle \right)} - \sum_{t \in \mathcal{U}} \eta_t$$

$$\leq L(\boldsymbol{u}) + X_T \sqrt{\frac{2}{\beta} f(\boldsymbol{u})} \sqrt{M + \sum_{t \in \mathcal{U}} \frac{\eta_t^2 \|\boldsymbol{x}_t\|_*^2 + 2\beta \eta_t y_t \langle \boldsymbol{w}_t, \boldsymbol{x}_t \rangle}{X_t^2}} - \sum_{t \in \mathcal{U}} \eta_t$$

where we have used the fact that $X_t \leq X_T$ for all $t = 1, \ldots, T$. Solving for M we get

$$M \leq L(\boldsymbol{u}) + \frac{1}{\beta} f(\boldsymbol{u}) X_T^2 + X_T \sqrt{\frac{2}{\beta} f(\boldsymbol{u})} \sqrt{\frac{1}{2\beta} X_T^2 f(\boldsymbol{u}) + L(\boldsymbol{u}) + D'} - \sum_{t \in \mathcal{U}} \eta_t \qquad (21)$$

with $\frac{1}{2\beta} X_T^2 f(\boldsymbol{u}) + L(\boldsymbol{u}) + D' \geq 0$, and

$$D' = \sum_{t \in \mathcal{U}} \left(\frac{\eta_t^2 \|\boldsymbol{x}_t\|_*^2 + 2\beta \eta_t y_t \langle \boldsymbol{w}_t, \boldsymbol{x}_t \rangle}{X_t^2} - \eta_t \right).$$

We further upper bound the right-hand side of (21) using the elementary inequality $\sqrt{a + b} \leq \sqrt{a} + \frac{b}{2\sqrt{a}}$ for all $a > 0$ and $b \geq -a$. This gives

$$M \leq L(\boldsymbol{u}) + \frac{1}{\beta} f(\boldsymbol{u}) X_T^2 + X_T \sqrt{\frac{2}{\beta} f(\boldsymbol{u})} \sqrt{\frac{1}{2\beta} X_T^2 f(\boldsymbol{u}) + L(\boldsymbol{u})}$$

$$+ \frac{X_T D' \sqrt{\frac{2}{\beta} f(\boldsymbol{u})}}{2\sqrt{\frac{1}{2\beta} X_T^2 f(\boldsymbol{u}) + L(\boldsymbol{u})}} - \sum_{t \in \mathcal{U}} \eta_t$$

$$\leq L(\boldsymbol{u}) + \frac{1}{\beta} f(\boldsymbol{u}) X_T^2 + X_T \sqrt{\frac{2}{\beta} f(\boldsymbol{u})} \sqrt{\frac{1}{2\beta} X_T^2 f(\boldsymbol{u}) + L(\boldsymbol{u})} + D' - \sum_{t \in \mathcal{U}} \eta_t.$$

Applying the inequality $\sqrt{a + b} \leq \sqrt{a} + \sqrt{b}$ and rearranging gives the desired bound. □

The p-norm Perceptron of Gentile (2003) is obtained by running OMD in conservative mode with $f_t = f = \frac{1}{2} \|\cdot\|_p^2$ for $1 < p \leq 2$. In this case we have $\mathcal{U} = \emptyset$, $\|\cdot\|_* = \|\cdot\|_q$ where $q = \frac{p}{p-1}$, and $\beta = p - 1$ because $\frac{1}{2} \|\cdot\|_p^2$ is $(p-1)$-strongly convex with respect to $\|\cdot\|_p$ for $1 < p \leq 2$, see Shalev-Shwartz (2007, Lemma 17). Hence Corollary 2 delivers the mistake bound of Gentile (2003).

However, the term D in the bound of Corollary 2 can also be negative. We can minimize it, subject to $0 \leq \eta_t \leq 1$, by setting

$$\eta_t = \max \left\{ \min \left\{ \frac{X_t^2 - \beta y_t \langle \boldsymbol{w}_t, \boldsymbol{x}_t \rangle}{\|\boldsymbol{x}_t\|_*^2}, 1 \right\}, 0 \right\}.$$

This tuning of η_t is quite similar to that of the passive–aggressive algorithm (type I) of Crammer et al. (2006). In fact for $f_t = f = \frac{1}{2} \|\cdot\|_2^2$ we would have

$$\eta_t = \max \left\{ \min \left\{ \frac{X_t^2 - y_t \langle \boldsymbol{w}_t, \boldsymbol{x}_t \rangle}{\|\boldsymbol{x}_t\|^2}, 1 \right\}, 0 \right\}$$

while the update rule for PA-I is

$$\eta_t = \max \left\{ \min \left\{ \frac{1 - y_t \langle \boldsymbol{w}_t, \boldsymbol{x}_t \rangle}{\|\boldsymbol{x}_t\|^2}, 1 \right\}, 0 \right\}.$$

The mistake bound of Corollary 2 is however better than the aggressive bounds for PA-I of Crammer et al. (2006) and Shalev-Shwartz (2007). Indeed, while the PA-I bounds are generally worse than the Perceptron mistake bound

$$M \leq L(\boldsymbol{u}) + \left(\|\boldsymbol{u}\| X_T \right)^2 + \|\boldsymbol{u}\| X_T \sqrt{L(\boldsymbol{u})}$$

as discussed by Crammer et al. (2006), our bound is better as soon as $D < 0$. Hence, it can be viewed as the first theoretical evidence in support of aggressive updates. It also improves over previous attempts to justify aggressive updates in (Orabona et al. 2009; Jie et al. 2010).

6.2 Second-order algorithms

We now move on to the analysis of second-order algorithms for binary classification. Here we use $\mathbb{X} = \mathbb{R}^d$ and let the inner product $\langle \boldsymbol{u}, \boldsymbol{x} \rangle$ be the standard dot product $\boldsymbol{u}^\top \boldsymbol{x}$.

Second-order algorithms for binary classification are online variants of Ridge regression. Recall that the Ridge regression linear predictor is defined by

$$\boldsymbol{w}_t = \underset{\boldsymbol{w} \in \mathbb{R}^d}{\operatorname{argmin}} \left(\sum_{s=1}^{t} (\boldsymbol{w}^\top \boldsymbol{x}_s - y_s)^2 + \|\boldsymbol{w}\|^2 \right).$$

The closed-form expression for $\boldsymbol{w}_t$, involving the design matrix $S_t = \begin{bmatrix} \boldsymbol{x}_1, \ldots, \boldsymbol{x}_t \end{bmatrix}$ and the label vector $\boldsymbol{y}_t = (y_1, \ldots, y_t)$, is given by $\boldsymbol{w}_t = \left(I + S_t^\top S_t \right)^{-1} S_t \boldsymbol{y}_t$. The second-order Perceptron (see below) uses this weight $\boldsymbol{w}_t$, but S_t and $\boldsymbol{y}_t$ only contain the examples $(\boldsymbol{x}_s, y_s)$ such that $y_s \boldsymbol{w}_s^\top \boldsymbol{x}_s \leq 0$. Namely, those previous examples on which a mistake occurred. In this sense, it is an online variant of the Ridge regression algorithm.

In practice, second-order algorithms may perform better than their first-order counterparts, such as the algorithms in the Perceptron family. There are two basic second-order algorithms: the second-order Perceptron of Cesa-Bianchi et al. (2005) and the AROW algorithm of Crammer et al. (2009). We show that both of them are instances of OMD and recover their mistake bounds as special cases of our analysis.

Let $z_t = \eta_t y_t \boldsymbol{x}_t$ and $f_t(\boldsymbol{x}) = \frac{1}{2} \boldsymbol{x}^\top A_t \boldsymbol{x}$, where $A_0 = I$ and $A_t = A_{t-1} + \frac{1}{r} \boldsymbol{x}_t \boldsymbol{x}_t^\top$ with $r > 0$. Each dual function f_t^* is given by $f_t^*(\boldsymbol{x}) = \frac{1}{2} \boldsymbol{x}^\top A_t^{-1} \boldsymbol{x}$. The functions f_t are 1-strongly convex with respect to the norm $\|\boldsymbol{x}\|_t = \sqrt{\boldsymbol{x}^\top A_t \boldsymbol{x}}$ with dual norm $\|\boldsymbol{x}\|_{t,*} = \sqrt{\boldsymbol{x}^\top A_t^{-1} \boldsymbol{x}}$.

The conservative version of OMD run with f_t chosen as above and $r = 1$ corresponds to the second-order Perceptron. The aggressive version corresponds instead to AROW with a minor difference. Let $m_t = \boldsymbol{\theta}_t^\top A_{t-1}^{-1} \boldsymbol{x}_t$. Then for AROW we have $\boldsymbol{w}_t^\top \boldsymbol{x}_t = m_t$ whereas for OMD it holds that $\boldsymbol{w}_t^\top \boldsymbol{x}_t = \boldsymbol{\theta}_t^\top A_t^{-1} \boldsymbol{x}_t = m_t \frac{r}{r+\chi_t}$, where we used the Woodbury identity and set $\chi_t = \boldsymbol{x}_t^\top A_{t-1}^{-1} \boldsymbol{x}_t$. Note that the sign of $\boldsymbol{w}_t^\top \boldsymbol{x}_t$ is the same for both algorithms, but OMD updates when $y_t m_t \frac{r}{r+\chi_t} \leq 1$ while AROW updates when $y_t m_t \leq 1$. Typically, for t large the value of χ_t is small and the two algorithms behave similarly.

In order to derive a mistake bound for OMD run with $f_t(\boldsymbol{x}) = \frac{1}{2} \boldsymbol{x}^\top A_t \boldsymbol{x}$, first observe that using the Woodbury identity we have

$$f_t^*(\theta_t) - f_{t-1}^*(\theta_t) = -\frac{\left(\boldsymbol{x}_t^\top A_{t-1}^{-1} \boldsymbol{\theta}_t \right)^2}{2 \left(r + \boldsymbol{x}_t^\top A_{t-1}^{-1} \boldsymbol{x}_t \right)} = -\frac{m_t^2}{2(r + \chi_t)}.$$

Hence, using the second bound of Lemma 3 in the "Appendix", and setting $\eta_t = 1$, we obtain

$$M + U - L(\boldsymbol{u}) \leq \sqrt{\boldsymbol{u}^\top A_T \boldsymbol{u}} \sqrt{\sum_{t \in \mathcal{M} \cup \mathcal{U}} \left(\boldsymbol{x}_t^\top A_t^{-1} \boldsymbol{x}_t + 2 y_t \boldsymbol{w}_t^\top \boldsymbol{x}_t - \frac{m_t^2}{r + \chi_t} \right)}$$

$$\leq \sqrt{\|\boldsymbol{u}\|^2 + \frac{1}{r} \sum_{t \in \mathcal{M} \cup \mathcal{U}} (\boldsymbol{u}^\top \boldsymbol{x}_t)^2} \sqrt{r \ln |A_T| + \sum_{t \in \mathcal{M} \cup \mathcal{U}} \left(2 y_t \boldsymbol{w}_t^\top \boldsymbol{x}_t - \frac{m_t^2}{r + \chi_t} \right)}$$

$$= \sqrt{r\|\boldsymbol{u}\|^2 + \sum_{t\in\mathcal{M}\cup\mathcal{U}} (\boldsymbol{u}^\top \boldsymbol{x}_t)^2} \sqrt{\ln|A_T| + \sum_{t\in\mathcal{M}\cup\mathcal{U}} \frac{m_t(2ry_t - m_t)}{r(r+\chi_t)}}$$

for all $\boldsymbol{u} \in \mathbb{R}^d$.

This bound improves slightly over the known bound for AROW in the last sum in the square root. In fact in AROW we have the term U, while here we have

$$\sum_{t\in\mathcal{M}\cup\mathcal{U}} \frac{m_t(2ry_t - m_t)}{r(r+\chi_t)} \le \sum_{t\in\mathcal{U}} \frac{m_t(2ry_t - m_t)}{r(r+\chi_t)} \le \sum_{t\in\mathcal{U}} \frac{r^2}{r(r+\chi_t)} \le U$$

where the first inequality holds because $t \in \mathcal{M}$ implies $y_t m_t \le 0$, which in turn implies $m_t(2ry_t - m_t) \le 0$. In the conservative case, when $\mathcal{U} \equiv \emptyset$, the bound specializes to the standard second-order Perceptron bound.

6.3 Diagonal updates

Computing f_t^* in AROW and the second-order Perceptron requires inverting A_t, which can be done from A_{t-1}^{-1} in time quadratic in d. A much better scaling, linear in d, can be obtained when the algorithm use a diagonalized version of A_t. We now use Corollary 3 to prove a mistake bound for the diagonal version of the second-order Perceptron. Denote $D_t = \text{diag}\{A_t\}$ be the diagonal matrix that agrees with A_t on the diagonal, where A_t is defined as before, and let $f_t(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top D_t \boldsymbol{x}$. Setting $\eta_t = 1$, using the second bound of Lemma 3, and using also Lemma 4, we have[3]

$$M + U \le \operatorname*{argmin}_{\boldsymbol{u}\in\mathbb{R}^d} L(\boldsymbol{u}) + \sqrt{\boldsymbol{u}^T D_T \boldsymbol{u} \left(r\sum_{i=1}^d \ln\left(\frac{1}{r}\sum_{t\in\mathcal{M}\cup\mathcal{U}} x_{t,i}^2 + 1\right) + 2U\right)}$$

$$= \operatorname*{argmin}_{\boldsymbol{u}\in\mathbb{R}^d} L(\boldsymbol{u})$$

$$+ \sqrt{\|\boldsymbol{u}\|^2 + \frac{1}{r}\sum_{i=1}^d u_i^2 \left(\sum_{t\in\mathcal{M}\cup\mathcal{U}} x_{t,i}^2\right)} \sqrt{r\sum_{i=1}^d \ln\left(\frac{1}{r}\sum_{t\in\mathcal{M}\cup\mathcal{U}} x_{t,i}^2 + 1\right) + 2U}. \tag{22}$$

This allows us to theoretically analyze the cases where this algorithm could be advantageous. For example, features of textual data are typically binary, and it is often the case that most of the features are zero most of the time. On the other hand, these "rare" features are usually the most informative ones—see, e.g., the discussion of Dredze et al. (2008), Crammer et al. (2012).

Figure 1 shows the number of times each feature (word) appears in two sentiment datasets versus the word rank. Clearly, there are a few very frequent words and many rare words. These exact properties originally motivated the CW and AROW algorithms, and now our analysis provides a theoretical justification. Concretely, the above considerations support the assumption that the optimal hyperplane $\boldsymbol{u}$ satisfies

$$\sum_{i=1}^d u_i^2 \sum_{t\in\mathcal{M}\cup\mathcal{U}} x_{t,i}^2 \approx \sum_{i\in\mathcal{I}} u_i^2 \sum_{t\in\mathcal{M}\cup\mathcal{U}} x_{t,i}^2 \le s \sum_{i\in\mathcal{I}} u_i^2 \approx s\|\boldsymbol{u}\|^2$$

---

[3] We did not optimize the constant multiplying U in the bound.
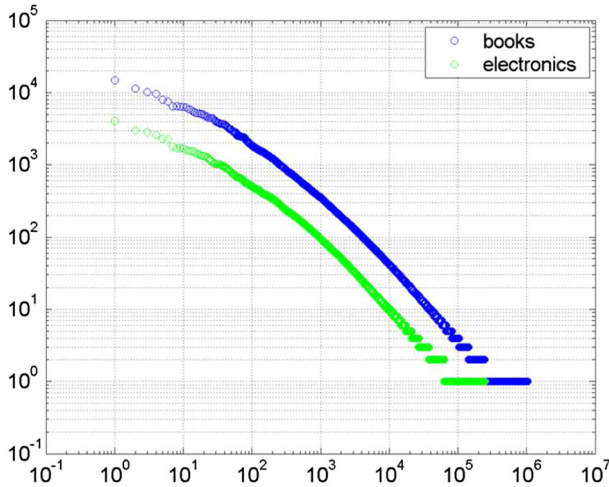
Fig. 1 Evidence of heavy tails for textual data. The plots show the number of words versus the word rank on two sentiment data sets

where $\mathcal{I}$ is the set of informative and rare features, and s is the maximum number of times these features appear in the sequence. Running the diagonal version of the second order Pereptron so that $\mathcal{U} = \emptyset$, and assuming that

$$\sum_{i=1}^{d} u_i^2 \sum_{t \in \mathcal{M} \cup \mathcal{U}} x_{t,i}^2 \leq s \|\boldsymbol{u}\|^2 \tag{23}$$

the last term in the mistake bound (22) can be re-written as

$$\sqrt{\|\boldsymbol{u}\|^2 + \frac{1}{r} \sum_{i=1}^{d} u_i^2 \sum_{t \in \mathcal{M}} x_{t,i}^2} \sqrt{r \sum_{i=1}^{d} \ln\left(\frac{1}{r} \sum_{t \in \mathcal{M}} x_{t,i}^2 + 1\right)}$$

$$\leq \|\boldsymbol{u}\| \sqrt{r + s} \sqrt{d \ln\left(\frac{M X_T^2}{dr} + 1\right)}$$

where we calculated the maximum of the sum, given the constraint

$$\sum_{i=1}^{d} \sum_{t \in \mathcal{M}} x_{t,i}^2 \leq X_T^2 M.$$

We can now use Corollary 4 in the "Appendix" to obtain

$$M \leq \operatorname*{argmin}_{\boldsymbol{u} \in \mathbb{R}^d} L(\boldsymbol{u}) + \|\boldsymbol{u}\| \sqrt{(r + s) d \ln\left(\frac{\sqrt{8} \|\boldsymbol{u}\|^2 (r + s) X_T^4}{edr^2} + 2L(\boldsymbol{u}) \frac{X_T^2}{dr} + 2\right)}.$$

Hence, when the hypothesis (23) is verified, the number of mistakes of the diagonal version of AROW depends on $\sqrt{\ln L(\boldsymbol{u})}$ rather than on $\sqrt{L(\boldsymbol{u})}$.

Diagonal updates for online convex optimization were also proposed and analyzed by (McMahan and Streeter 2010; Duchi et al. 2011; Ross et al. 2013). When instantiated to

the binary classification setting studied in this section, their analysis delivers regret bounds which are not comparable to ours.

7 Conclusions

We proposed a framework for online convex optimization combining online mirror descent with time-varying regularizers. This allowed us to view second-order algorithms (such as the Vovk–Azoury–Warmuth algorithm, the second-order Perceptron, and the AROW algorithm) and algorithms for composite losses as special cases of mirror descent. Our analysis also captures second-order variants that only employ the diagonal elements of the second order information matrix, a result which was not within reach of the previous techniques.

Within our framework, we also derived and analyzed new regularizers based on an adaptive weighted version of the p-norm Perceptron. These regularizers generate instances of OMD that are both efficient to implement and invariant to rescaling of individual coordinates in the data. In the case of sparse targets, the corresponding instances of OMD achieve performance bounds better than that of OMD with 1-norm regularization.

We also improved previous bounds for existing first-order algorithms. For example, we were able to formally explain the phenomenon according to which aggressive algorithms often exhibit better empirical performance than their conservative counterparts. Specifically, our refined analysis provides a bound for passive–aggressive (PA-I) that is never worse (and sometimes better) than the Perceptron bound.

One interesting direction to pursue is the derivation and analysis of algorithms based on time-varying versions of the entropic regularizers used by the EG and Winnow algorithms. A remarkable recent result along these lines is the work of Steinhardt and Liang (2014), in which a time-varying entropic regularizers is used to obtain an improved version of EG for the prediction with experts setting (a special case of online convex optimization).

More in general, it would be useful to devise a more systematic approach to the design of adaptive regularizers enjoying a given set of desired properties, such as invariance to rescalings. This should help in obtaining more examples of adaptation mechanisms that are not based on second-order information.

Appendix: Technical lemmas

Proof of Lemma 2   Define $h(\boldsymbol{w}) = g(A\boldsymbol{w})$, where A is an invertible matrix. First note that the Fenchel conjugate of $h(\boldsymbol{w})$ is $h^*(\boldsymbol{\theta}) = g^*(A^{-\top}\boldsymbol{\theta})$—see for example (Bauschke and Combettes 2011, Proposition 13.20 IV). Hence, the Fenchel conjugate of f is obtained by setting: $A = \operatorname{diag}\big(\{a_1^{1/q}, \ldots, a_d^{1/q}\}\big)$, $g(\boldsymbol{w}) = \frac{1}{2(p-1)}\|\boldsymbol{\theta}\|_p^2$, and by using the known Fenchel conjugate of g.

In order to show the second part, using (Shalev-Shwartz 2007, Lemma 14) we may prove strong convexity of h w.r.t. a norm $\|\cdot\|$ by showing that

$$\boldsymbol{x}^\top \nabla^2 h(\boldsymbol{w})\boldsymbol{x} = \boldsymbol{x}^\top A^\top \nabla^2 g(A\boldsymbol{w})A\boldsymbol{x} \geq \|\boldsymbol{x}\|^2 . \tag{24}$$

Moreover, (Shalev-Shwartz 2007, Lemma 17) proves that, for any $\boldsymbol{x}, \boldsymbol{w} \in \mathbb{R}^d$, we have

$$\boldsymbol{x} \nabla^2 g(\boldsymbol{w}) \boldsymbol{x} \geq \|\boldsymbol{x}\|_p^2. \tag{25}$$

Putting together (24) and (25) and the same setting of A and g used above, we have that the strong convexity of f.

We now prove that the dual norm of $\left(\sum_{i=1}^d |x_i|^q b_i\right)^{1/q}$ is $\left(\sum_{i=1}^d |\theta_i|^p b_i^{1-p}\right)^{1/p}$. By definition of dual norm,

$$\sup_{\boldsymbol{x}} \left\{ \boldsymbol{u}^\top \boldsymbol{x} : \left(\sum_{i=1}^d |x_i|^q b_i\right)^{1/q} \leq 1 \right\} = \sup_{\boldsymbol{x}} \left\{ \boldsymbol{u}^\top \boldsymbol{x} : \left(\sum_{i=1}^d \left| x_i b_i^{1/q} \right|^q \right)^{1/q} \leq 1 \right\}$$

$$= \sup_{\boldsymbol{y}} \left\{ \sum_i u_i y_i b_i^{-1/q} : \left(\sum_{i=1}^d |y_i|^q\right)^{1/q} \leq 1 \right\}$$

$$= \left\| \left(u_1 b_1^{-1/q}, \ldots, u_d b_d^{-1/q} \right) \right\|_p$$

where $1/q + 1/p = 1$. Writing the last norm explicitly and observing that $p = q/(q-1)$,

$$\left(\sum_i |u_i|^p b_i^{-p/q}\right)^{1/p} = \left(\sum_i |u_i|^p b_i^{1-q}\right)^{1/p}$$

which concludes the proof. □

Lemma 3 *Assume OMD is run with functions $f_1, f_2, \ldots, f_T$ defined on $\mathbb{X}$ and such that each f_t is β_t-strongly convex with respect to the norm $\|\cdot\|_{f_t}$ and $f_t(\lambda \boldsymbol{u}) \leq \lambda^2 f_t(\boldsymbol{u})$ for all $\lambda \in \mathbb{R}$ and all $\boldsymbol{u} \in S$. For each $t = 1, 2, \ldots, T$ let $\|\cdot\|_{t,*}$ be the dual norm of $\|\cdot\|_t$. Assume further the input sequence is $z_t = -\eta_t \ell_t'$ for some $\eta_t > 0$, where $\ell_t' \in \partial \ell_t(\boldsymbol{w}_t), \ell_t(\boldsymbol{w}_t) = 0$ implies $\ell_t' = \boldsymbol{0}$, and $\ell_t = \ell(\langle \cdot, \boldsymbol{x}_t \rangle, y_t)$ satisfies (20). Then, for all $T \geq 1$,*

$$\sum_{t \in \mathcal{M} \cup \mathcal{U}} \eta_t \leq L_\eta + \lambda f_T(\boldsymbol{u}) + \frac{1}{\lambda} \left(B + \sum_{t \in \mathcal{M} \cup \mathcal{U}} \left(\frac{\eta_t^2}{2\beta_t} \|\ell_t'\|_{t,*}^2 - \eta_t \langle \boldsymbol{w}_t, \ell_t' \rangle \right) \right) \tag{26}$$

for any $\boldsymbol{u} \in S$ and any $\lambda > 0$, where

$$L_\eta = \sum_{t \in \mathcal{M} \cup \mathcal{U}} \eta_t \, \ell_t(\boldsymbol{u}) \quad and \quad B = \sum_{t=1}^T \left(f_t^*(\boldsymbol{\theta}_t) - f_{t-1}^*(\boldsymbol{\theta}_t) \right).$$

In particular, choosing the optimal λ, we obtain

$$\sum_{t \in \mathcal{M} \cup \mathcal{U}} \eta_t \leq L_\eta + 2\sqrt{f_T(\boldsymbol{u})} \sqrt{\left[B + \sum_{t \in \mathcal{M} \cup \mathcal{U}} \left(\frac{\eta_t^2}{2\beta_t} \|\ell_t'\|_{t,*}^2 - \eta_t \langle \boldsymbol{w}_t, \ell_t' \rangle \right) \right]_+}.$$

Proof We apply Lemma 1 with $z_t = -\eta_t \ell_t'$ and using $\boldsymbol{u} = \lambda \tilde{\boldsymbol{u}}$ for any $\lambda > 0$,

$$\sum_{t=1}^T \eta_t \langle \ell_t', \boldsymbol{w}_t - \lambda \tilde{\boldsymbol{u}} \rangle \leq \lambda^2 f_T(\tilde{\boldsymbol{u}}) + \sum_{t=1}^T \left(\frac{\eta_t^2}{2\beta_t} \|\ell_t'\|_{t,*}^2 + f_t^*(\boldsymbol{\theta}_t) - f_{t-1}^*(\boldsymbol{\theta}_t) \right).$$

Since $\ell_t(\boldsymbol{w}_t) = 0$ implies $\boldsymbol{\ell}'_t = \boldsymbol{0}$, and using (20),

$$\sum_{t \in \mathcal{M} \cup \mathcal{U}} \left(\eta_t \langle \boldsymbol{\ell}'_t, \boldsymbol{w}_t \rangle + \eta_t - \eta_t \, \ell_t(\boldsymbol{u}) \right) \leq \sum_{t=1}^{T} \eta_t \langle \boldsymbol{\ell}'_t, \boldsymbol{w}_t - \lambda \tilde{\boldsymbol{u}} \rangle.$$

Dividing by λ and rearranging gives the first bound. The second bound is obtained by choosing the λ that makes equal the last two terms in the right-hand side of (26). □

Lemma 4 *For all $\boldsymbol{x}_1, \ldots \boldsymbol{x}_T \in \mathbb{R}^d$ let $D_t = \mathrm{diag}\{A_t\}$ where $A_0 = I$ and $A_t = A_{t-1} + \frac{1}{r}\boldsymbol{x}_t \boldsymbol{x}_t^\top$ for some $r > 0$. Then*

$$\sum_{t=1}^{T} \boldsymbol{x}_t D_t^{-1} \boldsymbol{x}_t \leq r \sum_{i=1}^{d} \ln \left(\frac{1}{r} \sum_{t=1}^{T} x_{t,i}^2 + 1 \right).$$

Proof Consider the sequence $a_t \geq 0$ and define $v_t = a_0 + \sum_{i=1}^{t} a_i$ with $a_0 > 0$. The concavity of the logarithm implies $\ln b \leq \ln a + \frac{b-a}{a}$ for all $a, b > 0$. Hence we have

$$\sum_{t=1}^{T} \frac{a_t}{v_t} = \sum_{t=1}^{T} \frac{v_t - v_{t-1}}{v_t} \leq \sum_{t=1}^{T} \ln \frac{v_t}{v_{t-1}} = \ln \frac{v_T}{v_0} = \ln \frac{a_0 + \sum_{t=1}^{T} a_t}{a_0}.$$

Using the above and the definition of D_t, we obtain

$$\sum_{t=1}^{T} \boldsymbol{x}_t D_t^{-1} \boldsymbol{x}_t = \sum_{i=1}^{d} \sum_{t=1}^{T} \frac{x_{t,i}^2}{1 + \frac{1}{r}\sum_{j=1}^{t} x_{j,i}^2} = r \sum_{i=1}^{d} \sum_{t=1}^{T} \frac{x_{t,i}^2}{r + \sum_{j=1}^{t} x_{j,i}^2}$$

$$\leq r \sum_{i=1}^{d} \ln \frac{r + \sum_{t=1}^{T} x_{t,i}^2}{r}.$$

□

We conclude the appendix by proving the results required to solve the implicit logarithmic equations of Sect. 6.3. We use the following result—see (Orabona et al. 2012, Lemma 2).

Lemma 5 *Let $a, x > 0$ be such that $x \leq a \ln x$. Then for all $n > 1$*

$$x \leq \frac{n}{n-1} a \ln \frac{na}{e}.$$

This allows to prove the following easy corollaries.

Corollary 3 *For all $a, b, c, d, x > 0$ such that $x \leq a \ln(bx + c) + d$, we have*

$$x \leq \frac{n}{n-1} \left(a \ln \frac{nab}{e} + d \right) + \frac{c}{b} \frac{1}{n-1}.$$

Corollary 4 *For all $a, b, c, d, x > 0$ such that*

$$x \leq \sqrt{a \ln(bx + 1) + c} + d \tag{27}$$

we have

$$x \leq \sqrt{a \ln \left(\frac{\sqrt{8}ab^2}{e} + 2b\sqrt{c} + 2db + 2 \right) + c} + d.$$

Proof Assumption (27) implies

$$x^2 \leq \left(\sqrt{a\ln(bx+1)+c}+d\right)^2 \leq 2a\ln(bx+1)+2c+2d^2 = a\ln(bx+1)^2+2c+2d^2$$

$$\leq a\ln(2b^2x^2+2)+2c+2d^2. \tag{28}$$

From Corollary 3 we have that if $f, g, h, i, y > 0$ satisfy $y \leq f\ln(gx+h)+i$, then

$$y \leq \frac{n}{n-1}\left(f\ln\frac{nfg}{e}+i\right)+\frac{h}{g}\frac{1}{n-1} \leq \frac{n}{n-1}\left(\frac{nf^2g}{e^2}+i\right)+\frac{h}{g}\frac{1}{n-1}$$

where we have used the elementary inequality $\ln y \leq \frac{y}{e}$ for all $y \geq 0$. Applying the above to (28) we obtain

$$x^2 \leq \frac{n}{n-1}\left(\frac{2na^2b^2}{e^2}+2c+2d^2\right)+\frac{1}{b^2}\frac{1}{n-1}$$

which implies

$$x \leq \sqrt{\frac{n}{n-1}}\left(\frac{\sqrt{2}nab}{e}+\sqrt{2c}+\sqrt{2}d\right)+\frac{1}{b}\frac{1}{\sqrt{n-1}}. \tag{29}$$

Note that we have repeatedly used the elementary inequality $\sqrt{x+y} \leq \sqrt{x}+\sqrt{y}$. Choosing $n = 2$ and applying (29) to (27) we get

$$x \leq \sqrt{a\ln(bx+1)+c}+d \leq \sqrt{a\ln\left(\frac{\sqrt{8}ab^2}{e}+2b\sqrt{c}+2db+2\right)+c}+d$$

concluding the proof.   □

References

Abernethy, J., Hazan, E., & Rakhlin, A. (2008). Competing in the dark: An efficient algorithm for bandit linear optimization. In R. A. Servedio & T. Zhang (Eds.), *COLT* (pp. 263–274). Madison: Omnipress.

Abernethy, J., Hazan, E., & Rakhlin, A. (2012). Interior-point methods for full-information and bandit online learning. *IEEE Transactions on Information Theory*, *58*(7), 4164–4175.

Auer, P., Cesa-Bianchi, N., & Gentile, C. (2002). Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, *64*(1), 48–75.

Azoury, K. S., & Warmuth, M. K. (2001). Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, *43*(3), 211–246.

Bartlett, P. L., Hazan, E., & Rakhlin, A. (2007). Adaptive online gradient descent. In Platt John C., Koller Daphne, Singer Yoram, & Roweis Sam T. (Eds.), *NIPS*. USA: Curran Associates, Inc.

Bauschke, H. H., & Combettes, P. L. (2011). *Convex analysis and monotone operator theory in hilbert spaces*. Berlin: Springer.

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

Cesa-Bianchi, N., & Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge: Cambridge University Press.

Cesa-Bianchi, N., Conconi, A., & Gentile, C. (2005). A second-order Perceptron algorithm. *SIAM Journal on Computing*, *34*(3), 640–668.

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., & Singer, Y. (2006). Online passive–aggressive algorithms. *Journal of Machine Learning Research*, *7*, 551–585.

Crammer, K., Dredze, M., & Pereira, F. (2008). Exact convex confidence-weighted learning. *Advances in Neural Information Processing Systems*, *22*, 345–352.

Crammer, K., Kulesza, A., & Dredze, M. (2009). Adaptive regularization of weight vectors. *Advances in Neural Information Processing Systems*, *23*, 414–422.

Crammer, K., Dredze, M., & Pereira, F. (2012). Confidence-weighted linear classification for text categorization. *Journal of Machine Learning Reseatch*, *13*, 1891–1926. ISSN: 1532-4435.

Dredze, M., Crammer, K., & Pereira, F. (2008). Online confidence-weighted learning. In *Proceedings of the 25th international conference on machine learning*.

Duchi, J., & Singer, Y. (2009). Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, *10*, 2899–2934.

Duchi, J., Shalev-Shwartz, S., Singer, Y., &Tewari, A. (2010). Composite objective mirror descent. In *Proceedings of the 23rd annual conference on learning theory* (pp. 14–26).

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, *12*, 2121–2159.

Forster, J. (1999). On relative loss bounds in generalized linear regression. In *FCT*, ISBN: 3-540-66412-2.

Gentile, C. (2003). The robustness of the p-norm algorithms. *Machine Learning*, *53*(3), 265–299.

Gentile, C., & Warmuth, M. K. (1998). Linear hinge loss and average margin. In M. J. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *NIPS* (pp. 225–231). Cambridge: MIT Press.

Grave, E., Obozinski, G., & Bach, F.R. (2011). Trace Lasso: a trace norm regularization for correlated designs. *CoRR*, abs/1109.1990.

Hazan, E. (2011). The convex optimization approach to regret minimization. *Optimization for machine learning* (pp. 287–303).

Hazan, E., Agarwal, A., & Kale, S. (2007). Logarithmic regret algorithms for online convex optimization. *Machine Learning*, *69*(2–3), 169–192.

Jie, L., Orabona, F., Fornoni, M., Caputo, B.,& Cesa-Bianchi, N. (2010). OM-2: An online multi-class multi-kernel learning algorithm. In *4th IEEE online learning for computer vision workshop (in CVPR10)*. IEEE Computer Society.

Kakade, S. M., Shalev-Shwartz, S., & Tewari, A. (2012). Regularization techniques for learning with matrices. *The Journal of Machine Learning Research*, *13*, 1865–1890.

Kivinen, J., & Warmuth, M. K. (1997). Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, *132*(1), 1–63.

Kivinen, J., & Warmuth, M. K. (2001). Relative loss bounds for multidimensional regression problems. *Machine Learning*, *45*(3), 301–329.

Kivinen, J., Warmuth, M. K., & Hassibi, B. (2006). The p-norm generalization of the LMS algorithm for adaptive filtering. *IEEE Transactions on Signal Processing*, *54*(5), 1782–1793.

Littlestone, N. (1988). Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, *2*(4), 285–318.

McMahan, B., & Orabona, F. (2014). Unconstrained online linear learning in Hilbert spaces: Minimax algorithms and Normal approximations. In *Conference on Learning Theory (COLT)*.

McMahan, H.B., & Streeter, M. J. (2010). Adaptive bound optimization for online convex optimization. In *Proceedings of the 23rd conference on learning theory* (pp. 244–256). Omnipress.

Nemirovsky, A., & Yudin, D. (1983). *Problem complexity and method efficiency in optimization*. New York: Wiley.

Orabona, F. (2013). Dimension-free exponentiated gradient. In *Advances in neural information processing systems (NIPS)* (Vol. 26).

Orabona, F., Keshet, J., & Caputo, B. (2009). Bounded kernel-based online learning. *Journal of Machine Learning Research*, *10*, 2571–2594.

Orabona, F., Cesa-Bianchi, N., & Gentile, C. (2012). Beyond logarithmic bounds in online learning. In *Proceedings of the 15th international conference on artificial intelligence and statistics* (pp. 823–831). JMLR W&CP.

Rakhlin, A.,& Tewari, A. (2008). Lecture notes on online learning. Technical report. http://stat.wharton.upenn.edu/~rakhlin/papers/online_learning.pdf .

Ross, S., Mineiro, P., & Langford, J. (2013). Normalized online learning. In *Proceedings of the 29th conference on uncertainty in artificial intelligence* (pp. 537–545). AUAI Press.

Shalev-Shwartz, S. (2007). *Online learning: Theory, algorithms, and applications*. Ph.D. thesis, The Hebrew University.

Shalev-Shwartz, S. (2012). Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, *4*(2), 107–194.

Shalev-Shwartz, S., & Kakade, S. M. (2009). Mind the duality gap: Logarithmic regret algorithms for online optimization. *Advances in Neural Information Processing Systems*, *21*, 1457–1464.

Shalev-Shwartz, S., & Singer, Y. (2007). A primal–dual perspective of online learning algorithms. *Machine Learning Journal*, *69*(2–3), 115–142.

Srebro, N., Sridharan, K., & Tewari, A. (2011). On the universality of online mirror descent. *Advances in Neural Information Processing Systems*, *24*, 2645–2653.

Sridharan, K., & Tewari, A. (2010a). Convex games in banach spaces. In *Proceedings of the 23rd conference on learning theory* (pp. 1–13). Omnipress.

Sridharan, K.,& Tewari, A. (2010b). Convex games in banach spaces. In *Proceedings of the 23rd annual conference on learning theory* (pp. 1–13). Citeseer.

Steinhardt, J., & Liang, P. (2014). Adaptivity and optimism: An improved exponentiated gradient algorithm. In *Proceedings of the 31st international conference on machine learning* (pp. 1593–1601).

Streeter, M., & McMahan, B. (2012). No-regret algorithms for unconstrained online convex optimization. In Bartlett, P., Pereira, F. C. N., Burges, C. J. C., Bottou, L., Weinberger, K. Q. (Eds.), *Advances in neural information processing systems* (Vol. 25, pp. 2411–2419).

Vovk, V. (2001). Competitive on-line statistics. *International Statistical Review*, *69*, 213–248.

Warmuth, M.K., & Jagota, A.K. (1997). Continuous and discrete-time nonlinear gradient descent: Relative loss bounds and convergence. In *Electronic proceedings of the 5th international symposium on artificial intelligence and mathematics*.

Xiao, L. (2010). Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, *11*, 2543–2596.