Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

A Generalized Online Mirror Descent with Applications to Classification and Regression

Francesco Orabona FRANCESCO@ORABONA.COM
Toyota Technological Institute at Chicago
60637 Chicago, IL, USA

Koby Crammer KOBY@EE.TECHNION.AC.IL
Department of Electrical Enginering
The Technion
Haifa, 32000 Israel

Nicolò Cesa-Bianchi NICOLO.CESA-BIANCHI@UNIMI.IT
Department of Computer Science
Università degli Studi di Milano
Milano, 20135 Italy

Editor: ...

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 other on multiplicative updates, like Winnow. Online convex optimization is a general framework to unify both the design and the analysis of online algorithms using a single prediction strategy: online mirror descent. Different first-order online algorithms are obtained by choosing the regularization function in online mirror descent. We generalize online mirror descent to sequences of time-varying regularizers. Our approach allows us to recover as special cases many recently proposed second-order algorithms, such as the Vovk-Azoury-Warmuth, the second-order Perceptron, and the AROW algorithm. Moreover, we derive a new second order adaptive p-norm algorithm, and improve bounds for some first-order algorithms, such as Passive-Aggressive (PA-I).

Keywords: Online learning, Convex optimization, Second-order algorithms

1. Introduction

Online learning provides a scalable and flexible approach for the solution of a wide range of prediction problems, including classification, regression, ranking, and portfolio management. Popular online algorithms for classification include the standard Perceptron and its many variants, such as kernel Perceptron (Freund and Schapire, 1999), p-norm Perceptron (Gentile, 2003), and Passive-Aggressive (Crammer et al., 2006). These algorithms have well known counterparts for regression problems, such the Widrow-Hoff algorithm and its p-norm generalization. Other online algorithms, with properties different from those of the standard Perceptron, are based on exponential (rather than additive) updates, such as Winnow (Littlestone, 1988) for classification and Exponentiated Gradient (Kivinen and

Warmuth, 1997) for regression. Whereas these online algorithms are all essentially variants of stochastic gradient descent (Tsypkin, 1971), in the last decade many algorithms using second-order information from the input features have been proposed. These include the Vovk-Azoury-Warmuth algorithm for regression (Vovk, 2001; Azoury and Warmuth, 2001), the second-order Perceptron (Cesa-Bianchi et al., 2005), the CW/AROW algorithms (Dredze et al., 2008; Crammer et al., 2009,?), and the algorithms proposed by Duchi et al. (2011), all for binary classification.

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 general online convex optimization algorithm which is parametrized by a regularizer, i.e., a strongly convex function. By appropriate choices of the regularizer, most first-order online learning algorithms are recovered as special cases of OMD. Moreover, performance guarantees can be also derived simply by instantiating the general OMD bounds to the specific regularizer being used. The theoretical study of OMD relies on convex analysis. Warmuth and Jagota (1997) and Kivinen and Warmuth (2001) pioneered the use of Bregman divergences in the analysis of online algorithms, as explained in the monography of Cesa-Bianchi and Lugosi (2006). Shalev-Shwartz and Singer (2007), Shalev-Shwartz (2007) in his disseration, and Shalev-Shwartz and Kakade (2009) showed a different analysis based on a primal-dual method. Starting from the work of Kakade et al. (2009), it is now clear that many instances of OMD can be analyzed using only a few basic convex duality properties. See the the recent survey by Shalev-Shwartz (2012) for a lucid description of these developments.

In this paper we extend and generalize the theoretical framework of Kakade et al. (2009). In particular, we allow OMD to use a sequence of time-varying regularizers. This is known to be the key to obtaining second-order algorithms, and indeed we recover the Vovk-Azoury-Warmuth, the second-order Perceptron, and the AROW algorithm as special cases, with a slightly improved analysis of AROW. Our generalized analysis also captures the efficient variants of these algorithms that only use the diagonal elements of the second order information matrix, a result which was not within reach of the previous techniques.

Besides being able to express second-order algorithms, time-varying regularizers can be used to perform other types of adaptation to the sequence of observed data. We give a concrete example by introducing a new adaptive regularizer corresponding to a weighted version of the p-norm regularizer. In the case of sparse targets, the corresponding instance of OMD achieves a performance bound better than that of OMD with 1-norm regularization, which is the standard regularizer for the sparse target assumption.

Even in case of first-order algorithms our framework gives improvements on previous results. 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 so far elusive. 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.

2. Online convex programming

Let $\mathbb{X}$ be some Euclidean space (a finite-dimensional linear space over the reals equipped with an inner product). In the online convex optimization protocol an algorithm sequentially chooses elements from $S \subseteq \mathbb{X}$, 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,

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

for all $\boldsymbol{u} \in S$ and for any sequence of convex loss functions ℓ_t. An important application domain for this protocol is sequential linear regression/classification. In this case, there is a fixed and given 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 $\ell_t(\boldsymbol{w}_t) = \ell\big(\langle \boldsymbol{w}, \boldsymbol{x}_t \rangle, y_t\big)$. For example, in regression $\ell\big(\langle \boldsymbol{w}, \boldsymbol{x}_t \rangle, y_t\big) = \big(\langle \boldsymbol{w}, \boldsymbol{x}_t \rangle - y_t\big)^2$. In classification, where $y_t \in \{-1, +1\}$, a typical loss function is the hinge loss $\big[1 - y_t \langle \boldsymbol{w}, \boldsymbol{x}_t \rangle\big]_+$, where $[a]_+ = \max\{0, a\}$. This is a convex upper bound on the true quantity of interest. Namely, the mistake indicator function $\mathbb{I}_{\{y_t \langle \boldsymbol{w}, \boldsymbol{x}_t \rangle \leq 0\}}$.

2.1 Further notation and definitions

We now introduce some basic notions of convex analysis that are used in the paper. We refer to Rockafellar (1970) for definitions and terminology. We consider functions $f : \mathbb{X} \to \mathbb{R}$ that are closed and convex. This is equivalent to say that their epigraph $\{(\boldsymbol{x}, y) : f(\boldsymbol{x}) \leq y\}$ is a convex and closed subset of $\mathbb{X} \times \mathbb{R}$. The (effective) domain of f, that is the set $\{\boldsymbol{x} \in \mathbb{X} : f(\boldsymbol{x}) < \infty\}$, is a convex set whenever f is convex. We can always choose any $S \subseteq \mathbb{X}$ as 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 as $f^*(\boldsymbol{u}) = \sup_{\boldsymbol{v} \in S}\big(\langle \boldsymbol{v}, \boldsymbol{u} \rangle - f(\boldsymbol{v})\big)$. 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 as $\|\boldsymbol{v}\|_* = \sup_{\boldsymbol{u}} \{\langle \boldsymbol{u}, \boldsymbol{v} \rangle : \|\boldsymbol{u}\| \leq 1\}$. The Fenchel-Young inequality states that $f(\boldsymbol{u}) + f^*(\boldsymbol{v}) \geq \langle \boldsymbol{u}, \boldsymbol{v} \rangle$ for all $\boldsymbol{v}, \boldsymbol{u}$.

A vector $\boldsymbol{x}$ is a subgradient of a convex function f at $\boldsymbol{v}$ if $f(\boldsymbol{u}) - f(\boldsymbol{v}) \geq \langle \boldsymbol{u} - \boldsymbol{v}, \boldsymbol{x} \rangle$ 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 $\boldsymbol{x} \in \partial f(\boldsymbol{u})$,

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

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 $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{X}$,

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

See also the paper of Kakade et al. (2009) and references therein. A further property of strongly convex functions $f : S \to \mathbb{R}$ is the following: for all $\boldsymbol{u} \in \mathbb{X}$,

$$\nabla f^*(\boldsymbol{u}) = \operatorname*{argsup}_{\boldsymbol{v} \in S} \Big(\big\langle \boldsymbol{v}, \boldsymbol{u} \big\rangle - f(\boldsymbol{v}) \Big) \ . \tag{1}$$

This implies the useful identity

$$f\big(\nabla f^*(\boldsymbol{u})\big) + f^*(\boldsymbol{u}) = \big\langle \nabla f^*(\boldsymbol{u}), \boldsymbol{u} \big\rangle \ . \tag{2}$$

Strong convexity and strong smoothness are key properties in the design of online learning algorithms. In the following, we often write $\|\cdot\|_f$ to denote the norm according to which f is strongly convex.

3. Online Mirror Descent

We now introduce our main algorithmic tool: a generalization of the standard OMD algorithm for online convex programming in which the regularizers may change over time.

Algorithm 1 Online Mirror Descent

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

Standard OMD —see, e.g., (Kakade et al., 2009)— uses $f_t = f$ for all t. Note the following remarkable property of Algorithm 1: while $\boldsymbol{\theta}_t$ moves freely in $\mathbb{X}$ as determined by the input sequence $\boldsymbol{z}_t$, because of (1) the property $\boldsymbol{w}_t \in S$ holds for all t.

The following lemma is a generalization of Corollary 4 of Kakade et al. (2009) and of Corollary 3 of Duchi et al. (2011).

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

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

where we set $f_0^*(\mathbf{0}) = 0$. Moreover, $f_t^*(\boldsymbol{\theta}_t) - f_{t-1}^*(\boldsymbol{\theta}_t) \leq f_{t-1}(\boldsymbol{w}_t) - f_t(\boldsymbol{w}_t)$ for all $t \geq 1$.

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 $\big(\|\cdot\|_{f_t}\big)_*$, and recalling that $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \boldsymbol{z}_t$,

$$\begin{aligned}
\Delta_t &= f_t^*(\boldsymbol{\theta}_{t+1}) - f_t^*(\boldsymbol{\theta}_t) + f_t^*(\boldsymbol{\theta}_t) - f_{t-1}^*(\boldsymbol{\theta}_t) \\
&\leq f_t^*(\boldsymbol{\theta}_t) - f_{t-1}^*(\boldsymbol{\theta}_t) - \langle \nabla f_t^*(\boldsymbol{\theta}_t), \boldsymbol{z}_t \rangle + \frac{1}{2\beta_t}\big(\|\boldsymbol{z}_t\|_{f_t}\big)_*^2 \\
&= f_t^*(\boldsymbol{\theta}_t) - f_{t-1}^*(\boldsymbol{\theta}_t) - \langle \boldsymbol{w}_t, \boldsymbol{z}_t \rangle + \frac{1}{2\beta_t}\big(\|\boldsymbol{z}_t\|_{f_t}\big)_*^2
\end{aligned}$$

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

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

We now prove the second statement. Recalling again the definition of $\boldsymbol{w}_t$ we have that (2) 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. ■

Next, we show a general regret bound for Algorithm 1.

Corollary 1 *Let $R : S \to \mathbb{R}$ be a convex function and let $g_1, g_2, \ldots$ be a sequence of nondecreasing convex functions $g_t : S \to \mathbb{R}$. Fix $\eta > 0$ and assume $f_t = g_t + \eta t R$ are β_t-strongly convex with respect to $\|\cdot\|$. 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*

$$\sum_{t=1}^{T} \big(\ell_t(\boldsymbol{w}_t) + R(\boldsymbol{w}_t)\big) - \sum_{t=1}^{T} \big(\ell_t(\boldsymbol{u}) + R(\boldsymbol{u})\big) \leq \frac{g_T(\boldsymbol{u})}{\eta} + \eta \sum_{t=1}^{T} \frac{\big(\big\|\boldsymbol{\ell}_t'\big\|_{f_t}\big)_*^2}{2\beta_t} \tag{3}$$

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

Moreover, if $f_t = g\sqrt{t} + \eta t R$ where $g : S \to \mathbb{R}$ is β-strongly convex, then

$$\sum_{t=1}^{T} \big(\ell_t(\boldsymbol{w}_t) + R(\boldsymbol{w}_t)\big) - \sum_{t=1}^{T} \big(\ell_t(\boldsymbol{u}) + R(\boldsymbol{u})\big) \leq \sqrt{T} \left(\frac{g(\boldsymbol{u})}{\eta} + \frac{\eta}{\beta} \max_{t \leq T} \big(\big\|\boldsymbol{\ell}_t'\big\|_{f_t}\big)_*^2 \right) \tag{4}$$

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

Finally, if $f_t = t\,R$, where R is β-strongly convex with respect to a norm $\|\cdot\|$, then

$$\sum_{t=1}^{T}\big(\ell_t(\boldsymbol{w}_t) + R(\boldsymbol{w}_t)\big) - \sum_{t=1}^{T}\big(\ell_t(\boldsymbol{u}) + R(\boldsymbol{u})\big) \leq \max_{t\leq T}\big(\big\|\boldsymbol{\ell}'_t\big\|_{f_t}\big)_*^2 \frac{(1+\ln T)}{2\beta} \tag{5}$$

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,

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

where we used the fact that the terms $g_{t-1}(\boldsymbol{w}_t) - g_t(\boldsymbol{w}_t)$ are nonpositive under the hypothesis that the functions g_t are nondecreasing. Reordering terms we obtain (3). In order to obtain (4) 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 (4). Finally, bound (5) is proven by observing that $f_t = t\,R$ is βt-strongly convex because R is β-strongly convex. The elementary inequality $\sum_{t=1}^{T}\frac{1}{t} \leq 1 + \ln T$ concludes the proof. ■

A special case of OMD is the Regularized Dual Averaging framework of Xiao (2010), where the prediction at each step is defined by

$$\boldsymbol{w}_t = \operatorname*{argmin}_{\boldsymbol{w}} \frac{1}{t-1}\sum_{s=1}^{t-1}\boldsymbol{w}^\top\boldsymbol{\ell}'_s + \frac{\beta_{t-1}}{t-1}g(\boldsymbol{w}) + R(\boldsymbol{w}) \tag{6}$$

for some $\boldsymbol{\ell}'_s \in \partial\ell_s(\boldsymbol{w}_s)$, $s = 1,\dots,t-1$. Using (1), it is easy to see that this update is equivalent[1] to

$$\boldsymbol{w}_t = \nabla f_t^*\left(\sum_{s=1}^{t-1}\boldsymbol{\ell}'_s\right)$$

where $f_t(\boldsymbol{w}) = \beta_{t-1}\,g(\boldsymbol{w}) + (t-1)\,R(\boldsymbol{w})$. 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. However, their framework is not flexible enough to include algorithms that update without using the gradient of the loss function with respect to which the regret is calculated. Examples of such algorithms are the Vovk-Azoury-Warmuth algorithm of the next section and the online binary classification algorithms of Section 6.

A bound similar to (3) has been recently presented by Duchi et al. (2011) and extended to the variable potential functions 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.

Note that the only hypothesis on R is convexity. Hence, R can be a nondifferentiable function as $\|\cdot\|_1$. Thus we recover the results about minimization of strongly convex and composite loss functions, and adaptive learning rates, in a simple unique framework. In the next sections we show more algorithms that can be viewed as special cases of this framework.

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

4. Square Loss

In this section we recover known regret bounds for online regression with the square loss via Lemma 1. Throughout this section, $\mathbb{X} = \mathbb{R}^d$ and the inner product $\langle \boldsymbol{u}, \boldsymbol{x} \rangle$ is the standard dot product $\boldsymbol{u}^\top \boldsymbol{x}$. We set $\ell_t(\boldsymbol{u}) = \frac{1}{2}\big(y_t - \boldsymbol{u}^\top \boldsymbol{x}_t\big)^2$ 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 by setting $\boldsymbol{z}_t = -y_t \boldsymbol{x}_t$ and $f_t(\boldsymbol{u}) = \frac{1}{2}\boldsymbol{u}^\top A_t \boldsymbol{u}$, where $A_1 = aI_d$ and $A_t = A_{t-1} + \boldsymbol{x}_t \boldsymbol{x}_t^\top$ for $t > 1$. The regret bound of this algorithm —see, e.g., Theorem 11.8 of Cesa-Bianchi and Lugosi (2006)— is recovered from Lemma 1 by noting that f_t is 1-strongly convex with respect to the norm $\|\boldsymbol{u}\|_{f_t} = \sqrt{\boldsymbol{u}^\top A_t \boldsymbol{u}}$. Hence,

$$\begin{aligned}
R_T &= \sum_{t=1}^T (y_t \boldsymbol{u}^\top \boldsymbol{x}_t - y_t \boldsymbol{w}_t^\top \boldsymbol{x}_t) - f_T(\boldsymbol{u}) + \frac{a}{2}\|\boldsymbol{u}\|^2 + \frac{1}{2}\sum_{t=1}^T (\boldsymbol{w}_t^\top \boldsymbol{x}_t)^2 \\
&\le f_T(\boldsymbol{u}) + \sum_{t=1}^T \left(\frac{y_t^2 \big(\|\boldsymbol{x}_t\|_{f_t}\big)_*^2}{2} + f_t^*(\boldsymbol{\theta}_t) - f_{t-1}^*(\boldsymbol{\theta}_t) \right) - f_T(\boldsymbol{u}) + \frac{a}{2}\|\boldsymbol{u}\|^2 + \frac{1}{2}\sum_{t=1}^T (\boldsymbol{w}_t^\top \boldsymbol{x}_t)^2 \\
&\le \frac{a}{2}\|\boldsymbol{u}\|^2 + \frac{Y}{2}\sum_{t=1}^T \boldsymbol{x}_t^\top A_t^{-1} \boldsymbol{x}_t
\end{aligned}$$

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

We can also generalize the p-norm LMS algorithm of Kivinen et al. (2006) for controlling the adaptive filtering regret

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

(The reader interested in the motivations behind the study of this regret is addressed to that paper.) This is achieved by setting $\boldsymbol{z}_t = -\big(y_t - \boldsymbol{w}_t^\top \boldsymbol{x}_t\big)\boldsymbol{x}_t$ and $f_t(\boldsymbol{u}) = \frac{X_t}{\beta} f(\boldsymbol{u})$ in OMD, where f is an arbitrary β-strongly convex function with respect to some norm $\|\cdot\|$, and $X_t = \max_{s \le t} \|\boldsymbol{x}_s\|_*$. We can then write

$$\begin{aligned}
R_T + \frac{1}{2}R_T^{\text{AF}} &= \sum_{t=1}^T \Big(\big(y_t - \boldsymbol{w}_t^\top \boldsymbol{x}_t\big)\boldsymbol{u}^\top \boldsymbol{x}_t - \big(y_t - \boldsymbol{w}_t^\top \boldsymbol{x}_t\big)\boldsymbol{w}_t^\top \boldsymbol{x}_t \Big) \\
&\le f_T(\boldsymbol{u}) + \frac{1}{2}\sum_{t=1}^T \big(y_t - \boldsymbol{w}_t^\top \boldsymbol{x}_t\big)^2
\end{aligned}$$

where in the last step we used Lemma 1, the X_t-strong convexity of f_t, and the fact the $f_t \ge f_{t-1}$. Simplifying the expression we obtain the following adaptive filtering bound

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

Compared to the bounds of Kivinen et al. (2006), our algorithm inherits the ability to adapt to the maximum norm of $\boldsymbol{x}_t$ without any prior knowledge. Moreover, instead of using a decreasing learning rate here we use an increasing regularizer.

5. A new algorithm for online regression

In this section we show the full power of our framework by introducing a new time-varying regularizer f_t generalizing the squared q-norm. Then, we derive the corresponding regret bound. 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 $\boldsymbol{u}^\top \boldsymbol{x}$.

Given $(b_1, \dots, b_d) \in \mathbb{R}_+$ and $q \in (1, 2]$ let the weighted q-norm of $\boldsymbol{w} \in \mathbb{R}^d$ be

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

Define the corresponding regularization function by

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

This function has the following properties (proof in appendix).

Lemma 2 *The Fenchel conjugate of f is*

$$f^*(\boldsymbol{\theta}) = \frac{1}{2(p-1)} \left(\sum_{i=1}^{d} |\theta_i|^p \, b_i^{1-p} \right)^{2/p} \qquad \text{for} \quad p = \frac{q}{q-1}. \tag{7}$$

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

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

whose dual norm is defined by

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

We can now prove the following regret bound for linear regression with absolute loss.

Corollary 3 *Let*

$$f_t(\boldsymbol{u}) = \frac{\sqrt{2et}}{2\sqrt{q_t - 1}} \left(\sum_{i=1}^{d} |u_i|^{q_t} b_{t,i} \right)^{2/q_t}$$

where $b_{t,i} = \max_{s=1,\dots,t} |x_{s,i}|$, and let

$$q_t = \left(1 - \frac{1}{2 \ln \max_{s=1,\dots,t} \|\boldsymbol{x}_s\|_0} \right)^{-1} .$$

If OMD is run using regularizers f_t on the input sequence $\boldsymbol{z}_t = -\eta\boldsymbol{\ell}'_t$, where $\boldsymbol{\ell}'_t \in \partial\ell_t(\boldsymbol{w}_t)$ for $\ell_t(\boldsymbol{w}) = \left|\boldsymbol{w}^\top\boldsymbol{x}_t - y_t\right|$ and $\eta > 0$, then

$$\sum_{t=1}^{T} |\boldsymbol{w}_t^\top\boldsymbol{x}_t - y_t| - \sum_{t=1}^{T} |\boldsymbol{u}^\top\boldsymbol{x}_t - y_t| \leq \sqrt{2eT}\sqrt{2\ln\max_{t=1,\dots,T}\|\boldsymbol{x}_t\|_0 - 1}\left(\frac{\left(\sum_{i=1}^{d}|u_i|B_{T,i}\right)^2}{\eta} + \eta\right)$$

for any $\boldsymbol{u} \in \mathbb{R}^d$, where $B_{T,i} = \max_{t=1,\dots,T}|x_{t,i}|$.

This bound has the interesting property to be invariant with respect to arbitrary scaling of individual coordinates of the data points $\boldsymbol{x}_t$. This is unlike running standard OMD with non-adaptive regularizers, which gives bounds of the form $\|\boldsymbol{u}\|\max_t\|\boldsymbol{x}_t\|_*\sqrt{T}$. In particular, by an appropriate tuning of η the regret in Corollary 3 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, that is $\|\boldsymbol{u}\|_1$ are small, this is always better than running standard OMD with a non-weighted q-norm regularizer, which for $q \to 1$ (the best choice for the sparse $\boldsymbol{u}$ case) gives bounds of the form

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

Indeed, 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.

6. Binary classification: aggressive and diagonal updates

In this section we show that several known algorithms for online binary classification are special cases of OMD. These algorithms include p-norm Perceptron (Gentile, 2003), Passive-Aggressive (Crammer et al., 2006), second-order Perceptron (Cesa-Bianchi et al., 2005), and AROW (Crammer et al., 2009). Besides recovering all previously known mistake bounds, we also show new bounds for Passive-Aggressive and for AROW with diagonal updates.

Fix any Euclidean space with inner product $\langle\cdot,\cdot\rangle$. 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 $\ell_t(\boldsymbol{w}) = \ell\big(\langle\boldsymbol{w}, \boldsymbol{x}_t\rangle, y_t\big)$ be the hinge loss $\big[1 - y_t\langle\boldsymbol{w}, \boldsymbol{x}_t\rangle\big]_+$. It is easy to verify that the hinge loss satisfies the following condition:

$$\text{if } \ell_t(\boldsymbol{w}) > 0 \quad \text{then} \quad \ell_t(\boldsymbol{u}) \geq 1 + \langle\boldsymbol{u}, \boldsymbol{\ell}'_t\rangle \quad \text{for all } \boldsymbol{u}, \boldsymbol{w} \in \mathbb{R}^d \text{ with } \boldsymbol{\ell}'_t \in \partial\ell_t(\boldsymbol{w}). \tag{8}$$

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

In the following, when T is understood from the context, we denote by $\mathcal{M}$ the set of steps t on which the algorithm made a mistake, $\widehat{y}_t \neq y_t$. Similarly, we denote by $\mathcal{U}$ the set of margin error steps; that is, steps where $\widehat{y}_t = y_t$ but $\ell_t(\boldsymbol{w}_t) > 0$. 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

If we run OMD in conservative mode, and let $f_t = f = \frac{1}{2}\|\cdot\|_p^2$ for $1 < p \leq 2$, then we recover the p-norm Perceptron of Gentile (2003). We now show how to use our framework to generalize and improve previous analyses for binary classification algorithms that use aggressive updates.

Corollary 4 *Assume OMD is run with $f_t = f$ where f, with 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 $\boldsymbol{z}_t = \eta_t y_t \boldsymbol{x}_t$, for some $0 < \eta_t \leq 1$ such that $y_t \langle \boldsymbol{w}_t, \boldsymbol{x}_t \rangle \leq 0$ implies $\eta_t = 1$. Then, for all $T \geq 1$ and for all $\boldsymbol{u} \in \mathbb{X}$,*

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

where $M = |\mathcal{M}|$, $X_T = \max_{t \leq T} \|\boldsymbol{x}_t\|_*$,

$$L(\boldsymbol{u}) = \sum_{t=1}^{T} \left[1 - y_t \langle \boldsymbol{u}, \boldsymbol{x}_t \rangle\right]_+ \qquad \text{and} \qquad D = \sum_{t \in \mathcal{U}} \eta_t \left(\frac{\eta_t \|\boldsymbol{x}_t\|_*^2 + 2\beta y_t \langle \boldsymbol{w}_t, \boldsymbol{x}_t \rangle}{X_t^2} - 2 \right) .$$

For the conservative p-norm Perceptron, 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 Lemma 17 of Shalev-Shwartz (2007). We therefore recover the mistake bound of Gentile (2003).

The term D in the bound of Corollary 4 can 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 4 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

$$M \leq L(\boldsymbol{u}) + \big(\|\boldsymbol{u}\| \, X_T\big)^2 + \|\boldsymbol{u}\| \, X_T \sqrt{L(\boldsymbol{u})}, \tag{9}$$

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.

Proof (of Corollary 4) Using (15) in Lemma 5 with the assumption $\eta_t = 1$ when $t \in \mathcal{M}$, we get

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

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

$$M \leq L(\boldsymbol{u}) + \frac{1}{\beta} f(\boldsymbol{u}) X_T^2 + X_T \sqrt{\frac{2}{\beta} f(\boldsymbol{u})} \sqrt{\frac{1}{2\beta} X_T^2 f(\boldsymbol{u}) + L(\boldsymbol{u}) + D'} - \sum_{t \in \mathcal{U}} \eta_t \tag{10}$$

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 (10) 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

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

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

6.2 Second-order algorithms

We now apply our framework to second-order algorithms for binary classification. Here, we let $\mathbb{X} = \mathbb{R}^d$ and 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

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

The closed-form expression for $\boldsymbol{w}_{t+1}$, which involves the design matrix $S_t = \left[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t\right]$ 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 $\boldsymbol{w}_{t+1}$, but S_t and $\boldsymbol{y}_t$ only contain the examples $(\boldsymbol{x}_s, y_s)$ on which a mistake occurred. In this sense, we call it an online variant of Ridge regression.

In practice, second-order algorithms perform typically 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 $f_t(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top A_t \boldsymbol{x}$, where $A_0 = I$ and $A_t = A_{t-1} + \frac{1}{r}\boldsymbol{x}_t\boldsymbol{x}_t^\top$ with $r > 0$. Each function f_t is 1-strongly convex with respect to the norm $\|\boldsymbol{x}\|_{f_t} = \sqrt{\boldsymbol{x}^\top A_t \boldsymbol{x}}$ with dual norm $\left(\|\boldsymbol{x}\|_{f_t}\right)_* = \sqrt{\boldsymbol{x}^\top A_t^{-1} \boldsymbol{x}}$. The dual function of $f_t(\boldsymbol{x})$ is $f_t^*(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top A_t^{-1}\boldsymbol{x}$. Now, the conservative version of OMD run with f_t chosen as above is the second-order Perceptron. The aggressive version corresponds instead to AROW with a minor difference. Indeed, in this case the prediction of OMD is the sign of $y_t \boldsymbol{w}_t^\top \boldsymbol{x}_t = m_t \frac{r}{r+\chi_t}$, where we use the notation $\chi_t = \boldsymbol{x}_t^\top A_{t-1}^{-1}\boldsymbol{x}_t$ and $m_t = y_t \boldsymbol{x}_t^\top A_{t-1}^{-1}\boldsymbol{\theta}_t$. On the other hand, AROW simply predicts using the sign of m_t. The sign of the predictions is the same, but OMD updates when $m_t \frac{r}{r+\chi_t} \leq 1$ while AROW updates when $m_t \leq 1$. Typically, for large t the value of χ_t is small, and thus the two update rules coincide in practice.

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

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

Hence, using (15) in Lemma 5, and setting $\eta_t = 1$, we obtain

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

for all $\boldsymbol{u} \in \mathbb{X}$, where

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

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

$$\sum_{t\in\mathcal{M}\cup\mathcal{U}} \frac{m_t(2r-m_t)}{r(r+\chi_t)} \le \sum_{t\in\mathcal{U}} \frac{m_t(2r-m_t)}{r(r+\chi_t)} \le \sum_{t\in\mathcal{U}} \frac{r^2}{r(r+\chi_t)} \le U \tag{11}$$

In the conservative case, when $\mathcal{U}\equiv\emptyset$, the bound specializes to the standard second-order Perceptron bound.

6.3 Diagonal updates

AROW and the second-order Perceptron can be run more efficiently using diagonal matrices. In this case, each update takes time linear in d. We now use Corollary 5 to prove a mistake bound for the diagonal version of the second-order Perceptron. Denote $D_t = \mathrm{diag}\{A_t\}$ be the diagonal matrix that agrees with A_t on the diagonal, where A_t is defined as before and $f_t(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top D_t\,\boldsymbol{x}$. Setting $\eta_t = 1$, using the second bound of Lemma 5, and Lemma 6, we have[2]

$$\begin{aligned} M + U &\le L(\boldsymbol{u}) + \sqrt{\boldsymbol{u}^T D_T \boldsymbol{u}\left(r\sum_{i=1}^{d}\ln\left(\frac{1}{r}\sum_{t\in\mathcal{M}\cup\mathcal{U}} x_{t,i}^2 + 1\right) + 2U\right)} \\ &= L(\boldsymbol{u}) + \sqrt{\|\boldsymbol{u}\|^2 + \frac{1}{r}\sum_{i=1}^{d} u_i^2\left(\sum_{t\in\mathcal{M}\cup\mathcal{U}} x_{t,i}^2\right)}\sqrt{r\sum_{i=1}^{d}\ln\left(\frac{1}{r}\sum_{t\in\mathcal{M}\cup\mathcal{U}} x_{t,i}^2 + 1\right) + 2U}\,. \end{aligned} \tag{12}$$

This allows us to theoretically analyze the cases where this algorithm could be advantageous. In particular, features of NLP 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).

Figure 1 shows the number of times each feature (word) appears in two sentiment datasets vs. 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

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

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,

$$\sum_{i=1}^{d} u_i^2 \sum_{t\in\mathcal{M}\cup\mathcal{U}} x_{t,i}^2 \le s\|\boldsymbol{u}\|^2 \tag{13}$$

2. We did not optimize the constant multiplying U in the bound.

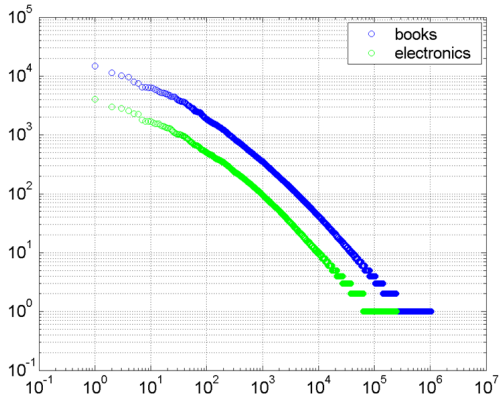

Figure 1: Evidence of heavy tails for NLP Data. The plots show the number of words vs. the word rank on two sentiment data sets.

the last term in the mistake bound (12) can be re-written as

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

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 3 in the appendix to obtain

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

Hence, when the hypothesis (13) is verified, the number of mistakes of the diagonal version of AROW depends on $\sqrt{\ln L(\boldsymbol{u})}$ rather than on $\sqrt{L(\boldsymbol{u})}$.

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 forecaster, the second-order Perceptron, and the AROW algorithm) 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 a new regularizer based on an adaptive weighted version of the p-norm Perceptron. In the case of sparse targets, the

corresponding instance of OMD achieves a performance bound 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 typically 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. More in general, it would be useful to devise a more systematic approach to the design of adaptive regularizers enojoying a given set of desired properties. This would help obtaining more examples of adaptation mechanisms that are not based on second-order information.

Acknowledgments

The third author gratefully acknowledges partial support by the PASCAL2 Network of Excellence under EC grant no. 216886. This publication only reflects the authors' views. The second author gratefully acknowledges partial support by an Israeli Science Foundation grant ISF-1567/10.

Technical lemmas

Proof (of Lemma 2) The Fenchel conjugate of f is $f^*(\boldsymbol{\theta}) = \sup_{\boldsymbol{v}} \left(\boldsymbol{v}^\top \boldsymbol{\theta} - f(\boldsymbol{v})\right)$. Set $\boldsymbol{w}$ equal to the gradient of $\frac{1}{2(p-1)} \left(\sum_{i=1}^d |\theta_i|^p \, b_i^{1-p}\right)^{2/p}$ with respect to $\boldsymbol{\theta}$. Easy calculations show that

$$\boldsymbol{w}^\top \boldsymbol{\theta} - f(\boldsymbol{w}) = \frac{1}{2(p-1)} \left(\sum_{i=1}^d |\theta_i|^p \, b_i^{1-p}\right)^{2/p} .$$

We now show that this quantity is indeed $\sup_{\boldsymbol{v}} \boldsymbol{v}^\top \boldsymbol{\theta} - f(\boldsymbol{v})$. Pick any $\boldsymbol{v} \in \mathbb{R}^d$. Applying Hölder inequality to the vectors $(v_1 b_1^{1/q}, \ldots, v_d b_d^{1/q})$ and $(\theta_1 b_1^{-1/q}, \ldots, \theta_d b_d^{-1/q})$ we get,

$$\boldsymbol{v}^\top \boldsymbol{\theta} \leq \left(\sum_{i=1}^d |v_i|^q \, b_i\right)^{1/q} \left(\sum_{i=1}^d |\theta_i|^p \, b_i^{-p/q}\right)^{1/p} = \left(\sum_{i=1}^d |v_i|^q \, b_i\right)^{1/q} \left(\sum_{i=1}^d |\theta_i|^p \, b_i^{1-p}\right)^{1/p} .$$

Hence

$$\boldsymbol{v}^\top \boldsymbol{\theta} - f(\boldsymbol{v}) \leq \left(\sum_{i=1}^d |v_i|^q \, b_i\right)^{1/q} \left(\sum_{i=1}^d |\theta_i|^p \, b_i^{1-p}\right)^{1/p} - \frac{1}{2(q-1)} \left(\sum_{i=1}^d |v_i|^q \, b_i\right)^{2/q} .$$

The right-hand side is a quadratic function in $\left(\sum_{i=1}^d |v_i|^q \, b_i\right)^{1/q}$. If we maximize it, we obtain

$$\boldsymbol{v}^\top \boldsymbol{\theta} - f(\boldsymbol{v}) \leq \frac{q-1}{2} \left(\sum_{i=1}^d |\theta_i|^p \, b_i^{1-p}\right)^{2/p} = \frac{1}{2(p-1)} \left(\sum_{i=1}^d |\theta_i|^p \, b_i^{1-p}\right)^{2/p}$$

which concludes the proof for f^*.

In order to show the second part, we follow Lemma 17 of Shalev-Shwartz (2007) and prove that $\left(\sum_{i=1}^d |x_i|^q b_i\right)^{2/q} \leq \boldsymbol{x}^\top \nabla^2 f(\boldsymbol{w}) \boldsymbol{x}$. Define $\Psi(a) = \frac{a^{2/q}}{2(q-1)}$ and $\phi(a) = |a|^q$, hence $f(w) = \Psi\big(\sum_{i=1}^d b_i \phi(x_i)\big)$. Clearly

$$\Psi'(a) = \frac{a^{2/q-1}}{q(q-1)} \qquad \text{and} \qquad \Psi''(a) = \frac{2/q-1}{q(q-1)} a^{2/q-2} \ .$$

Moreover, $\phi'(a) = q \operatorname{sign}(a)|a|^{q-1}$ and $\phi''(a) = q(q-1)|a|^{q-2}$. The (i,j) element of $\nabla^2 f(\boldsymbol{w})$ for $i \neq j$ is

$$\Psi'' \left(\sum_{k=1}^d b_k \phi(w_k)\right) b_i b_j \phi'(w_i) \phi'(w_j) \ .$$

The diagonal elements of $\nabla^2 f(\boldsymbol{w})$ are

$$\Psi'' \left(\sum_{k=1}^d b_k \phi(w_k)\right) b_i^2 \left(\phi'(w_i)\right)^2 + \Psi' \left(\sum_{k=1}^d b_k \phi(w_k)\right) b_i \phi''(w_i) \ .$$

Thus we have

$$\boldsymbol{x}^\top \nabla^2 f(\boldsymbol{w}) \boldsymbol{x} = \Psi'' \left(\sum_{k=1}^d b_k \phi(w_k)\right) \left(\sum_{i=1}^d b_i x_i \phi'(w_i)\right)^2 + \Psi' \left(\sum_{k=1}^d b_k \phi(w_k)\right) \sum_{i=1}^d b_i x_i^2 \phi''(w_i) \ .$$

The first term is non-negative since $q \in (1,2)$. Writing the second term explicitly we have,

$$\boldsymbol{x}^\top \nabla^2 f(\boldsymbol{w}) \boldsymbol{x} \geq \left(\sum_{k=1}^d b_k |w_k|^q\right)^{2/q-1} \sum_{i=1}^d b_i x_i^2 |w_i|^{q-2} \ .$$

We now lower bound this quantity using Hölder inequality. Let $y_i = b_i^\gamma |w_i|^{(2-q)q/2}$ for $\gamma = (2-q)/2$. We have

$$\begin{aligned}
\left(\sum_{i=1}^d x_i^q b_i\right)^{2/q} &= \left(\sum_{i=1}^d y_i \frac{x_i^q b_i}{y_i}\right)^{2/q} \leq \left(\left(\sum_{i=1}^d y_i^{2/(2-q)}\right)^{(2-q)/2} \left(\sum_{i=1}^d \frac{x_i^2 b_i^{2/q}}{y_i^{2/q}}\right)^{q/2}\right)^{2/q} \\
&= \left(\left(\sum_{i=1}^d \left(b_i^\gamma |w_i|^{(2-q)q/2}\right)^{2/(2-q)}\right)^{(2-q)/2} \left(\sum_{i=1}^d \frac{x_i^2 b_i^{2/q}}{\left(b_i^\gamma |w_i|^{(2-q)q/2}\right)^{2/q}}\right)^{q/2}\right)^{2/q}
\end{aligned}$$

$$
\begin{aligned}
&= \left(\sum_{i=1}^d \left(b_i^{2\gamma/(2-q)}|w_i|^q\right)\right)^{(2-q)/q} \left(\sum_{i=1}^d \frac{x_i^2 b_i^{2/q}}{b_i^{2\gamma/q}|w_i|^{(2-q)}}\right) \\
&= \left(\sum_{i=1}^d (b_i|w_i|^q)\right)^{(2-q)/q} \left(\sum_{i=1}^d x_i^2 |w_i|^{q-2} b_i^{2/q-2(2-q)/(2q)}\right) \\
&= \left(\sum_{i=1}^d (b_i|w_i|^q)\right)^{(2-q)/q} \left(\sum_{i=1}^d x_i^2 |w_i|^{q-2} b_i^{2/q-(2-q)/q}\right) \\
&= \left(\sum_{i=1}^d (b_i|w_i|^q)\right)^{(2-q)/q} \left(\sum_{i=1}^d x_i^2 |w_i|^{q-2} b_i\right) \ .
\end{aligned}
$$

We just showed that

$$
\boldsymbol{x}^\top \nabla^2 f(\boldsymbol{w}) \boldsymbol{x} \geq \left(\sum_{k=1}^d b_k |w_k|^q\right)^{2/q-1} \sum_{i=1}^d b_i x_i^2 |w_i|^{q-2} \geq \left(\sum_{i=1}^d x_i^q b_i\right)^{2/q} \ .
$$

This concludes the proof of the 1-strict 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,

$$
\begin{aligned}
\sup_{\boldsymbol{x}} \left\{ \boldsymbol{u}^\top \boldsymbol{x} \ : \ \left(\sum_{i=1}^d |x_i|^q b_i\right)^{1/q} \leq 1 \right\} &= \sup_{\boldsymbol{x}} \left\{ \boldsymbol{u}^\top \boldsymbol{x} \ : \ \left(\sum_{i=1}^d \left|x_i b_i^{1/q}\right|^q\right)^{1/q} \leq 1 \right\} \\
&= \sup_{\boldsymbol{y}} \left\{ \sum_i u_i y_i b_i^{-1/q} \ : \ \left(\sum_{i=1}^d |y_i|^q\right)^{1/q} \leq 1 \right\} \\
&= \left\| \left(u_1 b_1^{-1/q}, \dots, u_d b_d^{-1/q}\right) \right\|_p
\end{aligned}
$$

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 5 *Assume OMD is run with functions $f_1, f_2, \dots$ 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 $\boldsymbol{u} \in S$. Assume further the input sequence is $\boldsymbol{z}_t = -\eta_t \boldsymbol{\ell}'_t$ for some $\eta_t > 0$, where $\boldsymbol{\ell}'_t \in \partial \ell_t(\boldsymbol{w}_t)$, $\ell_t(\boldsymbol{w}_t) = 0$ implies $\boldsymbol{\ell}'_t = \boldsymbol{0}$, and $\ell_t = \ell\big(\langle \cdot, \boldsymbol{x}_t \rangle, y_t\big)$ satisfies (8). Then, for all $T \geq 1$,*

$$
\sum_{t \in \mathcal{M} \cup \mathcal{U}} \eta_t \leq L_\eta + \lambda f_T(\boldsymbol{u}) + \frac{1}{\lambda} \left(B + \sum_{t \in \mathcal{M} \cup \mathcal{U}} \left(\frac{\eta_t^2}{2\beta_t} \big(\|\boldsymbol{\ell}'_t\|_{f_t}\big)_*^2 - \eta_t \langle \boldsymbol{w}_t, \boldsymbol{\ell}'_t \rangle \right) \right) \tag{14}
$$

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

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

In particular, choosing the optimal λ*, we obtain*

$$\sum_{t \in \mathcal{M} \cup \mathcal{U}} \eta_t \leq L_\eta + 2\sqrt{f_T(\boldsymbol{u})} \sqrt{\left[B + \sum_{t \in \mathcal{M} \cup \mathcal{U}} \left(\frac{\eta_t^2}{2\beta_t} \big(\big\| \boldsymbol{\ell}_t' \big\|_{f_t} \big)_*^2 - \eta_t \langle \boldsymbol{w}_t, \boldsymbol{\ell}_t' \rangle \right) \right]_+} \ . \tag{15}$$

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

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

Since $\ell_t(\boldsymbol{w}_t) = 0$ implies $\boldsymbol{\ell}_t' = \boldsymbol{0}$, and using (8),

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

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 (14). ■

Lemma 6 *For all* $\boldsymbol{x}_1, \dots \boldsymbol{x}_T \in \mathbb{R}^d$ *let* $D_t = \operatorname{diag}\{A_t\}$ *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*

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

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

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

Using the above and the definition of D_t, we obtain

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

■

We conclude the appendix by proving the results required to solve the implicit logarithmic equations of Section 6.3. We use the following fact of Orabona et al. (2012).

Lemma 7 *Let $a, x > 0$ be such that $x \leq a \ln x$. Then for all $n > 1$*

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

Corollary 2 *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 3 *For all $a, b, c, d, x > 0$ such that*

$$x \leq \sqrt{a \ln(bx + 1) + c} + d \tag{16}$$

we have

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

Proof Assumption (16) implies

$$\begin{aligned} x^2 &\leq \left(\sqrt{a \ln(bx + 1) + c} + d \right)^2 \\ &\leq 2a \ln(bx + 1) + 2c + 2d^2 \\ &= a \ln(bx + 1)^2 + 2c + 2d^2 \\ &\leq a \ln(2b^2 x^2 + 2) + 2c + 2d^2 \ . \end{aligned} \tag{17}$$

From Corollary 2 we have that if $f, g, h, i, y > 0$ satisfy $y \leq f \ln(gx + h) + i$, then

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

where we have used the elementary inequality $\ln y \leq \frac{y}{e}$ for all $y \geq 0$. Applying the above to (17) we obtain

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

which implies

$$x \leq \sqrt{\frac{n}{n-1}} \left(\frac{\sqrt{2n} ab}{e} + \sqrt{2c} + \sqrt{2} d \right) + \frac{1}{b} \frac{1}{\sqrt{n-1}} \ . \tag{18}$$

Note that we have repeatedly used the elementary inequality $\sqrt{x + y} \leq \sqrt{x} + \sqrt{y}$. Choosing $n = 2$ and applying (18) to (16) we get

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

concluding the proof. ■

References

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

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

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

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

K. Crammer, M. Dredze, and F. Pereira. Exact convex confidence-weighted learning. *Advances in Neural Information Processing Systems*, 21:345–352, 2009.

K. Crammer, A. Kulesza, and M. Dredze. Adaptive regularization of weight vectors. *Advances in Neural Information Processing Systems*, 22:414–422, 2009.

M. Dredze, K. Crammer, and F. Pereira. Online confidence-weighted learning. *Proceedings of the 25th International Conference on Machine Learning*, 2008.

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

J. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *Proceedings of the 23rd Annual Conference on Learning Theory*, pages 14–26, 2010.

Y. Freund and R. E. Schapire. Large margin classification using the Perceptron algorithm. *Machine Learning*, pages 277–296, 1999.

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

E. Grave, G. Obozinski, and F.R. Bach. Trace Lasso: a trace norm regularization for correlated designs. *Advances in Neural Information Processing Systems*, 24:2187–2195, 2011.

S.M. Kakade, S. Shalev-Shwartz, and A. Tewari. Regularization techniques for learning with matrices. *CoRR*, abs/0910.0610, 2009.

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

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

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

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

F. Orabona, N. Cesa-Bianchi, and C. Gentile. Beyond logarithmic bounds in online learning. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, pages 823–831. JMLR W&CP, 2012.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

S. Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, The Hebrew University, 2007.

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

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

S. Shalev-Shwartz and Y. Singer. A primal-dual perspective of online learning algorithms. *Machine Learning Journal*, 2007.

Y. Tsypkin. *Adaptation and Learning in Automatic Systems*. Academic Press, 1971.

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

M.K. Warmuth and A.K. Jagota. 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*, 1997.

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