General Algorithmic Frameworks for Online Problems

Yair Censor[1], Simeon Reich[2] and Alexander J. Zaslavski[2]

[1]Department of Mathematics, University of Haifa
Mt. Carmel, 31905 Haifa, Israel
(yair@math.haifa.ac.il)

[2]Department of Mathematics
The Technion – Israel Institute of Technology
32000 Haifa, Israel
({sreich, ajzasl}@techunix.technion.ac.il)

April 17, 2008

Abstract

We study general algorithmic frameworks for online learning tasks. These include binary classification, regression, multiclass problems and cost-sensitive multiclass classification. The theorems that we present give loss bounds on the behavior of our algorithms which depend on general conditions on the iterative step sizes.

Rrunning title: Online Algorithmic Frameworks.
Key-words: Online learning, general algorithms, classification, regression, multiclass, convex feasibility.

1 Introduction

Online learning algorithms for various prediction tasks differ fundamentally from batch learning algorithms. The online learning process assumes that the instances that need to be classified and their correct labels are not all available to the algorithm at the start of the training, but rather that they are unveiled sequentially. Moreover, the algorithm starts to offer predicted labels from its exposure to the first instance-label pair, and it subsequently learns and makes further predictions simultaneously. The theoretical aspect of online learning algorithms analysis is to provide tight bounds on their performance. Very often, these algorithms can be analyzed and shown to work quite well even when no statistical assumptions of any kind are made about the process producing the observed data. Many of the algorithms and methods of analysis used in this area can trace their roots to the work of Littlestone, Vovk and Warmuth, see [6, 7, 8]. Inspired and influenced by [4], we formulate general sets of conditions under which many algorithmic variants of online passive-aggressive algorithms can be analyzed. More precisely, we answer the following question: under what conditions on the choice of iterative steps can one obtain results analogous to those of [4]?

Thus, our work contributes to the development of new analytical frameworks that advance theoretical studies of practical learning methods. All the algorithms of [4] can be obtained as special cases of our algorithmic framework, but the framework is wide enough to encompass many more variants. Our way of looking at the subject will lead to additional developments of a similar nature. In particular, there are many links between online learning algorithms and projection algorithms for solving convex feasibility problems, see, e.g., [1, 2, 5, 3], which can lead to new studies of the latter that will concentrate on providing tight bounds on their performance as online algorithms, rather then on their asymptotic convergence.

We structured the paper so that the sections order follows closely that of [4], successively handling binary classification (Section 2), regression (Section 3), multiclass problems (Section 4) and cost-sensitive multiclass classification (Section 4).

2 Binary classification

We denote the *instance* presented to the algorithm on *round* t by $x^t \in R^n$, where R^n is the n-dimensional Euclidean space. We assume that x^t is associated with a unique *label* $y_t \in \{+1, -1\}$ and refer to each instance-label pair (x^t, y_t) as an *example*. The algorithms discussed in this paper make predictions using a *classification function*. We restrict our discussion to classification functions that are based on a vector of weights $w \in R^n$, and which take the form sign $\langle w, x \rangle$. We denote by w^t the weight vector used by the algorithm on round t, and refer to the expression $y_t \langle w^t, x^t \rangle$ as the *(signed) margin* attained on round t. Whenever sign $\langle w^t, x^t \rangle = y_t$ the algorithm has made a correct *prediction.* The *loss* is defined by the following hinge-loss function:

$$\ell(w; (x, y)) := \begin{cases} 0, & \text{if } y \langle w, x \rangle \geq 1, \\ 1 - y \langle w, x \rangle, & \text{otherwise,} \end{cases} \tag{2.1}$$

and, clearly,

$$\ell(w; (x, y)) = \max\{0, 1 - y \langle w, x \rangle\}. \tag{2.2}$$

We assume henceforth that for any number $c > 0$, $c/0 := +\infty$.

Algorithm 2.1 ***General Online Passive-Aggressive Algorithmic Framework for Binary Classification***

Initialization: *Set* $w^1 = (0, 0, \ldots, 0)$ *and choose parameters* γ_1 *and a sufficiently small* $\kappa > 0$ *such that*

$$0 < \gamma_1 < 2. \tag{2.3}$$

Iterative step: *(1) Given the weight* w^t *and receiving the instance* x^t*, predict:*

$$\hat{y}_t = \text{sign} \left\langle w^t, x^t \right\rangle. \tag{2.4}$$

(2) Receive the correct label $y_t \in \{+1, -1\}$ *and calculate the loss* $\ell_t = \ell(w^t; (x^t, y_t))$.

(3) Choose a nonnegative parameter τ_t *for which*

$$\tau_t \leq \gamma_1 \ell_t / \|x^t\|^2 \textit{ and if } \ell_t \geq 1 \textit{ then } \tau_t \geq \kappa. \tag{2.5}$$

(4) Update:

$$w^{t+1} = w^t + \tau_t y_t x^t. \tag{2.6}$$

We now turn to the analysis of our algorithmic framework. For any set E, denote by card(E) its cardinality. As before, we denote by ℓ_t the instantaneous loss suffered by Algorithm 2.1 on round t. In addition, we denote by $\widehat{\ell_t}$ the loss suffered by an arbitrary fixed predictor to which we are comparing our performance. Formally, let u be an arbitrary vector in R^n, and denote

$$\ell_t = \ell(w^t; (x^t, y_t)) \text{ and } \widehat{\ell_t} = \ell(u; (x^t, y_t)). \tag{2.7}$$

For any natural number t, define

$$\Delta_t := \|w^t - u\|^2 - \|w^{t+1} - u\|^2. \tag{2.8}$$

Lemma 2.2 *Let $\{(x^1, y_1), (x^2, y_2), \ldots, (x^T, y_T)\}$ be a sequence of examples, where $x^t \in R^n$ and $y_t \in \{+1, -1\}$ for all t. Let τ_t satisfy (2.5) for all t. Then*

$$\sum_{t=1}^{T} \tau_t \left(2\ell_t - \tau_t\|x^t\|^2 - 2\widehat{\ell}_t\right) \le \|u\|^2. \tag{2.9}$$

Proof. Clearly,

$$\sum_{t=1}^{T} \Delta_t = \sum_{t=1}^{T} \left(\|w^t - u\|^2 - \|w^{t+1} - u\|^2\right) = \|w^1 - u\|^2 - \|w^{T+1} - u\|^2, \tag{2.10}$$

and hence,

$$\sum_{t=1}^{T} \Delta_T \le \|u\|^2. \tag{2.11}$$

By (2.6) and (2.8) we have, for $t = 1, 2, \ldots, T$,

$$\begin{aligned}
\Delta_t &= \|w^t - u\|^2 - \|w^{t+1} - u\|^2 = \|w^t - u\|^2 - \|w^t - u + y_t\tau_t x^t\|^2 \\
&= \|w^t - u\|^2 - \left(\|w^t - u\|^2 + \left\langle 2\tau_t y_t(w^t - u), x^t\right\rangle + \tau_t^2\|x^t\|^2\right) \\
&= -2\tau_t y_t \left\langle w^t - u, x^t\right\rangle - \tau_t^2\|x^t\|^2.
\end{aligned} \tag{2.12}$$

By (2.1), (2.5), (2.6) and (2.8) we also have for $t = 1, 2, \ldots, T$, that

$$\text{if } \ell_t = 0, \text{ then } \tau_t = 0 \text{ and } \Delta_t = 0. \tag{2.13}$$

Assume that

$$t \in \{1, 2, \ldots, T\} \text{ and } \ell_t > 0. \tag{2.14}$$

Applying (2.1), we get

$$\ell_t = 1 - y_t \left\langle w^t, x^t\right\rangle \text{ and } \widehat{\ell}_t \ge 1 - y_t \left\langle u, x^t\right\rangle. \tag{2.15}$$

By (2.12) and (2.15),

$$\Delta_t \ge 2\tau_t((1 - \widehat{\ell}_t) - (1 - \ell_t)) - \tau_t^2\|x^t\|^2 = 2\tau_t(\ell_t - \widehat{\ell}_t) - \tau_t^2\|x^t\|^2, \tag{2.16}$$

which, in view of (2.11), (2.14) and (2.13), yields

$$\|u\|^2 \geq \sum_{t=1}^{T} \Delta_t \geq \sum_{t=1}^{T} \tau_t \left(2\ell_t - \tau_t \|x^t\|^2 - 2\widehat{\ell}_t\right), \tag{2.17}$$

proving the lemma. ■

Set

$$E_1 := \{t \in \{1, 2, \ldots, T\} \mid \ell_t \geq 1\}. \tag{2.18}$$

Theorem 2.3 *Let $\{(x^1, y_1), (x^2, y_2), \ldots, (x^T, y_T)\}$ be a sequence of examples, where $x^t \in R^n$ and $y_t \in \{+1, -1\}$ for all t and assume that τ_t satisfies (2.5) for all t. Assume that there exists a vector u such that $\widehat{\ell}_t = 0$ for all t. Then* $\mathrm{card}(E_1) \leq \kappa^{-1}(2-\gamma_1)^{-1}\|u\|^2$, *i.e., the number of indices $t \in \{1, 2, \ldots, T\}$ for which $\ell_t \geq 1$ does not exceed $\kappa^{-1}(2-\gamma_1)^{-1}\|u\|^2$.*

Proof. By Lemma 2.2, (2.9) holds. Since $\widehat{\ell}_t = 0$ for all t, (2.9) implies that

$$\sum_{t=1}^{T} \tau_t \left(2\ell_t - \tau_t \|x^t\|^2\right) \leq \|u\|^2 \tag{2.19}$$

and that $x^t \neq 0$ for all t. In view of (2.5) and (2.19),

$$\begin{aligned} \|u\|^2 &\geq \sum_{t=1}^{T} \|x^t\|^{-2} \left(2\ell_t\tau_t\|x^t\|^2 - \tau_t^2\|x^t\|^4\right) = \sum_{t=1}^{T} (2\ell_t\tau_t - \tau_t^2\|x^t\|^2) \\ &\geq \sum_{t=1}^{T} (2\ell_t\tau_t - \tau_t\gamma_1\ell_t) = \sum_{t=1}^{T} \tau_t\ell_t(2-\gamma_1). \end{aligned} \tag{2.20}$$

By (2.18), (2.20), (2.3) and (2.5),

$$\|u\|^2 \geq \sum_{t \in E_1} (2-\gamma_1)\tau_t\ell_t \geq \sum_{t \in E_1} (2-\gamma_1)\tau_t \geq \kappa(2-\gamma_1)\,\mathrm{card}(E_1) \tag{2.21}$$

and the required result follows. ■

Theorem 2.4 *Let $\{(x^1, y_1), (x^2, y_2), \ldots, (x^T, y_T)\}$ be a sequence of examples, where $x^t \in R^n$ and $y_t \in \{+1, -1\}$ for all t and assume that τ_t satisfies (2.5) for all t. Let $u \in R^n$ and assume that there is a number $c > 0$ such that $\tau_t \leq c$ for all t. Then*

$$\mathrm{card}(E_1) \leq \sum_{t \in E_1} \ell_t \leq \kappa^{-1}(2-\gamma_1)^{-1}(\|u\|^2 + 2\sum_{t=1}^{T} c\widehat{\ell}_t). \tag{2.22}$$

Proof. By Lemma 2.2, (2.9) holds and implies that

$$\sum_{t=1}^{T} \tau_t \left(2\ell_t - \tau_t \|x^t\|^2\right) \le \|u\|^2 + \sum_{t=1}^{T} 2\tau_t \widehat{\ell}_t. \tag{2.23}$$

Together with (2.5) this implies that

$$\begin{aligned} \|u\|^2 + \sum_{t=1}^{T} 2\tau_t \widehat{\ell}_t &\ge \sum_{t=1}^{T} (2\ell_t \tau_t - \tau_t \gamma_1 \ell_t) = \sum_{t=1}^{T} \tau_t \ell_t (2 - \gamma_1) \\ &\ge (2 - \gamma_1) \sum_{t \in E_1} \tau_t \ell_t \ge (2 - \gamma_1) \kappa \sum_{t \in E_1} \ell_t. \end{aligned} \tag{2.24}$$

Since $\tau_t \le c$ for all t, it follows from (2.24) that

$$\begin{aligned} \operatorname{card}(E_1) \le \sum_{t \in E_1} \ell_t &\le \kappa^{-1} (2 - \gamma_1)^{-1} (\|u\|^2 + \sum_{t=1}^{T} 2\tau_t \widehat{\ell}_t) \\ &\le \kappa^{-1} (2 - \gamma_1)^{-1} (\|u\|^2 + 2 \sum_{t=1}^{T} c \widehat{\ell}_t), \end{aligned} \tag{2.25}$$

which completes the proof. ■

2.1 Special cases

We show that all three variants ((PA), (PA-I) and (PA-II)) of the online passive-aggressive learning algorithm of Crammer et al. [4, Figure 1] are special cases of Algorithm 2.1 when the sequence $\{x^t\}$ is bounded. To see this, assume that there is an $r_0 > 0$ such that

$$\|x^t\| \le r_0 \quad \text{for all integers } t \ge 1. \tag{2.26}$$

Consider the algorithmic variant (PA) of [4, Figure 1] with $\tau_t = \ell_t \|x^t\|^{-2}$. Clearly, the first half of (2.5) holds with $\gamma_1 = 1$. We show that its second half holds with $\kappa = r_0^{-2}$. Assume that $\ell_t \ge 1$. By definition,

$$\tau_t \ge \|x^t\|^{-2} \ge r_0^{-2}. \tag{2.27}$$

Thus, (PA) is indeed a particular case of Algorithm 2.1. Consider now the algorithmic variant (PA-I) of [4, Figure 1] with $\tau_t = \min\{C,\ \ell_t \|x^t\|^{-2}\}$. (Here

C is a positive constant.) Clearly, the first half of (2.5) holds with $\gamma_1 = 1$. If $\ell_t \geq 1$, then

$$\tau_t \geq \min\{C, \|x^t\|^{-2}\} \geq \min\{C, r_0^{-2}\} \tag{2.28}$$

and the second half of (2.5) holds with $\kappa = \min\{C, r_0^{-2}\}$.

Next, consider the algorithm variant (PA-II) of [4, Figure 1] with

$$\tau_t = \ell_t(\|x^t\|^2 + (2C)^{-1})^{-1}. \tag{2.29}$$

Clearly, the first half of (2.5) holds with $\gamma_1 = 1$. If $\ell_t \geq 1$, then

$$\tau_t \geq (\|x^t\|^2 + (2C)^{-1})^{-1} \geq (r_0^2 + (2C)^{-1})^{-1} \tag{2.30}$$

and the second half of (2.5) holds with $\kappa = (r_0^2 + (2C)^{-1})^{-1}$.

3 Regression

Each instance x^t is associated with a real target value $y_t \in R$ which the online algorithm tries to predict. On every round, the algorithm receives an instance $x^t \in R^n$ and predicts a target value $\widehat{y}_t \in R$ using its interval regression function $\widehat{y}_t = \langle w^t, x^t \rangle$, where w^t is the incrementally-learned vector.

We use the ε-insensitive hinge-loss functions

$$\ell_\varepsilon(w; (x, y)) := \begin{cases} 0, & \text{if } |\langle w, x \rangle - y| \leq \varepsilon, \\ |\langle w, x \rangle - y| - \varepsilon, & \text{otherwise}, \end{cases} \tag{3.1}$$

where ε is a positive parameter.

Algorithm 3.1 ***General Online Passive-Aggressive Algorithmic Framework for Regression***

Initialization: *Fix $\varepsilon > 0$. Set $w^1 = (0, 0, \ldots, 0)$ and choose parameters γ_1 and a sufficiently small $\kappa > 0$ such that*

$$0 < \gamma_1 < 2. \tag{3.2}$$

Iterative step: *(1) Given the weight w^t and receiving the instance x^t, predict:*

$$\hat{y}_t = \langle w^t, x^t \rangle . \tag{3.3}$$

(2) Receive the correct label $y_t \in R$ and calculate the loss $\ell_t = \ell_\varepsilon(w^t; (x^t, y_t))$.

(3) Choose a nonnegative parameter τ_t for which

$$\tau_t \leq \gamma_1 \ell_t / \|x^t\|^2, \text{ and if } \ell_t \geq \varepsilon \text{ then } \tau_t \geq \kappa. \tag{3.4}$$

(4) Update:

$$w^{t+1} = w^t + \operatorname{sign}(y_t - \widehat{y}_t)\tau_t x^t. \tag{3.5}$$

Again we denote by $\widehat{\ell}_t$ the loss suffered by an arbitrary fixed predictor to which we are comparing our performance. Formally, let u be an arbitrary vector in R^n, and denote

$$\ell_t = \ell_\varepsilon(w^t; (x^t, y_t)) \text{ and } \widehat{\ell}_t = \ell_\varepsilon(u; (x^t, y_t)). \tag{3.6}$$

We also re-use the definition

$$\Delta_t := \|w^t - u\|^2 - \|w^{t+1} - u\|^2. \tag{3.7}$$

Lemma 3.2 *Let $\{(x^1, y_1), (x^2, y_2), \ldots, (x^T, y_T)\}$ be a sequence of examples, where $x^t \in R^n$ and $y_t \in R$ for all t. Let τ_t satisfy (3.4) for all t. Then*

$$\sum_{t=1}^{T} \tau_t \left(2\ell_t - \tau_t \|x^t\|^2 - 2\widehat{\ell}_t\right) \leq \|u\|^2. \tag{3.8}$$

Proof. By (3.7),

$$\sum_{t=1}^{T} \Delta_t = \|w^1 - u\|^2 - \|w^{T+1} - u\|^2 \leq \|u\|. \tag{3.9}$$

Let $t \in \{1, \ldots, T\}$. By both (3.7) and (3.5),

$$\begin{aligned}\Delta_t = {} & \|w^t - u\|^2 - \|w^t - u + \operatorname{sign}(y_t - \widehat{y}_t)\tau_t x^t\|^2 = \\ & - \operatorname{sign}(y_t - \widehat{y}_t) 2\tau_t (w_t - u) x^t - \tau_t^2 \|x^t\|^2. \end{aligned} \tag{3.10}$$

We now add and subtract the term $\operatorname{sign}(y_t - \widehat{y}_t)2\tau_t y_t$ from the right-hand side in the above equation to get the bound

$$\begin{aligned}\Delta_t \geq {} & - \operatorname{sign}(y_t - \widehat{y}_t) 2\tau_t (\langle w^t, x^t \rangle - y_t) \\ & + \operatorname{sign}(y_t - \widehat{y}_t) 2\tau_t (\langle u, x^t \rangle - y_t) - \tau_t^2 \|x^t\|^2. \end{aligned} \tag{3.11}$$

From (3.3) we obtain

$$-\operatorname{sign}(y_t - \widehat{y}_t)(\langle w^t, x^t\rangle - y_t) = \left|\langle w^t, x^t\rangle - y_t\right|. \tag{3.12}$$

Assume that $\ell_t \neq 0$. We then get, by (3.1),

$$\ell_t = \left|\langle w^t, x^t\rangle - y_t\right| - \varepsilon. \tag{3.13}$$

By (3.11), (3.12), (3.13), (3.6) and (3.1),

$$\begin{aligned}\Delta_t &\geq 2\tau_t(\ell_t + \varepsilon) + \operatorname{sign}(y_t - \widehat{y}_t)2\tau_t(\langle u, x^t\rangle - y_t) - \tau_t^2\|x^t\|^2 \\ &\geq 2\tau_t(\ell_t + \varepsilon) - 2\tau_t(\widehat{\ell}_t + \varepsilon) - \tau_t^2\|x^t\|^2 = \tau_t(2\ell_t - \tau_t\|x^t\|^2 - 2\widehat{\ell}_t). \end{aligned}\tag{3.14}$$

When combined with (3.9), this implies that

$$\sum_{t=1}^{T} \tau_t(2\ell_t - \tau_t\|x^t\|^2 - 2\widehat{\ell}_t) \leq \sum_{t=1}^{T} \Delta_t \leq \|u\|^2, \tag{3.15}$$

which completes the proof. ■

Now set

$$E_\varepsilon := \{t \in \{1, 2, \ldots, T\} \mid \ \ell_t \geq \varepsilon\}. \tag{3.16}$$

Theorem 3.3 *Let $\{(x^1, y_1), (x^2, y_2), \ldots, (x^T, y_T)\}$ be a sequence of examples, where $x^t \in R^n$ and $y_t \in R$ for all t, let the nonnegative parameters τ_t satisfy (3.4) for all t, and let $\varepsilon > 0$ be fixed. Assume that there exists a vector u such that $\widehat{\ell}_t = 0$ for all t. Then* $\operatorname{card}(E_\varepsilon) \leq \|u\|^2(\varepsilon(2 - \gamma_1)\kappa)^{-1}$.

Proof. From Lemma 3.2 we have (3.8) which together with (3.4) gives

$$\|u\|^2 \geq \sum_{t=1}^{T}(2\ell_t\tau_t - \tau_t\gamma_1\ell_t) = \sum_{t=1}^{T} \tau_t\ell_t(2 - \gamma_1). \tag{3.17}$$

This yields, in view of (3.2), (3.16) and (3.4),

$$\|u\|^2 \geq \sum_{t=1}^{T} \ell_t(2 - \gamma_1)\tau_t \geq \sum_{t \in E_\varepsilon} \varepsilon(2 - \gamma_1)\kappa. \tag{3.18}$$

Hence,

$$\operatorname{card}(E_\varepsilon) \leq \|u\|^2(\varepsilon(2 - \gamma_1)\kappa)^{-1},$$

as asserted. ■

Theorem 3.4 *Let $\{(x^1, y_1), (x^2, y_2), \ldots, (x^T, y_T)\}$ be a sequence of examples, where $x^t \in R^n$ and $y_t \in R$ for all t, let the nonnegative parameters τ_t satisfy (3.4) for all t, let $\varepsilon > 0$ be fixed and let $u \in R^n$. Assume that there is a number $c > 0$ such that $\tau_t \le c$ for all t. Then*

$$\varepsilon \,\mathrm{card}(E_\varepsilon) \le \sum_{t \in E_\varepsilon} \ell_t \le ((2-\gamma_1)\kappa)^{-1})(\|u\|^2 + \sum_{t=1}^{T} 2c\widehat{\ell_t}). \tag{3.19}$$

Proof. From Lemma 3.2 we have (3.8) which, together with (3.4), implies that

$$\begin{aligned} \|u\|^2 + \sum_{t=1}^{T} 2\tau_t \widehat{\ell_t} &\ge \sum_{t=1}^{T} \tau_t (2\ell_t - \gamma_1 \ell_t) = (2-\gamma_1) \sum_{t=1}^{T} \tau_t \ell_t \\ &\ge (2-\gamma_1)\kappa \sum_{t \in E_\varepsilon} \ell_t. \end{aligned} \tag{3.20}$$

The last inequality leads to

$$\begin{aligned} \varepsilon \,\mathrm{card}(E_\varepsilon) \le \sum_{t \in E_\varepsilon} \ell_t &\le ((2-\gamma_1)\kappa)^{-1}(\|u\|^2 + \sum_{t=1}^{T} 2\tau_t \widehat{\ell_t}) \\ &\le ((2-\gamma_1)\kappa)^{-1}(\|u\|^2 \sum_{t=1}^{T} 2c\widehat{\ell_t}), \end{aligned} \tag{3.21}$$

which completes the proof. ■

4 Multiclass problems

In multiclass problems every instance x^t is associated with a set of labels Y_t. Denoting by $Y := \{1, 2, \ldots, k\}$ the set of all possible labels, Y_t is a subset of Y. We say that $y \in Y$ is *relevant* for the instance x^t if $y \in Y_t$. The online algorithm receives instances $x^1, x^2, \ldots$ sequentially, where x^t belongs to an *instance space* X.

Assume that we are provided with a set of functions $\phi_1, \phi_2, \ldots, \phi_d : X \times Y \to R$ and $\phi = (\phi_1, \phi_2, \ldots, \phi_d)$. On round t, the prediction of the algorithm is the k-dimensional vector

$$(\langle w^t, \phi(x^t, 1)\rangle, \langle w^t, \phi(x^t, 2)\rangle, \ldots, \langle w^t, \phi(x^t, k)\rangle). \tag{4.1}$$

We define the *margin* attained by the algorithm on round t for the *example* (x^t, Y_t) by

$$\gamma(w^t;(x^t,Y_t)) := \min\{\langle w^t, \phi(x^t,r)\rangle \mid r \in Y_t\} - \max\{\langle w^t, \phi(x^t,s)\rangle \mid s \notin Y_t\}. \tag{4.2}$$

The instantaneous loss suffered after receiving Y_t is defined by the following hinge-loss function:

$$\ell_{MC}(w;(x,Y)) := \begin{cases} 0, & \text{if } \gamma(w;(x,Y)) \geq 1, \\ 1-\gamma(w;(x,Y)), & \text{otherwise,} \end{cases} \tag{4.3}$$

and we define

$$\ell_t = \ell_{MC}(w^t;(x^t,Y_t)) \text{ and } \widehat{\ell}_t = \ell_{MC}(u;(x^t,Y_t)), \tag{4.4}$$

where $u \in R^n$.

Algorithm 4.1 ***General Online Passive-Aggressive Algorithmic Framework for Multiclass Classification***

Initialization: *Set* $w^1 = (0,0,\ldots,0)$ *and choose parameters* γ_1, γ_2 *and a sufficiently small* $\kappa > 0$ *such that*

$$0 < \gamma_1 < 2, \ \ \gamma_2 \in (0,1]. \tag{4.5}$$

Iterative step: *(1) Given the weight* w^t *and receiving the instance* x^t*, predict the associated set of labels* $\hat{Y}_t$*.*

(2) Receive the correct associated set of labels Y_t *and calculate the loss* $\ell_t = \ell_{MC}(w^t;(x^t,Y_t))$*.*

(3) Calculate

$$\begin{aligned} r_t &:= \mathit{argmin}\{\langle w^t, \phi(x^t,r)\rangle \mid r \in Y_t\}, \\ s_t &:= \mathit{argmax}\{\langle w^t, \phi(x^t,s)\rangle \mid s \notin Y_t\}. \end{aligned} \tag{4.6}$$

(4) Choose a nonnegative parameter τ_t *such that* $\tau_t = 0$ *if* $l_t = 0$*; otherwise*

$$\tau_t \leq \gamma_1 \ell_t / \left\|\phi(x^t,r_t) - \phi(x^t,s_t)\right\|^2, \ \textit{and if } \ell_t \geq \gamma_2 \textit{ then } \tau_t \geq \kappa. \tag{4.7}$$

(5) Update:

$$w^{t+1} = w^t + \tau_t\left(\phi(x^t,r_t) - \phi(x^t,s_t)\right). \tag{4.8}$$

Again, it can be shown that the three algorithmic variants that appear in [4, Section 7] are particular cases of our Algorithm 4.1 if there is an m_0 such that $\|\phi(x^t, r_t) - \phi(x^t, s_t)\|^2 \leq m_0$ for all t.

Lemma 4.2 *Let $\{(x^1, Y_1), (x^2, Y_2), \ldots, (x^T, Y_T)\}$ be a sequence of examples with $x^t \in R^n$, $Y_t \subseteq \{1, 2, \ldots, k\}$, let $w^1 = (0, 0, \ldots, 0)$, and let $u \in R^n$. Then*

$$\sum_{t=1}^{T} \tau_t \left(2\ell_t - 2\widehat{\ell}_t - \tau_t \left\|\phi(x^t, r_t) - \phi(x^t, s_t)\right\|^2\right) \leq \|u\|^2. \tag{4.9}$$

Proof. Set again

$$\Delta_t := \|w^t - u\|^2 - \|w^{t+1} - u\|^2 \tag{4.10}$$

for all t. Then

$$\sum_{t=1}^{T} \Delta_T = \|w^1 - u\|^2 - \|w^{T+1} - u\|^2 \leq \|u\|^2. \tag{4.11}$$

For $t = 1, 2, \ldots, T$ with $\ell_t > 0$ it follows from (4.10) and (4.8) that

$$\begin{aligned} \Delta_t &= \|w^t - u\|^2 - \|w^t - u + \tau_t \left(\phi(x^t, r_t) - \phi(x^t, s_t)\right)\|^2 \\ &= -2\tau_t \left\langle w^t - u, \phi(x^t, r_t) - \phi(x^t, s_t)\right\rangle - \tau_t^2 \left\|\phi(x^t, r_t) - \phi(x^t, s_t)\right\|^2. \end{aligned} \tag{4.12}$$

Assume that $t \in \{1, 2, \ldots, T\}$ and the loss $\ell_t > 0$. By (4.3),

$$\ell_t = 1 - \gamma(w^t; (x^t, Y_t)) \text{ and } \widehat{\ell}_t \geq 1 - \gamma(u; (x^t, Y_t)). \tag{4.13}$$

Together with (4.2) and (4.6) this implies that

$$\begin{aligned} (1 - \widehat{\ell}_t) - (1 - \ell_t) &\leq \gamma(u; (x^t, Y_t)) - \gamma(w^t; (x^t, Y_t)) \\ &= \gamma(u; (x^t, Y_t)) - \left(\left\langle w^t, \phi(x^t, r_t)\right\rangle - \left\langle w^t, \phi(x^t, s_t)\right\rangle\right) \\ &\leq \left\langle u, \phi(x^t, r_t)\right\rangle - \left\langle u, \phi(x^t, s_t)\right\rangle \\ &\quad - \left(\left\langle w^t, \phi(x^t, r_t)\right\rangle - \left\langle w^t, \phi(x^t, s_t)\right\rangle\right) \\ &= \left\langle u - w^t, \phi(x^t, r_t) - \phi(x^t, s_t)\right\rangle. \end{aligned} \tag{4.14}$$

By (4.12) and (4.14),

$$\Delta_t \geq 2\tau_t(\ell_t - \widehat{\ell}_t) - \tau_t^2 \left\|\phi(x^t, r_t) - \phi(x^t, s_t)\right\|^2. \tag{4.15}$$

Together with (4.11) this implies (4.9) as asserted. ■

Recall that

$$E_{\gamma_2} := \{t \in \{1,2,\ldots,T\} \mid \ \ell_t \geq \gamma_2\}. \tag{4.16}$$

Theorem 4.3 *Let* $\{(x^1, Y_1), (x^2, Y_2), \ldots, (x^T, Y_T)\}$ *be a sequence of examples with* $x^t \in R^n$*,* $Y_t \subseteq \{1, 2, \ldots, k\}$*, let* $w^1 = (0, 0, \ldots, 0)$*, and let* $u \in R^n$ *be such that* $\widehat{\ell}(u; (x^t, Y_t)) = 0$ *for all* t*. Then* $\mathrm{card}(E_{\gamma_2}) \leq \|u\|^2(\kappa(2-\gamma_1)\gamma_2)^{-1}$.

Proof. It follows from (4.9) and (4.7) that, for $t = 1, 2, \ldots, T$,

$$2\ell_t - \tau_t \left\|\phi(x^t, r_t) - \phi(x^t, s_t)\right\|^2 \geq (2-\gamma_1)\ell_t. \tag{4.17}$$

Together with (4.9) this implies that

$$\|u\|^2 \geq \sum_{t=1}^{T} \tau_t(2-\gamma_1)\ell_t. \tag{4.18}$$

By (4.18) and (4.7),

$$\|u\|^2 \geq \sum_{t \in E_{\gamma_2}} \tau_t(2-\gamma_1)\gamma_2 \geq \kappa(2-\gamma_1)\gamma_2 \,\mathrm{card}(E_{\gamma_2}), \tag{4.19}$$

and the theorem follows. ■

Theorem 4.4 *Let* $\{(x^1, Y_1), (x^2, Y_2), \ldots, (x^T, Y_T)\}$ *be a sequence of examples with* $x^t \in R^n$*,* $Y_t \subseteq \{1, 2, \ldots, k\}$ *and let* $u \in R^n$*. Assume that there is a number* $c > 0$ *such that* $\tau_t \leq c$ *for all* t*. Then*

$$\gamma_2 \,\mathrm{card}(E_{\gamma_2}) \leq \sum_{t \in E_{\gamma_2}} \ell_t \leq ((2-\gamma_1)\kappa)^{-1}(\|u\|^2 + 2c\sum_{t=1}^{T} \widehat{\ell}_t). \tag{4.20}$$

Proof. By (4.9),

$$\sum_{t=1}^{T} \tau_t \left(2\ell_t - \tau_t \left\|\phi(x^t, r_t) - \phi(x^t, s_t)\right\|^2\right) \leq \|u\|^2 + \sum_{t=1}^{T} 2\tau_t\widehat{\ell}_t. \tag{4.21}$$

Together with (4.7) this implies that

$$\|u\|^2 + \sum_{t=1}^{T} 2\tau_t\widehat{\ell}_t \geq \sum_{t=1}^{T} \tau_t\left(2\ell_t - \gamma\ell_t\right) = (2-\gamma_1)\sum_{t=1}^{T} \tau_t\ell_t \geq (2-\gamma_1)\kappa \sum_{t \in E_{\gamma_2}} \ell_t. \tag{4.22}$$

This implies that

$$\gamma_2 \operatorname{card}(E_{\gamma_2}) \leq \sum_{t \in E_{\gamma_2}} \ell_t \leq ((2-\gamma_1)\kappa)^{-1}(\|u\|^2 + \sum_{t=1}^{T} 2\tau_t \widehat{\ell}_t)$$

$$\leq ((2-\gamma_1)\kappa)^{-1}(\|u\|^2 + 2c \sum_{t=1}^{T} \widehat{\ell}_t) \tag{4.23}$$

and the result follows. ■

5 Cost-sensitive multiclass classification

With Y and ϕ as in Section 4, in cost-sensitive multiclass classification each instance x^t is associated with a single label $y_t \in Y$ and the prediction extended by the online algorithm is simply

$$\widehat{y}_t = \operatorname{argmax}\{\langle w^t, \phi(x^t, y)\rangle \mid y \in Y\}. \tag{5.1}$$

A prediction error occurs if $y_t \neq \widehat{y}_t$. More specifically, for every pair of labels $(y, \bar{y})$ there is a cost $\rho(y, \bar{y})$. We assume that $\rho(y, y) = 0$ for all $y \in Y$ and that $\rho(y, \bar{y}) > 0$ whenever $y \neq \bar{y}$. The goal is to minimize $\sum_{t=1}^{T} \rho(y_t, \widehat{y}_t)$. Define the cost sensitivity loss

$$\ell_{PB}(w; (x, y)) := \langle w, \phi(x, \widehat{y})\rangle - \langle w, \phi(x, y)\rangle + \rho(y, \widehat{y})^{1/2}. \tag{5.2}$$

Algorithm 5.1 ***General Online Passive-Aggressive Algorithmic Framework for Cost-Sensitive Multiclass Classification***

Initialization: *Set* $w^1 = (0, 0, \ldots, 0)$ *and choose parameters* γ_1, γ_2 *and a sufficiently small* $\kappa > 0$ *such that*

$$0 < \gamma_1 < 2, \ \gamma_2 \in (0, 1]. \tag{5.3}$$

Iterative step: *(1) Given the weight* w^t *and receiving the instance* x^t, *predict the label* $\widehat{y}_t$.

(2) Receive the correct label y_t *and calculate the loss* $\ell_t := \ell_{PB}(w^t; (x^t, y_t))$.

(3) Choose a nonnegative parameter τ_t *such that if* $\ell_t = 0$, *then* $\tau_t = 0$; *otherwise*

$$\tau_t \leq \gamma_1 \ell_t / \left\|\phi(x^t, y_t) - \phi(x^t, \widehat{y}_t)\right\|^2, \ \textit{and if } \ell_t \geq \gamma_2 \textit{ then } \tau_t \geq \kappa. \tag{5.4}$$

(4) Update:

$$w^{t+1} = w^t + \tau_t \left(\phi(x^t, y_t) - \phi(x^t, \widehat{y}_t)\right). \tag{5.5}$$

Again, it can be shown that the three algorithmic variants that appear in [4, Section 8] are particular cases of our Algorithm 5.1 if there is a constant m_0 such that $\|\phi(x^t, y_t) - \phi(x^t, \widehat{y}_t)\| \le m_0$ for all t.

To analyze Algorithm 5.1, let $\tilde{y} = \tilde{y}(w; x, y) \in Y$ be defined, for any given w, x and y, by

$$\begin{aligned} \tilde{y} &:= \ \operatorname{argmax}\{\langle w, \phi(x, r)\rangle - \langle w, \phi(x, y)\rangle + \rho(y, r)^{1/2} \mid r \in Y\}, \\ \tilde{y}_t &:= \tilde{y}(w^t; x^t, y_t). \end{aligned} \tag{5.6}$$

Define the loss for the max-loss update by

$$\ell_{ML}(w; (x, y)) := \langle w, \phi(x, \tilde{y})\rangle - \langle w, \phi(x, y)\rangle + \rho(y, \tilde{y})^{1/2}. \tag{5.7}$$

By (5.2), (5.6) and (5.7),

$$\ell_{PB}(w^t; (x^t, y_t)) \le \ell_{ML}(w^t; (x^t, y_t)). \tag{5.8}$$

Lemma 5.2 *Let $\{(x^1, y_1), (x^2, y_2), \ldots, (x^T, y_T)\}$ be a sequence of examples, where $x^t \in R^n$ and $y_t \in Y$ for all t. Let u be an arbitrary vector in R^n. If $\tau_t \ge 0$ satisfies (5.4), then, for any sequence $\{w^t\}$ generated by Algorithm 5.1,*

$$\sum_{t=1}^{T} [\tau_t (2\ell_{PB}(w^t; (x^t, y_t))) - \tau_t^2 \left\|\phi(x^t, y_t) - \phi(x^t, \widehat{y}_t)\right\|^2 - 2\ell_{ML}(u; (x^t, y_t))\tau_t] \le \|u\|^2. \tag{5.9}$$

Under the same conditions, if in Algorithm 5.1 $\widehat{y}_t$ is replaced by $\tilde{y}_t$, then

$$\sum_{t=1}^{T} [\tau_t (2\ell_{ML}(w^t; (x^t, y_t))) - \tau_t^2 \left\|\phi(x^t, y_t) - \phi(x^t, \tilde{y}_t)\right\|^2 - 2\ell_{ML}(u; (x^t, y_t))\tau_t] \le \|u\|^2. \tag{5.10}$$

Proof. As usual, set

$$\Delta_t := \|w^t - u\|^2 - \|w^{t+1} - u\|^2. \tag{5.11}$$

Then

$$\sum_{t=1}^{T} \Delta_t \leq ||w^1 - u||^2 - ||w^{T+1} - u||^2 \leq \|u\|^2. \tag{5.12}$$

Let $t \in \{1, 2, \ldots, T\}$ with

$$\ell_{PB}(w^t; (x^t, y_t)) > 0. \tag{5.13}$$

Then, by (5.11) and (5.5),

$$\begin{aligned}\Delta_t &= \|w^t - u\|^2 - \|w^t - u + \tau_t \left(\phi(x^t, y_t) - \phi(x^t, \widehat{y}_t)\right)\|^2 \\ &= -2\tau_t \left\langle w^t - u, \phi(x^t, y_t) - \phi(x^t, \widehat{y}_t)\right\rangle - \tau_t^2 \|\phi(x^t, y_t) - \phi(x^t, \widehat{y}_t)\|^2. \end{aligned} \tag{5.14}$$

By definition (see (5.6) and (5.7)),

$$\ell_{ML}(u; (x^t, y_t)) = \max\{\left\langle u, \phi(x^t, r) - \phi(x^t, y_t)\right\rangle + \rho(y_t, r)^{1/2} \mid r \in Y\}. \tag{5.15}$$

Therefore,

$$\widehat{\ell}_t := \ell_{ML}(u; (x^t, y_t)) \geq \left\langle u, \phi(x^t, \widehat{y}_t) - \phi(x^t, y_t)\right\rangle + \rho(y_t, \widehat{y}_t)^{1/2}. \tag{5.16}$$

By (5.14) and (5.16),

$$\begin{aligned}\Delta_t \geq -2\tau_t \left\langle w^t, \phi(x^t, y_t) - \phi(x^t, \widehat{y}_t)\right\rangle + 2\tau_t(\rho(y_t, \widehat{y}_t)^{1/2} \\ - \ell_{ML}(u; (x^t, y_t))) - \tau_t^2 \|\phi(x^t, y_t) - \phi(x^t, \widehat{y}_t)\|^2. \end{aligned} \tag{5.17}$$

By the definition of ℓ_{PB} (see (5.2)),

$$\left\langle w^t, \phi(x^t, y_t) - \phi(x^t, \widehat{y}_t)\right\rangle = \rho(y_t, \widehat{y}_t)^{1/2} - \ell_{PB}(w^t; (x^t, y_t)). \tag{5.18}$$

By (5.17) and (5.18),

$$\begin{aligned}\Delta_t &\geq -2\tau_t(\rho(y_t, \widehat{y}_t)^{1/2} - \ell_{PB}(w^t; (x^t, y_t))) + 2\tau_t(\rho(y_t, \widehat{y}_t)^{1/2} \\ &- \ell_{ML}(u; (x^t, y_t))) - \tau_t^2 \|\phi(x^t, y_t) - \phi(x^t, \widehat{y}_t)\|^2 \\ &= \tau_t(2\ell_{PB}(w^t; (x^t, y_t))) - \tau_t^2 \|\phi(x^t, y_t) - \phi(x^t, \widehat{y}_t)\|^2 - 2\ell_{ML}(u; (x^t, y_t))\tau_t. \end{aligned} \tag{5.19}$$

Relations (5.19) and (5.12) show that

$$\begin{aligned}\|u\|^2 \geq \sum_{t=1}^{T} \Delta_t \geq \sum_{t=1}^{T}[\tau_t(2\ell_{PB}(w^t; (x^t, y_t)) - \tau_t^2 \|\phi(x^t, y_t) - \phi(x^t, \widehat{y}_t)\|^2 \\ - 2\ell_{ML}(u; (x^t, y_t))\tau_t], \end{aligned} \tag{5.20}$$

which proves the first case of the lemma. Considering the second case, where in Algorithm 5.1 $\widehat{y}_t$ is replaced by $\tilde{y}_t$, we define Δ_t again by (5.11). Clearly,

$$\sum_{t=1}^{T} \Delta_t \leq \|u\|^2. \tag{5.21}$$

Let $t \in \{1, 2, \ldots, T\}$ with

$$\ell_{PB}(w^t; (x^t, y_t)) > 0. \tag{5.22}$$

As in (5.14) we can show that

$$\Delta_t = -2\tau_t \left\langle w^t - u, \phi(x^t, y_t) - \phi(x^t, \tilde{y}_t) \right\rangle - \tau_t^2 \|\phi(x^t, y_t) - \phi(x^t, \tilde{y}_t)\|^2. \tag{5.23}$$

By definition (see (5.6) and (5.7)),

$$\widehat{\ell}_t := \ell_{ML}(u; (x^t, y_t)) \geq \left\langle u, \phi(x^t, \tilde{y}_t) - \phi(x^t, y_t) \right\rangle + \rho(y_t, \tilde{y}_t)^{1/2}. \tag{5.24}$$

By (5.24) and (5.23),

$$\begin{aligned} \Delta_t \geq{}& -2\tau_t \left\langle w^t, \phi(x^t, y_t) - \phi(x^t, \tilde{y}_t) \right\rangle + 2\tau_t (\rho(y_t, \tilde{y}_t)^{1/2} \\ &- \ell_{ML}(u; (x^t, y_t))) - \tau_t^2 \|\phi(x^t, y_t) - \phi(x^t, \tilde{y}_t)\|^2). \end{aligned} \tag{5.25}$$

Again, by definition (see (5.6) and (5.7)),

$$\left\langle w^t, \phi(x^t, y_t) - \phi(x^t, \tilde{y}_t) \right\rangle = \rho(y_t, \tilde{y}_t) - \ell_{ML}(w^t; (x^t, y_t)). \tag{5.26}$$

By (5.25) and (5.26),

$$\begin{aligned} \Delta_t \geq{}& -2\tau_t(\rho(y_t, \tilde{y}_t)^{1/2} - \ell_{ML}(w^t; (x^t, y_t))) + 2\tau_t(\rho(y_t, \tilde{y}_t)^{1/2} \\ &- \ell_{ML}(u; (x^t, y_t))) - \tau_t^2 \|\phi(x^t, y_t) - \phi(x^t, \tilde{y}_t)\|^2 \\ ={}& \tau_t(2\ell_{ML}(w^t; (x^t, y_t)) - 2\ell_{ML}(u; (x^t, y_t))) - \tau_t^2 \|\phi(x^t, y_t) - \phi(x^t, \tilde{y}_t)\|^2. \end{aligned} \tag{5.27}$$

Together with (5.21) this implies that

$$\begin{aligned} \|u\|^2 \geq \sum_{t=1}^{T} \Delta_t \geq \sum_{t=1}^{T} [&\tau_t(2\ell_{ML}(w^t; (x^t, y_t)) - 2\ell_{ML}(u; (x^t, y_t))) \\ &- \tau_t^2 \|\phi(x^t, y_t) - \phi(x^t, \tilde{y}_t)\|^2], \end{aligned} \tag{5.28}$$

completing the proof. ■

Consider the set

$$E_{\gamma_2^2} := \{t \in \{1, 2, \ldots, T\} \mid \rho(y_t, \widehat{y}_t) \geq \gamma_2^2\}. \tag{5.29}$$

Theorem 5.3 *Let* $\{(x^1, y_1), (x^2, y_2), \ldots, (x^T, y_T)\}$ *be a sequence of examples, where* $x^t \in R^n$ *and* $y_t \in R$ *for all* t, *and let* u *be an arbitrary vector in* R^n. *Assume that*

$$\ell_{ML}(u; (x^t, y_t)) = 0 \tag{5.30}$$

for all t. *Then*

$$\operatorname{card}(E_{\gamma_2^2}) \leq (\kappa\gamma_2(2-\gamma_1))^{-1}\|u\|^2. \tag{5.31}$$

Proof. By (5.1), (5.4), (5.6) and Lemma 5.2,

$$\begin{aligned}\|u\|^2 &\geq \sum_{t=1}^{T} \tau_t(2\ell_{PB}(w^t; (x^t, y_t)) - \tau_t\|\phi(x^t, y_t) - \phi(x^t, \widehat{y}_t)\|^2) \\ &\geq \sum_{t=1}^{T} \tau_t(2-\gamma_1)\ell_{PB}(w^t; (x^t, y_t)) \geq \sum_{t=1}^{T} \tau_t(2-\gamma_1)\rho(y_t, \widehat{y}_t)^{1/2}.\end{aligned} \tag{5.32}$$

This and (5.4) yield

$$\|u\|^2 \geq \sum_{t \in E_{\gamma_2^2}} \tau_t(2-\gamma_1)\gamma_2 \geq \kappa\gamma_2(2-\gamma_1)\operatorname{card}(E_{\gamma_2^2}), \tag{5.33}$$

proving the theorem. ■

Theorem 5.4 *Let* $\{(x^1, y_1), (x^2, y_2), \ldots, (x^T, y_T)\}$ *be a sequence of examples, where* $x^t \in R^n$ *and* $y_t \in R$ *for all* t, *and let* u *be an arbitrary vector in* R^n. *Assume that there exists a number* $c > 0$ *such that* $\tau_t \leq c$ *for all* t. *Then*

$$\gamma_2 \operatorname{card}(E_{\gamma_2^2}) \leq \sum_{t \in E_{\gamma_2^2}} \rho(y_t, \widehat{y}_t)^{1/2} \leq ((2-\gamma_1)\kappa)^{-1}(\|u\|^2 + \sum_{t=1}^{T} 2c\ell_{ML}(u; (x^t, y_t))). \tag{5.34}$$

Proof. Inequality (5.9), when combined with (5.4), implies that

$$\begin{aligned}\|u\|^2 + \sum_{t=1}^{T} 2\ell_{ML}(u; (x^t, y_t))\tau_t &\geq \sum_{t=1}^{T} \tau_t(2\ell_{PB}(w^t; (x^t, y_t)) \\ -\gamma_1\ell_{PB}(w^t; (x^t, y_t))) &\geq (2-\gamma_1)\sum_{t=1}^{T} \tau_t\ell_{PB}(w^t; (x^t, y_t)) \\ &\geq (2-\gamma_1)\sum_{t=1}^{T} \tau_t\rho(y_t, \widehat{y}_t)^{1/2}.\end{aligned} \tag{5.35}$$

This inequality, (5.4), (5.1) and (5.2) show that

$$\begin{aligned}\gamma_2 \operatorname{card}(E_{\gamma_2^2}) &\leq \sum_{t\in E_{\gamma_2^2}} \rho(y_t,\widehat{y}_t)^{1/2} \leq (2-\gamma_1)^{-1}(2-\gamma_1)\kappa^{-1} \sum_{t\in E_{\gamma_2^2}} \tau_t\rho(y_t,\widehat{y}_t)^{1/2} \\ &\leq ((2-\gamma_1)\kappa)^{-1}(\|u\|^2 + \sum_{t=1}^{T} 2\ell_{ML}(u;(x^t,y_t))\tau_t) \\ &\leq ((2-\gamma_1)\kappa)^{-1}(\|u\|^2 + \sum_{t=1}^{T} 2\ell_{ML}(u;(x^t,y_t))c), \qquad (5.36)\end{aligned}$$

concluding the proof. ■

Acknowledgments. We gratefully acknowledge illuminating discussions with Shai Shalev-Shwartz and Yoram Singer. This work was partially supported by grant No. 2003275 of the United States-Israel Binational Science Foundation (BSF) and by a National Institutes of Health (NIH) grant No. HL70472. The second author was partially supported by the Fund for the Promotion of Research at the Technion and by the Technion VPR Fund - B. and G. Greenberg Research Fund (Ottawa).

References

[1] H.H. Bauschke and J.M. Borwein, On projection algorithms for solving convex feasibility problems, *SIAM Review* **38** (1996), 367–426.

[2] D. Butnariu, R. Davidi, G.T. Herman and I.G. Kazantsev, Stable convergence behavior under summable perturbations of a class of projection methods for convex feasibility and optimization problems, *IEEE Journal of Selected Topics in Signal Processing* **1** (2007), 540–547.

[3] Y. Censor and S.A. Zenios, *Parallel Optimization: Theory, Algorithms, and Applications*, Oxford University Press, New York, NY, USA, 1997.

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

[5] M. Jiang and G. Wang, Convergence studies on iterative algorithms for image reconstruction, *IEEE Transactions on Medical Imaging* **22** (2003), 569–579.

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

[7] N. Littlestone and M.K. Warmuth, The weighted majority algorithm, *Information and Computation* **108** (1994), 212–261.

[8] V.G. Vovk, Aggregating strategies, in: *Proceedings of the Third Annual Workshop on Computational Learning Theory*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990, 371–383.