Problems of Information Transmission, Vol. 41, No. 4, 2005, pp. 368–384. Translated from Problemy Peredachi Informatsii, No. 4, 2005, pp. 78–96.

METHODS OF SIGNAL PROCESSING

Recursive Aggregation of Estimators by the Mirror Descent Algorithm with Averaging[1]

A. B. Juditsky[⋆], A. V. Nazin[⋆⋆][2], A. B. Tsybakov[⋆⋆⋆], and N. Vayatis[⋆⋆⋆]

[⋆] *Laboratoire de Modélisation et Calcul, Université Grenoble I, France*
anatoli.iouditski@imag.fr
[⋆⋆] *Institute of Control Sciences, RAS, Moscow*
nazine@ipu.rssi.ru
[⋆⋆⋆] *Laboratoire de Probabilités et Modèles Aléatoires, Université Paris VI, France*
tsybakov@ccr.jussieu.fr vayatis@ccr.jussieu.fr

Received March 16, 2005; in final form, July 26, 2005

Abstract—We consider a recursive algorithm to construct an aggregated estimator from a finite number of base decision rules in the classification problem. The estimator approximately minimizes a convex risk functional under the ℓ_1-constraint. It is defined by a stochastic version of the mirror descent algorithm which performs descent of the gradient type in the dual space with an additional averaging. The main result of the paper is an upper bound for the expected accuracy of the proposed estimator. This bound is of the order $C\sqrt{(\log M)/t}$ with an explicit and small constant factor C, where M is the dimension of the problem and t stands for the sample size. A similar bound is proved for a more general setting, which covers, in particular, the regression model with squared loss.

1. INTRODUCTION

The methods of generalized portrait (i.e., support vector machines, SVM) and boosting recently became widely used in classification practice (see, e.g., [1–4]). These methods are based on minimization of a convex empirical risk functional with a penalty. Their statistical analysis is given, for instance, in papers [5–8] (see also references therein). Note that the provided analysis is only approximate since numerical boosting and SVM algorithms do not necessarily minimize the empirical risk functional exactly. Moreover, it is assumed that the whole data sample is available, but often it is interesting to consider the on-line setting where observations come one-by-one and recursive methods need to be implemented.

There exists an extensive literature on recursive classification starting from Perceptron and its various modifications (see, e.g., [9–11] and references therein, as well as overviews in [12, 13]). We mention here only methods which use the same loss functions as boosting and SVM, and which may thus be viewed as their on-line analogs. Probably, the first technique of such kind is the method of potential functions, some versions of which can be considered as on-line analogs of SVM (see [10, 11] and [12, ch. 10]). Recently, on-line analogs of SVM and boosting-type methods using convex losses have been proposed in [14]. We also point out the paper [15], where the stochastic gradient algorithm with averaging is studied for the general class of loss functions (cf. [16]). All these papers use the standard stochastic gradient method for which the descent takes place in the initial parameter space.

[1] The work was made within the framework of Projects ACI NIM "BIOCLASSIF" and ACI MD "OPSYC," France.

[2] The research was made during visits to the Paris–VI and Grenoble–I Universities (France) in 2004–2005.

In this paper, we also suggest on-line versions of boosting and SVM, but based on a different principle: the gradient descent is performed in the dual space. Algorithms of this kind are known as mirror descent methods [17], and they were initially introduced for deterministic optimization problems. Their advantage, as compared to the standard gradient methods, is that the convergence rate depends logarithmically on the dimension of the problem. Therefore, they turn out to be very efficient in high-dimensional problems [18].

Some versions of the original mirror descent method of [17] (see also [19]) were derived independently in the learning community and have been applied to classification and other learning problems in the papers [20–22], where bounds for the relative risk criterion were obtained. However, these results are formulated in a deterministic setting, and they do not extend straightforwardly to the standard stochastic analysis with a mean risk criterion (see [20, 23, 24] for insights on connections between the two types of results). Below we propose a novel version of the mirror descent method, which attains the optimal bounds of the mean risk accuracy. Its main difference from the previous methods is the additional step of averaging of the updates.

The goal of this paper is to construct an aggregated decision rule: we introduce a fixed and finite base class of decision functions, and we choose weights in their convex or linear combination in an optimal way. The optimality of weights is understood in the sense of minimization of a convex risk function under the ℓ_1-constraints on the weights. This aggregation problem is similar to those considered, for instance, in [25, 26] for the regression model with squared loss. To solve the problem, we propose a recursive, on-line algorithm of the mirror descent type with averaging of updates. We prove that the algorithm converges with a rate of the order $C\sqrt{(\log M)/t}$ with an explicit and small constant factor C, where M is the dimension of the problem and t stands for the sample size.

The paper is organized as follows. First, we give the problem statement and formulate the main result on the convergence rate (Section 2). Then, the algorithm is described (Section 3) and the proof of the main result is given (Section 4). In Section 5, the result is extended to general loss functions and to the general estimation problem. Conclusive remarks are given in Section 6.

2. PROBLEM SETTING AND THE MAIN RESULT

We consider the problem of binary classification. Let (X, Y) be a pair of random variables with values in $\mathcal{X} \times \{-1, +1\}$, where $\mathcal{X}$ is a feature space. A decision rule $g_f \colon \mathcal{X} \to \{-1, +1\}$ corresponding to a measurable function $f \colon \mathcal{X} \to \mathbb{R}$ is defined as $g_f(x) = 2\mathbb{I}_{[f(x)>0]} - 1$, where $\mathbb{I}_{[\cdot]}$ denotes the indicator function. A standard measure of quality of a decision rule g_f is its risk, which equals the probability of misclassification: $R(g_f) = \mathbb{P}\{Y \neq g_f(X)\}$. An optimal decision rule is defined as g_{f^*}, where f^* is a minimizer of $R(g_f)$ over all measurable f. The optimal rule is not implementable in practice since the distribution of (X, Y) is unknown. In order to approximate g_{f^*}, one looks for empirical decision rules $\widehat{g}_n$ based on a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, where (X_i, Y_i) are independent random pairs having the same distribution as (X, Y).

An abstract approach to construction of empirical decision rules [4,27] prescribes to search for $\widehat{g}_n$ in the form $\widehat{g}_n = g_{\widehat{f}_n}$, where $\widehat{f}_n$ is a minimizer of the empirical risk (empirical classification error)

$$R_n(g_f) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\left[Y_i \neq g_f(X_i)\right]} \tag{1}$$

over all f from a given class of decision rules. Conditions of statistical optimality of the minimization method for empirical classification error (1) were extensively studied (see, in particular, [4,12,27]). However, this method is not computationally tractable since the risk functional $R_n(g_f)$ is neither convex nor even continuous in f. Until recently, this was the reason for an inconsistency between

the classification theory, proving the optimality of the method of empirical classification error (1) (see [4,12,27]), and practice, where quite other approaches were applied, such as, for instance, SVM and boosting based on numerical minimization of convex functionals of the empirical risk different from (1), as was noticed in [28, 29]. It is demonstrated in [5, 6, 8] that the inconsistency can be removed; namely, a scheme is proposed to prove that many of the practical methods have small classification errors. A key argument used in these works is that, under rather general assumptions, the optimal decision rule g_{f^*} coincides with g_{f^A}, where f^A is an optimal decision function in the sense that it minimizes a convex risk functional called the φ-risk and defined by

$$A(f) = \mathbf{E}\varphi(Yf(X)),$$

where $\varphi\colon \mathbb{R} \to \mathbb{R}_+$ is a convex loss function and $\mathbf{E}$ denotes the expectation. Typical choices of a loss function are the hinge loss function $\varphi(x) = (1-x)_+$ (used in SVM) as well as the exponential and logit losses (used in boosting): $\varphi(x) = \exp(-x)$ and $\varphi(x) = \log_2(1+\exp(-x))$, respectively.

Thus, to find an empirical decision rule $\widehat{g}_n$ which approximates the optimal g_{f^*}, we do not necessarily have to minimize the classification error (1) with respect to f; instead, we may consider minimization of the empirical φ-risk

$$A_n(f) = \frac{1}{n}\sum_{i=1}^{n} \varphi(Y_i f(X_i)),$$

which is an unbiased estimate for $A(f)$. This minimization problem is simpler than the original one because it can be solved by standard numerical procedures, the functional A_n being convex. When relevant penalty functions are added, it leads to some versions of boosting and SVM algorithms. At the same time, one needs the whole sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ for their implementation, i.e., these are *batch procedures*.

In this paper, we consider the problem of minimization of the φ-risk A on a finite-parametric class of functions f when the data (X_i, Y_i) come sequentially (on-line setting).

Let us introduce a parametric class of functions in which f is selected. Assume that a finite set of base functions $\{h_1, \ldots, h_M\}$ is given, where $h_j\colon \mathcal{X} \to [-K, K]$, $j = 1, \ldots, M$, $K \in (0, \infty)$ is a constant, and $M \geq 2$. We denote by H the vector function whose components are the base elements:

$$H(x) = (h_1(x), \ldots, h_M(x))^T. \tag{2}$$

A typical example is the one where the functions h_j are decision rules, i.e., they take values in $\{-1, 1\}$. Furthermore, for a fixed $\lambda > 0$ we denote by $\Theta_{M,\lambda}$ the λ-simplex in $\mathbb{R}^M$:

$$\Theta_{M,\lambda} = \left\{ \theta = (\theta^{(1)}, \ldots, \theta^{(M)})^T \in \mathbb{R}_+^M : \sum_{i=1}^{M} \theta^{(i)} = \lambda \right\}.$$

Introduce a family of λ-convex combinations of functions $h_1, \ldots, h_M$, i.e.,

$$\mathcal{F}_{M,\lambda} = \left\{ f_\theta = \theta^T H : \ \theta \in \Theta_{M,\lambda} \right\}.$$

That is the class of functions over which we want to minimize $A(f)$. Minimization of $A(f)$ over all $f \in \mathcal{F}_{M,\lambda}$ is equivalent to minimization of $A(f_\theta)$ over all $\theta \in \Theta_{M,\lambda}$, so we simplify the notations and in what follows write

$$A(\theta) \triangleq A(f_\theta).$$

Define the vector of optimal weights of a λ-convex combination of the base functions as a solution to the minimization problem

$$\min_{\theta \in \Theta_{M,\lambda}} A(\theta). \tag{3}$$

We assume that the distribution of (X, Y) is unknown, hence the function A is also unknown, and its direct minimization is impossible. However, we have access to a learning sample of independent random pairs (X_i, Y_i) having the same distribution as (X, Y) that are delivered sequentially and may be used for estimation of the optimal weights.

In Section 3, we propose a stochastic algorithm based on the mirror descent principle which, at the tth iteration, yields the estimate $\widehat{\theta}_t = \widehat{\theta}_t((X_1, Y_1), \dots, (X_{t-1}, Y_{t-1}))$ of the solution to problem (3). The estimate $\widehat{\theta}_t$ is measurable with respect to $(\widehat{\theta}_{t-1}, X_{t-1}, Y_{t-1})$, which means that the algorithm fits with the on-line setting. To obtain updates of the algorithm, it suffices to have random realizations of the subgradient of A which have the form

$$u_i(\theta) = \varphi'(Y_i \theta^T H(X_i)) Y_i H(X_i) \in \mathbb{R}^M, \quad i = 1, 2, \dots, \tag{4}$$

where φ' represents an arbitrary monotone version of the derivative of φ (one may take, for instance, the right continuous version).

Given $\widehat{\theta}_t$, a λ-convex combination $\widehat{\theta}_t^T H(\cdot)$ of the base functions can be constructed, which defines an aggregated decision rule

$$\widetilde{g}_t(x) = 2\mathbb{I}_{\left[\widehat{\theta}_t^T H(x) > 0\right]} - 1.$$

Statistical properties of this decision rule are described by the following result, which establishes the convergence rate for the expected accuracy of the estimator $\widehat{\theta}_t$ with respect to the φ-risk.

Theorem 1. *For a given convex loss function φ, for a fixed number $M \geq 2$ of base elements* (2) *and a fixed value of $\lambda > 0$, let the estimate $\widehat{\theta}_t$ be defined by the algorithm of Section* 3.4 (*see below*). *Then, for any integer $t \geq 1$,*

$$\mathbf{E} A(\widehat{\theta}_t) - \min_{\theta \in \Theta_{M,\lambda}} A(\theta) \leq C \frac{(\ln M)^{1/2} \sqrt{t+1}}{t}, \tag{5}$$

where $C = C(\varphi, \lambda) = 2\lambda L_\varphi(\lambda)$ and $L_\varphi(\lambda) = K \sup_{|x| \leq K\lambda} |\varphi'(x)|$.

For example, Theorem 1 holds with constant $C = 2$ in a typical case where we deal with convex ($\lambda = 1$) aggregation of base classifiers h_j taking values in $\{-1, 1\}$ and we use the hinge loss $\varphi(x) = (1 - x)_+$. We also note that Theorem 1 is distribution free: there is no assumption on the joint distribution of X and Y except that Y takes values in $\{-1, 1\}$ since we deal with the classification problem.

Remark 1 (efficiency). The rate of convergence of order $\sqrt{(\ln M)/t}$ is typical without low noise assumptions (introduced in [30]). Batch procedures based on minimization of the empirical convex risk functional present a similar rate. Hence, from the statistical point of view, there is no substantial difference between batch methods and our mirror descent procedure. On the other hand, from the computational point of view, our procedure is quite comparable with the direct stochastic gradient descent. However, the mirror descent algorithm presents two major advantages: (i) its behavior with respect to the cardinality of the base class is better than for the direct stochastic gradient descent (of the order of $\sqrt{\ln M}$ in Theorem 1, instead of M or $\sqrt{M}$ for the direct stochastic gradient); (ii) mirror descent presents a higher efficiency, especially in high-dimensional problems, since its algorithmic complexity and memory requirements are of order strictly smaller than for the corresponding batch procedures (see [25] for a comparison).

Remark 2 (optimality of the convergence rate). Using the techniques of [25, 26], it is easy to prove the minimax lower bound on the excess risk $\mathbf{E} A(\widehat{\theta}_t) - \min_{\theta \in \Theta_{M,\lambda}} A(\theta)$ having the order $\sqrt{(\ln M)/t}$ for $M \geq t^{1/2+\delta}$ with some $\delta > 0$. This indicates that the upper bound of Theorem 1 is rate optimal for such values of M.

Remark 3 (choice of the base class). The good behavior of this method crucially relies on the choice of the base class of functions $\{h_j\}_{1\le j\le M}$. A natural choice would be to consider a symmetric class in the sense that, if an element h is in the class, then $-h$ is also in the class. For a practical implementation, some initial data set should be available in order to preselect a set of M functions h_j. Another choice, which is practical and widely spread, is to choose very simple and not so efficient decision rules h_j, such as decision stumps (see, e.g., [3]); nevertheless, aggregation can lead to good performance if their cardinality M is very large. As far as theory is concerned, in order to provide a complete statistical analysis, one should establish approximation error bounds on the quantity $\inf_{f\in\mathcal{F}_{M,\lambda}} A(f) - \inf_f A(f)$ showing that the richness of the base class is reflected both by diversity (orthogonality or independence) of the functions h_j and by its cardinality M. For example, one can take h_j as the eigenfunctions associated to some positive definite kernel. We refer to [31] for related results (see also [7]). The choice of λ can be motivated by similar considerations. In fact, if the approximation error is taken into account, it might be useful to take λ depending on the sample size t and tending to infinity with a slow rate (cf. [6]). A balance between the stochastic error as given in Theorem 1 and the approximation error would then determine the optimal choice of λ. These considerations are left beyond the scope of the paper since here we focus on the aggregation problem.

3. DEFINITION AND DISCUSSION OF THE ALGORITHM

In this section, we introduce the proposed algorithm. It is based on the mirror descent idea going back to [17]. We first give some definitions and recall some facts from convex analysis.

3.1. Proxy Functions

We denote by $E = \ell_1^M$ the space $\mathbb{R}^M$ equipped with the 1-norm

$$\|z\|_1 = \sum_{j=1}^M |z^{(j)}|,$$

and denote by $E^* = \ell_\infty^M$ the dual space, which is $\mathbb{R}^M$ equipped with the sup-norm

$$\|z\|_\infty = \max_{\|\theta\|_1=1} z^T\theta = \max_{1\le j\le M} |z^{(j)}|, \quad \forall z\in E^*,$$

with the notation $z = (z^{(1)},\dots,z^{(M)})^T$.

Let Θ be a convex, closed set in E. For a given parameter $\beta > 0$ and a convex function $V\colon \Theta\to\mathbb{R}$, by the *$\beta$-conjugate* function of V we call the Legendre–Fenchel type transform of βV:

$$\forall z\in E^*, \quad W_\beta(z) = \sup_{\theta\in\Theta}\left\{-z^T\theta - \beta V(\theta)\right\}. \tag{6}$$

Now we introduce the key assumption (Lipschitz condition in the conjugate norms $\|\cdot\|_1$ and $\|\cdot\|_\infty$), which will be used in the proof of Theorem 1.

Assumption (L). *A convex function $V\colon\Theta\to\mathbb{R}$ is such that its β-conjugate W_β is continuously differentiable on E^* and its gradient ∇W_β satisfies the inequality*

$$\|\nabla W_\beta(z) - \nabla W_\beta(\widetilde{z})\|_1 \le \frac{1}{\alpha\beta}\|z-\widetilde{z}\|_\infty, \quad \forall z,\widetilde{z}\in E^*, \quad \beta>0,$$

where $\alpha > 0$ is a constant independent of β.

As is known (see, e.g., [19,32]), this assumption is related to the notion of strong convexity with respect to the $\|\cdot\|_1$-norm.

Definition 1. Fix $\alpha > 0$. A convex function $V\colon \Theta \to \mathbb{R}$ is said to be *α-strongly convex* with respect to the norm $\|\cdot\|_1$ if

$$V(sx + (1-s)y) \le sV(x) + (1-s)V(y) - \frac{\alpha}{2}s(1-s)\|x-y\|_1^2 \tag{7}$$

for all $x, y \in \Theta$ and any $s \in [0,1]$.

The following proposition sums up some properties of β-conjugates and, in particular, yields a sufficient condition for Assumption (L).

Proposition 1. *Let a function $V\colon \Theta \to \mathbb{R}$ be convex and parameter β be positive. Then the β-conjugate, W_β, of V has the following properties.*

1. *The function $W_\beta\colon E^* \to \mathbb{R}$ is convex and has a conjugate βV, i.e.,*

$$\forall \theta \in \Theta, \quad \beta V(\theta) = \sup_{z \in E^*} \left\{ -z^T\theta - W_\beta(z) \right\}.$$

2. *If V is α-strongly convex with respect to the norm $\|\cdot\|_1$, then*
   (i) *Assumption* (L) *holds true*;
   (ii) $\arg\max\limits_{\theta\in\Theta} \left\{ -z^T\theta - \beta V(\theta) \right\} = -\nabla W_\beta(z) \in \Theta$.

For a proof of this proposition, see [19,32].

Definition 2. We call a function $V\colon \Theta \to \mathbb{R}_+$ a *proxy function* if it is convex and
   (i) There exists a point $\theta_* \in \Theta$ such that $\min\limits_{\theta\in\Theta} V(\theta) = V(\theta_*)$;
   (ii) Assumption (L) holds true.

Example. Let $\Theta = \Theta_{M,\lambda}$. Consider an entropy-type proxy function

$$\forall \theta \in \Theta_{M,\lambda}, \quad V(\theta) = \lambda \ln\left(\frac{M}{\lambda}\right) + \sum_{j=1}^{M} \theta^{(j)} \ln \theta^{(j)}, \tag{8}$$

(where $0 \ln 0 \triangleq 0$) which has a single minimizer $\theta_* = (\lambda/M, \dots, \lambda/M)^T$ with $V(\theta_*) = 0$. It is easy to check that this function is α-strongly convex with respect to the norm $\|\cdot\|_1$, with the parameter $\alpha = 1/\lambda$. Hence, Assumptions (L) holds true by Proposition 1 (see Appendix for a direct proof). An important property of this choice of V is that the optimization problem (6) can be solved explicitly, so that W_β and ∇W_β are given by the following formulas:

$$\forall z \in E^*, \quad W_\beta(z) = \lambda\beta \ln\left(\frac{1}{M}\sum_{k=1}^{M} e^{-z^{(k)}/\beta}\right), \tag{9}$$

$$\frac{\partial W_\beta(z)}{\partial z^{(j)}} = -\lambda e^{-z^{(j)}/\beta} \left(\sum_{k=1}^{M} e^{-z^{(k)}/\beta}\right)^{-1}, \quad j = 1, \dots, M. \tag{10}$$

Note that for $\lambda = 1$ we have the following:

– The proxy function (8) equals the Kullback information divergence between the uniform distribution on the set $\{1, \dots, M\}$ and the distribution on the same set defined by probabilities $\theta^{(j)}$, $j = 1, \dots, M$;
– In view of (10), the components of the vector $-\nabla W_\beta(z)$ define a Gibbs distribution on the coordinates of vector z, with β interpreted as a temperature parameter.

3.2. Algorithm

Mirror descent algorithms are recursive optimization procedures achieving a stochastic gradient descent in the dual space. The proposed algorithm is of this kind with a peculiarity that it uses *stochastic* subgradients and averaging of iteration results. At each iteration i, a new data pair (X_i, Y_i) is observed, and there are two updates: one is the variable ζ_i defined by the stochastic subgradients $u_k(\theta_{k-1})$, $k = 1, \dots, i$, as the result of the descent in the dual space E^*; the other update is the parameter θ_i, which is the "mirror image" of ζ_i in the initial space E. In order to tune the algorithm properly, we will also need two fixed positive sequences $(\gamma_i)_{i\geq 1}$ (step size) and $(\beta_i)_{i\geq 1}$ ("temperature") such that $\beta_i \geq \beta_{i-1}$, $\forall i \geq 1$. The algorithm is defined as follows:

- Fix the initial values $\theta_0 \in \Theta$ and $\zeta_0 = 0 \in \mathbb{R}^M$.
- For $i = 1, \dots, t-1$, do the recursive update

$$\begin{aligned} \zeta_i &= \zeta_{i-1} + \gamma_i u_i(\theta_{i-1}), \\ \theta_i &= -\nabla W_{\beta_i}(\zeta_i). \end{aligned} \tag{11}$$

- At iteration t, output the following convex combination:

$$\widehat{\theta}_t = \frac{\sum\limits_{i=1}^{t} \gamma_i \theta_{i-1}}{\sum\limits_{i=1}^{t} \gamma_i}. \tag{12}$$

Note that the components $\theta_i^{(j)}$ of the vector θ_i from (11) have the form

$$\theta_i^{(j)} = \frac{\lambda \exp\left(-\beta_i^{-1} \sum\limits_{m=1}^{i} \gamma_m u_{m,j}(\theta_{m-1})\right)}{\sum\limits_{k=1}^{M} \exp\left(-\beta_i^{-1} \sum\limits_{m=1}^{i} \gamma_m u_{m,k}(\theta_{m-1})\right)},$$

where $u_{m,j}(\theta)$ represents the jth component of $u_m(\theta)$, $j = 1, \dots, M$.

3.3. Heuristics

Assume that we want to minimize a convex function $\theta \mapsto A(\theta)$ over a convex set Θ. If $\theta_0, \dots, \theta_{t-1}$ are available search points at iteration t, we can provide affine approximations ϕ_i of the function A, which are defined for $\theta \in \Theta$ by

$$\phi_i(\theta) = A(\theta_{i-1}) + (\theta - \theta_{i-1})^T \nabla A(\theta_{i-1}), \quad i = 1, \dots, t.$$

Here $\theta \mapsto \nabla A(\theta)$ is a vector function belonging to the subdifferential of $A(\cdot)$. Taking a convex combination of the functions ϕ_i, we obtain an averaged approximation of $A(\theta)$:

$$\bar{\phi}_t(\theta) = \frac{\sum\limits_{i=1}^{t} \gamma_i \left(A(\theta_{i-1}) + (\theta - \theta_{i-1})^T \nabla A(\theta_{i-1})\right)}{\sum\limits_{i=1}^{t} \gamma_i}.$$

At first glance, it seems reasonable to choose, as the next search point, a vector $\theta_t \in \Theta$ minimizing the approximation $\bar{\phi}_t$, i.e.,

$$\theta_t = \arg\min_{\theta\in\Theta} \bar{\phi}_t(\theta) = \arg\min_{\theta\in\Theta} \theta^T \left(\sum_{i=1}^{t} \gamma_i \nabla A(\theta_{i-1})\right).$$

However, this does not make any progress because our approximations are "good" only in the vicinity of search points $\theta_0, \dots, \theta_{t-1}$. Therefore, it is necessary to modify the criterion, for instance, by adding some penalty $B_t(\theta, \theta_{t-1})$ to the target function in order to keep the next search point θ_t in the vicinity of the previous one, θ_{t-1}. Thus, one chooses the point

$$\theta_t = \arg\min_{\theta\in\Theta} \left[\theta^T \left(\sum_{i=1}^{t} \gamma_i \nabla A(\theta_{i-1}) \right) + B_t(\theta, \theta_{t-1}) \right]. \tag{13}$$

Our algorithm corresponds to a specific type of penalty $B_t(\theta, \theta_{t-1}) = \beta_t V(\theta)$, where V is the proxy function. Note also that in our problem the vector-function $\nabla A(\cdot)$ is not available. Therefore, we replace in (13) the unobservable gradients $\nabla A(\theta_{i-1})$ by the stochastic subgradients $u_i(\theta_{i-1})$. This yields a new definition of the tth search point:

$$\theta_t = \arg\min_{\theta\in\Theta} \left[\theta^T \left(\sum_{i=1}^{t} \gamma_i u_i(\theta_{i-1}) \right) + \beta_t V(\theta) \right] = \arg\max_{\theta\in\Theta} \left[-\zeta_t^T \theta - \beta_t V(\theta) \right], \tag{14}$$

where

$$\zeta_t = \sum_{i=1}^{t} \gamma_i u_i(\theta_{i-1}).$$

Observe that, by Proposition 1, the value of (14) coincides with $-\nabla W_{\beta_t}(\zeta_t)$; it is now easy to deduce the iterative scheme (11).

3.4. Particular Case of the Algorithm

A special case of the mirror descent method with averaging, for which Theorem 1 is proved, is defined as follows. It is the algorithm described in Section 3.2 with the entropy-type proxy function V as defined in (8) and with the sequences $(\gamma_i)_{i\ge1}$ and $(\beta_i)_{i\ge1}$ of the form

$$\gamma_i \equiv 1, \quad \beta_i = \beta_0\sqrt{i+1}, \quad i = 1, 2, \dots, \tag{15}$$

where

$$\beta_0 = L_\varphi(\lambda)(\ln M)^{-1/2}. \tag{16}$$

Thus, the algorithm becomes simpler and can be implemented in the following recursive form:

$$\zeta_i = \zeta_{i-1} + u_i(\theta_{i-1}), \tag{17}$$

$$\theta_i = -\nabla W_{\beta_i}(\zeta_i), \tag{18}$$

$$\widehat{\theta}_i = \widehat{\theta}_{i-1} - \frac{1}{i}\left(\widehat{\theta}_{i-1} - \theta_{i-1}\right), \quad i = 1, 2, \dots, \tag{19}$$

with initial values $\zeta_0 = 0$, $\theta_0 \in \Theta$, and $(\beta_i)_{i\ge1}$ from (15) and (16).

3.5. Comparison with Other Mirror Descent Methods

The versions of mirror descent method proposed in [17] are somewhat different from our iterative scheme (11). One of them, which is the closest to (11), is studied in detail in [19]. It is based on the recursive relation

$$\theta_i = -\nabla W_1\Big(-\nabla V(\theta_{i-1}) + \gamma_i u_i(\theta_{i-1})\Big), \quad i = 1, 2, \dots, \tag{20}$$

where the function V is strongly convex with respect to the norm of an initial space E (which is not necessarily the space ℓ_1^M) and W_1 is the ordinary conjugate to V. If $\Theta = \mathbb{R}^M$ and $V(\theta) = \frac{1}{2}\|\theta\|_2^2$,

the scheme of (20) coincides with the ordinary gradient method. For the unit simplex $\Theta = \Theta_{M,1}$ and the entropy-type proxy function V from (8), the components $\theta_i^{(j)}$ of the vector θ_i from (20) are explicit:

$$\theta_i^{(j)} = \frac{\theta_{i-1}^{(j)} \exp\left(-\gamma_i u_{i,j}(\theta_{i-1})\right)}{\sum\limits_{k=1}^{M} \theta_{i-1}^{(k)} \exp\left(-\gamma_i u_{i,k}(\theta_{i-1})\right)} = \frac{\theta_0^{(j)} \exp\left(-\sum\limits_{m=1}^{i} \gamma_m u_{m,j}(\theta_{m-1})\right)}{\sum\limits_{k=1}^{M} \theta_0^{(k)} \exp\left(-\sum\limits_{m=1}^{i} \gamma_m u_{m,k}(\theta_{m-1})\right)}, \tag{21}$$

$j = 1, \ldots, M$. Algorithm (21) is also known as the exponentiated gradient (EG) method [20]. The differences between (20) and our algorithm are the following:

– The initial iterative scheme (11) is different from that of (20); in particular, it includes the second tuning parameter β_i; moreover, algorithm (21) uses the initial value θ_0 in a different manner;
– Along with (11), our algorithm contains an additional step of averaging of the updates (12).

The papers [21, 22] study convergence properties of the EG method (21) in a deterministic setting. Namely, they show that, under certain assumptions, the difference $A_t(\theta_t) - \min\limits_{\theta \in \Theta_{M,1}} A_t(\theta)$ is bounded by a constant depending on M and t. If this constant is small enough, these results show that the EG method provides good numerical minimizers of the empirical risk A_t. However, they do not apply to the expected risk. In particular, they do not imply that the expected risk $\mathbf{E}A(\theta_t)$ is close to the minimal possible value $\min\limits_{\theta \in \Theta_{M,1}} A(\theta)$, which, as we prove, is true for the algorithm with averaging proposed here.

Finally, we point out that algorithm (20) can be deduced from the ideas mentioned in Section 3.3 and which are studied in the literature on proximal methods of convex optimization (see, e.g., [33,34] and references therein). Namely, under rather general conditions, the variable θ_i from (20) is a solution to the minimization problem

$$\theta_i = \arg\min_{\theta \in \Theta} \left(\theta^T \gamma_i u_i(\theta_{i-1}) + B(\theta, \theta_{i-1}) \right),$$

where the penalty $B(\theta, \theta_{i-1}) = V(\theta) - V(\theta_{i-1}) - (\theta - \theta_{i-1})^T \nabla V(\theta_{i-1})$ represents the Bregman divergence between θ and θ_{i-1} related to a strongly convex function V.

4. PROOFS

In this section, we provide technical details leading to the result of Theorem 1. They will be given in a more general setting than that of Theorem 1. Namely, we will consider an arbitrary proxy function V and use the notations and assumptions of Section 3.2. Propositions 2 and 3 below are valid for an arbitrary closed convex set Θ in E and for the estimate sequences (θ_i) and $(\widehat{\theta}_i)$ defined by the algorithm (11) and (12). The arguments up to relation (26) in the proof of Theorem 1 are valid under the assumption that Θ is a convex compact set in E.

Introduce the notations

$$\nabla A(\theta) = \mathbf{E} u_i(\theta),$$
$$\xi_i(\theta) = u_i(\theta) - \nabla A(\theta), \quad \forall \theta \in \Theta,$$

where the random functions $u_i(\theta)$ are defined in (4). Note that the mapping $\theta \mapsto \mathbf{E}u_i(\theta)$ belongs to the subdifferential of A (which explains the notation ∇A). This fact and the inequality $\mathbf{E}\|u_i(\theta)\|_\infty^2 \le L_\varphi^2(\lambda)$, valid for all $\theta \in \Theta$, are the only properties of u_i that will be used in the proofs, other specific features of definition (4) being of no importance.

Proposition 2. *For any $\theta \in \Theta$ and any integer $t \geq 1$, we have the inequality*

$$\sum_{i=1}^{t} \gamma_i (\theta_{i-1} - \theta)^T \nabla A(\theta_{i-1}) \leq \beta_t V(\theta) - \beta_0 V(\theta_*) - \sum_{i=1}^{t} \gamma_i (\theta_{i-1} - \theta)^T \xi_i(\theta_{i-1}) + \sum_{i=1}^{t} \frac{\gamma_i^2}{2\alpha\beta_{i-1}} \|u_i(\theta_{i-1})\|_\infty^2.$$

Proof. By continuous differentiability of $W_{\beta_{t-1}}$, we have

$$W_{\beta_{i-1}}(\zeta_i) = W_{\beta_{i-1}}(\zeta_{i-1}) + \int_0^1 (\zeta_i - \zeta_{i-1})^T \nabla W_{\beta_{i-1}}(\tau\zeta_i + (1-\tau)\zeta_{i-1})\, d\tau.$$

Put $v_i = u_i(\theta_{i-1})$. Then $\zeta_i - \zeta_{i-1} = \gamma_i v_i$, and by Assumption (L) we have

$$\begin{aligned} W_{\beta_{i-1}}(\zeta_i) &= W_{\beta_{i-1}}(\zeta_{i-1}) + \gamma_i v_i^T \nabla W_{\beta_{i-1}}(\zeta_{i-1}) \\ &\quad + \gamma_i \int_0^1 v_i^T \Big[\nabla W_{\beta_{i-1}}(\tau\zeta_i + (1-\tau)\zeta_{i-1}) - \nabla W_{\beta_{i-1}}(\zeta_{i-1})\Big]\, d\tau \\ &\leq W_{\beta_{i-1}}(\zeta_{i-1}) + \gamma_i v_i^T \nabla W_{\beta_{i-1}}(\zeta_{i-1}) \\ &\quad + \gamma_i \|v_i\|_\infty \int_0^1 \|\nabla W_{\beta_{i-1}}(\tau\zeta_i + (1-\tau)\zeta_{i-1}) - \nabla W_{\beta_{i-1}}(\zeta_{i-1})\|_1\, d\tau \\ &\leq W_{\beta_{i-1}}(\zeta_{i-1}) + \gamma_i v_i^T \nabla W_{\beta_{i-1}}(\zeta_{i-1}) + \frac{\gamma_i^2 \|v_i\|_\infty^2}{2\alpha\beta_{i-1}}. \end{aligned}$$

Using the last inequality and the facts that $(\beta_i)_{i\geq 1}$ is a nondecreasing sequence and that, for z fixed, $\beta \mapsto W_\beta(z)$ is a nonincreasing function, we get

$$W_{\beta_i}(\zeta_i) \leq W_{\beta_{i-1}}(\zeta_i) \leq W_{\beta_{i-1}}(\zeta_{i-1}) - \gamma_i \theta_{i-1}^T v_i + \frac{\gamma_i^2 \|v_i\|_\infty^2}{2\alpha\beta_{i-1}}.$$

Summing up, we obtain

$$\sum_{i=1}^{t} \gamma_i \theta_{i-1}^T v_i \leq W_{\beta_0}(\zeta_0) - W_{\beta_t}(\zeta_t) + \sum_{i=1}^{t} \frac{\gamma_i^2 \|v_i\|_\infty^2}{2\alpha\beta_{i-1}}.$$

Using the representation $\zeta_t = \sum_{i=1}^{t} \gamma_i v_i$, we get that, for any $\theta \in \Theta$,

$$\sum_{i=1}^{t} \gamma_i (\theta_{i-1} - \theta)^T v_i \leq W_{\beta_0}(\zeta_0) - W_{\beta_t}(\zeta_t) - \zeta_t^T \theta + \sum_{i=1}^{t} \frac{\gamma_i^2 \|v_i\|_\infty^2}{2\alpha\beta_{i-1}}.$$

Finally, since $v_i = \nabla A(\theta_{i-1}) + \xi_i(\theta_{i-1})$, we find

$$\sum_{i=1}^{t} \gamma_i (\theta_{i-1} - \theta)^T \nabla A(\theta_{i-1}) \leq W_{\beta_0}(\zeta_0) - W_{\beta_t}(\zeta_t) - \zeta_t^T \theta - \sum_{i=1}^{t} \gamma_i (\theta_{i-1} - \theta)^T \xi_i(\theta_{i-1}) + \sum_{i=1}^{t} \frac{\gamma_i^2 \|v_i\|_\infty^2}{2\alpha\beta_{i-1}}.$$

Thus, the desired inequality follows from the fact that

$$W_{\beta_0}(\zeta_0) = W_{\beta_0}(0) = \beta_0 \sup_{\theta\in\Theta}\{-V(\theta)\} = -\beta_0 V(\theta_*)$$

and $\beta V(\theta) \geq -W_\beta(\zeta) - \zeta^T\theta$, for all $\zeta \in \mathbb{R}^M$. $\triangle$

Now we derive the main result of this section.

Proposition 3. *For any integer $t \geq 1$, we have the inequality*

$$\mathbf{E}A(\widehat{\theta}_t) \leq \inf_{\theta\in\Theta}\left[A(\theta) + \frac{\beta_t V(\theta) - \beta_0 V(\theta_*)}{\sum\limits_{i=1}^t \gamma_i}\right] + L_\varphi^2(\lambda)\left(\sum_{i=1}^t \gamma_i\right)^{-1} \sum_{i=1}^t \frac{\gamma_i^2}{2\alpha\beta_{i-1}}. \tag{22}$$

Hence, the expected accuracy of the estimate $\widehat{\theta}_t$ with respect to the φ-risk satisfies the following upper bound:

$$\mathbf{E}A(\widehat{\theta}_t) - \min_{\theta\in\Theta} A(\theta) \leq \frac{1}{\sum\limits_{i=1}^t \gamma_i}\left(\beta_t V(\theta_A^*) - \beta_0 V(\theta_*) \ + \ L_\varphi^2(\lambda)\sum_{i=1}^t \frac{\gamma_i^2}{2\alpha\beta_{i-1}}\right), \tag{23}$$

where $\theta_A^* \in \arg\min\limits_{\theta\in\Theta} A(\theta)$.

Proof. For any $\theta \in \Theta$, by convexity of $A(\theta)$, we get

$$\mathbf{E}A(\widehat{\theta}_t) - A(\theta) \leq \frac{\sum\limits_{i=1}^t \gamma_i \left(\mathbf{E}A(\theta_{i-1}) - A(\theta)\right)}{\sum\limits_{i=1}^t \gamma_i} \leq \frac{\sum\limits_{i=1}^t \gamma_i \mathbf{E}[(\theta_{i-1} - \theta)^T \nabla A(\theta_{i-1})]}{\sum\limits_{i=1}^t \gamma_i}. \tag{24}$$

Conditioning on θ_{i-1} and then using both the definition of $\xi_i(\theta_{i-1})$ and the independence between θ_{i-1} and (X_i, Y_i), we obtain $\mathbf{E}\xi_i(\theta_{i-1}) = 0$. We now combine (24) and the inequality of Proposition 2, making use of the bound $\sup\limits_{\theta\in\Theta} \mathbf{E}\|u_i(\theta)\|_\infty^2 \leq L_\varphi^2(\lambda)$. This leads to (22). Inequality (23) is straightforward in view of (22). $\triangle$

Note that simultaneous (and the same) change of scale in the definition of the sequences (β_i) and (γ_i) (i.e., multiplication by the same positive constant factor) does not affect the upper bounds in Propositions 2 and 3; moreover, the sequences (θ_i) and $(\widehat{\theta}_i)$ of algorithm (11) and (12) do not change (for $\zeta_0 = 0$).

Proof of Theorem 1. We have $V(\theta_*) = 0$ and

$$V(\theta_A^*) \leq \max_{\theta\in\Theta} V(\theta) \triangleq V^*.$$

Using (23) with the choice $\gamma_i \equiv 1$ and $\beta_i = \beta_0\sqrt{i+1}$ for an arbitrary $\beta_0 > 0$, we get

$$\mathbf{E}A(\widehat{\theta}_t) - A(\theta_A^*) \leq \frac{\sqrt{t+1}}{t}\left(\beta_0 V^* + \frac{L_\varphi^2(\lambda)}{\alpha\beta_0}\right). \tag{25}$$

The minimum of this bound over β_0 is attained at

$$\beta_0 = \frac{L_\varphi(\lambda)}{\sqrt{\alpha V^*}},$$

which gives the bound

$$\mathbf{E}A(\widehat{\theta}_t) - A(\theta_A^*) \le \frac{2L_\varphi(\lambda)}{t}\sqrt{\frac{V^*}{\alpha}(t+1)}. \tag{26}$$

Let now $\Theta = \Theta_{M,\lambda}$. Recall that, for the function V defined in (8), we have $\alpha = \lambda^{-1}$. Furthermore, this proxy function attains its maximum at each vertex of the λ-simplex $\Theta_{M,\lambda}$ and satisfies

$$V^* = \max_{\theta\in\Theta_{M,\lambda}} V(\theta) = \lambda \ln M.$$

Therefore, the optimal value β_0 equals $L_\varphi(\lambda)(\ln M)^{-1/2}$. This gives the accuracy bound as in the statement of Theorem 1. $\triangle$

5. GENERALIZATION

The proof given above does not actually use the fact that both the loss function and proxy function have a specific form. The required properties of these functions are summarized at the beginning of Section 4. Therefore, the result of the Theorem 1 can be stated under more general assumptions. This is done in the present Section.

Let a random variable Z take its values in a set $\mathcal{Z}$. Let Θ be a convex closed set in $\mathbb{R}^M$, and let a loss function $Q\colon \Theta \times \mathcal{Z} \to \mathbb{R}_+$ be such that the random function $Q(\cdot\,, Z)\colon \Theta \to \mathbb{R}_+$ is convex almost surely. Define the convex risk function $A\colon \Theta \to \mathbb{R}_+$ as follows:

$$A(\theta) = \mathbf{E}Q(\theta, Z).$$

Assume that a learning sample is given in the form of an i.i.d. sequence $(Z_1, \dots, Z_{t-1})$, where each Z_i has the same distribution as Z. Our aim now consists in *criterial minimization* of A over Θ (see, e.g., [35]), which means that we characterize the accuracy of the estimate $\widehat{\theta}_t = \widehat{\theta}_t(Z_1, \dots, Z_{t-1}) \in \Theta$ (minimizing A) by the difference

$$\mathbf{E}A(\widehat{\theta}_t) - \min_{\theta\in\Theta} A(\theta)$$

(we assume that $\min\limits_{\theta\in\Theta} A(\theta)$ is attainable). We denote by

$$u_i(\theta) = \nabla_\theta Q(\theta, Z_i), \quad i = 1, 2, \dots, \tag{27}$$

the stochastic subgradients which are measurable functions defined on $\Theta \times \mathcal{Z}$ such that, for any $\theta \in \Theta$, the expectation $\mathbf{E}u_i(\theta)$ belongs to the subdifferential of the function $A(\theta)$.

Theorem 2. *Let Θ be a convex closed set in $\mathbb{R}^M$, and let $Q(\cdot\,,\cdot)$ be a loss function which meets the conditions mentioned above. Moreover, assume that*

$$\sup_{\theta\in\Theta} \mathbf{E}\|\nabla_\theta Q(\theta, Z)\|_\infty^2 \le L_{\Theta,Q}^2, \tag{28}$$

where $L_{\Theta,Q}$ is a finite constant. Let V be a proxy function on Θ satisfying Assumption (L) *with a parameter $\alpha > 0$, and assume that there exists $\theta_A^* \in \arg\min\limits_{\theta\in\Theta} A(\theta)$. Then, for any integer $t \ge 1$, the estimate $\widehat{\theta}_t$ defined in Section* 3.2 *with stochastic subgradients* (27) *and with sequences $(\gamma_i)_{i\ge1}$ and $(\beta_i)_{i\ge1}$ from* (15) *with an arbitrary $\beta_0 > 0$ satisfies the inequality*

$$\mathbf{E}A(\widehat{\theta}_t) - \min_{\theta\in\Theta} A(\theta) \le \left(\beta_0 V(\theta_A^*) + \frac{L_{\Theta,Q}^2}{\alpha\beta_0}\right)\frac{\sqrt{t+1}}{t}.$$

Furthermore, if $\overline{V}$ *is a constant such that* $V(\theta_A^*) \le \overline{V}$ *and we set* $\beta_0 = L_{\Theta,Q}(\alpha\overline{V})^{-1/2}$, *then*

$$\mathbf{E}A(\widehat{\theta}_t) - \min_{\theta\in\Theta} A(\theta) \le 2L_{\Theta,Q}\left(\alpha^{-1}\overline{V}\right)^{1/2}\frac{\sqrt{t+1}}{t}. \tag{29}$$

In particular, if Θ *is a convex compact set, we can take* $\overline{V} = \max_{\theta\in\Theta} V(\theta)$.

This theorem follows from the proofs of Section 4 (cf. (23), (25), and (26)), where $L_{\Theta,Q}$ should replace the constant $L_\varphi(\lambda)$. It generalizes Theorem 1 and encompasses different statistical models, including the one described in Section 2, where the role of Z is played by the pair of random variables (X, Y), sets $\Theta = \Theta_{M,\lambda}$ and $\mathcal{Z} = \mathcal{X}\times\{-1,+1\}$, and loss function $Q(\theta, Z) = \varphi(Y\theta^T H(X))$. In the same way, Theorem 2 is also applicable to the standard regression model with squared loss $Q(\theta, Z) = (Y - \theta^T H(X))^2$, in which case a similar result was proved in [25] for another method.

Finally note that the constant $L_{\Theta,Q}$ from Theorem 2 can be raised, aiming at its simpler calculation, by taking the supremum in (28) over a wider subset of Θ, for instance, in the case of $\Theta \subset \{\theta : \|\theta\|_1 \le \lambda\}$ for some $\lambda > 0$; one may also proceed in a similar manner when calculating the constant $\overline{V} \ge \max_{\theta\in\Theta} V(\theta)$.

Remark 4 (dependent data). Inspection of the proofs shows that Theorem 2 can easily be extended to the case of dependent data Z_i. In fact, instead of assuming that Z_i are i.i.d., it suffices to assume that they form a stationary sequence, where each Z_i has the same distribution as Z. Then Theorem 2 remains valid if we additionally assume that the conditional expectations $\mathbf{E}(\xi_i(\theta_{i-1})\,|\,\theta_{i-1}) = 0$ a.s.

6. CONCLUSIVE REMARKS

To conclude with, we discuss the choices of the proxy function V, parametric set Θ, and sequences $(\beta_i)_{i\ge1}$ and $(\gamma_i)_{i\ge1}$.

6.1. Choice of the Proxy Function V

The choice of the entropic proxy function defined in (8) is not the only possible. A key condition on V is the strong convexity with respect to the norm $\|\cdot\|_1$, which guarantees that Assumption (L) holds true. One may also consider other proxy functions satisfying this condition, such as, for instance,

$$\forall\theta\in\mathbb{R}^M,\quad V(\theta) = \frac{1}{2\lambda^2}\|\theta\|_p^2 = \frac{1}{2\lambda^2}\left(\sum_{j=1}^M\left(\theta^{(j)}\right)^p\right)^{2/p}, \tag{30}$$

where $p = 1 + 1/\ln M$ (see [19]). In contrast with the function (8), the proxy function (30) can be used when Θ is any convex closed set in $\mathbb{R}^M$. For the simplex $\Theta_{M,\lambda}$, one may consider functions of the form

$$\forall\theta\in\Theta_{M,\lambda},\quad V(\theta) = C_0 + C_1\sum_{j=1}^M\left(\theta^{(j)}\right)^{s+1},\quad s = \frac{1}{\ln M}, \tag{31}$$

where the constants $C_0 = -\lambda^2/(es(s+1))$ and $C_1 = \lambda^{1-s}/(s(s+1))$ are adjusted in order to have $\min_{\theta\in\Theta_{M,\lambda}} V(\theta) = 0$. It is easy to see that the proxy function (31) is α-strongly convex in the norm $\|\cdot\|_1$. When $\lambda = 1$, this proxy function is a particular case of the Csiszár f-divergence (see the definition in [36]) between the uniform distribution on the set $\{1,\dots,M\}$ and the distribution on the same set defined by the probabilities $\theta^{(j)}$. Recall that, for $\lambda = 1$, the proxy function defined in (8) equals the Kullback divergence between these distributions. Presumably, other proxy functions can be based on some properly chosen Csiszár f-divergences.

On the other hand, if a proxy function V is such that the gradient of its β-conjugate ∇W_β cannot be written explicitly, numerical implementation of our algorithm might become time-consuming. From the upper bound (29), we can see that an important characteristic of V is the ratio $\overline{V}/\alpha$ (or, in particular, $\max_{\theta\in\Theta} V(\theta)/\alpha$ if the set Θ is bounded). One may define optimal proxy functions minimizing this ratio. We conjecture that for $\Theta = \Theta_{M,\lambda}$ such an optimal proxy function is the entropy-type function given in (8); however, we do not have a rigorous proof of this fact. For the latter, we have

$$\frac{1}{\alpha}\max_{\theta\in\Theta_{M,\lambda}} V(\theta) = \lambda^2 \ln M.$$

For other proxy functions, this ratio is of the same order. For instance, it is proved in [19, Lemma 6.1] that the proxy function (30) satisfies

$$\frac{1}{\alpha}\max_{\theta\in\Theta_{M,\lambda}} V(\theta) = O(1)\lambda^2 \ln M.$$

This relation holds true for the proxy function defined in (31) as well. Finally, note that such a widely used penalty function as $\|\cdot\|_1$ is not a proxy function in the sense of Definition 2 since it is not strongly convex with respect to $\|\cdot\|_1$. Another frequently used penalty function, $\|\theta\|_2^2$, is strongly convex. However, it can easily be verified that its "performance ratio" is extremely bad for large M: this function V satisfies

$$\frac{1}{\alpha}\max_{\theta\in\Theta_{M,\lambda}} V(\theta) = \frac{1}{2}\lambda^2 M.$$

6.2. Choice of the Set Θ

Theorem 2 holds for any convex closed set Θ contained in $\mathbb{R}^M$. However, for the general sets, the gradient ∇W_β usually cannot be computed explicitly, and the computation effort of implementing an iteration of the algorithm can become prohibitive. Hence, it is important to consider only sets Θ for which the solution $\theta_*(z) = -\nabla W_\beta(z)$ of the optimization problem (6) can be computed easily. Some examples of such "simple" sets are: (i) the λ-simplex $\Theta_{M,\lambda}$, (ii) the full-dimensional λ-simplex $\{\theta \in \mathbb{R}_+^M : \|\theta\|_1 \le \lambda\}$, and (iii) the symmetrized version of the latter, the hyper-octahedron $\{\theta \in \mathbb{R}^M : \|\theta\|_1 \le \lambda\}$.

6.3. Choice of the Sequences (β_i) and (γ_i)

The constant factor in the bound of Theorem 1 can only be slightly decreased. The sequences (β_i) and (γ_i) as described in (15) are close to optimal ones in the sense of the upper bound (23). Indeed, if we further bound $V(\theta_A^*)$ by $V^* = \max_{\theta\in\Theta} V(\theta)$ in (23) and minimize with respect to (γ_i) and (β_i) under the monotonicity condition $\beta_i \ge \beta_{i-1}$, we get that the minimum is obtained for sequences (γ_i) and (β_i) which are independent of i and such that $\beta_i/\gamma_i \equiv L_\varphi(\lambda)\sqrt{t/2\alpha V^*}$. We can take, for instance,

$$\gamma_i \equiv \frac{1}{\sqrt{t}}, \qquad \beta_i \equiv \frac{L_\varphi(\lambda)}{\sqrt{2\alpha V^*}}, \tag{32}$$

which leads to a better bound than that of (5) in Theorem 1:

$$\mathbf{E}A(\widehat{\theta}_t) - \min_{\theta\in\Theta_{M,\lambda}} A(\theta) \le \lambda L_\varphi(\lambda)\sqrt{\frac{2\ln M}{t}}.$$

Thus, we can improve the constant factor in the upper bound from 2 to $\sqrt{2}$. However, in order to make this improvement, one needs to know the sample size t in advance, and this is not compatible with the on-line framework.

APPENDIX

We now prove that Assumption (L) holds for the β-conjugate, W_β, of the entropy-type function V given by (8).

Evidently, the function W_β in (9) is twice continuously differentiable on $E^* = \ell_\infty^M$. Introduce

$$L = \sup_{\substack{z_1,z_2\in E^*\\ z_1\neq z_2}} \frac{\|\nabla W(z_1) - \nabla W(z_2)\|_1}{\|z_1 - z_2\|_\infty} \le \sup_{\substack{\|x\|_\infty=1\\ \|y\|_\infty\le 1}} \sup_{z\in E^*} x^T \nabla^2 W(z) y,$$

where the second-derivative matrix $\nabla^2 W(z)$ has entries

$$\frac{\partial^2 W(z)}{\partial z_i \partial z_j} = \frac{\lambda}{\beta}\left(\frac{e^{-z_i/\beta}\delta_{ij}}{\sum\limits_k e^{-z_k/\beta}} - \frac{e^{-z_j/\beta}e^{-z_i/\beta}}{\left(\sum\limits_k e^{-z_k/\beta}\right)^2}\right).$$

Here δ_{ij} stands for the Kronecker symbol. Denote $a_i = e^{-z_i/\beta}/\sum\limits_k e^{-z_k/\beta}$, which are evidently positive, with $\sum\limits_i a_i = 1$. Now,

$$\begin{aligned}\frac{\beta}{\lambda} x^T \nabla^2 W(z) y &= \sum_i x_i y_i a_i - \sum_i a_i x_i \sum_j a_j y_j = \sum_i x_i a_i \left(y_i - \sum_j a_j y_j\right)\\ &= \sum_i x_i a_i \sum_{j\neq i} a_j (y_i - y_j) \le \sum_i \sum_{j\neq i} a_i a_j |y_i - y_j|. \end{aligned} \tag{A.1}$$

Finally, the latter sum is bounded by 1 for any $|y_i| \le 1$ and $a_i \ge 0$, $\sum\limits_i a_i = 1$. To see this, note that the maximum of the convex (in $y \in \mathbb{R}^M$) function on the right-hand side of inequality (A.1) on the convex set $\{y \in \mathbb{R}^M : \|y\|_\infty \le 1\}$ is always attained at extreme points of the set, which are the vertices of the hypercube $\{y \in \mathbb{R}^M : y_i = \pm 1,\ i = 1, \dots, M\}$. Denote any of such extreme points by $y^* = (y_1^*, \dots, y_M^*)^T$. Let us split the index set $\{1, \dots, M\}$ into $I_+ = \{i : y_i^* = 1\}$ and $I_- = \{i : y_i^* = -1\}$. Then the maximal value of the sum can be decomposed, and we get

$$\frac{\beta}{\lambda} L \le 2 \sum_{i\in I_+, j\in I_-} a_i a_j + 2 \sum_{j\in I_+, i\in I_-} a_i a_j = 4 \sum_{i\in I_+} a_i \sum_{j\in I_-} a_j = 4a_+(1 - a_+) \le 1,$$

where $a_+ = \sum\limits_{i=I_+} a_i$. Hence, $\alpha = 1/\lambda$, which is independent of β.

REFERENCES

1. Schapire, R.E., The Strength of Weak Learnability, *Machine Learning*, 1990, vol. 5, no. 2, pp. 197–227.
2. Freund, Y., Boosting a Weak Learning Algorithm by Majority, *Inform. Comput.*, 1995, vol. 121, no. 2, pp. 256–285.
3. Schapire, R.E., Freund, Y., Bartlett, P.L., and Lee, W.S., Boosting the Margin: a New Explanation for the Effectiveness of Voting Methods, *Ann. Statist.*, 1998, vol. 26, no. 5, pp. 1651–1686.
4. Vapnik, V.N., *Statistical Learning Theory*, New York: Wiley, 1998.
5. Bartlett, P.L, Jordan, M.I., and McAuliffe, J.D., Convexity, Classification, and Risk Bounds, *Tech. Report of Dept. Statist., Univ. of California*, Berkeley, 2003, no. 638.
6. Lugosi, G. and Vayatis, N., On the Bayes-Risk Consistency of Regularized Boosting Methods (With Discussion), *Ann. Statist.*, 2004, vol. 32, no. 1, pp. 30–55.

7. Scovel, J.C. and Steinwart, I., Fast Rates for Support Vector Machines, *Los Alamos National Lab. Tech. Report*, 2003, no. LA-UR-03-9117.

8. Zhang, T., Statistical Behavior and Consistency of Classification Methods Based on Convex Risk Minimization (With Discussion), *Ann. Statist.*, 2004, vol. 32, no. 1, pp. 56–85.

9. Tsypkin, Ya.Z., *Osnovy teorii obychayushchikhsya sistem*, Moscow: Nauka, 1970. Translated under the title *Foundations of the Theory of Learning Systems*, New York: Academic, 1973.

10. Aizerman, M.A., Braverman, E.M., and Rozonoer, L.I., *Metod potential'nykh funktsyi v teorii obycheniya mashin* (Method of Potential Functions in the Theory of Learning Machines), Moscow: Nauka, 1970.

11. Aizerman, M., Braverman, E., and Rozonoer, L., Extrapolative Problems in Automatic Control and the Method of Potential Functions, *Am. Math. Soc. Transl.*, 1970, vol. 87, pp. 281–303.

12. Devroye, L., Györfi, L., and Lugosi, G., *A Probabilistic Theory of Pattern Recognition*, New York: Springer, 1996.

13. Cesa-Bianchi, N., Conconi, A., and Gentile, C., A Second-Order Perceptron Algorithm, *SIAM J. Comput.*, 2005, vol. 34, no. 3, pp. 640–668.

14. Kivinen, J., Smola, A.J., and Williamson, R.C., Online Learning with Kernels, *IEEE Trans. Signal Process.*, 2004, vol. 52, no. 8, pp. 2165–2176.

15. Zhang, T., Solving Large Scale Linear Prediction Problems Using Stochastic Gradient Descent Algorithms, in *Proc. 21st Int. Conf. on Machine Learning, Banff, Alberta, Canada, 2004 (ICML'04)*, New York: ACM, 2004, vol. 69, p. 116.

16. Polyak, B.T. and Juditsky, A.B., Acceleration of Stochastic Approximation by Averaging, *SIAM J. Control Optim.*, 1992, vol. 30, no. 4, pp. 838–855.

17. Nemirovskii, A.S. and Yudin, D.B., *Slozhnost' zadach i effektivnost' metodov optimizatsii*, Moscow: Nauka, 1979. Translated under the title *Problem Complexity and Method Efficiency in Optimization*, Chichester: Wiley, 1983.

18. Ben-Tal, A., Margalit, T., and Nemirovski, A., The Ordered Subsets Mirror Descent Optimization Method with Applications to Tomography, *SIAM J. Optim.*, 2001, vol. 12, no. 1, pp. 79–108.

19. Ben-Tal, A. and Nemirovski, A., The Conjugate Barrier Mirror Descent Method for Non-Smooth Convex Optimization, *Preprint of the Faculty of Industr. Eng. Manag., Technion – Israel Inst. Technol.*, Haifa, 1999. Available at `http://iew3.technion.ac.il/Labs/Opt/opt/Pap/CP_MD.pdf`.

20. Kivinen, J. and Warmuth, M.K., Additive Versus Exponentiated Gradient Updates for Linear Prediction, *Inform. Comput.*, 1997, vol. 132, no. 1, pp. 1–64.

21. Helmbold, D.P., Kivinen, J., and Warmuth, M.K., Relative Loss Bounds for Single Neurons, *IEEE Trans. Neural Networks*, 1999, vol. 10, no. 6, pp. 1291–1304.

22. Kivinen, J. and Warmuth, M.K., Relative Loss Bounds for Multidimensional Regression Problems, *Machine Learning*, 2001, vol. 45, no. 3, pp. 301–329.

23. Cesa-Bianchi, N. and Gentile, C., Improved Risk Tail Bounds for On-Line Algorithms, *Neural Information Processing Systems, NIPS 2004 Workshop on (Ab)Use of Bounds*, Whistler, BC, Canada, December 18, 2004. Available at `http://mercurio.srv.dsi.unimi.it/~cesabian/Pubblicazioni/iada.pdf`.

24. Cesa-Bianchi, N., Conconi, A., and Gentile, C., On the Generalization Ability of On-Line Learning Algorithms, *IEEE Trans. Inform. Theory*, 2004, vol. 50, no. 9, pp. 2050–2057.

25. Juditsky, A. and Nemirovski, A., Functional Aggregation for Nonparametric Estimation, *Ann. Statist.*, 2000, vol. 28, no. 3, pp. 681–712.

26. Tsybakov, A., Optimal Rates of Aggregation, *Computational Learning Theory and Kernel Machines*, Scholkopf, B. and Warmuth, M., Eds., Lecture Notes in Artificial Intelligence, Heidelberg: Springer, 2003, vol. 2777, pp. 303–313.

27. Vapnik, V. and Chervonenkis, A., *Teoriya raspoznavaniya obrazov*, Moscow: Nauka, 1974. Translated under the title *Theorie der Zeichenerkennung*, Berlin: Akademie-Verlag, 1979.

28. Breiman, L., Arcing the Edge, *Tech. Rep. of Statist. Dept., Univ. of California*, Berkeley, 1997, no. 486.

29. Friedman, J., Hastie, T., and Tibshirani, R., Additive Logistic Regression: a Statistical View of Boosting (With Discussion and a Rejoinder by the Authors), *Ann. Statist.*, 2000, vol. 28, no. 2, pp. 337–407.

30. Tsybakov, A., Optimal Aggregation of Classifiers in Statistical Learning, *Ann. Statist.*, 2004, vol. 32, no. 1, pp. 135–166.

31. Tarigan, B. and van de Geer, S.A., Adaptivity of Support Vector Machines with ℓ_1 Penalty, *Tech. Rep. of Math. Inst., Univ. of Leiden*, Leiden, 2004, no. MI 2004-14. Available at `http://www.math.leidenuniv.nl/~geer/svm4.pdf`.

32. Rockafellar, R.T. and Wets, R.J.B., *Variational Analysis*, New York: Springer, 1998.

33. Kiwiel, K.C., Proximal Minimization Methods with Generalized Bregman Functions, *SIAM J. Control Optim.*, 1997, vol. 35, no. 4, pp. 1142–1168.

34. Beck, A. and Teboulle, M., Mirror Descent and Nonlinear Projected Subgradient Methods for Convex Optimization, *Oper. Research Letters*, 2003, vol. 31, no. 3, pp. 167–175.

35. Polyak, B.T. and Tsypkin, Ya.Z., Criterial Algorithms of Stochastic Optimization, *Avtom. Telemekh.*, 1984, no. 6. pp. 95–104 [*Autom. Remote Contr.* (Engl. Transl.), vol. 45, no. 6, part 2, pp. 766–774].

36. Vajda, I., *Theory of Statistical Inference and Information*, Dordrecht: Kluwer, 1986.