Sequential Greedy Approximation for Certain Convex Optimization Problems

Tong Zhang

Abstract— **A greedy algorithm for a class of convex optimization problems is presented in this paper. The algorithm is motivated from function approximation using sparse combination of basis functions as well as some of its variants. We derive a bound on the rate of approximate minimization for this algorithm, and present examples of its application. Our analysis generalizes a number of earlier studies.**

Keywords— **Boosting, Incremental algorithms, Greedy approximation, Rate of approximate optimization, Sparse approximation.**

I. Introduction

The goal of many engineering applications is to estimate a solution that satisfies a certain optimality criterion. Furthermore, one would often like to obtain an approximation of the solution using a sparse combination of some basic solutions.

An example of this is to estimate a functional relationship between an input vector and its associated output. In statistical estimation, this is typically done by finding a function that optimizes a certain criterion from the data. In practice, it is also often desirable to find the simplest model that can explain the data. This is because simple models are often easier to understand and can have significant computational advantages over more complicated models. In addition, the philosophy of Occam's Razor implies that the simplest solution is likely to be the best solution among all possible solutions.

In this paper, we are interested in composite models that can be expressed as linear combinations of basic models. In this framework, it is natural to measure the simplicity of a composite model by the number of its basic model components. Since a composite model in our framework corresponds to a linear weight over the basic model space, our measurement of model simplicity corresponds to the sparsity of the linear weight representation.

Specifically we are interested in achieving sparsity through a greedy optimization algorithm which we propose later in the paper. This algorithm is closely related to a number of previous works. The basic idea was originated in [1], where Jones observed that if a target vector in a Hilbert space is a convex combination of a library of basic vectors, then using greedy approximation, one can achieve an error rate of $O(1/k)$ with k basic library vectors. The idea has been refined in [2] to analyze the approximation property of sigmoidal functions including neural networks. The same rate of approximation has also been derived earlier by Maurey using a probabilistic argument [3] (also see [2]).

T. Zhang is with IBM T.J. Watson Research Center, Yorktown Heights, NY, USA. E-mail: tzhang@watson.ibm.com.

The above methods can be regarded as greedy sparse algorithms for functional approximation, which is the noise-free case of regression problems. A similar greedy algorithm can also be used to solve general regression problems under noisy conditions [4]. In addition to regression, greedy approximation can also be applied to classification problems. The resulting algorithm is closely related to boosting [5] under the greedy function approximation point of view [6], [7], [8], [9]. This paper shows how to generalize the averaging technique in [1], [2] for analyzing greedy algorithms (in their case, for functional approximation problems) and apply it to boosting. Detailed analysis will be given in Section V. Specifically, our main theorem is an extension of the construction by Maurey, Jones and Barron to a class of convex optimization problems and that the same rate as in their works is obtained when the functional takes on the form $f(\cdot) = \|\cdot - v_*\|^2$ for some $v_* \in V$, where V is a Hilbert space with norm $\|\cdot\|$.

The analysis introduced in this paper can be compared with previous approaches for analyzing the boosting procedure (for example, see [10] and references therein) that are based on standard convergence analysis of numerical algorithms. Although such an analysis can lead to faster convergence rates in certain scenarios, it requires that the condition number[1] of the Hessian operator associated with the optimization problem[2] to be well-bounded. This assumption can be easily violated in practice. For example, the analysis is not suitable for a problem that contains an infinite number of basic models. Even for a problem with a finite number of basic models, such an analysis can fail simply because the Hessian is not invertible (for example, when a composite model can be represented as linear combinations of the basic models in more than one ways) or is ill-conditioned. Compared with those approaches, our analysis does not depend on the condition number of the associated parametric estimation problem. This advantage can be very important since many statistical estimation problems are known to be ill-conditioned. Consequently, the original analysis in [1], [2] has been successfully applied to problems that cannot be handled by the traditional condition number based convergence analysis for numerical iterative procedures.

The proposed greedy approximation method can also be applied to other prediction problems with different loss functions. For example, in density estimation, the goal is to

[1] This for example can be measured by the ratio of the largest and the smallest eigenvalues of the Hessian.

[2] In their case the problem should be written as a parametric estimation problem where the parameters are components of the linear weight representation.

find a model that has the smallest negative log-likelihood. Greedy algorithms have been studied in [11], [12]. Similar approximation bounds can be directly obtained under the general framework proposed in this paper.

We proceed as follows. Section II formalizes the general class of problems considered in this paper, and proposes a greedy algorithm to solve the formulation. The convergence rate of the algorithm is investigated in Section III and Section IV. In Section V we apply our analysis to a few concrete examples. Some concluding remarks are given in Section VI.

II. A General Greedy Approximation Algorithm

Let V be a linear vector space, and consider a subset $S \subset V$. Note that in the examples studied in this paper, V can be thought as a functional space. For any positive integer k, denote by $\mathrm{co}_k(S)$ the convex hull of S spanned by k elements in S:

$$\mathrm{co}_k(S) = \left\{ \sum_{j=1}^{k} \alpha_j u_j : \alpha_j \geq 0, \sum_{j=1}^{k} \alpha_j = 1, u_j \in S \right\}.$$

Define the convex hull of S as:

$$\begin{aligned} \mathrm{co}(S) &= \bigcup_{k \geq 1} \mathrm{co}_k(S) \\ &= \left\{ \sum_{j=1}^{k} \alpha_j u_j : \alpha_j \geq 0, \sum_{j=1}^{k} \alpha_j = 1, u_j \in S, k \in Z^+ \right\}, \end{aligned} \quad (1)$$

where we use Z^+ to denote the set of positive integers.

We consider the following optimization problem on $\mathrm{co}(S)$:

$$\inf_{v \in \mathrm{co}(S)} f(v), \quad (2)$$

where we assume that f is a real valued convex function. Although in convex analysis, a convex function f is allowed to take the $+\infty$ value, in this paper we require that $f(v) < +\infty$ for all $v \in \mathrm{co}(S)$. Throughout the paper, we also assume that the optimal value in (2) satisfies $\inf_{v \in \mathrm{co}(S)} f(v) > -\infty$: the condition implies that the quantity $\Delta f(v')$ given in (3) below is well-defined.

It is important to mention that we do not require that a solution of (2) exists. However, the value $\inf_{v \in \mathrm{co}(S)} f(v)$ can be arbitrarily approximated with a vector $v' \in \mathrm{co}(S)$. The quality of an approximate solution v' of (2) can be naturally measured by the following quantity:

$$\Delta f(v') = f(v') - \inf_{v \in \mathrm{co}(S)} f(v). \quad (3)$$

Therefore from the approximation point of view, one would like to find a sequence of vectors v^k such that $\Delta f(v^k) \to 0$.

In this paper, we investigate the convergence behavior of the following incremental algorithm that approximately solves (2). The algorithm also has the property that for each k, v^k can be expressed as the convex combination of $k+1$ basis vectors in S: $v^k \in \mathrm{co}_{k+1}(S)$. This means that the method leads to a sparse approximate solution of (2) when v^k is expressed as a sparse linear combination of basis vectors in S.

Algorithm II.1: (Sequential greedy approximation)
Given $v^0 \in S$
for $k = 1, 2, \ldots, N$
Find $\bar{u}_k \in S$ and $0 \leq \bar{\alpha}_k \leq 1$ to approximately minimize the function:
$(\alpha_k, u_k) \to f((1-\alpha_k)v^{k-1} + \alpha_k u_k)$ $(*)$
Let $v^k = (1-\bar{\alpha}_k)v^{k-1} + \bar{\alpha}_k \bar{u}_k$
end

Remark II.1: The approximate minimization of $(*)$ in Algorithm II.1 should be interpreted as finding $\bar{u}_k \in S$ and $0 \leq \bar{\alpha}_k \leq 1$ such that

$$\begin{aligned} &f((1-\bar{\alpha}_k)v^{k-1} + \bar{\alpha}_k \bar{u}_k) \\ &\leq \inf_{u_k, \alpha_k} f((1-\alpha_k)v^{k-1} + \alpha_k u_k) + \epsilon_{k-1}, \end{aligned} \quad (4)$$

where $\epsilon_{k-1} \geq 0$ is a sequence of non-negative numbers that converges to zero.

Remark II.2: The number N is a prefixed number of iterations at which the algorithm stops. For notational convenience, in the following, we shall just let $N = +\infty$, and study the behavior of the algorithm after any specific number of iterations k.

In Algorithm II.1, we are looking for a minimizing sequence $v^k \in \mathrm{co}_{k+1}(S)$ for f over $\mathrm{co}(S)$. In this paper, we show that under appropriate regularity conditions, $\Delta f(v^k) \to 0$ as $k \to +\infty$, where v^k is computed from Algorithm II.1. In addition, a convergence rate of $O(1/k)$ can be obtained in many cases. Specifically, a simplified form of the main result (see Theorem IV.2) obtained in this paper can be stated as follows:

Theorem II.1: Assume f is differentiable and

$$\sup_{v,w \in \mathrm{co}(S), \theta \in (0,1)} \frac{d^2}{d\theta^2} f((1-\theta)v + \theta w) \leq M < +\infty.$$

Assume further that the optimization in (4) can be performed exactly for all $k \geq 1$ (that is, $\epsilon_{k-1} = 0$), then $\Delta f(v^k) \leq 2M/(k+2)$.

Clearly our bound depends on the second derivative of the objective function in one-dimensional subspaces of V. It leads to a relatively slow convergence rate of the order $O(1/k)$. As a comparison, the analysis in [10] predicts a faster convergence rate of the form $\Delta f(v^k) \leq (1-1/\eta)^k \Delta f(v^0)$. However the parameter η depends heavily on properties of the basis set S. For example, one has an expression $\eta = O(|S|^4)$ ($|S|$ denotes the cardinality of S) and the constant in the $O(\cdot)$ notation has other dependencies on S (such as a properly defined condition number).

Observe that our problem formulation does not depend on any topological structure on the vector space V. In fact, our analysis does not require any specific topology of the

underlying space V. This treatment makes the analysis more general.

However, in many practical applications, a natural topological structure may exist so that the functional f is continuous. In this case, (2) can be equivalently replaced by the following problem

$$\inf_{v\in\overline{\mathrm{co}}(S)} f(v), \tag{5}$$

where we use $\overline{\mathrm{co}}(S)$ to denote the closure of $\mathrm{co}(S)$ with respect to the underlying topology. If V is a topological linear vector space, then $\overline{\mathrm{co}}(S)$ can also be written as the closure of

$$\{v\in V : v=\sum_{j=1}^{\infty}\alpha_j u_j, \alpha_j\geq 0, \sum_{j=1}^{\infty}\alpha_j = 1, u_j\in S\}.$$

That is, we can replace finite summations in the definition of $\mathrm{co}(S)$ in (1) by convergent (with respect to the underlying topology) infinite linear combinations over countably many elements in S.

For the sake of clarity, we do not further discuss the case with topology in the paper except when it is necessary. However it is easy to see that results in this paper can be trivially generalized to include topology.

III. One Step Convergence Analysis

We would like to study the one step convergence of Algorithm II.1. In particular, for all $v\in\mathrm{co}(S)$, we are interested in an upper bound for

$$f^+(v) = \inf_{\eta\in[0,1],u\in S} f((1-\eta)v+\eta u), \tag{6}$$

which clearly determines the convergence of the $(*)$ step in Algorithm II.1.

The analysis used in this paper is based on the following idea. Consider an arbitrary $w=\sum_{i=1}^m \alpha_i u_i\in\mathrm{co}(S)$, where $\alpha_i\geq 0$ and $\sum_{i=1}^m\alpha_i=1$. We consider the following quantity for all $\eta\in[0,1]$:

$$g(v,w,\eta) = \sum_{i=1}^m \alpha_i f((1-\eta)v+\eta u_i). \tag{7}$$

Note that with a little abuse of notation, we have used $w\in\mathrm{co}(S)$ to denote a representation $w=\sum_{i=1}^m\alpha_i u_i$. Although such a representation may not be unique, we may simply choose any one of them in our analysis.

From the definition of $g(v,w,\eta)$, it is clear that

$$f^+(v)\leq \inf_{w,\eta} g(v,w,\eta). \tag{8}$$

It follows that an upper bound for $g(v,w,\eta)$ can be used to obtain an upper bound for $f^+(v)$. In addition, bounds for $g(v,w,\eta)$ can be easily obtained under some regularity conditions of f.

To further analyze the quantity g, some smoothness properties of f are necessary. Let f be a convex function defined on $\mathrm{co}(S)\subset V$. Denote by $\mathrm{span}(S)\subset V$ the linear vector space spanned by S. Denote by $\mathrm{span}(S)'$ the dual space of $\mathrm{span}(S)$ (that is, the space of real valued linear functions on $\mathrm{span}(S)$). We say that f is differentiable with gradient $\nabla f\in\mathrm{span}(S)'$ if it satisfies the following Fréchet-like differentiability condition for all $v,w\in\mathrm{co}(S)$:

$$\lim_{h\to 0^+}\frac{1}{h}(f((1-h)v+hw)-f(v)) = \nabla f(v)^T(w-v),$$

where $\nabla f(v)^T(w-v)$ the value of linear functional $\nabla f(v)$ at $w-v$. Note that we use the notation $v_1^Tv_2$ from linear algebra, where it is just the scalar product of the two vectors.

It is easy to see from the definition of convexity that the following quantity, sometimes referred to as the Bregman divergence of f, is non-negative for all $v,w\in\mathrm{co}(S)$:

$$d_f(v,w) = f(w)-f(v)-\nabla f(v)^T(w-v)\geq 0. \tag{9}$$

This inequality plays a very important role in our analysis. Indeed, this is where the convexity assumption of f is used. We shall now rewrite (7) as:

$$\begin{aligned}
g(v,w,\eta) =& \sum_{i=1}^m \alpha_i f((1-\eta)v+\eta u_i)\\
=& f(v)+\sum_{i=1}^m\alpha_i\nabla f(v)^T(\eta u_i-\eta v)+\\
& \sum_{i=1}^m\alpha_i d_f(v,(1-\eta)v+\eta u_i)\\
=& f(v)-\eta(f(v)-f(w)+d_f(v,w))+\\
& \sum_{i=1}^m\alpha_i d_f(v,(1-\eta)v+\eta u_i) \qquad (10)\\
\leq& f(v)-\eta(f(v)-f(w)+d_f(v,w))+\\
& \sup_{u\in S} d_f(v,(1-\eta)v+\eta u). \qquad (11)
\end{aligned}$$

It can be seen that equation (11) is in a very attractive form. Observe that it does not depend on the representation $\sum_{i=1}^m\alpha_i u_i$ of w any more. This is one of the reasons why we do not explicitly put α_i and u_i in the definition of $g(v,w,\eta)$. The second term is first-order in η. The third term is usually second order in η under appropriate regularity conditions on f.

In order to obtain an estimate for the third term of (11), we need to impose second order differentiability condition on f: for all $v,w\in\mathrm{co}(S)$, let

$$f_{v,w}(h) = f((1-h)v+hw)$$

be a function defined on $[0,1]$. We assume that f is second order differentiable in the sense that the second derivative of $f_{v,w}(h)$ exists for all $v,w\in\mathrm{co}(S)$ and $h\in(0,1)$. Using Taylor expansion, it is easy to verify that the following bound holds:

$$\sup_{u\in S} d_f(v,(1-\eta)v+\eta u)\leq\frac{\eta^2}{2}\sup_{u\in S,\theta\in(0,\eta)} f''_{v,u}(\theta). \tag{12}$$

This gives the following one step bound.

Lemma III.1: Let f be a second order differentiable convex function. For all $v \in \mathrm{co}(S)$, consider $A(v)$ and $M(v)$ such that

$$0 \leq A(v) \leq \sup_{w \in \mathrm{co}(S)} (f(v) - f(w) + d_f(v,w)),$$

and

$$M(v) \geq \sup_{u \in S, \theta \in (0,1)} f''_{v,u}(\theta).$$

Then we have

$$f^+(v) \leq \begin{cases} f(v) - \frac{A(v)^2}{2M(v)} & \text{if } A(v) < M(v), \\ f(v) - A(v) + \frac{M(v)}{2} & \text{otherwise.} \end{cases}$$

Proof. Combine (11) and (12), we obtain

$$g(v,w,\eta) \leq f(v) - \eta(f(v) - f(w) + d_f(v,w)) + \frac{\eta^2}{2} \sup_{u \in (S), \theta \in (0,\eta)} f''_{v,u}(\theta).$$

Now, using (8), we have

$$f^+(v) \leq \inf_{w,\eta} g(v,w,\eta) \leq \inf_{\eta} \left[f(v) - \eta A(v) + \frac{\eta^2}{2} M(v) \right].$$

Now, by choosing $\eta = \min(1, A(v)/M(v))$ on the right hand side, we obtain the desired bound. □

The following lemma gives estimates of $A(v)$ using $\Delta f(v)$.

Lemma III.2: Recall $\Delta f(v')$ defined in (3). For all $v \in \mathrm{co}(S)$, we have

$$\sup_{w \in \mathrm{co}(S)} (f(v) - f(w) + d_f(v,w)) \geq \Delta f(v).$$

Furthermore, if $\inf_{v \in \mathrm{co}(S)} f(v)$ can be achieved at an algebraic interior point $v_* \in \mathrm{co}(S)$,[3] then we have the following bound:

$$\sup_{w \in \mathrm{co}(S)} (f(v) - f(w) + d_f(v,w)) \geq \left(1 + \frac{d_f(v,v_*)}{d_f(v_*,v)}\right) \Delta f(v).$$

Proof. Since $d_f(v,w) \geq 0$, we have

$$\sup_w (f(v) - f(w) + d_f(v,w)) \geq \sup_w (f(v) - f(w)) = \Delta f(v).$$

To prove the second part of the lemma, we first observe that the function $f((1-h)v_* + hv)$ of h achieves the minimum at the algebraic interior point $h = 0$, which implies that the derivative $\nabla f(v_*)^T(v - v_*) = 0$. This further implies that $d_f(v_*, v) = f(v) - f(v_*)$. Therefore

$$\begin{aligned} &\sup_w (f(v) - f(w) + d_f(v,w)) \\ \geq & f(v) - f(v_*) + d_f(v, v_*) \\ = & f(v) - f(v_*) + \frac{d_f(v,v_*)}{d_f(v_*,v)} (f(v) - f(v_*)). \end{aligned}$$

[3]We say that v_* is an algebraic interior point of $\mathrm{co}(S)$ if for all $w \in \mathrm{co}(S)$, there exists $h < 0$ such that $(1-h)v_* + hw \in \mathrm{co}(S)$. This assumption can also be replaced by the assumption $\nabla f(v_*) = 0$.

□

Remark III.1: Note that the algebraic interior concept does not require any topology. Therefore it is not an interior point in the topological sense. It has similarities with the concept of relative algebraic interior (i.e. interior with respect to the core topology; see, e.g., [13]), which plays a basic role in convex analysis and properties of related objects such as Minkowski functional. See [13] for further details.

Remark III.2: The technique introduced in this paper relies on the assumption that the objective function f is (Fréchet-like) differentiable. This assumption is crucial in the proof, and it is unclear how to relax it. Although we assume second order differentiability in Lemma III.1, one may still apply the same technique as long as $d_f(v, (1-\eta)v + \eta u)$ is super-linear in η — this is almost always true when f is continuously differentiable. For example, if we assume that $M_t(v)$ is such that

$$\sup_{u \in S} d_f(v, (1-\eta)v + \eta u) \leq \frac{\eta^t}{t} M_t(v), \tag{13}$$

where $t \in (1, 2]$, then similar to Lemma III.1 we have

$$f^+(v) \leq \begin{cases} f(v) - \frac{A(v)^s}{s M_t(v)^{s-1}} & \text{if } A(v) < M_t(v), \\ f(v) - A(v) + \frac{M_t(v)}{t} & \text{otherwise,} \end{cases} \tag{14}$$

where $1/t + 1/s = 1$.

Remark III.3: The analysis can be easily generalized to a non-convex objective function f that satisfies the following condition with a parameter $\kappa > 0$:

$$\forall f(w) \leq f(v) : \nabla f(v)^T (v - w) \geq \kappa (f(v) - f(w)).$$

We do not consider such functions in this paper since they do not appear to have much practical significance.

The $O(1/k)$ order convergence rate of Algorithm II.1 obtained from our analysis using (11) is usually tight in the worst case (for special cases, such as when S is finite, the $O(1/k)$ rate can clearly be improved). However, the associated constant may not be optimal. To understand what causes this sub-optimality, we consider the objective function $f(v) = \|v - v_*\|^2$ where V is an Hilbert space with norm $\|\cdot\|$. Assume also S is a subset of $\{v \in V : \|v\| \leq A\}$. It is easy to see that

$$d_f(v,w) = \|w - v_*\|^2 - \|v - v_*\|^2 - 2(v - v_*)^T(w - v) = \|w - v\|^2,$$

which is independent of v_*. Using this expression, we can rewrite (10) and (11) as:

$$\begin{aligned} g(v,w,\eta) = & f(v) - \eta(f(v) - f(w) + d_f(v,w)) \\ & + \eta^2 \sum_{i=1}^m \alpha_i \|v - u_i\|^2 \qquad (15) \\ \leq & f(v) - \eta(f(v) - f(w) + d_f(v,w)) + \\ & \eta^2 \sup_{u \in S} \|v - u\|^2. \qquad (16) \end{aligned}$$

In our analysis, the third term of (16) is bounded by using (12), which leads to

$$\eta^2 \sup_{u\in S} \|v-u\|^2 \le 4A^2\eta^2. \tag{17}$$

However the third term of (15) can be bounded by

$$\begin{aligned}&\eta^2 \sum_{i=1}^m \alpha_i \|v-u_i\|^2\\ =&\eta^2[\|v-w\|^2 + \sum_{i=1}^m \alpha_i\|u_i\|^2 - \|w\|^2]\\ \le&\eta^2[\|v-w\|^2 + A^2 - \|w\|^2].\end{aligned} \tag{18}$$

Let $w^* \in \overline{\mathrm{co}}(S)$ achieve the minimum of $\inf_{w\in\overline{\mathrm{co}}(S)} f(w)$, then $\|v-w_*\|^2 \le f(v)-f(w_*)$. Note that we do not assume that $v_* \in \overline{\mathrm{co}}(S)$ and hence it is possible that $w^* \neq v_*$. Now by picking $w \to w_*$ when bounding $\inf_w g(v,w,\eta)$ as in the proof of Lemma III.1, (18) becomes

$$\eta^2(\Delta f(v) + A^2 - \|w_*\|^2). \tag{19}$$

If we pick $\epsilon_k \to 0^+$ in (4), then $\Delta f(v^k) \to 0$ (the general case is proved in Theorem IV.1). Therefore using (19), we can asymptotically reduce the quantity $M(v)$ in Lemma III.1 (given by (17)) by a factor of four.

It is clear that the above derivation and the associated conclusion holds for any quadratic objective function f. Furthermore, it also holds in a more general setting. For example, consider the case that V is a topological linear vector space, and there exists $w^* \in \overline{\mathrm{co}}(S)$ that achieves the minimum of $\inf_{w\in\overline{\mathrm{co}}(S)} f(w)$. We further assume that the Fréchet-like Hessian of f exists in the sense that as $h \to 0$, we have uniformly over v and $v+\Delta v \in \overline{\mathrm{co}}(S)$ that

$$\begin{aligned}&f(v+h\Delta v)\\ =&f(v) + h\nabla f(v)^T\Delta v + \frac{h^2}{2}\Delta v^T\nabla^2 f(v)\Delta v + o(h^2),\end{aligned} \tag{20}$$

and that the Hessian is uniformly continuous on $\overline{\mathrm{co}}(S)$. In addition we assume that the condition that $\Delta f(v) \to 0$ implies that $v \to w_*$. Then by Theorem IV.1, we have $w_* = \lim_{k\to\infty} v^k$. It follows from the same derivation of (18) that asymptotically we can improve (12) by the following bound:

$$\begin{aligned}&\frac{2}{\eta^2}\sup_{u\in S} d_f(v,(1-\eta)v+\eta u)\\ \le&\inf_{z\in V}\sup_{u\in S}[(u-z)^T\nabla^2 f(w_*)(u-z) -\\ &\quad (w_*-z)^T\nabla^2 f(w_*)(w_*-z)] + o(1).\end{aligned} \tag{21}$$

The above discussion indicates that we may lose a factor of four in our analysis. This factor of four difference occurs when we compare previous Hilbert space functional approximation results in [2], [4] to the corresponding bounds obtained from Theorem IV.2. The above discussion explains where this difference comes from.

IV. Rate of convergence

For simplicity, we shall only derive bounds for problems that satisfy the following additional assumption:

$$\sup_{v,w\in\mathrm{co}(S),\theta\in(0,1)} f''_{v,w}(\theta) \le M < +\infty, \tag{22}$$

where M is a non-negative number that will appear in the convergence bound. We also choose a non-negative number ρ such that

$$\rho \le \begin{cases}\inf_{v,w\in\mathrm{co}(S)} \frac{d_f(v,w)}{d_f(w,v)} & \text{if (2) has an algebraic}\\ & \text{interior solution,}\\ 0 & \text{otherwise.}\end{cases} \tag{23}$$

Lemma III.2 implies that we can choose $A(v) = (1+\rho)\Delta f(v)$ in Lemma III.1. Also let $M(v) = M$ we obtain

$$\begin{aligned}&\Delta f(v^{k+1}) \le\\ &\begin{cases}\epsilon_k + \Delta f(v^k) - \frac{(1+\rho)^2\Delta f(v^k)^2}{2M} & \text{if } (1+\rho)\Delta f(v^k) \le M,\\ \epsilon_k + \frac{M}{2} - \rho\Delta f(v^k) & \text{otherwise,}\end{cases}\end{aligned} \tag{24}$$

where $\epsilon_k \ge 0$ is the parameter in (4) that controls the approximate minimization of $(*)$. By choosing $\Delta f(v^k)$ to maximize the right hand side of (24), we obtain for all $k \ge 0$:

$$\Delta f(v^{k+1}) \le \epsilon_k + \frac{M}{2(1+\rho)^2}. \tag{25}$$

In order to obtain a good convergence rate for Algorithm II.1, it is necessary to choose ϵ_k that converges to zero sufficiently fast. Before doing so, we show that Algorithm II.1 converges as long as $\epsilon_k \to 0$.

Theorem IV.1: Assume that

$$M = \sup_{v\in\mathrm{co}(S),u\in S,\theta\in(0,1)} f''_{v,u}(\theta) < +\infty.$$

Then Algorithm II.1 converges: $\lim_{k\to+\infty}\Delta f(v^k) = 0$ as long as $\epsilon_k \to 0$ in (4).

Proof. $\forall 0 < \delta < M/2(1+\rho)^2$, there exists $k_0 > 0$ such that $\epsilon_k < \frac{\delta^2(1+\rho)^2}{2M} < \delta$ when $k \ge k_0$. Using (25), we obtain $\forall k > k_0$: $\Delta f(v^k) \le M/(1+\rho)^2$. Hence by (24), $\forall k > k_0$ we have

$$\Delta f(v^{k+1}) \le \Delta f(v^k) - (1+\rho)^2\frac{\Delta f(v^k)^2 - \delta^2}{2M}.$$

This implies that if $\Delta f(v^k) > 2\delta$, we decrease $\Delta f(v^k)$ by a positive number bounded away from zero. Since $\Delta f(v^k)$ is bounded below by zero, there must exist $k_1 > k_0$ such that $\Delta f(v^{k_1}) \le 2\delta < M/(1+\rho)^2$. This implies that $\forall k \ge k_1$:

$$\begin{aligned}\Delta f(v^{k+1}) \le&\Delta f(v^k) - (1+\rho)^2\frac{\Delta f(v^k)^2 - \delta^2}{2M}\\ \le&2\delta - (1+\rho)^2\frac{3\delta^2}{2M} < 2\delta.\end{aligned}$$

Note that in the above derivation we have used the fact that function $x \rightarrow x - (1+\rho)^2 x^2/2M$ is increasing on $[0, M/(1+\rho)^2]$. □

The above theorem shows that the convergence can be achieved as long as $\epsilon_k \rightarrow 0$. The following theorem gives a bound on the rate of convergence when we choose ϵ_k that converges to zero sufficiently fast.

Theorem IV.2: Assume that M and ρ are given by (22) and (23) respectively. Assume also that we pick ϵ_k in (4) such that $\epsilon_{k-1} \leq c/(k+1)^2$, where $(1+\rho)^2 c \leq M/4$. Let

$$c' = \frac{2M}{(1+\rho)^2} + \frac{2c}{1+\sqrt{1+\frac{2c}{M}(1+\rho)^2}},$$

then $\forall k \geq 1$, we have the following bound for v^k obtained by Algorithm II.1:

$$\Delta f(v^k) \leq \frac{c'}{k+2}.$$

Proof. At $k=1$, from (25) we obtain an estimate

$$\Delta f(v^1) \leq \frac{c}{4} + \frac{M}{2(1+\rho)^2} < \frac{c'}{1+2} \leq \frac{M}{(1+\rho)^2}.$$

Now, we assume as an induction hypothesis that for some $k \geq 1$:

$$\Delta f(v^k) \leq \frac{c'}{k+2} \leq \frac{M}{(1+\rho)^2}.$$

From (24), we obtain

$$\Delta f(v^{k+1}) \leq \epsilon_k + \Delta f(v^k) - \frac{(1+\rho)^2 \Delta f(v^k)^2}{2M}.$$

Observe that function $x \rightarrow x - (1+\rho)^2 x^2/2M$ is increasing on $[0, M/(1+\rho)^2]$, and $\Delta f(v^k) \leq \frac{c'}{k+2} \leq \frac{M}{(1+\rho)^2}$, we have

$$\begin{aligned}\Delta f(v^{k+1}) \leq & \frac{c}{(k+2)^2} + \frac{c'}{k+2} - \frac{(1+\rho)^2 c'^2}{2M(k+2)^2} \\ = & \frac{c'}{k+2} - \frac{c'}{(k+2)^2} \leq \frac{c'}{k+3}.\end{aligned}$$

Note that in the above derivation, we used the equality

$$c - \frac{(1+\rho)^2}{2M} c'^2 = -c',$$

which can be easily verified by solving the above quadratic equation in c'. □

As being pointed out at the end of Section III, the constant in Theorem IV.2 expressed in term of M can be improved in some cases. For example, if we further assume that (20) holds and that $\Delta f(v) \rightarrow 0$ implies that $v \rightarrow w_*$ (where w_* achieves the minimum of $f(v)$ on $\overline{\text{co}}(S)$), then asymptotically we have a bound $\Delta f(v^k) \leq c'/k + o(1/k)$ as in Theorem IV.2 but with M in the definition of c' replaced by

$$\begin{aligned}\inf_{z \in V} \sup_{u \in S} & [(u-z)^T \nabla^2 f(w_*)(u-z) - \\ & (w_*-z)^T \nabla^2 f(w_*)(w_*-z)] + o(1).\end{aligned}$$

In addition, if (2) has an algebraic interior solution, then asymptotically the quantity ρ in the definition of c' can be replaced by 1.

V. Examples

In this section, we discuss some applications of Algorithm II.1. We show that the general formulation considered in this paper includes a number of previous formulations as special cases. We will also compare our results with similar results in the literature.

A. Function approximation in Hilbert spaces

The technique used in this paper is motivated from the greedy version of Maurey's approximation result in [1], [2], where the following problem of function approximation in Hilbert spaces was considered: find $v \in \text{co}(S)$ such that

$$f(v) = \|v - v_*\|^2$$

is as small as possible where we assume that V is a Hilbert space, and S is a subset of $\{v \in V : \|v\| \leq A\}$. Note that this problem has been briefly studied at the end of Section III.

In this case we have $f''_{v,w}(\theta) = 2\|v-w\|^2$. It is easy to verify that (22) and (23) becomes

$$M = \sup_{v,w \in \text{co}(S)} 2\|v-w\|^2 \leq 8A^2,$$

and

$$\rho = \begin{cases} 1 & \text{if (2) has an algebraic interior solution,} \\ 0 & \text{otherwise.} \end{cases}$$

It follows that the constant c' in Theorem IV.2 is given by (assuming $\epsilon_k = 0$):

$$c' = \begin{cases} 4A^2 & \text{if (2) has an algebraic interior solution,} \\ 16A^2 & \text{otherwise.} \end{cases}$$

The scenario that (2) has an algebraic interior solution has the same effect of the assumption that $v_* \in \overline{\text{co}}(S)$, which was the original problem considered in [1], [2]. To see that this is true, we just observe that if $v_* \in \overline{\text{co}}(S)$, then v_* is the unique solution of (2) in $\overline{\text{co}}(S)$ and $\nabla f(v_*) = 0$. As pointed out in Footnote 3, in the proof of Lemma III.2 this has the same effect as the assumption that (2) has an algebraic interior solution.

Under the assumption that $v_* \in \overline{\text{co}}(S)$, the approximation bound in [2] (which has a better constant than that of [1]) can be expressed asymptotically as $\Delta f(v^k) \sim A^2/k$ (assuming $\epsilon_k = o(1/k^2)$ for simplicity), which is a factor of four better than the bound given by Theorem IV.2. As pointed out at the end of Section III, this difference is caused by the more general treatment following (11), which can be easily removed for this problem if we replace the bound in (17) by (18).

Approximation bound in [2] was later generalized in [4], where the assumption that $v_* \in \overline{\text{co}}(S)$ was removed.

The resulting bound can be expressed asymptotically as $\Delta f(v^k) \sim 4A^2/k$ (we again assume that $\epsilon_k = o(1/k^2)$ for simplicity). This bound can be compared with the corresponding bound obtained from IV.2: $\Delta f(v^k) \sim 16A^2/k$ which is again a factor of four worse in the constant. This difference can be easily removed using the derivation outlined at the end of Section III.

B. L_p Regression

Another extension of [1], [2] was considered in [14], where the authors studied function approximation in L_p space ($1 < p < \infty$).

In this section, we generalize it to be the following L_p regression problem. We would like to approximate an output signal $y = v_*(x)$ using a function $v(x)$ of the input so that the loss defined by

$$L(v(\cdot)) = \int_x |v_*(x) - v(x)|^p d\mu(x)$$

is small, where we use $\mu(x)$ to denote a non-negative measure on x. In this case, the linear vector space V is the L_p space:

$$V = \left\{ v = v(x) : \|v(x)\|_p = \left(\int v(x)^p d\mu(x) \right)^{1/p} < +\infty \right\}.$$

We also assume that the $v_* \in V$.

Similar to the case of Hilbert space, we consider the set of basis S to be a subset of $\{v(x) \in V : \|v(x)\|_p \le A\}$. The derivation in [14] relied on the assumption that $v_*(x) \in \overline{\text{co}}(S)$, which is not required in our analysis.

Instead of working with $L(v(x))$ directly, we define the following objective function in (2):

$$f(v) = \|v - v_*\|_p^q = L(v(\cdot))^{q/p}, \tag{26}$$

where $q = \min(p, 2)$, $v = v(\cdot)$ and $v_* = v_*(\cdot)$.

Clearly the minimization of $f(v)$ can be easily related to the minimization of L. It is also not difficult to check that $f(v)$ is a convex function of v. One way to see this is by noticing that norm $\|v\|_p$ is convex in v and thus f is a composite of a non-decreasing convex function and a convex function. It is simple to verify by the definition of convexity that such a composite function is convex.

Now, for all $v(x)$, $w(x)$ in $\text{co}(S)$ and $\theta \in [0, 1]$, we have

$$f_{v,w}(\theta) = \|(1-\theta)v(\cdot) + \theta w(\cdot) - v_*(\cdot)\|_p^q.$$

For notational simplicity, let $z(\theta, x) = (1-\theta)v(x) + \theta w(x) - v_*(x)$ and $\Delta v(x) = w(x) - v(x)$. Now differentiate with respect to θ, we obtain

$$\begin{aligned} & f'_{v,w}(\theta) \qquad (27) \\ =& q\|z(\theta,\cdot)\|_p^{q-p} \int_x |z(\theta,x)|^{p-1} \text{sign}(z(\theta,x)) \Delta v(x) d\mu(x). \end{aligned}$$

Differentiate with respect to θ again we have

$$\begin{aligned} f''_{v,w}(\theta) =& q(q-p)\|z(\theta,\cdot)\|_p^{q-2p} \qquad (28) \\ & \cdot \left[\int_x |z(\theta,x)|^{p-1} \text{sign}(z(\theta,x)) \Delta v(x) d\mu(x) \right]^2 \\ & + q(p-1)\|z(\theta,\cdot)\|_p^{q-p} \int_x |z(\theta,x)|^{p-2} \Delta v(x)^2 d\mu(x). \end{aligned}$$

In the following, we shall treat $p \ge 2$ and $p \le 2$ separately. Consider first that $p \ge 2$. In this case, we have $q = 2$ and thus:

$$\begin{aligned} f''_{v,w}(\theta) \le& 2(p-1)\|z(\theta,\cdot)\|_p^{2-p} \int_x |z(\theta,x)|^{p-2} \Delta v(x)^2 d\mu(x) \\ \le& 2(p-1)\|z(\theta,\cdot)\|_p^{2-p} \left(\int_x |z(\theta,x)|^p d\mu(x) \right)^{(p-2)/p} \\ & \left(\int_x |\Delta v(x)|^p d\mu(x) \right)^{2/p} \\ =& 2(p-1)\|\Delta v(x)\|_p^2 \le 8(p-1)A^2. \end{aligned}$$

In the above, the first inequality follows from (28), and the second inequality follows from the Hölder's inequality with the dual pair $(p/(p-2), p/2)$. It follows that we may assign

$$M = 8(p-1)A^2 \qquad \text{and} \qquad \rho = 0$$

in Theorem IV.2. Ignoring ϵ_k (say, by assuming $\epsilon_k = 0$) for simplicity, we obtain the following rate when $p \ge 2$ and $k \ge 1$:

$$\|v^k - v_*\|_p^2 \le \inf_{v \in \overline{\text{co}}(S)} \|v - v_*\|_p^2 + 16(p-1)\frac{A^2}{k+2}.$$

The rate matches that in [14] which assumed $v_* \in \overline{\text{co}}(S)$.

We now consider the case $1 < p \le 2$ with $q = p$ in (26). It is clear that f is not second order differentiable. However, using Taylor expansion we know that $\exists s \in (0, \eta)$:

$$d_f(v, (1-\eta)v + \eta w) = (f'_{v,w}(s) - f'_{v,w}(0))\eta.$$

Combined with (27) and the fact that $||z_1|^{p-1} - |z_2|^{p-1}| \le |z_1 - z_2|^{p-1}$, we have

$$d_f(v, (1-\eta)v + \eta w) \le p\|\Delta v(x)\|_p^p \eta^p \le p(2A)^p \eta^p.$$

Now, using (14) and the same induction as in the proof of Theorem IV.2, we can obtain a convergence rate of the order $\Delta f(v^k) = O(1/k^{p-1})$, which matches the rate proved in [14] (the assumption $v_* \in \overline{\text{co}}(S)$ in [14] can be removed in our analysis). We shall skip the detailed proof for simplicity.

C. Classification and boosting

Our analysis can also be applied to classification problems. For simplicity, we shall focus only on binary classification, where our goal is to predict a binary output $y \in \{\pm 1\}$ based on an observed input vector x. Given a

real-valued model $p(x)$, we consider the following discrete prediction rule:

$$y = \begin{cases} 1 & \text{if } p(x) \geq 0, \\ -1 & \text{if } p(x) < 0. \end{cases} \tag{29}$$

The classification error (for simplicity, we ignore the point $p(x) = 0$, which is assumed to occur rarely) is given by

$$L_\gamma(p(x), y) = \begin{cases} 1 & \text{if } p(x)y \leq \gamma, \\ 0 & \text{if } p(x)y > \gamma \end{cases}$$

with $\gamma = 0$. The parameter $\gamma \geq 0$ is often referred to as margin, and we shall call the corresponding error function L_γ margin error. This margin concept is useful in some theoretical analysis of classification problems.

Being non-convex, the classification error function cannot be directly handled in our analysis. However there are many convex formulations of classification problems that are used in practice. In fact, these convex formulations have become very popular in recent years due to the simplicity of optimization. As a comparison, directly minimizing the classification error often leads to NP hard problems.

In this paper, we provide analysis for two methods: the recently proposed Adaboost method [5], and the classical logistic regression model. Both are of significant practical interests. It is not difficult to see that our analysis also applies to some other convex classification formulations (as long as they are sufficiently smooth) such as certain variants of support vector machines.

In order to relate our analysis to previous works, we shall consider the boosting framework [5], [7]. In this framework, we are given a probability measure on (x, y), and a set of basis predictors: $p(x) \in S$ (hypothesis space). In boosting, S is a subset of measurable functions with values in $[-1, 1]$. Each $p(x) \in S$ can be regarded as a classifier itself using the decision rule given by (29). The purpose of boosting is to find a linear combination of classifiers in S that has better classification performance than any single classifier in S.

The original theory of boosting was based on the existence of a weak learner that has the ability to produce a classifier with better than random classification error for any given set of reweighted samples. Algorithmically, boosting can be considered as a procedure to construct a strong learner through successive weak learning on reweighted samples (adaptive resampling). However Breiman later observed that a boosting procedure can also be viewed as a greedy optimization of a convex loss function [6]. This point of view has led to new boosting formulations [6], [7], [8], [9]. The connection between boosting and the sparse approximation algorithms in [1], [2], [4] has also been noticed [8].

In general boosting procedures can be regarded as sparse approximation methods to compute additive models. These additive models are essentially linear models spanned by the basis functions in S. Formally, additive models correspond to functions in $\text{span}(S)$. Computationally, boosting methods are related to greedy approximation to the following problem

$$\inf_{p(x) \in \text{span}(S)} E_{x,y} L(p(x), y),$$

where E denotes the expectation over (x, y) and L is the loss function. In order to apply our analysis, it is necessary to restrict the optimization in $\text{span}(S)$ to $\text{co}(S)$ as

$$\inf_{p(x) \in \text{co}(S)} E_{x,y} L(p(x), y). \tag{30}$$

Clearly (30) is a special case of (2). This slight modification does not change the nature of boosting since a positive scaling of a predictor $p(x)$ does not change the associated classification prediction in (29). Furthermore, the theoretical analysis of boosting from the margin point of view [15] also relies on the assumption that the combined classifier $p(x)$ belongs to $\text{co}(S)$. It is thus very natural to study boosting under the modified formulation (30).

In our framework, the basic boosting procedure associated with a loss function L is simply Algorithm II.1 applied to the formulation (30). From the boosting point of view, the procedure to obtain an approximate solution of $(*)$ can be regarded as weak learning.

C.1 Adaboost

We consider the following form of loss that corresponds to Adaboost [5]:

$$L(p, y) = \exp(-Apy), \tag{31}$$

where A is a positive scaling factor. Similar to the case of L_p regression, we do not apply Theorem IV.2 to (30) directly. Instead, we consider the following equivalent optimization problem:

$$\inf_{p(x) \in \text{co}(S)} \ln E_{x,y} \exp(-Ap(x)y). \tag{32}$$

For all $v(x), w(x) \in \text{co}(S)$ and $\theta \in [0, 1]$, we have

$$f_{v,w}(\theta) = \ln E_{x,y} e^{-A((1-\theta)v(x) + \theta w(x))y}.$$

Note that it is well known that $f(p)$ is a convex function of p. One way to see this is to check that $f''_{v,w}(\theta) \geq 0$, which we shall verify below.

Let $z(\theta, x) = (1-\theta)v(x) + \theta w(x)$, and $\Delta v(x) = w(x) - v(x)$. Differentiate $f_{v,w}(\theta)$, we obtain:

$$f'_{v,w}(\theta) = -A \frac{E_{x,y} [e^{-Az(\theta,x)y} \Delta v(x) y]}{E_{x,y} e^{-Az(\theta,x)y}}.$$

Differentiate again, we have

$$\begin{aligned} f''_{v,w}(\theta) =& A^2 \frac{E_{x,y} [e^{-Az(\theta,x)y} \Delta v(x)^2]}{E_{x,y} e^{-Az(\theta,x)y}} \\ & - A^2 \frac{[E_{x,y} (e^{-Az(\theta,x)y} \Delta v(x) y)]^2}{[E_{x,y} e^{-Az(\theta,x)y}]^2} \\ \leq & A^2 \frac{E_{x,y} [e^{-Az(\theta,x)y} \Delta v(x)^2]}{E_{x,y} e^{-Az(\theta,x)y}} \\ \leq & A^2 \sup_x \Delta v(x)^2 \leq 4A^2. \end{aligned} \tag{33}$$

From the first equality above, we can see that $f''_{v,w}(\theta)$ can be regarded as the variance of $\Delta v(x)y$ with respect to the probability density $e^{-Az(\theta,x)y}/E_{x,y}\, e^{-Az(\theta,x)y}$. It follows that $f''_{v,w}(\theta) \geq 0$, and thus f is convex.

Inequality (33) implies that we can take $M = 4A^2$ in Theorem IV.2. For simplicity, we also let $\rho = 0$. After $k \geq 1$-round of boosting, we have

$$\begin{aligned} & E_{x,y} \exp(-Av^k(x)y) \\ \leq & \inf_{p(x)\in \mathrm{co}(S)} E_{x,y} \exp(-Ap(x)y + \frac{c'}{k+2}), \end{aligned} \tag{34}$$

where c' is defined in Theorem IV.2 with $M = 4A^2$ and $\rho = 0$.

The original idea of Adaboost is developed under the assumption that the weak learning algorithm can always make reasonable progress at each round. Under some appropriate measurement of progress, it was shown in [5] that the expected classification error decreases exponentially. The result was later extended in [15], where the authors proved that under similar assumptions on the weaker learner, the expected margin error L_γ with a positive margin $\gamma > 0$ also decreases exponentially. It follows that regularity assumptions of weak learning for Adaboost imply the following margin condition: $\exists \gamma_0 > 0$:

$$\inf_{p(x)\in \mathrm{co}(S)} E_{x,y} L_{\gamma_0}(p(x)y) = 0.$$

Equivalently, this can be rewritten as the following condition:

$$\inf_{p(x)\in \mathrm{co}(S)} P(p(x)y \leq \gamma_0) = 0. \tag{35}$$

Analysis given in this paper can be used to show that under this assumption, Algorithm IV.2 leads to classifiers with expected margin errors decreasing exponentially for all margin $0 \leq \gamma < \gamma_0$ (with appropriately chosen A in (31)). Clearly such a result complement existing results in the boosting literature.

In fact, to prove that the expected margin error decreases exponentially, we only need to impose the following weaker assumption which is a direct consequence of (35): $\forall A > 0$,

$$\inf_{p(x)\in \mathrm{co}(S)} E_{x,y} \exp(-Ap(x)y) \leq \exp(-\gamma_0 A). \tag{36}$$

Consider $c \leq A^2$ in Theorem IV.2. We have $c' \leq 9A^2$, and thus

$$\begin{aligned} & E_{x,y} \exp(-Av^k(x)y) \\ \leq & \inf_{p(x)\in \mathrm{co}(S)} E_{x,y} \exp(-Ap(x)y + \frac{c'}{k+2}) \\ \leq & \exp(-\gamma_0 A + \frac{9A^2}{k+2}). \end{aligned}$$

This implies that $\forall \gamma < \gamma_0$:

$$\begin{aligned} & E_{x,y} L_\gamma(v^k(x)y) \\ \leq & E_{x,y} \exp(-A(v^k(x)y - \gamma)) \\ \leq & \exp\left(-(\gamma_0 - \gamma)A + \frac{9A^2}{k+2}\right). \end{aligned}$$

Now fix a number $k \geq 1$, we can choose $A = (\gamma_0 - \gamma)(k+2)/18$ to obtain

$$E_{x,y} L_\gamma(v^k(x)y) \leq \exp\left(-\frac{(\gamma_0 - \gamma)^2}{36}(k+2)\right). \tag{37}$$

This implies that the expected margin error decreases exponentially for all margin $\gamma < \gamma_0$. The exponential decay of misclassification error (margin error) is the original motivation of Adaboost [5]. This observation has also been combined with margin analysis [15] to give theoretical justification for boosting. From the margin point of view, it is clear that our analysis, which starts with the margin assumption (35), leads to a more direct approximation bound for boosting. As a comparison, the original Adaboost analysis given in [5], [15] also requires assumptions on the underlying weak learning algorithm in that the classification error of the combined classifier is assumed to decrease by a certain amount at each step. Note that the existence of weak learning algorithms with such a property was not investigated in [5], [15].

In the framework of additive models, Adaboost corresponds to the exponential loss (31) which is analyzed in this section. However as pointed out in [6], [7], [8], other loss functions can also be used. Sparse approximation bounds for these different formulations can be obtained using our analysis. However, it is also easy to observe that they will not lead to the exponential decay of classification error in the separable case. For comparison purpose, we will also study the logistic regression formulation of boosting. Although the Adaboost exponential loss (31) is attractive for separable problems due to the exponential decay of expected margin error, it is also very sensitive to outliers for non-separable problems. Therefore the logistic regression formulation of boosting can be preferable for noisy problems.

C.2 Logistic regression boosting

We consider logistic regression in the boosting framework proposed in [7]. The corresponding loss function is

$$L(p, y) = \ln(1 + \exp(-Apy)). \tag{38}$$

In this case, (30) becomes:

$$\inf_{p(x)\in \mathrm{co}(S)} E_{x,y} \ln(1 + \exp(-Ap(x)y)). \tag{39}$$

From a statistical point of view, logistic regression is desirable since it corresponds to the (conditional) maximum-likelihood estimate with the following conditional probability model:

$$P(y|x) = \frac{1}{1 + \exp(-Ap(x)y)}.$$

This also implies that if the model is correct, then the maximum-likelihood estimate is consistent and asymptotically efficient (under appropriate regularity conditions). Even if the true conditional probability does not lie in the

model family, the estimator computes an estimate that is closest to the true distribution in the KL-divergence distance (relative entropy). Because of these reasons, the logistic regression formulation is highly desirable for classification problems.

Although a boosting procedure based on the logistic regression formulation was proposed in [7], no sparse approximation bound was given. Therefore up to now there are no quantitative results on how well such a boosting procedure can perform. A sparse approximation bound for logistic regression can be of great interests. It is also easy to see that such a bound is relatively easy to obtain under the framework developed in this paper. In the following we measure the convergence using the negative log-likelihood of the computed model which is natural for logistic regression.

For all $v(x), w(x) \in \mathrm{co}(S)$ and $\theta \in [0,1]$, let $z(\theta,x) = (1-\theta)v(x)+\theta w(x)$ and $\Delta v(x) = w(x)-v(x)$. Differentiate

$$f_{v,w}(\theta) = E_{x,y} \ln(1 + e^{-Az(\theta,x)y})$$

with respect to θ, we obtain:

$$f'_{v,w}(\theta) = E_{x,y} \frac{-A\Delta v(x) y}{1 + e^{Az(\theta,x)y}},$$

and

$$f''_{v,w}(\theta) = E_{x,y} \frac{A^2 \Delta v(x)^2}{(1 + e^{Az(\theta,x)y})(1 + e^{-Az(\theta,x)y})} \le A^2.$$

This shows that we may take $M = A^2$ and $\rho = 0$ in Theorem IV.2, and obtain the following bound for Algorithm II.1:

$$\begin{aligned} &E_{x,y} \ln(1 + \exp(-Av^k(x)y)) \\ &\le \inf_{p(x)\in\mathrm{co}(S)} E_{x,y} \ln(1 + \exp(-Ap(x)y)) + \frac{9A^2}{4(k+2)} \end{aligned}$$

for all $k \ge 1$ if we pick $\epsilon_k \le A^2/4(k+2)^2$.

D. Mixture density estimation

We consider the negative log-likelihood loss for density estimation which corresponds to the standard maximum-likelihood estimate:

$$f(p(\cdot)) = -E_x \ln p(x),$$

where $p(x) \ge 0$ is a probability density function.

Consider a set S of basis density functions $p(x)$, which we call mixture components. A mixture density model is a convex combination of mixture components. That is, the space of mixture models formed from the mixture component set S is $\mathrm{co}(S)$.

The mixture density model is a very popular technique to enhance the approximation power of a basic mixture component model. It also has some Bayesian statistical interpretation. A frequently used method to estimate such a mixture model from data is the so-called expectation maximization (EM) algorithm. However, it is also interesting to analyze the performance of the greedy fitting method in Algorithm II.1, which were studied in [11], [12]. Furthermore, this analysis can naturally lead to approximation bounds for some EM algorithms that can be regarded as refinements (as far as sparse approximation is concerned) of Algorithm II.1.

In the following we apply Theorem IV.2 to this problem and compare with results in [12], [16], where a more elaborated analysis specific to this problem was given.[4] We shall mention that the basic idea used in their approach is the same as what we have used in this paper. For all $v(x), w(x) \in \mathrm{co}(S)$ and $\theta \in [0,1]$, we have

$$f_{v,w}(\theta) = -E_x \ln((1-\theta)v(x) + \theta w(x)).$$

Let $z(\theta,x) = (1-\theta)v(x)+\theta w(x)$, and $\Delta v(x) = w(x)-v(x)$. Differentiate $f_{v,w}(\theta)$, we obtain:

$$f''_{v,w}(\theta) = E_x \frac{\Delta v(x)^2}{z(x)^2} \le 4 \sup_{p(x),q(x)\in\mathrm{co}(S)} E_x \frac{q(x)^2}{p(x)^2}.$$

Observe that $E_x \frac{q(x)^2}{p(x)^2}$ is convex both in $p(x)$ and in $q(x)$. The definition of convexity implies that the supremum of a convex function in a convex hull equals the supremum of the function on the vertices of the convex hull:

$$\sup_{p(x),q(x)\in\mathrm{co}(S)} E_x \frac{q(x)^2}{p(x)^2} = \sup_{p(x),q(x)\in S} E_x \frac{q(x)^2}{p(x)^2}.$$

We can thus let

$$M = 4 \sup_{p(x),q(x)\in S} E_x \frac{q(x)^2}{p(x)^2} \qquad \text{and} \qquad \rho = 0$$

in Theorem IV.2, and obtain the following bound:

$$-E_x \ln v^k(x) \le -\inf_{v(x)\in\mathrm{co}(S)} E_x \ln v(x) + \frac{c'}{k+2}.$$

The bound given in [12] is of a different form, and it is also written under the assumption that $\epsilon_k = 0$ (this assumption can be easily removed with a more careful analysis). Specifically the following bound was proved for all $v(x) \in \mathrm{co}(S)$ under the presentation $v(x) = \sum_{i=1}^m \alpha_i p_i(x)$, where $\alpha_i \ge 0$, $\sum_i \alpha_i = 1$, and $p_i \in S$:

$$-E_x \ln v^k(x) \le -E_x \ln v(x) + \frac{c_v}{k+1}\gamma_S,$$

where $\gamma_S = 4(\ln(3\sqrt{e}) + \sup_{p,q,\in S,x} \ln \frac{p(x)}{q(x)})$, and

$$c_v = E_x \frac{\sum_{i=1}^m \alpha_i p_i(x)^2}{(\sum_{i=1}^m \alpha_i p_i(x))^2}.$$

It is clear that the term c_v corresponds to the second derivative $f''_{v,w}(\theta)$ in our analysis. However, their analysis is refined for this specific problem in the sense that they preserve the representation $v(x) = \sum_{i=1}^m \alpha_i p_i(x)$ in the final

[4] We shall not compare with [11] since their result was given in a different form.

bound. In our analysis, we take the supremum over all $v(x)$ and the corresponding α_i. However, their refinement also introduced an extra S-dependent factor γ_S, which could be replaced by an S-independent constant asymptotically. To see this, consider the uniform norm topology, and assume that $w^*(x) \in \overline{\text{co}}(S)$ is the optimum solution of (2). Then the discussion of asymptotic bound at the end of Section IV suggests that by restricting $v(x)$ in their bound so that $v(x) \to w^*(x)$, it should be possible to replace γ_S by an S-independent numerical constant asymptotically as $k \to \infty$.

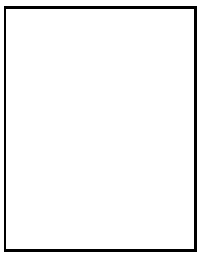
VI. Conclusion

The general greedy algorithm investigated in this paper is motivated by a number of recent studies. In particular, our analysis shows that the sparse functional approximation work of Maurey [3] and its greedy version [1], [2], and the boosting idea [5], [6], [7] can be studied under the same general framework. We have shown that for a variety of loss functions, a convergence rate of $O(1/k)$ can be achieved using a convex combination of k basic models. We have demonstrated the consequence of this general algorithm for a number of estimation methods, and related the resulting algorithms to previous studies.

The analysis given in this paper is quite general. Although it leads to convergence rates that have the tight order of convergence, examples in Section V indicate that in many cases, constants in the bounds resulted from Theorem IV.2 can be further improved using more refined analysis. Although such refinement has to be done in a case by case fashion, our general analysis still provides hints on how to approach this and the form of the final bound one expects to obtain. For example, this information can be obtained from the asymptotic analysis which we outlined at the end of Section III and Section IV.

Acknowledgments

The author would like to thank Andrew Barron, Wee Sun Lee, and the referees for useful comments and for pointing out related works, and Gunnar Rätsch for discussion on the relationship of the current work and the approach given in [10].



Tong Zhang Tong Zhang received a B.A. in mathematics and computer science from Cornell University in 1994 and a Ph.D. in computer Science from Stanford University in 1998. Since 1998, he has been with IBM Research, T.J. Watson Research Center, Yorktown Heights, New York, where he is now a research staff member in the Knowledge management department. His research interests include machine learning, numerical algorithms, and their applications.

References

[1] L.K. Jones, "A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training," *Ann. Statist.*, vol. 20, no. 1, pp. 608–613, 1992.

[2] A.R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 930–945, 1993.

[3] G. Pisier, "Remarques su un resultat non publié de b. maurey," in *Séminaire d'Analyse Fonctionelle*, École Polytechnique, Centre de Mathematiques,Palaiseau, 1980-1981, vol. 1.

[4] W.S. Lee, P.L. Bartlett, and R.C. Williamson, "Efficient agnostic learning of neural networks with bounded fan-in," *IEEE Transactions on Information Theory*, vol. 42, no. 6, pp. 2118–2132, 1996.

[5] Y. Freund and R.E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, 1997.

[6] L. Breiman, "Prediction games and arcing algorithms," *Neural Computation*, vol. 11, pp. 1493–1517, 1999.

[7] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: A statistical view of boosting," *The Annals of Statistics*, vol. 28, no. 2, pp. 337–407, 2000, With discussion.

[8] L. Mason, P. Bartlett, J. Baxter, and M. Frean, "Functional gradient techniques for combining hypotheses," in *Advances in Large Margin Classifiers*, B. Schölkopf A. Smola, P. Bartlett and D. Schuurmans, Eds. MIT Press, 2000.

[9] Robert E. Schapire and Yoram Singer, "Improved boosting algorithms using confidence-rated predictions," *Machine Learning*, vol. 37, pp. 297–336, 1999.

[10] G. Rätsch, S. Mika, and M.K. Warmuth, "On the convergence of leveraging," NeuroCOLT2 Technical Report NC-TR-01-098, Royal Holloway College, London, August 2001, A short version appeared in NIPS 14, MIT press, 2002.

[11] A. Zeevi and R. Meir, "Density estimation through convex combinations of densities; approximation and estimation bounds," *Neural Networks*, vol. 10, pp. 99–109, 1997.

[12] J.Q. Li and A.R. Barron, "Mixture density estimation," in *Advances in Neural Information Processing Systems 12*, S.A. Solla, T.K. Leen, and K.-R. Müller, Eds. 2000, pp. 279–285, MIT Press.

[13] V. Barbu and Th. Precupanu, *Convexity and Optimization in Banach Spaces*, Editura Academiei, Bucaresti,Romania, 1986.

[14] M. J. Donahue, L. Gurvits, C. Darken, and E. Sontag, "Rates of convex approximation in non-Hilbert spaces," *Constr. Approx.*, vol. 13, no. 2, pp. 187–220, 1997.

[15] R.E. Schapire, Y. Freund, P.L. Bartlett, and W.S. Lee, "Boosting the margin: a new explanation for the effectiveness of voting methods," *Ann. Statist.*, vol. 26, no. 5, pp. 1651–1686, 1998.

[16] J.Q. Li, *Estimation of Mixture Models*, Ph.D. thesis, The Department of Statistics. Yale University, 1999.