J. KSIAM Vol.22, No.1, 1–13, 2018 http://dx.doi.org/10.12941/jksiam.2018.22.001

# AN OPTIMAL BOOSTING ALGORITHM BASED ON NONLINEAR CONJUGATE GRADIENT METHOD

JOOYEON CHOI, BORA JEONG, YESOM PARK, JIWON SEO, CHOHONG MIN†

DEPARTMENT OF MATHEMATICS, EWHA WOMANS UNIVERSITY, KOREA
*E-mail address*: chohong@ewha.ac.kr

ABSTRACT. Boosting, one of the most successful algorithms for supervised learning, searches the most accurate weighted sum of weak classifiers. The search corresponds to a convex programming with non-negativity and affine constraint. In this article, we propose a novel Conjugate Gradient algorithm with the Modified Polak-Ribiera-Polyak conjugate direction. The convergence of the algorithm is proved and we report its successful applications to boosting.

## 1. INTRODUCTION

Boosting refers to constructing a strong classifier based on the given training set and weak classifiers, and has been one of the most successful algorithms for supervised learning [1, 8, 9]. A first and seminal boosting algorithm, named AdaBoost, was introduced by [3]. AbaBoost can be understood as a gradient descent algorithm to minimize the margin, a measure of confidence of the strong classifier [3, 7, 10].

Though simple and explicit, Adaboost is still one of the most popular boosting algorithms for classification and supervised learning. According to the analysis by [10], Adaboost tries to minimize a smooth margin. The hard margin refers to a direct sum of the confidence of each data, and the soft margin takes the log-sum-exponential function. LPBoost invented by [2, 4] minimizes the hard margin, resulting in a linear programming. It is observed that LPBoost does not perform well in most cases compared to Adaboost [11].

The strong classifier is a weighted sum of the weak classifiers. Adaboost determines the weight by the stagewise and unconstrained gradient descent. Adaboost increases the support of the weight one-by-one for each iteration. Due to the stagewise search and the stop of its search when the support is enough, Adaboost is not the optimal search.

The optimal solution needs to be sought among all the linear combinations of weak classifiers. The optimization becomes valid with a constraint that sum of the weights is bounded, and the bound was observed to be proportional to the support size of the weight [11].

Received by the editors January 14 2018; Accepted March 7 2018; Published online March 10 2018.
2000 *Mathematics Subject Classification.* 47N10,34A45.
*Key words and phrases.* convex programming, boosting, machine learning, convergence analysis.

† Corresponding author.

In this article, we propose a new and efficient algorithm that solves the constrained optimized problem. Our algorithm is based on the Conjugate-Gradient method with non-negativity constraint by [5]. They showed the convergence of CG with the modified Polak-Ribiera-Polyak (MPRP) conjugate direction.

The optimization that arise in Boosting has the non-negativity constraint and an affine constraint. Our novel algorithm extends the CG with non-negativity to hold the affine constraint. The addition of the affine constraint is a deal as big as adding the non-negative constraint.

We present a mathematical setting of boosting in section 2, introduce the novel CG and prove its convergence in section 3, and report its applications to bench mark problems of boosting in section 4.

## 2. Mathematical formulation of boosting

In boosting, one is given with a set of training examples $\{x_1, \cdots, x_M\}$ with binary labels $\{y_1, \cdots, y_M\} \subset \{\pm 1\}$, and weak classifiers $\{h_1, h_2, \cdots, h_N\}$. Each weak classifier $h_j$ gives a label to each example, and hence it is a function $h_j : \{x_1, \cdots, x_M\} \to \{\pm 1\}$.

A strong classifier F is made up of a weighted sum of the weak classifiers, so that $F(x) := \sum_{j=1}^{N} w_j h_j(x)$ for some $w \in \mathbb{R}^N$ with $w \geq 0$.

For each example $x_i$, a label $+1$ is put when $F(x_i) > 0$, and $-1$ otherwise. Hence the strong classifier is successful on $x_i$ if the sign of $F(x_i)$ matches the given label $y_i$, or $\operatorname{sign}(F(x_i)) \cdot y_i = +1$ and unsuccessful on $x_i$ if $\operatorname{sign}(F(x_i)) \cdot y_i = -1$.

The hard margin, which is a measure of the fidelity of the strong classifier, is thus given as

$$\text{(Hard margin)} : -\sum_{i=1}^{M} \operatorname{sign}(F(x_i)) \cdot y_i$$

When the margin is smaller, more of $\operatorname{sign}(F(x_i)) \cdot y_i$ are $+1$, and $F$ can be said to be more reliable. Due to the discontinuity present in the hard margin, the soft margin of Adaboost takes the form, via the monotonicity of log and exponential,

$$\text{(Soft margin)} : \log\left(\sum_{i=1}^{M} e^{-F(x_i)\cdot y_i}\right)$$

The composition of log-sum-exponential functions is referred to lse. Let us denote by $A \in \{\pm 1\}^{M \times N}$, the matrix whose entry is $a_{ij} = h_j(x_i) \cdot y_i$. Then the soft margin can be simply put to $\operatorname{lse}(-Aw)$, where $w = [w_1, \cdots, w_N]^T$.

The main goal of this work is to find out a weight that minimizes the soft margin, which is to solve the following optimization problem.

$$\text{minimize } \operatorname{lse}(-Aw) \text{ subject to } w \geq 0 \text{ and } w \cdot 1 = \frac{1}{T} \tag{2.1}$$

Here, $A \in \{\pm 1\}^{M \times N}$ is a given matrix from the training data and weak classifiers, and $T$ is a parameter to control the support size of $w$. We finish this section with the lemma that shows that the optimization is a convex programming, and we will introduce a novel algorithm to solve the optimization.

**Lemma 1.** $\operatorname{lse}(-Aw)$ *is a convex function with respect to* $w$.

*Proof.* Given any $w, \tilde{w} \in \mathbb{R}^N$ and any $\theta \in (0,1)$, let $z = -Aw$ and $\tilde{z} = -A\tilde{w}$.

$$
\begin{aligned}
&= (1-\theta)\,\mathrm{lse}\,(z) + \theta \mathrm{lse}\,(\tilde{z}) \\
&= \log \left( \left( \sum_{i=1}^{M} e^{z_i} \right)^{1-\theta} \cdot \left( \sum_{i=1}^{M} e^{\tilde{z_i}} \right)^{\theta} \right) \\
&= \log \left( \left( \sum_{i=1}^{M} \left( e^{z_i(1-\theta)} \right)^{\frac{1}{1-\theta}} \right)^{1-\theta} \cdot \left( \sum_{i=1}^{M} \left( e^{\tilde{z_i}\theta} \right)^{\frac{1}{\theta}} \right)^{\theta} \right) \\
&\leq \log \left( \sum_{i=1}^{M} e^{z_i(1-\theta)} \cdot e^{\tilde{z_i}\theta} \right) \text{ by the Hlder's inequality.} \\
&= \mathrm{lse}\,((1-\theta)\,z + \theta\tilde{z}) \\
&= \mathrm{lse}\,(-A\,((1-\theta)\,w + \theta\tilde{w})).
\end{aligned}
$$

□

## 3. Conjugate gradient method

In this section, we introduce a conjugate gradient method for solving the convex programming (2.1).

$$\min\ f\,(w) \text{ subject to } w \geq 0 \text{ and } w \cdot 1 = \frac{1}{T}$$

Throughout this section, $f\,(w)$ denotes the convex function $lse\,(-Aw)$, and $g\,(w)$ denotes its gradient $\nabla f\,(w)$. Let $d$ be the direction at a position $w$ to seek the next position. When $w$ is located on the boundary of the constraint, $w$ cannot be moved to a certain direction $d$ due to the constraints $\left\{ w \in \mathbb{R}^N \mid w \geq 0 \text{ and } w \cdot 1 = 1/T \right\}$.

We refer $d$ to be feasible at $w$, if $w + \alpha d$ stays in the constraint set for sufficiently small $\alpha > 0$.

**Definition 1.** *(Feasible direction) Given a direction $d \in \mathbb{R}^N$ at a position $w \in \mathbb{R}^N$ with $w \geq 0$ and $w \cdot 1 = \frac{1}{T}$, the feasible direction $d^f = d^f(d, w)$ associated with $d$ is defiend as the nearest vector to $d$ among the feasible directions at the position. Precisely, it is defined by the minimization*

$$d^f = \underset{y_{I(w)} \geq 0 \text{ and } y \cdot 1 = 0}{\operatorname{argmin}} \parallel d - y \parallel \tag{3.1}$$

*where $I(w) = \{i \mid w_i = 0\}$. The domain of the minimization is convex, and the functional is strictly convex and coercive, so that $d^f$ is determined uniquely.*

*Define the index set $J(w) = \{j \mid w_j > 0\}$.*

**Lemma 2.** *$\forall \omega$ with $w \cdot 1 = \frac{1}{T}$, $\forall d$, let $d^f = d^f\,(d, w)$, then $w + \alpha d^f \geq 0$ and $\left(w + \alpha d^f\right) \cdot 1 = \frac{1}{T}$ for sufficiently small $\alpha \geq 0$.*

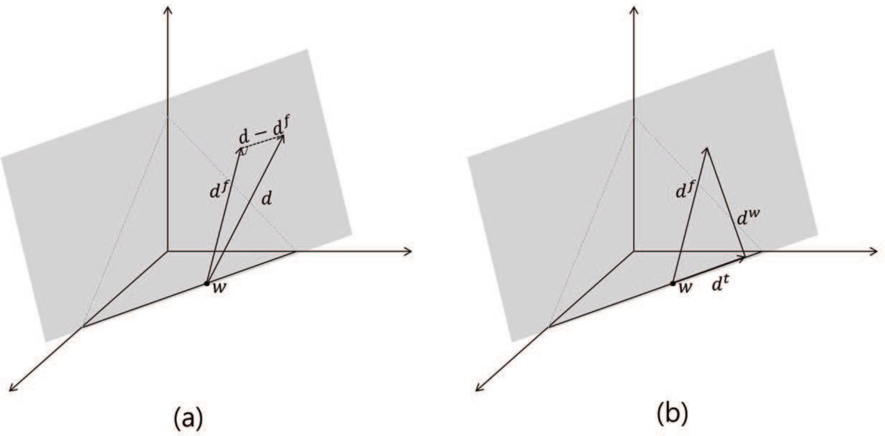

FIGURE 1. For a given direction $d$ at a position $w$, a colored region is a feasible region of $d$. Since $d^f$ is the nearest vector to $d$ among the feasible directions at $w$, it is the orthogonal projection of $d$ onto the colored region (a). $d^f$ is decomposed into two orthogonal components, $d^f = d^t + d^w$, where $d^t$ is the orthogonal projection of $d^f$ onto the tangent space (b).

*Proof.* Clearly, $\forall\alpha$, $\left(w + \alpha d^f\right) \cdot 1 = w \cdot 1 + 0 = 1/T$.

$$\begin{aligned} \forall\alpha \geq 0, \quad \text{if } i \in I\,(w)\,, &\quad w_i + \alpha d_i^f = 0 + \alpha d^f \geq 0, \text{ and} \\ \text{if } j \in J\,(w)\,, &\quad w_j + \alpha d_j^f \geq w_j - \alpha \left( \left| d_j^f \right| + 1 \right). \end{aligned}$$

Thus for any $\alpha \geq 0$ with $\alpha \leq \min_{j \in J(w)} \frac{w_j}{\left|d_j^f\right|+1}$, $w + \alpha d^f \geq 0$. □

**Proposition 1.** *(Calculation of the feasible direction) For a given direction $d$ at a position $w$, $d^f$ is calculated as*

$$\begin{cases} d_i^f & = (d_i - r)^+, \; i \in I \\ d_j^f & = d_j - r \quad , \; j \in J \end{cases}$$

*where $r$ is a zero of* $(d_J - r \cdot 1_J) \cdot 1_J + (d_{I_1} - r)^+ + \cdots + (d_{I_k} - r)^+, k = \mid I \mid$.

*Proof.* Since $d^f$ is the KKT point of $(Def.2)$, there exist $\lambda_I$ and $\mu$ such that

$$d^f - d = \begin{bmatrix} \lambda_I \\ 0 \end{bmatrix} - r \cdot 1, \text{ with } d_I^f \geq 0\,,\, \lambda_I \cdot d_I^f = 0\,,\, d^f \cdot 1 = 0\,.$$

From these conditions, we get $d_J^f = d_J - r \cdot 1_J$ and $d_i - r = d_i^f - \lambda_i$ , for $i \in I$.
If $d_i - r > 0$, then $d_i^f > 0$ and $\lambda_i = 0$. Thus, $d_i^f = d_i - r$.
If $d_i - r \leq 0$, then $d_i^f = 0$ and $\lambda_i \geq 0$.

**Algorithm 1** Computing the feasible direction, $d^f$.

```
Input : $w$, $d$
Output : $d^f$
Procedure :
1 : Make index sets $I(w) \coloneqq \{i|w(i)=0\}$ and $J(w) \coloneqq \{j|w(j)>0\}$
2 : Define a function $p(r) = \sum_{i\in I}(d_i - r)^+ + \sum_{j\in J}(d_j - r)$. And find
        $\alpha = \underset{i\in I, p(d_i)>0}{argmax\ i}$
        $\beta = \underset{i\in I, p(d_i)\le 0}{argmin\ i}$
3 : $r \leftarrow$ zero of $\sum_{j\in J}(d_j - r) + \sum_{i\in I, i\le\alpha}(d_i - r) - \sum_{i\in I, i>\beta}(d_i - r)$
4 : Compute $d^f$ as following : $d_i^f = \begin{cases} d_i - max\{0, -d_i + r\} + r, & i \in I \\ d_i - r, & i \in J \end{cases}$
```

By combining these two, we have $d_i^f = (d_i - r)^+$, for $i \in I$. Since $d^f \cdot 1 = 0$,

$$\begin{aligned} d^f \cdot 1 &= d_J^f \cdot 1_J + d_I^f \cdot 1_I \\ &= (d_J - r \cdot 1_J) \cdot 1_J + (d_{I_1} - r)^+ + (d_{I_2} - r)^+ + \cdots + (d_{I_k} - r)^+ = 0. \end{aligned}$$

$r$ is the root of the monotonically decreasing function. The monotone function is piecewisely linear, so that the root can be easily obtained by probing intervals between $\{d_{I_1}, \cdots, d_{I_k}\}$ where the monotone function changes the sign. After $r$ is obtained, $d^f$ is defined as stated. □

**Definition 2.** *(Tangent Space) The domain for $w$ is the simplex $\{w \mid w \ge 0$ and $w \cdot 1 = 0\}$. When $w > 0$, $w$ is inside and the tangent space $T = 1^\perp$.When $w_i = 0$ and $w_j > 0$ ($\forall j \neq i$), $w$ is on the boundary, and the tangent space becomes smaller $T_w = \{1, e_i \mid i \in I\}^\perp$. In general, we define the tangent space of $w$ as $T_w := [1 \cup \{e_i \mid w_i = 0\}]^\perp \subset \mathbb{R}^N$.*

**Definition 3.** *(Orthogonal decomposition of direction) Given a direction $d \in \mathbb{R}^N$ on a position $w \in \mathbb{R}^N$ with $w \ge 0$ and $1 \cdot w = \frac{1}{T}$, the direction is decomposed into three mutually orthogonal vectors; tangential, wall, and non-feasible components.*

$$\begin{aligned} d &= d^f + \left(d - d^f\right) \\ &= d^t + d^w + \left(d - d^f\right). \end{aligned}$$

Here, $d^f = d^f(d, w)$ is the feasible direction. $d^t$ is its orthogonal projection onto the tangent space $T_w$, and $d^w = d^f - d^t \in T_w^\perp$. Their mutual orthogonality is proved below.

**Lemma 3.** *The above vectors $d^t, d^w$, and $(d - d^f)$ are orthogonal to each other. Furthermore, $d - d^f \in T_w^\perp$.*

*Proof.* By the definition of the orthogonal projection, $d^t \perp d^w$. The KKT condition of the minimization (3.1) is

$$\left(d^f - d\right) = \begin{bmatrix} \lambda_I \\ 0 \end{bmatrix} + r1 \text{ for some } \lambda_I \geq 0 \text{ with } \lambda_I \cdot d_I^f = 0 \text{ and some } r \text{ with } r(d \cdot 1) = 0,$$

where $I = I(w)$. Since $d^t \in T_w = \{1, e_I\}^\perp$, $d^t \cdot \begin{bmatrix} \lambda_I \\ 0 \end{bmatrix} = 0$ and $d^t \cdot 1 = 0$, thus $d^t \perp d - d^f$.

From $d^f \cdot (d - d^f) = d_I^f \cdot \lambda_I + r(1 \cdot d) = 0$, we have $d^f \perp d - d^f$ and $d^w = d^f - d^t \perp d - d^f$ which completes the proof of their mutual orthogonalities.

Since $T_w = \{1, e_I\}^\perp$ and $d - d^f \in \text{span}\{1, e_I\}$, $d - d^f$ is orthogonal to the tangent space. □

**Definition 4.** *(MPRP direction) Let $w$ be a point with $w \geq 0$ and $w \cdot 1 = \frac{1}{T}$, and let $g = \nabla f(w)$. Putting tilde for the variable in the previous step : let $\tilde{g}$ be the gradient and $\tilde{d}$ be the search direction in the previous step, then the modified Polak-Ribiera-Polyak direction $d^{MPRP} = d^{MPRP}\left(w, \tilde{g}, \tilde{d}\right)$ is defined as*

$$d^{MPRP} = (-g)^f - \frac{(-g)^t \cdot (g - \tilde{g})^t}{\tilde{g} \cdot \tilde{g}} \tilde{d}^t + \frac{(-g)^t \cdot \tilde{d}^t}{\tilde{g} \cdot \tilde{g}} (g - \tilde{g})^t$$

**Theorem 1.** *(KKT condition) $\forall w \geq 0$ with $w \cdot 1 = \frac{1}{T}, \forall \tilde{g}, \forall \tilde{d}$, let $g = \nabla f(w)$ and $d = d^{MPRP}\left(w, \tilde{g}, \tilde{d}\right)$, then $(-g)^f \cdot d \geq 0$. Moreover $(-g)^f \cdot d = 0$ if and only if $w$ is a KKT point of the minimization problem (2.1).*

*Proof.*

$$\begin{aligned} (-g)^f \cdot d &= (-g)^f \cdot \left[(-g)^f - \frac{(-g)^t \cdot (g - \tilde{g})^t}{\tilde{g} \cdot \tilde{g}} \tilde{d}^t + \frac{(-g)^t \cdot \tilde{d}^t}{\tilde{g} \cdot \tilde{g}} (g - \tilde{g})^t\right] \\ &= \| (-g)^f \|^2 + \frac{1}{\tilde{g} \cdot \tilde{g}} \left[- \left[(-g)^t \cdot (g - \tilde{g})^t\right] \left[(-g)^f \cdot \tilde{d}^t\right] \right. \\ &\quad \left. + \left[(-g)^f \cdot (g - \tilde{g})^t\right] \left[(-g)^t \cdot \tilde{d}^t\right]\right] \end{aligned}$$

Since $(-g)^w \perp T_w$, $(-g)^w \cdot (g - \tilde{g})^t = 0$ and $(-g)^f \cdot (g - \tilde{g})^t = (-g)^t \cdot (g - \tilde{g})^t$.
Similarly, $(-g)^f \cdot \tilde{d}^t = (-g)^t \cdot \tilde{d}^t$, and we have $(-g)^f \cdot d = \| (-g)^f \|^2 \geq 0$.
The KKT condition for 2.1 is that

$$g = \lambda + r \cdot 1 \text{ for some } \lambda \geq 0 \text{ with } \lambda \cdot \text{w=0}$$
$$\text{and some } r \text{ with } r\left(w \cdot 1 - \frac{1}{T}\right) = 0.$$

**Algorithm 2** Algorithm based on nonlinear conjugate gradient

```
Input : Given constants $\rho \in (0,1)\,, \delta > 0, \epsilon > 0$. Initial point $w_0 \succeq 0$. Let $k = 0$, and
$g = \nabla f\left(w_0\right)$ where $f = lse\left(-Aw\right)$.
Output : $w$
Procedure :
1 : Compute $d = \left(d_I, d_J\right)$ by Algorithm 1.
        If $\left|(-g)^f \cdot d\right| \leq \epsilon,$ then stop.
        Otherwise, go to the next step.
2 : Determine $\alpha = \max\left\{\frac{-d_k\cdot\nabla f(w)}{d_k\cdot(\nabla^2 f(w) d_k)}\rho^j, j = 0, 1, 2, \cdots\right\}$ satisfying $w + \alpha d \geq 0$
        and $f\left(w + \alpha d\right) \leq f\left(w\right) - \delta\alpha^2 \parallel d \parallel^2$
3 : $w \leftarrow w + \alpha d$
4 : $k \leftarrow k + 1$, and go to step 2.
```

Since $w_J > 0$ and $\lambda \geq 0$, $\lambda_J = 0$. Since $w \cdot 1 = \frac{1}{T}$, and $w_I = 0$, the condtions $r\left(w \cdot 1 - \frac{1}{T}\right) = 0$ and $\lambda \cdot w = \lambda_I \cdot w_I + \lambda_J \cdot w_J = 0$ are unnecessary. Therefore, the KKT condition is simplified as

$$g = \left[\begin{array}{c}\lambda_I\\ 0\end{array}\right] + r \cdot 1 \text{ for some } \lambda_I \geq 0 \text{ and some } r.$$

On the other hand, $(-g)^f \cdot d = \parallel (-g)^f \parallel^2 = 0$ if and only if $0 = (-g)^f = argmin_{y_I \geq 0 \text{ and } y\cdot 1 = 0} \parallel (-g) - y \parallel$, whose KKT condition is that

$$g = \left[\begin{array}{c}\lambda_I\\ 0\end{array}\right] + r \cdot 1 \text{ for some } \lambda_I \geq 0 \text{ and some } r.$$

Each of the two minimization problems has a unique minimum point, accordingly a unique KKT condition. Since their KKT conditions are same, we have

$$(-g)^f \cdot d = 0 \iff w \text{ is the KKT point of the minimization problem 2.1.}$$

□

Next, we introduce some properties of $f(w)$ and Algorithm 2 to prove the global convergence of Algorithm 2.

*Properties*

Let $V = \left\{w \in \mathbb{R}^N \mid w \geq 0 \text{ and } w \cdot 1 = \frac{1}{T}\right\}$.

(1) Since the feasible set V is bounded, the level set $\left\{w \in \mathbb{R}^N \mid f(w) \leq f(w_0)\right\}$ is bounded. Thus, $f$ is bounded from below.

(2) The sequence $\{w_k\}$ generated by Algorithm 2 is a feasible point sequence and the function value sequence $\{f(w_k)\}$ is decreasing. In addition, since $f(w)$ is bounded below,

$$\sum_{k=0}^{\infty} \alpha_k^2 \parallel d_k \parallel^2 < \infty.$$

Thus we have

$$\lim_{k\to\infty} \alpha_k \parallel d_k \parallel = 0.$$

(3) $f$ is continuously differentiable, and its gradient is the Lipschitz continuous; there exists a constant $L > 0$ such that

$$\parallel \nabla f(w) - \nabla f(y) \parallel \leq \parallel x - y \parallel, \forall x, y \in \text{V}$$

These imply that there exists a constant $\gamma_1$ such that

$$\parallel \nabla f(w) \parallel \leq \gamma_1, \forall x \in V.$$

**Lemma 4.** *If there exists a constant $\epsilon \geq 0$ such that*

$$\parallel g\left(x_k\right) \parallel \geq \epsilon,\ \forall k,$$

*then there exists a constant $M > 0$ such that*

$$\parallel d_k \parallel \leq M,\ \forall k.$$

*Proof.*

$$\begin{aligned}\parallel d_k^{MPRP} \parallel &\leq \parallel (-g)^f \parallel + \frac{2 \parallel (-g)^t \parallel \cdot \parallel (g-\tilde{g})^t \parallel \cdot \parallel \tilde{d}_k^t \parallel}{\parallel \tilde{g} \parallel^2} \\ &\leq \gamma_1 + \frac{2\gamma_1 L \alpha_k \parallel \tilde{d}_k^t \parallel}{\epsilon^2} \parallel \tilde{d}_k^t \parallel\end{aligned}$$

Since $\lim_{k\to\infty} \alpha_k \parallel d_k \parallel = 0$, $\exists$a constant $\gamma \in (0,1)$ and $k_0 \in \mathbb{Z}$ such that

$$\frac{2L\gamma_1}{\epsilon^2} \alpha_{k-1} \parallel \tilde{d}_k^t \parallel \leq \gamma \text{ for all } k \geq k_0.$$

Hence, for any $k \geq k_0$,

$$\begin{aligned}\parallel d_k^{MPRP} \parallel &\leq 2\gamma_1 + \gamma \parallel d_{k-1} \parallel \\ &\leq 2\gamma_1 \left(1 + \gamma + \cdots + \gamma^{k-k_0-1}\right) + \gamma^{k-k_0} \parallel d^{k_0} \parallel \\ &\leq \frac{2\gamma_1}{1-\gamma} + \parallel d_{k_0} \parallel\end{aligned}$$

Let $M = max\left\{\parallel d_1 \parallel, \parallel d_2 \parallel, \cdots, \parallel d_{k_r} \parallel, \frac{2\gamma_1}{1-\gamma} + \parallel d_{k_0} \parallel\right\}$. Then $\parallel d_k^{MPRP} \parallel \leq M, \forall k$. □

**Lemma 5.** *(Success of Line search) In Algorithm 2, the line search step is guaranteed to succeed for each $k$. Precisely speaking,*

$$f\left(w_k+\alpha_k d_k\right) \leq f\left(w_k\right)-\delta \alpha_k^2 \parallel d_k \parallel^2$$

*for all sufficiently small $\alpha_k$.*

*Proof.* By the Mean Value Theorem,

$$f\left(w_k+\alpha_k d_k\right)-f\left(w_k\right)=\alpha_k g\left(w_k+\alpha_k \theta_k d_k\right) \cdot d_k,$$

for some $\theta_k \in (0,1)$. The line search is performed only if $\left(-g\left(w_k\right)\right)^f \cdot d_k > \epsilon$. In Lemma7, we showed that $(-g(w_k))-(-g(w_k))^f \perp T_w$ and $(-g(w_k))-(-g(w_k))^f \perp (-g(w_k))^f$. Since $d_k \in (-g(w_k))^f + T_w$, $\left[(-g(w_k))-(-g(w_k))^f\right] \cdot d_k = 0$ and we have

$$-g(w_k) \cdot d_k = (-g(w_k))^f \cdot d_k > \epsilon.$$

From the continuity of $g(w)$,

$$-g\left(w_k+\alpha_k \theta_k d_k\right) \cdot d_k > \frac{\epsilon}{2}$$

for sufficiently small $\alpha_k$. Choosing $\alpha_k \in \left(0, \frac{\epsilon}{2\delta \|d_k\|^2}\right)$, we get

$$\begin{aligned} f\left(w_k+\alpha_k d_k\right) &= f\left(w_k\right)+\alpha_k g\left(w_k+\alpha_k \theta_k d_k\right) \cdot d_k \\ &< f\left(w_k\right)-\frac{\epsilon}{2}\alpha_k \\ &\leq f\left(w_k\right)-\delta \alpha_k^2 \parallel d_k \parallel^2 . \end{aligned}$$

□

**Theorem 2.** *Let $\{w_k\}$ and $\{d_k\}$ be the sequence generated by Algorithm 2, then*

$$\liminf_{k\to\infty} \left(-g_k\right)^f \cdot d_k = 0.$$

*Thus the minimum point $w^*$ of our main problem* (2.1) *is a limit point of the set $\{w_k\}$and Algorithm 2 is convergent.*

*Proof.* We first note that $(-g_k)^f \cdot d_k = -g_k \cdot d_k$ that appeared in the proof of Lemma 11. We prove the theorem by contradiction. Assume that the theorem is not true, then there exists an $\epsilon > 0$ such that

$$\parallel \left(-g_k\right)^f \parallel^2 = \left(-g_k\right)^f \cdot d_k > \epsilon, \text{ for all } k$$

By Lemma 10, there exists a constant $M$ such that

$$\parallel d_k \parallel \leq M, \text{ for all } k.$$

If $\liminf_{k\to\infty} \alpha_k > 0$, then $\lim_{k\to\infty} \parallel d_k \parallel = 0$. Since $\parallel g \parallel_\infty < -r$, $\lim_{k\to\infty} \left(-g_k\right)^f \cdot d_k = 0$. This contradicts assumption.

If $\liminf_{k\to\infty} \alpha_k = 0$, then there is an infinite index set $K$ such that

$$\lim_{k\in K, k\to\infty} \alpha_k = 0.$$

It follows from the step 2 of Algorithm 2, that when $k \in K$ is sufficiently large, $\rho^{-1}\alpha_k$ does not satisfy $f\left(w_k + \alpha_k d_k\right) \leq f\left(w_k\right) - \delta\alpha_k^2 \parallel d_k \parallel ^2$, that is

$$f\left(w_k + \rho^{-1}\alpha_k d_k\right) - f\left(w_k\right) > -\delta\rho^{-2}\alpha_k^2 \parallel d_k \parallel ^2 \tag{3.2}$$

By the Mean Value Theorem and Lemma 10, there is $h_k \in (0,1)$ such that

$$\begin{aligned} f\left(w_k\right) - f\left(w_k + \rho^{-1}\alpha_k d_k\right) &= \rho^{-1}\alpha_k g\left(w_k + h_k\rho^{-1}\alpha_k d_k\right) \cdot d_k \\ &= \rho^{-1}\alpha_k g\left(w_k\right) \cdot d_k + \rho^{-1}\alpha_k\left(g\left(w_k + h_k\rho^{-1}\alpha_k d_k\right) - g\left(w_k\right)\right) \cdot d_k \\ &\leq \rho^{-1}\alpha_k g\left(w_k\right) \cdot d_k + L\rho^{-2}\alpha_k^2 \parallel d_k \parallel ^2 \end{aligned}$$

Substitute the last inequality into (3.2) and applying $-g(w_k) \cdot d_k = (-g)^f(w_k) \cdot d_k$, we get for all $k \in K$ sufficiently large,

$$0 \leq (-g)^f\left(w_k\right) \cdot d_k \leq \rho^{-1}\left(L + \delta\right)\alpha_k \parallel d_k \parallel ^2.$$

Taking the limit on both sides of the equation, then by combining $\parallel d_k \parallel \leq M$ and recalling $\lim_{k \in K, k \to \infty} \alpha_k = 0$, we obtain the $\lim_{k \in K, k \to \infty} \mid (-g)^f\left(x_k\right) \cdot d_k \mid = 0$.

This also yields a contradiction. □

**Remark 1.** *To say the existence of* $k$ *which satisfies (3.2), we should verify that* $w_k + \rho^{-1}\alpha_k d_k$ *is feasible. Since* $d_k \cdot 1 = 0$, $(w_k + \rho^{-1}\alpha_k d_k) \cdot 1 = w_k \cdot 1 = \frac{1}{T}$*. So, we should check* $w_k + \rho^{-1}\alpha_k d_k \geq 0$*. Since* $\lim_{k \in K, k \to \infty} \alpha_k = 0$, $\alpha_k$ *is near to zero for sufficiently large* $k$*. Thus,* $w_k + \rho^{-1}\alpha_k d_k \geq 0$ *except very special cases.*

## 4. NUMERICAL RESULTS

In this section, we test our proposed CG algorithm on two boosting examples of non-negligible size. Through the tests, we check if their numerical results match the analyses presented in section 3.

Our algorithm is supposed to generate a sequence $\{w_k\}$ on which the soft margin monotonically decreases, which is the first check point. According to Theorem (2), the stopping criteria $(-g_k)^f \cdot d_k < \epsilon$ should be satisfied after a finite number of iterations for any given threshold $\epsilon > 0$, which is the second check point. According to Theorem(1), the solution $w_k$ with the stopping criteria satisfied is the KKT point, which is the third one. The KKT point is the global minimizer of the soft margin, the optimal strong classifier, which is the final one.

### 4.1. **Low dimensional example.**

We solve a boosting problem that minimizes lse$(-Aw)$ with $w \geq 0$ and $w \cdot 1 = \frac{1}{2}$, where $A$ is a $4 \times 3$ matrix given below.

$$A = \begin{bmatrix} -1 & 1 & 1 \\ -1 & 1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & -1 \end{bmatrix}$$

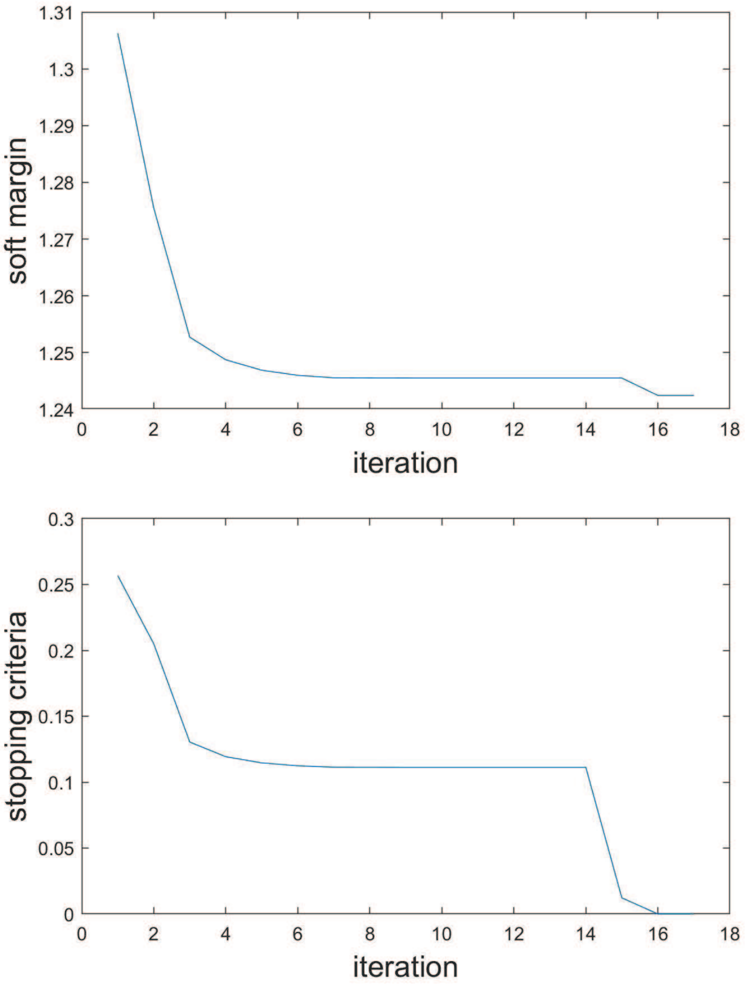

FIGURE 2. the convergence of the CG method for example 4.1

As shown in Figure 2, the soft margin lse$(-Aw)$ monotonically decreases and the stopping criteria $(-g_k)^f \cdot d_k$ drops to a very small number in finite iterations, which is equivalent to the statement of Theorem 2, $\liminf_{k\to\infty} (-g_k)^f \cdot d_k = 0$.

4.2. **Classifying win/loss of sports games.** One of the primal applications of boosting is to classify win/loss of sports games [6]. As an example, we take the vast amount of statistics from the basketball league of a certain country*(for a patent issue, we do not disclose the details).

The statistics of each game is represented by the following 36 numbers.

TABLE 1. Statistics form the basketball league

| | | | | |
|---|---|---|---|---|
| Home $(+1)$ | Field Goals Made | Field Goals Attempted | Field Goals Percentage | 3 Point Goals |
| | 3 Point Goals Attempted | 3 Point Goals Percentage | Free Throws Made | Free Throws Attempted |
| | Free Throws Percentage | Offensive Rebounds | Defensive Rebounds | Total Rebounds |
| | Assists | Personal Fouls | Steals | Turnovers |
| Road $(-1)$ | Field Goals Made | Field Goals Attempted | Field Goals Percentage | 3 Point Goals |
| | 3 Point Goals Attempted | 3 Point Goals Percentage | Free Throws Made | Free Throws Attempted |
| | Free Throws Percentage | Offensive Rebounds | Defensive Rebounds | Total Rebounds |
| | Assists | Personal Fouls | Steals | Turnovers |

In a whole year, there were 538 number of games with the win/loss results, from which we take a training data $\{x_1, \cdots, x_{M=269}\}$ with the win/loss of the home team $\{y_1, \cdots, y_M\} \subset \{\pm 1\}$. Each $x_i$ represents the statistics of a game, and $x_i \in \mathbb{R}^{269\times 36}$.

Similarly to the previous example, Figure 3 shows that the soft margin monotonically decreases and the stopping criteria drops to a very small number in finite iterations, matching the analyses in Section 3.

## 5. CONCLUSION

We proposed a new Conjugate Gradient method for solving convex programmings with the non-negative constraints and a linear constraint, and successfully applied the method to the boosting problems. We also presented a convergence analysis for the method. Our analysis shows that the method is convergent in a finite iteration for any small stopping threshold. The solution with the stopping criteria satisfied is shown to be the KKT point of the convex programming and hence the global minimizer of the programming. We solved two benchmark boosting problems by the CG method, and obtained numerical results that completely cope with the analysis. Our algorithm with the guaranteed convergence can be successful in other boosting problems as well as other convex programmings.

## REFERENCES

[1] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168. ACM, 2006.

[2] Ayhan Demiriz, Kristin P Bennett, and John Shawe-Taylor. Linear programming boosting via column generation. *Machine Learning*, 46(1):225–254, 2002.

[3] Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995.

[4] Adam J Grove and Dale Schuurmans. Boosting in the limit: Maximizing the margin of learned ensembles. In *AAAI/IAAI*, pages 692–699, 1998.

[5] Can Li. A conjugate gradient type method for the nonnegative constraints optimization problems. *Journal of Applied Mathematics*, 2013, 2013.

[6] B. Loeffelholz, B. Earl, and B.W. Kenneth. Predicting nba games using neural networks. *Journal of Quantitative Analysis in Sports*, pages 1–15, 2009.

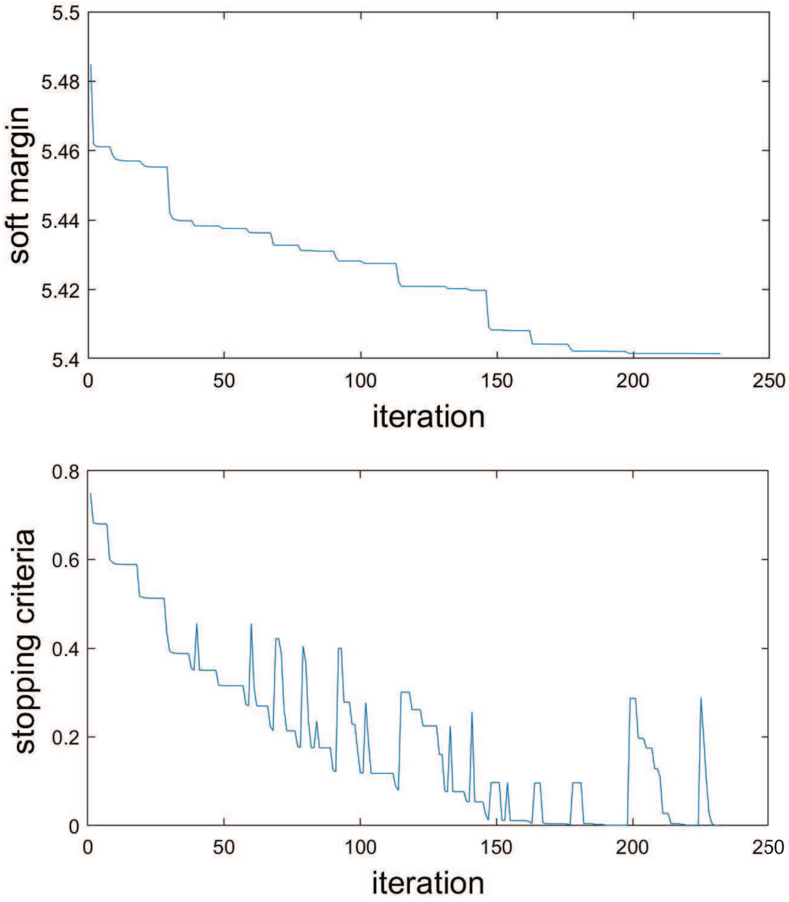

FIGURE 3. the convergences of the CG method for example 4.2

[7] N.Duffy and D.Helmbold. A geometric approach to leveraging weak learners. In *Computational Learning Theory, Lecture Notes in Comput. Sci.*, pages 18–33. Springer, 1999.

[8] R.E.Schapire. The boosting approach to machine learning: An overview. In *Nonlinear Estimation and Classification. Lecture Notes in Statist.*, volume 171, pages 149–171. Springer, 2003.

[9] R.Meir and G.Ratsch. An introduction to boosting and leveraging. In *Advanced Lectures on Machine Learning*, volume 2600, pages 119–183. Springer, 2003.

[10] Cynthia Rudin, Robert E Schapire, Ingrid Daubechies, et al. Analysis of boosting algorithms using the smooth margin function. *The Annals of Statistics*, 35(6):2723–2768, 2007.

[11] Chunhua Shen and Hanxi Li. On the dual formulation of boosting algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(12):2216–2231, 2010.