数理解析研究所講究録
第1829巻 2013年 88-101

An Inexact Coordinate Descent Method for the Weighted l_1-regularized Convex Optimization Problem

Xiaoqin Hua* Nobuo Yamashita†

October 26, 2012

Abstract. The purpose of this paper is to propose an inexact coordinate descent (ICD) method for solving the weighted l_1-regularized convex optimization problem with a box constraint. The proposed algorithm solves a one dimensional subproblem inexactly at each iteration. We give criteria of the inexactness under which the sequence generated by the proposed method converges to an optimal solution and its convergence rate is at least R-linear without assuming the uniqueness of solutions.

Keywords. l_1-regularized convex optimization, inexact coordinate descent method, linear convergence, error bound.

AMS Subject Classification (2000). 65K05,90C25,90C30.

1 Introduction

We consider the following weighted l_1-regularized convex optimization problem:

$$\begin{aligned} &\text{minimize} \quad F(x) := g(Ax) + \langle b, x\rangle + \sum_{i=1}^{n} \tau_i |x_i| \\ &\text{subject to} \quad l \leq x \leq u, \end{aligned} \tag{1.1}$$

where $g : \mathcal{R}^m \to (-\infty, \infty]$ is a strictly convex function on $\operatorname{dom} g$, $A \in \mathcal{R}^{m\times n}$ and $b \in \mathcal{R}^n$. Moreover, τ, l and u are n-dimensional vectors such that $l_i \in [-\infty, \infty)$, $u_i \in (-\infty, \infty]$, $\tau_i \in [0, \infty)$ and $l_i < u_i$ for each $i = 1, \cdots, n$. The nonnegative scalar constant τ_i is called weight and the term $\sum_{i=1}^{n} \tau_i |x_i|$ is called the l_1-regularization function. For convenience, we denote the differentiable term of F by f, that is, $f(x) := g(Ax) + \langle b, x\rangle$.

*Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University, Kyoto 606-8501, JAPAN. E-mail: hua.xiaoqin.22r@st.kyoto-u.ac.jp

†Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University, Kyoto 606-8501, JAPAN. E-mail:nobuo@i.kyoto-u.ac.jp

The problem (1.1) contains many well-known problems as special cases [7, 17, 15]. When $\tau_i = 0$ for all index i, the problem (1.1) is reduced to the differentiable convex problem with a box constraint. When $l_i = -\infty$ and $u_i = \infty$ for each i, it is reduced to the unconstrained l_1-regularized convex problem. When τ_i is a fixed positive constant $\hat{\tau}$ for all i, $b = 0$ and $g(y) := \frac{1}{2}\|y - z\|^2$ with some $z \in \mathcal{R}^m$, it is the famous l_1-l_2 problem. Another important special case is the l_1-regularized logistic regression problem where g is given by $g(y) := \frac{1}{m}\sum_{i=1}^{m} \log(1 + \exp(-y_i))$. Each special case has wide applications in the real life such as the compressed sensing [15], the feature selection in the data classification [7], the data mining [9], geophysics [1] and so on. Typically, the scales of these weighted l_1-regularized convex optimization problems are very large and the objective functions are not differentiable everywhere due to the regularization function. Moreover, the optimal solutions are possibly not unique because the matrix A may not have full column rank. Thus the Newton-type methods such as the interior point method can not be applied directly.

In the past, the coordinate descent (CD) method is verified to be one of the feasible methods for the large scale optimization problems [4, 11, 14, 17]. The CD method minimizes the objective function with respect to one variable while all the others are fixed at each iteration. The idea of this method is very simple and the storage of calculations is little. In some special cases, it can be implemented in parallel. Luo and Tseng [17] proved its global and linear convergence for a smooth problem, that is, $\tau_i = 0$ for all i. Note that the problem (1.1) can be reformulated as a smooth problem. However, the reformulated problem has twice variables. In 2001, Tseng [11] showed the global convergence of a block coordinate descent (BCD) method for minimizing an nondifferentiable function with certain separability. But the exact minimizers of the subproblem must be found on each iteration in [11, 17]. It is possible for the l_1-l_2 problem, while usually it is hard for the general l_1-regularized convex problem.

In this paper, we propose an inexact coordinate descent (ICD) method and extend the results of Luo and Tseng [17]. Roughly speaking, we extend in the following three aspects:

- The smooth convex problem is extended to that with the l_1-regularized function.
- On each iteration step, we accept an inexact solution of a subproblem instead of the exact solution.
- The linear convergence rate is extended to the nonsmooth problem.

In this paper, under the same basic assumptions as in [17], we show that the proposed ICD method is not only globally convergent but also with at least R-linear convergence rate under the almost cycle rule.

This paper is organized as follows. In Section 2, we derive the optimal conditions of the problem (1.1) and also define the ε-optimality conditions which are related to an inexact solution. In Section 3, we present a framework of the ICD method and make some assumptions for the "inexact solutions". The global convergence and linear convergence rate are established in Section 4. Finally, we conclude this paper in Section 5.

Throughout this paper, we use the following notations. For a differentiable function h, ∇h denotes the gradient of h and $\nabla^2 h$ denotes the Hessian matrix of h. $\nabla_i h$ denotes the ith coordinate of the gradient vector ∇h. If h is convex and nondifferentiable, ∂h denotes the subdifferential of h. For any real number x, $|x|$ denotes the absolute value of x. For a given vector $x \in \mathcal{R}^n$, we denote by x_i the ith coordinate of x. We denote the 2-norm of x by $\|x\|$. For any matrix A, the transpose of A is denoted by A^T and A_j denotes the jth column. For the function $F : \mathcal{R}^n \to \mathcal{R}$ and a vector $x \in \mathcal{R}^n$, we sometimes use a notation $F(x_1, \cdots, x_n)$ instead of $F(x)$.

Since the length of the paper must be within 17 pages, we omit all proofs of theorems in the subsequent sections. The full proofs can be found in [16].

2 Preliminaries

Throughout the paper, we make the following basic assumptions for the problem (1.1).

Assumption 2.1. *For the problem (1.1), we assume that*

(a) A_j *is a nonzero vector for all* $j \in \{1, 2, \cdots, n\}$.

(b) $l_i < 0 < u_i$ *for all* $i \in \{1, 2, \cdots, n\}$.

(c) The set of the optimal solutions, denoted by X^**, is nonempty.*

(d) The effective domain of g*, denoted by* $\mathrm{dom}\, g$*, is nonempty and open.*

(e) g *is twice continuously differentiable on* $\mathrm{dom}\, g$.

(f) $\nabla^2 g(Ax^*)$ *is positive definite for every optimal solution* $x^* \in X^*$.

We make a few remarks on these assumptions. In Part (a), if A_j is zero, then x_j^* of the optimal solution x^* can be easily determined. Thus we can remove x_j from the problem (1.1). Part (b) is just for simplification. If both l_i and u_i are positive for some $i \in \{1, 2, \cdots, n\}$, we may replace x_i, l_i and u_i by $\bar{x}_i + \frac{l_i+u_i}{2}$, $\frac{l_i-u_i}{2}$ and $\frac{u_i-l_i}{2}$. Then the problem (1.1) is reformulated into the case without l_1-regularized term for the index i. If g is strongly convex and twice differentiable on $\mathrm{dom}\, g$, then Part (e) and (f) are satisfied automatically. For example, a quadratic

function, an exponential function, and even some complicate functions in the l_1-regularized logistic regression problem satisfy (e) and (f). Note that we do not assume the boundness of the optimal solution set X^*.

Next, we present some properties under Assumption 2.1 that are used in the subsequent sections. From Assumption 2.1(e) and (f), there exists a sufficiently small closed neighborhood $B(Ax^*)$ of Ax^* such that $B(Ax^*) \subseteq \operatorname{dom} g$ and $\nabla^2 g$ is positive definite in $B(Ax^*)$. Furthermore, it implies that g is strongly convex in $B(Ax^*)$, i.e., there exists a scalar $\sigma > 0$ such that

$$g(y) - g(z) - \langle \nabla g(z), y - z \rangle \geq \sigma \|y - z\|^2, \ \forall y, z \in B(Ax^*). \tag{2.1}$$

2.1 Optimality conditions

The KKT conditions [13] for the problem (1.1) are described as follows.

$$\begin{aligned} &\nabla_i f(x) + \tau_i \partial |x_i| - \mu_i + \nu_i \ni 0, \\ &x_i \geq l_i, \mu_i \geq 0, \mu_i(x_i - l_i) = 0, \quad i = 1, \cdots, n, \\ &x_i \leq u_i, \nu_i \geq 0, \nu_i(u_i - x_i) = 0, \end{aligned} \tag{2.2}$$

where $\partial |\cdot|$ is the subdifferential of the absolute value function. Since the problem (1.1) is convex, x satisfying (2.2) is a solution of the problem (1.1). The KKT conditions (2.2) can be rewritten as follows.

Lemma 2.1. *A vector x is an optimal solution of the problem (1.1) if and only if one of the following statements holds for each i.*

(i) $\nabla_i f(x) \geq \tau_i$ *and* $x_i = l_i$.

(ii) $\nabla_i f(x) = \tau_i$ *and* $l_i \leq x_i \leq 0$.

(iii) $|\nabla_i f(x)| \leq \tau_i$ *and* $x_i = 0$.

(iv) $\nabla_i f(x) = -\tau_i$ *and* $0 \leq x_i \leq u_i$.

(v) $\nabla_i f(x) \leq -\tau_i$ *and* $x_i = u_i$.

Next, we represent these conditions as a fixed point of some operator. To this end, we first define a mapping $T_\tau : \mathcal{R}^n \to \mathcal{R}^n$ as

$$T_\tau(x)_i := (|x_i| - \tau_i)_+ \operatorname{sgn}(x_i), \tag{2.3}$$

where the scalar function $(a)_+$ is defined by $(a)_+ := \max(0, a)$ and $\mathrm{sgn}(a)$ is a sign function defined as follows:

$$\mathrm{sgn}(a) := \begin{cases} -1 & \text{if } a < 0, \\ 0 & \text{if } a = 0, \\ 1 & \text{if } a > 0. \end{cases}$$

It can be verified that T_τ is nonexpensive, i.e., $\|T_\tau(y) - T_\tau(z)\| \leq \|y - z\|$, for any $y, z \in \mathrm{dom}\, F$.

Let $[x]^+_{[l,u]}$ denote the orthogonal projection of a vector x onto the box $[l, u]$. This projection is also nonexpensive and its ith coordinate can be written as $[x_i]^+_{[l_i,u_i]} := \mathrm{mid}\{x_i, l_i, u_i\}$, where $\mathrm{mid}\{x_i, l_i, u_i\}$ is defined by $\mathrm{mid}\{x_i, l_i, u_i\} := \max\{l_i, \min\{u_i, x_i\}\}$.

By using the mappings T_τ and $[\cdot]^+_{[l,u]}$, we define a mapping $P_{\tau,l,u}(x) : \mathcal{R}^n \to \mathcal{R}^n$ by

$$P_{\tau,l,u}(x) := [T_\tau(x - \nabla f(x))]^+_{[l,u]}. \tag{2.4}$$

Since $[x]^+_{[l,u]}$ and T_τ are nonexpensive, we have that

$$\|P_{\tau,l,u}(y) - P_{\tau,l,u}(z)\| \leq \|y - z - \nabla f(y) + \nabla f(z)\|, \ \forall y, z \in \mathrm{dom}\, F. \tag{2.5}$$

Now, the optimal solutions can be described as a fixed point of the mapping $P_{\tau,l,u}$.

Theorem 2.1. *For the problem (1.1), a vector x belongs to the optimal solution set X^* if and only if $x = P_{\tau,l,u}(x)$, i.e., $X^* = \{x |\ x \in \mathrm{dom}\, g, x = P_{\tau,l,u}(x)\}$.*

Since the solution set X^* is not necessarily bounded, the level set of F may be not bounded. Nevertheless, as an extension of [17, Lemma 3.3], we can show the compactness of the set $\Omega(\zeta) := \{t |\ t = Ax, F(x) \leq \zeta, x \in [l, u]\}$.

Lemma 2.2. *For a given constant value ζ, the set $\Omega(\zeta)$ is a compact subset of* dom g.

Next, we show that ∇g is Lipschitz continuous on some compact set including $\Omega(\zeta)$. For this purpose, we define a set $\Omega(\zeta) + B(\epsilon_0)$ as $\Omega(\zeta) + B(\epsilon_0) := \{p + v |\ p \in \Omega(\zeta), \ \|v\| \leq \epsilon_0\}$, where ϵ_0 is a positive constant. It is easy to see that the set $\Omega(\zeta) + B(\epsilon_0)$ is a compact set.

Lemma 2.3. *There exist constants $L > 0$ and $\epsilon_0 > 0$ such that $\Omega(\zeta) + B(\epsilon_0) \subseteq \mathrm{dom}\, g$ and $\|\nabla g(y) - \nabla g(z)\| \leq L\|y - z\|$ for all $y,\ z \in \Omega(\zeta) + B(\epsilon_0)$.*

Similar to [18, Lemma 2.1], we can prove the following invariant property of the optimal solution set X^*.

Lemma 2.4. *For any $x^*, y^* \in X^*$, $Ax^* = Ay^*$.*

2.2 ε-optimality conditions

In this subsection, we give a definition of a relaxed optimality conditions and a relation between the conditions and the mapping $P_{\tau,l,u}$.

Definition 2.1. *We say that the ε-optimality conditions for the problem (1.1) hold at x if one of the following statements holds for each i.*

(i) $\nabla_i f(x) - \tau_i \geq -\varepsilon$ *and* $|x_i - l_i| \leq \varepsilon$.

(ii) $|\nabla_i f(x) - \tau_i| \leq \varepsilon$ *and* $l_i - \varepsilon \leq x_i \leq \varepsilon$.

(iii) $|\nabla_i f(x)| \leq \tau_i + \varepsilon$ *and* $|x_i| \leq \varepsilon$.

(iv) $|\nabla_i f(x) + \tau_i| \leq \varepsilon$ *and* $-\varepsilon \leq x_i \leq u_i + \varepsilon$.

(v) $\nabla_i f(x) + \tau_i \leq \varepsilon$ *and* $|x_i - u_i| \leq \varepsilon$.

Definition 2.2. *We say x is an ε-approximate solution of the problem (1.1) if the ε-optimality conditions hold at x.*

Note that the optimality condition in Lemma 2.1 can be obtained by Definition 2.1 with $\varepsilon = 0$.

For convenience, we define the following five index sets:

$$
\begin{aligned}
J_1(x,\varepsilon) &:= \{i|\ \nabla_i f(x) - \tau_i \geq -\varepsilon, |x_i - l_i| \leq \varepsilon\};\\
J_2(x,\varepsilon) &:= \{i|\ |\nabla_i f(x) - \tau_i| \leq \varepsilon, l_i - \varepsilon \leq x_i \leq \varepsilon\};\\
J_3(x,\varepsilon) &:= \{i|\ |\nabla_i f(x)| \leq \tau_i + \varepsilon, |x_i| \leq \varepsilon\};\\
J_4(x,\varepsilon) &:= \{i|\ |\nabla_i f(x) + \tau_i| \leq \varepsilon, -\varepsilon \leq x_i \leq u_i + \varepsilon\};\\
J_5(x,\varepsilon) &:= \{i|\ \nabla_i f(x) + \tau_i \leq \varepsilon, |x_i - u_i| \leq \varepsilon\}.
\end{aligned}
$$

Then the ε-optimality conditions hold at x if and only if $\bigcup_{i=1}^{5} J_i(x,\varepsilon) = \{1, 2, \cdots, n\}$.

Throughout the paper, for simplicity, we assume that

$$\varepsilon < \frac{1}{2}\min_i\{-l_i, u_i\}. \tag{2.6}$$

By the definition of $T_\tau(x)$ and $P_{\tau,l,u}(x)$ in (2.3) and (2.4), we give an equivalent description of the ε-optimality conditions by the next theorem , which will be used for constructing an inexact CD method and investigating its convergence properties, where the proof is omitted here.

Theorem 2.2. *The ε-optimality conditions hold at x if and only if $|x_i - P_{\tau,l,u}(x)_i| \leq \varepsilon$ holds for each i.*

3 Inexact coordinate descent (ICD) method

In this section, we will first present a framework for the ICD method, and then give some assumptions for the "inexact solutions".

A general framework of the ICD method can be described as follows.

Inexact coordinate descent (ICD) method:

Step 0: Choose an initial point $x^0 \in [l, u]$ and let $r := 0$.

Step 1: If some termination condition holds, then stop.

Step 2: Choose an index $i(r) \in \{1, \cdots, n\}$, get an approximate solution $x_{i(r)}^{r+1}$ of the following one dimensional subproblem:

$$\underset{x_{i(r)} \in \{l_{i(r)} \le x_{i(r)} \le u_{i(r)}\}}{\text{minimize}} F(x_1^r, x_2^r, \cdots, x_{i(r)-1}^r, x_{i(r)}, x_{i(r)+1}^r, \cdots, x_n^r). \tag{3.1}$$

Step 3: Set $x_j^{r+1} := x_j^r$ for all $j \in \{1, \cdots, n\}$ such that $j \neq i(r)$ and let $r := r + 1$. Go to Step 1.

Note that the exact solution of the subproblem (3.1) is unique from Assumption 2.1(a) and the strict convexity of g. We use the notation $i(r)$ for the index chosen at the rth iteration. For simplicity, we will use i instead of $i(r)$ when $i(r)$ is clear from the context.

For the global convergence of the ICD method, it is important to define the inexactness of the approximate solutions of the subproblem (3.1) and choose an appropriate index $i(r)$ in Step 2.

For the inexactness, we require the following assumptions:

Assumption 3.1. *We assume that the following statements hold:*

(i) $F(x_1^r, x_2^r, \cdots, x_{i-1}^r, x_i^{r+1}, x_{i+1}^r, \cdots, x_n^r) \le \min_{x_i \in \{l_i, 0, u_i, x_i^r\}} F(x_1^r, x_2^r, \cdots, x_{i-1}^r, x_i, x_{i+1}^r, \cdots, x_n^r)$.

(ii) x_i^{r+1} *is feasible, i.e.,* $x_i^{r+1} \in [l_i, u_i]$.

(iii) x_i^{r+1} *is an* ε^{r+1}*-approximate solution of the subproblem (3.1).*

(iv) ***Conditions on*** ε^{r+1}*:* $\varepsilon^{r+1} \le \min\{\delta_r, \alpha_r |x_i^{r+1} - x_i^r|, \varepsilon^r\}$*, where* $\{\delta_r\}$ *is a monotonically decreasing sequence such that* $\lim_{r \to \infty} \delta_r = 0$*, and* $\alpha_r \in [0, \bar{\alpha}]$ *with a positive constant* $\bar{\alpha}$.

(v) ***Conditions on*** α_r*:* $\alpha_r < \dfrac{\sigma \min_j \|A_j\|^2}{L \max_j \|A_j\|^2 + 1}$ *holds for sufficiently large* r*, where* σ *is a positive constant defined in (2.1), and* L *is the Lipschitz constant of* ∇g *given in Lemma 2.3.*

Here we make a simple explanation. Part (i) enforces not only that $\{F(x^r)\}$ is decreasing but also that $\{F(x^{r+1})\}$ is less than $F(x_1^r, x_2^r, \cdots, x_{i-1}^r, x_i, x_{i+1}^r, \cdots, x_n^r)$ at a point where F is nonsmooth. This condition is easy to check when computing. It also plays a key role for the convergence of $\{x^r\}$ when the objective function is not differentiable. In Part (iii), recall that the ε-optimality conditions for the one dimensional subproblem (3.1) is that one of (i)-(v) in Definition 2.1 holds at $x_{i(r)}$. The assumptions(i)-(iv) are necessary for the global convergence while the assumption (v) for α_r is used to guarantee the linear convergence rate of $\{x^r\}$.

Note that if we obtain the exact solution of the subproblem (3.1) on each iteration, then the sequence $\{x^r\}$ satisfies Assumption 3.1 automatically. Hence, the classical CD method is a special case of the ICD method.

For the choice of the coordinate $i(r)$ in Step 2 , we adopt the following almost cycle rule, which is an extension of the classical cyclic rule in [3].

Almost cyclic rule:
There exists an integer $B \geq n$, such that every coordinate is iterated upon at least once every B successive iterations.

In the next section, we will show the ICD method with almost cycle rule converges R-linearly to a solution under Assumption 2.1 and 3.1.

4 Global and linear convergence

In this section, we will show the global and linear convergence of the ICD method. Compared with the classical exact CD method, the ICD method has many "inexact" factors. Thus we need some preparations.

First of all, we illustrate a brief outline of the proof.

(1) $\lim_{r\to\infty}\{x^{r+1} - x^r\} = 0$. (Lemma 4.3)

(2) $Ax^r \to Ax^*$ where x^* is one of the optimal solutions. (Theorem 4.1)

(3) Sufficient decreasing: $F(x^r) - F(x^{r+1}) \geq \eta\|x^r - x^{r+1}\|^2$ for some positive constant η. (Lemma 4.8)

(4) Error bound: $\|Ax^r - Ax^*\| \leq \kappa\|x^r - P_{\tau,l,u}(x^r)\|$ for some κ. (Lemma 4.9)

(5) Linear convergence. (Theorems 4.2 and 4.3)

Note that since it is not necessary for the matrix A to have full column rank, $Ax^r \to Ax^*$ (Theorem 4.1) does not imply $x^r \to x^*$.

For convenience, we define two vectors $\tilde{x}^{r+1}$ and x^{r+1} as follows.

$$\tilde{x}^{r+1} := (x_1^r, x_2^r, \cdots, x_{i(r)-1}^r, \tilde{x}_{i(r)}^{r+1}, x_{i(r)+1}^r, \cdots, x_n^r) \tag{4.1}$$

and

$$x^{r+1} := (x_1^r, x_2^r, \cdots, x_{i(r)-1}^r, x_{i(r)}^{r+1}, x_{i(r)+1}^r, \cdots, x_n^r), \tag{4.2}$$

where $x_{i(r)}^{r+1}$ and $\tilde{x}_{i(r)}^{r+1}$ are an ε^{r+1}-approximate solution and the exact solution of the subproblem (3.1), respectively.

In the first part of this section, we show $\lim_{r\to\infty}\{F(\tilde{x}^r) - F(x^r)\} = 0$ and $\lim_{r\to\infty}\{x^{r+1} - x^r\} = 0$. To this end, we need the following function $h_i : \mathcal{R}^n \times \mathcal{R}^n \to \mathcal{R}$ and Lemma 4.1.

$$\begin{aligned} h_i(y,z) :=& \nabla_i f(z)(y_i - z_i) + \tau_i(|y_i| - |z_i|) \\ =& \begin{cases} (\nabla_i f(z) + \tau_i)(y_i - z_i) & \text{if } y_i \geq 0, z_i \geq 0, \\ \nabla_i f(z)(y_i - z_i) + \tau_i(y_i + z_i) & \text{if } y_i \geq 0, z_i \leq 0, \\ \nabla_i f(z)(y_i - z_i) + \tau_i(-y_i - z_i) & \text{if } y_i \leq 0, z_i \geq 0, \\ (\nabla_i f(z) - \tau_i)(y_i - z_i) & \text{if } y_i \leq 0, z_i \leq 0. \end{cases} \end{aligned} \tag{4.3}$$

Lemma 4.1. *There exists a positive constant M such that $|x_{i(r)}^{r+1} - \tilde{x}_{i(r)}^{r+1}| \leq \frac{2M}{\|A_{i(r)}\|}$ for all r.*

Lemma 4.2. $\lim_{r\to\infty}\{F(\tilde{x}^r) - F(x^r)\} = 0.$

Using the above lemmas, we can show that $\{x^{r+1} - x^r\}$ converges to 0 by a similar way of [18, Lemma 4.1].

Lemma 4.3. *For the sequence $\{x^r\}$ generated by the ICD method, we have* $\lim_{r\to\infty}\{x^{r+1} - x^r\} = 0.$

In the second part of this section, we will show the convergence of $\{Ax^r\}$. Since $\{Ax^r\}$ is bounded, there exist $t^\infty \in \mathcal{R}^n$ and an infinite set $\mathcal{X}$ such that

$$\lim_{r\to\infty,\, r\in\mathcal{X}} Ax^r = t^\infty. \tag{4.4}$$

Then with the continuity of ∇g, we have

$$\lim_{r\to\infty,\, r\in\mathcal{X}} \nabla f(x^r) = d^\infty, \tag{4.5}$$

where

$$d^\infty := A^T \nabla g(t^\infty) + b. \tag{4.6}$$

For the set $\mathcal{X}$, we have the following result with Lemma 4.3, which provides an interesting property associated with $\{\nabla f(x^r)\}$.

Lemma 4.4. *For any $s \in \{0, 1, \cdots, B-1\}$, where B is the integer defined in the almost cycle rule, we have* $\lim_{r\to\infty,\, r\in\mathcal{X}} \nabla f(x^{r-s}) = d^{\infty}$.

Lemma 4.4 implies that for each $i \in \{1, 2, \cdots, n\}$, for any $s \in \{0, 1, \cdots, B-1\}$, we have

$$\lim_{r\to\infty,\, r\in\mathcal{X}} \nabla_i f(x^{r-s}) = d_i^{\infty}. \tag{4.7}$$

Fixed coordinate i, let , (r, i) denote the largest integer $\bar{r}$, which does not exceed r, such that the ith coordinate of x is iterated upon at the $\bar{r}$th iteration, that is to say, for all $r \in \mathcal{X}$, we have

$$x_i^r = x_i^{\varphi(r,i)}. \tag{4.8}$$

Since the coordinate is chosen by the almost cycle rule, the relation $r - B + 1 \leq , (r, i) \leq r$ holds for all $r \in \mathcal{X}$. From (4.7), we further obtain

$$\lim_{r\to\infty,\, r\in\mathcal{X}} \nabla_i f(x^{\varphi(r,i)}) = d_i^{\infty}. \tag{4.9}$$

Now we define the following six index sets associated with d_i^{∞} as

$$\begin{aligned}
J_1^{\infty} &:= \{i|\ d_i^{\infty} > \tau_i\};\\
J_2^{\infty} &:= \{i|\ d_i^{\infty} < -\tau_i\};\\
J_3^{\infty} &:= \{i|\ |d_i^{\infty}| < \tau_i\};\\
J_4^{\infty} &:= \{i|\ d_i^{\infty} = \tau_i,\ \tau_i > 0\};\\
J_5^{\infty} &:= \{i|\ d_i^{\infty} = -\tau_i,\ \tau_i > 0\};\\
J_6^{\infty} &:= \{i|\ d_i^{\infty} = 0,\ \tau_i = 0\}.
\end{aligned}$$

Note that $\bigcup_{i=1}^{6} J_i^{\infty} = \{1, 2, \cdots, n\}$. Next two lemmas give sufficient conditions under which $\{x_i^r\}_{\mathcal{X}}$ is fixed or lies in some interval.

Lemma 4.5. *Suppose that Assumption 3.1(i) and (iii) hold, and that $x^{\varphi(r,i)}$ is a vector, where the i-th coordinate is chosen on the , (r, i)-th iteration. Let L and ε_0 be the constants given in Lemma 2.3. If $\varepsilon^{\varphi(r,i)} < \varepsilon_0$, then the following statements hold for any fixed i:*

(i) *If $\nabla_i f(x^{\varphi(r,i)}) - \tau_i > L\|A_i\|^2\varepsilon^{\varphi(r,i)}$ and $x_i^{\varphi(r,i)} \leq \varepsilon^{\varphi(r,i)} + l_i$, then $x_i^{\varphi(r,i)} = l_i$.*

(ii) *If $\nabla_i f(x^{\varphi(r,i)}) + \tau_i < -L\|A_i\|^2\varepsilon^{\varphi(r,i)}$ and $u_i - \varepsilon^{\varphi(r,i)} \leq x_i^{\varphi(r,i)}$, then $x_i^{\varphi(r,i)} = u_i$.*

(iii) *If $\nabla_i f(x^{\varphi(r,i)}) + \tau_i > L\|A_i\|^2\varepsilon^{\varphi(r,i)}$ and $|x_i^{\varphi(r,i)}| \leq \varepsilon^{\varphi(r,i)}$, then $x_i^{\varphi(r,i)} \leq 0$.*

(iv) *If $\nabla_i f(x^{\varphi(r,i)}) - \tau_i < -L\|A_i\|^2\varepsilon^{\varphi(r,i)}$ and $|x_i^{\varphi(r,i)}| \leq \varepsilon^{\varphi(r,i)}$, then $x_i^{\varphi(r,i)} \geq 0$.*

Lemma 4.6. *Suppose that Assumption 3.1 holds. Then, for sufficiently large r, we have*

$$\{x_i^r\}_{\mathcal{X}} = l_i, \forall i \in J_1^\infty; \tag{4.10}$$

$$\{x_i^r\}_{\mathcal{X}} = u_i, \forall i \in J_2^\infty; \tag{4.11}$$

$$\{x_i^r\}_{\mathcal{X}} = 0, \forall i \in J_3^\infty; \tag{4.12}$$

$$l_i \leq \{x_i^r\}_{\mathcal{X}} \leq 0, \forall i \in J_4^\infty; \tag{4.13}$$

$$0 \leq \{x_i^r\}_{\mathcal{X}} \leq u_i, \forall i \in J_5^\infty; \tag{4.14}$$

$$l_i \leq \{x_i^r\}_{\mathcal{X}} \leq u_i, \forall i \in J_6^\infty. \tag{4.15}$$

Next, we will show $Ax^r \to Ax^*$, where x^* is an arbitrary optimal solution of the problem (1.1). For this purpose, we recall Hoffman's error bound [2].

Lemma 4.7. *Let $B \in \mathcal{R}^{k\times n}$, $C \in \mathcal{R}^{k\times n}$ and $e \in \mathcal{R}^k$, $d \in \mathcal{R}^k$. Suppose that the linear system $By = e, Cy \leq d$ is consistent. There exists a scalar $\theta > 0$ depending only on B and C, and such that, for any $\bar{x} \in [l, u]$, $l,\ u \in \mathcal{R}^n$, there is a point $\bar{y} \in \mathcal{R}^n$ satisfying $B\bar{y} = e, C\bar{y} \leq d$ and $\|\bar{x} - \bar{y}\| \leq \theta(\|B\bar{x} - e\| + \|(C\bar{x} - d)_+\|$, where $(x_i)_+ := \max\{0, x_i\}$.*

Theorem 4.1. *Let x^* be an optimal solution of the problem (1.1). Then we have $\lim_{r\to\infty} Ax^r = Ax^*$.*

Theorem 4.1 implies that there exists a scalar $\bar{r} > 0$, for any $r \geq \bar{r}$, such that $Ax^r \in B(Ax^*)$, where $B(Ax^*)$ is the closed ball defined just before (2.1). Note that g is strongly convex.

In the third part of this section, we show the sufficient decreasing of $\{F(x^r)\}$ for sufficiently large r.

Lemma 4.8. *Under Assumption 3.1, there exists a scalar $\eta > 0$ such that $F(x^r) - F(x^{r+1}) \geq \eta\|x^r - x^{r+1}\|^2$ for sufficiently large r.*

In the last part of this section, before showing the global and linear convergence of $\{x^r\}$, we first recall a kind of the Lipschitz error bound in [10, 11, 18].

Lemma 4.9. *There exists a scalar constant $\kappa > 0$ such that for any $Ax^r \in B(Ax^*)$,*

$$\|Ax^r - Ax^*\| \leq \kappa\|x^r - P_{\tau,l,u}(x^r)\|. \tag{4.16}$$

The following result is a direct extension of [17, Lemma 4.5(a)] to the problem (1.1).

Lemma 4.10. *Under Assumption 3.1, there exists a constant $. > 0$ such that the inequality $\|Ax^r - Ax^*\| \leq . \sum_{h=r}^{r+B-1} \|x^h - x^{h+1}\|$ holds for sufficiently large r.*

Now we are ready to show the linear convergence of $\{F(x^r)\}$ and $\{x^r\}$.

Theorem 4.2. *Suppose that $\{x^r\}$ is generated by the ICD method with the almost cycle rule. Let F^* denote the optimal value of the problem (1.1). Then $\{F(x^r)\}$ converges to F^* at least B-step Q-linearly.*

Theorem 4.3. *Suppose that $\{x^r\}$ is generated by the ICD method with the almost cycle rule. Then there exists an optimal solution x^* of the problem (1.1) such that $\{x^r\}$ converges to x^* at least R-linearly.*

5 Conclusions

In this paper, we have presented a framework of the ICD method for solving l_1-regularized convex optimization (1.1). We also have established the R-linear convergence rate of this method under the almost cycle rule. The key to the ICD method lies in Assumption 3.1 for the "inexact solution". On each iteration step, we only need to find an approximate solution, that raises the possibility to solve general l_1-regularized convex problem.

The proposed ICD method solves an one-dimensional subproblem on each iteration. The Block Coordinate Descent method, which solves a small scale multi-dimensional subproblem, is efficient for some practical problems. Thus it is interesting to extend the proposed ICD method to the "inexact" block CD method.

References

[1] A. Gholami and H. R. Siahkoohi, Regularization of linear and non-linear geophysical ill-posed problems with joint sparsity constraints, Geophysical Journal International, Vol.180, No.2, pp. 871–882, 2010.

[2] A. J. Hoffman, On approximate solutions of systems of linear inequalities, Journal of Research of the National Bureau of standards, Vol. 49, No.4, pp. 263–265, 1952.

[3] D. G. Luenberger, Linear and nonlinear programming, Addision-Wesley, Reading, Massachusetts, 1973.

[4] H. Liu, M. Palatucci, and J. Zhang, Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery, Proceedings of the 26th Annual International Conference on Machine Learning, pp. 649-656, 2009.

[5] J. M. Borwein and A. S. Lewis. Convex analysis and nonlinear optimization: theory and examples, second edition, New York: spinger-verlag, 2000, Canadian mathematical society books in mathematics.

[6] J. M. Ortega and W. C. Rheinboldt, Iterative solution of nonlinear equations in several variables, Reprinted by SIAM, Philadelphia, 2000.

[7] K. Koh, S. J. Kim, and S. Boyd, An interior-point method for large- scale l_1-regularized logistic regression, Journal of Machine Learning Research 8, pp.1519-1555, 2007.

[8] M. A. T. Figueiredo, R. D. Nowak and S. J. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems, IEEE Journal of Selected Topics in Signal Processing, Vol.1, No.4, pp. 586-597, 2007.

[9] M. Y. Park, and T. Hastie, L1-regularization path algorithm for generalized linear models, Journal of the Royal Statistical Society: Series B (Statistical Methodology), Vol. 69, No.4, pp. 659–677, 2007.

[10] P. Tseng, Approximation accuracy, gradient methods, and error bound for structured convex optimization, Mathematical Programming: Series A and B - 20th International Symposium on Mathematical Programming – ISMP 2009, Vol. 125 N0. 2, pp. 263–295, 2010.

[11] P. Tseng, Convegence of a block coordinate descent method for nondifferentiable minimization, Journal of Optimization and Applications, Vol.109, pp. 475-494, 2001.

[12] P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization, Math Programming, Vol. 117, pp. 387-423, 2009.

[13] R. T. Rockafellar, Convex analysis, Princeton University Press, Princeton, New Jersey, 1970.

[14] T. T. Wu, and K. Lange, Coordinate descent algorithms for lasso penalized regression, The Annals of Applied Statistics, Vol.2, No.1, pp. 224-244, 2008.

[15] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, Bregman iterative algorithms for l_1-minimization with applications to compressed sensing, SIAM J. Imaging Sciences, Vol.1, No.1, pp. 143-168, 2008.

[16] X. Q. Hua, N. Yamashita, An Inexact Coordinate Descent Method for the Weighted l_1-regularized Convex Optimization Problem, Technical Reports 2012, http://www.amp.i.kyoto-u.ac.jp/tecrep/.

[17] Z. Q. Luo and P. Tseng, On the convergence of the coordinate descent method for convex differentiable minimization, Journal of Optimization and Applications, Vol.72, pp. 7-35, 1992.

[18] Z. Q. Luo and P. Tseng, On the linear convergence of descent methods for convex essentially smooth minimization, SIAM J. Control and Optimization, Vol.30, pp. 408-425, 1992.

[19] Yu. Nesterov, Introductory lectures on convex optimization: a basic course, 2004.