A Rate of Convergence for Two-Block Coordinate Descent

Seyyed Mohammad Rouzban
University of Tehran
Reshad Hosseini
reshad.hosseini@ut.ac.ir
University of Tehran

Abstract

This short report gives a non-asymptotic rate of convergence proof for solving a two-block coordinate descent problem. This non-asymptotic proof is a simple result that can be derived easily from available results in the literature. We give the results in this report because in this general form we have not seen being stated in the literature.

1 Introduction

In this short report, we show the global convergence guarantee for a class of two-block coordinate descent (BCD) method. We give a theorem for the global convergence rate of the proposed method. Our results can be derived easily from available results [1, 3] but we have not seen these results being stated clearly in those papers and other related literatures. Because of its wide applicability and importance, we give the results here.

2 A Convergence Theorem for Coordinate Descent Method

In this part, we give the following convergence theorem for two-block coordinate descent. As a corollary, we show that the coordinate descent method, where we find the solution with respect to one block analytically.

**Theorem 1.** Consider, we use a block coordinate descent algorithm for minimizing the differentiable function $f(x, y)$ with two blocks of variables $x$ and $y$. The block coordinate descent algorithm decreased the cost function with respect to block $y$ and finds a stationary point with respect to this block, i.e., $\nabla_y f(x_t, y_t) = 0$. In the algorithm, the solution for the block $x$ satisfies the following condition

$$f(x_t, y_t) - f(x_{t+1}, y_t) \geq \frac{1}{2E_x} \|\nabla_x f(x_t, y_t)\|^2. \quad (1)$$

Then,

$$\sum_{t=0}^{T-1} \frac{1}{E_x} \|\nabla f(x_t, y_t)\|^2 \leq f(x_0, y_0) - f(x_T, y_T). \quad (2)$$

Furthermore, assume $f(x_t, y_t)$ is bounded from below and $E_x \geq E > 0$ for all $t$ then every limit point is a stationary point. The rate of convergence to the stationary point is $O(1/\sqrt{T})$, where $T$ is the number of iterations.

**Proof.** From (1), $\nabla_y f(x_t, y_t) = 0$, and $f(x_{t+1}, y_{t+1}) \leq f(x_{t+1}, y_t)$, we have

$$\frac{1}{2E_x} \|\nabla f(x_t, y_t)\|^2 \leq f(x_t, y_t) - f(x_{t+1}, y_{t+1}). \quad (3)$$

The summation of inequality (3) over $t$ from 0 to $T - 1$, yields (2). The right side of equation (2) is bounded and therefore if $T \to \infty$ then $\nabla f(x^*, y^*) \to 0$, where $(x^*, y^*)$ is a limit point. From (2), it is easy to see that

$$\min_{0 \leq t \leq T - 1} \|\nabla f(x_t, y_t)\|^2 \leq E_x \frac{f(x_0, y_0) - f(x_T, y_T)}{T},$$

where $E_x = \max_{1 \leq t \leq T} E_x$. Therefore, we obtain the rate of convergence $O(1/\sqrt{T})$. \qed
The condition (1) is very easy to be satisfied. For example, it is enough to use gradient descent with fixed step-size, and the gradient for the block is Lipschitz continuous (see [2]).

**Proposition 2.** Assume the gradient function \( g_t(x) = f(x, y_t) \) is Lipschitz continuous, i.e.,

\[
\|\nabla_x f(x', y_t) - \nabla_x f(x, y_t)\| \leq L(y_t)\|x' - x\|.
\]

For the gradient descent step \( x_{t+1} = x_t - \frac{1}{L(y_t)} \nabla_x f(x_t, y_t) \), the following condition holds:

\[
f(x_t, y_t) - f(x_{t+1}, y_t) \geq \frac{1}{2L(y_t)} \|\nabla_x f(x_t, y_t)\|^2.
\]

If the optimization with respect to second block is solved analytically, \( x_{t+1} = \arg\min_x f(x, y_t) \). Then, if the function satisfies certain properties, the condition (1) is satisfied. The following corollary is straightforward application of the previous proposition.

**Corollary 3.** Assume the gradient function \( g_t(x) = f(x, y_t) \) is Lipschitz continuous. If for the \( x \) block we have \( x_{t+1} = \arg\min_x f(x, y_t) \), then

\[
f(x_t, y_t) - f(x_{t+1}, y_t) \geq \frac{1}{2L(y_t)} \|\nabla_x f(x_t, y_t)\|^2.
\]

**Proof.** Defining \( x' = x_t - \frac{1}{L(y_t)} \nabla_x f(x_t, y_t) \) and using the result of Proposition 2 together with the fact that \( f(x_{t+1}, y_t) \leq f(x', y_t) \), the proof is immediate.

The previous result shows that two-block coordinate-descent method has a good convergence behavior far from optimum when the gradient is large. The convergence of gradient-descent is similar but its constant can be much larger depending on the Lipschitz constant of the whole function, while here it depends only on the Lipschitz constant for one block. Another important advantage also observed empirically for coordinate-descent method in compare to gradient decent is that here for any region of data it behaves like a gradient descent with best step-size while in practice the best step-size for gradient descent method is not available. With the following remark we finish this section.

**Remark 4.** Without additional structure on the objective function, it seems hard to get better constants in the bound obtained. If the function \( f(x, y_t) \) is Lipschitz continuous gradient. Then if the function is twice differentiable, it is equal to say \( \nabla_{xx} f(x, y_t) \leq L(y_t) \) [3]. For the following quadratic function, the bound becomes equality:

\[
f(x, y_t) = c(y_t) + \nabla_x f(x_t, y_t)^T (x - x_t) + \frac{L(y_t)}{2}\|x - x_t\|^2.
\]

Therefore with the common assumption that assumes a Lipschitz constant for the set containing the whole iterations, it is hard to obtain better constant for the convergence rate.

**References**

[1] A. Beck and L. Tetrovashvili. On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization*, 23(4):2037–2060, 2013.

[2] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2004.

[3] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.