Journal of Computational Mathematics
Vol.38, No.3, 2020, 528–546.

http://www.global-sci.org/jcm
doi:10.4208/jcm.1912-m2016-0634

A STOCHASTIC MOVING BALLS APPROXIMATION METHOD OVER A SMOOTH INEQUALITY CONSTRAINT*

Leiwu Zhang
Department of Mathematics, Nanjing University, Nanjing 210023, China
Email: zaleiwu@sina.com

Abstract

We consider the problem of minimizing the average of a large number of smooth component functions over one smooth inequality constraint. We propose and analyze a stochastic Moving Balls Approximation (SMBA) method. Like stochastic gradient (SG) methods, the SMBA method's iteration cost is independent of the number of component functions and by exploiting the smoothness of the constraint function, our method can be easily implemented. Theoretical and computational properties of SMBA are studied, and convergence results are established. Numerical experiments indicate that our algorithm dramatically outperforms the existing Moving Balls Approximation algorithm (MBA) for the structure of our problem.

Mathematics subject classification: 65C20, 90C15, 90C25.
Key words: Smooth convex constrained minimization, Large scale problem, Moving Balls Approximation, Regularized logistic regression.

1. Introduction

In this article, we consider the following smooth convex optimization problem:

$$f_* := \min\{f(x) : x \in C\}, \tag{1.1}$$

where

$$f(x) = \frac{1}{p}\sum_{i=1}^{p} f_i(x), \quad C := \{x \in \mathbb{R}^n : \ g(x) \le 0\},$$

and $f_i, g : \mathbb{R}^n \mapsto \mathbb{R}$ are smooth convex functions. Problems of this form often arise in machine learning and statistics. A classical example is least-squares regression,

$$\min_{x\in C} \frac{1}{p}\sum_{i=1}^{p}(a_i^T x - b_i)^2,$$

where $a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$ are the data samples associated with a regression problem and $C := \{x \in \mathbb{R}^n : \ \|x\|^2 - \tau \le 0, \ \tau > 0\}$. Another important example is logistic regression,

$$\min_{x\in C} \frac{1}{p}\sum_{i=1}^{p} \log\Big(1 + \exp(-b_i a_i^T x)\Big),$$

where $a_i \in \mathbb{R}^n$ and $b_i \in \{-1, 1\}$ are the data samples associated with a binary classification problem and the constraint can be the same as C mentioned above.

* Received May 13, 2016 / Revised version received August 7, 2019 / Accepted December 24, 2019 / Published online March 4, 2020 /

For a function f, we denote by ∇f its gradient, let $F_l^{1,1}(C)$ be the class of convex functions which are continuously differentiable on C with lipschitz constant $l > 0$, i.e.,

$$\|\nabla f(x) - \nabla f(y)\| \leqslant l\|x - y\|, \quad \forall x, y \in C.$$

Further more if f is two times continuously differentiable, then we can choose (see [14]) $l = \max_{x \in C} \|\nabla^2 f(x)\|$. We denote $f \in S_{\mu,l}^{1,1}(C)$ if $f \in F_l^{1,1}(C)$ and for any $x,\ y \in C,\ \mu \geq 0$ we have:

$$f(x) \geq f(y) + \langle \nabla f(y), x - y\rangle + \frac{\mu}{2}\|x - y\|^2,$$

namely f is strongly convex on C. When $f \in S_{\mu,l}^{1,1}(C)$ in problem (1.1), a standard method of solving (1.1) is the full gradient method [14] for simple sets (here we say C is a simple set means that we can solve the following problem (1.2) easily). Given an initial point $x_0 \in C$, set

$$x_C(x_k; l) = \arg\min_{x \in C} \Big\{ f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + \frac{l}{2}\|x - x_k\|^2 \Big\}, \tag{1.2}$$

$$g_C(x_k; l) = l(x_k - x_C(x_k; l)),$$

then the full gradient method uses the following update rule for $k = 1, 2, \ldots$

$$x_{k+1} = x_k - \frac{1}{l} g_C(x_k; l).$$

The full gradient method satisfies [14, Theorem 2.2.8]

$$\|x_k - x_*\|^2 = \mathcal{O}((1 - \frac{\mu}{l})^k),$$

here we write $a_k = \mathcal{O}(b_k)$ means for nonnegative scalars $\{a_k\},\ \{b_k\}$, there exist constants $c_1 > c_2$ such that $c_2 b_k \leq a_k \leq c_1 b_k$ for every k. The accelerated full gradient method can be found in [15], as well as other extensions, variants and applications, see [3,6,18]. A shortcoming of the full gradient method is that its iteration cost of computing $\nabla f(x_k)$ scales linearly in p.

An effective alternative is the stochastic gradient (SG) method. The main advantage of SG is that they have an iteration cost which is independent of p, this is very suited for modern problems where p can be very large. The basic SG method uses the following form

$$x_{k+1} = \Pi_C(x_k - \alpha_k \nabla f_{i_k}(x_k)),$$

where Π_C denotes the Euclidean orthogonal projection onto C, $\alpha_k \geq 0$ is the step-size and the index i_k is sampled uniformly from $\{1, \ldots, p\}$. The randomly chosen gradient $\nabla f_{i_k}(x_k)$ obtains an unbiased estimate of the full gradient $\nabla f(x_k)$. Under standard assumptions [13], and for a properly chosen decreasing step-size sequence $\{\alpha_k\}$, the SG methods have an expected sub-optimality for convex objectives of

$$E[f(x_k)] - f(x_*) = \mathcal{O}(\frac{1}{\sqrt{k}}),$$

and for strongly convex objective functions we have

$$E[f(x_k)] - f(x_*) = \mathcal{O}(\frac{1}{k}).$$

We note that in these rate the expectations are taken with respect to the selection of i_k.