Journal of Computational Mathematics
Vol.36, No.3, 2018, 404–425.

http://www.global-sci.org/jcm
doi:10.4208/jcm.1708-m2017-0113

PARALLEL STOCHASTIC NEWTON METHOD*

Mojmír Mutný
Department of Informatics, ETH Zürich, Zürich, Switzerland
Email: mojmir.mutny@inf.ethz.ch
Peter Richtárik
School of Mathematics, University of Edinburgh, Edinburgh ETH9 3FD, UK and
Computer, Electrical and Mathematical Sciences & Engineering Department, KAUST, Saudi Arabia
Email: peter.richtarik@ed.ac.uk, peter.richtarik@kaust.edu.sa

Abstract

We propose a parallel stochastic Newton method (PSN) for minimizing unconstrained smooth convex functions. We analyze the method in the strongly convex case, and give conditions under which acceleration can be expected when compared to its serial counterpart. We show how PSN can be applied to the large quadratic function minimization in general, and empirical risk minimization problems. We demonstrate the practical efficiency of the method through numerical experiments and models of simple matrix classes.

Mathematics subject classification: 65K05, 65Y05, 68W10, 68W20.
Key words: optimization, parallel methods, Newton's method, stochastic algorithms.

1. Introduction

This work presents a novel parallel algorithm for minimizing a strongly convex function without constraints. This work is motivated by the possibility of better leveraging the structure in surrogate approximation, and the need for efficient optimization methods of high dimensional functions. The age of "Big Data" demands efficient algorithms to solve optimization problems that arise, for example, in fitting of large statistical models or large systems of equations. These new demands define open questions in algorithm design that make previously efficient algorithms obsolete.

For example, in this context, classical second order methods such as Newton method are not applicable as the inversion step of the algorithm is too costly ($O(n^3)$) to be performed in big data settings. Due to this reason, first-order algorithms enjoy huge popularity in the field of practicing optimizers, mainly in the field of machine learning. Recent years have shown that randomization and use of second-order information can lead to better convergence properties of algorithms. A prime example of this utilization are coordinate methods; to mention a few: [3, 14, 18, 20]. Another school, more traditionally grouped under term second-order, has seen a plethora of algorithms in recent year with modified LBFGS [6,8] methods to sub-sampled Newton methods [2, 11, 21–23], which coincide with the direction of this work.

In the current trend, computations are increasingly becoming parallelized, and the increase in performance is usually achieved by including more computing units solving a problem in parallel. Such architectures demand an efficient design of parallel algorithms that are able to exploit the parallel nature of computing clusters. An effort has been undertaken to provide

* Received May 3, 2017 / Revised version received July 4, 2017 / Accepted August 7, 2017 / Published online March 28, 2018 /

theoretical certificates on convergence of parallel optimization algorithms, to name a few, [19, 20], or from class of stochastic methods [16, 21, 29].

We chose to extend an existing algorithm that utilizes curvature information, called SDNA [13], which improves on standard coordinate methods such as SDCA [24] (of which parallel versions exist [20], [17]), and present theoretical certificates on parallelization efficiency of this algorithm along with analysis of special matrix classes. These analyses hint to better theoretical and practical than parallel coordinate descent method (PCDM) [20].

We apply the algorithm to general quadratic problems that arise in finite differences, and further we focus on big data application in machine learning, namely, Empirical Risk Minimization (ERM)

$$\min_{w\in\mathbb{R}^d}\left[P(w)=\frac{1}{n}\sum_{i=1}^{n}\phi_i(a_i^\top w)+\lambda g(w)\right], \tag{1.1}$$

which fits many of the statistical estimation models such as Ridge Regression. We present modified PSN for this type of problems that track dual and primal variables and gives ability to asses duality gap.

1.1. Contributions

The main contribution of this paper is the *design of a novel parallel algorithm* and its subsequent *novel theoretical analysis*. In the case of a smooth objective function, we present convergence analysis with proofs. The method in its simple serial case reduces to variants of algorithms introduced in [13] or [24]. For a different parallelization strategy in the case of convex quadratic optimization, see [21].

We identify parameters of the problem that determine its parallelizability and analyze them in special cases. To do this, we generalize two classes of quadratic optimization problems parametrized by one parameter and analytically calculate the convergence rates for them.

This work utilizes the research on sampling analyzed in paper [12], and is contrasted mainly with another parallel algorithm - parallel coordinate method (PCDM) analyzed in [18]. Furthermore, it generalizes further the class of coordinate methods beyond the generalization of blocks. In this work, the sampled blocks of the over-approximation are not fixed and can overlap. The choice of sampling leading to non-overlapping and fixed blocks has been analyzed previously in [4, 9] and mainly in [17].

1.2. Notation

Vectors. In this work, we use the convention that vectors in $\mathbb{R}^n$ are labeled with lowercase Latin letters. By $e_1, e_2, \dots e_n$ we denote the standard basis vectors in $\mathbb{R}^n$. The ith element of a vector $x \in \mathbb{R}^n$ therefore is $x_i = e_i^\top x$. The standard Euclidean inner product between vectors in $\mathbb{R}^n$ is given by $\langle x, y\rangle := x^\top y = \sum_{i=1}^n x_i y_i$.

Matrices. We use the convention that matrices in $\mathbb{R}^{n\times n}$ are labeled with uppercase bold Latin letters. By $\mathbf{I}$ we denote the identity matrix in $\mathbb{R}^{n\times n}$. The diagonal matrix with vector $w \in \mathbb{R}^n$ on the diagonal is denoted by $\mathbf{D}(w)$. We write $\mathbf{M} \succeq 0$ (resp. $\mathbf{M} \succ 0$) to indicate that $\mathbf{M}$ is symmetric positive semi-definite (resp. symmetric positive definite). Elements of a matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ are denoted in the natural way: $\mathbf{A}_{ij} := e_i^\top \mathbf{A} e_j$.