Journal of Computational Mathematics
Vol.40, No.2, 2022, 294–334.

http://www.global-sci.org/jcm
doi:10.4208/jcm.2012-m2020-0144

STOCHASTIC TRUST-REGION METHODS WITH TRUST-REGION RADIUS DEPENDING ON PROBABILISTIC MODELS*

Xiaoyu Wang[1)]
Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China; University of Chinese Academy of Sciences, Beijing 100049, China
Email: wxy@lsec.cc.ac.cn

Ya-xiang Yuan
State Key Laboratory of Scientific/Engineering Computing, Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
Email: yyx@lsec.cc.ac.cn

Abstract

We present a stochastic trust-region model-based framework in which its radius is related to the probabilistic models. Especially, we propose a specific algorithm termed STRME, in which the trust-region radius depends linearly on the gradient used to define the latest model. The complexity results of the STRME method in nonconvex, convex and strongly convex settings are presented, which match those of the existing algorithms based on probabilistic properties. In addition, several numerical experiments are carried out to reveal the benefits of the proposed methods compared to the existing stochastic trust-region methods and other relevant stochastic gradient methods.

Mathematics subject classification: 65K05, 65K10, 90C60.
Key words: Trust-region methods, Stochastic optimization, Probabilistic models, Trust-region radius, Global convergence.

1. Introduction

In this paper, we are concerned with the following unconstrained optimization problem

$$\min_{x\in\mathbb{R}^d} f(x), \tag{1.1}$$

where the objective function f is assumed to be smooth and bounded from below. But we only have access to the function value and its derivative information with some noise. In recent years, the expected risk minimization (ERM) problem, which is fundamental in the field of machine learning and statistics, has become the focus of many researchers. The ERM problems can be formulated as follows:

$$\min_{x\in\mathbb{R}^d} f(x) = \mathbb{E}\left[f(x;\xi)\right], \tag{1.2}$$

where $\mathbb{E}\left[\cdot\right]$ denotes the expectation taken with respect to the random noise variable ξ. However, because the probability distribution of ξ is unknown in advance, solving (1.2) is intractable

* Received June 1, 2020 / Revised version received September 16, 2020 / Accepted December 15, 2020 / Published online September 16, 2021 /
1) Corresponding author

directly. Usually only noisy information about the gradient of f is available. The empirical risk problem with a fixed amount of data (possibly very large) or the on-line setting problem where the data is flowing in sequentially, which involves an estimate of problem (1.2), is more often considered in practice. Throughout the whole paper, we mainly consider stochastic optimization methods to solve such kind of problems.

1.1. Related work

The classic stochastic optimization method is stochastic gradient descent (SGD) method, which dates back to the work by [1]. The method is prominent in large-scale machine learning due to its simplicity and low-cost computing. However, because of the variance introduced by random sampling, the sequence of learning rate (step-size) progressively diminishes both in theoretical analysis and practical implementation, which leads to slow convergence. Thus finding an appropriate learning rate is critical for the performance of the SGD method, but it is not easy in practice. There are many existing methods to deal with aforementioned issues, and most of them can be classified into the following categories.

Adaptive stochastic gradient methods, e.g., AdaGrad [2], RMSProp [3], Adam [4], AMSGrad [5], have emerged to alleviate the burden to tune step size, which are widely used in deep learning.

Variance reduction methods have been proposed to improve the convergence of the SGD method, such as SVRG [6], SAGA [7] and SARAH [8]. Especially, they have achieved linear convergence rate when solving strongly convex problems, which is a stronger result than that of the SGD method. Furthermore, these methods have also been extended to solve nonconvex problems such as the deep neural networks [9–11]. The variance reduction technique is applicable to the problem with a large but fixed sample set, for which the full gradient has to be calculated as a compromise to achieve the significant variance reduction. Hence they are not easy to fit on the on-line setting like the SGD method and adaptive gradient methods.

Trust-region based algorithms. Recently, with the success of deep neural networks, the development and analysis of methods for nonconvex problems have attracted tremendous attention. As we know, the traditional trust-region method is a class of well-established and effective method in nonlinear optimization [12, 13]. An advantage of trust-region methods is that they can be efficient on nonconvex and ill-conditioned problems since they can make use of curvature information. Besides, due to the boundedness of the trust-region, the Hessian approximation matrix is not required to be positive definite. The trust-region based framework has already been considered to solve machine learning problems [14–20]. A saddle free Newton (SFN) method [16] is proposed to exploits the exact Hessian information to escape saddle points. However its computation is high cost for large-scale and high-dimension problems. A two-stage subspace trust-region approach [17] is proposed to train deep neural networks, in which the local second-order model is conducted based on the partial information computed from a subset of the data. But the approach lacks theoretical guarantees. Some literatures, for example, [14, 15], focus a specific class of machine learning problems, which need to utilize the accurate derivative information to construct good models, and the accurate function values to obtain good estimators. In [18], inexact Hessian information is incorporated into the trust-region framework but the exact gradient and function values are required to be computed per iteration. In [19] and [20], the authors construct the inexact models to satisfy some first-order accurate conditions with sufficiently high probability when building the trust-region subproblem. The complexity of line search and cubic regularization algorithms is analysed [21]