Journal of Computational Mathematics
Vol.43, No.5, 2025, 1045–1062.

http://www.global-sci.org/jcm
doi:10.4208/jcm.2505-m2025-0004

# MINI-BATCH STOCHASTIC CONJUGATE GRADIENT ALGORITHMS WITH MINIMAL VARIANCE*

Caixia Kou
*Key Laboratory of Mathematics and Information Networks, Beijing University of Posts and Telecommunications, Ministry of Education;*
*School of Mathematical Sciences, Beijing University of Posts and Telecommunications, Beijing 100876, China*
*Email: koucx@bupt.edu.cn*
Feifei Gao
*School of Mathematical Sciences, Beijing University of Posts and Telecommunications, Beijing 100876, China;*
*Department of Statistics, Zhejiang Gongshang University Hangzhou College of Commerce, Hangzhou 310018, China*
Yu-Hong Dai[1)]
*LSEC, ICMSEC, AMSS, Chinese Academy of Sciences, Beijing 100190, China;*
*School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China*
*Email: dyh@lsec.cc.ac.cn*

**Abstract**

Stochastic gradient descent (SGD) methods have gained widespread popularity for solving large-scale optimization problems. However, the inherent variance in SGD often leads to slow convergence rates. We introduce a family of unbiased stochastic gradient estimators that encompasses existing estimators from the literature and identify a gradient estimator that not only maintains unbiasedness but also achieves minimal variance. Compared with the existing estimator used in SGD algorithms, the proposed estimator demonstrates a significant reduction in variance. By utilizing this stochastic gradient estimator to approximate the full gradient, we propose two mini-batch stochastic conjugate gradient algorithms with minimal variance. Under the assumptions of strong convexity and smoothness on the objective function, we prove that the two algorithms achieve linear convergence rates. Numerical experiments validate the effectiveness of the proposed gradient estimator in reducing variance and demonstrate that the two stochastic conjugate gradient algorithms exhibit accelerated convergence rates and enhanced stability.

*Mathematics subject classification:* 49M37, 90C25.
*Key words:* Stochastic gradient descent, Minimal variance, Stochastic conjugate gradient, Stochastic gradient estimator.

## 1. Introduction

Consider the empirical risk minimization (ERM) problem

$$\min_{\omega\in\mathbb{R}^d} f(\omega) = \frac{1}{n}\sum_{i=1}^{n} f_i(\omega), \tag{1.1}$$

* Received January 3, 2025 / Revised version received March 31, 2025 / Accepted May 12, 2025 / Published online June 23, 2025 /
1) Corresponding author

where $\omega \in \mathbb{R}^d$ represents the decision vector, $f_i(\omega) : \mathbb{R}^d \to \mathbb{R}$ is the loss function for the $i$-th sample, $n$ is the number of samples. This problem is widely raised from machine learning for building and training models to minimize some loss on a training dataset.

In the case when the number of samples $n$ is extremely large, the calculating of full gradients $\nabla f(\omega)$ becomes expensive. Stochastic gradient descent, proposed in the seminal work [23], randomly generates one sample of the loss function to update the iteration. As the gradient direction oscillates in the SGD method, the mini-batch SGD was introduced in [19,25], which is to select some subset of samples instead of a single sample at each iteration. This technique is able to reduce gradient variance and hence increases the stability of the method. The stochastic average gradient (SAG) method [24] needs to store the gradient of each sample and utilizes the average of historical gradients to reduce variance. The SAG method achieves a linear convergence rate in strongly convex problems, which is much faster than SGD. The stochastic variance reduced gradient (SVRG) [15] does not need to maintain all gradients in memory, but performs gradient replacement every $m$ iterations and achieves the linear convergence as well. The stochastic average gradient amélioré (SAGA) method [6], like SAG, needs to store the gradients of all samples, yet retains unbiasedness. It is worth noting that, although the variance produced by the gradient estimation in SAG is $1/n^2$ times that of SAGA, this reduction in variance comes at the expense of introducing a non-zero bias [6]. Other variants of variance reduction algorithms, as well as through adaptive learning rates, can be found in [7, 16, 20, 26, 28, 30, 32–34]. For more SGD algorithms, we refer to [2, 18, 27, 29] and the references therein.

This leads us to focus on how to estimate the gradients in the best way. We believe that an ideal gradient estimator should be unbiased and exhibit low variance. Therefore, this paper mainly focuses on designing unbiased stochastic gradient estimators that minimize variance, and applies them to stochastic conjugate gradient algorithms, aligning with our research interests.

Recalling the classical conjugate gradient method, it is widely used for large-scale optimization problems due to its fast convergence and low storage requirement. Well-known formulas for the conjugate gradient parameter are Fletcher-Reeves (FR) [8], Polak-Ribière-Polyak (PRP) [21, 22], Hestenes-Stiefel (HS) [12], and Dai-Yuan (DY) [5] ones. A range of hybrid conjugate gradient methods have been developed, including the TAS method by Touati-Ahmed and Storey [31], the PRP-FR method (a hybrid version of PRP and FR) proposed by Hu and Storey [13], and the GN method by Gilbert and Nocedal [9]. These hybrid methods combine the properties of standard conjugate gradient methods to acquire new characteristics, facilitating rapid convergence to the solution [1]. By exploring the second-order information and analyzing the relationship between conjugate gradient directions and quasi-Newton directions, more efficient conjugate gradient methods have been developed in Dai and Kou [4], and Hager and Zhang [10, 11]. For more details on conjugate gradient methods, we refer to [1, 3].

Applying the conjugate gradient algorithms to solve ERM problem (1.1), two stochastic conjugate gradient algorithms, the conjugate gradient with variance reduction (CGVR) [14] and the stochastic conjugate gradient algorithm (SCGA) [17], have been proposed. It is found that these stochastic conjugate gradient methods demonstrate a faster decrease in the objective function value compared to SGD. Motivated by these findings, in the second part of this paper, we introduce two stochastic conjugate gradient algorithms with minimal variance by integrating the new proposed gradient estimator.

The rest of this paper is organized as follows. Section 2 proposes a new family of stochastic gradient estimators with unbiasedness and a new stochastic gradient estimator with minimal variance. Section 3 describes the details of the two stochastic conjugate gradient algorithms