NUMERICAL MATHEMATICS: Theory, Methods and Applications
Numer. Math. Theor. Meth. Appl., Vol. **1**, No. 4, pp. 460-474 (2008)

A Stochastic Approximation Frame Algorithm with Adaptive Directions

Zi Xu[1] and Yu-Hong Dai[2]*

[1] *Department of Mathematics, College of Sciences, Shanghai University, Shanghai, 200444, P.R.China.*
[2] *State Key Laboratory of Scientific and Engineering Computing, Institute of Computational Mathematics and Scientific/Engineering Computing, The Academy of Mathematics and Systems Science, Chinese Academy of Sciences, P.O.Box 2719, Beijing, 100190, P.R.China.*

Received 24 October, 2007; Accepted (in revised version) 17 April, 2008

Abstract. Stochastic approximation problem is to find some root or extremum of a nonlinear function for which only noisy measurements of the function are available. The classical algorithm for stochastic approximation problem is the Robbins-Monro (RM) algorithm, which uses the noisy evaluation of the negative gradient direction as the iterative direction. In order to accelerate the RM algorithm, this paper gives a frame algorithm using adaptive iterative directions. At each iteration, the new algorithm goes towards either the noisy evaluation of the negative gradient direction or some other directions under some switch criterions. Two feasible choices of the criterions are proposed and two corresponding frame algorithms are formed. Different choices of the directions under the same given switch criterion in the frame can also form different algorithms. We also proposed the simultanous perturbation difference forms for the two frame algorithms. The almost surely convergence of the new algorithms are all established. The numerical experiments show that the new algorithms are promising.

AMS subject classifications: O242.1
Key words: Stochastic approximation, conjugate gradient, adaptive directions.

1. Introduction

Stochastic approximation algorithm provides a simple and effective approach for finding root or minimum of function whose evaluations are contaminated with noise. Consider a n-dimensional loss function $f : \mathscr{R}^n \to \mathscr{R}$, with gradient $g : \mathscr{R}^n \to \mathscr{R}^n$. We have that $g(x) = 0$ if and only if $x = x^*$ when f has a unique local minimizer $x^* \in \mathscr{R}^n$. If the direct noisy estimate of the gradient function $\tilde{g}_k$ is available, the Robbins-Monro(RM) algorithm [1](extended by Blum [2] to multidimensional systems) estimates a root of g with the following recursion:

$$x_{k+1} = x_k - \alpha_k \tilde{g}_k, \tag{1.1}$$

*Corresponding author. *Email addresses:* `xuzi@lsec.cc.ac.cn` (Z. Xu), `dyh@lsec.cc.ac.cn` (Y.-H. Dai)

$$\tilde{g}_k = g(x_k) + \varepsilon_k, \tag{1.2}$$

where ε_k is the noise and α_k is a sequence that satisfies

$$\alpha_k > 0, \quad \sum_{k\geq 1} \alpha_k = \infty, \quad \sum_{k\geq 1} \alpha_k^2 < \infty. \tag{1.3}$$

Since the direct noisy measurements $\tilde{g}_k$ are sometimes not available, Kiefer and Wolfowitz [3] introduced the finite difference form of the RM algorithm, which employs an estimator for the gradient denoted by $\hat{g}(x_k)$, whose ith component is given by

$$\hat{g}_i(x_k) = \frac{\tilde{f}(x_k + c_k e_i) - \tilde{f}(x_k - c_k e_i)}{2c_k}, \tag{1.4}$$

where e_i is the unit vector along the ith axis and $\tilde{f}$ is a noisy measurement of the function value f. We can call this algorithm finite difference stochastic approximation (FDSA) algorithm or KW algorithm simply. The almost surely convergence of the KW algorithm is also given by Kiefer and Wolfowitz [3]. The major disadvantage of the KW algorithm is that it requires $2n$ measurements of the function value per iteration. By contrast, the random direction stochastic approximation (RDSA) algorithm first given by Kushner and Clark [4], needs only two measurements per iteration. It has the following recursion:

$$x_{k+1} = x_k - \alpha_k \left[\frac{\tilde{f}(x_k + c_k \xi_k) - \tilde{f}(x_k - c_k \xi_k)}{2c_k} \right] \xi_k. \tag{1.5}$$

A special case of the RDSA algorithm is the simultaneous perturbation stochastic approximation (SPSA) algorithm introduced by Spall [5] which employs the estimator:

$$\hat{g}(x_k) = \left[\frac{\tilde{f}(x_k + c_k \xi_k) - \tilde{f}(x_k - c_k \xi_k)}{2c_k} \right] \zeta_k, \tag{1.6}$$

where ξ_k is chosen from a distribution that has to satisfy some particular constraints, and the ith component of ζ_k are given by

$$\zeta_k^{(i)} = 1/\xi_k^{(i)}. \tag{1.7}$$

Since in fact the Bernoulli distribution is the only choice that has ever been advocated for SPSA, SPSA is a special case of RDSA, though it does bear remarking that the use of a Bernoulli distribution with RDSA had not been suggested until after SPSA had been introduced. The FDSA, RDSA and SPSA algorithm exhibit similar convergence properties.

The RM algorithm is a classical stochastic approximation algorithm and exhibits the property that it converges to a stationary point almost surely. The major disadvantage of RM algorithm and its difference forms including the FDSA, RDSA and SPSA algorithms are their slow speed of convergence. There have been many efforts to accelerate the RM algorithm. Most of them consist in the choice of the step size α_k, such as the Kesten algorithm