12th IFAC/IEEE Workshop on Discrete Event Systems
Cachan, France. May 14-16, 2014

A New Hybrid Stochastic Approximation Algorithm ⋆

Marie Chau * **Huashuai Qu** ** **Michael C. Fu** ***

* *Department of Mathematics, University of Maryland, College Park, College Park, MD 20742 USA (e-mail: mchau@math.umd.edu)*
** *Department of Mathematics, University of Maryland, College Park, College Park, MD 20742 USA (e-mail: huashuai@math.umd.edu)*
*** *Robert H. Smith School of Business & Institute for Systems Research, University of Maryland, College Park, College Park, MD 20742 USA, (e-mail: mfu@isr.umd.edu)*

Abstract: We introduce Secant-Tangents AveRaged (STAR) Stochastic Approximation (SA), a new SA algorithm that estimates the gradient using a hybrid estimator, which is a convex combination of a symmetric finite difference and an average of two direct gradient estimators. For the deterministic weight sequence that minimizes the variance of the STAR gradient, we prove that for quadratic functions, the mean squared error (MSE) of the STAR-SA algorithm using this weight sequence is strictly less than that of the classical SA methods of Robbins-Monro (RM) and Kiefer-Wolfowitz (KW). We also prove convergence of the STAR-SA algorithm for general concave functions. Furthermore, we illustrate its effectiveness through numerical experiments by comparing the MSE of the STAR-SA algorithm against RM and KW for simple quadratic functions with various steepness and noise levels.

Keywords: stochastic approximation, simulation, stochastic optimization, direct gradients, algorithm, convergence, finite difference

1. INTRODUCTION

Consider the stochastic optimization problem

$$\max_{\mathbf{x}\in\Theta} f(\mathbf{x}) = E[\tilde{f}(\mathbf{x})], \tag{1}$$

where $\tilde{f}(\mathbf{x})$ is a noisy observation of $f(\mathbf{x})$, $\Theta \in \mathbb{R}^d$ is a continuous parameter space, and the objective is to find $\mathbf{x}^*$ maximizing f. In this paper, we only consider the case when $d = 1$. Stochastic Approximation (SA) is a classic iterative method used to solve stochastic optimization problems with the recursion

$$x_{n+1} = \Pi_\Theta\left(x_n + a_n\widehat{\nabla} f(x_n)\right), \tag{2}$$

where $\Pi_\Theta(x)$ is a projection of x back into the feasible region Θ if $x \notin \Theta$, a_n is called the gain size or step size, and $\widehat{\nabla} f(x_n)$ is an estimate of the gradient $\nabla f(x_n)$. Two prototypical methods, namely the Robbins-Monro (RM) and Kiefer-Wolfowitz (KW) algorithms, introduced in the early 1950's, estimate $\nabla f(x_n)$ using unbiased direct gradient estimates and finite differences, respectively. Under certain conditions, RM and KW have the respective asymptotic convergence rates $O(n^{-1/2})$ and $O(n^{-1/3})$.

There has been extensive research conducted to improve the performance and robustness of these algorithms. Although the recursion in (2) appears to be simple, the choice of step size sequence $\{a_n\}$, gradient estimator $\widehat{\nabla} f(x_n)$, and projection operator Π_Θ has a significant impact on the performance of the algorithm. It is widely-known that the practical performance of the original RM and KW is extremely sensitive to the initial choice of step size; therefore, it is imperative to choose it appropriately. The gain sequence $\{a_n\}$ could be deterministic or adaptive. A common step size considered is $a_n = \theta_a/n$ where $\theta_a \in \mathbb{R}^+$. Adaptive sequences are more sophisticated and dynamically adjust according to the ongoing performance of the algorithm. One well-known example is the rule proposed by Kesten (1958), which decreases the step size only when there is a directional change in the iterates. The concept behind this is that if the estimates continue to move in the same direction, it suggests that the estimates are not in close proximity of the optimum; therefore, the step sizes should not decrease. Sardis (1970) extended this notion and increases the step size when this occurs to increase the momentum towards the optimum.

Originally, the asymptotic theory for SA only considered functions satisfying specific global conditions; however, later research has shown that it is only necessary for the requirements to hold on a compact set Θ that contains the optimum. Therefore, the projection operator Π_Θ is particularly important in the constrained optimization setting. Since the optimum is unknown, the compact set must be large enough to increase the likelihood that $x^* \in \Theta$; however, enlarging the search space may deteriorate the performance of an algorithm. Chen and Zhu (1986) proposed the idea to project the iterates onto a predetermined

⋆ This work was supported in part by the National Science Foundation (NSF) under Grants CMMI-0856256 and EECS-0901543, and by the Air Force Office of Scientific Research (AFOSR) under Grant FA9550-10-1-0340.

fixed point once the magnitude of the estimator surpassed a specific value, which increases after each occurrence. This method converges asymptotically, but in practice, it has its drawbacks. In particular, when the iterates are projected onto an arbitrary fixed point, the algorithm loses all of the progress gained prior to the projection, in addition to reducing the step sizes. In a sense, the algorithm restarts with an even smaller step size, so the iterates move at a slower rate. Yin and Zhu (1989) later modified this operator and proposed to project the iterates onto a predetermined compact set, so at least the iterates do not start from the same initial position after a projection. Andradóttir (1995) extended this idea even further and introduced a projection operator that projects the iterates onto an adaptively increasing sequence of compact sets (i.e., $\Theta_k \subseteq \Theta_{k+1}$) such that $\Theta = \cup\Theta_k$.

Various gradient estimators have been proposed to increase the accuracy and robustness. Andradóttir (1990) introduced a new gradient estimator to SA, which allows the algorithm to converge under more general assumptions than KW and RM. If unbiased estimators are employed using $a_n = \theta_a/n$ where $\theta_a > 0$, then this algorithm has an asymptotic convergence rate of $O(n^{-1/2})$. Spall (1992) developed the Simultaneous Perturbation Stochastic Approximation (SPSA) method specifically for multivariate stochastic optimization problems. Similar to KW, SPSA estimates the gradient using finite differences; however, it randomly perturbs all of the parameters simultaneously (hence, the name of the method).

Asymptotically, RM converges faster than KW; however, in practice, it is difficult to determine which algorithm should be implemented when both are applicable. RM requires more information about the system, so if the direct gradient is unavailable or unreliable, KW is the choice by default. KW is easy to implement, but it requires an additional task of picking a sequence of finite difference step sizes $\{c_n\}$. However, as a rule of thumb, if the direct gradient is available, it should be incorporated into the algorithm. Since the performance of one algorithm is not always superior to the other, we propose an algorithm that is a hybrid of the two. The new hybrid algorithm, called Secant-Tangents AveRaged (STAR)-SA, follows the recursion in (2) and uses a convex combination of a symmetric finite difference estimator (secant) and the average of two unbiased direct gradient estimators (tangents) to estimate $\widehat{\nabla} f(x_n)$.

The new STAR gradient estimator involves function and gradient estimates on two points, $x_n + c_n$ and $x_n - c_n$, for each $\widehat{\nabla} f(x_n)$. We assume that direct gradients are available when we apply KW; therefore, we obtain two direct gradient estimators in addition to the two performance measurements from simulation. As a result, aside from the cost of gradient estimation, the STAR-SA algorithm is computationally equivalent to KW since both require two sample paths and twice the cost of the original RM algorithm since RM only uses one. Although each iteration is more expensive, the variance of the STAR gradient is less than the variance of both the symmetric finite difference estimator and direct gradient estimator with the appropriate weights. We choose the weights that minimize the variance of the gradient estimator, and this is shown to be theoretically optimal in terms of MSE for quadratic cases. More importantly, we prove convergence in MSE of the STAR-SA algorithm under mild conditions.

We conduct two sets of numerical experiments to test the performance of the new STAR-SA algorithm against the classic KW and RM methods when the underlying function is of the form $f(x) = -ax^2$, where $a > 0$. In our analysis, we double the number of iterations for RM, since it only requires one sample path for each run, whereas STAR-SA and KW require two. We implement all three algorithms under a constrained setting using various parameter settings, including different combinations of function and gradient noise levels and estimate the MSE using 1000 sample runs.

The rest of the paper is organized as follows. We briefly review the two SA algorithms, RM and KW, and provide the original convergence results in Section 2. In Section 3, we introduce our new STAR gradient and show that if the weights are chosen to minimize the variance, then it can be strictly less than the variance of the gradient estimator in KW and RM. We also present MSE convergence results of the STAR-SA algorithm for general functions. Analytically, we calculate the exact MSE of STAR-SA, KW, and RM for quadratic functions and show that the MSE of STAR-SA is strictly less than that of KW and RM under certain conditions in Section 3.2. In Section 4, we describe our empirical tests and discuss the numerical results. Finally, we conclude in Section 5.

2. CLASSIC ALGORITHMS

In this section, we will summarize the classic RM and KW methods.

2.1 The Robbins-Monro Method

Originally, the Robbins-Monro method was introduced to solve the root-finding problem

$$M(x) = \alpha$$

for $x \in \mathbb{R}$ where $M(x)$ is a monotone function and $\alpha \in \mathbb{R}$. However, it was later applied to a specific case of root-finding in the stochastic optimization setting, where the objective is to optimize a stochastic objective function $f(x)$ in (1) by setting $M(x) = \nabla f(x)$ and $\alpha = 0$. RM solves this problem iteratively as in (2) by replacing $\widehat{\nabla} f(x_n)$ in (2) with an unbiased estimator and the output is taken as the last iterate, x_n, where N is the number of iterations. RM is similar to deterministic steepest descent and the Newton-Raphson method. However, in RM, the direct gradient measurements are still approximations to the actual gradient because of the presence of noise. This algorithm has the potential to converge asymptotically at a rate of $O(n^{-1/2})$ for concave functions.

Robbins and Monro (1951) established mean squared convergence of the RM algorithm, assuming $\nabla f(x)$ has a unique root at x^*. Suppose $\widehat{\nabla} f(x)$ is an unbiased gradient estimator, i.e., $\nabla f(x) = E[\widehat{\nabla} f(x)]$ Assume

1. $\{a_n\}$ is a fixed sequence of positive constants such that $\sum_{n=1}^{\infty} a_n^2 < \infty$.
2. $\nabla f(x) \leq 0$ for $x > x^*$ and $\nabla f(x) \geq 0$ for $x < x^*$.