Proceedings of the 1991 Winter Simulation Conference
Barry L. Nelson, W. David Kelton, Gordon M. Clark (eds.)

A PROJECTED STOCHASTIC APPROXIMATION ALGORITHM

Sigrún Andradóttir

Department of Industrial Engineering
University of Wisconsin – Madison
1513 University Avenue
Madison, WI 53706

ABSTRACT

Classical stochastic approximation algorithms often diverge because of boundedness problems. The standard approach to preventing this is to project the sequence generated by the algorithm onto a predetermined compact set K. However, in the typical application, the approximate location of the solution is not known. To minimize the probability that the solution lies outside the set K, it is therefore necessary to let K be large. This can seriously curtail the efficiency of the algorithm. We propose a new stochastic approximation algorithm which bounds the sequence of estimates of the solution to an increasing sequence of sets. This eliminates the possibility of bounding the algorithm to a set which doesn't contain the solution. Furthermore, it is possible to let the initial set be small, which can result in improved empirical performance.

1 INTRODUCTION

Stochastic approximation algorithms are concerned with the problem of finding the root of a function h whose values are not known analytically and therefore have to be estimated or measured. Their goal is to obtain a sequence of iterates $\{\theta_n\}$ that converges almost surely to the solution. The function h is typically evaluated using simulation. It is not assumed to have any particular structure and the distribution of the estimates of the function values is unknown.

Stochastic approximation algorithms can be used to optimize the performance of a complex stochastic system with respect to continuous decision parameters. If $f : \Re^d \to \Re$ is the performance measure of interest, then $h = \nabla f$ if we are interested in determining the minimum of f, and $h = -\nabla f$ if we want to find the maximum of f. Estimates of the gradient ∇f can be obtained by using finite differences, or if the system is being analyzed through simulation, gradient estimates can be obtained by using either the likelihood ratio method (see for instance Glynn (1989) and Andradóttir (1990a, 1991b)), or perturbation analysis (see for instance Suri (1989) and Glasserman (1991)). The likelihood ratio method and perturbation analysis have several advantages over finite differences. In particular, gradient estimates can be obtained by observing a single sample path of the system under study, and when the gradient estimates produced by these methods are unbiased, the stochastic approximation algorithm will have a faster asymptotic convergence rate.

This paper is organized as follows: classical stochastic approximation algorithms are discussed in Section 2. The projected algorithm is introduced in Section 3 and Section 4 contains a comparison of the empirical performance of these algorithms. Finally, Section 5 contains some concluding remarks.

2 CLASSICAL STOCHASTIC APPROXIMATION ALGORITHMS

Classical stochastic approximation algorithms obtain a sequence $\{\theta_n\}$ of estimates of the solution as described below:

Algorithm 1

Step 0: *Choose* $\theta_1 \in \Re^d$.

Step 1: *Given* θ_n, *generate an estimate* Y_n *of* $h(\theta_n)$.

Step 2: *Compute*

$$\theta_{n+1} = \theta_n - a_n Y_n. \tag{1}$$

Step 3: *Let* $n = n + 1$ *and go to step 1.*

Here $\{a_n\}$ is a predetermined sequence of positive constants. We assume that $a_n \to 0$ and $\sum_{n=1}^{\infty} a_n = \infty$. These assumptions are needed to guaranty the convergence of Algorithm 1. If $\sum_{n=1}^{\infty} a_n < \infty$ and

the sequence Y_n is bounded, then the iterates $\{\theta_n\}$ will all lie in a compact sphere with center θ_1. This means that the algorithm can converge only if θ_1 is close enough to the solution of the problem. The assumption that $a_n \to 0$ is needed to dampen out the effect of the errors in the function evaluations to obtain almost sure convergence to the solution of the problem as the number of iterations goes to infinity. Typically, a_n is chosen as a/n, where a is a positive constant.

The Robbins-Monro algorithm (Robbins and Monro 1951) and the Kiefer-Wolfowitz algorithm (Kiefer and Wolfowitz 1952) are the two most commonly used stochastic approximation algorithms. They are both special cases of Algorithm 1. The Robbins-Monro algorithm is more general than the Kiefer-Wolfowitz algorithm in that the Kiefer-Wolfowitz algorithm can only be applied to solve optimization problems (it is designed to determine the root of the gradient of the objective function), whereas the Robbins-Monro algorithm can be used to solve more general root-finding problems. When applied to stochastic optimization, the two algorithms differ in how they estimate the gradient of the objective function. The Robbins-Monro algorithm estimates the gradient directly, whereas the Kiefer-Wolfowitz algorithm uses finite differences to estimate the gradient. These algorithms are far from ideal. It is well known that they converge extremely slowly when the objective function is very flat and it is possible show that they do not necessarily converge when the objective function is steep. Indeed, if $d = 1$, $|\theta_1| > \sqrt{3/a}$, and for all n, $a_n = a/n$ and $Y_n = \theta_n^3$, where $a > 0$, then one can show that the sequence generated by Algorithm 1 satisfies $|\theta_n| \geq |\theta_1| n!$ for all n (Andradóttir 1990a, 1990b). This shows that Algorithm 1 does not necessarily converge when applied to finding the root of the function $h(\theta) = \theta^3$. The problem is that the function h is quite steep, so the length of the step taken in an iteration ($a_n|h(\theta_n)|$) can be very large. The algorithm generates a sequence of estimates of the solution that goes to infinity in norm, fluctuating around the solution.

The standard approach to ensure that the sequence generated by Algorithm 1 is bounded is to project it onto a predetermined compact set K, so equation (1) is replaced by

$$\theta_{n+1} = \pi_K(\theta_n - a_n Y_n), \tag{2}$$

where, for all $\theta \in \Re^d$, $\pi_K(\theta)$ denotes the point in K which is closest to θ. (We can assume, without loss of generality, that K is convex, so the projection π_K is well defined.) The problem with this approach is that in the typical application, the approximate location of the solution is not known. To minimize the probability that the solution lies outside the set K, it is therefore necessary to let K be large. However, as shown in Section 4, this can seriously curtail the efficiency of the algorithm. In Section 3, we propose a new algorithm which is based on the idea of bounding the sequence $\{\theta_n\}$, but we will bound it to an increasing sequence of sets. This approach was proposed earlier by Chen and Zhu (1986). Their method uses the following equation to update θ_n:

$$\theta_{n+1} = (\theta_n - a_n Y_n) I_{\{\|\theta_n - a_n Y_n\| \leq M_{\sigma(n)}\}} + \overline{\theta} I_{\{\|\theta_n - a_n Y_n\| > M_{\sigma(n)}\}},$$

where I_A is the indicator random variable, $\overline{\theta} \in \Re^d$ is fixed, $\{M_n\}$ is an increasing sequence of positive real numbers such that $M_n \to \infty$ as $n \to \infty$, and $\sigma(1) = 1$ and for $n > 1$,

$$\sigma(n) = 1 + \sum_{j=1}^{n-1} I_{\{\|\theta_j - a_j Y_j\| > M_{\sigma(j)}\}}.$$

The main problem with this approach is that whenever $\|\theta_n - a_n Y_n\| > M_{\sigma(n)}$, the algorithm returns to a fixed point $\overline{\theta}$. This means that it is possible to return arbitrarily often to $\overline{\theta}$ before the algorithm converges. This can make the convergence of the algorithm very slow, particularly when the function h is very flat and $\|\theta^*\| \gg M_1$. The algorithm proposed in Section 3 does not have this property.

Another stochastic approximation algorithm that deserves to be mentioned is the scaled algorithm proposed by Andradóttir (1990a, 1990b, 1991a). This algorithm uses the following equation to update θ_n:

$$\theta_{n+1} = \theta_n - a_n \left[\frac{Y_n^1}{\max\{\epsilon, \|Y_n^2\|\}} + \frac{Y_n^2}{\max\{\epsilon, \|Y_n^1\|\}} \right]$$

where ϵ is a positive real number and Y_n^1 and Y_n^2 are conditionally independent estimates of $h(\theta_n)$ (conditional on $\theta_1, \ldots, \theta_n$). This algorithm converges under much more general conditions on the function h than Algorithm 1, while maintaining the same asymptotic rate of convergence ($n^{-1/2}$). Moreover, empirical evidence shows that it sometimes approaches the solution much more rapidly than Algorithm 1.

3 A PROJECTED STOCHASTIC APPROXIMATION ALGORITHM

In the previous section, we discussed the problems associated with classical stochastic approximation algorithms: they converge extremely slowly when applied to flat functions and they often diverge when applied

to steep functions. To make these algorithms more robust, the sequence that they generate is often projected onto a fixed compact set K. The problem with this approach is that usually K is large, which may reduce the efficiency of the algorithm. We propose instead to bound θ_n to K_n, for all n, where $\{K_n\}$ is a sequence of compact sets such that $K_n \subseteq K_{n+1}$ for all n and $\cup_{n=1}^{\infty} K_n = \Re^d$. This approach has several advantages: since $\cup_{n=1}^{\infty} K_n = \Re^d$, we have eliminated the possibility of not obtaining convergence because the solution lies outside the set K. Furthermore, it is possible to let K_1 be small and to increase the size of the sets K_n slowly. This often results in an improved empirical performance (see Section 4).

Let $\{b_n\}$ be a non-decreasing sequence of positive constants such that $\lim_{n\to\infty} b_n = \infty$, let $K_n = \{\theta \in \Re^d : ||\theta - \theta_1|| \leq b_n\}$ and let π_n denote the orthogonal projection onto the set K_n ($\pi_n(\theta) = \theta$ when $||\theta - \theta_1|| \leq b_n$ and $\pi_n(\theta) = \theta_1 + b_n(\theta - \theta_1)/||\theta - \theta_1||$ when $||\theta - \theta_1|| > b_n$). Consider the following algorithm:

Algorithm 2

Step 0: *Choose* $\theta_1 \in \Re^d$.

Step 1: *Given* θ_n, *generate an estimate* Y_n *of* $h(\theta_n)$.

Step 2: *Compute*

$$\theta_{n+1} = \pi_n(\theta_n - a_n Y_n).$$

Step 3: *Let* $n = n + 1$ *and go to step 1.*

Algorithm 2 ensures that $||\theta_n - \theta_1|| \leq b_{n-1}$ for $n \geq 2$. It is possible to show that when the sequence $\{b_n\}$ is chosen appropriately, this algorithm converges for a large class of functions h. In particular, if $b_n = b\log(n+1)$ for all n, where $b > 0$, then it is possible to show that Algorithm 2 converges under much more general conditions on the function h than Algorithm 1. As was the case for Algorithm 1, the sequence $\{a_n\}$ of Algorithm 2 satisfies $a_n \to 0$ as $n \to \infty$ and $\sum_{n=1}^{\infty} a_n = \infty$.

4 EMPIRICAL WORK

In this section, we compare the empirical behavior of the projected version of Algorithm 1 with that of Algorithm 2 when applied to solve the problem of finding $\theta \in \Re$ such that $E_\theta\{X\} = 0$, where X has a normal distribution with mean θ^3 and variance 1. The solution of this problem is $\theta^* = 0$. As discussed in Section 2, Algorithm 1 does not converge on this problem in general. We can however obtain convergence by projecting the sequence $\{\theta_n\}$ generated by this algorithm to a bounded interval containing θ^* (see equation (2)). We therefore want to compare the performance of the projected version of Algorithm 1 with that of Algorithm 2, when applied to solving this problem. For this purpose, we conducted the experiment described below. Both algorithms were started at $\theta_1 = 10$ and run for 2000 iterations. This process was repeated 1000 times and Figure 1 shows the average performance of the two algorithms (the x-axis shows the number of iterations (n) and the y-axis shows the distance ($|\theta_n|$) of the estimate θ_n to the solution $\theta^* = 0$). To be able to compare the performance of the two algorithms more meaningfully, we used common random numbers to estimate the errors in the function evaluations. This means that for $1 \leq m \leq 1000$ and $1 \leq n \leq 2000$, the error in the nth function evaluation in the mth replication is the same for both algorithms. For the purposes of this experiment, we let $a_n = 1/n$ and $b_n = 10 \times \log(n+1)$, for all n. To evaluate the sensitivity of the projected version of Algorithm 1 to the bounds on its variables, we ran the algorithm three times, bounding the sequence $\{\theta_n\}$ to the intervals $[-M, M]$, where $M = 20, 50, 100$.

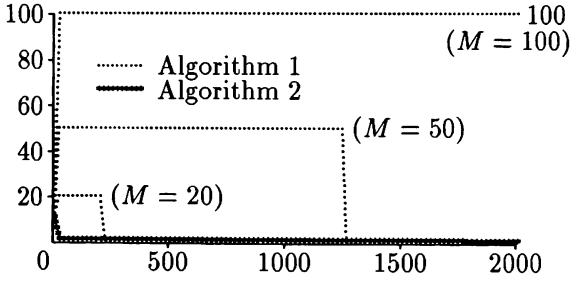

Figure 1: A Comparison of the Performance of the Projected Version of Algorithm 1 with that of Algorithm 2

Algorithm 2 converges very fast on this problem: after 20 iterations, the average estimate of the solution is -0.28 (the 90% confidence interval is $-0.28 \pm 0.16 \times 10^{-1}$) and after 2000 iterations, the average estimate of the solution is -0.14 (the 90% confidence interval is $-0.14 \pm 0.10 \times 10^{-1}$). The performance of Algorithm 1 depends heavily on the bounds on its variables. If the sequence of iterates $\{\theta_n\}$ is restricted to the set $[-M, M]$, then it oscillates between the upper and lower bounds until the length of the step taken in iteration n ($|Y_n|/n \simeq |h(\theta_n)|/n \simeq M^3/n$) is less than the length of the interval that the sequence $\{\theta_n\}$ is bounded to ($2M$), or until $n > M^2/2$. After that it converges quickly to the solution. We ran the algorithm with $M = 20, 50, 100$. After 2000

iterations, the 90% confidence intervals were $-0.46 \pm 0.13 \times 10^{-2}$ when $M = 20$, $-0.45 \pm 0.19 \times 10^{-1}$ when $M = 50$ and 100 ± 0 when $M = 100$.

5 CONCLUSION

Classical stochastic approximation algorithms have severe problems associated with them: they converge extremely slowly when applied to flat functions, and they often diverge when applied to steep functions. We have developed a new stochastic optimization algorithm that is more robust than the classical algorithms in that it is guaranteed to converge on a larger class of problems. At the same time, we have observed that it sometimes converges significantly faster in practice than the classical algorithms.

REFERENCES

Andradóttir, S. 1990a. Stochastic Optimization with Applications to Discrete Event Systems. Ph.D. Dissertation, Department of Operations Research, Stanford University, Stanford, California.

Andradóttir, S. 1990b. A new algorithm for stochastic optimization. In *Proceedings of the 1990 Winter Simulation Conference*, eds. O. Balci, R. P. Sadowski, and R. E. Nance, 364 – 366.

Andradóttir, S. 1991a. A stochastic approximation algorithm with bounded iterates. Technical Report 91-2, Department of Industrial Engineering, University of Wisconsin – Madison, Madison, Wisconsin.

Andradóttir, S. 1991b. Optimization of the steady-state behavior of discrete event systems. Technical Report 91-5, Department of Industrial Engineering, University of Wisconsin – Madison, Madison, Wisconsin.

Chen, H. F., and Y. M. Zhu. 1986. Stochastic approximation procedures with randomly varying truncations. *Scientia Sinica* (Series A) **29**: 914 – 926.

Glasserman, P. 1991. *Gradient Estimation via Perturbation Analysis.* Norwell: Kluwer Academic Publishers.

Glynn, P. W. 1989. Likelihood Ratio Derivative Estimators for Stochastic Systems. In *Proceedings of the 1989 Winter Simulation Conference*, eds. E. A. MacNair, K. J. Musselman, and P. Heidelberger, 374 – 380.

Kiefer, J., and J. Wolfowitz. 1952. Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics* **23**: 462 – 466.

Robbins, H., and S. Monro. 1951. A stochastic approximation method. *Annals of Mathematical Statistics* **22**: 400 – 407.

Suri, R. 1989. Perturbation analysis: the state of the art and research issues explained via the GI/G/1 queue. *Proceedings of the Institute of Electrical and Electronics Engineers* **77**: 114 – 137.

AUTHOR BIOGRAPHY

SIGRÚN ANDRADÓTTIR is an Assistant Professor of Industrial Engineering at the University of Wisconsin – Madison. She received a B.S. in Mathematics from the University of Iceland in 1986, an M.S. in Statistics from Stanford University in 1989, and a Ph.D. in Operations Research from Stanford University in 1990. Her research interests include stochastic optimization, simulation and stochastic processes. She is presently a member of ORSA and TIMS.