Proceedings of the 36th
Conference on Decision & Control
San Diego, California USA • December 1997

SOME GUIDELINES FOR USING ITERATE AVERAGING IN STOCHASTIC APPROXIMATION

John L. Maryak

The Johns Hopkins University
Applied Physics Laboratory
Laurel, Maryland 20723-6099, U.S.A.
e-mail: john.maryak@jhuapl.edu

Abstract

Averaging of the output (iterates) from a stochastic approximation (SA) recursion has been shown to be a useful technique for the gradient-based Robbins-Monro form of SA. For the gradient-free (e.g., Kiefer-Wolfowitz) form, iterate averaging can produce an improvement in the stability of the algorithm and competitive mean-square errors relative to the standard (unaveraged) recursion. We discuss guidelines on how and when to use averaging in this context

Keywords: Stochastic Approximation, Iterate Averaging, Mean Square Error, Finite-difference Stochastic Approximation, Simultaneous Perturbation Stochastic Approximation.

1. Introduction

Averaging techniques have received much attention for their potential to enhance the performance of stochastic approximation (SA) algorithms for optimization of some loss function (see Polyak and Juditsky (1992) and the references therein). The efficacy of the averaging approach for Robbins-Monro SA has been established by Polyak and Juditsky (1992), who developed the asymptotic distribution theory for averaged SA iterates, and by others (Kushner and Yang (1993, 1995), Ljung (1993)).

These techniques are useful to practitioners because they can lead to optimal results without the need to know certain quantities, such as optimal SA gains or particular values of the Hessian of the loss function. These quantities are usually unknown in practical applications, but knowledge of them may be necessary to satisfy theoretical requirements for the optimal performance of the standard SA recursion (i.e., the SA recursion without averaging).

This work was partially supported by the JHU/APL IRAD Program

For the gradient-free form of SA (e.g., the Kiefer-Wolfowitz finite-difference method, FDSA), the situation is not so clear. Such methods construct gradient approximations based on measurements of the loss function; they differ from the Robbins-Monro setting, which requires a direct measurement of the gradient of the loss function with respect to the parameters being optimized. Dippon and Renz (1996, 1997) have investigated the asymptotic distributions of various weighted averages of the output (iterates) in the gradient-free setting. Dippon and Renz (1997) find that one can run an averaged FDSA, without knowledge of (usually) unknown quantities such as the optimal gain or a lower bound on the smallest eigenvalue of the Hessian matrix, with confidence that the result will have a reasonably small error. The reason for this is that the asymptotic mean squared error (AMSE) for the averaged iterates in FDSA is (under hypotheses) less than four times the AMSE of the standard algorithm run in optimal mode. Since the optimal setup for the standard algorithm is seldom known, and since bad choices of algorithm parameters can result in arbitrarily large AMSE, the averaging technique can provide a competitive error level without requiring such special knowledge

2. Guidelines

Given the above situation, it would be useful to have some guidelines on when to use iterate averaging in the gradient-free setting. To investigate this, we consider the Simultaneous Perturbation SA (SPSA) algorithm introduced by Spall (1988). This algorithm is based on a special ("simultaneous perturbation" (SP)) form of the gradient approximation required in the computations, and is especially efficient in high-dimensional problems, relative to FDSA

Let us introduce some notation. From Spall (1992), the (standard) SPSA iteration is $\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \hat{g}_k(\hat{\theta}_k)$, where $\hat{\theta}_k$ is the k^{th} estimate of the optimizing vector θ^* of a loss function $L(\theta)$ having derivative $g(\theta) \equiv \partial L(\theta)/\partial\theta$ such that $g(\theta^*) = 0$, $\hat{g}_k(\bullet)$ is the SP

0-7803-3970-8/97 $10.00 © 1997 IEEE

estimate of the gradient at the k^{th} iteration, $a_k = a/(k+1)^{\alpha}$, $a \geq 0$, $0 \leq \alpha \leq 1$, and $\hat{g}_k(\bullet)$ is estimated based on evaluating $L(\hat{\theta}_k \pm c_k \Delta_k)$, which is observed with additive noise, where Δ_k is a particular random perturbation and $c_k = c/(k+1)^{\gamma}$ is a scale factor, with $c \geq 0$ and $0 \leq 2\gamma \leq \alpha \leq 6\gamma$.

In order to compare the AMSEs of the standard and averaged versions of the algorithm, we examine the asymptotic distribution of the error in the estimate of θ^*. In the above setup, with $\alpha = 1$ (chosen for optimal asymptotic performance) and $\gamma = 1/6$ (chosen to provide the same rate of convergence, i.e., a scale factor of $k^{1/3}$, as the averaged algorithm) the asymptotic distribution of the error in the standard SPSA iterate is (from Spall (1992)) given by:

$$k^{\beta/2}(\hat{\theta}_k - \theta^*) \xrightarrow{D} N\left(c^2(H - \beta I/2a)^{-1}b, PMP^T\right), \quad (1)$$

where $\beta = \alpha - 2\gamma = 2/3$ (with the choices of α and γ as above), P is an orthogonal matrix such that PHP^T is diagonal, H is the Hessian matrix, $\partial^2 L(\theta)/\partial\theta\partial\theta^T$, evaluated at θ^*, $M = a\sigma^2\rho^2(PHP^T - \beta I/2a)^{-1}/8c^2$, σ and ρ are parameters associated with the observation noise and random perturbations, respectively, and b is a vector related to the third derivative of the loss function (superscript T denotes matrix transpose). Of course, the asymptotic results depend on regularity conditions, and on conditions related to the choice of the random perturbations, Δ_k.

For the averaged iterate, we examine the simple average, defined as $\bar{\theta}_k = k^{-1}\sum_{i=1}^{k}\hat{\theta}_i$, although various other weighted averages have also been proposed. The asymptotic distribution of the error in the averaged iterate (from Dippon and Renz (1997)) with $\alpha \in (2/3, 1)$ and $\gamma = 1/6$ (standard settings for averaging) is given by

$$k^{1/3}(\bar{\theta}_k - \theta^*) \xrightarrow{D} N\left(3c^2H^{-1}b/2, 3\sigma^2\rho^2H^{-2}/16c^2\right). \quad (2)$$

A few facts of interest may be noted regarding expressions (1) and (2), which were developed independently of each other. First, the dependencies on the quantities a and c are shown explicitly. Second, the asymptotic means are not zero, as they are in Robbins-Monro SA; this complicates the analysis of the AMSE. Finally, the parameters of the asymptotic distribution of the averaged iterates do not depend on the choice of a. This is important to the practitioner, since it is often not easy to determine the optimal choice of a

Taking the AMSE of an asymptotic $N(\mu, \Sigma)$ distribution to be $\|\mu\mu^T + \Sigma\|$, where $\|\bullet\|$ is a matrix norm, we have that the AMSE of the averaged SPSA iterates is less than that of the standard SPSA asymptotic iterate if:

$$\|9c^4H^{-1}bb^TH^{-1}/4 + 3\sigma^2\rho^2H^{-2}/16c^2\| < \|c^4(H - \beta I/2a)^{-1}bb^T(H - \beta I/2a)^{-1} + a\sigma^2\rho^2P(PHP^T - \beta I/2a)^{-1}P^T/8c^2\|. \quad (3)$$

It is easy to see that there are cases where expression (3) would or would not hold, i.e., where either the averaged or the standard iteration would be preferable.

In order to obtain an intuitive understanding of when averaging can be most successful, we examine expression (3) in the case when θ is a scalar. The analysis of this case is simplified by the fact that $P = 1$. We focus on two possibilities, the case where the Hessian is large compared to $\beta/2a$, and the case of smaller values of H, where other terms are held constant. If H is large, it is easy to see that the inequality (3) will tend to hold, because of the exponent of -1 in the variance term of expression (1) compared to the squared inverses in the other terms. In that case, averaging will tend to improve the estimate. On the other hand, smaller values of H will tend to favor the unaveraged iterates. This finding also makes sense from the viewpoint of geometric intuition. That is, larger values of H (the second derivative of the loss function) are associated with a more rapidly varying loss function, which would tend to make the iterates jump back and forth around the minimizing value θ^*, a situation that should favor averaging.

3. Numerical Study

In order to test the idea that larger values of H tend of favor iterate averaging, we undertook the problem of minimizing the following loss function for $\theta \in R^6$:

$$L(\theta) = \eta\left(\sum_{i=1}^{6}\tau_i^2 + (.1)\sum_{i=1}^{6}\tau_i^3 + (.01)\sum_{i=1}^{6}\tau_i^4\right),$$

where τ_i is the i^{th} component of the vector $\tau = B\theta$, and B is the 6×6 upper triangular matrix with all entries = 1/6 in the upper part, and η is a scale factor to control the size of H (see the discussion below).This function was chosen to ensure significant variable interaction and to provide a reasonably challenging loss function, where the ratio of the maximum to minimum eigenvalues of H is 65. For this function, the value of θ^* is zero in all components and $H = 2\eta B^TB$.

Several runs were made with various settings of the parameters, and changing seeds for the random variables. The parameters and settings that were *fixed* for all of the

runs were: $a=1$, $\Delta_k \sim$ Bernoulli(± 1), $\gamma = 1/6$, $c=.001$, the observation noise added to the value of $L(\theta)$ was $N(0,.0001^2)$, and the number of iterations within SPSA was 10,000.

Using the above parameters and settings and a fixed value of η, we ran one SPSA recursion with $\alpha=1$ (to implement the classical $O(1/k)$ SA gain) and another recursion with $\alpha=.68$ (this type of "slowly decaying" gain is required for use with averaging). After 10,000 iterations, we computed the squared norm of the final θ values (i.e., the error values, since $\theta^* = 0$) from the $\alpha=1$ recursion and from the $\alpha=.68$ recursion, and the squared norm of the average of the last 1000 iterates of the $\alpha=.68$ recursion. These three squared error norms were computed for 10 independent trials for both $\eta=10$ (large H) and $\eta=.005$ (small H). Means of these results over the 10 trials are shown in Table 1 in normalized form, i.e., the squared norms in each row are divided by the squared norm for the $\alpha=1$ run. This normalization is done to allow more direct comparisons and in an effort to remove variation due to suboptimalities in the SA recursion brought about by using a fixed value of a for both small and large values of H (necessary since the size of H in the equations is relative to the size of a).

Table 1.

Normalized Squared Error Norms for Standard and Averaged Iterates at Various Values of α and η, after 10,000 iterations (values shown are means over 10 replications). Averaging used the final 1000 iterations

	$\alpha=1$ Final θ, unaveraged	$\alpha=.68$ Average of last 1000	$\alpha=.68$ Final θ, unaveraged
$\eta=10$ (Large H)	1	.0110	.0094
$\eta=.005$ (Small H)	1	.9665	.9609

Table 1 shows that, as predicted, for larger H, iterate averaging provides a strong improvement over the unaveraged final result of the $\alpha=1$ recursions. For smaller values of H, this study finds that averaging provides a small improvement over the (unaveraged) $\alpha=1$ result, which is not contrary to the discussion in Section 3 above (although we might have predicted that averaging would actually be harmful; our small H results are affected by the fact that the initial guess of θ was not much changed by the SA processing due to the extreme flatness of the loss function in the vicinity of θ^*).

Another result of interest is shown by comparing the last two columns of the table, which show that the final (unaveraged) squared error for the $\alpha=.68$ recursion was smaller than that of the averaged iterate, despite the (presumably) asymptotically suboptimal value of α. This result, which we have noted rather often in experimenting with averaging, suggests that, for non-asymptotic (i.e., real-world) recursions, at least of the SPSA type, averaging can sometimes degrade the output of SA relative to the standard (i.e., unaveraged) iterate produced using a slowly decaying gain. Of course, since our only theory of these recursions is asymptotic, it is difficult to provide a theoretical rationale for this finding.

4. Conclusions

In gradient-free SA, averaging of the iterates can provide convenience for the user and stability of the output of the algorithm. Because of the non-zero mean in the asymptotic distribution of the error in the estimate of the optimizing value of θ, however, averaging does not guarantee an optimal asymptotic result, as is the case with the Robbins-Monro SA algorithm. We have examined the asymptotic theory of a form of gradient-free SA, the SPSA algorithm, which is particularly efficient in many complex (large dimensional) problems. This theory indicates that, in accordance with geometric intuition, averaging will tend to be most successful in cases where the Hessian of the loss function is large relative to certain parameters of the algorithm, and will tend to be unsuccessful in some cases where the Hessian is smaller. We have completed a small-scale numerical study that supports the general trend of these ideas. Further guidelines for when averaging is useful can stem from expression (3), which indicates that averaging will be useful when the inequality holds.

Although not considered here, gains of the form $a_k = a/(k+A+1)^\alpha$ with $A>0$ (used in Spall (1997)), provide a simple means for ensuring stability of the algorithm in the early iterations (via the $A>0$ term) and a larger step size (via a larger choice of a, allowed by the use of A) in the later iterations. The author has found that an unaveraged algorithm with such a simple gain often yields superior finite-sample performance, competitive with or better than an averaged solution. Perhaps an averaging approach using $A>0$ (untried by the author as of this writing) would yield even better results.

References

Dippon, J. and J. Renz (1997), "Weighted Means in Stochastic Approximation of Minima," *SIAM J. Control & Optimization,* vol. 35, in press.

Dippon, J. and J. Renz (1996), "Weighted Means of Processes in Stochastic Approximation," *Math. Methods of Statistics,* vol. 5, pp. 32-60.

Kushner, H.J. and J. Yang (1995), "Stochastic Approximation with Averaging and Feedback: Rapidly Convergent 'On-Line' Algorithms," *IEEE Trans. Auto. Control,* vol. 40, pp.24-34.

Kushner, H.J. and J. Yang (1993), "Stochastic Approximation with Averaging of the Iterates: Optimal Asymptotic Rate of Convergence for General Processes," *SIAM J. Control & Optimization,* vol. 31, pp.1045-1062.

Ljung, L. (1993), "Aspects of Accelerated Convergence in Stochastic Approximation Schemes," *Proc. IEEE Conf. Decision & Control,* pp. 1649-1652.

Polyak, B.T. and A.B. Juditsky (1992), "Acceleration of Stochastic Approximation by Averaging," *SIAM J. Control & Optimization,* vol. 30, pp. 838-855.

Spall, J.C. (1992), "Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation," *IEEE Trans. Auto. Control,* vol. 37, pp.332-341.

Spall, J.C. (1988), "A Stochastic Approximation Algorithm for Large-Dimensional Systems in the Kiefer-Wolfowitz Setting," *Proc. IEEE Conf. Decision & Control,* pp. 1544-1548.

Spall, J.C. (1997), "Implementation of the Simultaneous Perturbation Algorithm for Stochastic Optimization," JHU/APL Technical Report.