Pergamon

Neural Networks, Vol. 7, No. 3, pp. 573–574, 1994

0893-6080/94 $6.00 + .00

0893-6080(93)E0014-X

LETTER TO THE EDITOR

A More Efficient Global Optimization Algorithm Based on Styblinski and Tang

Editor:

The Styblinski and Tang (1990) (S&T) paper published in this journal applied the stochastic approximation algorithm with function smoothing (SAS) to find the global minimum of loss functions $f(\cdot)$ in practical applications. They showed that SAS was much more efficient than simulated annealing. This letter will show that another type of algorithm, simultaneous-perturbation stochastic approximation (SPSA) discussed in Spall (1988, 1992), can also be used for global optimization and typically requires a significantly lower number of function evaluations (compared to SAS) to achieve the same level of accuracy [Spall (1988, 1992) only considers the use of SPSA in a local optimization context].

To use SPSA for global optimization, one follows algorithms 2 and 3 in S&T modified by a new gradient approximation. Letting the elements of $\eta^* \in R^n$ be the vector of inverses of the n elements in the random vector η, the SPSA-based gradient approximation for use in eqn 12a of S&T (with $N = 1$) is

$$\hat{\nabla}\tilde{f}(x, \beta) = \frac{1}{2\beta}\eta^*[f(x + \beta\eta) - f(x - \beta\eta)],$$

$$\beta > 0, \quad x \in R^n.$$

This form is slightly (but critically!) different from the SAS gradient approximation in S&T. One of the convergence conditions for SPSA is that the distribution of the elements of η must have finite inverse moments (Spall, 1992). The distributions discussed in S&T for η (Gaussian and Cauchy) are not applicable for SPSA because they violate this inverse moment condition; Spall (1992) and Chin (1992) used a Bernoulli distribution with outcomes ± 1.

The 10-dimensional loss function of Example 3 in S&T was used to compare the SPSA and SAS global optimization techniques; the results are tabulated below. There are two cases studied for SPSA using two different sequences of SA gains. Let τ^m and β^m be the gains for the mth cycle of the global optimization process with each cycle corresponding to a new kernel approximation to the loss function and having a maximum number of allowable SA iterations (i.e., the gains were kept fixed for all SA iterations within a cycle). One case in the table uses the standard decaying (per cycle) gains, $\tau^m = 0.007/m$ for the mth cycle; the other uses optimal τ^m, which were estimated as suggested in Fabian (1971). In both cases, $\beta^m = \{5, 2.5, 1.25, \ldots\}$, no smoothing is used ($\rho_k = 1$ in S&T), and the initial conditions are the same as S&T. The values shown in the table are the number of function evaluations for the iteration processes of SPSA and SAS when the maximum difference between iterated values and the known global minimum is reduced to the 0.24% level as in S&T. The table shows that global optimization using SPSA is three to five times as efficient as that using SAS for this 10-dimensional problem.

TABLE 1
Relative Performance of the SPSA and SAS Algorithms When the Maximum Errors Are Reduced to 0.24% Level of S&T

	No. of Funct. Evals. (Mean No. for SPSA)	Range of Funct. Evals. in 20 Trials
SAS from S&T	9228	Not available
SPSA standard gains	2764	2696, 2884
SPSA optimal gains	1762	1614, 1890

More generally, based on theory in Chin (1992) we expect SPSA to only need about $\frac{1}{3}$ the number of function evaluations of SAS for global optimization problems of any dimension when the number of cycles is small and $< \frac{1}{3}$ when the number of cycles is large [Chin (1992) refers to SAS as random-directions stochastic approximation]. To demonstrate the applicability of SPSA in higher dimensions, we considered a 50-dimensional version of Example 3 in S&T and found the number of function evaluations averaged to a quite reasonable 7460 to achieve the 0.24% level of accuracy above. Moreover, Spall and Cristion (1992) successfully use SPSA in a 332-dimensional optimization problem for neural network-based adaptive control.

Daniel C. Chin
The Johns Hopkins University
Applied Physics Laboratory
Laurel, MD 20723-6099

REFERENCES

Chin, D. (1993). Performance of several stochastic approximation algorithms in the multivariate Kiefer-Wolfowitz setting, *Proceedings of the 25th Symposium on the Interface: Computing Science and Statistics*, 289–295.

Fabian, V. (1971). Stochastic approximation. In J. J. Rustagi, (ed.), *Optimizing methods in statistics* (pp. 439–470). New York: Academic.

Spall, J. C. (1988). A stochastic approximation algorithm for large-dimensional systems in the Kiefer-Wolfowitz setting. *Proceedings IEEE Conference Decision Control*, 1544–1548.

Spall, J. C. (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Auto Control*, **37**, 332–341.

Spall, J. C., & Cristion, J. A. (1992). Direct adaptive control of nonlinear systems using neural networks and stochastic approximation. *Proceedings IEEE Conference Decision Control*, 878–883.

Styblinski, M. A., & Tang, T. S. (1990). Experiments in nonconvex optimization: Stochastic approximation with function smoothing and simulated annealing. *Neural Networks*, **3**, 467–483.