In: Brain Theory Newsletter, Vol. 3, # 3/4
March/May 1979, 69-71

Application of Stochastic Averaging to Learning Systems

Stuart Geman

Division of Applied Mathematics
Brown University
Providence, Rhode Island 02912

Mathematical descriptions of learning systems often involve the following elements: 1. An input process representing the systems environment; 2. A set of state parameters which, at any fixed time, define the input-out relation of the system; 3. A rule, in the form of a differential or difference equation, for updating these parameters in response to the input process. I will describe here a technique for ascertaining the relation between the development in time of the state parameters and the statistics of the environment, when the latter is modelled as a stochastic process. The point of the technique is to relate the statistical structure of an environment to the performance of a given model. Let me first discuss this technique in a general context, and then give an example of its application.

Assume that the algorithm (3, above) has been formulated in continuous time, i.e. using a differential equation for its description. Let S(t) be an n-vector whose components are the set of state parameters (e.g. the vector of all relevant synaptic strengths, or the vector of coefficients in a linear regression). I(t) will be an m-vector representing input from the environment. The components of I(t) may, for example, be the activities of retinal cells, or perhaps some "higher-up" coded description of those activities. (Of course the precise nature of this code is critical, but not the topic of this letter.)

Now suppose that the algorithm ("$\mathscr{A}$") for modification of S(t) is described as follows:

$$\dot{S}(t) = \varepsilon\mathscr{A}(S(t),I(t)) \qquad \text{(I)}$$

where "ε" is a small parameter modelling slow changes in S(t) relative to the fluctuations in I(t). What I will say holds for a class of equations much more general than (I), but most neural-net learning models, or on-line pattern classification schemes can be formulated as in (I). (If I(t) involves the activities of pre and post synaptic neurons, which in turn involve (possibly time-delayed) inputs and S(t) itself, then it is simplest to replace these activities by their equilibrium for a given input and state S; the assumption being that these activities are fast relative to S(t).)

My goal is to describe the relation between the development of S(t) and the distribution of I(t). The main requirement is that I(t) be a strongly mixing process. This means, loosely, that events in the distant past of I(t) and the distant future of I(t) are nearly independent. In particular, "past" and "future" are asymptotically independent as their separation goes to infinity. In the pattern recognition literature, for example, the much stronger assumption of independent samplings is typical. It is common in physics to assume ergodicity, a property closely related to mixing. In any case, mixing processes form a very broad class, and can accommodate most realistic modes of the environment.

The idea is to replace (I) by the deterministic equation

$$\dot{S}'(t) = \varepsilon E[\mathscr{A}(S'(t),I(t))], \qquad \text{(II)}$$

where for fixed x, $E[\mathscr{A}(x,I(t))]$ is the expected value of $\mathscr{A}(x,I(t))$ with respect to the distribution of I(t). Under suitable conditions

$$\lim_{\varepsilon\to o}\ \sup_{t\geq o} E|S(t)-S'(t)| = 0, \qquad \text{(III)}$$

so that, for ε small, the development of S'(t) approximates that of S(t) (in the sense of (III)), for all time. Through this simple averaging technique the relation between the performance of a system and the statistics of its environment is available, to within an approximation. What we have, of course, is a stochastic version of the classical method of averaging developed in mechanics (Cf. 1)(although that method did not generally permit averaging on the entire interval, [0,∞)).

Uttley (3,4, and 5) recently published a series of papers describing a model for the recognition of patterns by neural systems. The model has attractive theoretical properties, performs effectively in simulation, and suggests numerous anatomical and physiological hypotheses. Let us apply the averaging technique to Uttley's system and see if any further insights can be gained. Focus on the basic unit of the model - an individual "informon". A continuous time formulation of that system can be expressed as follows. Let x(t) be an n dimensional feature vector, with each component indicating the presence or absence of some feature at time t. During a learning period, the feature vector is continuously classified with respect to a particular category, and this classification is expressed by y(t), a scalar function of time. At each time t, the value of y indicates whether or not the feature vector, x(t), belongs to that category. The components of x, and y, can be 0-1 valued, or more generally, continuously valued variables indicating the amount of a given quantity. (There may be many categories, in which case we assign a classifier, y, to each.) Finally, let z(t) be an n dimensional state vector, whose components are to be modified by the experience of the model.

The dynamics of the model are most conveniently defined in the vector notation. The output is a scalar, indicating the model's guess as to the appropriate value of y, given a feature vector, x. The output, at time t, is

$$x^T(t)z(t),$$

the inner product of z and x. The system learns according to the following rule:

$$\dot{z}(t) = \varepsilon x(t)\{y(t)-x^T(t)z(t)\}. \qquad \text{(IV)}$$

(I should mention that Uttley's notation is considerably different, but what is here is an equivalent formulation - as a simple check with (3) will verify. Also, refer to that article for a neural interpretation of these dynamics.)

If ε is such that z changes slowly relative to the fluctuations in x(t) and y(t), then the averaging technique is appropriate, and we can approximate z(t) by z'(t):

$$\dot{z}'(t) = \varepsilon\{E[xy]-E[xx^T]z'(t)\},$$

leaving out the t, since we assume that these expectations are, at least approximately, constant (although averaging is justified whether or not this is true). Now as long as $E[xx^T]$ is non-singular (i.e. as long as we have chosen the components of x such that no one is a completely deterministic linear function of the others), this deterministic equation approaches, exponentially fast, the solution

$$z' = E[xx^T]^{-1}E[xy].$$

The averaging principle then says that z(t) will also approach

$$E[xx^T]^{-1}E[xy],$$

and thereafter remain close. Finally, then, the asymptotic output of the model is well approximated by

$$x^T(t)E[xx^T]^{-1}E[xy], \qquad \text{(V)}$$

given the input x(t).

Now (V) is an encouraging result, for it is precisely the best choice of z, in the mean square error sense - as is easily demonstrated, and very well known in the theory of linear regression. (In fact, if we replaced the constant "gain", ε, in (IV) by, say, 1/t, we would have a stochastic version of the gradient descent algorithm, and this has been studied (c.f. (6) and (7)).) Uttley's model, then, inherits the known advantages, as well as difficulties, of this solution. Since the square error criterion is an entirely reasonable measure of performance, the solution is, in a strong sense, optimal. But this optimality is within the context of linear solutions, and we can not be sure that the optimal linear solution is anywhere near the unconstrained mean square optimal (unless, for example, the feature components and classification are jointly Gaussian). In fact, the actual performance of this solution is highly dependent on the particular choice of features, and in this sense the model recasts the perception problem into the problem of developing, possibly through experience, an appropriate feature set. Also, unmodified stochastic descent procedures are typically very slow (8), especially when the dimension of the feature set is large.

Although not always explicitly, a number of authors have exploited the close relation between equations (I) and (II) in their analysis of learning systems. Look at the case where ε is an appropriately decreasing function of time, such as $1/t$. If "$\mathscr{A}$" is the negative of the gradient of a suitable criterion function (such as the square of the error in Uttley's case) then, with some regularity conditions, $S(t)$ will converge to the value which minimizes the mean of the criterion function. A number of general theorems of this sort are around, and Pfaffelhuber (9) has discussed their relation to memory models. The relation to the averaging principle is as follows. When $\varepsilon \to 0$ as $t \to \infty$ (as with $\varepsilon = 1/t$), (III) can be replaced by

$$\lim_{t\to\infty} E|S(t)-S'(t)| = 0. \qquad \text{(VI)}$$

Now in the present context, (II) is a deterministic gradient descent procedure, implying that S'(t) converges to the minimizing value for the mean of the criterion function. Hence, by (VI), S(t) also converges to this "optimal" value.

On the other hand, when the gain, ε, is fixed but small, and when "$\mathscr{A}$" is, again, the negative of the gradient of a suitable criterion function, then S(t) will come to lie "close" to the optimal value. (For some specific criterion functions, the result is discussed in (10). More generally, it is a special case of (III).) See Amari (11) for an application to neural modelling. The work by Kohonen (for example (12), sec. 3.2.2) on recursive computation of linear filters is in a similar vein, albeit in a deterministic setting. There, the input patterns are chosen from a finite set in arbitrary sequence, but such that each pattern occurs infinitely often - a kind of deterministic version of the mixing assumption. Indeed, the resulting limit is the limit of the averaged equation, (II), where expectation is taken with respect to <u>any</u> distribution on patterns, provided that each pattern has positive probability.[1]

[1] It may seem odd that the asymptotic behavior of a filter would be independent of the marginal distribution of the inputs. But Kohonen assumes that each pattern is deterministically associated with a fixed classification, and that the number of patterns does not exceed the dimension of the feature space. Hence there exists an exact solution

> to the pattern classification problem, and the mean square error of this solution is zero independent of the underlying distribution.

The deterministic interpretation has the advantage that convergence is obtained without requiring that the "gain", ε, approach zero (there is no variance in a given input, hence no "noise" to be damped out).

What I wish to emphasize is that the approximation of the solution of equation (I) by that of equation (II) is appropriate in a broad variety of contexts (for precise conditions, see (2)). "$\mathscr{A}$" can be nonlinear, and need not be the gradient of any criterion function. With regards to the input process, if, as is often assumed, it is a sequence of independent and identically distributed observations, then it is certainly mixing. But many Markov processes, for example, are also mixing, and in fact a mixing process need not even be stationary. When it is justified, averaging simplifies the relation between the dynamics of a model and the statistics of its environment. For example, it is particularly simple to obtain an approximate equilibrium for the state vector in terms of the distribution of the input. Hence, the asymptotic performance of a model can be evaluated.

A more general application is to utilize averaging to define the statistical structure of those environments in which a given model is effective. The "real-world" character of these structures is one measure of the appropriateness of that model.

References

1. Mitropolsky, Iu. A., Averaging Method in Non-Linear Mechanics. International Journal of Non-Linear Mechanics, 2, 1967, 69-96.

2. Geman, S., Some Averaging and Stability Results for Random Differential Equations. Submitted for publication (Manuscript is available as: Reports in Pattern Analysis No. 59, Div. Appl. Math., Brown University, Prov. R.I. 02912).

3. Uttley, A.M., A Two-Pathway Informon Theory of Conditioning and Adaptive Pattern Recognition. Brain Research, 102 (1976), 23-35.

4. Uttley, A.M., Simulation Studies of Learning in an Informon Network. Brain Research, 102 (1976) 37-53.

5. Uttley, A.M., Neurophysiological Predictions of a Two-Pathway Informon Theory of Neural Conditioning. Brain Research, 102 (1976) 55-70.

6. Blum, J., Multidimensional Stochastic Approximation Methods. Ann. Math. Stat., 1958, 373-407.

7. Blaydon, C.C., Recursive Algorithms for Pattern Classification. Technical Rept. #520, Div. Engrg. and Appl. Physics, Harvard Univ., Cambridge, Mass., March 1967. (Also; ONR Contract Nonr 1866(16), NR-372-012.)

8. Duda, R.O. and Hart, P.E., Pattern Classification and Scene Analysis. John Wiley & Sons, New York, 1973.

9. Pfaffelhuber, E., Correlation Memory Models - A First Approximation in a General Learning Scheme. Biol. Cybernetics 18, (1975) 217-223.

10. Wasan, M.T., Stochastic Approximation. Cambridge University Press, Cambridge, 1969.

11. Amari, S.I., Neural Theory of Association and Concept-Formation. Biol. Cybernetics 26, (1977) 175-185.

12. Kohonen, T., Associative Memory, A System-Theoretical Approach. Springer-Verlag, New York, 1977.