SIAM-AMS Proceedings
Volume 13
1981

# The Law of Large Numbers in Neural Modelling

STUART GEMAN[1]

Put loosely, the law of large numbers (LLN) says that the average of a large number of independent, or nearly independent, random variables is usually close to its mean. For some of the mathematics that typically arise in neural modelling, this simple principle has a natural and rewarding application. In one version of this application, equations for the development of long term memory traces (usually modelled as changes in "synaptic efficacies") are well approximated by more elementary equations, and from these the performance of the model can be more easily anticipated. In a second version, a large system of equations modelling the individual activities of interconnected homogeneous populations of neurons is replaced by a small number of prototype equations which accurately describe the macroscopic dynamics of the network. Models of this latter type might be relevant, for example, to the generation of phrenic nerve activity by the brainstem respiratory centers.

What I mean to present is more a point of view than a strict mathematical technique. It is another, more simple, way of looking at models which may be very complex, or even intractable, in their first formulation. For this purpose, I feel that a presentation completely by example will be most effective. A reader interested in a more formal and rigorous development, and a more general context, is referred to [9], [10], [12], and [14], and the references therein to other authors.

**Time averaging: the behavior of models for the development of long term memory.** In the three examples of this section, the LLN takes the form of a stochastic "method of averaging" for differential equations, through which method the behavior of a complex neural or cognitive model can often be anticipated with surprising ease. The method applies to differential equations in which the solution is slowly varying relative to the other time dependent terms

1980 *Mathematics Subject Classification.* Primary 60F99.
[1]Supported by the National Science Foundation under grant MCS76-80762.

of the equation. Equations modelling the development of a long term memory are typically of this form: the dependent variable represents some component of the long term trace, and it is slowly varying relative to the stimuli which effect changes in that trace. It may be that a particular model has been formalized using a system of integral or difference equations, but such models typically have a natural reformalization using differential equations. And, conversely, a method of averaging can be formulated for these other settings as well.

As both an introduction and a good demonstration of the method's utility, I will first apply it to a neural network memory model proposed by Uttley (in [22]–[24]). (My discussion of averaging in Uttley's model is, by and large, a repeat of what was said in [8].) The model consists of a network of units called "informons". These are adaptive neuron-like elements which can learn to signal whether or not a vector input belongs to a particular classification. The dynamics of the informon and the rule by which it self-organizes are defined by the following equations

$$F(Y) = \sum_{i=1}^{n} F(X_i)\gamma_i + F(Z)\gamma_z, \qquad \Delta\gamma_i = -bF(X_i)F(Y) \tag{i}$$

where

$F(Y)$ is the output (firing rate) of the informon (neuron) labelled "$Y$";

$F(X_i)$, $i = 1, 2, \ldots, n$, are outputs from other units in the network, and comprise the input to the unit $Y$;

$F(Z)$ is a binary classifying signal which indicates, during training, which inputs $(F(X_1), \ldots, F(X_n))$ belong to a particular category;

$\gamma_i$, $i = 1, \ldots, n$, are modifiable conductivities, which determine the extent to which the signals $F(X_i)$, $i = 1, \ldots, n$, contribute to the output at $Y$;

$\gamma_z$ is a fixed and negative conductivity transmitting the classifying signal to $Y$;

$\Delta\gamma_i$ is the change in the conductance $\gamma_i$ due to a simultaneous appearance of an input $F(X_i)$ with an output $F(Y)$;

$b$ is a positive constant determining the rate at which this latter modification proceeds.

The connection between this example and those which follow will be more transparent if I use, as nearly as is possible, a unified notation. For this purpose, let $y(t)$ be the output at time $t$ of the unit designated "$Y$" in the Uttley model (replacing "$F(Y)$"). $x_1(t), \ldots, x_n(t)$ will replace $F(X_1), \ldots, F(X_n)$ as input signals to $Y$, and $\gamma_1(t), \ldots, \gamma_n(t)$ will, again, denote the corresponding conductivities. Since $\gamma_z$ is fixed and negative, it will be convenient to denote the net classifying signal, $\gamma_z F(Z)$, simply by $-z(t)$. Finally, writing $\varepsilon$ in place of $b$, a continuous time formulation for (i) is

$$y(t) = \sum_{i=1}^{n} \gamma_i(t)x_i(t) - z(t),$$

$$\frac{d}{d\tau}\gamma_i(t) = -\varepsilon x_i(t)y(t) = \varepsilon x_i(t)\left\{ z(t) - \sum_{j=1}^{n} \gamma_j(t)x_j(t) \right\}. \tag{ii}$$

Uttley does not analyze this system directly, but instead replaces it by a new system which is intended to be a more tractable approximation. The analysis of this approximating system, together with simulation results, indicates that the informon and networks of interconnected informons have properties suggestive of classical and operant conditioning as well as a capability for pattern classification. We will see that the method of averaging yields some additional insights, and a more direct and precise analysis of the behavior of (ii).

This is a long term memory model, and Uttley assumes that $\gamma_i(t)$ reflects only the long term behavior of $x_i(t)$ and $y(t)$, not their most immediate fluctuations. The translation, for (ii), is the assumption that $\varepsilon$ is small; changes in $\gamma_i(t)$ are slow relative to those in $x_i(t)$ and $y(t)$. This assumption is an important one, because it means that, essentially, $d\gamma_i(t)/dt$ "sees only the average" of the right-hand side of the differential equation in (ii). To make this a little more precise and to see why it should be true, consider the change in $\gamma_i(t)$ over a period of time $\Delta t = \delta/\varepsilon$, where $\delta$ is small, but not nearly as small as $\varepsilon$. In this period of time, $\gamma_i(t)$ will not appreciably change (its derivative being of order $\varepsilon$), but this is a considerable interval relative to the time course of $x_i(t)$ and of $y(t)$. From (ii)

$$\gamma_i\left(t + \frac{\delta}{\varepsilon}\right) - \gamma_i(t) = -\varepsilon \int_t^{t+\delta/\varepsilon} x_i(s)y(s)\, ds$$

i.e.

$$\gamma_i(t + \Delta t) - \gamma_i(t) = \frac{\delta}{\Delta t}\int_t^{t+\Delta t} x_i(s)\left\{ z(s) - \sum_{j=1}^{n} \gamma_j(s)x_j(s)\right\} ds$$

$$\approx \frac{\delta}{\Delta t}\int_t^{t+\Delta t} x_i(s)z(s)\, ds - \sum_{j=1}^{n} \gamma_j(t)\frac{\delta}{\Delta t}\int_t^{t+\Delta t} x_i(s)x_j(s)\, ds. \quad \text{(iii)}$$

The latter (very rough) approximation is because $\gamma_j(s)$ is nearly constant over the interval $[t, t + \Delta t]$.

Now, let us take the point of view (deferring discussion on this) that $(x_1(t), \ldots, x_n(t), z(t))$ is a random process, and that over large periods of time it is essentially "independent of itself" (current observations of the process tell us very little about its distant future). Then, $(\Delta t)^{-1}\int_t^{t+\Delta t} x_i(s)z(s)\, ds$ and $(\Delta t)^{-1}\int_t^{t+\Delta t} x_i(s)x_j(s)\, ds$ are long run averages (since $\Delta t = \delta/\varepsilon$ is large) of nearly independent random variables and should be well approximated by means (i.e. expected values)

$$\frac{1}{\Delta t}\int_t^{t+\Delta t} x_i(s)z(s)\, ds \approx \frac{1}{\Delta t}\int_t^{t+\Delta t} E\left[ x_i(s)z(s)\right] ds,$$

$$\frac{1}{\Delta t}\int_t^{t+\Delta t} x_i(s)x_j(s)\, ds \approx \frac{1}{\Delta t}\int_t^{t+\Delta t} E\left[ x_i(s)x_j(s)\right] ds.$$

Finally, put this back in (iii) and take derivatives

$$\frac{d}{dt}\gamma_i(t) \approx \varepsilon E\left[ x_i(t)z(t)\right] - \varepsilon \sum_{j=1}^{n} \gamma_j(t)E\left[ x_i(t)x_j(t)\right], \quad \text{(iv)}$$

and this is what I meant when I said that $d\gamma_i(t)/dt$ sees only the average of the right-hand side in (ii). The point is this: (iv) relates the "memory trace", $\gamma_i(t)$, to the statistical structure of the environment (as revealed by the operator, $E$). When the solution to (iv) (with $\approx$ replaced by $=$) is close to that in (ii), we will be able to infer from (iv) the most important features of the model's behavior.

In fact, under very general conditions the solution to (ii) is well approximated by the solution to (iv). The smaller $\varepsilon$, the better the approximation, and, in particular, the error goes to zero with $\varepsilon$. Details about the conditions, as well as a precise statement of the sense in which the approximation holds, can be found in [**9**] or [**10**].

The next step, then, is to develop the consequences of (iv). But before this, we should briefly examine in a nontechnical manner the two most important assumptions implicit in this approximation procedure. Above all, the reader may question the use of a random process model for $x_i(t)$, $i = 1, 2, \ldots, n$, and $z(t)$, the "environment" of the conductivity, $\gamma_i(t)$. In fact, we did not have to take this approach at all, since the "method of averaging" is, originally, a technique for approximating *deterministic* equations (see, for example, Mitropolsky [**21**]). Thus, a deterministic model would lead us to a version of (iv) in which a certain time average plays the role of expectation, $E$, and we could then proceed to analyze instead this analogue to (iv). But I believe that the probabilistic point of view has something special to offer, and the further discussion of this example together with the examples below should convincingly support this position. Whether the environment is in some sense truly random is of no importance; the probability model offers a convenient framework in which to describe characteristics of that environment. It does not in any sense suggest that the environment is *unstructured*. Indeed, a determinstic model is merely a special case.

There has also been made a "mixing" assumption: that the past and future are, asymptotically, independent. Mixing is an ergodic-like property that, practically speaking, puts very little constraint on potential models of the environment. (Any deterministic model, for example, is mixing. But then (iv) is (ii), and the method offers no simplification.) For example, a wide variety of Markov processes, and in particular those which would be most appropriate in representing a model's environment (i.e. bounded and obeying some mild regularity conditions), are mixing in a way suitable for application of the method of averaging. In the pattern recognition literature, a much stronger assumption is typical: successive scenes or patterns are statistically independent.

In short, (iv) will approximate (ii) under assumptions which are natural for the system being modelled. What, then, can (iv) tell us about the behavior of Uttley's model? A good place to start is with the asymptotic behavior: How does the model perform after a theoretically infinite period of time? If there is to be an "asymptotic behavior", then we must first assume that something like an equilibrium for $\gamma_i(t)$ exists, and this amounts to making an assumption of stationarity. Or, at the least, an assumption that the expectations appearing in

(iv) do not depend on time (but averaging is appropriate whether or not this is the case). Really, this is not much of an additional assumption, since there would be no point in a long term memory if the environment did not possess some degree or stationarity. Let us assume, then, that $E[x_i(t)z(t)] = E[x_i z]$ and $E[x_i(t)x_j(t)] = E[x_i x_j]$ do not depend on $t$ (certainly *not*, however, that $x_i(t)$ or $z(t)$ are constant). Then, the equilibrium for (iv), and therefore the approximate equilibrium for (ii), is immediately available. Simply set the derivative in (iv) equal to 0 and solve for $\gamma_i$,

$$\sum_{j=1}^{n} E[x_i x_j]\gamma_j = E[x_i z], \qquad i = 1, 2, \ldots, n. \tag{v}$$

Define $n \times 1$ column vectors $X(t)$ and $\Gamma(t)$ by $X(t) = (x_1(t), \ldots, x_n(t))^T$ and $\Gamma(t) = (\gamma_1(t), \ldots, \gamma_n(t))^T$ (using $T$ to denote transpose). In vector-matrix notation, (v) is $E[XX^T]\Gamma = E[Xz]$. And therefore, assuming that $E[XX^T]$ is nonsingular,[2]

$$\Gamma = E[XX^T]^{-1}E[Xz]. \tag{vi}$$

Since (ii) behaves like (iv), the conclusion is that $\Gamma(t)$ will approach and remain close to $E[XX^T]^{-1}E[Xz]$.

The reader familiar with multivariate analysis will recognize (vi) as the solution to the linear regression problem: Choose $\gamma_1, \ldots, \gamma_n$ so as to minimize the mean square error in approximating $z(t)$ by the linear combination $\sum_{j=1}^{n} \gamma_j x_j(t)$, i.e. minimize

$$E\left|z - \sum_{j=1}^{n} \gamma_j x_j\right|^2 \tag{vii}$$

over all possible values of $\Gamma = (\gamma_1, \ldots, \gamma_n)^T$. In words, the conductivities of the informon modify in such a way that the output of $Y$ in *absence* of the classifying signal (i.e. $\sum_{j=1}^{n} \gamma_j(t)x_j(t)$) approaches the best linear predictor of $z(t)$ (the classification) given $x_1(t), \ldots, x_n(t)$. Actually, we know much more. (iv) is an autonomous system of linear differential equations, and its exact solution is well known. Then, since (ii) stays close to (iv), we have available essentially the entire time course of $\Gamma(t)$. Roughly, $\Gamma(t)$ approaches $E[XX^T]^{-1}E[Xz]$ exponentially with rate determined by the eigenvalues of the positive definite matrix $E[XX^T]$.

The method of averaging, really an application of the LLN, gives us a virtually complete description of the dynamics of the informon. It reveals details about the unit's behavior not obviously apparent in (i) and not found in the system which Uttley offers as a more tractable alternative. Thus we know that the informon is asymptotically a nearly optimal classifier–at least among linear machines. In fact, $\sum_{j=1}^{n} \gamma_j x_j(t)$ will predict $z(t)$, in an approximately minimum

[2]Equivalent is the assumption that no component of $X(t)$, say $x_i(t)$, is a *deterministic* linear combination of the remaining components $\{x_j(t)\}, j \neq i$. Any such deterministic relation would be undone by "noise" in a real system.

mean square error sense, whether or not $z(t)$ is the binary signal assumed in the model. That is, asymptotically, the solution to (iv) minimizes (vii) whatever the nature (discrete or continuous) of the "classifying signal", $z(t)$. And, if $x_1(t), \ldots, x_n(t), z(t)$ jointly form a Gaussian process, then the best linear predictor of $z(t)$ is also the best unconstrained predictor.

The analysis, then, supports (ii) as an appropriate system for learning to predict a "classification", $z(t)$, from the information contained in the channels $x_1(t), \ldots, x_n(t)$. But, by making the connection to some well-studed areas of statistics and pattern recognition, the analysis also suggests some possibly unattractive features of the model. For example, unless $(x_1(t), \ldots, x_n(t), z(t))$ is a Gaussian process, the best linear predictor of $z(t)$ may be quite inferior to the overall best predictor. Although the inevitable noise present in neural activity is probably well approximated by a Gaussian model, I would doubt that the signals themselves are anything like a Gaussian process. If these signals are not Gaussian, would the nevous system employ a suboptimal solution? Also, there is reason to question the efficiency of the modification procedure defined in (ii). It is, essentially, a stochastic approximation procedure for finding the least mean square error linear predictor of $z(t)$ given $x_1(t), \ldots, x_n(t)$ (see, for example, Duda and Hart [6], or Wasan [25]). We must, then, ask why the nervous system would utilize this particular version of stochastic approximation when there are other versions known to perform more efficiently. Again, there is raised a question of optimality. There may, of course, be good answers for these questions, and it may be that the model is entirely appropriate. But, at the least, we have established a framework in which the model can be meaningfully compared to already existing theory.

Uttley points out that since $\gamma_i(t)$ may be positive or negative, its neural realization would require both excitatory and inhibitory synapses. Amari (in [2]) has proposed a model quite similar in spirit to Uttley's, but one which more explicitly addresses the problem of achieving a net conductivity which may be positive or negative, out of couplings which are individually constrained to be excitatory or inhibitory. In [2], Amari is already aware of the method of averaging and applies it, much as we did above, to determine the equilibrium behavior of his model. I will retrace some of Amari's analysis, and interpret the conclusions with special attention to the close relationship between the Uttley and the Amari theories.

The fundamental unit in Amari's model is a neuron-like device, which I will again call "$Y$", receiving inputs $x_1(t), \ldots, x_n(t)$, possibly from other units or possibly from an external source. Each of these inputs $x_i(t)$ influences $Y$ through both an excitatory and an inhibitory coupling; let us denote the strengths of these couplings by $\gamma_i^+(t)$ and $\gamma_i^-(t)$ respectively. The unit learns in the sense that these coupling strengths are modified by its experience. The net input to $Y$ through this variable pathway is $\sum_{j=1}^n \gamma_j^+(t)x_j(t) - \sum_{j=1}^n \gamma_j^-(t)x_j(t)$. Or, in terms of the corresponding vector quantities (for notation, refer back to the discussion of Uttley's model): $\Gamma^+(t)^T X(t) - \Gamma^-(t)^T X(t)$. There is also at $Y$ an unmodifiable

channel which receives a "teacher" input, $z(t)$. Learning is by modification of the $\gamma$ connectivities, as is described (in its continuous time formulation) by the following equations:

$$\frac{d}{dt}\gamma_i^+(t) = \varepsilon(\alpha_1 z(t)x_i(t) - \alpha_2\gamma_i^+(t))$$

$$\frac{d}{dt}\gamma_i^-(t) = \varepsilon\left\{\alpha_3 x_i(t)\left(\sum_{j=1}^{n}\gamma_j^+(t)x_j(t) - \sum_{j=1}^{n}\gamma_j^-(t)x_j(t)\right) - \alpha_4\gamma_i^-(t)\right\}.$$

Here again $\varepsilon$ is a small positive constant. $\alpha_1$, $\alpha_2$, $\alpha_3$ and $\alpha_4$ are for the time being arbitrary positive constants. Writing $\Gamma(t)$ for $\Gamma^+(t) - \Gamma^-(t)$, the system is rewritten in more convenient vector-matrix notation as

$$\frac{d}{dt}\Gamma^+(t) = \varepsilon(\alpha_1 z(t)X(t) - \alpha_2\Gamma^+(t)),$$

$$\frac{d}{dt}\Gamma^-(t) = \varepsilon(\alpha_3 X(t)X(t)^T\Gamma(t) - \alpha_4\Gamma^-(t)). \tag{viii}$$

Now let us apply the LLN. When $\varepsilon$ is small, a good approximation to (viii) (making all the necessary assumptions, as discussed in the previous example) is:

$$\frac{d}{dt}\Gamma^+(t) = \varepsilon(\alpha_1 E[zX] - \alpha_2\Gamma^+(t)),$$

$$\frac{d}{dt}\Gamma^-(t) = \varepsilon(\alpha_3 E[XX^T]\Gamma(t) - \alpha_4\Gamma^-(t)). \tag{ix}$$

This is an autonomous linear system, and we could if we wished analyze it in complete detail. But the asymptotics (equilibrium) are the most revealing:

$$\Gamma^+(t) \to (\alpha_1/\alpha_2)E[zX], \qquad \Gamma^-(t) \to (\alpha_3/\alpha_4)E[XX^T]\Gamma(t).$$

Therefore, at equilibrium,

$$\Gamma = \Gamma^+ - \Gamma^- = \frac{\alpha_1}{\alpha_2}E[zX] - \frac{\alpha_3}{\alpha_4}E[XX^T]\Gamma$$

$$\Rightarrow \Gamma = \frac{\alpha_1\alpha_4}{\alpha_2\alpha_3}\left(\frac{\alpha_4}{\alpha_3}I + E[XX^T]\right)^{-1}E[zX] \tag{x}$$

where $I$ is the $n \times n$ identity matrix.

Let us examine the information that $Y$ receives after learning (i.e. with $\Gamma$ given by (x)) and in the absence of the teacher signal, $z(t)$. Then,

$$\Gamma^{+T}X(t) - \Gamma^{-T}X(t) = X(t)^T\Gamma$$

$$= \frac{\alpha_1\alpha_4}{\alpha_2\alpha_3}X^T(t)\left(\frac{\alpha_4}{\alpha_3}I + E[XX^T]\right)^{-1}E[zX]. \tag{xi}$$

The constant, $\alpha_1\alpha_4/\alpha_2\alpha_3$, is obviously unimportant; the interpretation of (xi) will be clearest if we choose $\alpha_2/\alpha_1 = \alpha_4/\alpha_3 = \delta$ so that

$$X(t)^T\Gamma = X(t)^T(\delta I + E[XX^T])^{-1}E[zX]. \tag{xii}$$

Notice: *If* $\delta$ were 0 (it cannot be) this would be exactly the asymptotic output of Uttley's informon in the absence of the "classifying signal", $z(t)$. Thus (xii) approximates the minimum mean square error linear predictor of the classifying signal (here called the teacher signal), $z(t)$. The term $\delta I$, which may at first appear to be a nuisance, actually represents a potentially important improvement over the unmodified "optimal" solution. In fact, (xii) is a "ridge estimator" for $z(t)$, introduced by Hoerl and Kennard [**19**], and since then analyzed in some detail (see for example [**20**]). When $E[XX^T]$ is "nearly singular" (more precisely, "ill-conditioned"), the addition of $\delta I$ stabilizes the inverse in (xii), making it more accurately computable in a real system.

To appreciate the relevance of this in the present context, consider again the Uttley system (ii), but when $E[XX^T]$ is nearly singular (as would be the case, for example, if two of the channels $x_i(t)$ and $x_{i'}(t)$ were essentially redundant). For fixed $\varepsilon$, the approximation of (ii) by (iv) (i.e. the method of averaging) is made less accurate as (iv) is brought closer to instability–which is just what happens when $E[XX^T]$ is brought closer to singularity. Although the solution to (iv) will still asymptotically approach the desired (optimal) equilibrium, the solution to (ii) the *real system* will behave erratically, wandering far from the course predicted for it by (iv). Hence the system is, under these circumstances, unreliable. In contrast, the relative stability of (ix) (the "averaged system" for the Amari theory) is essentially unaffected by an ill-conditioned matrix $E[XX^T]$. As a consequence, the method of averaging remains in force and the desired solution, (xii), is still realized to within a good approximation.

Amari and Uttley, in the papers reviewed here, have each proposed neural-like mechanisms capable of learning pattern classifications. Thus modelled neurons in these theories learn to predict a one dimensional "classifying signal" based on the evidence available in an $n$ dimensional pattern. There is also the problem of postulating mechanisms by which the nervous system can commit to memory patterns themselves, both motor and sensory, as it is evidently capable of this task as well. Grossberg has proposed a theory for pattern learning in which individual neuron-like units learn to reproduce an entire pattern of activity (see, for example, [**17**] and [**18**]). Excitation of one of these units elicits an activity pattern in the "postsynaptic" units, and this pattern is identical (in the sense of relative, "figure to ground" activities) to a practice pattern arriving at these postsynasptic units during learning. Grossberg's analysis is deterministic and rather sophisticated. Although I will not add to the conclusions reached by that analysis, I will show how the essential properties of the learned behavior of the system can be anticipated by an application of the method of averaging.

All units (modelled neurons) of the network belong to one or both of two subpopulations: "$I$" represents the collection of subscripts belonging to those units in one of these subpopulations, and "$J$" represents the collection of subscripts associated with the other subpopulation. The units in the subpopulation "$J$" receive a pattern of input from outside the immediate network. Each unit in "$I$" contacts all units in "$J$", and, under appropriate conditions, will

learn to reproduce the pattern seen at "$J$". What makes this model particularly complex is that $I$ and $J$ are not assumed to be disjoint; the intersection, $I \cap J$, is arbitrary. In other words, the "receptor cells" in "$J$" may themselves be "sampling cells" in "$I$", realizing a feedback, rather than feedforward, system.

Let us represent by $b_i(t)$, $i \in I$, the output of the $i$ element of the "$I$" subpopulation at time $t$. The activity of the $j$ element of the "$J$" subpopulation, call it $x_j(t)$, is determined by the outputs of the units in "$I$", and an exogenous input, $c_j(t)$. Formally,

$$\frac{d}{dt}x_j(t) = -a(t)x_j(t) + \sum_{i \in I} b_i(t)\gamma_{ij}(t) + c_j(t) \tag{xiii}$$

for each $j \in J$, where $1/a(t)$ is an "instantaneous decay time" ($a(t) \geqslant 0$ for all $t$), and $\gamma_{ij}(t)$ is the synaptic or coupling strength for the $i \in I$ to $j \in J$ contact. The exogenous input to "$J$" is a pattern in the sense that it takes the form

$$c_j(t) = \psi(t)\theta_j, \quad \text{where } \sum_{j \in J} \theta_j = 1 \tag{xiv}$$

and $\theta_j \geqslant 0$ for all $j \in J$. $\psi(t)$ is the input intensity at time $t$, and may vary arbitrarily during the learning period.

As in the previous examples, learning is by modification of the "synaptic weights", $\gamma_{ij}(t)$, $i \in I, j \in J$,

$$d\gamma_{ij}(t)/dt = -d_i(t)\gamma_{ij}(t) + e_i(t)x_j(t). \tag{xv}$$

$e_i(t)$ plays a role analogous to $b_i(t)$, representing the signal from $i \in I$ available to effect change in $\gamma_{ij}(t)$. In the absence of a correlated $i \in I$ and $j \in J$ activity, i.e. when $e_i(t)x_j(t) = 0$, $\gamma_{ij}(t)$ decays towards 0 ($d_i(t) \geqslant 0$ for all $t$). Observe that (xv) is physically "realizable", in the sense that modification of $\gamma_{ij}(t)$ depends only on signals locally available, i.e. it depends only on pre- and postsynaptic activities. (Actually, the theory in [17] is developed for a system somewhat more general than (xiii) and (xv). The slightly specialized version here will serve for better illustration. The general system can be discussed in much the same way.)

What sort of results should we be looking for? Grossberg gives conditions under which the system demonstrates, asymptotically, the following learned behavior: With or without the exogenous pattern of input (xiv), activity in the "$I$" subpopulation leads to a reproduction of the learned pattern at $J$. In particular, after learning, "$I$" activity will produce a relative activity at $x_j$ equal to $\theta_j$

$$\frac{x_j(t)}{\sum_{k \in J} x_k(t)} \to \theta_j, \tag{xvi}$$

the relative strength of the exogenous signal at $j$. (See [17] for a precise formulation of results.) Let us see how we might anticipate this behavior by taking a probabilistic point of view and applying an LLN.

$x_j(t)$ models the activity of the $j \in J$ neuron, and should be "fast" relative to the other time dependent terms appearing in the right-hand side of (xiii). As a

first approximation then, it is not unreasonable to replace $x_j(t)$ by its "instantaneous equilibrium value", determined by setting $dx_j(t)/dt = 0$ in (xiii),

$$x_j(t) = \sum_{i \in I} \frac{b_i(t)}{a(t)} \gamma_{ij}(t) + \frac{c_j(t)}{a(t)}. \tag{xvii}$$

With this substitution, plus the one in (xiv), (xv) becomes

$$\frac{d}{dt}\gamma_{ij}(t) = -d_i(t)\gamma_{ij}(t) + \sum_{k \in I} \frac{e_i(t)b_k(t)}{a(t)} \gamma_{kj}(t) + \frac{e_i(t)\psi(t)}{a(t)} \theta_j. \tag{xviii}$$

Now let us again make the assumption that $\gamma_{ij}(t)$ is slowly varying, representing a long term memory trace. If, in equation (xv), we write $\varepsilon\tilde{d}_i(t)$ for $d_i(t)$ and $\varepsilon\tilde{e}_i(t)$ for $e_i(t)$, then with this assumption, we may take $\varepsilon$ to be small while $\tilde{d}_i(t)$ and $\tilde{e}_i(t)$ are still of order 1. (xviii) then becomes

$$\frac{d}{dt}\gamma_{ij}(t) = \varepsilon\left(-\tilde{d}_i(t)\gamma_{ij}(t) + \sum_{k \in I} \frac{\tilde{e}_i(t)b_k(t)}{a(t)} \gamma_{kj}(t) + \frac{\tilde{e}_i(t)\psi(t)}{a(t)} \theta_j\right),$$

which should be well approximated by the "averaged equation"

$$\begin{aligned}\frac{d}{dt}\gamma_{ij}(t) = \varepsilon\Bigg(&-E\left[\tilde{d}_i(t)\right]\gamma_{ij}(t) \\ &+ \sum_{k \in I} E\left[\frac{\tilde{e}_i(t)b_k(t)}{a(t)}\right]\gamma_{kj}(t) + E\left[\frac{\tilde{e}_i(t)\psi(t)}{a(t)}\right]\theta_j\Bigg).\end{aligned} \tag{xix}$$

Although (xix) is deterministic, it is not at all simple. In Grossberg's theory, $a(t)$, $b_i(t)$, $d_i(t)$, $e_i(t)$, and $\psi(t)$ may themselves depend on $\{x_i(t)\}$, $i \in I \cup J$, and $\{\gamma_{ij}(t)\}$, $i \in I$, $j \in J$, as long as the subscript conditions (e.g. $a(t)$ does not depend on $j$) are not violated. The possible dependence on the $\gamma_{ij}(t)$'s in particular prevents us from assuming that the expectations in (xix) are constant or that (xix) is linear. However, at any equilibrium point for (xix) the $\gamma_{ij}(t)$'s entering into these expectations are constant (although unknown), and in this case these expectations themselves may be assumed constant (just as in the previous two examples). At equilibrium, then, we may write

$$\gamma_{ij} = \sum_{k \in I} \frac{E\left[\tilde{e}_i b_k / a\right]}{E\left[\tilde{d}_i\right]} \gamma_{kj} + \frac{E\left[\tilde{e}_i \psi / a\right]}{E\left[\tilde{d}_i\right]} \theta_j, \tag{xx}$$

where the expectations, which may depend on $\{\gamma_{ij}\}$, $i \in I, j \in J$, do not depend on time.

Fix $j$. Then (xx) is a linear system of equations for $\gamma_{kj}$, $k \in I$, in which the only dependence on $j$ appears in the inhomogeneous terms (due to $\theta_j$). Hence, its solution (which I will assume exists) has the form

$$\gamma_{ij} = \sum_{k \in I} m_{ik} \frac{E\left[\tilde{e}_k \psi / a\right]}{E\left[\tilde{d}_k\right]} \theta_j,$$

where $M = \{m_{ik}\}$, $i, k \in I$, is a square matrix determined by the coefficients of $\gamma_{kj}$ in (xx) and does not depend on $j$. The point is that, at an equilibrium, $\gamma_{ij}$ (of the averaged equation) must have the form $\gamma_{ij} = \alpha_i \theta_j$, in which case (from (xvii))

$$x_j(t) = \left( \sum_{i \in I} \frac{b_i(t)}{a(t)} \alpha_i + \frac{\psi(t)}{a(t)} \right) \theta_j,$$

and this obviously implies that (xvi) holds. In other words, activity in "$I$" reproduces the practiced pattern at "$J$", even when that pattern is no longer present as an exogenous input (i.e. even when $\psi(t) = 0$). In fact, a direct (but much more involved) analysis of the *unapproximated* system, (xiii) and (xv), shows that under suitable conditions (xvi) holds there as well, and without a slowly varying assumption for $\gamma_{ij}(t)$ (see [17]).

Certain generalizations are immediately available, "free of charge". Since (xv) is well approximated by the averaged equation (xix), any substitution for the system (xiii) and (xv) that leaves (xix) unmodified will demonstrate essentially the same learned behavior. This includes, for example, allowing $a(t)$, $b_i(t)$, $d_i(t)$, and $e_i(t)$ to depend on $j$, provided that $E[d_{ij}]$, $E[e_{ij} b_{kj}/a_j]$, and $E[e_{ij}\psi/a_j]$ are still independent of $j$. Of course, this is as true in the previous two examples: the average equation represents a class of systems that must all exhibit approximately the same asymptotic behavior.

There are in the literature many other examples which can be, or already have been, treated in very much the same way. Although the three which I have discussed above should serve as a good introduction to this application of the method of averaging, I would also recommend (3), (4), (5), and (11), each of which contains an example of the explicit use of this technique in problems of neural or cognitive modelling.

**Population averaging: stable oscillations in a large system of modelled neurons.** In mathematical models of neural network activity, considerable use has been made of equations for the average activity of homogeneous collections of modelled neurons (some examples are in [7], [13], [15], [16], and [26]). The implicit assumption is that "macroscopic" (average) activity has a description which does not involve "microscopic" (individual neuronal) activities, in close analogy to the situation in statistical mechanics. Simulation experiments (see especially [1]) indicate that such a description is in fact broadly available, but there have been very few rigorous analytic results. This "population averaging" can again be viewed as an application of the LLN; here I will quote some analytic results (from [12]) which, in some instances, rigorously justify this application.

The discussion will be through a specific example. For this purpose, I will use essentially the model proposed by Miller and me (in [13]) for the generation of periodic phrenic nerve activity by the brainstem respiratory centers. Our analysis was based on the hypothesis that the LLN was operating in the proposed system. As I will indicate, for the equations discussed here the hypothesis is

indeed correct, meaning that the behavior of the entire (very large) system can be accurately described by a small number of prototype (averaged) equations.

In [13], Miller and I argue for a respiratory model based on reciprocating activities of negatively coupled inspiratory and expiratory populations of neurons, each of which is capable of independent stable oscillation if (theoretically) isolated from the other. The model postulates that these populations are further divided into excitatory and inhibitory subpopulations, and that the interaction between these subpopulations is responsible for the inspiratory and expiratory oscillations. For the discussion here, it will be enough to examine just one population, let us say the inspiratory population of neurons. Suppose that there are $n$ excitatory inspiratory neurons and $m$ inhibitory inspiratory neurons. $x_i(t)$ will denote the cell body membrane potential of the $i$th excitatory neuron at time $t$, and $y_i(t)$ will denote this potential for the $i$th inhibitory neuron. The dynamics of the modelled inspiratory population are described by the following systems of equations

$$\begin{aligned}\frac{d}{dt}x_i(t) &= -\alpha x_i(t) + \frac{1}{n}\sum_{j=1}^{n}\gamma_{ji}^{++}f(x_j(t)) \\ &\quad - \frac{1}{m}\sum_{j=1}^{m}\gamma_{ji}^{-+}g(y_j(t)), \qquad 1 \leqslant i \leqslant n, \\ \frac{d}{dt}y_i(t) &= -\beta y_i(t) + \frac{1}{n}\sum_{j=1}^{n}\gamma_{ji}^{+-}f(x_j(t)) \\ &\quad - \frac{1}{m}\sum_{j=1}^{m}\gamma_{ji}^{--}g(y_j(t)), \qquad 1 \leqslant i \leqslant m. \end{aligned} \tag{xxi}$$

Here,

$\alpha$ ($\beta$) is the inverse of the membrane decay time of excitatory (inhibitory) neurons;

$f(x_i(t))$ ($g(y_i(t))$) is the frequency of action potentials generated in the axon of the $i$th excitatory (inhibitory) neuron by a cell body membrane potential of $x_i(t)$ ($y_i(t)$). $f(x)$ and $g(x)$ are assumed to be bounded and increasing functions.

All $\gamma_{ji}$'s are nonnegative. $n^{-1}\gamma_{ji}^{++}$ is the coupling strength ("synaptic weight") from the $j$th excitatory to the $i$th excitatory neuron, $m^{-1}\gamma_{ji}^{-+}$ is the coupling strength from the $j$th inhibitory to the $i$th excitatory neuron, etc. When there is no synaptic connection between two neurons, the corresponding $\gamma_{ji}$ is zero. "$1/n$" and "$1/m$" embody the assumption that the total synaptic contribution to a neuron's input is "order 1", regardless of the number of these synapses.

Let us suppose that the system (xxi) represents a reasonable first approximation of the dynamics of the inspiratory population of brainstem neurons. With $n$ and $m$ of order $10^5$ (conservatively), there is the ontological problem of choosing $10^{10}$ or more parameters in (xxi) so as to achieve a stable oscillation with a specified period and wave form. If it were possible to specify that all synaptic weights of a given type of synapse (such as excitatory to excitatory) were

identical, then (xxi) would be perfectly described by just two prototype equations, one for the excitatory and one for the inhibitory subpopulations. Then, the biological problem would be entirely manageable, requiring the appropriate specification of only a very small number of parameters. But it is not a tenable proposition that this level of precision is achieved in a developing nervous system.

Miller and I (in [13]) have argued that the ontological problem is limited to the specification of target values for each of the four types of connections in (xxi), and that random fluctuations about these target values (means) will not influence the dynamics of the network as a whole. We reasoned heuristically, as follows: For each type of connection, take the case of excitatory to excitatory, let us model the synaptic weights, $\gamma_{ji}^{++}$, $1 \leq j, i \leq n$, as independent and identically distributed random variables, chosen from a distribution in which the mean only is genetically specified. Let $\gamma^{++}$ be the mean strength of an excitatory to excitatory synapse. When $n$ is very large, the dependence between any two excitatory neurons, or between any excitatory neuron and any synaptic weight, should be very small. Then, the excitatory input to the $i$th excitatory cell,

$$\frac{1}{n}\sum_{j=1}^{n}\gamma_{ji}^{++}f(x_j(t)), \tag{xxii}$$

should "look like" an average of independent random variables. The LLN would replace (xxii) by its mean

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n}\gamma_{ji}^{++}f(x_j(x)) &\approx \frac{1}{n}\sum_{j=1}^{n}E\left[\gamma_{ji}^{++}f(x_j(t))\right]\\ &\approx \frac{1}{n}\sum_{j=1}^{n}E\left[\gamma_{ji}^{++}\right]E\left[f(x_j(t))\right]\end{aligned}$$

(because $\gamma_{ji}^{++}$ and $x_j(t)$ are "nearly independent")

$$= E\left[\gamma_{ii}^{++}\right]E\left[f(x_i(t))\right]$$

(because all of the $\gamma_{ji}^{++}$'s and all of the $x_j(t)$'s are identically distributed)

$$= \gamma^{++}E\left[f(x_i(t))\right].$$

If this makes sense, then it applies as well to the other three input terms in (xxi). For each $i$, then, we expect that

$$\begin{aligned}\frac{d}{dt}x_i(t) &\approx -\alpha x_i(t) + \gamma^{++}E\left[f(x_i(t))\right] - \gamma^{-+}E\left[g(y_i(t))\right],\\ \frac{d}{dt}y_i(t) &\approx -\beta y_i(t) + \gamma^{+-}E\left[f(x_i(t))\right] - \gamma^{--}E\left[g(x_i(t))\right].\end{aligned} \tag{xxiii}$$

Since the right-hand side of (xxiii) is deterministic, $x_i(t)$ and $y_i(t)$ are nearly deterministic. In this case, $E[f(x_i(t))] \approx f(x_i(t))$ and $E[g(y_i(t))] \approx g(y_i(t))$. Put

this back in (xxiii) and conclude that the behavior of the entire system (xxi) should be well described by the two dimensional prototype system

$$\frac{d}{dt}x(t) = -\alpha x(t) + \gamma^{+}{}^{+}f(x(t)) - \gamma^{-}{}^{+}g(y(t)),$$
$$\frac{d}{dt}y(t) = -\beta y(t) + \gamma^{+}{}^{-}f(x(t)) - \gamma^{-}{}^{-}g(y(t)). \qquad \text{(xxiv)}$$

Miller and I assumed that (xxiv) provided an adequate description for (xxi), and analyzed the dynamics of (xxiv) as a possible model for the generation of inspiratory neuronal activity. In [12] it has been shown that (xxiv) does in fact provide an arbitrarily good approximation for (xxi) as $n$ and $m$ go to $\infty$. More specifically, as $n \to \infty$ and $m \to \infty$ all excitatory activities $x_i(t)$, $1 \leqslant i \leqslant n$, and all inhibitory activities $y_i(t)$, $1 \leqslant i \leqslant m$, will remain, respectively, arbitrarily close to the trajectories of $x(t)$ and $y(t)$, as defined by the prototype equations in (xxiv) (see [12] for details). In other words, the LLN is in force in (xxi), and the consequence is that the behavior of the entire system is determined by the parameters in the two dimensional system, (xxiv).

Simulations of this averaging effect can be quite striking. For example, we may choose the functions $f$ and $g$ and the six parameters in (xxiv) so that $(x(t), y(t))$ has a globally stable limit cycle. Then, in a typical experiment, with the standard deviation of each $\gamma$ larger than 50% of its mean, (xxi) already oscillates when $n = m = 7$. (For smaller $n$ and $m$, all activities approach an equilibrium.) But, for this still small system, the oscillation is quite different from the limit cycle trajectory predicted by (xxiv). When $n$ and $m$ are 80, however, the $x_i(t)$ and $y_i(t)$, $1 \leqslant i \leqslant 80$, trajectories are virtually indistinguishable from the prototype $x(t)$ and $y(t)$ trajectories. See [12] for phase portraits from one such experiment.

If the output of the respiratory centers can be simulated by a four dimensional system (two subpopulations of inspiratory and two subpopulations of expiratory neurons), why commit large numbers of neurons to the generation of this activity? One obvious reason is reliability. It is widely appreciated that there is advantage to redundancy in the nervous system, especially when promoting vital functions such as breathing. There may also be a second purpose, as is suggested by the averaging effect discussed here. The dynamics of a low dimensional system will be more critically dependent on individual parameters. The relevance of this is that each component of a small system generating respiratory activity would need to be specified with extreme accuracy, if the system is precisely to achieve a desired output. The alternative is to reach this precision by averaging: when in force, the LLN guarantees arbitrary precision in arbitrarily large systems. Of course, nothing said here needs to be limited to models for the control of breathing. Averaging is an available mechanism for the reliable and precise generation of activity by a homogeneous collection of neurons, whatever the physiological application.

## REFERENCES


1. S. I. Amari, *Characteristics of random nets of analog neuron-like elements*, IEEE Trans. on Systems, Man and Cybernetics, vol. SMC-2, 1972, pp. 643–657.

2. _____, *Neural theory of association and concept-formation*, Biol. Cybernet. **26** (1977), 175–185.

3. _____, *Topographic organization of nerve fields*, Bull. Math. Biol. **42** (1980), 339–364.

4. S. I. Amari and A. Takeuchi, *Mathematical theory on formation of category detecting nerve cells*, Biol. Cybernet. **29** (1978), 127–136.

5. E. L. Bienenstock, L. N. Cooper and P. Munro, *On the development of neuronal selectivity: orientation specificity and binocular interaction in visual cortex* (in preparation).

6. R. O. Duda and P. E. Hart, *Pattern classification and scene analysis*, Wiley, New York, 1973.

7. J. L. Feldman and J. D. Cowan, *Large-scale activity in neural nets. II. A model for the brainstem respiratory oscillator*, Biol. Cybernet. **17** (1975), 39–51.

8. S. Geman, *Application of stochastic averaging to learning systems*, Brain Theory Newsletter **3** (1978), 69–71.

9. _____, *Some averaging and stability results for random differential equations*, SIAM J. Appl. Math. **36** (1979), 86–105.

10. _____, *A method of averaging for random differential equations with applications to stability and stochastic approximations*, Approximate Solution of Random Equations (A. T. Bharucha-Reid, ed.), North-Holland Series in Probability and Appl. Math., North-Holland, Amsterdam, 1979, pp. 49–85.

11. _____, *Notes on a self-organizing machine*, Parallel Models of Associative Memory (G. Hinton and J. Anderson, eds.), Erlbaum Associates, Hillsdale, N. J., 1980.

12. _____, *Almost sure stable oscillations in a large system of randomly coupled equations*, Reports on Pattern Analysis, no. 97, Div. Appl. Math., Brown University, 1980 (submitted).

13. S. Geman and M. Miller, *Computer simulation of brainstem respiratory activity*, J. Appl. Physiol. **41** (1976), 931–938.

14. S. Geman and C. R. Hwang, *A chaos hypothesis for some large systems of random equations*, Reports on Pattern Analysis, no. 82, Div. Appl. Math., Brown University, 1980 (submitted).

15. S. Grossberg, *A neural theory of punishment and avoidance. I. Qualitative theory*. Math. Biosci. **15** (1972), 39–67.

16. _____, *A neural theory of punishment and avoidance. II. Quantitative theory*. Math. Biosci. **15** (1972), 253–285.

17. _____, *Pattern learning by functional-differential neural networks with arbitrary path weights*, Delay and Functional Differential Equations and Their Applications (K. Schmitt, ed.), Academic Press, New York, 1972, pp. 121–160.

18. _____, *Classical and instrumental learning by neural networks*, Progress in Theoret. Biol. **3** (1974), 51–141.

19. A. E. Hoerl and R. W. Kennard, *Ridge regression. Biased estimation for non-orthogonal problems*, Technometrics **12** (1970), 55–67.

20. K. Kadiyala, *Operational ridge regression estimator under the prediction goal*, Comm. Statist. A-Theory Methods **8** (1979), 1377–1391.

21. Iu. A. Mitropolsky, *Averaging method in non-linear mechanics*, Internat. J. Non-Linear Mech. **2** (1967), 69–96.

22. A. M. Uttley, *A two-pathway informon theory of conditioning and adaptive pattern recognition*, Brain Res. **102** (1976), 23–35.

23. _____, *Simulation studies of learning in an informon network*, Brain Res. **102** (1976), 37–53.

24. _____, *Neurophysiological predictions of a two-pathway informon theory of neural conditioning*, Brain Res. **102** (1976), 55–70.

25. M. T. Wasan, *Stochastic approximation*, Cambridge Univ. Press, Cambridge, 1969.

26. H. R. Wilson and J. D. Cowan, *A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue*, Kybernetik **13** (1973), 55–80.


DIVISION OF APPLIED MATHEMATICS, BROWN UNIVERSITY, PROVIDENCE, RHODE ISLAND 02912