

Real Expectations: A Harmony Theoretic Approach to Decision Making Under  
Uncertainty

El-Gama, Mahmoud A.

Working Paper No. 176  
February 1989

University of  
Rochester

**Real Expectations:  
A Harmony Theoretic Approach to Decision Making Under  
Uncertainty**

Mahmoud A. El-Gamal

Rochester Center for Economic Research  
Working Paper No. 176

REAL EXPECTATIONS: A HARMONY THEORETIC APPROACH TO  
DECISION MAKING UNDER UNCERTAINTY

Mahmoud A. El-Gamal  
Department of Economics  
University of Rochester  
Rochester, NY 14627

Working Paper No. 176

First Draft November 1988

Revised February 1989



**Real Expectations:  
A Harmony Theoretic Approach to  
Decision Making Under Uncertainty**

Mahmoud A. El-Gamal  
Department of Economics  
University of Rochester  
Rochester, N.Y. 14627

First Draft November 2<sup>nd</sup>, 1988

Revised February 1<sup>st</sup>, 1989

**Abstract**

The prevailing methodology of economic modelling of decision making under uncertainty referred to as “rational expectations” assumes the maximization of expected payoff where some ad hoc assumption about the formation of those expectations is made. Whether a simple weighted average of past values (Nerlove (1958)) or the true distribution in equilibrium (Muth (1961) and Lucas and Prescott (1971)) is used to form those expectations, there seems to be little justification from probability theory for such procedures. This paper uses techniques of statistical mechanics that were adapted to information theory by Shannon (1962) and more recently to cognitive science in what Smolensky (1986) calls “Harmony Theory”. We justify the use of those methods from the developments of the foundations of probability theory by Bernoulli, Bayes, Laplace, Jeffreys, and Jaynes; briefly tracing the evolution of Bayesian inference and the development of the maximum entropy formalism. The generation of heat baths for continuous state spaces is shown to give the same asymptotic distribution as the maximum entropy formalism. This yields a biologically feasible method of achieving initial distributions of relevant unknown variables. The Harris ergodicity of the heat bath is also shown to make the computation of expectations of functions of those variables easy and inexpensive due to von-Neumann and Birkhoff type ergodic theorems. Finally, annealing schedules in continuous state spaces are shown to converge to the mode of the distribution of those variables whose use in decision making is briefly argued to be a more “real” model for most economic agents even though it might not be “rational” in the Bayesian sense.

*Keywords:* Rational Expectations methodology, parallel distributed processing, cognitive systems, chaotic algorithms, simulated annealing.



## 1. Introduction

There is little doubt that the most dominating method of economic modelling of decision making under uncertainty for the past two decades has been that of “rational expectations”. That method’s development is well studied in the economic literature, and therefore, no attempt to survey that development will be made in this paper. It is of interest, however, to see how the first steps towards the development of that paradigm were taken. The consensus among economists seemed to be that generalizing the method of maximizing payoff in deterministic settings to the stochastic setting can best be made by letting the agents maximize expected payoff. The problem that naturally arose out of that concept was the way to form expectations of unknown future variables. The initial response was to simply take some weighted average of past values of that variable as the agent’s estimate of the relevant expectations. This method was viewed by many as truly ad hoc, and hence, a more systematic method for the formation of expectations was required.

Almost all economists will agree that the definition of rational expectations due to Muth (1961) was the first popularly known attempt at such a systematization of the formation of expectations, and that it inspired most of the work that followed although the exact definition of Muth’s was not necessarily uniformly used. Muth’s main idea was that “...expectations, since they are informed predictions of future events, are essentially the same as the predictions of the relevant economic theory”. Most subsequent work took the view that that justifies taking the mathematical expectation of the relevant variables using the true distribution of those variables in equilibrium. It is the view of this author that that procedure is no less ad hoc than the method of taking some arbitrary convex combination of past values.

There is no foundation in probability theory or cognitive science for the use of the rational expectations hypothesis. The next section will quickly review the development of

probability theoretic foundations from Bernoulli, through Bayes and Laplace to the contemporary work of Jeffreys and Jaynes which discusses the choice of initial (or prior) distributions that a rational agent will use in decision making under uncertainty. If the reader is skeptical about Bayesian foundations of probability theory and prefers the frequentist approach (in which case he/she is in very good company), one can still argue against the use of the true (or - as a frequentist would say - objective) probabilities obtained from the equilibrium conditions of the theory. One can first urge the reader to refer to the studies by Jeffreys (1967) and Cox (in Levine and Tribus 1979) for a philosophical critique of the frequentist approach. Since this is not the arena for such a debate, however, we raise another objection to the rational expectations hypothesis, namely, that from what we know in cognitive and neural science, there is no justification of such a hypothesis. One major criticism that one can launch is the very serial nature of the computation of expectations à la rational expectations which makes it virtually impossible to take place in a human brain whose comparative advantage lies mainly in performing jobs that are of a highly parallel nature. The other major criticism that we wish to raise in this paper is the independence of the distribution used from the amount of information the agent has. This latter criticism obviously does not apply to models that take into account asymmetric information, but it seems that those models also introduce that asymmetry in a rather ad hoc fashion. We would like to have a systematic procedure by means of which the presumed distribution of the unknown variables is determined from the amount of information that the agent has. This will obviously lead us into the realm of information theory where Shannon (1962) adapted the notion of entropy from statistical mechanics and used it as a measure of information. Later, Jaynes (see papers in Rosenkrantz 1983) developed what came to be known as the maximum entropy formalism which offers the natural basis for a consistent choice of initial distributions (priors) of agents based on their respective knowledge.

This paper has the modest objective of circumventing the objections in the previous paragraph by providing an algorithm which is physically feasible, and which results in the



maximum entropy distribution of relevant unknown variables. The results that we shall achieve are generalization of recent results in cognitive science by Smolensky (1986), and Geman and Geman (1984) to continuous state spaces. We shall introduce a variant of the energy function in statistical mechanics which Smolensky called a “harmony function”, and then we shall show that a “heat bath” (the algorithm that we shall assume to take place in the human brain, and we use that name due to an old tradition in statistical mechanics) converges to the maximum entropy distribution. We also demonstrate that the heat bath exhibits the statistical property of Harris ergodicity which allows us to use von-Neumann and Birkhoff type ergodic theorems to easily compute expectations by taking time averages from the heat bath, thus avoiding the previous criticism that expectations are hard to compute in the human brain. Our final objective is to offer an alternative for theorists who find expected payoff as an unacceptable objective function. Expected payoff has been shown to miserably fail almost all experimental tests, and it is technically questionable since expectations need not exist and even if they exist, they may not be in the support of the distribution in which case it is very unreasonable to maximize expected payoff. One can argue that the majority of economic agents are not even aware of what a stochastic structure to a problem means. They simply know some relevant variables and do not know others, and we argue that they would proceed in much the same way they do when they see blurred road signs or characters in a word, namely by trying to “complete the picture” by choosing the most likely values for those unknown variables. Statistically, this will correspond to using the mode of the maximum entropy distribution. We show that the result of Geman and Geman (1984) which shows that lowering the “temperature” in the heat bath at a slow enough rate will yield convergence to the mode of the maximum entropy distribution. There is another added dimension of realism in that latter procedure since the agents are even relieved from the burden of having to compute time averages that is required for “real expectations” (the name we give to the formation of expectations using the maximum entropy distribution).

The paper will only attempt to establish the plausibility and prove the validity of the mathematical concepts involved without attempting to apply “real expectations” to any particular economic models. One of the advantages of the proposed modelling methods is its ability to accommodate for learning as you play in experimental settings where it is not clear how fast the participants actually learn the structure of the game. Another of the main targets of this research is the ability to efficiently and cheaply be able to simulate decision making on parallel computers, a program whose full advantage should be realized once we know how to model whole economies as parallel processors thus introducing two levels of parallelism. But now, let us begin with the objective at hand. Section 2 will briefly trace the development of probability theory and justify the maximum entropy formalism. Section 3 will then introduce a variant of Smolensky’s (1986) version of statistical mechanical heat baths in what he calls “Harmony Theory” and argue for its realism in economic settings as well. Section 4 will then state the major theorems from Smolensky (1986) and Geman and Geman (1984) generalized to the continuous state space and provide proofs where they differ from the ones for discrete state spaces. Finally, Section 5 will conclude the paper.

## **2. Bayesian theory and the maximum entropy formalism<sup>1</sup>**

The study of decision making under uncertainty and the nature of probability has occupied scientists, philosophers, and mathematicians for a number of centuries. A long tradition of recognizing probabilities as measures of the state of knowledge dates back at least to Jacob Bernoulli (1713) in his discussion of the principle of insufficient reason (in modern terms this will be referred to as uncertainty). All of his work was based on the existence of many “equally likely” events that may occur. In the equally likely framework, Bernoulli obtained his well known weak law of large numbers, and then DeMoivre and

---

<sup>1</sup> This section is based on Jaynes’s “Where do we stand on maximum entropy” (1978), and “prior probabilities” (1968) in Rosenkrantz (ed. 1983). For a more detailed discussion, refer to those papers.

Laplace produced what is generally cited as the first of a long and rich family of central limit theorems. The question of inferring the true probabilities from observed frequencies then preoccupied many mathematicians of the time. The first attempt at inverting the binomial probability model was due to Thomas Bayes (1763) but the first general derivation of what came to be known as Bayes's theorem or Bayes's rule was due to Laplace in 1774. Whether to ascribe the discovery to Bayes or to Laplace is only of interest to historians of scientific thought (and perhaps to people who indulge in British versus continental greatness debates). The important and essential idea to note in that development is the choice of priors for application in Bayes updating in real problems. The principle of insufficient reason suggested to the scientists and mathematicians of the time to use uniform priors; i.e. where no further knowledge is available, one should start with a prior that all possible events are equally likely, and then use accumulated evidence to update that prior.

In the following two centuries, scientists, mathematicians, and philosophers engaged in a long and fruitless debate about the definition of probability and whether the Bernoulli-Bayes-Laplace view of probability as a measure of ignorance or the modern frequentist view of probability is the true one. As Jaynes (1978) notes, there is truly no substantive content to that argument other than a fight over the right to use a term. In this paper, we have decided that for economic agents with lack of information about certain variables of interest, the distributions they ascribe are a measure of their ignorance. But if we agree with Laplace that with no information at all, one should assume that all events are equally likely, how does one decide on the best prior to use if there was a certain amount of knowledge possessed by the agents? Jaynes and others have argued for the use of Shannon's (1962) measure of information

$$S(p) = - \sum_{i=1}^n p_i \cdot \log(p_i)$$

which is mathematically equivalent to entropy as used in Gibbs's statistical mechanics. We shall avoid arguments here about the use of the term entropy in Shannon's information theory, and whether Gibbs or Maxwell entropies should be used. We simply note that with

no further information, maximizing Shannon's measure of information (or more concisely maximizing entropy) yields the uniform  $p_i = 1/n$  for all  $i$  which is the outcome of Laplace's principle of insufficient reason. If more information is available, for instance, if we know that the expected value of some variable  $X$  takes a certain value, then we can solve the problem

$$\begin{aligned} \max_p \quad & - \sum_{i=1}^n p_i \cdot \log(p_i) \\ \text{s.t.} \quad & \sum_{i=1}^n p_i \cdot X_i = x \end{aligned}$$

The general philosophy behind the maximum entropy formalism, therefore, is to choose the distribution that is consistent with all the information that we have (satisfying the constraints of the maximization problem) and which does not presuppose any more information. Many uses of the maximum entropy formalism in the physical, social, and cognitive sciences have been found, and the method received great success (e.g. see Levine and Tribus (1979), and Justice (1986)).

One problem we encounter in extending the maximum entropy formalism to continuous state spaces is that the expression  $-\int p \cdot \log(p)$  does not arise as the limit of Shannon's entropy as  $n \uparrow \infty$  as postulated by Shannon (1962). Jaynes ((1963) in Rosenkrantz (1983)) showed quite easily that with  $x_1, \dots, x_n$  as the possible realizations, as  $n \uparrow \infty$ ,  $[n(x_{i-1} - x_i)] \rightarrow [m(x_i)]^{-1}$ , where  $m(\cdot)$  is a well defined density. But we can write the discrete probability distribution  $q_1, \dots, q_n$  as  $q_i = p(x_i)(x_{i+1} - x_i)$  where  $p(\cdot)$  is the limiting density of the distribution as  $n \uparrow \infty$ . This tells us that  $q_i \rightarrow p(x_i) \cdot [n \cdot m(x_i)]^{-1}$ . The discrete entropy  $S^d(q)$  will therefore converge to

$$S^d(q) \rightarrow - \int p(x) \cdot \log\left(\frac{p(x)}{n \cdot m(x)}\right) \cdot dx$$

which after subtracting the term  $\log(n)$  gives us the continuous version of Shannon's entropy

$$- \int p(x) \cdot \log\left(\frac{p(x)}{m(x)}\right) \cdot dx$$

One of the important and essential features of Shannon's entropy in the discrete case was its invariance to many transformations of the data, which contributed to the sense in which maximizing entropy was optimal. The choice of the measure defined by  $m(x)$  will have to be used to ensure such an invariance in the continuous case (since  $p(x)$  and  $m(x)$  will transform in the same way leaving  $S$  invariant to the change of variables). In general, one may need to discuss the choice of those prior measures  $m(x)$  according to group theoretic criteria (e.g. see Jaynes (1968) in Rosenkrantz (1983) for examples). For the purposes of this paper, we shall simplify the problem by letting our space be  $X = [0, 1]^n$  and choose the uniform prior  $m(x) = 1$  for  $x \in X$  (which corresponds to Lebesgue measure which by being a Haar measure is invariant under shift and rotation; for this state space, the use of Shannon's own postulated generalization of entropy to continuous state spaces is, therefore, justified), in which case unconstrained maximization of entropy will yield

$$p(x) = \frac{m(x)}{\int_X m(x).dx} = 1/1 = 1$$

which is the principle of insufficient reason revisited. We reiterate that generalizing the results of the next two sections where  $m(x)$  is more generally chosen for other state spaces is possible, but there does not seem to be a need to use it at this stage.

### **3. Introduction to harmony theory:<sup>2</sup>**

We start with a complete probability space in the strict sense  $(E, \mathfrak{F}, P)$  where  $E$  is the economic environment in which the agent is making his/her decisions,  $\mathfrak{F}$  is a  $\sigma$ -algebra of subsets of  $E$ , and  $P$  is a probability measure on  $(E, \mathfrak{F})$ . The environment is not directly observable to the agent, but there is a transduction map  $T: E \rightarrow R$  where  $R = [0, 1]^n$  is a

---

<sup>2</sup> This section is an adaptation of Smolensky's framework as in Smolensky (1986), where a number of changes in the interpretation and the functional and probabilistic specification have been made to resemble the framework of economic modelling.

representation space. This map will not play any role in our analysis, but it is used to make sense of the definition of representations of reality as random variables. The map  $T$  will therefore induce a probability measure  $p$  on  $R$  by  $p(\cdot) = P \circ T^{-1}(\cdot)$ . The agent's environment (to be distinguished from the Economic environment or the true environment) is completely specified by  $R$  and  $p$ , hence, we shall be referring to  $p$  as the agent's environment. Given a representation  $r \in R$ , the agent will make his/her decision in a deterministic fashion. In other words, if the agent knew the actual environment (i.e. the realization in  $E$ ), he would be able to solve his problem with certainty. The general representation space is introduced to define how he/she will act when the actual outcome of the economic environment is not observed but some measurable function on it is; and  $R$  is assumed to be a sufficient statistic in the sense that knowing  $R$  is all the agent needs to make a deterministic decision. An example of an element of  $R$  may be a price of some commodity, the agent's decision to purchase may depend on the weather, its effect on the crop, etc. but it is sufficient for the agent's purposes to see the price for him to be able to make his decision.

The uncertainty in the model is introduced via an input function  $D: E \rightarrow I$  where  $I = R$  is the same space. An individual's input is assumed to cover some elements of  $R$  but not others. Hence, the agent will know with certainty some of the random variables that are sufficient to make his decision, but he will have to guess about the others. The default input for a variable on which no data is available is 0. Since all the measures we shall consider will be absolutely continuous with respect to Lebesgue measure, the probability of an actual data point being 0 is itself 0, and hence, we need not worry about such cases. Now, the agent attempts to fill out the 0 entries in his input vector to achieve a representation vector on which to base his decision; for this 'fill in the blanks' mission, the agent has to find a completion function. A completion function  $c: I \rightarrow 2^R$  will then consider all representations of the environment which agree with the input vector where data is available. When  $r \in R$  is a completion of  $i \in I$ , we write  $r \supset i$ .

Now the agent may consider a number of possible representation vectors that are compatible with the input he has received. For each such possible representation, we define a basic event as the event that the some of the elements of the representation vector is equal to some value. Therefore, a basic event  $\alpha$  in our environment will be an assignment of values to some of the elements of  $R$ , i.e.  $\alpha = (r_{i_1} = b_1), \dots, (r_{i_\beta} = b_\beta)$ , where  $(b_1, \dots, b_\beta) \in [0, 1]^\beta$ . The default value for entries in the representation vector that are not known under that basic event will be set to 0 to agree with the input vector which has the same non-zero values and the same unknown entries. We shall construct a gain function which determines under any particular event  $\alpha$  the distance between  $\alpha$  and our representation vector  $r$ . A metric  $\rho(r, k_\alpha)$  is now introduced as a measure of the usefulness of the knowledge vector  $k_\alpha$ , to be defined below, one example of such a metric would be

$$\xi_\alpha(r) = \rho(r, k_\alpha) = 1 - \frac{1}{\beta} \sum_{\mu=1}^{\beta} (r_{i_\mu} - b_\mu)^2 \quad (2.1)$$

which is 1 - mean squared error; none of the results in following sections depend on the particular form of the gain function; and we will just keep it defined as a general metric. We impose some restrictions on the metric, however, only as a normalization. We require that if we have a perfect match, then  $\xi_\alpha(r) = 1$ , and the largest mismatch will give a  $\xi_\alpha(r) = 0$ . Now, for each event  $\alpha$ , there corresponds a knowledge vector  $k_\alpha = (0, \dots, 0, b_{i_1}, 0, \dots, 0, b_{i_2}, \dots)$ , and an atom  $\alpha$  in the agent's mind which decides what to do if  $\alpha$  was the case. That is reminiscent of the standard way of modeling decision making under uncertainty by considering each possible configuration of reality and what the utility of such an outcome would be. Here, we actually assign physical pseudo-entities (which we call atoms) that correspond each to such a possible configuration. Each of the atoms  $\alpha \in O$  where  $O$  is the set of observables (i.e. all possible events that the agent can witness), has two parameters assigned to it from the agent's cognitive system. The atom may be activated or not according to the activation parameter  $a_\alpha \in \{0, 1\}$ , and the strength of the connection to the atom may be increased or reduced by means of the strength parameter  $\sigma_\alpha$ . So, if a particular atom corresponds to a configuration of reality which has a negative impact on the overall understanding of the

cognitive system (this overall understanding will soon be rigorously defined by the harmony function), it will not be activated which means that that particular configuration will not be considered by the system. Depending on how the different configurations that have a positive contribution to harmony interact, their respective strengths will be set to maximize that harmony.

Now, to choose a representation that is in some sense optimal, we rigorously define a “harmony function” which assigns to each representation vector  $r \in R$  the real number

$$H_k(r) = \sum_{\alpha \in O} a_\alpha \sigma_\alpha h_k(r, k_\alpha) \quad (2.2)$$

where  $h_k(r, k_\alpha)$  is the contribution of atom  $\alpha$  to the harmony of representation  $r$ . We choose for this function the form

$$h_k(r, k_\alpha) = \frac{1}{n} \sum_{i=1}^n J(r_i - k_{\alpha_i}) - \kappa \quad (2.3)$$

where  $J(x) = 1$  if  $x = 0$  and  $J(x) = -1$  otherwise, and  $\kappa \in [-1, 1]$ . Now, it is only reasonable that an atom will be activated if and only if it has a non-negative contribution to the harmony function. So if  $\kappa = 1$ , then atom  $\alpha$  will only be activated if there is a perfect match. At  $\kappa = 0$ , half or more matches are required, and at  $\kappa = -1$ , the atom will always be activated. As with the choice of  $\xi_\alpha$ , we can choose an arbitrary form for the function  $h_k$  which satisfies certain regularity conditions and all the following results will still hold.

We now turn back to the issue of assigning a probability distribution to the various representation vectors that we may consider. We start with the maximum entropy distribution whose merits were discussed in the previous section. Given an environment  $p$ , and a set of observables  $O$ , the observable statistics of the environment are defined as the set  $\{p(\alpha)\}_{\alpha \in O}$ . The entropy of a distribution  $p$  is taken to be the continuous version of the Shannon entropy (Shannon, 1962)

$$S(p) = - \int_{r \in [0,1]^n} p(r) \cdot \ln(p(r)) \cdot dr$$



The maximum entropy estimate of  $p$  with observables  $O$  is the probability distribution with maximum entropy among the class of distributions that have the same observable statistics as  $p$ . This notion will give us the class of distributions that add the least possible amount of information over what is observed. Given any distribution  $p$  on  $R$ , the maximum likelihood completion of an input vector  $i \in I$  will then be

$$c(i) = \{r \in R: \forall r' \in R \text{ with } r' \supset i; p(r) \geq p(r')\}$$

This completion function will seem to be the one that we would highly recommend investigating given the discussion in section 2 of the paper. The result of Theorem 1 in section 4 will be to show that this completion function will be equivalent to the completion function arising from a natural and physically feasible algorithm based on the harmony function defined above, to which we now turn.

We shall investigate distributions of representation vectors that depend on the harmony function, and the analytic form of the densities of those distribution will turn out to be of the familiar Gibbs form. We shall offer two different methods to arrive at that form of the densities in question. We first start with the more heuristic version due to Smolensky (1986). We assume that the agent will use his harmony function  $H_k$  to estimate the density of the relevant variables summarized by his/her representation vector. The form of that distribution function can be achieved in one of two ways. The first way à la Smolensky (1986) is to impose the restriction that if the cognitive system was to be split into two unconnected parts, the probability density of the two corresponding sets of observable vectors should be independent. In other words, if we were to split the space of observable atoms  $O$  in two parts  $O_1$ , and  $O_2$ , then we require that the density

$$f(H_k(r)) = f(H_k^{O_1}(r) + H_k^{O_2}(r)) = f(H_k(r)) \cdot f(H_k(r))$$

The interpretation of this condition is that if we were to have two people collect the information instead of one, with each of them considering only some of the knowledge atoms, where

there is no overlap between their investigations, then we should expect their outcomes to be independent of each other. It is well known that the only family of functions that map addition into multiplication is the exponential family, and hence,  $p(r) = A \cdot a^{H(r)} = A \cdot e^{H(r)/T}$ , where  $T = \ln(a)$ . The constant  $A$  will obviously be chosen to normalize the integral of  $p(r)$  to unity. This density is commonly known in statistical physics as the Gibbs density, and in the statistical physics framework,  $H$  will be known as the energy, and  $T$  as the temperature of the system.

The more formally rigorous way of achieving the Gibbs distribution as the proper distribution of the representation vectors is to utilize the equivalence between Gibbs distributions and Markov random fields. This equivalence relationship is discussed in Geman and Geman (1984) for the case with a discrete state space. The same equivalence result for continuous state spaces, and the intuition behind it is quite similar in many ways to Smolensky's derivation as well as Geman and Geman's. The idea of a Markov random field was first developed by Dobrushin (1968) to generalize the notion of Markovian processes in one dimension (usually time) to the multi-dimensional case. The relevance of this equivalence result for Geman and Geman (1984) was due to the nature of their problem where the neighboring pixels in a blurred image were the important ones in determining the value for any particular pixel. A similar argument could be used here provided that the variables of interest are arranged in the agent's representation vector so that adjacent entries are entries of variables with strong mutual information, and distant entries do not influence each other. The other important aspect of this equivalence for our purposes in this paper is its implication that during the process of inference (to be studied shortly), only neighboring atoms need communicate with each other which makes the implementation of the algorithm to be studied in a parallel processor such as the human brain or the Boltzmann machine feasible and fast. On a more technical note, the equivalence also figures heavily in the proofs of Geman and Geman (1984), and hence, the analogue of their results for the continuous state space requires the availability of that equivalence result for continuous state spaces.

Now, we are ready to state the rigorous derivation of the Gibbs density, the remainder of this section is based mostly on Rozanov (1982, pp. 55-74), and then we shall be ready to proceed with the estimation procedures. We say that a  $\sigma$ -algebra  $B$  splits the two  $\sigma$ -algebras  $A_1$  and  $A_2$  if

$$Pr(a_1 \cup a_2 | B) = Pr(a_1 | B) \cdot Pr(a_2 | B)$$

for all  $a_1 \in A_1$  and  $a_2 \in A_2$ . In the standard uni-dimensional Markov case, we think of a random variable  $\eta_t$  and let  $A_1 = \mathfrak{F}(-\infty, t)$ ,  $A_2 = \mathfrak{F}(t, \infty)$ , and  $B = \mathfrak{F}(t)$ , where as usual,  $\mathfrak{F}(i, j)$  is the  $\sigma$ -algebra generated by  $\eta_i, \dots, \eta_j$ , and the Markovian condition states, conditional on the present, the past and the future are independent which is equivalent to saying that  $B$  splits  $A_1$  and  $A_2$ . For a general Markov random field, we shall be thinking of  $T$  to be a general locally compact metric space instead of the unidimensional "time" dimension. We now proceed to define more of the mathematical primitives needed for the understanding of Markov Random Fields.

*Definition:* Let a collection of  $\sigma$ -algebras  $A(S)$  be connected with domains  $S \subset T$  where  $T$  is some locally compact metric space. Consider all open domains  $S \subset T$ , then we call  $A(S)$  a *random field* (RF) if  $A(S' \cup S'') = A(S') \vee A(S'')$

*Definition:* Let  $G$  be a system of open domains of  $S \subset T$ , and for all  $S$  define  $S_1 = S$ ,  $\Gamma = \delta(S)$ ,  $S_2 = T \setminus \bar{S}$  where as usual  $\delta(\cdot)$  is the boundary, and  $\bar{\cdot}$  is the closure of the set. We call  $S_2$  the complement domain of  $S$ , and we call  $\Gamma$  the boundary between  $S_1$  and  $S_2$ . For the next definitions, we shall take  $\Gamma^\epsilon$  to be an  $\epsilon$  neighborhood of  $\Gamma$  in the proper topology.

*Definition:* The random field  $A(S)$  is a *Markov Random Field* (MRF) w.r.t. a system  $G$  defined above if  $\forall S \in G$ , the boundary  $\Gamma$  splits  $S_1$  and  $S_2$ ; i.e. if the  $\sigma$ -algebras  $A(S_1)$ ,  $A(\Gamma^\epsilon)$ , and  $A(S_2)$  form a Markov sequence for  $\epsilon$  small enough.

Now, given a MRF, we can consider additive functionals on  $A(S)$ , i.e., real valued

functions on  $A(S)$  s.t. for any  $S_1$  and  $S_2 \in A(S)$ ,

$$\eta(S_1 \cup S_2) = \eta(S_1) + \eta(S_2) - \eta(S_1 \cap S_2)$$

*Lemma:*

Given  $\eta$  an additive functional on a Markov Random Field, if  $\int \eta(S) < \infty$  for all  $S \subset T$ , has a density

$$p = \frac{1}{\int e^{\eta(T)}} e^{\eta(T)}$$

*proof* is trivial since the set  $S_1 \cap S_2 = \emptyset$  and, as with the heuristic development in Smolensky, we notice that the only function that maps addition into multiplication is the exponential.

#### 4. Estimation of the distribution, and the maximum likelihood completion

An agent's problem is fully characterized by the quintuple  $(R, p, O, \pi, c)$  where  $R = [0, 1]^n$  is the representation space,  $p$  is the environment,  $O$  is a set of statistical observables,  $\pi$  is the maximum entropy estimate of the environment  $p$  which agrees with it at the observables  $O$ , and  $c$  is the maximum likelihood completion function with the density  $\pi$ . We also define the generic Gibbs density

$$p_V(x) = Z^{-1} e^{V(x)} \quad (3.1)$$

where

$$Z = \int_{x \in X} e^{V(x)} . dx \quad (3.2)$$

The first theorem we state here is a generalization of what Smolensky (1986) referred to as the competence theorem of the agent with the above defined problem. Part 1 of the

theorem states that for appropriate coefficients (which come naturally out of the constrained optimization problem of finding the maximum entropy density  $\pi$ ), the maximum entropy density will be a Gibbs density with “energy”  $U$  equal to a linear combination of the penalty function  $\xi_\alpha$  at all observable events. The second part of the theorem states that the maximum likelihood completion function with density  $\pi$  is equal to the maximum likelihood completion function with the Gibbs density defined by the energy  $U$  of part 1 of the theorem, which in turn is equal to the maximum likelihood completion with the Gibbs density with energy equalling the Harmony function defined in the previous section. The statement and proof of the theorem below are adaptations of the aforementioned theorem in Smolensky (1986) after the necessary changes in functional forms and algebraic manipulations have been made to generalize to the continuous representation space.

*Theorem 1:*

1. *The maximum entropy density  $\pi$  of the agent’s problem  $(R, p, O, \pi, c)$  is the Gibbs density (as in (3.1) and (3.2)) defined by  $U(r) = \sum_{\alpha \in O} \lambda_\alpha \xi_\alpha(r)$  for appropriate weights  $\lambda_\alpha$ , where  $\xi_\alpha$  is defined in (2.1).*
2. *The maximum likelihood completion function  $c$  corresponding to  $\pi$  of part 1. is the completion function defined by the Gibbs distribution with energy  $H_k(r)$  as defined by equations (2.2) and (2.3) for suitable choices of  $\sigma_\alpha$ , and  $\kappa$ .<sup>3</sup> ○*

*Proof of Theorem 1:*

The constrained optimization problem defined by the maximum entropy density determination

---

<sup>3</sup> The existence of a simple algorithm which converges to the proper values of  $\lambda$  and  $\sigma$  needed for this theorem will be discussed in the end of this paper after the two major results in Theorems 1 & 2 are stated and proofs are provided.

is

$$\begin{aligned} \max_{\pi} \quad & S(\pi) = - \int_{r \in R} \pi(r) \cdot \ln(\pi(r)) \cdot dr \\ \text{s.t.} \quad & \int_{r \in R} \pi(r) \cdot dr = 1 \\ \text{and} \quad & \int_{r \in R} \pi(r) \cdot \xi_{\alpha}(r) \cdot dr = p_{\alpha} \end{aligned}$$

The first constraint insures that the estimate is a density and the second insures that the maximum entropy estimate agrees with the environment for all observable events. When we write the Lagrangian for that problem, and find the first order condition

$$0 = \frac{\partial}{\partial \pi(r)} \left[ \int_{s \in R} \pi(s) \cdot \ln(\pi(s)) \cdot ds - \lambda \left( \int_{s \in R} \pi(s) \cdot ds - 1 \right) - \sum_{\alpha \in O} \lambda_{\alpha} \left( \int_{s \in R} \xi_{\alpha}(s) \cdot \pi(s) \cdot ds - p_{\alpha} \right) \right]$$

and solving that FOC, we get

$$0 = 1 + \ln(\pi(r)) - \lambda - \sum_{\alpha \in O} \lambda_{\alpha} \xi_{\alpha}(r)$$

which after rearranging terms, setting  $U(r) = \sum_{\alpha \in O} \lambda_{\alpha} \cdot \pi_{\alpha}(r)$ , and  $Z^{-1} = e^{\lambda-1} = 1 / \left[ \int_{r \in R} e^{U(r)} \cdot dr \right]$ , we get

$$\pi(r) = Z^{-1} e^{-U(r)}$$

which ends the proof of part 1.

For the proof of part 2, we wish to rewrite

$$\xi_{\alpha}(r) = \rho(r, k_{\alpha})$$

as

$$\xi_{\alpha}(r) = \max_{a_{\alpha} \in \{0,1\}} \left[ \frac{a_{\alpha}}{1-\kappa} h_k(r, k_{\alpha}) \right]$$

where  $\kappa$  is chosen to equate the two expressions. The existence of such a value of  $\kappa$  is obvious since we have bounded  $\rho(r, k_{\alpha})$  between 0 and 1. Once  $\kappa$  is chosen to equate the two expressions, it is clear that by setting the atom strengths  $\sigma_{\alpha} = \lambda_{\alpha} / (1 - \kappa)$ , we get

$$\max_{r \supset i} U(r) = \max_{r \supset i} H_k(r)$$

and with the result of part 1, it follows that the completion with the maximum entropy density is equal to that with the Gibbs density defined by  $U(r)$ , which is equal to that with the Gibbs density defined by  $H_k(r)$ , and that finishes the proof of the theorem. ■

Now, given the result of theorem 1, we know that the plausible distributions arising from maximum entropy estimation (for a full study of the importance of this class of density estimators, see Christensen 1981, 1982, and Levine & Tribus, 1979) and harmony theory as described in the previous section coincide. The next step is to consider mechanisms that may converge to those estimates starting from some randomly chosen initial distribution of representations. The next theorem was designed to achieve that result; the first part has been in the literature for a long time - see for instance Metropolis et al (1953) - and the second part is mainly due to Geman and Geman (1984); in both cases, the results mentioned have been proven for the discrete state space, and we generalize them here to continuous state spaces.

We start by defining a “heat bath” process resulting from a sequence of absolutely continuous probability measures on the representation space which is characterized by the sequence of densities  $\{p_t\}_{t=0}^{\infty}$ . The process starts with an initial density  $f_0$ , and an initial representation vector  $r_0$  which is randomly drawn from  $f_0$ . At each step of the algorithm, we shall randomly choose one of the coordinates of the representation vector  $r$ . Let the set  $S = \{1, 2, \dots, n\}$  index the coordinates of the representation space, and let the sequence of visits to the different elements of  $S$  be denoted by  $\{n_1, n_2, \dots\}$ . Then, at each point in time, we randomly choose an element of  $S$ , and update the density  $f_{t+1}$  by the Chapman-Kolmogorov type equation

$$f_{t+1}(r_s = x_s; s \in S) = p_U(r_{n_{t+1}} = x_{n_{t+1}} | r_s = x_s; s \neq n_{t+1}) \cdot f_t(r_s = x_s; s \neq n_{t+1}) \quad (3.3)$$

where  $p_U(x)$  is again defined by (3.1), (3.2), and the conditional density in (3.3) above is

defined according to

$$p_U(r_s | r_l; l \neq s) = \frac{p_U(r)}{\int_{x_s \in [0,1]} p_U(r) \cdot dx_s} \quad (3.4)$$

where  $s \in S$ , and  $r \in R = [0, 1]^n$ . The time series of representations  $r_t; t = 1, 2, \dots$  is then generated by randomly drawing from  $f_t$  after each iteration of the type described above.

The first part of the next theorem tells us that for a heat bath process where  $p_U$ , which is defined according to equations (3.1), (3.2) for a fixed temperature  $T$ , is used to define the constant sequence  $\{p_t\}$  of our heat bath, and starting from any initial density, the sequence of densities  $f_t$  converges to the maximum entropy density  $\pi$  of the agent's problem. It is clear that the equivalence of  $p_U$  and  $\pi$  proved in part 1 of Theorem 1 plays a major role in achieving this result. The main step to prove part 1 of theorem 2 is to recognize the Harris ergodicity of the heat bath process described above, together with the stationarity of the density  $p_U$ , which therefore is the unique stationary density (see El-Gamal (1988), or Nummelin (1984) for definition of Harris ergodicity/ exactness and some interesting results). This result allows us to think of a variant of the representative agent. We know that asymptotically, as each individual agent  $s$  goes through their heat bath stochastic process for representations of the economy, the profile of representation vectors in the economy will be a random draw from  $p_U$ . The "representative" or "mean" agent may thus be thought of as the one with the mean representation vector according to the density  $p_U$ . If our main interest is say the optimal action for that average person, then under the assumption made so far that the agent's optimal action is a deterministic function of their representation vector, the mean optimal action can also be achieved by taking the time average in one individual's heat bath. The ergodicity result stated above, and used in proving part 1 of Theorem 2 below assures us that that average or mean representation vector in the economy is asymptotically equivalent to the time average of the representation vectors through the heat bath process for any one individual with any initial density. This result is viewed as a formalization of a variant on the idea of the representative agent which is based on



biologically feasible and easily implementable model. It also allows us to simulate models of that average individual's behavior based on his/her representation vector without having to simulate many replications from many different initial densities since the time series from one heat bath simulation can be used to cheaply compute the average that would result from such a panel exercise.

The second part of Theorem 2 deals with computing the mode of the density  $\pi = p_U = p_H$  which will give us the maximum likelihood completion function for the agent's problem. The idea here is to consider the heat bath as above, but let it be generated by the sequence  $\{p_{T_t}\}$ , where  $p_{T_t}$  is the Gibbs density defined by (3.1), (3.2) where the temperature at time  $t$  is taken to be  $T_t$ . The idea, which was copied from models of statistical physics and convergence to equilibrium in chemical processes, is that starting from a large temperature, the stochastic process  $r_t$  will be evolving almost completely stochastically which is what is required by the first part of the theorem for convergence to  $p_U$ . As time progresses, the temperature is slowly lowered until ultimately, as  $T_t \rightarrow 0$ , the process becomes more and more like steepest ascent towards the mode of the density. This process is known as simulated annealing due to the way it simulates the progress of a chemical process towards its equilibrium, and it has been suggested for numerical analysis applications of searching for the global maximum of a function (see, for example, Anily, and Federruen 1987) where greedy algorithms like steepest ascent can get trapped in a local maximum. As in the chemical processes, lowering the temperature too fast can "freeze" the system in an unwarranted local maximum; hence, we should make sure that  $T_t \rightarrow 0$  at a slow enough rate. The first known result which gives a rate for the convergence of  $T_t$  (referred to as the annealing schedule) is due to Geman and Geman (1984), and it is stated in part 2 of Theorem 2 below. The rate provided by the authors is very slow, and their simulation results show that a much faster rate of convergence achieves in practice, which suggests the possibility of improving on their bounds. At this stage, however, we just state their result as it appears in Geman and Geman (1984), and note that their proof extends directly to

the case with a continuous state space without any necessary changes. For the purposes of the theorem, let us define

$$R_0 = \{r \in R: U(r) = \max_{y \in R} U(y)\}$$

and let  $\pi_0$  be the uniform measure on  $R_0$ . Also define  $U^* = \max_{r \in R} U(r)$ ,  $U_* = \min_{r \in R} U(r)$ , and  $\Delta = U^* - U_*$ .

*Theorem 2:*

1. *The heat bath process determined by  $p_U$  converges for all initial densities  $f_0$  to  $\pi$ , the maximum entropy density of the agent's problem.*
2. *Consider the annealing process determined by  $\pi_{T_t}$  with  $T_t \rightarrow 0$  and  $T(t) \geq N \cdot \frac{\Delta}{\ln(t)}$  for all  $t \geq t_0 \geq 2$ . Then, for all initial densities  $f_0$ ,*

$$\lim_{t \uparrow \infty} f_t(r(t) = x | r(0) = y) = \pi_0(x)$$

○

*Proof of Theorem 2:*

It is clear that the stochastic process  $r_t$  is a Markov process. Now, for a fixed  $t$ , let  $r \in R = [0, 1]^n$ , and for any  $x \in [0, 1]$ ; let  $r^x$  be the representation vector such that  $r_{n_t}^x = x$ , and  $r_s^x = r_s$  for all  $s \neq n_t$ . Then, the transition kernel for the Markov process  $r_t$  is defined for all  $r' \in r^x$

$$K(r'|r) = P_{n_t} p_U(r'_{n_t} = x_{n_t} | r'_s = r_s; s \neq n_t)$$

where  $p_U(\cdot)$  is defined by (3.1), (3.2), and  $p_U(\cdot)$  is defined by (3.3), (3.4), and where  $P_{n_t}$  is the probability of choosing coordinate  $n_t$  to change at time  $t$  (this will typically be  $\frac{1}{n}$ ). But that kernel easily shows us that

$$p_U(r') \cdot K(r|r') = p_U(r) \cdot K(r'|r)$$

where  $r$  and  $r'$  differ in at most one component, and  $K$  is a stationary kernel. Hence, if we start with the density  $f_t(\cdot) = p_U(\cdot)$ , we get

$$\begin{aligned} f_{t+1}(r) &= \int_{r' \in r^x} K(r|r') \cdot p_U(r') \cdot dr' \\ &= \int_{r' \in r^x} K(r'|r) \cdot p_U(r) \cdot dr' = p_U(r) = f_t(r) \end{aligned}$$

and hence the Gibbs density  $p_U(\cdot)$  is a stationary density. Notice moreover, that the way we constructed the heat bath process, all coordinates will be visited infinitely often with probability 1, and by the positivity of  $p_U(\cdot)$  and hence of  $p_U(\cdot|\cdot)$ ,  $K(\cdot|\cdot)$  is irreducible, and  $p_U$  defines an irreducibility measure. Now, there exists an irreducibility measure of  $K(\cdot|\cdot)$  (by proposition 2.4, p.13, Nummelin 1984), and it is clear that by the positivity of  $p_U$  everywhere, the measure defined by  $p_U$  is a maximal irreducibility measure (i.e. all other irreducibility measures are absolutely continuous with respect to it). We also know that all sets of  $\int p_U$  measure non-zero will be visited infinitely often with probability 1. But that is equivalent to positive Harris recurrence (Nummelin 1984, Ch.3), and that implies the Harris ergodicity of the process  $r_t$ , and since we know that  $p_U$  is a stationary density for  $K(\cdot|\cdot)$ , it follows that  $p_U$  is the unique stationary density and the limiting density of the process. This concludes the proof of part 1. The proof of part 2 of theorem in Geman and Geman (1984) extends to the continuous state space without requiring any changes other than replacing all transition probabilities with transition kernel densities, and replacing all sums with integrals where applicable, and so on. ■

We therefore have the two major results that we set out to achieve. Theorem 1 gave us the relationship between the maximum entropy distribution and those based on the distributions arising from the intuitive Harmony theoretic approach, and between the completions of representations based on them. Then, the results of Theorem 2 assured us of the convergence of heat baths to the distributions in question, and the convergence of annealing schedules with “temperatures” falling slowly enough to the mode of that distribution. This gives us the full framework needed to write simple algorithms which converge to the

desired quantities that we use in models of economic behavior. The choice of  $\lambda$ 's for part 1 of Theorem 1 and of  $\sigma$ 's for part 2 have not been discussed, however. In the remainder of this section, we shall discuss the "learnability" result adapted from Hinton and Sejnowski (1983) by Smolensky (1986), which offers an algorithm for teaching a cognitive system (which in our case will be a program or machine that simulates the behavior of an economic agent) the proper values of  $\lambda$  and  $\sigma$  for Theorem 1. The algorithm suggested is Hinton and Sejnowski's (1983) "trace learning procedure" where we start with all the  $\lambda_\alpha$ 's = 0, and sample a representations  $r$  from the true distribution  $p$ , and store an increment  $i_\alpha =$  sample mean of  $\xi_\alpha(r)$  for each  $\lambda_\alpha$ . Also compute a decrement  $d_\alpha =$  sample mean of  $\xi_\alpha(s)$  where the  $s$ 's are sampled from the distribution  $p_U$  defined by the current  $\lambda$ 's. Then, we update each  $\lambda_\alpha = \lambda_\alpha + i_\alpha - d_\alpha$ . Repeat this algorithm ad infinitum incrementing the  $\lambda$ 's with the penalty function from the true distribution and decrementing it with the penalty function from the current distribution until the two distributions converge, and the increments and decrements start to offset (which by the law of large numbers are good approximations of the expected losses corresponding to the two distributions provided a large number of  $r$ 's and  $s$ 's are sampled at each iteration) and the  $\lambda$ 's converge. At each step in the iterations, set  $\sigma_\alpha = \frac{\lambda_\alpha}{1-\kappa}$ .

Theorem 3 in Smolensky (1986; proof in pp. 278-280) proves that the above algorithm converges to the  $\lambda$ 's and  $\sigma$ 's necessary for Theorem 1 above. His proof generalizes immediately to the continuous case, and hence we will only briefly discuss the main steps in it. The proof requires showing that the Lagrange multipliers  $\lambda$  in Theorem 1 minimize the function

$$F(\lambda) = \ln \int_{r \in R} e^{[\sum_{\alpha \in O} \lambda_\alpha (\xi_\alpha(r) - p_\alpha)]}$$

and then showing that  $F(\lambda)$  is a convex function. The next step is to show that the "trace learning algorithm" defined above is equivalent to gradient descent along  $F$  which by the convexity of  $F$  should clearly converge to the argmin of  $F(\lambda)$ . Hence  $\lambda$  and  $\sigma$  (which is a scalar multiple of  $\lambda$  in each iteration) converge to the values needed in Theorem 1. This

completes all the computational ingredients needed to provide a full algorithm for simulating economic agents going through the decision making problem be that the completion task (searching for the completion function corresponding to the mode achieved as the limit of the stochastic annealing algorithm) or the computation of expectations of variables of interest via time averages of their values in the heat bath.

## 5. Conclusion

In this paper, we have introduced the models of harmony theory and stochastic relaxation used in the cognitive sciences by Smolensky (1986) and Geman and Geman (1984) to the framework of rational economic agents making decisions under uncertainty. The form of uncertainty allowed is lack of data on important variables, and the inference process was estimation and maximization of the density of representation vectors. The competence results of Smolensky, and the relaxation and annealing results of Geman and Geman were shown to extend to continuous state spaces which are prevalent in economic modelling. Those results can then be used to discuss the representation vector of the average or representative agent by taking the time average of the representation vectors of one simulated individual through a heat bath. That result followed immediately from the Harris ergodicity of the heat bath process developed in section 4, and the well known result of Birkhoff's ergodic theorem. An alternative was also offered which is to use the mode of the density (which is viewed as the most likely values for the variables of interest in the economic problem) that naturally arises as the limit of the annealing schedule with temperatures converging to zero sufficiently slowly. Properties of decision rules based on those "real expectations" remain to be studied in future research, and it would be interesting to see if they avoid the well known experimental fallacies that "rational expectations" assumptions continually perpetrate.

More research also needs to be done to generalize the simulation programs of Boltzmann machines and harmony processes provided in McClelland and Rumelhart (1988) to the

continuous state space. Once such simulations are available, one has to start considering models with many individuals where the variables of ignorance differ among individuals. In particular, a subject of great interest would be to consider models with many individuals where one individual's action is one of the elements of the representation vector of another individual on which he does not have data. This cross inference and guessing procedure will then immediately become a dynamic game rich enough to offer a number of economic models for analysis. The multi-layer parallelism this framework (parallel processes in one individual brain, and the collection of individuals as nodes in a market which serves as a higher level parallel process) offers a number of challenges to researchers, but also offers the promise of the possibility of simulating very complex economic structures that are heretofore too difficult to analyze.

## References

Anily, S. and Dedergruen, A. "Simulated Annealing Methods with General Acceptance Probabilities", *Journal of Applied Probability*, vol. 24, 1987.

Christensen, R. *Belief and Behavior*, Entropy Ltd., Lincoln, MA, 1982.

Christensen, R. *Entropy Minimax Sourcebook, vols. I-III*, Entropy Ltd., Lincoln, MA 1981.

Dobrushin, R. "The Description of a Random Field by means of Conditional Probabilities and Conditions of its Regularity", *Theory Prob. Appl.*, vol.13, 1968.

Dunford, N. and Schwartz, J. *Linear Operators, Part I*, Interscience Publishers Inc., N.Y. 1957.

El-Gamal, M. "Simple Methods of Estimation and Inference in Systems Characterized by Deterministic Chaos", RCER working paper # 147, University of Rochester, 1988.

Geman, S. and Geman, D. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images", *IEEE Transactions of Pattern Analysis & Machine Intelligence*, vol. PAMI-6, no.6, November 1984.

Georgii, H. *Canonical Gibbs Measures*, Springer-Verlag, Berlin, 1979.

Hinton, G. and Sejnowski, T. "Analyzing Cooperative Computation", *Proceedings of the Fifth Annual Conference of the Cognitive Science Society*, 1983.

Hinton, G. and Sejnowski, T. "Learning and Relearning in Boltzmann Machines", in Rumelhart and McLelland, *Parallel Distributed Processing*, M.I.T. press, Camb. MA, 1986, vol.I.

Jeffreys, H. *Scientific Inference*, Cambridge University Press, Cambridge, 1973.

Jeffreys, H. *Theory of Probability*, Oxford University Press, 1967.

Justice, J. *Maximum Entropy and Bayesian Methods in Applied Statistics*, Cambridge Univ. Press, 1986.

Keynes, J. M. *A Treatise on Probability*, MacMillan and Co., London, 1921.

Levine, R. and Tribus, M. (eds.) *The Maximum Entropy Formalism*, M.I.T. press, Camb. MA, 1981.

Lucas, R. and Prescott, E. "Investment Under Uncertainty",

*Econometrica*, vol. 39, no. 5, 1971.

McClelland, J. and Rumelhart, D. *Explorations in Parallel Distributed Processing*, M.I.T. Press, Camb. MA, 1988.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E. "Equations of State Calculations By Fast Computing Machines", *Journal Chem. Phys.*, vol. 21, 1953.

Muth, J. "Rational Expectations and the Theory of Price Movement", *Econometrica*, vol. 24, July 1961.

Nerlove, M. *The Dynamics of Supply: Estimation of Farmers' Response to Price*, Johns Hopkins Press, Baltimore, 1958.

Nummelin, E. *General Irreducible Markov Chains and Non-Negative Operators*, Cambridge University Press, Cambridge, 1984.



*Pollard, D. Convergence of Stochastic Processes, Springer-Verlag, N.Y., 1984.*

*Preston, C. Gibbs States on Countable Sets, Cambridge University Press, Cambridge, 1974.*

*Rosenkrantz, R. (ed.) E. T. Jaynes: Papers of Probability, Physics, and Statistical Physics, D. Reidel Pub. Co., Dordrecht, Holland, 1983.*

*Rumelhart, D. and McClelland, J. Parallel Distributed Processing, M.I.T. Press, Camb. MA, 1986, vols. I & II.*

*Shannon, C. and Weave, W. The Mathematical Theory of Communications, University of Illinois Press, Urbana, 1962.*

*Shiryayev, A. Probability, Springer-Verlag, N.Y., 1984.*

*Smolensky, P. "Information Processing in Dynamical Systems: Foundations of Harmony Theory", in Rumelhart and McLelland, Parallel Distributed Processing, M.I.T. press, Camb. MA, 1986, vol.I.*

*Treves, F. Topological Vector Spaces, Distributions, and Kernels, Academic Press, N.Y., 1967.*