Comment

S. Amari

First of all, I would like to thank the Editor for giving me an opportunity to present my personal view on this interesting paper connecting the interdisciplinary field of neural networks and statistics. I also congratulate the authors for their excellent job of reviewing this difficult field in a very compact and comprehensive way.

The brain is an enormously complex system in which distributed information is processed in parallel by mutual dynamical interactions of neurons. It is still difficult, and challenging, to understand the mechanisms of the brain. Recently, the importance and effectiveness of brain-style computation has been widely recognized by the name of neural networks. Roughly speaking, there are three different research areas concerning neural networks. One is the experimental area based on physiology and molecular-biology, which is progressing rapidly and steadily. The second area is engineering applications of neural networks inspired by the brain-style computation where information is distributed as analog pattern signals, parallel computations are dominant and learning guarantees flexibility and robustness of computation. This area has opened new practical methods of pattern recognition, control systems, time-series analysis, optimization, memories, etc. The third area is concerned with theoretical (or mathematical) foundations of neurocomputing, which search for the fundamental principles of parallel distributed information systems with learning capabilities. From this standpoint, the actual brain is a biological realization of these principles through a long history of evolution.

Statistics has a close relation with the second applications area of neural networks, as the present authors have so clearly shown (also see Ripley, 1993a). Statistical methodology is indeed a very important tool for analyzing neural networks. On the other hand, neural networks provides statistics with tractable multivariate nonlinear models to be studied further. It also inspires statistical sciences with the notions of learning, self-organization, dynamics, field theory, etc. which statistics has so far paid little attention to. On the other hand, statistical sciences provides one of the crucial methods for constructing theoretical foundations of neurocomputing (e.g., Amari, 1990, 1993a). Without these foundations, it is difficult for neural network technology to take off from the present rather "easy and shallow" technology to a more fundamental one.

Artificial neural networks research has experienced ups and downs; up in the early sixties where the perceptron and the adaline were proposed and again a big up in the middle of the eighties until now. It is said that the dark period was around the seventies where little attention had been paid to ANN and that the Minsky-Papert critique gave rise to this down. However, I believe this prevailing story is merely a myth. We can point out the lack of supporting technology as the background of this fall. Computer technology had developed greatly through the sixties and seventies. Researchers on pattern recognition and artificial intelligence thought that it was easier and more powerful to use symbol processing in modern computers rather than to use neural networks technology. This was true, and information processing technology including artificial intelligence had been constructed successfully upon modern computers. However, hardware technology had further developed in the eighties such that it could support neural parallel computation. It was not a dream to construct neurochips or even neurocomputers. There are, of course, many other intellectual reasons to support the resuscitation in the eighties.

In the seventies, most researchers did not think that engineering applications of neural networks were realizable. The background technology was not yet matured at the time. However, it was not a dark period in theoretical study because many of the ideas were proposed in the "dark period" that were rediscovered or developed further to be the fundamental methods supporting the neural network methods today.

For example, the generalized delta rule for a multilayer perceptron was proposed in 1967 (Amari, 1967) where analog neurons were used and the stochastic descent algorithm was applied. The idea was also introduced in a Russian book (Tsypkin, 1973). I believe that there were not a few researchers who knew the idea at that time. It was

S. Amari is Professor, Department of Mathematical Engineering, University of Tokyo, Bunkyo-ku, Tokyo 113, Japan.

the great achievement of Rumelhart, Hinton and Williams (1986b) who not only rediscovered the old idea but have shown its effectiveness in practical problems.

The idea of associative memory of the Hopfield type was intensively studied in 1972 by Kohonen (1972), Nakano (1972) and Anderson (1972). Amari (1972) studied its dynamical characteristics, including both the symmetric connections case where memorized patterns are fixed points of dynamics and the asymmetric connections case where sequences of patterns are memorized and recalled. Hopfield introduced the new notion of the "energy" or Lyapunov function to analyze the associative memory model and opened the new approach of spin-glass analogy to this field. A lot of fundamental studies appeared by using the statistical-physical (spin-glass) method, although the statistical-mechanical theory of neural networks itself appeared in the seventies (Little, 1974; Amari, Yoshida and Kanatani, 1977) the latter of which treated more general non-equilibrium dynamics.

A fundamental idea of self-organizing neural networks was proposed by Von der Malsburg (1973). It was applied to the formation of neural topological maps (Willshaw and Von der Malsburg, 1976). The dynamical instability of such neural field dynamics was studied (Takeuchi and Amari, 1979), which guarantees the formation of patch structure and columns existing in the brain. Based on these works, Kohonen (1982) proposed an excellent idea of learning vector quantization (LVQ) and neural topological maps which are much more simple and efficient compared with the previous models. The possibility of neural principal component analyzer was also pointed out in the seventies (Amari, 1977). Grossberg's adaptive resonance theory (ATR) was proposed in 1976 (Grossberg, 1976).

The achievements in the seventies should not be too exaggerated. Not only old ideas were developed to be applied to practical problems, but a lot of new ideas emerged in the eighties. I would like to emphasize that we need much more fundamental new ideas and mathematical foundations in order to elucidate principles of neurocomputing. Statistical and probabilistic methods are very important for this purpose. The current applications have proved the usefulness of neurocomputing but are still superficial even though they have provided a strong impact on various fields of science and technology with novel nonlinear modeling.

Here, I would like to point out two more interesting topics related to statistics. One is the learning curve that shows how fast a learning machine can improve its behavior as the number of training examples increases. This problem is closely related to the asymptotic theory of statistical inference, but the behavior of a network is measured by the generalization error, not by the squared error of estimated parameters. The estimate of the generalization error can be applied to the model selection problem in which the statistical methods such as Akaike information criterion (AIC) and minimum description length (MDL) are useful. There are a number of approaches to this problem, for example, the computational learning theory approach (Baum and Hausler, 1989), statistical-mechanical approach (Levin, Tishby and Solla, 1990), information-theoretic one and statistical approach (Amari and Murata, 1993; Amari, 1993b). When a network behaves stochastically, the statistical asymptotic theory can easily be applied to this problem. However, when the underlying model is deterministic (or the 0 temperature case in physicists' terminology), the underlying model becomes nonregular in the sense that the Fisher information becomes infinitely large. Therefore, the regular statistical theory cannot be applied. However, we can still construct a universal theory (Amari, 1993b). This is one interesting fact about neural networks.

Another interesting topic concerns the expectation and maximization (EM) algorithm and information geometry. The EM algorithm is the technique of estimation when only partial data are observed. When a neural network includes hidden neurons, only input and output signals are observable as learning data and desired signals on the hidden neurons should be generated or estimated by some means. The EM algorithm is used in learning of hidden units of the Boltzmann machine (Amari, Kurato and Nagoako, 1992; Byrne, 1992). It is interesting that the procedures of the EM algorithm correspond to the e-geodestic projection and m-geodesic projection in the manifold of probability distributions, in the sense of differential geometry of statistical inference (Amari, 1985).

Recently, Jordan and Jacobs (1993) proposed a model called the mixture of expert networks in which one of the component networks is responsible for its own specific tasks. This enables parallel and distributed sharing of tasks. The missing or hidden data is which task should be processed by which network. This model is represented by a mixture of exponential families, and the EM algorithm as well as information geometry plays an essential role in such models.