Bulletin of the Seismological Society of America, Vol. 80, No. 5, pp. 1346–1373, October 1990

SEISMIC DISCRIMINATION WITH ARTIFICIAL NEURAL NETWORKS: PRELIMINARY RESULTS WITH REGIONAL SPECTRAL DATA

By Farid U. Dowla, Steven R. Taylor, and Russell W. Anderson

Abstract

An application of artificial neural networks (ANN) for discrimination between natural earthquakes and underground nuclear explosions has been studied using distance corrected spectral data of regional seismic phases. *Pn*, *Pg*, and *Lg* spectra have been analyzed from 83 western U.S. earthquakes and 87 Nevada Test Site explosions recorded at the four broadband seismic stations operated by Lawrence Livermore National Laboratory. Distance corrections are applied to the raw spectra using existing frequency-dependent *Q* models for the Basin and Range. The spectra are sampled logarithmically at 41 points between 0.1 and 10 Hz for each phase and checked for adequate signal-to-noise ratios ($S/N > 2$). The ANN was implemented on a SUN 4/110 workstation using a backpropagation-feedforward architecture. We find that, using even simple ANN architectures (82 input units, 1 hidden unit, and 2 output units), powerful discrimination systems can be designed. In order to regionalize the data characteristics, a separate neural network was assigned to each station. For this data set, the rate of correct recognition for untrained data is over 93 per cent for both earthquakes and explosions at any single station. Using a majority voting scheme with a network of four stations, the rate of correct recognition is over 97 per cent. Although the performance of the ANN is similar to that of the Fisher linear discriminant, the ANN exhibits a number of computational advantages over the conventional method. Finally, examination of the network weights suggests that, in addition to spectral shape, a criterion that the ANN utilized to discriminate between the two populations was the *Lg*/*Pg* spectral amplitude ratios.

Introduction

Discrimination of seismic records from natural earthquakes and underground nuclear explosions is an important problem in test ban treaty verification research (Dahlman and Israelson, 1977). Recent developments indicate that Artificial Neural Networks (ANNs) might be appropriate for solving difficult problems in signal discrimination and classification (Lippman, 1987). ANNs are computational systems consisting of a large number of simple processing units, or neurons, which are interconnected in a parallel structure. The parallel computational architecture of an ANN, besides resembling the biological brain, has potential in application areas like seismic discrimination where multiple hypotheses are pursued in parallel, where the number of input parameters is often large, and where well-defined solutions are not available. An ANN *learns* to solve a problem by training on examples of real data. For example, with ANNs it is not necessary to explicitly specify classification rules or algorithms. We studied ANNs in the context of a seismic discrimination problem using regional spectral data. The methodology and the results of this study are described in this paper.

The problem of distinguishing underground nuclear explosions from natural earthquakes using seismic data has been studied for a long time. Currently discrimination of regional data is an important research topic and a variety of regional discriminants have been proposed by many researchers (cf. Pomeroy *et al.*, 1982; Taylor *et al.*, 1989). Discrimination of small magnitude events, however, is still a

difficult problem. For small magnitude events ($m_b < 4$), spectral discrimination using multiple regional phases has recently received much attention (Bennett and Murphy, 1986; Taylor *et al.*, 1988). It is generally believed that both spectral shapes and ratios of the regional phases (*Pn*, *Pg*, and *Lg*) might be quite useful for distinguishing earthquakes from explosions (Pomeroy *et al.*, 1982). Generalization and regionalization of these discriminants is, however, important for optimum performance.

Since ANNs can classify populations by generating complex discriminant functions by training on real data, we used as input to the ANN the full broadband distance-corrected spectra of the regional seismic phases. During the learning phase, the ANN automatically extracted and learned the relationships among the discrete frequency components of the multiple regional phases for correct discrimination between earthquakes and explosions. The ANN was developed and tested with a large number of real seismic events, consisting of 83 earthquakes and 87 underground nuclear explosions recorded at each station of a network of four stations located in the western United States. Results of this study based on regional spectral data indicate that ANNs can indeed generate excellent discriminant functions. The rate of correct recognition for untrained data is over 93 per cent at any single station and is over 97 per cent for a network of four stations.

Our primary goal at the outset of this preliminary study was to gain an understanding of the performance of neural networks for seismic event discrimination with a set of real seismic data. As the work progressed, we realized that the engineering aspects of neural networks are still at an elementary stage and a number of the important issues in ANNs are unresolved. In this report we discuss some of the expertise in seismic ANNs that was developed during the course of this study. In particular, we discuss the important problems of the representation, preprocessing, normalization, and training of ANNs with a database of real seismic signals.

We begin with an introductory discussion of ANNs and then discuss the problems of pre-processing and data representation. This is followed by a discussion of the seismic spectral data for discrimination and the performance of the ANN for discrimination between earthquakes and explosions. We then apply the same data to the conventional Fisher discriminant (Tjøstheim, 1981), a linear method which utilizes covariance matrix information, and compare its performance with that of the ANN. Finally, we conclude with a discussion of the implications of our results and areas of future research in seismic neural networks.

Artificial Neural Net for Discrimination

The current interest in ANNs is an attempt at building a new class of powerful computers capable of solving cognitive tasks in recognition, discrimination, combinatorial optimization, and others. While these tasks are routinely performed by the human brain, they are still beyond the reach of conventional methods of computation. Part of the problem with conventional approaches is that the computational architecture might be inadequate (Rumelhart *et al.*, 1986). Conventional computers simply do not have the power of a massive interconnected network of nonlinear processing nodes, or neurons, which might be necessary for cognitive tasks (Hopfield, 1982). From this viewpoint, researchers are considering information processing techniques and devices, such as neural networks, which approximately resemble the brain.

In this study we have used a type of ANN called the *Multi-Layered Perceptron* (MLP) (also called the *backpropagation network*). The MLP has proven to be most useful in engineering applications (see, for example, DARPA Neural Network Study, 1988). However, because ANNs are quite new to the seismological community, we provide a tutorial on MLPs in this section. Because there are many different types of ANNs, an adequate discussion of all these networks is beyond the scope of this paper. For a more complete treatment of ANNs, interested readers are referred to the popular journal article by Lippman (1987) and the book by Rumelhart *et al.* (1986).

The Basic Model for Discrimination: Discriminant Functions

In order to motivate the structure of a MLP, we begin by considering a general two-category discrimination problem: suppose we need to classify a given real vector $\mathbf{x} = [x_1, x_2, \ldots, x_N]$, as belonging to either *class A* or *class B*. The discrimination problem is then to map any given point in R^N into one of two classes according to some desired criterion. As an example, consider the mapping shown in Figure 1a, where any point in R^2 plane is mapped into class A or class B according to its membership. How can we construct discriminant function, $D(\mathbf{x})$, that, given $\mathbf{x}$, $D(\mathbf{x})$ would correctly classify $\mathbf{x}$ as either belonging to class A or to class B?

If we construct two functions $D_A(\mathbf{x})$ and $D_B(\mathbf{x})$ with properties

$$D_A(\mathbf{x}) > D_B(\mathbf{x}) \quad \text{if } \mathbf{x} \text{ belongs to class A}$$

$$D_B(\mathbf{x}) > D_A(\mathbf{x}) \quad \text{if } \mathbf{x} \text{ belongs to class B,}$$

then, as shown in Figure 1b, we have a basic model for a discrimination system. Given $\mathbf{x}$ we simply need to compute the output of $D(\mathbf{x}) = D_A(\mathbf{x}) - D_B(\mathbf{x})$. The following discrimination rule is sufficient: if $D(\mathbf{x})$ is positive then $\mathbf{x}$ belongs to class A, else $\mathbf{x}$ belongs to class B. For example, it is useful to view $D_A(\cdot)$ and $D_B(\cdot)$ as elemental functions with maximum values of 1 when the input vectors are from class A and class B, respectively. From a conventional detection theory viewpoint, the discrimination function is an interconnection of *matched filters* and a threshold decision system.

Using the basic model of Figure 1b, we see that the essential problem for a discrimination system is the specification of the discrimination function $D(\cdot)$ using an interconnection of elemental functions, $D_A(\cdot)$ and $D_B(\cdot)$. The discrimination function constitutes a mapping from the input data to the output decision space. An example of a form of linear discriminant function is $D(\mathbf{x}) = \mathbf{w}_A{}^T\mathbf{x} - \mathbf{w}_B{}^T\mathbf{x}$ where $\mathbf{w}_A$ and $\mathbf{w}_B$ are weight vectors of length two. The major limitation of such a linear discrimination system is that it can only discriminate between classes of objects that are *linearly separable*; i.e., only when the two classes can be separated by a hyperplane in R^N (Rumelhart *et al.*, 1986). On the other hand, the MLPs with multiple nonlinear discriminant functions cascaded together can generate arbitrarily complex discrimination functions (Lippman, 1987). In other words, because the architecture of multi-layered perceptrons allows construction of discriminant functions for arbitrary mapping from the input data to the output decision space, these systems are quite powerful in pattern recognition and discrimination problems.

Terminology

The elemental functions of an ANN are the *artificial neurons* (also called units, processing elements, or nodes). The structure of a neuron is shown in Figure 2.

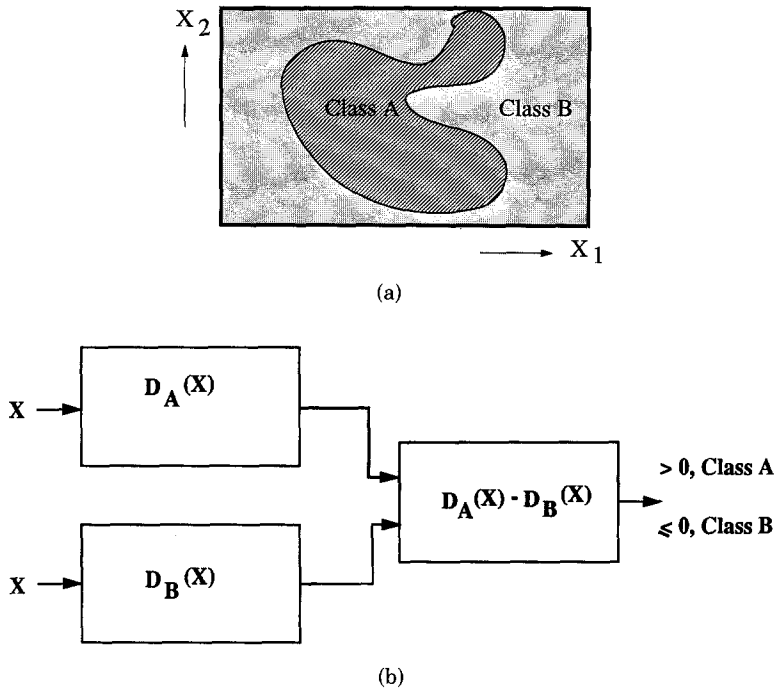


FIG. 1. (a) Example of a two-category classification problem where an input vector $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ is to be classified according to its membership in class A or class B as denoted by the shaded regions. (b) A classification system can be constructed by an interconnection of elementary discriminant functions. If $D_A(\mathbf{x})$ and $D_B(\mathbf{x})$ have maximum values of 1 when $\mathbf{x}$ belongs to class A and class B, respectively, then a positive output indicates $\mathbf{x}$ belongs to class A, otherwise $\mathbf{x}$ belongs to class B.

We represent the input pattern to the neuron as a vector of N elements, $\mathbf{x} = (x_1, x_2, \ldots, x_N)$. The vector $\mathbf{x}$ could be different measurements from a given seismogram or spectrum. The neuron weights and sums the input elements to form an intermediate scalar s given by

$$s = \sum_{n=1}^{N} w_n \cdot x_n = \mathbf{w}^T\mathbf{x}, \tag{1}$$

where $\mathbf{w} = (w_1, w_2, \cdots, w_N)$, denotes the weight vector of that neuron. The result of the vector dot product, s, is passed through a nonlinear transfer function (also called the *squashing function*), $f(\cdot)$, to obtain the output, $y = f(s)$, of the neuron for the input vector $\mathbf{x}$. The nonlinear transfer function is usually the sigmoid function defined by

$$f(s) = \frac{1}{1 + e^{-s}}. \tag{2}$$

A one-to-one correspondence between the artificial neuron and the biological neuron is commonly drawn: the inputs correspond to the dendrites of a biological neuron; the weights to the synapses; the summation and the transfer function unit corresponds to the cell body; and the output of the transfer function fanning out to other units to axons. Further, the firing rate of a biological neuron as a function of the input is well-approximated by the nonlinear sigmoid function (Hopfield, 1982).

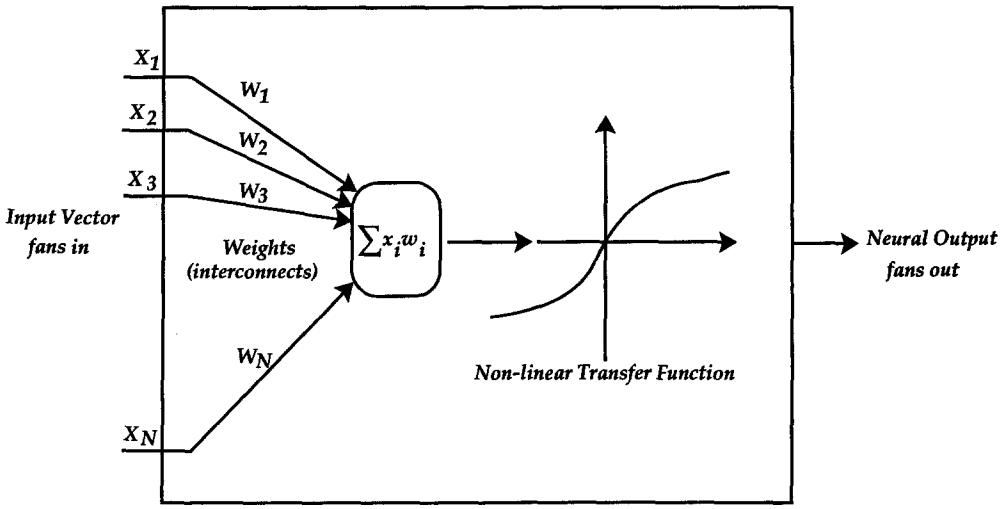


FIG. 2. The architecture of a single neuron consisting of N input units and 1 output unit.

The structure of perceptron-like ANNs is quite general and imposes no restriction on the number of inputs and outputs. For example, Figure 3 shows an example of a *multi-layered perceptron* consisting of four neurons in the input layer, the layer that receives the input directly, four neurons in the hidden layer, the layer that is not directly connected to either the inputs or the outputs, and two neurons in the output layer, the layer directly connected to the output of the system. Note that in these networks, the output of one layer constitutes the input to the next layer. Finally, because there are no feedback or lateral connections in the network, the network is also called a *multi-layered feedforward network.*

The knowledge in an ANN is encoded in the hundreds of interconnecting weights. For example consider a *trained network*, an ANN which has learned to solve a problem correctly for all the elements in the training set; we discuss learning in ANNs later. In the operational use of the trained ANN all that occurs is the *forward propagation* of the input vector from the input layer to the output layer. For the sake of clarity, let us consider the behavior of the network shown in Figure 3 when it is presented with an input pattern at the input layer. The first layer, the input layer, accepts the individual components of the input vector **x** and distributes them, without modification, to all the units of the next layer. Each of the units in hidden layer then computes a weighted sum of the received inputs and performs a nonlinear squashing operation and then fans out its output to each of the two processing elements of the output layer. The output layer, comprised of two processing units, produces the network estimate of the output vector (y_1, y_2). The exact architecture of an ANN is highly problem dependent and the number of hidden units is usually based on some ad hoc rules of thumb, some of which are discussed later in this paper.

The Neural Network Approach

The approach to solving a discrimination problem with a MLP can be summarized by the following steps:

- Select an architecture suitable for the problem. The number of units in the input and output layers is usually determined by the structure of the problem.

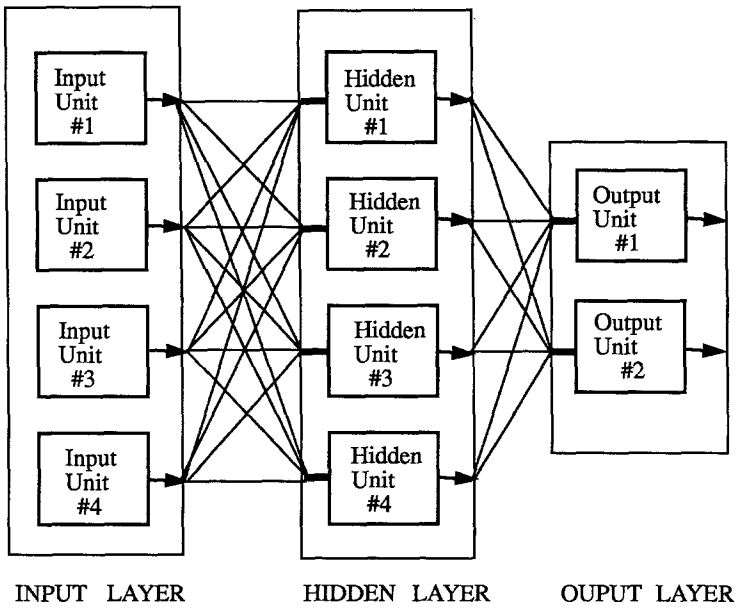


FIG. 3. Structure of a four-unit input layer, one four-unit hidden layer, and a two-unit output layer multi-layered perceptron. Note that the total number of weights or interconnects in this network is 24.

However, the number of hidden layers and the number of units in a hidden layer is a parameter decided by the designer. In network learning, generalization is increased and memorization is reduced by limiting the number of weights, or interconnects; as a general rule, the total number of weights in the network should be less than the number of elements in the training set.

- Collect and pre-process the data to form a database consisting of a large number of training events, called the training set. For example, a training set of P events might be denoted by $[(I_1, T_1), (I_2, T_2), \ldots, (I_P, T_P)]$ where I_k represents the input vector, and T_k represents the target output (or, the desired output). One important requirement for MLPs is the need for a large number of training elements in the training set. It is important that the training database spans the input-output space. In fact, one of the concerns in ANNs is what constitutes an adequate training set.
- Use a learning algorithm on the network such that the network learns to discriminate all or most of the elements in the training set. The most common way of training a multi-layered feedforward network is to use the backpropagation algorithm. This algorithm uses a gradient descent method to systematically modify all the weights in the network such that the network is able to discriminate correctly all the elements in the training set.
- Use the trained network to discriminate unknown input vectors. This is the forward propagation step that we described earlier.

Learning by Gradient Descent Optimization

In MLPs, the concept of *learning* is synonymous with the concept of *error minimization*. When a machine *learns* a task, it is able to perform the task by minimizing an *error* defined in terms of that task. Just as inversion problems minimize a least-squares error, the backpropagation learning algorithm also uses a

least-squares error minimization criterion. As the knowledge of the network is stored in the network weights, the learning algorithm utilizes the training set to adapt these weights.

Before we discuss the details of the learning algorithm, we note a historical point: although variations of MLPs have been known for about 30 years or so, it was not used extensively primarily because there was no well-known efficient learning algorithm. Recently, the backpropagation algorithm has changed all that. The widespread use of ANNs in applications today is partly due to the popularization of the backpropagation algorithm (cf. Rumelhart *et al.*, 1986; Dowla *et al.*, 1988). Although convergence issues of backpropagation are still a research topic, the algorithm has proven to be quite successful in many practical problems.

The backpropagation algorithm can be viewed as an elegant solution to a problem in unconstrained nonlinear optimization where an objective function $\mathrm{E}(\mathbf{w})$ is to be minimized with respect to the independent variable $\mathbf{w}$, the vector representing the network weights. The normalized sum of output errors over the entire training set is the objective function. Note that in order to apply the gradient descent method, we need to compute $\mathrm{E}(\mathbf{w})$ and its gradient vector, $\nabla\mathrm{E}(\mathbf{w}) = [\partial E/\partial w_1, \partial E/\partial w_2, \ldots, \partial E/\partial w_L]$, where L represents the total number of weights in the network. Formally, we can define

$$\mathrm{E}(\mathbf{w}) = \frac{1}{P}\sum_{p=0}^{P-1} E_p, \tag{3}$$

where P is the number of patterns in the training set, and E_p is the output error for the pth training pattern. E_p is defined as

$$E_p(\mathbf{w}) = \frac{1}{2}\sum_{j=0}^{N-1} (T_{pj} - O_{pj}(\mathbf{w}))^2 \tag{4}$$

where $O_{pj}(\mathbf{w})$ and T_{pj} are the network and the desired (or target) outputs, respectively, of the jth output neuron for the pth pattern, respectively. We have denoted O_{pj} as a function of $\mathbf{w}$ to emphasize the dependency of the network outputs on the weight vector, $\mathbf{w}$.

The error reduction rule in the backpropagation algorithm is basically described by

$$\mathbf{w} \leftarrow \mathbf{w} - \mu\nabla\mathrm{E}(\mathbf{w}). \tag{5}$$

In words, the rule modifies the present weight by an amount which is proportional to the negative derivative of the error with respect to that weight (see Fig. 4). The constant term μ, called the *learning rate*, is a small number which determines the rate at which the weight should be modified after each iteration.

The major contribution of the backpropagation algorithm is that it provides an efficient rule for determining the weights of the hidden units. We derive the mathematical details of backpropagation in the Appendix and summarize the mechanics of the backpropagation learning in the following.

Step 1. Initialize the weights in the network. Set all weights in the network to some small positive and negative random values.

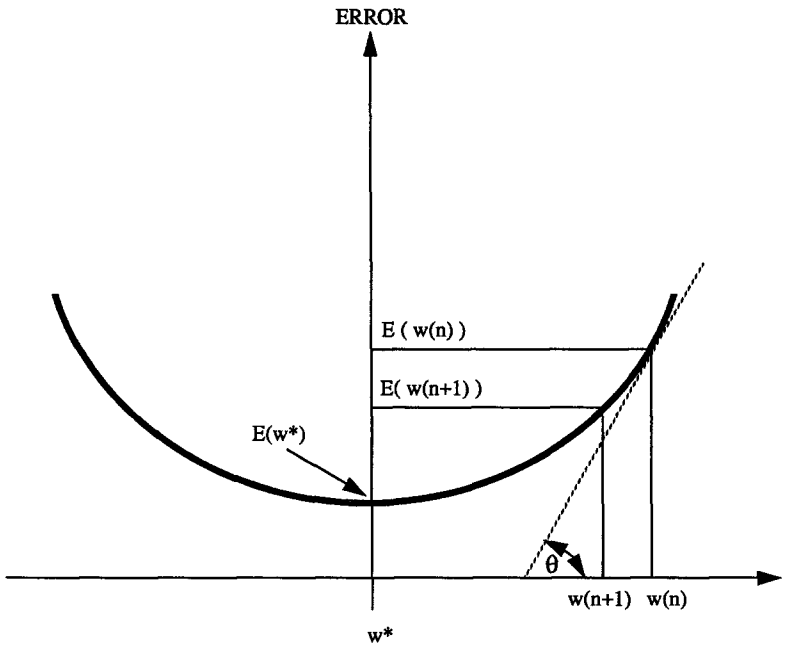


FIG. 4. The backpropagation learning algorithm is a gradient descent method whose principle can be illustrated using a simple one-dimensional error versus weight plot. Network error on the training set is minimized by adapting the weights. The goal is to determine a set of weights, w^*, for which the error, $E(w^*)$, is minimized. Using the present estimate of the weight, $w(n)$, the weight is modified by the rule $w(n+1) \leftarrow w(n) - \mu[\partial E/\partial w(n)]$, where $\partial E/\partial w(n) = \tan(\theta)$ and μ is a small number representing the learning rate. The learning rate determines how fast the weights are modified using the gradient of the error with respect to the present weight.

Step 2. Present input and the desired outputs. Present a new input: $(x_{p1}, x_{p2}, \ldots, x_{pN})$ and the desired outputs $(T_{p1}, T_{p2}, \ldots, T_{pM})$, for the pth element in the training set. The desired output with one output neuron, for example, might be 1 or 0, corresponding to the class of the input.

Step 3. Compute the actual outputs of the network. This is accomplished simply by forward propagating the input vector through the network that we described earlier.

Step 4. Adapt the weights in the network. Compute the error for the network output and update the weights appropriately. Starting from the output units of the network and working backwards layer by layer:

$$w_k \leftarrow w_k - \mu x \delta_k \tag{6}$$

where $\delta_k = f'(s)(\partial E/\partial x)$, and x is the input to the weight w_k. For the output units,

$$\delta_k = O_{pk}(1 - O_{pk})[(T_{pk} - O_{pk})]. \tag{7}$$

For a unit that is not in the output layer, the δ term of the above equation can be derived in terms of the δ's of those units whose inputs are the outputs of the unit. For the hidden units, as derived in the Appendix,

$$\delta_k = f(s)(1 - f(s)) \frac{\partial E}{\partial w_k} = f(s)(1 - f(s)) \sum_j \delta_j \cdot w_j, \tag{8}$$

where $f(s)$ is the output of the hidden unit k, s is the weighted sum of the inputs of the unit, and j is over all units receiving input from the unit and scaling them with weights w_j's.

Step 5. Repeat the iteration by going back to Step 2. The learning algorithm consists of presenting to the network all the elements of the training set and then adapting the weights according to the rule described by equation (6). This process is repeated until the error at the output of the network is sufficiently small.

In summary, each iteration of the backpropagation algorithm consists of two stages. In the forward propagation stage, the input is propagated from the input layer to the output layer. In the backward propagation stage, the error is propagated from the output to the input layer. In particular, during backward propagation δ's are computed layer by layer; δ's of the output layer are computed first and those of the input layer are computed last.

Although in the seismic discrimination problem we used a MLP, the architecture of the network which we eventually used had the structure of a very simple network with only one hidden unit and two output units. We found that, for this data set, increasing the number of layers or hidden units did not improve the performance significantly. In order to gain insight into the network, we chose to keep the architecture to a minimum level of complexity.

To gain further insight into the learning method consider an explosion spectrum presented at the input of the ANN shown in Figure 5. In the terminology of ANN, this network has just one hidden neuron with N units at the input layer and two units at the output layer. Input to the network are spectral values, i.e., in this problem the network input layer units corresponds to spectral frequency components of the three phases. The input layer is thus represented by a vector of $(N = 3 \cdot M)$ elements,

$$\mathbf{x} = [S_{Lg}(1), \ldots, S_{Lg}(M), S_{Pg}(1), \ldots, S_{Pg}(M), S_{Pn}(1), \ldots, S_{Pn}(M)], \quad (9)$$

corresponding to the $(M = 41)$ frequency components for each of the three phases, sampled logarithmically from 0.1 to 10 Hz. The mechanics of how we computed the spectra is described in the next section.

In order to teach the network to discriminate between earthquakes and explosions, a training set consisting of a large number of spectra for both earthquakes and explosions was applied to the network. The backpropagation algorithm was then used on the network where the objective was to adjust and determine the weights w_i such that for each member of the training set the network classified the spectra according to its appropriate category: an output of 1 for explosions, and an output 0 for earthquakes. When the network was tested, for example, an activation level at the output of around 0.5 means that the network is unable to decide on the event type.

Seismic Spectral Discriminants

Characteristics of the Spectral Data

Examples of regional seismic *phases* from earthquakes or underground nuclear explosions that are recorded by the Lawrence Livermore National Laboratory (LLNL) seismic observatory stations are illustrated in Figure 6. For the Basin and Range region, the regional phases *Pn*, *Pg*, and *Lg* are the principal arrivals corresponding to distinct modes and paths of seismic wave propagation. The first arrival

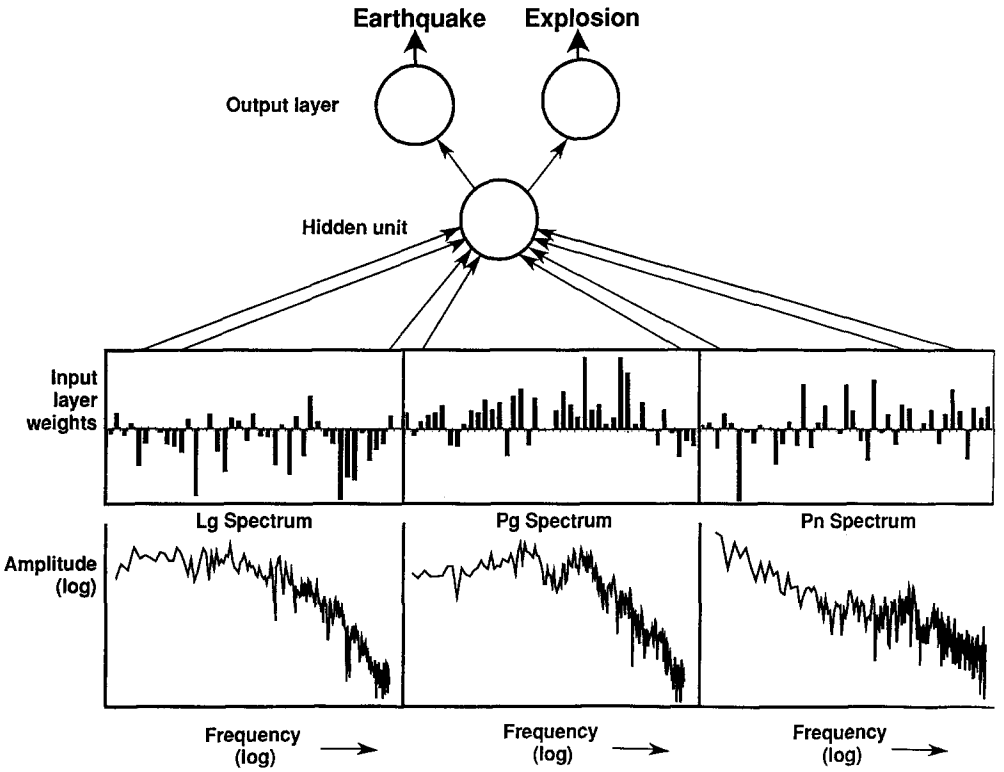


FIG. 5. The architecture of a neural net which was used to discriminate between explosion and earthquake spectra. The ANN had 123 input units (corresponding to the 41 spectral values of the three phases), 1 hidden unit, and 2 output units. The backpropagation training algorithm was used to teach the network to discriminate between the explosion and earthquake spectra at the input.

in our data is usually *Pn*, a body wave with longitudinal particle motion. *Pn* is usually followed by *Pg*, also a longitudinal wave with frequencies slightly higher than *Pn*. However, for the instruments we used the range of frequencies for both *Pn* and *Pg* were from 0.1 to 10 Hz. In this geologic region, *Pg* is observed to have higher amplitudes than *Pn*. The final principal arrival is *Lg*, a regional phase whose properties are not fully understood in spite of the fact that the *Lg* usually has the largest amplitude of the three phases. *Lg* has frequencies approximately in the same range as that of *Pn* and *Pg*. It is generally believed that the *Lg* phase is a superposition of higher modes of Love and Rayleigh waves propagating in a crustal wave guide. Typical amplitude spectra of the principal regional phases are illustrated in Figure 7.

In our notation we represent the amplitude spectrum of a phase at frequency f_k by $S_{\text{phase}}(f_k)$. For example, for the *Lg* phase, the amplitude spectrum will be represented by $S_{Lg}(f_k)$, the magnitude of the discrete-time Fourier transform of $s_{Lg}(n)$:

$$S_{Lg}(f_k) = \left| \sum_{n=0}^{N-1} w(n) s_{Lg}(n) e^{-j2\pi f_k n \Delta t} \right|, \tag{10}$$

where $s_{Lg}(n)$ represents the discrete-time windowed sequence corresponding to the

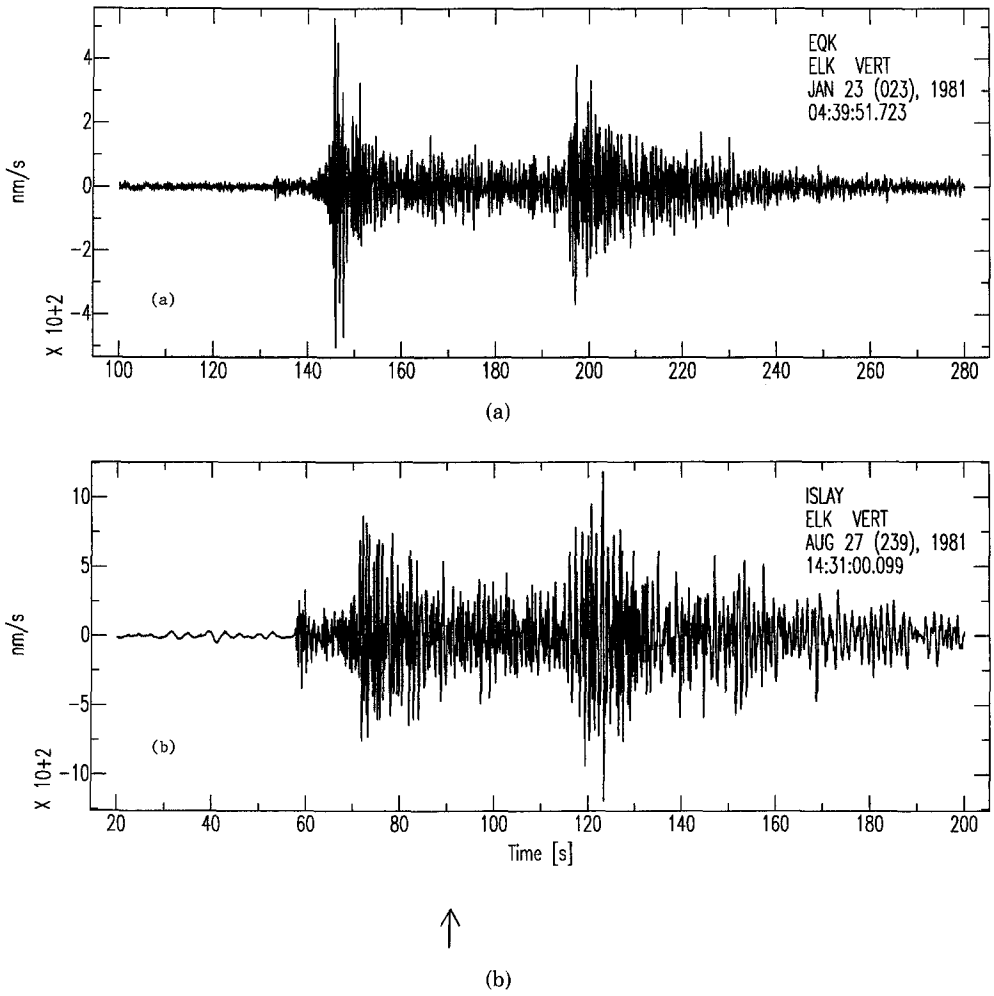


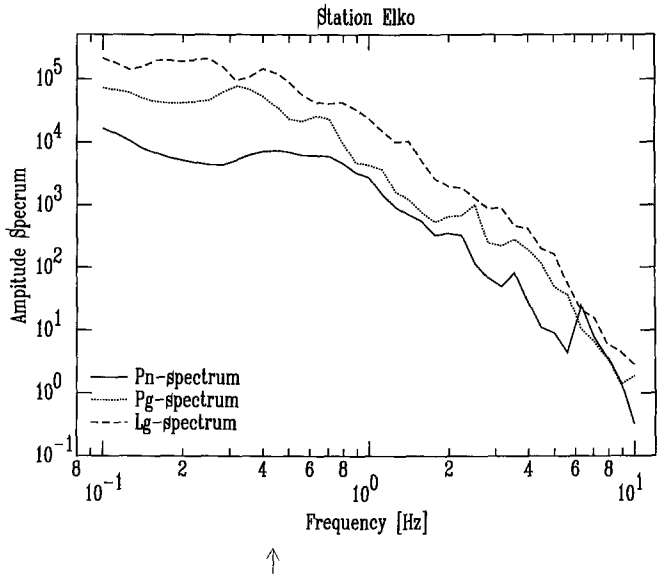
FIG. 6. Examples of seismograms for (a) an earthquake and (b) an underground nuclear explosion recorded at Elko, one of four stations of a seismic network operated by LLNL.

phase *Lg*, $w(n)$ represents a cosine tapering window, and Δt represents the sampling interval. For most of the data in this study, Δt is 0.025 sec.

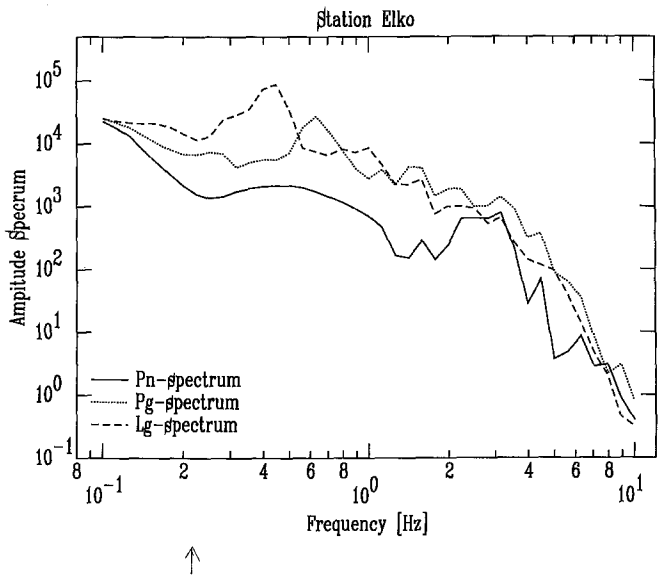
With these brief introductory comments on the seismic waves observed at the LLNL network, we state more precisely the seismic event discrimination problem which we studied: Given a seismogram with detected regional phases, *Pn*, *Pg*, and *Lg*, extract and use the characteristics of these phases to determine whether the source responsible for generating these waves was an earthquake or an underground nuclear explosion.

Seismic Discriminant Functions

In a review paper addressing the above problem, Pomeroy *et al.* (1982) summarize 15 classes of regional discriminants that could be used in the discrimination problem. They concluded that, while these discriminants had differing degrees of success in separating earthquakes from explosions, more research was required to determine clearly the most promising discriminants. Pomeroy *et al.* (1982) give an extensive



(a)



(b)

FIG. 7. Examples of displacement amplitude spectra (nm/Hz) of the *Pn*, *Pg*, and *Lg* phases of an earthquake (7a) and an underground nuclear explosion (7b) recorded at Elko. Note that the smooth characteristics of the spectral plots is due to the fact that the spectra were sampled in a log scale and plotted with linear interpolation between the sampled points.

discussion on the *Lg/P* amplitude ratio discriminant and conclude that, while *Lg/P* is a promising discriminant, earthquakes and explosions often, based on this discriminant, overlap significantly. A number of studies have extended the time-domain discriminants of Pomeroy *et al.* (1982) into the spectral domain and have conducted systematic comparisons on the performance of various discriminants. In a recent study, Taylor *et al.* (1988) extended the spectral ratio discriminant of

Bennett and Murphy (1986) to higher frequency bands for the same phase. For example, Taylor *et al.* (1988) define for the *Lg* phase a spectral ratio

$$D(Lg) = \sum_{k=K_{11}}^{K_{12}} S_{Lg}(f_k) \Big/ \sum_{k=K_{21}}^{K_{22}} S_{Lg}(f_k) \tag{11}$$

where (K_{11}, K_{12}) and (K_{21}, K_{22}) define upper and lower frequency bands of the spectra. They found that for a certain threshold level, a low value for $D(Lg)$ indicates a waveform was due to an earthquake. On the other hand, a high value for $D(Lg)$ would indicate an explosion. By applying their discriminants on real data, Taylor *et al.* (1988) concluded that spectral ratios at certain frequency bands ($[K_{11}, K_{12}]$, $[K_{21}, K_{22}]$) of a phase might be an important discriminant. Their results indicate that *Lg*, in comparison with *Pn* and *Pg*, performs the best for single-phase event discrimination. Performance in terms of misclassification probabilities ranged from 4 to 33 per cent, depending on the phase and the station, and from 7 to 16 per cent for the network of four stations of the LLNL Nevada Test Site (NTS) stations. In summary, the study by Taylor *et al.* (1988) shows clearly that the broadband spectral characteristics of the regional phases might be quite important for the discrimination of regional seismic events.

In view of the studies of Bennett and Murphy (1986) and Taylor *et al.* (1988) on spectral discriminants, it is reasonable to guess than an optimum discriminant might be a weighted spectral ratio of the form

$$D(Lg) = \sum_{\text{all } k} W_1(f_k) S_{Lg}(f_k) \Big/ \sum_{\text{all } k} W_2(f_k) S_{Lg}(f_k) \tag{12}$$

where $W_1(f_k)$ and $W_2(f_k)$ are spectral weightings at frequency f_k. In equation (11), $W_i(k) = 1$ for $i = 1, 2$. Another discriminant might be of the form of a ratio of spectral components of the different phases, a *multi-phase weighted spectral ratio*:

$$D(Pg, Lg) = \sum_{\text{all } k} W_{Lg}(f_k) S_{Lg}(f_k) \Big/ \sum_{\text{all } k} W_{Pg}(f_k) S_{Pg}(f_k) \tag{13}$$

In fact, the discriminant function need not be ratios but could conceivably be some unknown functional of the various frequency components of the three principal phases. In any event, the problem reduces to one of determining these unknown relationships and the optimum weighting functions $W_x(f_k)$. Given the large number of unknown variables like geology and source characteristics, analytical solution of this problem proves to be difficult.

The application of a "learning machine" like a neural network that determines discriminant functions automatically by systematically training on real data might be an alternative method for constructing the discriminant function. In this study, we focus on spectral discrimination by ANN using the frequency components of the principal seismic phases.

Seismic Database and Pre-processing

The results of this study are based on data from the four LLNL seismic stations at Elko (NV), Kanab (UT), Landers (CA), and Mina (NV), located at distances of

200 to 400 km from the Nevada Test Site (see Fig. 8). The database consisted of 83 earthquakes and 87 nuclear explosions at each of the four stations of the LLNL NTS network.

In order that discrimination is made on the basis of *source type*, and not unduly influenced by background noise, event distance, or event magnitude, for each event in the database we checked for the signal-to-noise ratio, exercised a *distance correction* and a *magnitude normalization* on the spectral data as routine pre-processing steps, i.e., before training and testing. It is important to check and account for many uncontrollable factors associated with the recordings of real data. We explain the pre-processing of the data in the following section.

Spectral Estimation and Signal-to-Noise Ratio Checking

For each event-station pair, the spectra were calculated from windowed *Pn*, *Pg*, and *Lg* phases. Group velocity windows were defined by t_1 and t_2, where $t_1 = \Delta/6.0$ and $t_2 = \Delta/5.0$ for the *Pg* phase, and $t_1 = \Delta/3.6$ and $t_2 = \Delta/3.0$ for the *Lg* phase. The *Pn* window was selected manually and generally ranged in length from 4 to 5 sec, starting from about 1 sec prior to the *Pn* arrival time. To get a smoother spectrum, the *Pn* window was extended to 20 sec by zero-padding of the data. Noise

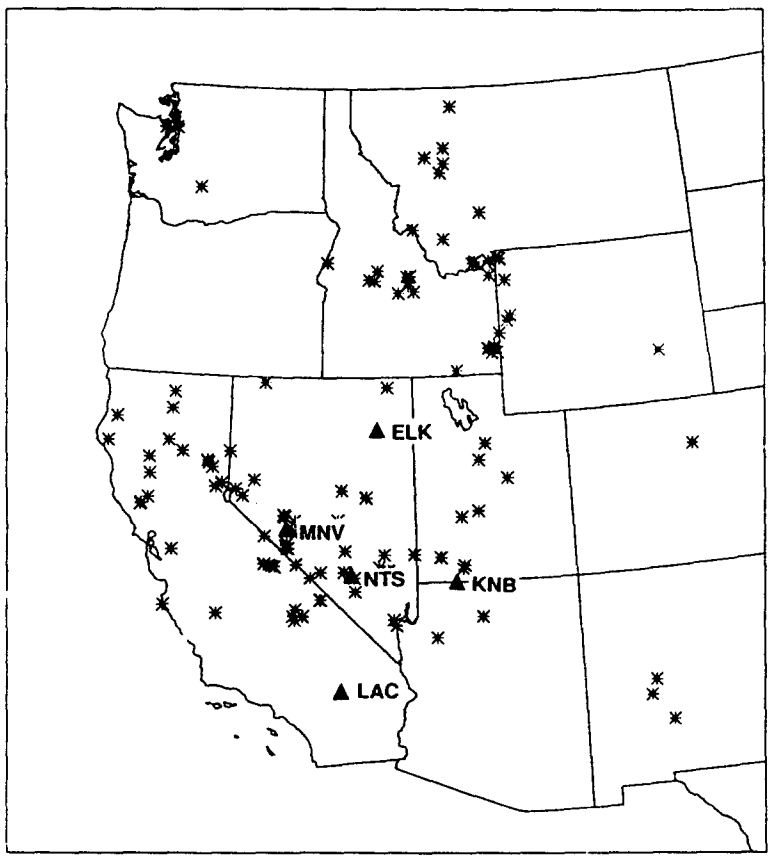


FIG. 8. Map of the LLNL Nevada Test Site seismic network. The network consists of four stations, indicated by the filled-triangles, located from 200 to 400 km from the Nevada Test Site. The locations of the earthquakes used in this study are indicated by asterisks.

spectra were calculated in a 30-sec window preceding the *Pn* arrival. The signals were differentiated to acceleration and windowed using a 10 per cent cosine taper between the limits defined above. The resulting acceleration spectra were divided by f^2 to convert them to displacement spectra. If three-component data were available, the *Pg* and *Lg* spectra were each averaged using the vertical, radial, and transverse components.

It is well known that spectral characteristics of seismic signals are strongly influenced by the background noise characteristics. In order to reduce the effect of noise on the signal spectra, only those frequencies for which the S/N level was greater than 2 was used. Frequencies at which the pre-event noise spectra exceeded the signal spectra by a factor of 2 were not used. Instead, these spectral points were obtained by fitting straight lines to the log spectra from the known neighboring points in the spectra.

The problem of missing data is a common problem in seismic analysis, and sophisticated methods of dealing with missing data are important when discrimination is based with a vector of parameters (Glaser *et al.*, 1986). However, in this study, since we used smooth spectral functions for inputs (rather than discrete measured parameters like $m_b : M_s$), a line fit to obtain missing points in the spectra was a simple solution which worked well.

Correction of Distance Effects on the Spectra

The distribution of the earthquakes in terms of distance is shown in Figure 9 for the station Elko. The explosions were all from approximately the same site, about 400 km from the station Elko. Because the members of the explosion population set were approximately from the same location, and the earthquake data were not, we made a first-order approximation to account for the effect of epicentral distance, d, on the source spectra. At frequency f, the observed spectrum $P(f, d)$, is given by

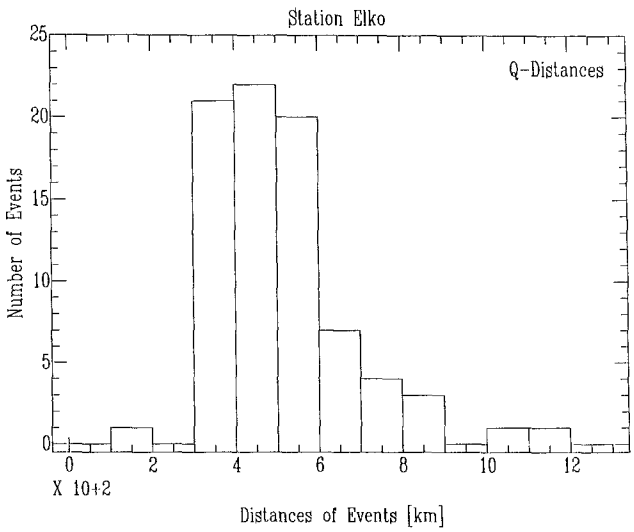


FIG. 9. Histogram of the distances for the earthquake events from the station Elko. Note that distances of the earthquakes varied from about 200 to 1200 km for this station.

the relation

$$P(f, d) = S(f) \cdot e^{-(\pi f d/Q(f)v)} \cdot d^{-\kappa} \tag{14}$$

where, $S(f)$ is the source spectrum, $d^{-\kappa}$ for some constant κ is the frequency-independent geometrical spreading, v the group velocity, and $Q(f)$ is the quality factor which is frequency dependent

$$Q(f) = \alpha \cdot f^{\beta}, \tag{15}$$

where α and β are constants which depend on the phase and on the regional geology. Since we assume that the $d^{-\kappa}$ term is frequency independent (and because we normalize the log spectra), a first-order distance correction function for the spectra is simply

$$C(f, d) = e^{+(\pi f d/Q(f)v)}. \tag{16}$$

Using results from studies applicable to the Basin and Range, we used $\alpha = 206$; $\beta = 0.60$ for *Pg*, and $\beta = 0.68$ for *Lg*; the velocity for *Pg* and *Lg* were 6.0 and 3.5 km/sec, respectively (Chavez and Priestly, 1986). (We did not use any distance correction for the *Pn* phase, and our final results are based on using only *Pg* and *Lg* phases as input to the ANN.) A distance-corrected spectrum, obtained from the raw spectrum is given by

$$S(f) = P(f, d) \cdot C(f, d). \tag{17}$$

A plot of the distance correction function versus frequency and distance is shown in Figure 10. From Figure 10 we see that distance correction effects are stronger at higher frequencies and for distant events.

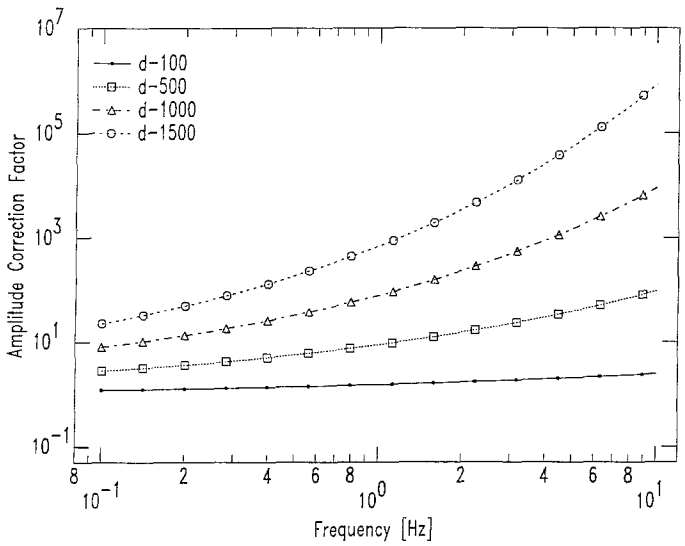


FIG. 10. Spectral correction function for the *Lg* spectrum plotted as a function of frequencies and distances (D-km). Note that large corrections are made only for distant events.

Normalization of Spectral Levels for Magnitude Invariance

As shown in Figure 11, on the average the explosions were of larger magnitudes than the earthquakes and certain pre-processing for reducing the effect of magnitude information in the data was appropriate. In order that discrimination is made on the basis of spectral shape and ratio, the input data to the ANN was formed by the distance-corrected log spectra of the three phases, *Pn*, *Pg*, and *Lg*, normalized such that the maximum value of the input data was 1; i.e., the distance-corrected normalized input vector for the ANN was obtained by first applying distance correction to the spectra, then taking the logarithm, and finally normalizing the spectra such that its maximum value of the three spectral phases was 1. In summary, we wanted to be able to discriminate events by type, explosion or earthquake, and not by event magnitude; for a good test, the magnitude histograms of the two population sets should be similar. However, since the magnitudes of the explosions were on the average larger than those of the earthquakes, we attempted to reduce the effect of event magnitude by normalizing the maximum value of each input pattern (the vector formed by the spectra of the three regional phases) to be unity.

Figures 12a and b show the statistical characteristics of the spectral data for the two populations. We draw attention to the fact that, for any single phase, there might be overlap. However, some distinct differences are observed between the two populations. For the explosions, the spectral shape of the *Lg* phase appears to have higher corner frequency and a steeper decay than the earthquakes. Because the maximum amplitude was generally taken from the *Lg* phase, there is more overlap in *Lg* than for *Pg*. The normalized *Pg* spectra from the explosions are typically greater than the earthquakes, which is consistent with a smaller *Lg*/*Pg* ratio for explosions over a broad range of frequencies, particularly between 1 and 5 Hz.

Discrimination Performance with Neural Nets

The multi-layered neural net used in this study was implemented on a SUN 4/110 workstation and we used a variety of architectures in terms of number of hidden layers and units. We found that even with a simple network architecture (see

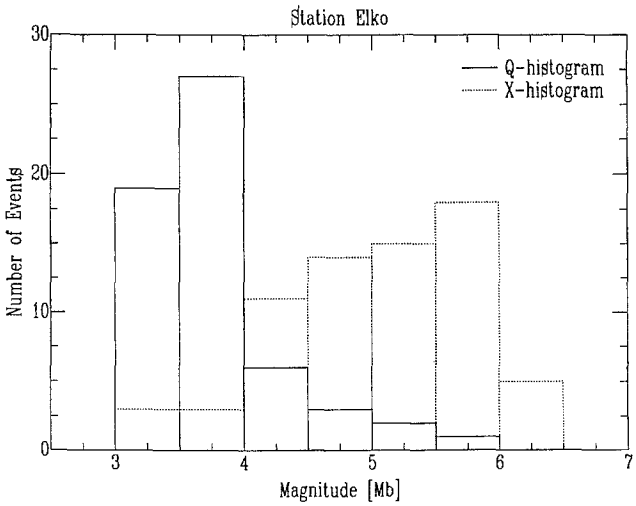
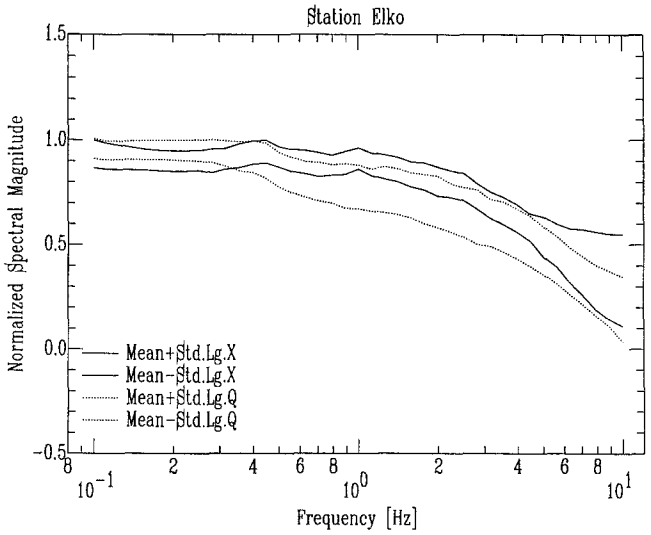
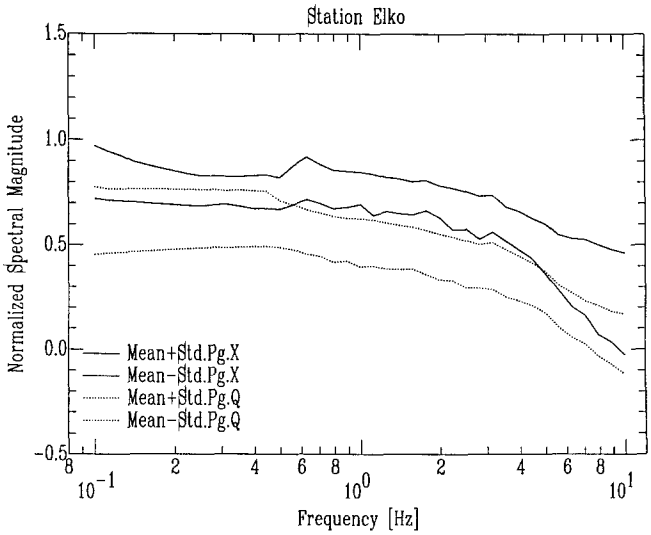


FIG. 11. Distribution of the two populations (earthquakes and explosions) of the data base plotted as a function of the event magnitude in m_b for the station Elko.



(a)



(b)

FIG. 12. Plots of the normalized (a) *Lg* and (b) *Pg* spectra for events at the station Elko. Like lines are the *mean* ± *standard deviation* obtained from the complete data base for an event type. Solid-dashed lines denote the explosion spectra and dotted-dashed lines denote earthquake spectra.

Fig. 5) consisting of one input layer, one hidden unit, and two output units, we could get very good performance. To regionalize the data characteristics, we assigned a separate network to each station.

We used a strategy which is called the "leave-one-out" (Lachenbruch and Mickey, 1968) in order to determine the performance of the ANN. According to this strategy, the ANN was trained with all but one of the events. Once trained, the ANN was tested with the event that was left out of the training set. To obtain a statistical measure of the discrimination performance, the process of leaving one event out,

TABLE 1

PERFORMANCE OF THE ARTIFICIAL NEURAL NETWORK

	Number of Events (Q/X)	Correct Identification (%)	Mis-Identification (%)	Undecided Classification (%)
Elko	80/79	97.5/97.5	0.00/2.50	2.50/0.00
Kanab	86/83	96.5/96.4	0.00/1.20	3.50/2.40
Landers	86/94	93.0/95.7	3.50/2.10	3.50/2.20
Mina	76/95	93.4/93.7	5.30/4.20	1.30/2.10
Network	79/83	100 /97.6	0.00/2.40	0.00/0.00

Performance of the ANN for discrimination expressed in terms of percentage. In the table the entries Q/X represent results for earthquakes (Q) and explosions (X), respectively.

training with the rest of the events, and testing with the left out event was repeated for all of the events in the database.

As shown in Table 1, preliminary results of this data set indicate that the rate of correct recognition for untrained data is from 93 to 97.5 per cent for earthquakes and explosions at any single station of the four station network. When using a majority voting scheme with the network of four stations, the rate of correct recognition is over 97 per cent. These results are based on using only the *Pg* and *Lg* spectra as input to the ANN. We found that inclusion of *Pn* data did not significantly change the performance, and since the *Pn* had lower signal-to-noise ratio it was left out in the final analysis.

Out of a total of 679 (earthquakes and explosions at four stations) spectra, the ANN either could not classify or mis-classified 29 spectra. The mean magnitude of these 29 spectra was 3.74 m_b with a standard deviation of 0.42. We can say, therefore, most of the events that were not correctly identified were among the small magnitude events.

While we are encouraged by these preliminary results, we note two defects in our present database. Figure 11 shows the histograms of the events in terms of their magnitude. Since we want to distinguish events by their type (explosion or earthquake) and not by event magnitude, ideally we want the histograms of the two populations to be similar. As the histogram shows, the magnitudes of the explosions were on the average larger than those of the earthquakes. We attempted to reduce the effect of event magnitude by normalizing the maximum value of each input pattern (the vector formed by the spectra of the three regional phases) to be unity. Because this normalization may not completely remove the effect of magnitude in the explosion and earthquake spectra, we cannot at present confirm whether the ANN was a magnitude-independent discriminator. The second peculiarity in our database was that the explosions were from the same site, while the earthquake locations had significant variations (see Fig. 9). We tried to reduce the distance effect by a simple distance correction algorithm explained earlier. In any event, to test the validity of these preliminary results, we need to verify them against other databases. Finally, we note that appropriate pre-processing (e.g., careful spectral analysis, distance correction, and normalization) was important for good discrimination performance.

Comparison with the Fisher Discriminant

Since empirical results showed that a very simple architecture for ANNs obtained results as good as those from complex structures with many hidden units, we applied

the conventional Fisher discrimination method (Tjøstheim, 1981) to the same data set for a first-order comparison of the results with conventional methods.

In order to use the Fisher discriminant, we must estimate a sample covariance matrix of the population and be able to invert this matrix. When the dimension of the input vector (≈80) is less than or equal to the number of example events in each population set, it is numerically difficult to invert the sample covariance matrix. By assuming equal covariance distribution for explosion and earthquake populations, however, we were able to compute a better conditioned covariance matrix because of more sample events, approximately 160. Under equal covariance Gaussian distribution of the population sets we have

$$\text{prob}(\mathbf{x} \mid \text{explosion}) = \frac{1}{(2\pi)^{n/2} \mid \Sigma \mid^{1/2}} \exp\left[\frac{1}{2}(\mathbf{x} - \mathbf{m}_\mathbf{x})^T \Sigma^{-1} (\mathbf{x} - \mathbf{m}_\mathbf{x})\right] \quad (18)$$

$$\text{prob}(\mathbf{x} \mid \text{earthquake}) = \frac{1}{(2\pi)^{n/2} \mid \Sigma \mid^{1/2}} \exp\left[\frac{1}{2}(\mathbf{x} - \mathbf{m}_q)^T \Sigma^{-1} (\mathbf{x} - \mathbf{m}_q)\right], \quad (19)$$

where $\mathbf{m}_\mathbf{x}$ and $\mathbf{m}_q$ are the sample mean vectors of the explosion and earthquake spectra, respectively, and Σ is the sample spectral covariance matrix. Assuming that the populations have an equal covariance Gaussian distribution as above, the log likelihood test for discrimination is given by

$$\text{Log}\left[\frac{\text{prob}(\mathbf{x} \mid \text{explosion})}{\text{prob}(\mathbf{x} \mid \text{earthquake})}\right] = \mathbf{x}^T \Sigma^{-1} (\mathbf{m}_\mathbf{x} - \mathbf{m}_q) - \frac{1}{2}(\mathbf{m}_\mathbf{x} + \mathbf{m}_q)^T \Sigma^{-1} (\mathbf{m}_\mathbf{x} - \mathbf{m}_q). \quad (20)$$

Results upon application of Fisher discriminant are listed in Table 2. Because we used the criteria that for positive likelihood ratio the event is an explosion and for negative likelihood ratio the event is an earthquake, there were no undecided events with the Fisher discriminant. (With an ANN, in contrast, high, medium, or low excitations in *both output neurons* meant that the classification was undecided. Because of these differences, a fair comparison over the network for the two methods becomes more complicated. As such, for this preliminary study, we do not compare the network performance between the two methods.) For any single station, the results on the same data set indicate that for the Fisher method rate of correct

TABLE 2

PERFORMANCE OF THE FISHER DISCRIMINANT

	Number of Events (Q/X)	Correct Identification (%)	Mis-Identification (%)
Elko	80/79	96.3/96.2	3.70/3.80
Kanab	86/83	93.0/90.4	7.00/9.60
Landers	86/94	83.9/98.9	16.1 /1.1
Mina	76/95	88.2/98.9	11.8 /1.1

Performance of the Fisher linear discriminant expressed in terms of percentage. As in Table 1, the entries Q/X represents results for earthquakes (Q) and explosions (X), respectively.

identification ranged from 88.2 to 98.9 per cent over the network for both earthquakes and explosions.

The results of this simple multivariate Fisher discrimination method are close to that of the ANN, although perhaps not as good as those of the ANN. The comparable performance of the two methods is not surprising because with a very simple architecture for the ANN we obtained results as good as those from complex ANN architectures with many hidden units meaning that the two populations are almost linearly separable and therefore the performance of the Fisher technique should be quite close to optimal.

In summary, both ANN and Fisher methods perform quite well with regional spectral data, indicating perhaps that given complete spectral information, events with magnitude 4 or higher in this geologic region can be discriminated rather well. Whether the same discrimination method will perform equally well in other geophysical environments or at lower m_b values is unknown.

Discussion and Conclusions

In our study we deliberately kept the architecture of the ANN to a simple form in order to examine the weights and understand the features in the spectral data which the network utilized for discrimination of the two populations.

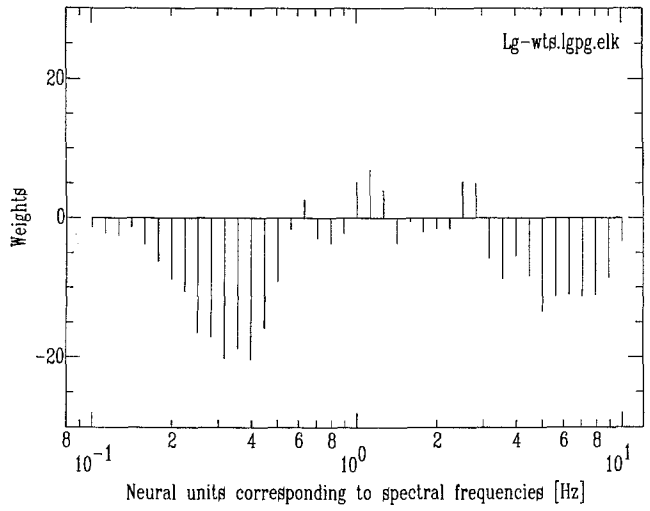
In Figure 13, we compare the *Lg* and *Pg* weights of the ANN for the station Elko when the input are the *Lg* and *Pg* spectra. Note that most of the weights for *Pg* are positive and those for *Lg* are negative. Since the log of the spectra was input to the ANN, this implies that the ANN has learned to distinguish an explosion from an earthquake by using the *Lg*/*Pg* spectral ratio. It is interesting, although perhaps not surprising, that the ANN through its learning process automatically developed a discriminant which had been previously recommended by experts in the field.

We investigated the variations of the input weights with the different training sets (using the leave-one-out approach, the number of training sets is equal to the total number of earthquakes and explosions at a station). The standard deviation of the weight vector about its mean showed that there are no significant variations in the weight distribution with different training sets of the leave-one-out method. Another intesting result was the fact that the weight vectors exhibited variations in the detailed shape between the four stations—although they had the same general shape. The *Pg* weight vectors for the other three stations are compared in Figure 14.

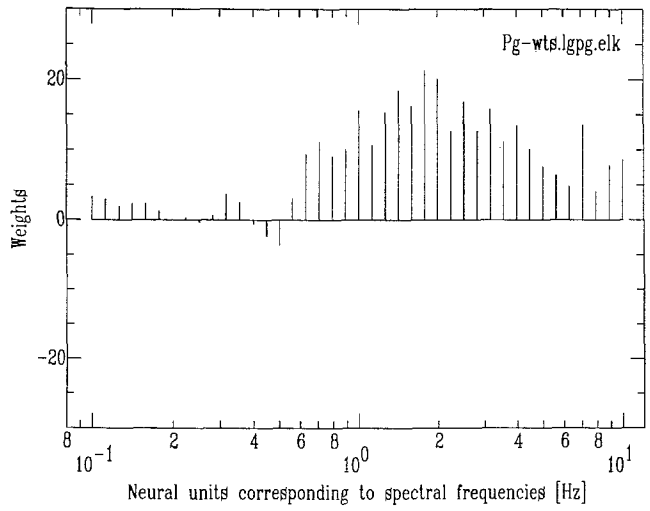
There are many interesting questions and issues that we have not addressed in this preliminary study. Some of the problems we are currently studying are: How important are the detailed spectral shape of the various phases? Was SNR the main reason that the *Pn* data did not significantly contribute to the solution? How important is it to train the ANN station by station, or would it be better to use all the data from the seismic network and apply it to the ANN? We intend to address these important questions in our next report.

We are also conducting further studies to extend the ANN technique for the discrimination of complete seismograms or waveforms, instead of using detected windowed phases. This would make the ANN more independent of analysts and render it an important tool in automatic platforms for seismic signal analysis and discrimination.

Although the performance of the ANN for discrimination between explosions and earthquakes was quite similar to that of the Fisher linear discriminant, we found



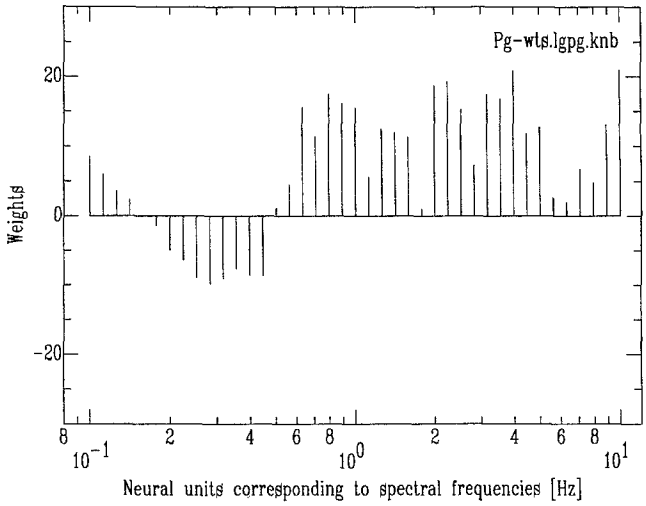
(a)



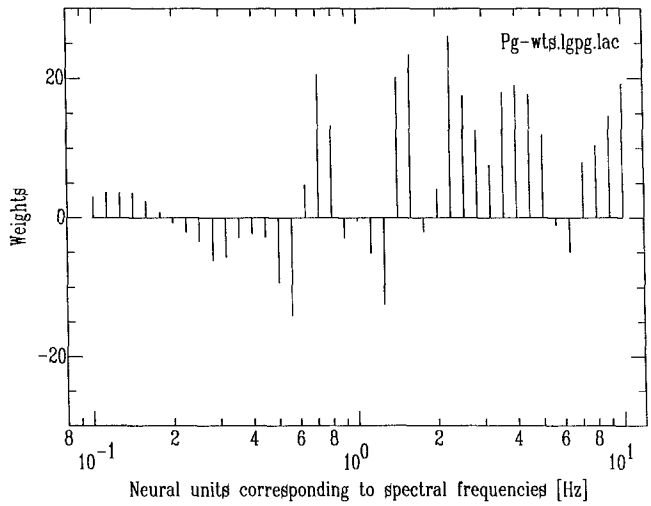
(b)

FIG. 13. The weights from the input to a hidden unit for a trained network at Elko. Because the net had a simple architecture and the input units correspond to the frequencies in the spectrum, examination of the weights is useful for understanding how the ANN used features in the *Lg* and *Pg* spectra for discrimination. Since the log spectra of *Pg* has mostly positive weights and the log spectra of *Lg* has mostly negative weights, the ANN is using the *Lg/Pg* spectral ratio for discrimination, in addition to the spectral shape.

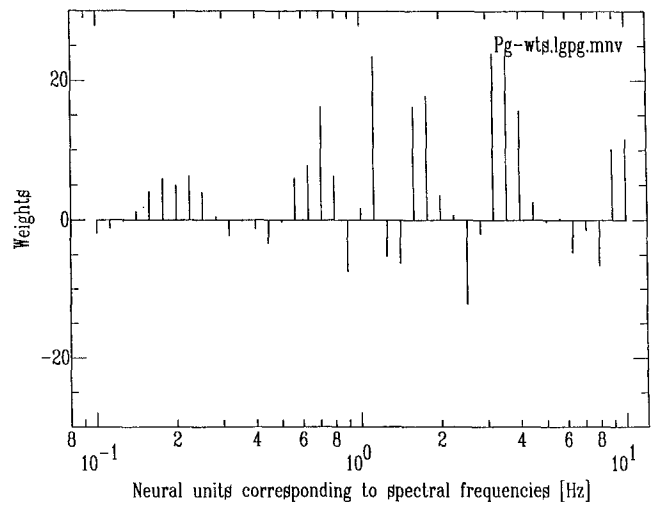
that the ANN exhibits a number of computational advantages over conventional methods. When the dimensions of the data is large (in this study the length of the input vector was 82), conventional methods can become numerically unstable and slow. For example, the Fisher technique requires inversion of a covariance matrix, and it is well-known that inversion of a large dimensional matrix can be numerically difficult. Furthermore, from many other experiments that we performed we found that the ANN is quite robust to missing data points. Finally, the ANN provides a computational environment that is significantly more flexible and simple from the viewpoint of the application scientist.



(a)



(b)



(c)

FIG. 14. The weights corresponding to the *Pg* spectra for the other three stations. The weights of the trained ANN differ from station to station, supporting the fact that regional differences are important for determining optimal discriminant functions.

ACKNOWLEDGMENTS

The authors wish to thank L. Hutchings, S. Lu, P. Kasameyer, R. Searfus, and particularly one of the reviewers for careful review and helpful comments. This work is performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract W-7405-ENG-48.

REFERENCES

Bennett, T. J. and J. R. Murphy (1986). Analysis of seismic discrimination using regional data from western United States events, *Bull. Seism. Soc. Am.* **76,** 1069–1086.

Chavez, D. E. and K. F. Priestley (1986). Measurement of frequency dependent Lg attenuation in the Great Basin, *Geophys. Res. Lett.* **13,** 551–554.

Dahlman, O. and E. Israelson (1977). *Monitoring Underground Nuclear Explosions*, Elsevier, Amsterdam, 440 pp.

DARPA Neural Network Study (1988). AFCEA International Press, Fairfax, Virginia.

Dowla, F. U., E. B. Talbot, and V. Vemuri (1988). Artificial neural networks: an application for the classification of seismic signals, Technical Report no. UCID-21552, Lawrence Livermore National Laboratory.

Glaser, R. E., S. R. Taylor, M. D. Denny, and E. S. Vergino (1986). Regional discriminants of NTS explosions and western U.S. earthquakes: multivariate discriminants, Technical Report no. UCID-20930, Lawrence Livermore National Laboratory.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational properties, *Proc. Natl. Acad. Sci.* **79,** 2554–2558.

Lachenbruch, P. A. and R. M. Mickey (1968). Estimation of error rates in discriminant analysis, *Technometrics* **10,** 1–11.

Lippman, R. P. (1987). Introduction to computing with neural nets, *IEEE ASSP Magazine* **4,** 4–22.

Pomeroy, P. W., W. J. Best, and T. V. McEvilly (1982). Test ban treaty verification with regional data: a review, *Bull. Seism. Soc. Am.* **72,** S89–S129.

Rumelhart, D. E., J. L. McClelland, and the PDP Research Group (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1, MIT Press, Boston.

Taylor, S. R., N. W. Sherman, and M. V. Denny (1988). Spectral discrimination between NTS explosions and western United States earthquakes at regional distances, *Bull. Seism. Soc. Am.* **78,** 1563–1579.

Taylor, S. R., M. V. Denny, E. S. Vergino, and R. E. Glaser (1989). Regional discrimination between NTS explosions and western U.S. earthquakes, *Bull. Seism. Soc. Am.* **79,** 1142–1176.

Tjøstheim, D. (1981). Multidimensional discrimination techniques: theory and application, in *Identification of Seismic Sources: Earthquake or Underground Explosion*, E. S. Husebye and S. Mykkeltveit (Editors), Reidel, Dordrecht, pp. 663–694.

APPENDIX: DERIVATION OF THE BACKPROPAGATION ALGORITHM

Notation

The backpropagation learning algorithm is a gradient descent method and uses the rule

$$w_k \leftarrow w_k - \mu \frac{\partial E}{\partial w_k} \tag{1}$$

iteratively to update the weights in the network until the weights converge to the desired solution. Before we derive the backpropagation rule, it is helpful to explain the notation first. In Figure 15, consider the mth neuron or unit in the nth layer. In this neuron, $w_{km}^{(n)}$ represents the kth weight, $s_m^{(n)}$ the sum of the weighted inputs, and $x_m^{(n)} = f(s_m^{(n)})$ the output of the neuron. As an example, $x_k^{(n-1)}$ is the output of the kth neuron in layer $(n-1)$; $x_k^{(n-1)}$ is also the kth input of each neuron in layer

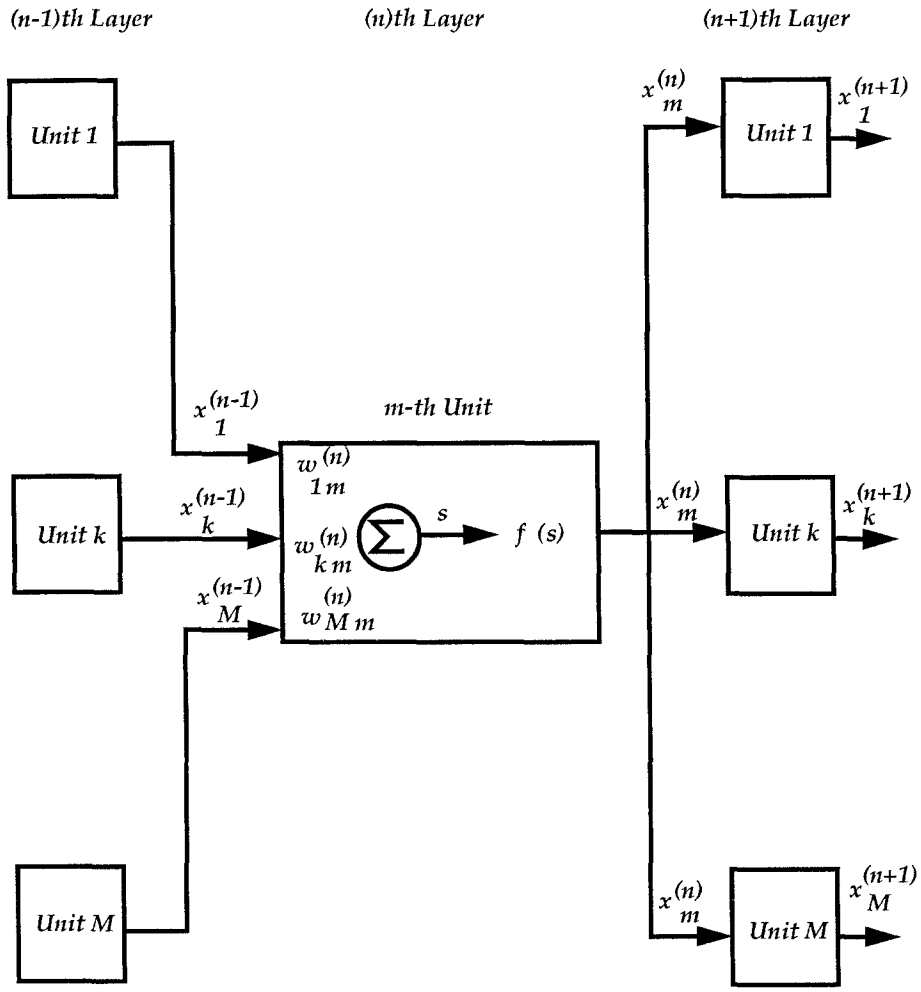


FIG. 15. Diagram illustrating the notation for the inputs, weights, and outputs of a general multilayered perceptron.

(n); i.e., $x_k^{(n-1)}$ is connected to weight $w_{km}^{(n)}$, the kth weight of the mth neuron in layer (n).

In order to derive the weight update rule of the backpropagation algorithm we use the chain rule to express the derivative of the error w.r.t. any weight as a product of the derivative of the error w.r.t. the sum of the weighted inputs, s, and the derivative of sum of the weighted inputs w.r.t. that weight

$$\frac{\partial E}{\partial w_{km}^{(n)}} = \frac{\partial E}{\partial s_m^{(n)}} \cdot \frac{\partial s_m^{(n)}}{\partial w_{km}^{(n)}}. \tag{2}$$

Since the sum of the weighted inputs of the nth neuron before applying the nonlinear transfer function is

$$s_m^{(n)} = \sum w_{jm}^{(n)} x_j^{(n-1)}, \tag{3}$$

we can write

$$\frac{\partial s_m^{(n)}}{\partial w_{km}^{(n)}} = x_k^{(n-1)}. \tag{4}$$

By substituting (4) in (2) we get

$$\frac{\partial E}{\partial w_{km}^{(n)}} = x_k^{(n-1)} \frac{\partial E}{\partial s_m^{(n)}}, \tag{5}$$

and using the chain rule again we have

$$\frac{\partial E}{\partial w_{km}^{(n)}} = x_k^{(n-1)} \cdot \frac{\partial E}{\partial x_m^{(n)}} \cdot \frac{\partial x_m^{(n)}}{\partial s_m^{(n)}}. \tag{6}$$

The nonlinear transfer function used in the network is the sigmoid function given by

$$f(s) = \frac{1}{1 + e^{-s}}, \tag{7}$$

and it can be shown that

$$f'(s) = f(s) \cdot (1 - f(s)). \tag{8}$$

Hence, when $f(s)$ is known then $f'(s)$ can be easily computed, an important property used in the derivation of the algorithm. Now, since $f(s_m^{(n)}) = x_m^{(n)}$, equation (5) reduces to

$$\frac{\partial E}{\partial w_{km}^{(n)}} = x_k^{(n-1)} f'(s_m^{(n)}) \cdot \frac{\partial E}{\partial x_m^{(n)}}. \tag{9}$$

Equation (9) is an important relation in backpropagation as it expresses derivative of the error w.r.t. any weight in terms of the derivative of the error w.r.t. the neural output.

The next key step to the derivation is a recursive expression for $\partial E/\partial x_m^{(n)}$ in terms of $\partial E/\partial x_k^{(n+1)}$, where the k's represent the neurons of the $(n + 1)$th layer. This is done by again making use of the chain rule. First, we write

$$\frac{\partial E}{\partial x_m^{(n)}} = \sum_j \frac{\partial E}{\partial s_j^{(n+1)}} \cdot \frac{\partial s_j^{(n+1)}}{\partial x_m^{(n)}}. \tag{10}$$

Now since

$$s_j^{(n+1)} = \sum w_{ij}^{(n+1)} x_i^{(n)}, \tag{11}$$

we have

$$\frac{\partial s_j^{(n+1)}}{\partial x_m^{(n)}} = w_{mj}^{(n+1)}. \tag{12}$$

By substitution of (12) in (10) we have

$$\frac{\partial E}{\partial x_m^{(n)}} = \sum_j \frac{\partial E}{\partial s_j^{(n+1)}} \cdot w_{mj}^{(n+1)}. \tag{13}$$

Since $x_j^{(n+1)} = f(s_j^{(n+1)})$,

$$\frac{\partial E}{\partial s_j^{(n+1)}} = f'(s_j^{(n+1)}) \frac{\partial E}{\partial x_j^{(n+1)}}. \tag{14}$$

Finally, substituting (14) in (13) we get the desired relationship:

$$\frac{\partial E}{\partial x_m^{(n)}} = \sum_j w_{mj}^{(n+1)} f'(s_j^{(n+1)}) \frac{\partial E}{\partial x_j^{(n+1)}}. \tag{15}$$

In summary, there are two key equations to the derivation of the method. These are equations (9) and (15). Equation (9) expresses, $\partial E/\partial w_{km}^{(n)}$, the derivative of the error w.r.t. a weight in terms of $\partial E/\partial x_m^{(n)}$, the derivative of the error w.r.t. the output of that neuron. And equation (15) shows that $\partial E/\partial x_m^{(n)}$ can be recursively expressed in terms of $\partial E/\partial x_m^{(n+1)}$, the derivative of the error w.r.t. the outputs of the next layer.

By defining

$$\delta_m^{(n)} = f'(s_m^{(n)}) \frac{\partial E}{\partial x_m^{(n)}} \tag{16}$$

and using equation (9), we have a concise recursive update formula for the network weights after presentation of each pattern

$$w_{km}^{(n+1)} \leftarrow w_{km}^{(n-1)} - \mu \delta_m^{(n-1)} x_m^{(n-2)}, \tag{17}$$

where μ is a small constant, the learning rate. And using equation (16), for the mth neuron in the output layer (the output layer is the Nth layer),

$$\delta_m^{(N)} = (T_m - O_m) f'(s_m^{(N)}). \tag{18}$$

For the hidden units in layer $(n-1)$, using equations (16) and (15) we have

$$\delta_k^{(n-1)} = f'(s_k^{(n-1)}) \sum_m w_{km}^{(n)} \delta_m^{(n)}, \tag{19}$$

where m is over all neurons in layer (n).

In conclusion, after presentation of each pair of input and desired output elements to the network and by using (17), (18) and (19), we can update the weights of the network until the network error has converged to a minimum value. The convergence property of the algorithm is an important research topic.

TREATY VERIFICATION PROGRAM
LAWRENCE LIVERMORE NATIONAL LABORATORY
LIVERMORE, CALIFORNIA 94550

Manuscript received 28 November 1989