Constrained Clustering as an Optimization Method

Kenneth Rose, *Member, IEEE,* Eitan Gurewitz, and Geoffrey C. Fox

Abstract—**Our deterministic annealing approach to clustering is derived on the basis of the principle of maximum entropy, is independent of the initial state, and produces natural hierarchical clustering solutions by going through a sequence of phase transitions. This approach is modified here for a larger class of optimization problems by adding constraints to the free energy. The concept of constrained clustering is explained, and then, three examples are given in which it is used as means to introduce deterministic annealing. First, the previous clustering method is improved by adding cluster mass variables and a total mass constraint. Second, the traveling salesman problem (TSP) is reformulated as constrained clustering, yielding the elastic net (EN) approach to the problem. More insight is gained by identifying a second Lagrange multiplier that is related to the tour length and can also be used to control the annealing process. Finally, the "open path" constraint formulation is shown to relate to dimensionality reduction by self-organization in unsupervised learning. A similar annealing procedure is applicable in this case as well.**

Index Terms—**Annealing, clustering, maximum entropy, neural networks, nonconvex optimization, self-organization.**

I. Introduction

The problem of clustering is an important optimization problem because it is encountered in many applications involving data analysis without prior knowledge of the probability distributions. In particular, it is a major problem in the fields of pattern recognition, unsupervised learning, and data compression. Clustering is usually formulated as an optimization problem by defining a cost function to be minimized. Most of these cost functions are nonconvex and have several local minima. Traditional techniques [1]–[4] are essentially descent algorithms because the cost is reduced at each iteration. They therefore tend to get trapped in local minima. For a review and discussion of standard clustering methods, see [5], [6], or [7].

Simulated annealing or stochastic relaxation [8] is a known technique for avoiding local minima of nonconvex optimization problems. A sequence of random moves is generated, and the decision to accept a move depends on the probability of the resulting configuration. Thus, the cost is not always reduced, and the process may escape local minima. However, this process requires very slow schedules [9] that are not realistic for many practical applications. An excellent review of the theory and applications of simulated annealing can be found in [10]. For an extension of stochastic relaxation to incorporate hard constraints, see [11].

In our previous work [12] [13], we proposed the concept of deterministic annealing for the problem of clustering and vector quantization. Although strongly motivated by the physical analogy, our approach is based on principles of information theory. Jaynes [14] applied Shannon's theory [15] in his information theoretical formulation of statistical mechanics and stated the principle of maximum entropy for statistical inference; this is the basis on which we formulated our probabilistic framework. Within this framework, we obtained an effective cost (free energy) that is parameterized by the "temperature," which is a parameter determined by the average cost (energy). We have shown that this free energy is minimized by the most probable set of cluster representatives at a given temperature. We use the term cluster representative for a vector of parameters of a particular cluster. In a simple context (such as for K-means), it is the cluster mean. If more complex distortion measures are used, it may be the corresponding cluster "centroid," and it may also include other parameters such as the cluster population, etc. At high temperatures, there is only one local minimum (the global minimum), and the annealing method tracks the minimum while gradually lowering the temperature. A hierarchy of clustering solutions at decreasing average cost is obtained as the process goes through a sequence of phase transitions.

Clustering belongs to a large family of optimization problems that can be stated in terms of the associations between a set of fixed data and a set of variables. In clustering, these are the set of data points and the set of representatives, respectively. At the limit of low temperature, these associations are no longer fuzzy (each point is fully associated with one representative). Many other optimization problems in this family are concerned with finding either fuzzy or nonfuzzy associations, which would meet some further requirements. We suggest a reformulation of these problems as constrained clustering and apply deterministic annealing methods to solve them.

We have chosen to demonstrate this approach with three examples. The first is an improvement of our clustering method. By adding "mass" variables that, in fact, measure the cluster population and a corresponding total mass constraint, we eliminate a weakness of the original algorithm and its dependence on the number of representatives, as will be explained in Section IV.

Manuscript received April 15, 1991; revised September 6, 1991. This work was carried out when the authors were with the Caltech Concurrent Computation Program, California Institute of Technology, Pasadena, CA 91125. It was supported by the Joint Tactical Fusion Program Office—Reference A:49-288-82201-0-8420 and by McDonnell Douglas. Recommended for acceptance by Editor-in-Chief A. K. Jain.

K. Rose is with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106.

E. Gurewitz is with the Department of Physics, Nuclear Research Centre—Negev, Beer-Sheva, Israel.

G. C. Fox is with the Northeast Parallel Architectures Center, Syracuse University, Syracuse, NY 13244.

IEEE Log Number 9208013.

0162-8828/93$03.00 © 1993 IEEE

Authorized licensed use limited to: UNIVERSIDADE FEDERAL DO RIO DE JANEIRO. Downloaded on April 28,2010 at 11:00:12 UTC from IEEE Xplore. Restrictions apply.

Another example of constrained clustering is the elastic net (EN) method [16] for the traveling salesman problem (TSP). In TSP, we want to associate an ordered set of variables (representatives) with a given fixed set of cities so that the sum of distances between consecutive cities is minimized. This is a typical example of a problem that is concerned with nonfuzzy associations. We show that by reformulating the problem as constrained clustering, where the constraint takes care of the distances between consecutive representatives, our deterministic annealing method becomes equivalent to EN. Moreover, in this new formulation, the coefficient used by EN to weigh the relative importance of terms in the cost function is identified as the Lagrange multiplier related to the tour length constraint. In other words, an additional annealing parameter (another type of "temperature") exists in the process. It is therefore also used as control parameter, yielding a more powerful annealing procedure, as will be explained in Section V.

Finally, we show that by removing the periodic boundary requirement from TSP, i.e., searching for the shortest open path, we obtain a process that is directly related to self-organization in learning. The objectives suggested by Kohonen [17] are that the representatives will "fit" the data distribution (clustering) and be "ordered" (the constraint). The approach applied in this example is neither stochastic nor sequential as in [17] and searches for the most probable set of ordered representatives for a given training set by minimizing the effective energy at each temperature.

Let us briefly comment on the convergence of our constrained clustering method. It essentially consists of repeated local optimization while gradually lowering the temperature. The formulation is independent of the actual choice of the local optimization algorithm, and the rate of convergence at a given temperature will depend on this choice. At the limit of zero temperature, the method usually becomes a known descent method (e.g., basic ISODATA for the clustering example), which has known convergence properties. Convergence to the global minimum, however, is not assured, and many issues concerning the annealing schedule have not yet been addressed.

A word of caution is due concerning some terminology used in this paper. Although our approach is truly probabilistic, we use "fuzzy-sets terminology," where it seems to convey more intuition or to emphasize the relation to fuzzy clustering. This "fuzzy terminology" should be understood in the following sense. In our probabilistic framework, a cluster is a set whose members are determined as realizations of random variables under the corresponding association probabilities. We conveniently introduce fuzzy clusters by setting a data point's degree of membership in a particular fuzzy cluster to be exactly its probability of belonging to the corresponding nonfuzzy cluster. In this light, fuzziness is directly related to uncertainty, and annealing can be viewed as gradually decreasing the fuzziness of the associations. However, we make no essential use of fuzzy-sets theory.

II. Clustering by Deterministic Annealing

In our previous work [12], we suggested a deterministic annealing approach to clustering. In this section, we briefly summarize the approach. A probabilistic formulation is used where each point is associated *in probability* with each cluster. The state of the system is given by the set of probability distributions for associating points with clusters. Assuming that the set of representatives $Y = \{y_j\}$ is given, then the expected energy (distortion) is

$$E = \sum_x \sum_j P(x \in C_j) d(x, y_j), \tag{1}$$

where $d(x, y_j)$ is the distortion or dissimilarity measure for representing data point x by the vector y_j, and $P(x \in C_j)$ is the probability that x belongs to the cluster of points represented by y_j. Since we do *not* have any prior knowledge about the probability distribution, we apply the principle of maximum entropy. As is well known, the probability distributions that maximize the entropy under an expectation constraint are Gibbs distributions. For the constraint (1), we get

$$P(x \in C_j) = \frac{e^{-\beta d(x,y_j)}}{Z_x(Y)} \tag{2}$$

where Z_x is the partition function

$$Z_x(Y) = \sum_k e^{-\beta d(x,y_k)}. \tag{3}$$

The parameter β is the Lagrange multiplier determined by the given value of E in (1). In the process of annealing, it is inversely proportional to the "temperature." For a given set of representatives, it is assumed that the associations of different data points to their clusters are independent. Hence, the total partition function is

$$Z(Y) = \prod_x Z_x(Y). \tag{4}$$

Instead of considering the association probability of a data point, we consider the probability of an entire instance (Y, V) given by a set of representatives $Y = \{y_j\}$ and a partition via the set of associations $V = \{v_{xj}\}$, where

$$v_{xj} = \begin{cases} 1 & \text{if } x \in C_j \\ 0 & \text{otherwise} \end{cases}$$

By the principle of maximum entropy, the probability of this instance is given by the Gibbs distribution

$$P(Y, V) = \frac{e^{-\beta D(Y,V)}}{\sum_{Y',V'} e^{-\beta D(Y',V')}} \tag{5}$$

where

$$D(Y, V) = \sum_x \sum_j v_{xj} d(x, y_j) \tag{6}$$

is the distortion associated with the instance. The most probable instance is the one that maximizes the probability in (5) and, hence, minimizes D. However, the most probable set of representatives is the one maximizing the marginal probability

$$P(Y) = \sum_V P(Y, V). \tag{7}$$

Authorized licensed use limited to: UNIVERSIDADE FEDERAL DO RIO DE JANEIRO. Downloaded on April 28,2010 at 11:00:12 UTC from IEEE Xplore. Restrictions apply.

It can be shown [12] that

$$P(Y) = \frac{Z(Y)}{\sum_{Y'} Z(Y')} = \frac{e^{-\beta F(Y)}}{\sum_{Y'} e^{-\beta F(Y')}} \tag{8}$$

where

$$F(Y) = -\frac{1}{\beta} \log Z(Y). \tag{9}$$

Maximizing the marginal probability $P(Y)$ in (8) requires minimizing the free energy F, which is given explicitly by

$$F = -\frac{1}{\beta} \sum_x \log \sum_k e^{-\beta d(x, y_k)}. \tag{10}$$

Note that the free energy is exactly the quantity that is minimized at isothermal equilibrium in statistical mechanics. This shows clearly the relation between this deterministic approach and simulated annealing. The set Y of vectors that optimizes the free energy satisfies

$$\sum_x P(x \in C_j) \frac{\partial}{\partial y_j} d(x, y_j) = 0. \tag{11}$$

At $\beta = 0$, all data points are equally associated with all representatives (2). Assuming that there is a unique $\hat{y}$ that minimizes

$$\sum_x d(x, y)$$

we get that there is only one solution $y_j = \hat{y} \forall j$, and regardless of their number, the representatives will all converge to the same point (the global minimum). Note that no matter how many representatives we throw in, i.e., regardless of the size of the set Y, the *natural* number of clusters, which is the number of *distinct* representatives, will emerge at every given β. At $\beta = 0$, we have one natural cluster consisting of the entire data set. However, at some positive β, the cluster will split into smaller clusters and will thus undergo a phase transition. In the case that our distortion measure is the sum of squared distances, the critical value for β satisfies

$$(I - 2\beta_c C_{xx}) y_j = 0 \tag{12}$$

where C_{xx} is the covariance matrix of the training set. The critical temperature is thus determined by the variance along the largest principal axis of the distribution. In addition, the split is initiated along this axis. A sequence of such phase transitions is observed as β is increased and reaches critical values related to the covariance of the fuzzy clusters.

III. Constrained Clustering

In the above formulation of clustering, no explicit constraint has been put on the set of representatives. It was only implicitly assumed that there were at least two identical representatives at each natural cluster to allow phase transitions. By adding explicit constraints, one can use our annealing mechanism to solve other optimization problems as well as improve the clustering solution.

There is a large family of optimization problems that may be viewed as looking for the optimal associations between two sets: one set of variables and one set of fixed data. In clustering, these are the set of representatives and the set of data points, respectively. In TSP, we want to associate an *ordered* set of variables with a given fixed set of cities to minimize the sum of consecutive distances. Image segmentation clearly belongs to this family as well because we want to optimally associate pixels with an appropriate set of meaningful labels. The deterministic annealing clustering method offers a tool for obtaining such associations. Annealing is obtained as the system starts at very fuzzy associations (high temperature), and then, the fuzziness is gradually reduced as the temperature is lowered. Thus, many association problems may be reformulated as constrained clustering, where the constraint incorporates the requirements for the optimal associations. This gives rise to new applications for our deterministic annealing method for clustering. It should, however, be noted that these optimization problems are divided into two groups. The objective of one group is to find the optimal nonfuzzy associations (e.g., TSP). For these problems, deterministic annealing is merely a tool for avoiding local minima. The second group is typically concerned with generalizing from a training data set, or related to input density parameter estimation, and therefore, fuzzy associations are sought. In this case, the free energy not only is a useful approximation for avoiding local minima but is apparently the right cost function to minimize to obtain the most probable solution at a given temperature.

Let us formulate the approach to constrained clustering, based on the general principle of maximum entropy. As before, an instance of the system (Y, V) is given by Y, which is the set of cluster representatives, and V, which is a hard partition. Over the set of instances, we define a probability distribution that will maximize the entropy subject to the following two constraints: First, we have our familiar average clustering distortion constraint

$$\langle D(Y, V) \rangle = E. \tag{13}$$

where D is as in (6)

$$D(Y, V) = \sum_x \sum_j v_{xj} \, d(x, y_j) \tag{14}$$

and then some extra constraint that only concerns the cluster representatives

$$\langle T(Y) \rangle = L. \tag{15}$$

The maximum entropy probability distribution is

$$P(Y, V) = \frac{e^{-\beta D(Y, V) - \lambda T(Y)}}{\sum_{Y', V'} e^{-\beta D(Y', V') - \lambda T(Y')}}. \tag{16}$$

By summing over all possible hard partitions, similarly to the derivation for clustering, we obtain the marginal probability

$$P(Y) = \frac{e^{-\beta F(Y, \beta) - \lambda T(Y)}}{\sum_{Y'} e^{-\beta F(Y', \beta) - \lambda T(Y')}} \tag{17}$$

where F is given in (9).

Authorized licensed use limited to: UNIVERSIDADE FEDERAL DO RIO DE JANEIRO. Downloaded on April 28,2010 at 11:00:12 UTC from IEEE Xplore. Restrictions apply.

The most probable Y is the one minimizing $\beta F + \lambda T$. Equivalently, we could say that it minimizes

$$F + \frac{\lambda}{\beta} T \tag{18}$$

which, by noting its Lagrangian form, can be conveniently viewed as minimizing $F(Y, \beta)$ subject to $T(Y) = L'$ for some appropriate L'. Furthermore, in many cases, and in all our examples, the actual value of L' will not be important. All that will matter is the way it is varied for the annealing process.

In constrained clustering, we shall therefore be optimizing the free energy subject to our constraint. As is often done in such optimization problems, it is useful to minimize the Lagrangian; therefore, let us rewrite (18) as

$$F' = F + qT \tag{19}$$

where F is given in (10), q is a Lagrange multiplier, and T is the constraint. The Lagrangian is normally optimized as functions of q (the Lagrange multiplier), which is then determined by satisfying the constraint.

Three examples of constrained clustering are given in this paper. The first is an improvement of our method for the clustering problem. Recall that at each temperature, we had groups of representatives converging to the same points and defined the number of natural clusters to be the number of distinct representatives. The number of identical representatives at a given natural cluster will be referred to as the multiplicity of representatives in the cluster. Although the number of natural clusters at a given temperature is independent of the total number of representatives, their actual location depends on the number of representatives and their multiplicity in the clusters. This weakness is eliminated by reformulating the problem as constrained clustering or, equivalently, by taking into account the mass (or population) of each natural cluster. As a second example of constrained clustering, we take TSP. We show how TSP can be viewed as constrained clustering at the limit of low temperature and obtain the EN method [16], [18], which is an important intuitive method that has been shown to obtain near-optimal solutions for relatively complicated configurations of cities. Moreover, our constrained-clustering formulation leads us to identify a second Lagrange multiplier and to propose an annealing scheme that exploits both annealing parameters. The last example is related to self-organization in unsupervised learning [17]. It is explained how an appropriate constrained clustering formulation leads to a deterministic annealing method to search for the optimal solution given a finite training set.

IV. Mass- Constrained Clustering

Let us reformulate our clustering method in terms of the natural clusters (or distinct representatives). Let λ_k denote the multiplicity of identical representatives in the kth cluster. Equation (3) for the partition function is rewritten as

$$Z_x = \sum_k \lambda_k e^{-\beta d(x, y_k)} \tag{20}$$

the association probability (2) is for a natural cluster

$$P(x \in C_j) = \frac{\lambda_j e^{-\beta d(x, y_j)}}{Z_x} \tag{21}$$

and the free energy (9) is now

$$F = -\frac{1}{\beta} \sum_x \log Z_x = -\frac{1}{\beta} \sum_x \log \sum_k \lambda_k e^{-\beta d(x, y_k)}. \tag{22}$$

The free energy is to be minimized under the constraint of a fixed total number of representatives. The Lagrangian to be minimized (19) is thus

$$F' = F + q\left(\sum_k \lambda_k - M\right). \tag{23}$$

In this formulation, we do not require λ_k to be integers. One should therefore visualize M as the total mass of representatives, which is divided between the natural clusters.

The set of representatives $\{y_j\}$ should satisfy

$$\frac{\partial}{\partial y_j} F' = 0. \tag{24}$$

Since the constraint is independent of y_j, this again yields (11), i.e.,

$$\sum_x P(x \in C_j) \frac{\partial}{\partial y_j} d(x, y_j) = 0 \tag{25}$$

with the distinction that now, the association probabilities are according to (21).

On the other hand, the corresponding set $\{\lambda_k\}$, which minimizes F', satisfies

$$\frac{\partial}{\partial \lambda_j} F' = -\frac{1}{\beta} \sum_x \frac{e^{-\beta d(x, y_j)}}{Z_x} + q = 0 \tag{26}$$

which yields

$$q\beta = \sum_x \frac{e^{-\beta d(x, y_j)}}{Z_x}. \tag{27}$$

Multiplying by the appropriate λ_k and summing over all natural clusters, we get

$$\sum_k \lambda_k q \beta = \sum_k \lambda_k \sum_x \frac{e^{-\beta d(x, y_j)}}{Z_x} \tag{28}$$

which by applying our total mass constraint, and (20), yields

$$q\beta = \frac{N}{M} \tag{29}$$

where N is the total number of data points in the training set. Substituting (29) in (27), we see that the optimal set of λ_k must satisfy

$$\sum_x \frac{e^{-\beta d(x, y_j)}}{Z_x} = \frac{N}{M} \tag{30}$$

where the λ_k are implicit in Z_x (20). Equation (30) is thus the equation we solve while optimizing over $\{\lambda_k\}$.

Authorized licensed use limited to: UNIVERSIDADE FEDERAL DO RIO DE JANEIRO. Downloaded on April 28,2010 at 11:00:12 UTC from IEEE Xplore. Restrictions apply.

Moreover, by using (30) and (21), we obtain

$$\lambda_j = \frac{M}{N}\sum_x \frac{\lambda_j e^{-\beta d(x,y_j)}}{Z_x} = \sum_x \mu P(x \in C_j) \tag{31}$$

where $\mu = M/N$ is the mass of one data point. This is intuitively appealing because the optimal representatives partition is, in fact, the training data set mass partition in the clusters. It also makes explicit the relation between this formulation at a given β and maximum likelihood estimation of parameters in mixtures, where the class prior probabilities are unknown (see [5]). Our formulation is more general and does not necessarily assume *a priori* knowledge on, or modeling of, the data distribution. The association probabilities are derived from the distortion function, which is not necessarily related to assumptions on the data distribution (for example, vector quantization in data compression).

Note that although μ is constant above, it could be made to depend on x to generalize the method to the case where the given data points are not equally important. In particular, this could apply to clustering of gray-scale images that are low-resolution representations of high-resolution binary sets. In other words, this enables a direct multiscale implementation of the method.

In the mass-constrained formulation, the process is independent of the number of representatives (as long as it is greater than the number of natural clusters). In order to see this, let the natural clusters be represented by $\{y_j\}$ and $\{\lambda_j\}$, which are the solution sets of centroids and masses, respectively. Now, let us raise the number of representatives and consider the case where the jth natural cluster is represented by m_j representatives $y_j^{(n)}$, whereas the cluster's mass is arbitrarily divided between them, i.e.

$$y_j^{(n)} = y_j. \qquad n = 1. \dots. m_j \tag{32}$$

$$\sum_{n=1}^{m_j} \lambda_j^{(n)} = \lambda_j. \tag{33}$$

By (20), Z_x is invariant to any such division. Furthermore, the probability of association with the natural cluster is unchanged as

$$\begin{aligned}\sum_n P(x \in C_j^{(n)}) = \sum_n \frac{\lambda_j^{(n)} e^{-\beta d(x,y_j)}}{Z_x} = \frac{\lambda_j e^{-\beta d(x,y_j)}}{Z_x} \\ = P(x \in C_j).\end{aligned} \tag{34}$$

It is therefore clear that the same representative locations will satisfy (11) and will thus be obtained by our method, regardless of the multiplicity m_j or the mass division $\{\lambda_j^{(n)}\}$.

It should also be noted that at the limit of low temperature ($\beta \rightarrow \infty$), both the nonconstrained method and the mass-constrained method converge to the same descent process, namely, LBG [4] (or basic ISODATA [1] for the sum of squared distances). This is so because the association probabilities in these deterministic annealing methods become identical at the limit and associate each data point to the nearest representative with probability one. The difference between the two is in their behavior at intermediate β, where the mass-constrained clustering method takes the cluster populations into account (Fig. 1).

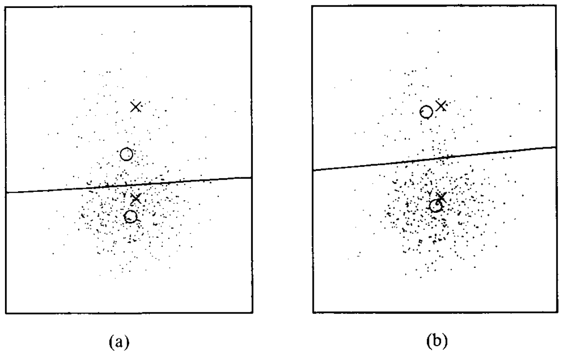


Fig. 1. Effect of cluster mass (population) at intermediate β. The data is sampled from two normal distributions whose centers are marked by X. The computed representatives are marked by O: (a) Nonconstrained clustering; (b) mass-constrained clustering.

To further illustrate the performance of mass-constrained clustering, we consider the example shown in Fig. 2. This is a mixture of six Gaussian densities of different masses and variances. We compare the result of our method with the well-known basic-ISODATA method. Since basic-ISODATA yields results that depend on the initialization, we have run it 25 times, each time with a different initial set of representatives (randomly extracted from the training set). In Fig. 2(a), we show the best result obtained by basic-ISODATA, where the mean squared error is 6.4. This result was obtained only once, whereas for $\approx 80\%$ of the runs, it got trapped in local optima of $\approx$ 12.5 MSE. In Fig. 2(b), we show the result obtained by our method. The MSE is 5.7, and this solution is, of course, independent of the initialization. The process of "annealing" is illustrated in Fig. 3. Here, we have a mixture of nine overlapping Gaussian densities. The process undergoes a sequence of phase transitions as β is increased. We show the results at some of the phases. Equiprobability contours are used to emphasize the fuzzy nature of the results at intermediate β. At the limit of high β, the MSE is 32.854. Repeated runs of basic-ISODATA on this example yielded a variety of local optima with MSE from 33.5 to 40.3.

V. The Traveling Salesman Problem (TSP)

In the deterministic annealing clustering algorithm, if we throw in enough representatives and let $\beta \rightarrow \infty$, then each data point will become a natural cluster. This can be viewed as a process of data association, where each data point is exclusively associated with a natural representative. As it stands, there is no preference as to which representative is associated with which data point. However, by adding a constraint, we can encourage the process to obtain associations thta would satisfy additional requirements. As an example, the EN approach to TSP [16], [18]–[20] is considered here.

The problem statement is stated as follows: Given a set of data points (usually called cities), find the shortest closed path that passes through all of them. In order to derive the EN

Authorized licensed use limited to: UNIVERSIDADE FEDERAL DO RIO DE JANEIRO. Downloaded on April 28,2010 at 11:00:12 UTC from IEEE Xplore. Restrictions apply.

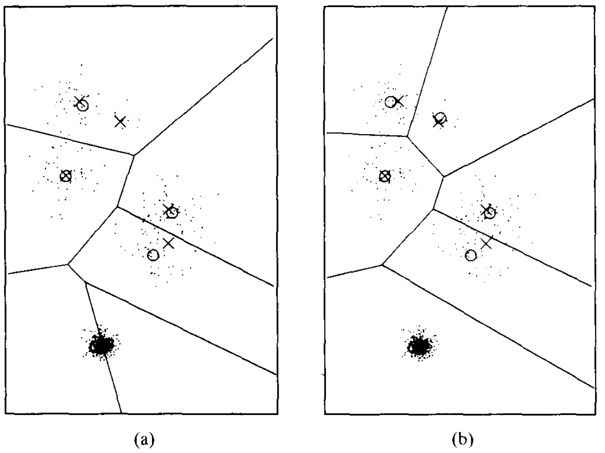


Fig. 2. Basic-ISODATA versus mass-constrained clustering: (a) Best result of basic-ISODATA out of 25 runs with random initialization: MSE=6.4; (b) mass-constrained clustering: MSE=5.7.

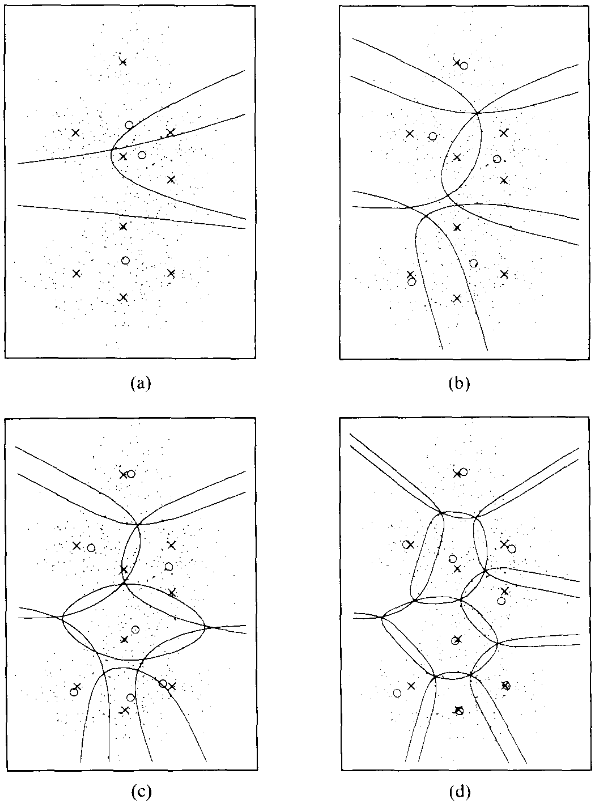


Fig. 3. Various phases in the annealing process. We use contours of constant association probability (p) to emphasize the fuzzy nature of the results at intermediate β: (a) Three clusters at $\beta = 0.005$ contours at $p = 0.45$; (b) five clusters at $\beta = 0.009$, $p = 0.33$; (c) seven clusters at $\beta = 0.013$, $p = 0.33$; (d) nine clusters at $\beta = 0.03$, $p = 0.33$.

method, we shall assume that the sum of squared distances between consecutive cities on the path is to be minimized, as is, in fact, done in [16]. This will be loosely referred to as "tour length." The basic optimization problem to solve is that of minimizing F subject to the constraint of a given tour length. Controlling the mean squared distance from the cities via β, as well controlling the required tour length, will be the essence of the annealing process. Hence, we add the appropriate constraint to the free energy to obtain the Lagrangian (19)

$$F' = F + \lambda\left(\sum_{k=1}^{N} |y_k - y_{k-1}|^2 - L\right), \qquad y_o = y_N. \quad (35)$$

Here, L is the tour length, and λ is the Lagrange multiplier related to it.

As an aside, note that we could start by defining a modified instance distortion (see [19])

$$D'(Y, V) = \sum_x \sum_k v_{xk} d(x, y_k) + \lambda \sum_k |y_k - y_{k-1}|^2$$

instead of D, as defined in (6). Then, deriving the effective cost similarly to the derivation of (5) to (10), we would get F' of (35) instead of F of (10) (except for a term that does not depend on the representatives) as the function to be minimized to obtain the most probable set Y. This approach, however, gives λ the interpretation of a coefficient weighing the relative importance of the second term in the instance distortion. This obscures the annealing role it should have, which will be explained in the sequel.

The optimal set Y must satisfy the condition

$$\frac{\partial}{\partial y_j} F' = 0 \qquad \forall j \quad (36)$$

which by substituting (35) yields

$$\sum_x P(x \in C_j) \frac{\partial}{\partial y_j} d(x, y_j) + 2\lambda(2y_j - y_{j-1} - y_{j+1}) = 0. \quad (37)$$

If we also choose the squared distance as our clustering distortion measure $d(x, y)$, then we obtain an EN formulation for the optimum

$$\sum_x P(x \in C_j)(y_j - x) + \lambda(2y_j - y_{j-1} - y_{j+1}) = 0. \quad (38)$$

Note that this equation depends on β through the association probabilities (2). As we have seen in the clustering method derivation, β controls the mean squared distance to the cities. By making $\beta \to \infty$, we make each representative converge to a city. The second Lagrange multiplier λ is related to the tour length.

An important question at this point is whether and how λ should be varied with β. In [16], the formulation implies $\lambda \propto 1/\sqrt{\beta}$, whereas in [19], it seems to be kept constant. It is instructive to first consider the tour length L (instead of λ) as the control parameter. Obviously, for small β, the representatives are close to the center of mass of the distribution, and the tour length is small. As β is increased, so, normally, is the tour length. If we do not constrain the length, then we obtain our clustering solution for each β. By constraining the tour length to be shorter than the free tour length, we maintain some "tension" in the elastic net. This is

Authorized licensed use limited to: UNIVERSIDADE FEDERAL DO RIO DE JANEIRO. Downloaded on April 28,2010 at 11:00:12 UTC from IEEE Xplore. Restrictions apply.

particularly important at the vicinity of phase transitions where separating representatives should be ordered to minimize the length.

The procedure suggested here is as follows: i) At a given β, gradually increase L and optimize until L reaches some appropriate value below the free tour length. ii) *Keeping L constant*, update β and optimize; return to i). Such an approach can be implemented directly using methods for nonlinear optimization; for example, one may consider using the generalized Hopfield network [21]. It is, however, more convenient and simpler to control the Lagrange multiplier λ rather than the tour length L directly.

Our problem at given β and L is to minimize $F(Y)$ subject to a constraint that we shall conveniently write as $h(Y) = L$. A necessary condition for an optimum is that the derivatives of the Lagrangian vanish, i.e.

$$\frac{\partial F}{\partial y_j} + \lambda\frac{\partial h}{\partial y_j} = 0 \qquad \forall j. \tag{39}$$

Now, let (Y^*, λ^*) be the optimum, and let F^* be the free energy at the optimum

$$F^* = F'(Y^*, \lambda^*) = F(Y^*).$$

It can be shown (for example, [22]) that for such constrained optimization

$$\lambda^* = -\frac{\partial F^*}{\partial L}. \tag{40}$$

This gives our Lagrange multiplier the interpretation of the rate of decrease of the minimal free energy with respect to increase in the tour length. Clearly, for $\lambda^* = 0$, we get the nonconstrained clustering solution and the free tour length. Our suggested procedure can thus be controlled as follows. At a given β, gradually decrease λ and optimize until a small positive value λ_{min} that maintains some "tension" in the net is reached. The next step is to update β and simultaneously find a new initial value for λ so that the tour length at the optimum is kept constant. Then again, λ is gradually decreased to λ_{min}, etc.

The next step in our derivation is therefore to determine an initial value for λ when updating β such that it will keep L constant. For this purpose, let us compute the following partial derivative, given that L is constant.

$$\frac{\partial \lambda^*}{\partial \beta} = -\frac{\partial}{\partial \beta}(\frac{\partial F^*}{\partial L}) = -\frac{\partial}{\partial L}(\frac{\partial F^*}{\partial \beta}) \tag{41}$$

where use was made of (40). Next, we note that

$$\frac{\partial F^*}{\partial \beta} = \frac{\partial F}{\partial \beta}(Y^*, \beta) + \sum_k \frac{\partial F}{\partial y_k}(Y^*.\beta)\frac{\partial y_k^*}{\partial \beta}. \tag{42}$$

Differentiating the constraint, we obtain

$$\sum_k \frac{\partial h}{\partial y_k}\frac{\partial y_k^*}{\partial \beta} = \frac{\partial L}{\partial \beta} = 0 \tag{43}$$

where 0 is obtained by the constant L assumption. Adding λ^*·(43) to (42), we get

$$\frac{\partial F^*}{\partial \beta} = \frac{\partial F}{\partial \beta}(Y^*, \beta) + \sum_k (\frac{\partial F}{\partial y_k} + \lambda^*\frac{\partial h}{\partial y_k})\frac{\partial y_k^*}{\partial \beta}. \tag{44}$$

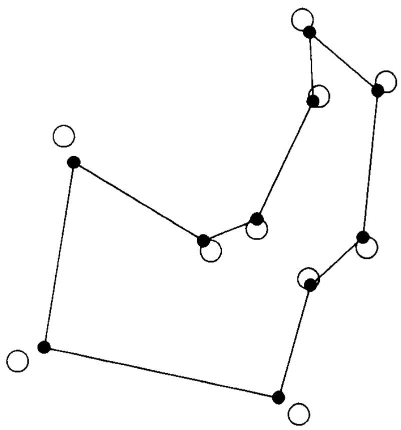


Fig. 4. Ten-cities problem solved by deterministic annealing. This is the optimal tour for both the sum of distances and the sum of squared distances.

By (39), the second term equals 0; therefore, (44) reduces to

$$\frac{\partial F^*}{\partial \beta} = \frac{\partial F}{\partial \beta}(Y^*, \beta). \tag{45}$$

Let us now make the following observation:

$$\frac{\partial}{\partial \beta}(\beta F) = \sum_x \sum_j |y_j - x|^2 P(x \in C_j) = E. \tag{46}$$

Therefore

$$\frac{\partial F}{\partial \beta} = \frac{E - F}{\beta} \tag{47}$$

and by (45), we have

$$\frac{\partial F^*}{\partial \beta} = \frac{E^* - F^*}{\beta}$$

which when substituted into (41) yields

$$\frac{\partial \lambda^*}{\partial \beta} = -\frac{1}{\beta}(\frac{\partial E^*}{\partial L} - \frac{\partial F^*}{\partial L}) = -\frac{1}{\beta}(\frac{\partial E^*}{\partial L} + \lambda^*). \tag{48}$$

In practice, this allows the use of the following approximation:

$$\Delta\lambda^*(\beta) \approx -\frac{\Delta\beta}{\beta}(\frac{\Delta E^*}{\Delta L} + \lambda^*) \tag{49}$$

where $\Delta E^*/\Delta L$ may be estimated using the last two iterations in λ (before the moment to update β arrived).

Fig. 4 shows our result for the ten-cities problem [18], which according to Durbin *et al.* is the optimal solution and is slightly better than the one obtained by them. In this simple example, one can show by exhaustive search that indeed, this path minimizes both the sum of squared distances and the sum of distances.

Fig. 5 shows our result for the first 50-cities problem [16]. Here, the resulting path is longer than the one obtained by Durbin and Willshaw, but the sum of squared distances is smaller, and indeed, this is the quantity that is actually minimized by the method.

More serious investigation of the annealing schedule is needed if one wants to optimize the method. In our simulations, β was increased exponentially, as had been done in

Authorized licensed use limited to: UNIVERSIDADE FEDERAL DO RIO DE JANEIRO. Downloaded on April 28,2010 at 11:00:12 UTC from IEEE Xplore. Restrictions apply.

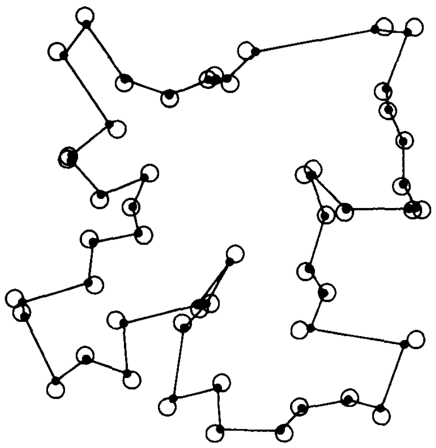


Fig. 5. Deterministic annealing result for the (first) 50-cities problem.

[16]. This is very convenient (note that logarithmic schedules are suggested for stochastic relaxation) but may compromise the results. We have also experimented with annealing while keeping λ constant. For both examples, it was possible to find values for λ empirically, such that the same results were obtained, only this required the annealing schedule to be extremely slow. For example, in the 50-cities problem, this required the rate $\Delta\beta/\beta = 0.0001$, as compared with $\Delta\beta/\beta = .01$ used in the proposed annealing method (the extra computations for the iterations in λ were negligible with respect to this). Moreover, experimentation with various values for constant λ were needed to find the best choice.

VI. Dimensionality Reduction by Self- Organization

Kohonen [17] suggested a sequential procedure for self-organization of neural nets. This procedure tends to make the representatives (neurons) "fit" the probability distribution of the input data while remaining "ordered." It is intuitively obvious that these two objectives fall nicely under "clustering" and "constraint," respectively. We shall refer to this procedure as Kohonen's learning procedure (KLP). KLP allows learning nets of low topological dimensions to deal with inputs of higher dimensions. It can be viewed as defining the net on an optimal hypersurface within the multidimensional input space. In other words, it is a nonlinear projection of a multidimensional data set onto a discretized space of lower dimensionality. Finding the optimal nonlinear projection is commonly called the dimensionality reduction problem (see [17] for a detailed treatment of the subject).

There exist vector quantization methods that are related to or based on KLP. They basically degenerate KLP to remove the topological (ordering) constraint. As a matter of fact, such algorithms are forms of competitive learning [23]–[26]. Chang and Gray [27] have independently developed a technique called stochastic gradient, which is a special case of KLP. Although it performed slightly better than LBG when parameters were optimized empirically, problems with the step-size adaptation, which is not well understood, led to their conclusion that LBG may be practically preferable. In this section, we attempt to close the circle by suggesting an extention of the deterministic annealing approach to self-organization via an appropriate constraint.

Instead of searching for the shortest closed path, as we have done for solving TSP, one may be looking for the shortest open path, i.e., the shortest way to traverse all cities. It can be shown by reduction that for the sum of squared distances, this problem is at least as hard as TSP. One can formulate this problem as constrained clustering in exactly the same way as we formulated TSP, except that the periodic boundary condition is removed from the constraint. Instead of (35), we now have

$$F' = F + \lambda\left(\sum_{k=2}^{N} |y_k - y_{k-1}|^2 - L\right). \tag{50}$$

If the data is 1-D and the number of representatives equals the number of data points, then what we get at the limit is a deterministic annealing method for sorting since the shortest open path is simply the ordered sequence of data points. Of course, deterministic annealing is not suggested as a practical method for sorting, but its use on the sorting problem does give a useful intuition for dealing with ordering in higher dimensions. As a vector equation, (50) deals with unsupervised learning of a linear network (linear topology) but can be extended to networks of higher topologies simply by defining the corresponding neighborhoods and adding the appropriate distances to the summation in the constraint.

Since all the derivation of the annealing procedure for TSP in the previous section (39)–(49) was, in fact, for a general constraint (denoted $h(Y)$), we can use the results directly without repeating the underlying mathematics. In particular, it is obvious that the Lagrange multiplier λ has a similar meaning here, and the same annealing procedure is applicable in this case as well.

Fig. 6 shows an example of the self-organization of a ten-unit linear network, given the same 50-cities example we used for TSP. The results show a behavior similar to that of the stochastic method documented in [17] (the differences are discussed below). This demonstrates how a linear network tries to cope with 2-D input. Thus, dimensionality reduction from 2-D data to a 1-D representation is obtained. The biological plausibility of a stochastic version of such self-organization in cortical maps is discussed in [28]. In [19], it is suggested how to obtain unique matching at the nonfuzzy limit by appropriately modifying the cost function. However, it seems that unsupervised learning belongs to the category of *fuzzy* association problems because the objective is to generalize from a training set and estimate parameters of clusters.

The distinctions between KLP and the deterministic annealing method proposed here are mainly as follows. KLP is sequential (stepwise) and, thus, may enable adaptation to nonstationary data. It may also be more biologically plausible. On the other hand, it suffers from the disadvantages of sequential algorithms. In particular, convergence in nontrivial cases is difficult to analyze, and step-size adaptation schemes are typically heuristic. Moreover, the results may depend on the order of presentation of data points. A major distinction to

Authorized licensed use limited to: UNIVERSIDADE FEDERAL DO RIO DE JANEIRO. Downloaded on April 28,2010 at 11:00:12 UTC from IEEE Xplore. Restrictions apply.

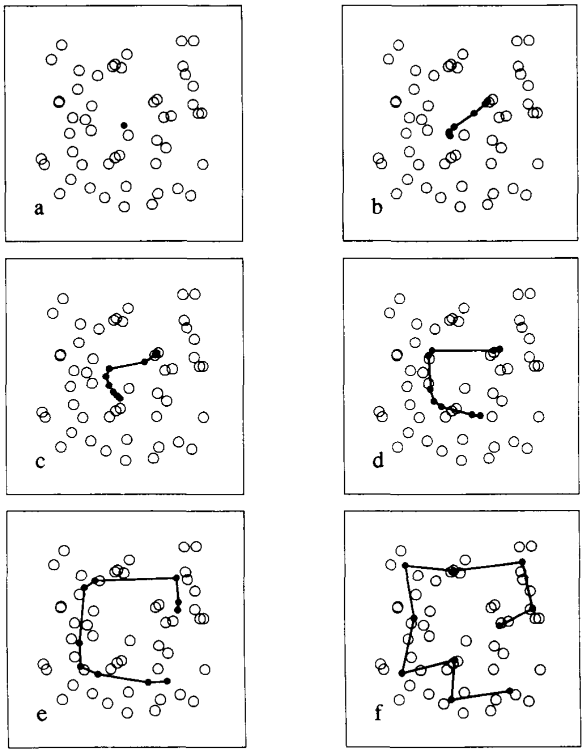


Fig. 6. Self-organization of a linear network of ten units, given the 50-cities problem data: (a) $\beta = 0$; (b) $\beta = 0.064$; (c) $\beta = 0.067$; (d) $\beta = 0.08$; (e) $\beta = 0.2$; (f) $\beta = 5$.

keep in mind is that if KLP converges to a local minimum, it is *only at the limit.* Intermediate results are stochastic and are not locally optimal. Deterministic annealing, on the other hand, converges to a local minimum of the Lagrangian *at each* β and thus yields reliable fuzzy solutions at intermediate β. This is obtained within the above mathematical framework, where the process virtually always stays at a "statistical equilibrium" of its stochastic counterpart. We strongly believe that fuzzy solutions are important in these applications for two reasons: the need to generalize from a given training set and the need to estimate cluster parameters correctly.

Given the known advantages of "batch" algorithms over sequential algorithms in the field of clustering, as well as the inherent fuzzy nature of the problem, we believe that deterministic annealing shows promise for self-organization, given a finite training set. Pursuit of this approach to unsupervised learning is intended in a future study. Here, we were mainly interested in this subject as an example for a constrained-clustering application.

VII. Concluding Remarks

Constrained clustering can be used to improve the deterministic annealing method for clustering. Moreover, it allows applying annealing to various other optimization problems that associate a data set with a set of variables and can be formulated as constrained clustering. Annealing is obtained by gradually making the associations less fuzzy and helps avoiding local minima.

Finally, let us note that this probabilistic clustering that brings about the annealing in the process can indeed be viewed as some special form of averaging over the data set. Therefore, at high temperatures, we get a "low-resolution" solution that evolves into the "high-resolution" solution as the temperature is lowered. This implies that for problems that deal with associating a small subset of the data (e.g., navigation-related problems where the solution involves only points on the path and not all the terrain data), annealing may not be useful unless there is a significant multiscale structure in the *data* so that low-resolution solutions contain information leading to good high-resolution solutions.

References

[1] G. Ball and D. Hall, "A clustering technique for summarizing multivariate data," *Behavioral Sci.*, vol. 12, pp. 153–155, 1967.

[2] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *J. Cybern.*, vol. 3, pp. 32–57, 1974.

[3] J. C. Bezdek, "A convergence theorem for the fuzzy ISODATA clustering algorithms," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. PAMI-2, pp. 1–8, 1980.

[4] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84–95, 1980.

[5] R. O. Duda and P. E. Hart, *Patt. Classification Scene Anal.* New York: Wiley-Interscience, 1974.

[6] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data.* Englewood Cliffs, NJ: Prentice Hall, 1988.

[7] M. Anderberg, *Cluster Analysis for Applications.* New York: Academic, 1973.

[8] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Sci.*, vol. 220, pp. 671–680, 1983.

[9] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. PAMI-6, pp. 721–741, 1984.

[10] P. J. M. van Laarhoven and E. H. L. Aarts, *Simulated Annealing: Theory and Applications.* Dordrecht, The Netherlands: D. Reidel, 1987.

[11] D. Geman, S. Geman, C. Graffigne, and P. Dong, "Boundary detection by constrained optimization," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. 12, pp. 609–628, 1990.

[12] K. Rose, E. Gurewitz, and G. C. Fox, "Statistical mechanics and phase transitions in clustering," *Phys. Rev. Lett.*, vol. 65, pp. 945–948, 1990.

[13] ——, "Vector quantization by deterministic annealing," *IEEE Trans. Inform. Theory*, vol. 38, no. 4, pp. 1249–1257, July 1992.

[14] E. T. Jaynes, "Information theory and statistical mechanics," in *Papers on Probability, Statistics and Statistical Physics* (R. D. Rosenkrantz, Ed.). Dordrecht, The Netherlands: Kluwer, 1989.

[15] C. Shannon and W. Weaver, *The Mathematical Theory of Communication.* Urbana, IL: Univ. of Illinois Press, 1949.

[16] R. Durbin and D. Willshaw, "An analogue approach to the travelling salesman problem using an elastic net method," *Nature*, vol. 326, pp. 689–691, 1987.

[17] T. Kohonen, *Self Organization and Associative Memory.* Berlin: Springer-Verlag, 1984.

[18] R. Durbin, R. Szeliski, and A. Yuille, "An analysis of the elastic net approach to the travelling salesman problem," *Neural Comput.*, vol. 1, pp. 348–358, 1989.

[19] A. L. Yuille, "Generalized deformable models, statistical physics, and matching problems," *Neural Comput.*, vol. 2, pp. 1–24, 1990.

[20] P. D. Simic, "Statistical mechanics as the underlying theory of elastic and neural optimization," *Network*, vol. 1, pp. 89–103, 1990.

[21] A. G. Tsirukis, G. V. Reklaitis, and M. F. Tenorio, "Nonlinear optimization using generalized Hopfield networks," *Neural Comput.*, vol. 1, pp. 511–521, 1989.

[22] G. V. Reklaitis, A. Ravindran, and K. M. Ragsdell, *Eng. Optimization.* New York: Wiley-Interscience, 1983.

[23] S. Grossberg, "Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors," *Biolog. Cybern.*, vol. 23, pp. 121–134, 1976.

Authorized licensed use limited to: UNIVERSIDADE FEDERAL DO RIO DE JANEIRO. Downloaded on April 28,2010 at 11:00:12 UTC from IEEE Xplore. Restrictions apply.

[24] S. Grossberg, "Adaptive pattern classification and universal recoding: II. Feedback, expectation, olfaction, illusions," *Biolog. Cybern.*, vol. 23, pp. 187–202, 1976.

[25] D. E. Rumelhart and D. Zipser, "Feature discovery by competitive learning," *Cognitive Sci.*, vol. 9, pp. 75–112, 1985.

[26] S. Grossberg, "Competitive learning: From interactive activation to adaptive resonance," *Cognitive Sci.*, vol. 11, pp. 23–63, 1987.

[27] P. -C. Chang and R. M. Gray, "Gradient algorithms for designing predictive vector quantizers," *IEEE Trans. Acoustics Speech Signal Processing*, vol. ASSP-34, pp. 679–690, 1986.

[28] R. Durbin and G. Mitchison, "A dimension reduction framework for understanding cortical maps," *Nature*, vol. 343, pp. 644–647, 1990.

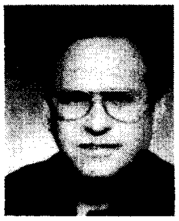
Kenneth Rose (S'85–M'91) received the B.Sc. (summa cum laude) and M.Sc. degrees in electrical and electronics engineering from Tel-Aviv University, Israel, in 1983 and 1987, respectively, and the Ph.D. degree in electrical engineering from the California Institute of Technology (Caltech), Pasadena, in 1990.

From July 1983 to July 1988, he was employed by Tadiran, Ltd., Israel, where he carried out research in the areas of image coding, image transmission through noisy channels, and general image processing. From September 1988 to December 1990, he was a graduate student at Caltech. In January 1991, he joined the Department of Electrical and Computer Engineering, University of California at Santa Barbara, as an assistant professor. His research interests are in pattern recognition, information theory, neural networks, image coding and processing, and nonconvex optimization in general.

Dr. Rose received the 1990 William R. Bennett Prize Paper Award of the IEEE Communications Society for his paper "Enhancement of One-Dimensional Variable-Length DPCM Images Corrupted by Transmission Errors" (with Arie Heiman) in the April 1989 IEEE TRANSACTIONS ON COMMUNICATIONS.

Geoffrey C. Fox received the Ph.D. degree in physics from Cambridge University.

He is a Professor of Physics and Computer Science at Syracuse University, Syracuse, NY, and Director of the Northeast Parallel Architectures Center there. From 1970 to 1990, he held various positions at the California Instsitute of Technology, Pasadena, where the work described in this paper was performed. His interests are computational science education with particular research projects in parallel computing software and physical computation. He is coauthor of the book *Solving Problems on Concurrent Processors* (Englewood Cliffs: Prentice-Hall, vols. 1 and 2).



Eitan Gurewitz received the B.Sc. degree in mathematics and physics from the Hebrew University, Jerusalem, Israel, in 1964 and the M.Sc. and Ph.D. degrees in physics from the Weizmann Institute of Science, Rehovoth, Israel, in 1966 and 1976, respectively.

Since 1967, he has been at the Nuclear Research Center–Negev, Beer Sheva, Israel, doing research on neutron scattering, magnetic structures group theory, and phase transitions. During the past five years, his research has been concentrated in the areas of computational physics, optimization, parallel processing, and image processing.

Dr. Gurewitz is a member of the Israel Physical Society.

Authorized licensed use limited to: UNIVERSIDADE FEDERAL DO RIO DE JANEIRO. Downloaded on April 28,2010 at 11:00:12 UTC from IEEE Xplore. Restrictions apply.