Deterministic Annealing, Clustering, and Optimization

Kenneth Rose

CRPC-TR91114
1991

Center for Research on Parallel Computation
Rice University
P.O. Box 1892
Houston, TX 77251-1892

SCCS - 32
C^3P - 950
CRPC - TR91114

"Deterministic Annealing, Clustering, and Optimization"

Thesis By:

Kenneth Rose

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

1991
(Defended December 6, 1990)

Syracuse Center for Computational Science
Syracuse University
111 College Place
Syracuse, New York 13244-4100
<sccs@npac.syr.edu>
(315) 443-1723

DETERMINISTIC ANNEALING, CLUSTERING, AND OPTIMIZATION

Thesis by
Kenneth Rose

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

1991
(Defended December 6, 1990)

If a man will begin with certainties,
he shall end in doubts;
but if he will be content to begin with doubts,
he shall end in certainties.
FRANCIS BACON,
Advancement of Learning

Acknowledgements

It is my great pleasure to thank my adviser Professor Geoffrey C. Fox for his help, support, and for his gift of seeing farther by refusing to be caught in details. I am deeply indebted to Professor Edward C. Posner for agreeing to be my "surrogate" adviser, for his help, and for many valuable suggestions and enjoyable discussions. I cannot overstate the contribution of Dr. Eitan Gurewitz, with whom I had almost daily discussions on this work, as well as on many other subjects of interest. I thank him for his contribution, inspiration, and friendship. I wish to acknowledge some basic comments made by Professor Robert J. McEliece at an early stage of this work, which triggered shifting the presentation from statistical physics analogies to information theory. My appreciation goes to Professors Yaser S. Abu-Mostafa, Alexander S. Kechris, Robert J. McEliece, Edward C. Posner, and P. P. Vaidyanathan for serving on my oral examination committee.

I thank my parents and all those who believed and helped me grow up and become myself. My very very special thanks go to my wife Monique, and to my daughters Emmanuelle and Armelle, who had to make all the sacrifices related to being dependents of a foreign graduate student, and still managed to have some fun.

Abstract

This work introduces the concept of deterministic annealing (DA) as a useful approach to clustering and other related optimization problems. It is strongly motivated by analogies to statistical physics, but is formally derived within information theory and probability theory. This approach enables escaping local optima that plague traditional techniques, without the extremely slow schedules typically required by stochastic methods. The clustering solutions obtained by DA are totally independent of the choice of initial configuration.

A probabilistic framework is constructed, which is based on the principle of maximum entropy. The association probabilities at a given average distortion are Gibbs distributions parametrized by the corresponding Lagrange multiplier β, which is inversely proportional to the temperature in the physical analogy. By computing marginal probabilities within the framework, an effective cost is obtained, which is minimized to find the most probable set of cluster representatives at a given temperature. This effective cost is the free energy in statistical mechanics, which is indeed optimized at isothermal, stochastic equilibrium.

Within the probabilistic framework, annealing is introduced by controlling the Lagrange multiplier β. This annealing is interpreted as gradually reducing the "fuzziness" of the associations. Phase transitions are identified in the process, which are, in fact, cluster splits. A sequence of phase transitions produces a hierarchy of fuzzy-clustering solutions. Critical β are computed exactly for the first phase transition and approximately for the following ones.

Specific algorithms are derivable from the general approach, to address different aspects of clustering in the large variety of application fields. Here, algorithms are derived, and simulation results are presented for the three major classes, namely, hard clustering, fuzzy clustering, and hierarchical clustering. From the experimental results it appears that DA is substantially superior to traditional techniques.

The last part of the work extends the approach to deal with a larger family of optimization problems that can be reformulated as constrained clustering. A probabilistic framework for constrained clustering is derived based on the principle of maximum entropy. It is shown that for our annealing purpose, the constraint can be directly applied to the free energy. Three examples of constrained clustering are discussed. Mass-constrained clustering is formulated and yields an improvement of the clustering procedure. The process is now independent of the number of representatives and their multiplicity in the clusters. Secondly, the travelling salesman problem (TSP) is reformulated as constrained clustering, yielding the elastic net approach. A second Lagrange multiplier is identified, which is used to obtain a more powerful annealing method. Finally, self-organization of neural networks is shown to be closely related to TSP, and a similar annealing method is suggested. A fuzzy solution is sought to obtain the optimal net for a given training data set.

Contents

List of Figures

Chapter 1

Introduction

1.1 What is Clustering?

Clustering can be informally stated as partitioning a given set of data points into subgroups, each of which should be as homogeneous as possible. The problem of clustering is an important optimization problem in a large variety of fields, such as pattern recognition, learning, source coding, image and signal processing. The exact definition of the clustering problem differs slightly from field to field, but in all of them it is a major tool for the analysis or processing of data without knowledge of a priori distributions. In pattern recognition, clustering helps the discerning of the underlying structure of the distribution, and identifying natural classes or components of the data [6][20]. In learning, clustering is usually related to unsupervised learning, where it groups together input data, producing a smaller number of representatives [25] which may then be dealt with by a supervised learning scheme [6]. In source coding, clustering is mainly used for vector quantization, or minimum distortion representation of the data by a small number of quantization levels [12][13] (for a review of vector quantization for image coding see [30]). In image and signal processing there are many diverse applications of clustering, from the obvious image segmentation [19][24][42][23],

to adaptation to nonstationary signals via approximation by piecewise stationary signals, which requires appropriate partitioning of the signals (for example, [38]).

The clustering-problem statement is usually made mathematically precise by defining a cost (energy) criterion to be minimized. The actual criterion is chosen according to the application. An important example, which is by far the most extensively used, is the sum of squared distances cost,

$$\sum_j \sum_{x \in C_j} |x - y_j|^2,$$

where C_j is the j-th cluster, represented by the vector y_j. Virtually all cost functions are *not* convex, and have several local minima [14]. Thus, clustering is a nonconvex optimization problem. While exhaustive search will find the global minimum, it is extremely impractical for all nontrivial and reasonably large data sets.

1.2 Traditional Methods

As clustering is important in many fields, methods for clustering have been suggested in different disciplines. In the communications or information-theory literature, an early clustering method was suggested for scalar quantization, which is known as the Lloyd algorithm [27] or the Max quantizer [28]. This method was later generalized to vector quantization, and to a large family of distortion functions [26]. The resulting algorithm is commonly referred to as the generalized Lloyd algorithm (GLA), or the Linde-Buzo-Gray algorithm (LBG). In the pattern-recognition literature, the ISODATA algorithm [1] was proposed, which is also known in its sequential version as the k-means algorithm. Later, fuzzy relatives to these algorithms were derived [7][2].

The main principles in the above methods are identical. Let us describe the basic nonfuzzy algorithm, using the terminology of LBG. We shall assume at this point that the number of clusters is given. Let $d(x, y)$ be a distortion measure,

i.e., the distortion introduced by representing the vector x by the vector y. For example, the squared distance distortion measure is

$$d(x,y) = |x-y|^2.$$

The total distortion for a given distortion measure is

$$D = \sum_j \sum_{x \in C_j} d(x, y_j). \tag{1.1}$$

Basic Nonfuzzy Algorithm

1. Select an initial configuration $Y^o = \{y_j^o\}$, $n = 0$.

2. Classify each point with the "nearest" representative, i.e.,

$$x \in C_j \qquad \text{if} \qquad d(x, y_j^n) \leq d(x, y_k^n) \qquad \forall k.$$

3. Compute new representatives satisfying the "centroid condition"

$$\sum_{x \in C_j} d(x, y_j^{n+1}) \leq \sum_{x \in C_j} d(x, y) \qquad \forall y.$$

4. Check convergence to stop.

5. Increment n; go to 2.

Convergence to a local minimum can be easily proved. Each iteration consists of step 2 (the nearest neighbor rule) and step 3 (the centroid rule). Both steps can only decrease the total distortion (1.1), or not change it at all. If the distortion is unchanged by the iteration, then the process has converged. Since the distortion is decreased at each step, and there are a finite number of possible partitions, the process is convergent in finite time.

However, the resulting local minimum is not necessarily the global minimum. In fact, the result depends directly on the choice of initial configuration. Many heuristic additions to the basic algorithm, as suggested in the literature, relate

to this fundamental shortcoming, either directly or indirectly. The distortion at local minima, where traditional methods get trapped, may be considerably higher than the distortion at the global minimum, as will be shown in the simulation results section.

Fuzzy relatives of the previous methods have been suggested [7][2] to overcome problems of overlapping clusters. The partition of the overlapping region by a decision boundary distorts the true form of the underlying cluster, and shifts the representatives away from the true centroid of the distribution. In cluster analysis applications these are exactly the parameters one wants to estimate.

Fuzzy methods define clusters as fuzzy sets; i.e., each data point has partial membership in different clusters. The fuzzy distortion is

$$D_f = \sum_j \sum_x u_{xj}^q d(x, y_j),$$

where u_{xj} is the partial membership of data point x in cluster C_j, and

$$\sum_j u_{xj} = 1.$$

The fuzzy algorithm consists of optimizing over the representatives *and the partial membership distributions.* If the parameter q took the value 1, then D_f could be interpreted as an average distortion, where the contribution of each data point is weighted according to its partial membership in the cluster. However, in this case the solution will always be nonfuzzy, as it will be advantageous to assign each point fully to the nearest representative. In order to enforce fuzzy solutions, q takes values greater than one, and is normally viewed as controlling the "fuzziness" of the solution. Two objections can be made at this point. First, the distortion function modification is arbitrary and is not directly justified in terms of the application. Secondly, the cost lost its appealing interpretation as average distortion, as the weights no longer sum up to one.

Fuzzy clustering methods can also be shown to converge to a local minimum [2][43], and also suffer from the plague of nonglobal minima.

To summarize, we briefly reviewed some basic traditional clustering techniques and identified a fundamental shortcoming, namely, the tendency to be trapped in local minima, and the dependence on the choice of initial configuration. We next consider more recent methods that offer means to escape local minima.

1.3 Stochastic Relaxation

The observation of annealing processes in physical chemistry motivates the use of similar concepts to avoid local minima of the distortion. Certain chemical systems can be driven to their low-energy states by annealing, which is a gradual reduction of temperature, spending a long time at the vicinity of the phase transition points. In the corresponding probabilistic framework, a Gibbs distribution is defined over the set of all possible configurations, and assigns higher probability to configurations of lower energy. This distribution is parametrized by the temperature, and as the temperature is lowered, it becomes more discriminating (concentrates most of the probability in a smaller subset of low-energy configurations). At the limit of low temperature it assigns nonzero probability only to global-minimum configurations.

A known technique for nonconvex optimization is stochastic relaxation or simulated annealing [22][47], based on the Metropolis algorithm [29] for atomic simulations. A sequence of random moves are generated and the decision to accept a move depends on the probability of the resulting configuration. More specifically, given a cost (energy) function $E(v)$, define the Gibbs distribution over the set of states $\{v_j\}$ as

$$P(v_j) = \frac{e^{-\beta E(v_j)}}{\sum_k e^{-\beta E(v_k)}}.$$

Obviously, states of lower energy have higher probability, and as β is increased, more probability is concentrated at a smaller subset of low-energy states. Now, let $v^{(n)}$ denote the state at the n'th iteration. By a random step we generate a

new state w, and let

$$q = \frac{P(w)}{P(v^{(n)})} = e^{-\beta \Delta E}$$

control the probability of accepting w. The decision is random according to

$$P(v^{(n+1)} = w) = \min(q, 1).$$

Thus, if w is of lower energy, it is accepted, while if it is of higher energy, it is accepted in probability q. Note that for finite β there is always some positive probability of escaping a local minimum. Annealing is obtained by gradually increasing β.

This is a powerful approach, and has been tried with many nonconvex optimization problems, including vector quantization [3][49]. However, one must be very careful with the annealing schedule, the rate at which the temperature is lowered, as the system has to be kept close to stochastic equilibrium. In their work on image restoration, Geman and Geman [11] have shown that in theory, convergence in probability to the global minimum can be achieved if the schedule obeys $\beta \propto \log n$, where n is the number of the current iteration. Cooling schedules are also analyzed in [18]. Such schedules are not realistic in many (if not most) applications.

1.4 Deterministic Annealing - Motivation

As its name suggests, deterministic annealing tries to enjoy the best of both worlds. On the one hand it is deterministic, meaning that we do not want to be wandering randomly on the energy surface, while making some incremental progress on the average, as is the case for stochastic relaxation. On the other hand, it is still an annealing method and aims at the global minimum, instead of going directly to a near local minimum.

Deterministic annealing can be intuitively understood as follows. Instead of making noisy moves on the given energy surface, we "incorporate" the noise into

the energy function. We derive a sequence of effective energy functions that are parametrized by the temperature, the level of the noise. These effective cost functions will be very smooth at low β, where even very large barriers can be easily traversed, and as β is increased they become more ragged and converge to the original energy at $\beta \rightarrow \infty$. In fact, at $\beta = 0$ the cost function will usually be convex, so that the global minimum of this function is easily obtained. Thus the deterministic annealing approach can be viewed as locating the global minimum at $\beta = 0$, and tracking the minimum while gradually increasing β, by using conventional convex optimization methods at each temperature.

This is an intuitive description of the concept of deterministic annealing. The mathematical derivation of the approach is an important part of this work.

1.5 Contributions of This Thesis

The concept of deterministic annealing

A primary contribution of this thesis is the introduction of the concept of deterministic annealing as a useful approach to clustering and other related optimization problems. This approach enables escaping nonglobal minima without the extremely slow schedules typically required in stochastic relaxation.

A probabilistic framework

A probabilistic framework is derived, which is based on principles of information theory. In particular, probability distributions are obtained using the principle of maximum entropy. These are Gibbs distributions, which are parametrized by the temperature. By computing marginal probabilities within this framework, we obtain an effective cost, the free energy. This effective cost is minimized to find the most probable set of representatives at a given temperature.

Annealing, phase transitions and hierarchical clustering

Within the probabilistic framework, the annealing process is obtained by controlling the Lagrange multiplier β. Phase transitions are identified in the process, which in our formulation are interpreted as cluster splits. A sequence of phase transitions produces a hierarchy of fuzzy clustering solutions. Critical temperatures are computed exactly for the first phase transition, and approximately for the following transitions.

Specific algorithms derived from the approach

From the general approach, specific algorithms are derivable in a straightforward manner. Contributions are thus made to virtually all aspects of clustering encountered in the different fields of application. In particular, hard clustering, fuzzy clustering, and hierarchical clustering are explained, and simulation results are presented. Deterministic annealing is shown to be substantially superior to traditional techniques, and is totally independent of the choice of initial configuration.

Extension to a larger family of optimization problems

It is shown that a family of association problems can be formulated as constrained clustering. Such a formulation allows the application of deterministic annealing to solve them. First, the framework for constrained clustering is derived based on the principle of maximum entropy. It is then shown that for our annealing purposes, the constraint can be directly applied to the free energy.

Constrained clustering formulation of three examples

First, the mass-constrained clustering problem is formulated and an improvement of the clustering procedure is obtained. The annealing process is now independent

of the number of representatives and their multiplicity in clusters. Secondly, the travelling salesman problem (TSP) is reformulated as constrained clustering yielding the elastic net (EN) approach. Within our framework a second Lagrange multiplier is identified, which is also used to control the process, resulting in a more powerful annealing method. Finally, self-organization of networks is shown to be closely related to TSP. A similar annealing method is suggested here, within the corresponding constrained clustering formulation. A fuzzy solution is sought to obtain the"optimal" net for a given training set.

1.6 Notes

Fuzzy vs. Probabilistic Framework

In fuzzy-sets theory, the distinction between association probability and partial membership is emphasized. The confusion may arise because both are normalized distributions that sum up to one. The probabilistic framework implies that each element belongs fully to a given set as a realization of an appropriate random variable, with the given underlying association probability. Fuzzy sets, on the other hand, consist of elements that belong to them at different degrees of membership.

In this work, the framework is fundamentally probabilistic. However, fuzzy-sets terminology will be used frequently. The reasons are that sometimes this seems to convey more intuition, and more importantly, this makes explicit the contribution to the field of fuzzy clustering.

To make this terminology mathematically correct, let us specify the meaning of our "fuzzy" terms and their relation to the probabilistic framework. We shall have probability distributions for associating each point with different clusters. These clusters are regular (nonfuzzy) sets whose contents will be the result of the realizations of all the corresponding random variables. It is, nevertheless, very convenient to define fuzzy clusters, where each data point's degree of membership

in the fuzzy cluster is exactly its probability of belonging to the corresponding nonfuzzy cluster. These fuzzy clusters are indeed fuzzy sets by definition. At the nonfuzzy limit, each point belongs fully to a specific cluster. If the associations become fuzzier, each point is less clearly associated with a specific fuzzy cluster. This reflects the informal relation between the "degree of fuzziness" and the degree of uncertainty, which in our probabilistic framework we shall measure by the entropy.

Hereafter, when a fuzzy term is used with respect to a cluster, it should be given the above interpretation of the fuzzy cluster.

References for the material covered

All the material presented in this thesis appears in a number of journal papers (either published or to be published), [32][33][34][35]. They, however, have been edited here to produce a more fluent style, by taking a more tutorial approach, and by eliminating overlapping parts that were necessary to make the papers self-contained.

Chapter 2

A Probabilistic Framework for Clustering

2.1 Association Probabilities by Maximum Entropy

Our point of departure in defining a probabilistic framework for clustering will be that each data point will belong to each cluster *in probability*. This viewpoint is much in the spirit of fuzzy clustering, where each data point has partial membership in clusters (see note in the introduction). The traditional framework for clustering is the marginal special case where all association probabilities are either zero or one. In the pattern recognition literature this is called "hard" clustering, as opposed to the more recent ("soft") fuzzy clustering methods.

Assuming some underlying probability distribution, we may consider the expected distortion (energy)

$$E = \sum_x \sum_j P(x \in C_j)\, d(x, y_j), \tag{2.1}$$

where $d(x, y_j)$ is the distortion 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 have no prior knowledge on the association probabilities, we apply the principle of maximum entropy to estimate them.

The principle of maximum entropy was suggested by Jaynes [21] as a general statistical inference procedure. This principle remains controversial to this very day because of its informal justification (or at best, axiomatic derivation [44]). This controversy has its roots in the intuitive (or axiomatic) equivalence between entropy and uncertainty. Nonetheless, it is undisputed that it is successfully applied in a great variety of fields. We shall regard entropy as the ultimate measure of uncertainty, and shall simply apply entropy maximization to our problem.

The principle of maximum entropy: Of all the probability distributions that satisfy a given set of constraints, choose the one that maximizes the entropy. The informal justification is that while this choice agrees with what is known (the given constraints), it maintains maximum uncertainty with respect to everything else. By choosing another distribution satisfying the constraints, we reduce the uncertainty and therefore implicitly make some extra restrictive assumption.

Hence, to determine the association probabilities at a given expected distortion, we maximize the entropy subject to the constraint (2.1). For the time being we shall make the assumption that the set of representatives $Y = \{y_j\}$ is fixed. This assumption is not realistic as we intend to optimize over this set, and it will be discarded later. For a fixed Y, we make the reasonable assumption that the association probabilities of different data points are independent. The entropy is thus

$$H = -\sum_x \sum_j P(x \in C_j) \log P(x \in C_j). \tag{2.2}$$

As is well known, the probability distributions that maximize the entropy under an expectation constraint are Gibbs distributions. For the constraint (2.1) we get

$$P(x \in C_j) = \frac{e^{-\beta d(x,y_j)}}{Z_x}, \tag{2.3}$$

where Z_x is the partition function

$$Z_x = \sum_k e^{-\beta d(x,y_k)}. \tag{2.4}$$

The parameter β is the Lagrange multiplier determined by the given value of E in (2.1). In our physical analogy β is inversely proportional to the temperature. At this point we can get a first glimpse of the annealing procedure to be described later. Note that decreasing the expected distortion is equivalent to increasing β. This Lagrange multiplier will be conveniently used to control the annealing process, and to drive the distortion down. Also consider the effect of β on the association probabilities. At $\beta = 0$, these are uniform distributions; i.e., each data point is equally associated with all clusters. We have thus extremely fuzzy associations. By increasing β we reduce the fuzziness, as the distribution becomes more discriminating. At $\beta \to \infty$ we get hard classification, and each data point is assigned to the nearest representative with probability one (or more precisely, is uniformly associated with the set of equidistant nearest representatives). This is the condition in which traditional techniques work. It is easy to visualize how the system in this case cannot "sense" a better optimum farther away, as each data point exercises local influence only on the nearest representative. On the other hand, by starting at low β, and slowly increasing it, we start with each data point equally influencing all representatives, and gradually localize the influence. This gives us some intuition as to how the system senses and settles into a better optimum.

2.2 The Free Energy as Effective Cost

Let us now discard our impossible assumption that the representatives set Y is fixed and given, and extend our derivation to include optimization over Y. Instead of considering the association probability of a data point, we consider the probability of an entire instance. An instance of the system is given by a set of

representatives $Y = \{y_j\}$, and a partition via the set of associations $V = \{v_{xj}\}$, where

$$v_{xj} = \begin{cases} 1 & \text{if } x \in C_j; \\ 0 & \text{otherwise.} \end{cases} \tag{2.5}$$

With every instance we associate a distortion

$$D(Y,V) = \sum_x \sum_j v_{xj}\, d(x, y_j), \tag{2.6}$$

which is the distortion of this specific, hard-clustering solution. To estimate the instance probability distribution at a given expected distortion

$$E = \langle D(Y,V) \rangle = \sum_{Y,V} P(Y,V)\, D(Y,V), \tag{2.7}$$

we apply the principle of entropy maximization subject to the constraint (2.7). The resulting distribution is

$$P(Y,V) = \frac{e^{-\beta D(Y,V)}}{\sum_{Y',V'} e^{-\beta D(Y',V')}}. \tag{2.8}$$

Note that the representatives y_j are continuous random variables. We shall avoid going into the ailments of entropy in the continuous case, and shall state that for our purpose Y could have been finely discretized and the integral replaced by summation.

The most probable instance is the one that maximizes the probability in (2.8), i.e., the instance of smallest distortion. This is the result one would seek if one wanted the optimal, hard-clustering solution for *the training set.* However, we may be more interested in estimating the most probable set of representatives, and also in generalizing from the given set of training samples. Consider, therefore, the following marginal probability

$$P(Y) = \sum_V P(Y,V), \tag{2.9}$$

where the summation is performed over all legal association sets. A legal association set V defines a partition, and is such that each data point is assigned to *exactly* one cluster. By using (2.6) and then (2.4), we obtain the identity

$$\sum_V e^{-\beta D(Y,V)} = \prod_x \sum_k e^{-\beta d(x,y_k)} = \prod_x Z_x(Y) = Z(Y). \tag{2.10}$$

$Z(Y)$ will be called here the total partition function for the given representatives set, and is indeed derivable by the independent associations assumption made in the previous section.

The marginal probability of (2.9) now becomes

$$P(Y) = \frac{Z(Y)}{\sum_{Y'} Z(Y')}. \tag{2.11}$$

This can be rewritten in an explicit Gibbs form

$$P(Y) = \frac{e^{-\beta F(Y)}}{\sum_{Y'} e^{-\beta F(Y')}}, \tag{2.12}$$

where

$$F(Y) = -\frac{1}{\beta} \log Z(Y). \tag{2.13}$$

F as defined here is exactly the free energy in our statistical mechanics analogy. Maximizing the marginal probability $P(Y)$ in (2.12) requires minimizing the free energy F. This is therefore our effective cost to be minimized at a given β. Note that the free energy is exactly what is minimized to obtain isothermal equilibrium in statistical physics. Thus, on the one hand we have the distortion D, which is minimized to obtain the optimal, hard-clustering solution for the training set. On the other hand, at a given β, we have an effective distortion, the free energy F, which is minimized to obtain the most probable set of representatives and the optimal "fuzzy" solution. Moreover, for $\beta \to \infty$, both D and F are minimized by the same Y, which is given probability one. This is easily seen by observing that at the limit, $P(Y,V)$ takes nonzero values only at global minimum configurations.

The marginal probability $P(Y)$ can obviously take nonzero values only at the same values of Y. In this perspective, solving the hard-clustering problem for the training set is a special case of the second problem which is parametrized by β.

2.3 Local Optimum

In the previous section we derived an effective cost to be minimized to find the most probable representatives set Y. This cost is the free energy (2.13), which we rewrite here explicitly in terms of a given distortion measure by using (2.4) and (2.10):

$$F = -\frac{1}{\beta}\sum_{x}\log\left(\sum_{k} e^{-\beta d(x,y_k)}\right). \tag{2.14}$$

The set Y of vectors that optimizes the free energy satisfies

$$\frac{\partial}{\partial y_j}F = 0, \qquad \forall j, \tag{2.15}$$

where this is shorthand notation for differentiation with respect to each component separately, or a gradient with respect to y_j. Differentiating (2.14) we obtain

$$\sum_{x} P(x \in C_j)\frac{\partial}{\partial y_j}d(x,y_j) = 0, \tag{2.16}$$

where $P(x \in C_j)$ is the association probability as given in (2.3),(2.4). Equivalently, after normalization we get

$$\langle \frac{\partial}{\partial y_j}d(x,y_j)\rangle_j = 0, \tag{2.17}$$

where the expectation is over the distribution of training samples in C_j. Note that if we could interchange the expectation and the differentiation operators in (2.17), we would get that y_j minimizes the average cluster distortion, which is a fuzzy formulation of the "centroid" condition in LBG [26]! However, the probability distribution over which we perform the expectation depends on y_j in general. There are two special cases where it does not. The first is $\beta = 0$,

which yields the uniform association distribution. The second is $\beta \to \infty$, where the distribution is piecewise constant (a region of 0 and a region of 1); i.e., its derivative vanishes almost everywhere, so

$$\lim_{\beta\to\infty} \frac{\partial F}{\partial y_j} = \frac{\partial}{\partial y_j} \sum_{x\in C_j} d(x, y_j). \tag{2.18}$$

The LBG algorithm is indeed a method for solving (2.17) at $\beta \to \infty$. At one step of the iteration it is assumed that the probabilities are locally constant, and the "centroid" condition (2.18) is enforced. The other step consists of checking to see if any of the associations moved from a region of one to a region of zero, or in other words, reclassifying the data set. This illustrates how our proposed approach is a generalization of traditional approaches, and how it converges to them at the limit of $\beta \to \infty$. This issue will be further discussed in the next chapter, within our treatment of annealing.

2.4 The ν-th Law Distortion

In this section we apply the general results to the important example of the ν-th law distortion measure

$$d_\nu(x, y) = \sum_i |x(i) - y(i)|^\nu, \tag{2.19}$$

which is the ν-th power of the l_ν norm. The squared distance distortion, $\nu = 2$, is apparently the most extensively used distortion measure, and will be particularly discussed here.

The necessary condition for minimizing the free energy (2.16) or (2.17) for the i-th component of the j-th vector becomes

$$\begin{aligned} \sum_{\{x|x(i)<y_j(i)\}} & P(x \in C_j)|x(i) - y_j(i)|^{\nu-1} = \\ \sum_{\{x|x(i)>y_j(i)\}} & P(x \in C_j)|x(i) - y_j(i)|^{\nu-1}. \end{aligned} \tag{2.20}$$

This shows that the optimal y_j is the cluster's symmetry (or antisymmetry) point of the $(\nu - 1)$-th moment.

In the case of the squared-distance distortion, (2.20) is rewritten $\forall i$ as

$$\sum_x P(x \in C_j)(x - y_j) = 0, \tag{2.21}$$

or in the form of

$$y_j = \frac{\sum_x xP(x \in C_j)}{\sum_x P(x \in C_j)}. \tag{2.22}$$

Each representative is interpreted as the center of mass of the fuzzy cluster, or the average over all data points, where to each data point we assign its relative weight in the cluster. This is a generalization of the center-of-mass condition in basic ISODATA [1], which is the centroid condition of LBG for the squared-distance distortion. Here also, the basic-ISODATA center-of-mass condition is obtained from (2.22) at $\beta \to \infty$

$$\lim_{\beta \to \infty} y_j = \frac{1}{n_j} \sum_{x \in C_j} x, \tag{2.23}$$

where n_j is the population of (number of data points in) the cluster.

It is also worth noting that in the squared-distance distortion case, the association probabilities take the Gaussian form

$$P(x \in C_j) = \frac{e^{-\beta |x - y_j|^2}}{Z_x}. \tag{2.24}$$

2.5 A Note on Maximum Likelihood

The relation between some of the above results and maximum likelihood estimation of parameters in density mixtures should be discussed at this point. Note, for example, that maximum likelihood estimation of means in normal mixtures often uses fixed-point iterations based on (2.22) and (2.24) [6][20]. To focus on the distinction between the approaches, recall that we did *not* make any assumption on the data distribution. The association probabilities were derived from the

distortion criterion to be minimized. If the sum of squared distances distortion is used, these probabilities become Gaussian. If the data happen to be a normal mixture, it may be intuitively satisfying to learn that we wind up doing maximum likelihood estimation (with the advantage of using annealing to avoid nonglobal minima as will be shown in the next chapter). While it may be reasonably argued for certain clustering problems in pattern recognition, in particular cluster analysis of mixtures, that selecting a distortion function is equivalent to assuming (or modelling) the data distribution, it is certainly not so in many other situations. An important example is vector quantization in source coding. In this case the distortion is usually related to the application requirements, not to the data distribution; e.g., the distortion to be minimized in image coding should reflect what the viewer is sensitive to and not assumptions on the input distribution.

Thus we conclude that the suggested probabilistic framework is more general than the maximum likelihood approach, and may be equivalent to it in special circumstances at a given "temperature," i.e., excluding the important aspect of annealing. Moreover, we shall see similar relations between our approach and maximum likelihood estimation with unknown class priors [6] when we introduce the mass-constrained clustering in a later chapter.

Chapter 3

Annealing and Phase Transitions

3.1 Assumptions

We restrict the class of allowed distortion (dissimilarity) measures by making the following assumptions.

d-1. $d : R^s \times R^s \to [0, \infty)$ is continuous.

d-2. $d(x, y)$ is a convex function of y for fixed x.

d-3. $d(x, y) \to \infty$ at $|y| \to \infty$ for fixed x.

The first assumption was made implicitly in the previous chapter. These assumptions are a subset of the assumptions made in the literature [40] [41].

3.2 Annealing

In the previous chapter we obtained the necessary condition for a local optimum at a given β (2.16), which is reproduced here for convenience,

$$\sum_x P(x \in C_j) \frac{\partial}{\partial y_j} d(x, y_j) = 0. \tag{3.1}$$

This local minimum of the free energy can be obtained by using one's favorite iterative method (e.g., gradient descent).

At $\beta = 0$, the association probabilities are uniform, and (3.1) becomes

$$\frac{\partial}{\partial y_j} \sum_x d(x, y_j) = 0. \tag{3.2}$$

This is the centroid condition for the entire data set viewed as one cluster!

Claim: There is a unique solution to (3.2).

Proof: Consider the function

$$f(y) = \sum_x d(x, y).$$

By assumption (d-2) it is a finite sum of functions that are convex in y. Hence, $f(y)$ is convex and the claim follows by assumption (d-3). □

Thus, at $\beta = 0$ our effective cost function is convex, and all representatives converge to the *same point*, which is the *global minimum*. This point is the centroid of the cluster consisting of the entire data set. At $\beta > 0$, however, a set Y of vectors satisfying (3.1) corresponds to a *local* minimum of the free energy. In order to avoid arbitrary local minima depending on the initialization of the iterations, we introduce deterministic annealing. Annealing can be viewed as starting at the global minimum at $\beta = 0$, and tracking the minimum while gradually increasing β.

The basic DA method for *hard* clustering is as follows. Set the required number of representatives. Initialize β to zero or to some small positive value. At each iteration minimize the free energy by solving (3.1), and then increase β. When β reaches a value that makes all associations practically hard – stop. Some questions concerning the annealing schedule (or the β increase rate) remain unanswered, such as what is the fastest schedule that still ensures the best minimum obtainable by the method. More insight in this matter will be gained in a later section.

Let us consider an algorithm for the special case of the squared-distance distortion measure. At each β we optimize by fixed-point iterations based on (2.22) and (2.24); i.e.,

$$y_j^{(n+1)} = \frac{\sum_x x P(x \in C_j)}{\sum_x P(x \in C_j)}, \tag{3.3}$$

where

$$P(x \in C_j) = \frac{e^{-\beta |x - y_j^{(n)}|^2}}{\sum_k e^{-\beta |x - y_k^{(n)}|^2}}. \tag{3.4}$$

Note that for $\beta \to \infty$, this becomes exactly the basic ISODATA algorithm, where each iteration is composed of two steps. First, each data point is assigned to the nearest representative, i.e., evaluating (3.4) at the limit. Then new representatives are computed as the centers of the resulting clusters, i.e., evaluating (3.3).

The DA algorithm performs fuzzy clustering at various degrees of fuzziness. It starts by finding an extremely fuzzy solution, and then the fuzziness is gradually reduced until at the limit it evolves into a known, hard-clustering method. This annealing process can also be understood as gradual localization of the influence of data points. Before we further analyze the annealing process, let us present some simulation results for the hard clustering problem.

3.3 Hard-Clustering Results

This version of the algorithm performs hard clustering, given a fixed number of representatives. This is a well-posed problem, and it is in this context that many traditional techniques are proved to converge to a local optimum. Regardless of the initial configuration, at $\beta = 0$ all the representatives will be at the center of mass of the entire data set. According to our annealing process we gradually increase β and reoptimize by solving (3.1). At the limit of $\beta \to \infty$, the associations (2.3) become hard, and each sample point is associated with exactly one representative. At this limit the algorithm becomes exactly the LBG algorithm; i.e., at each iteration every point is assigned to the "nearest" representative (in

the sense of minimal distortion), and then a new set of "centroids," minimizing the cluster average distortion (3.2), are computed. Since in our simulations the distortion measure is the squared distance, then at the limit we get exactly the basic ISODATA algorithm [1].

The simulation example demonstrates the performance of our annealing algorithm as compared to basic ISODATA. The training set was generated from a normal mixture whose density centers are marked by X in the figures. The output of basic ISODATA depends on the choice of initial configuration. Figure 3.1(a) shows the ISODATA result for an initialization consisting of placing the means on a small circle around the center of mass of the distribution. The final location of the representatives is marked by O, and the distortion is 9024. The output of the DA algorithm is shown in Figure 3.1(b), and is independent of the initial configuration. The final distortion here is 7635.

Finally, the ISODATA algorithm has been given an extremely "good" initial configuration by placing the representatives at the exact centers of the normal densities from which the data were generated. At this initial configuration the distortion was 7769. The ISODATA algorithm converged to a better local minimum of 7644, *which is still slightly higher than the minimum found by the DA method* (which is, of course, independent of the initial configuration). Similar results were obtained in simulations performed with other data sets. This also suggests the curious fact that there are many local minima in the vicinity of the global minimum.

I have not been able to prove that our algorithm will always find the global minimum. In fact, it is not improbable that there are situations when it will fail to do so. However, in all the simulations very good minima were obtained, and experiments such as the last one described above have never shown the existence of lower minima. This is especially encouraging, and shows that the proposed method consistently outperforms traditional methods, and leads me to conjecture that the conditions for finding the global minimum are not very restrictive.

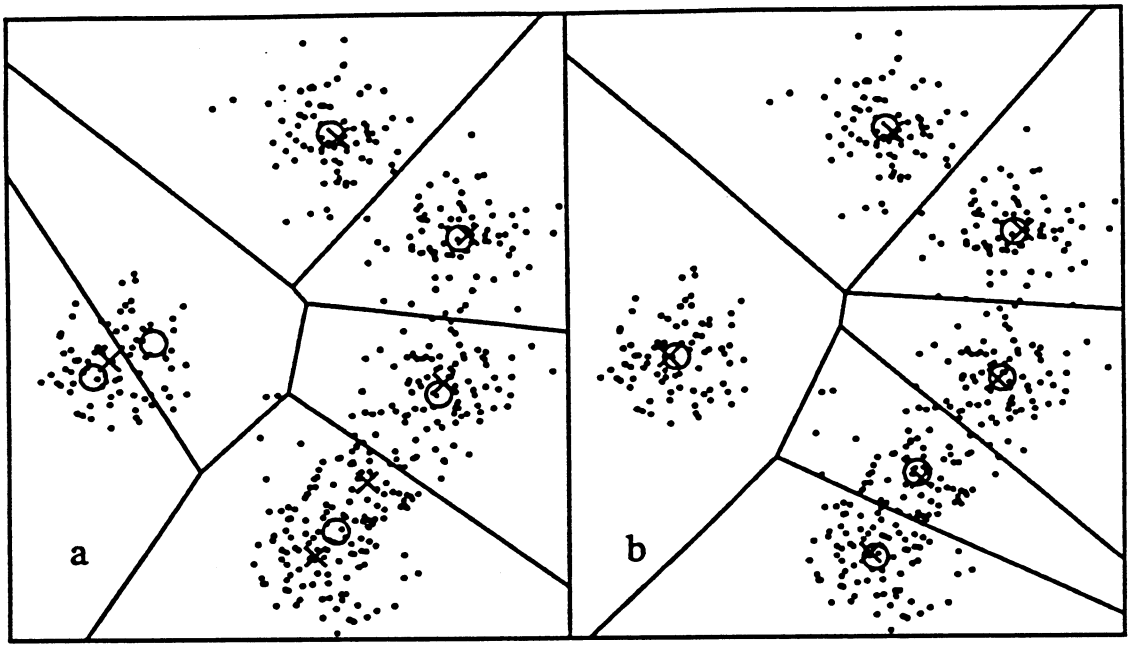


Figure 3.1: Hard-clustering results. The data are generated from six Gaussian distributions centered at the location marked by "X." The calculated cluster centroids are marked by "O." The lines are the decision boundaries. (a) Basic ISODATA clustering. The distortion is 9024. (b) Deterministic annealing. The distortion is 7635.

3.4 Phase Transitions and Hierarchical Clustering

We have already seen that at $\beta = 0$ all data points are equally associated with all representatives. We have further seen that for all distortion measures satisfying assumptions (d-1,2,3), there is a unique solution, and regardless of the number of representatives they will all converge to the same point. This point is the global minimum at $\beta = 0$.

We shall interpret identical representatives as representing the same cluster and shall consider the set Y of vectors *without repetitions* as the set of natural clusters. Thus, at $\beta = 0$ we actually have *one* natural cluster that is representable

by one representative. Mathematically, the single solution of $\beta = 0$ will be a solution of (3.1) for all β, but at some positive β it will change from a stable solution (local minimum) into a nonstable one (a saddle point or a local maximum). At this moment it becomes advantageous to split into subgroups of representatives. Each of these subgroups is a newly formed natural cluster. Note that this cluster split corresponds to a phase transition in our physical analogy.

As long as the number of vectors is not limited a priori, since we avoid repetitions, the *natural* number of clusters will emerge at every given β. At $\beta = 0$ we have one natural cluster consisting of the entire data set, but as β is increased, the system will undergo a sequence of phase transitions, where each phase corresponds to a certain number of natural clusters. This process results in a natural hierarchy of clustering solutions. The term "natural hierarchy" is used to reflect the fact that cluster splits happen naturally as we increase β. None of the common heuristics for introducing new representatives and placing them in "good" initial locations are needed.

Recalling that β is the Lagrange multiplier related to the average distortion, we realize that what we have is a hierarchy of clustering solutions at decreasing levels of average distortion. This is reminiscent of the basic philosophy of rate-distortion theory, but not exactly analogous, as will be discussed later when we consider the vector quantization application. The obtained hierarchy can be regarded as "looking at the problem at different scales." As an example to illustrate this, suppose we look at a picture and say, "This is a village surrounded by woods." We have just used a clustering solution at very low β. But we could have said, "Here is a house, here is a house, there is a tree, etc." This would be a clustering solution at high β, with many clusters. Both solutions are perfectly acceptable, only they correspond to looking at the problem at different scales. This explains why hierarchical clustering is very useful, especially when dealing with real-world, complex problems which often tend to have an inherent multiscale structure.

Let us return to the analysis of the annealing process. We have identified phase transitions in the process. In our physical analogy, phase transitions occur at certain critical temperatures. The rest of this section is devoted to computing critical β for our phase transitions. We shall derive the first critical β for the ν-th law distortion family, and then consider the special case of the squared-distance distortion, which gives the solution an interesting interpretation, and yields a better understanding of the process.

At $\beta = 0$ we have one cluster represented by the symmetry point of the $(\nu - 1)$-th moment. Without loss of generality we shall take this point to be the origin. The phase transition occurs when the Hessian at the origin is no longer positive-definite; i.e, it evolves from a local minimum into a nonstable solution. At $\beta = 0$ and $y_k = 0 \quad \forall k$, (2.20) becomes

$$\sum_{\{x|x(i)<0\}} |x(i)|^{\nu-1} = \sum_{\{x|x(i)>0\}} |x(i)|^{\nu-1}. \tag{3.5}$$

Let us use the following binomial series expansion for each component of the distortion

$$|a-b|^\nu = [(a-b)\mathrm{sgn}(a-b)]^\nu = \mathrm{sgn}^\nu(a-b)[a^\nu - \nu a^{\nu-1}b + O(b^2)],$$

where $\mathrm{sgn}(a-b)$ is the signum function, and for $b \to 0$ we replace it by $\mathrm{sgn}(a)$. Using this when differentiating (2.14) and discarding the $O(y^2(i))$ terms, we get

$$\frac{\partial F}{\partial y_j(i)} = -\nu \sum_x \frac{\mathrm{sgn}^\nu[x(i)][x^{\nu-1}(i) - (\nu-1)x^{\nu-2}(i)y_j(i)]\phi(x,y_j)}{\sum_k \phi(x,y_k)}, \tag{3.6}$$

where

$$\phi(x,y) = \exp\{\beta\nu \sum_m x^{\nu-1}(m)y(m)\mathrm{sgn}^\nu[x(m)]\}, \tag{3.7}$$

and where we also cancelled the common factor $\exp\{-\beta\sum_m |x(m)|^\nu\}$ from the numerator and the denominator. For convenience we rewrite the above as

$$\frac{\partial F}{\partial y_j(i)} = -\nu \sum_x \frac{A_x(i)}{B_x}, \tag{3.8}$$

where $A_x(i)$ and B_x stand for the numerator and denominator, respectively. Note that at the origin

$$B_x(Y=0) = n_r,$$

where n_r is the number of representatives. Therefore, for computing the Hessian at the origin,

$$H_{jj}[i,l] = \frac{\partial^2 F}{\partial y_j(i)\partial y_j(l)} = -\frac{\nu}{n_r}\sum_x \frac{\partial A_x(i)}{\partial y_j(l)} + \frac{\nu}{n_r^2}\sum_x A_x(i)\frac{\partial B_x}{\partial y_j(l)}. \tag{3.9}$$

By expanding the exponential function in y_j, we get

$$\sum_x A_x(i) = \sum_x \mathrm{sgn}^\nu[x(i)][x^{\nu-1}(i) - (\nu-1)x^{\nu-2}(i)y_j(i)] \cdot \{1 + \nu\beta\textstyle\sum_m x^{\nu-1}(m)y_j(m)\mathrm{sgn}^\nu[x(m)]\}. \tag{3.10}$$

Noting that

$$\sum_x \mathrm{sgn}^\nu[x(i)]x^{\nu-1}(i) = \sum_x \mathrm{sgn}[x(i)]|x(i)|^{\nu-1} = 0$$

is obtained by (3.5), and discarding terms containing second or higher powers of y, we get

$$\sum_x A_x(i) = \beta\nu\sum_x \mathrm{sgn}^\nu[x(i)]x^{\nu-1}(i)\textstyle\sum_m x^{\nu-1}(m)y_j(m)\mathrm{sgn}^\nu[x(m)] - (\nu-1)\sum_x \mathrm{sgn}^\nu[x(i)]x^{\nu-2}(i)y_j(i). \tag{3.11}$$

This result can be rewritten in matrix-vector notation

$$\sum_x A_x = N\left[\beta\nu C_{zz} - (\nu-1)\Lambda\right] y_j, \tag{3.12}$$

where N is the number of data points, Λ is the diagonal positive-semidefinite matrix

$$\Lambda[i,i] = \frac{1}{N}\sum_x |x(i)|^{\nu-2}, \tag{3.13}$$

and C_{zz} is the covariance matrix of the vectors z defined by

$$z(i) = \mathrm{sgn}[x(i)]|x(i)|^{\nu-1}. \tag{3.14}$$

Next let us consider the derivative of B_x at the origin

$$\frac{\partial B_x}{\partial y_j(l)}(Y=0) = \beta\nu|x(l)|^{\nu-1}\mathrm{sgn}[x(l)],$$

which implies that at the origin,

$$\sum_x A_x(i)\frac{\partial B_x}{\partial y_j(l)} = \beta\nu\sum_x |x(l)|^{\nu-1}|x(i)|^{\nu-1}\mathrm{sgn}[x(l)]\mathrm{sgn}[x(i)]. \tag{3.15}$$

From (3.9), (3.12) and (3.15) we obtain the sub-Hessian with respect to y_j at the origin,

$$H_{jj} = \frac{N\nu}{n_r}[(\nu-1)\Lambda - \beta\nu C_{zz}] + \frac{N\beta\nu^2}{n_r^2}C_{zz}. \tag{3.16}$$

The cross components of the Hessian consist only of the second term in (3.16); i.e.,

$$H_{ij} = \frac{N\beta\nu^2}{n_r^2}C_{zz}, \qquad i \neq j. \tag{3.17}$$

The Hessian H is the large matrix obtained by placing copies of H_{jj} on the diagonal, and copies of H_{ij} elsewhere. We are interested in analyzing the conditions for positive-definite H. Noting that C_{zz} is a covariance matrix, and therefore positive-semidefinite, it is quite straightforward to show that H is positive-definite *if and only if* the first term in (3.16) is. We shall use this claim here, and prove it later.

We thus consider the matrix $H_1 = (\nu-1)\Lambda - \beta\nu C_{zz}$. Both Λ and C_{zz} depend only on the data set, not on β, so that varying β only modifies the balance between these two fixed matrices. At $\beta = 0$, H_1 is positive-definite since Λ is (we ignore here pathologies such as $\det\Lambda = 0$). It stays positive definite until it reaches the point where its determinant vanishes.

The critical value for β is thus

$$\beta_c = \frac{\nu-1}{\nu\lambda_{max}}, \tag{3.18}$$

where λ_{max} is the largest eigenvalue of $\Lambda^{-1}C_{zz}$. Moreover, bifurcation occurs along the eigenvector corresponding to the Hessian's zero eigenvalue. This means

that the split will be in the direction of the eigenvector of $\Lambda^{-1}C_{zz}$ corresponding to λ_{max}.

Now that we have derived the critical β for the ν-th law family, let us see its interpretation for the special case of the squared distance distortion measure ($\nu = 2$). It is easy to see that in this case

$$H_1 = (I - 2\beta_c C_{xx}), \tag{3.19}$$

where C_{xx} is the covariance matrix of the training set, and (3.18) reduces to

$$\beta_c = \frac{1}{2\lambda_{max}}, \tag{3.20}$$

where λ_{max} is now the largest eigenvalue of C_{xx}. We see that in this case the critical temperature is determined by the variance along the major principal axis of the distribution. Furthermore, the split will be initiated in the direction of this principal axis. Finally, as long as we may neglect intercluster influences, this derivation will hold for the following phase transitions, and every cluster will split at the critical temperature corresponding to its variance.

We now have an approximate idea of how the annealing process works. As β is increased, whenever it reaches a value corresponding to the variance along a cluster's major principal axis, this cluster splits into smaller clusters. These clusters stay intact until β reaches values corresponding to their (smaller) variances, etc. This also indicates how β relates to the cluster variances in the solution. Note that this description is approximate after the first phase transition, because we neglected intercluster influences on the phase transition. It should be a good approximation when the phase transitions are well spaced.

Just as in the physical analogy, the critical moments in the process are the phase transitions. Knowing to predict the next critical β may allow us to accelerate the process between phase transitions, while being more careful during the transition.

Before ending this section, let us prove our claim regarding the necessary and sufficient condition for H_1 to be positive-definite.

Claim: Let Q be a positive-semidefinite $n \times n$ matrix. Let S be the $kn \times kn$ matrix constructed using $n \times n$ submatrices according to

$$S[i,j] = \begin{cases} P+Q & \text{if } i = j; \\ Q & \text{if } i \neq j. \end{cases}$$

Then S is positive-definite *iff* P is.

Proof: Consider the decomposition $S = S_1 + S_2$, where

$$S_1[i,j] = Q$$

and

$$S_2[i,j] = \begin{cases} P & \text{if } i = j; \\ 0 & \text{if } i \neq j. \end{cases}$$

Let v be an kn-tuple which we can view as the concatenation of k n-tuples: $(v_1 v_2 \ldots v_k)$. Now,

$$v^T S_1 v = (\sum_j v_j)^T Q (\sum_j v_j) \geq 0,$$

where the result is nonnegative because Q is positive-semidefinite.

$\Rightarrow$) If P is positive-definite, then so is S_2, and

$$v^T S v = v^T S_1 v + v^T S_2 v \geq v^T S_2 v > 0.$$

Hence, S is positive-definite.

$\Leftarrow$) If P is not positive-definite, then $\exists$ nonzero w such that $w^T P w \leq 0$. Construct v using n-tuples of the form $v_j = a_j w$, such that

$$\sum_j v_j = 0,$$

and therefore $v^T S_1 v = 0$. Now,

$$v^T S v = v^T S_1 v + v^T S_2 v = v^T S_2 v \leq 0.$$

Hence, S is not positive-definite. □

3.5 Hierarchical-Clustering Results

The algorithm used here produces a hierarchy of clustering solutions. The goal will be to find the most probable set of representatives (or optimal fuzzy solution) at different scales. Thus, each clustering result will show fuzzy clustering with the underlying Gibbs distribution defining the fuzzy membership in clusters, and the representatives will be the fuzzy cluster centroids. The results will illustrate the phase transitions in the process, as well as the fuzzy solutions for intermediate β.

The clustering hierarchy is shown in Figure 3.2. The training set is generated from a mixture of six normal distributions, and we see the solutions obtained at different phases. The process starts with one natural cluster containing all the training set (shown in Figure 3.2(a)). At the first phase transition it splits into two clusters (Figure 3.2(b)), and passes through a sequence of phase transitions until it reaches the "natural" set of six clusters. The next phase transition results in an "explosion" where all clusters split. This is predictable by our analysis of phase transitions. Here we have a set of identical isotropic clusters. By (3.19) we know that the critical temperature will be the same for all these clusters. Moreover, since their covariance matrices are isotropic ($C_{xx} = \lambda I$), every vector is an eigenvector, so that the split may be initiated in all directions. A phase diagram is given in Figure 3.3. It shows the behavior of the average distortion throughout the annealing process, and the number of natural clusters at each phase.

3.6 A Note on the Applications

As stated in the introduction, clustering problems are encountered in a large variety of fields. The objectives of these applications are not always exactly the same. Some fields, such as classical quantization, require hard clustering by definition. Others, such as estimation of parameters in mixtures, require fuzzy solutions

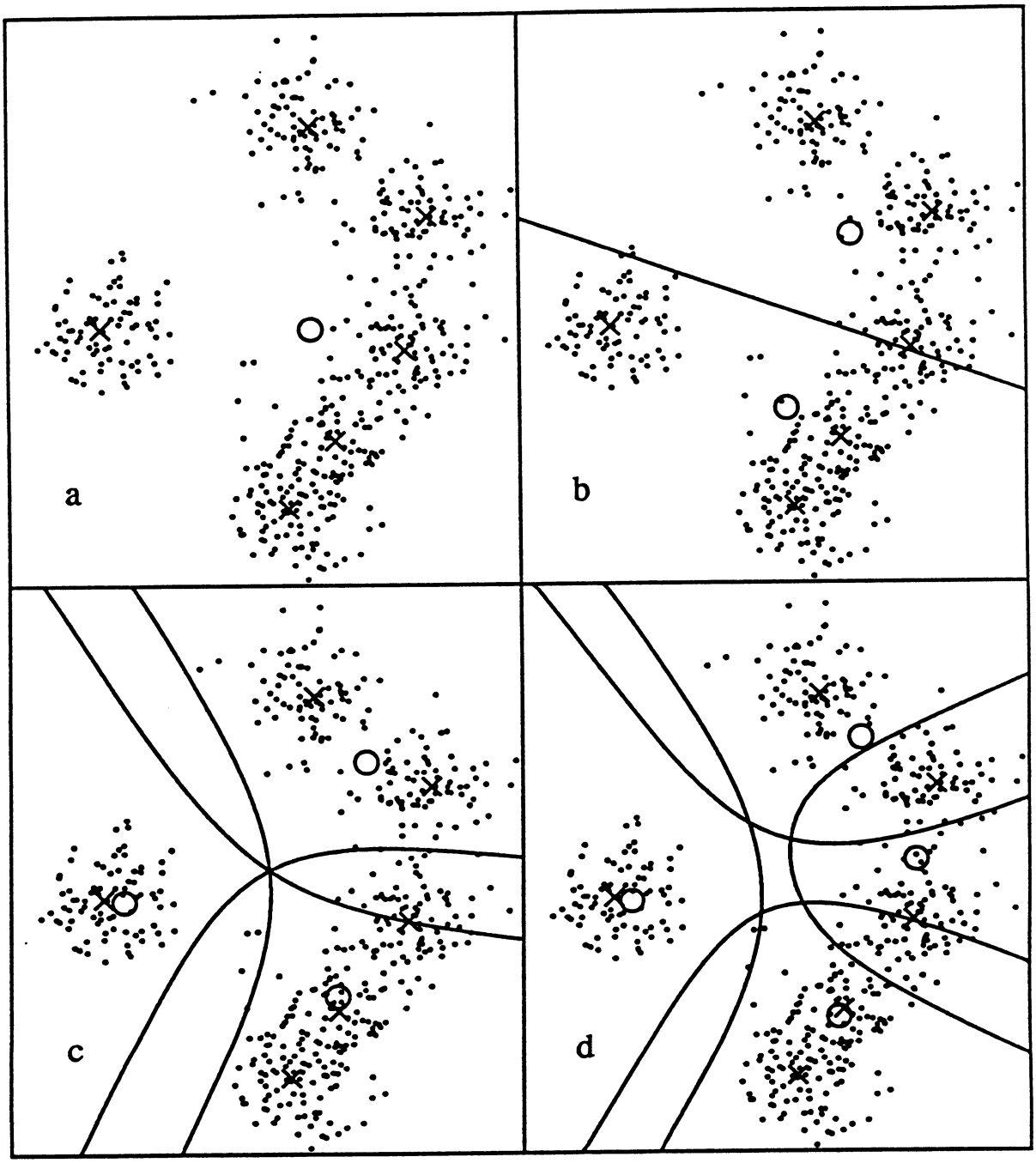
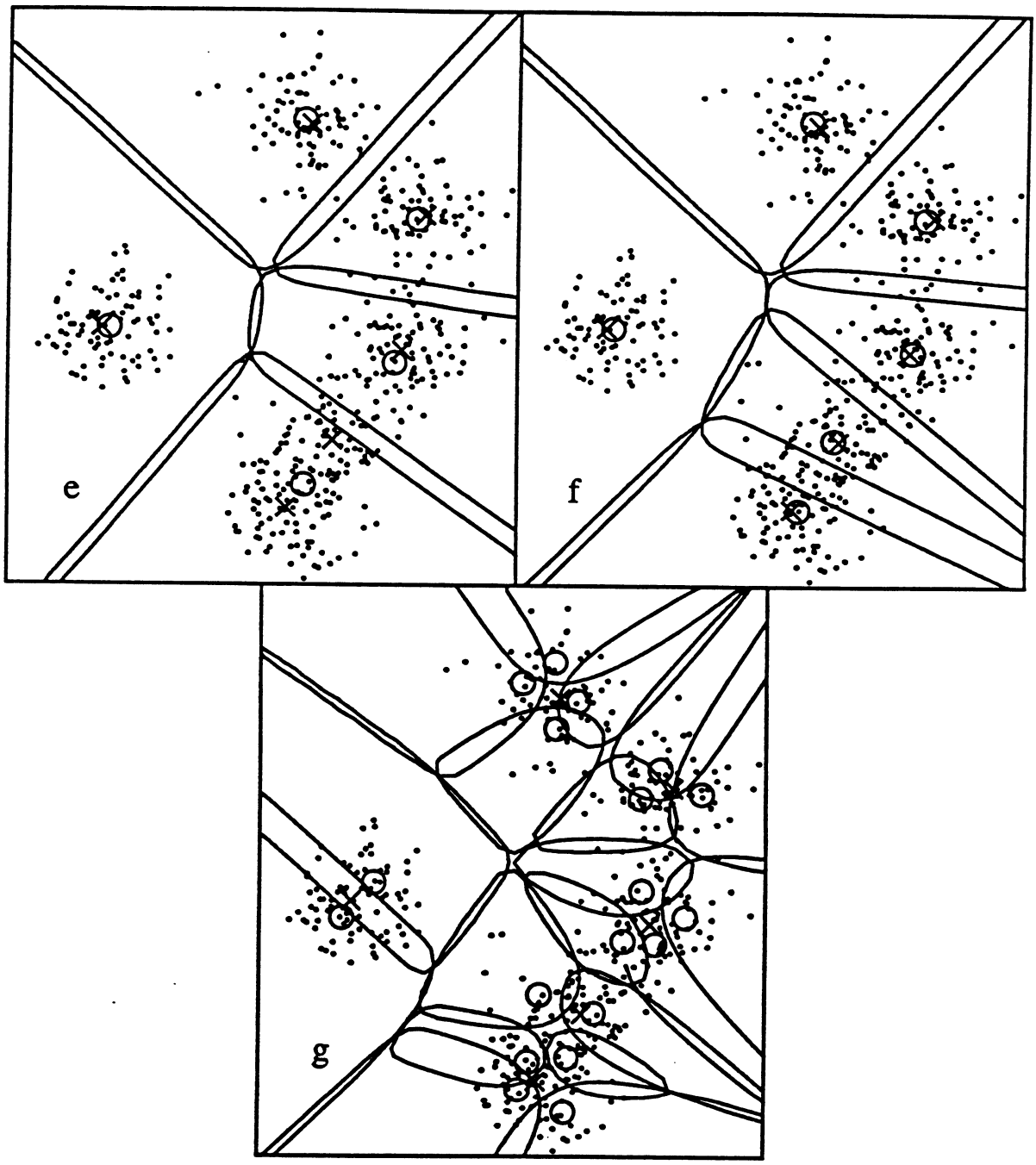


Figure 3.2: Clustering at different phases corresponding to Fig. 3.3. The lines are equiprobability contours, $p = 1/2$ in (b), and $p = 1/3$ elsewhere. (a) 1 cluster ($\beta = 0$), (b) 2 clusters ($\beta = 0.0049$), (c) 3 clusters ($\beta = 0.0056$), (d) 4 clusters ($\beta = 0.0100$), (e) 5 clusters ($\beta = 0.0156$), (f) 6 clusters ($\beta = 0.0347$), and (g) 19 clusters ($\beta = 0.0605$).


e
f
g

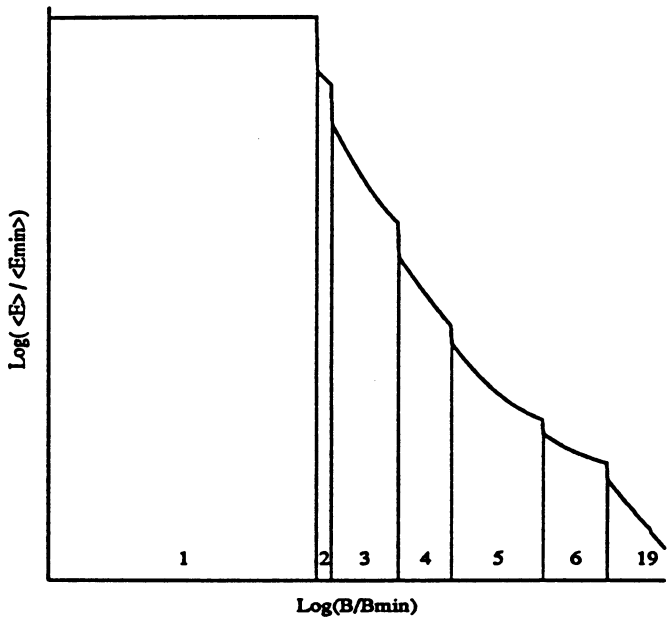


Figure 3.3: Phase diagram for the distribution shown in Fig. 3.2. The number of actual clusters is shown for each phase.

by definition. This is the reason that we have developed the deterministic annealing *approach* and have shown in the simulation results sections how differing algorithms are derived from the approach.

For the vector quantization application, we usually want decision boundaries, and partition the space into regions, each associated with a quantization level (representative). This is hard clustering. When demonstrating hierarchical clustering solutions we mentioned the relation to rate-distortion theory. However, in order to obtain hard-clustering solutions, we have to take the fuzzy solution at each phase and let $\beta \rightarrow \infty$ while fixing the number of representatives. This yields the minimal distortion at each rate (quantizer size). Moreover, to obtain full analogy with a rate-distortion function, we need to consider the representatives' entropy instead of their number. This leads to entropy-constrained clustering [5],

which is beyond the scope of this work. It is nevertheless an interesting problem which will be investigated in the future using generalizations of the methods for constrained clustering, which are discussed in the next chapter.

Having said that vector quantization is essentially a hard-clustering problem, we now point out that fuzzy-clustering solutions are still of interest in this application. To illustrate this issue consider the quantization of image pixel values to a small number of, say, two levels. Using hard clustering you simply threshold the picture. The resulting binary picture is seldom pleasing to look at. If you use fuzzy clustering, and randomly generate binary levels to pixels according to their association probabilities, you get a much better result. This is, in fact, the way pictures are printed in newspapers using binary (black/white) pixels. The common use of dithering in image quantization is directly related to this approach.

On the other side of the spectrum, for cluster analysis of mixtures, while fuzzy clustering is essentially required, skepticism still exists in the field. Skeptics are not convinced that fuzzy clustering offers advantages over the classical and better understood methods, and point out that without significantly overlapping clusters, it is obviously not needed [20].

We conclude by stating that the DA approach to clustering was developed in general. It offers a contribution to virtually each of the applications, by deriving an appropriate algorithm in a straightforward manner.

Chapter 4

Optimization by Constrained Clustering

4.1 Constrained Clustering

In our formulation of the DA approach to clustering, no explicit constraint has been put on the set of representatives. It was only assumed implicitly that there were at least two 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 the Travelling Salesman Problem (TSP) we want to associate an *ordered* set of variables with a given fixed set of cities, so as to minimize the sum of consecutive distances. Image segmentation clearly belongs to this family as well, as we want to associate pixels optimally with an appropriate set of meaningful labels. The DA 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 DA method for clustering. It should, however, be noted that similarly to the clustering applications, 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 DA is merely a tool for avoiding local minima. The second group is typically concerned with generalizing from a training data set, or is 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, the set of cluster parameters, and V, 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,

$$\langle D(Y,V) \rangle = E, \tag{4.1}$$

where D is given as in (2.6),

$$D(Y,V) = \sum_x \sum_j v_{xj}\, d(x, y_j), \tag{4.2}$$

and then the extra constraint, which concerns only the cluster parameters

$$\langle T(Y) \rangle = L. \tag{4.3}$$

The maximum entropy probability distribution is

$$P(Y,V) = \frac{e^{-\beta D(Y,V)-\lambda T(Y)}}{\sum_{Y',V'} e^{-\beta D(Y',V')-\lambda T(Y')}}. \tag{4.4}$$

By summing over all possible hard partitions, similarly to the derivation for clustering (2.8-2.12), we obtain the marginal probability

$$P(Y) = \frac{e^{-\beta F(Y,\beta)-\lambda T(Y)}}{\sum_{Y'} e^{-\beta F(Y',\beta)-\lambda T(Y')}}, \tag{4.5}$$

where F is given in (2.13).

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

$$F + \frac{\lambda}{\beta} T,$$

which, by noting its Lagrangian form, can be conveniently viewed as minimizing $F(Y,\beta)$ over Y 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 the constraint. Let us start again and state that as is often done in such optimization problems, it is useful to minimize the Lagrangian

$$F' = F + qT, \tag{4.6}$$

where F is given in (2.13), 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 chapter. The first is an improvement of our method for the clustering problem. Although the maximal number of natural clusters at a given temperature is independent of the number of representatives, their actual location does depend on the number of

representatives and their multiplicity in the clusters. This weakness is eliminated by reformulating the problem as constrained clustering, or equivalently, taking into account the mass (or population) of each natural cluster. As a second example of constrained clustering we take TSP. It is shown how TSP can be viewed as constrained clustering at the limit of low temperature, and obtain the EN method [10,9], 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 the second Lagrange multiplier, and to propose a more powerful annealing scheme. The last example is related to self-organization in unsupervised learning [25]. It is explained how an appropriate, constrained-clustering formulation leads to a DA method to search for the optimal solution, given a finite training set.

4.2 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 k-th cluster. Equation (2.4) for the partition function is rewritten as

$$Z_x = \sum_k \lambda_k e^{-\beta d(x,y_k)}; \tag{4.7}$$

the association probability (2.3) is for a natural cluster

$$P(x \in C_j) = \frac{\lambda_j e^{-\beta d(x,y_j)}}{Z_x}, \tag{4.8}$$

and the free energy (2.13) is now

$$F = -\frac{1}{\beta}\sum_x \log Z_x = -\frac{1}{\beta}\sum_x \log \sum_k \lambda_k e^{-\beta d(x,y_k)}. \tag{4.9}$$

The free energy is to be minimized under the constraint of a fixed total number of representatives. The Lagrangian to be minimized (4.6) is thus

$$F' = F + q\left(\sum_k \lambda_k - M\right). \tag{4.10}$$

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

$$\frac{\partial}{\partial y_j} F' = 0. \tag{4.11}$$

Since the constraint is independent of y_j, this yields again (2.16); i.e.,

$$\sum_x P(x \in C_j) \frac{\partial}{\partial y_j} d(x, y_j) = 0, \tag{4.12}$$

with the distinction that now the association probabilities are according to (4.8).

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

$$\frac{\partial}{\partial \lambda_j} F' = -\frac{1}{\beta} \sum_x \frac{e^{-\beta d(x,y_j)}}{Z_x} + q = 0, \tag{4.13}$$

which yields

$$q\beta = \sum_x \frac{e^{-\beta d(x,y_j)}}{Z_x}. \tag{4.14}$$

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

$$\sum_k \lambda_k q\beta = \sum_k \lambda_k \sum_x \frac{e^{-\beta d(x,y_j)}}{Z_x}, \tag{4.15}$$

which, by applying our total mass constraint and using (4.7), yields

$$q\beta = \frac{N}{M}, \tag{4.16}$$

where N is the total number of data points in the training set. Substituting (4.16) in (4.14) we see that the optimal set of λ_k must satisfy

$$\sum_x \frac{e^{-\beta d(x,y_j)}}{Z_x} = \frac{N}{M}, \tag{4.17}$$

where the λ_k are implicit in Z_x (4.7). Equation (4.17) is thus the equation we solve while optimizing over $\{\lambda_k\}$.

Moreover, by using (4.17) and (4.8), we obtain

$$\lambda_j = \frac{M}{N} \sum_x \frac{\lambda_j e^{-\beta d(x,y_j)}}{Z_x} = \sum_x \mu P(x \in C_j), \tag{4.18}$$

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 [6]). Our formulation is more general, and does not necessarily assume a priori knowledge on the data distribution, as explained in the note on maximum likelihood in Chapter 2.

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, which 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 $\{\lambda_j\}$, the solution sets of centroids and masses, respectively. Now, let us raise the number of representatives and consider the case where the j-th natural cluster is represented by m_j representatives $y_j^{(n)}$, while the cluster's mass is arbitrarily divided between them; i.e.,

$$y_j^{(n)} = y_j, \qquad n = 1, ..., m_j \tag{4.19}$$

$$\sum_{n=1}^{m_j} \lambda_j^{(n)} = \lambda_j. \tag{4.20}$$

By (4.7), Z_x is invariant to any such division. Furthermore, the probability of association to the natural cluster is unchanged as

$$\sum_n P(x \in C_j^{(n)}) = \sum_n \frac{\lambda_j^{(n)} e^{-\beta d(x,y_j)}}{Z_x} = \frac{\lambda_j e^{-\beta d(x,y_j)}}{Z_x} = P(x \in C_j). \tag{4.21}$$

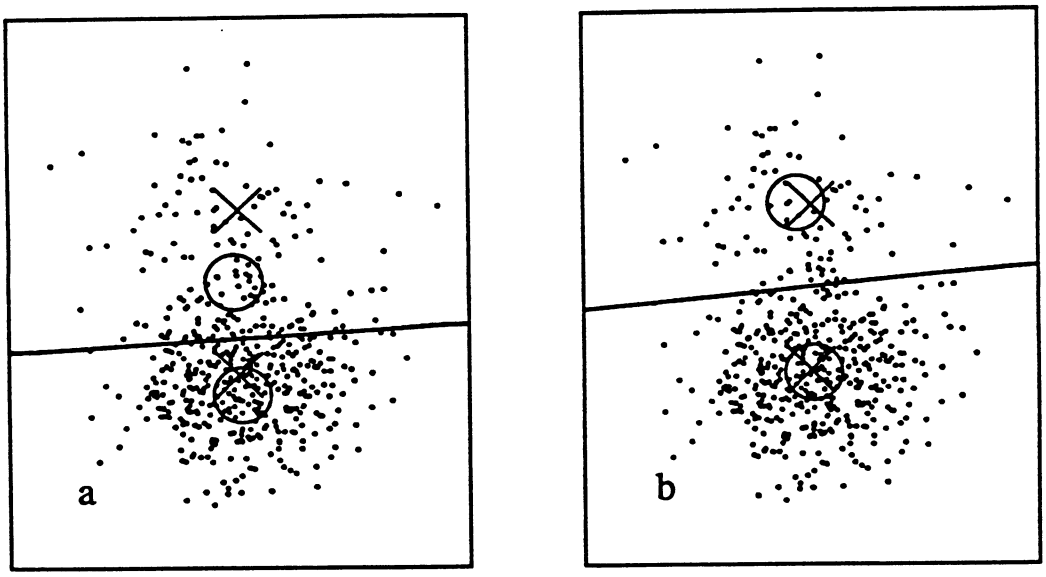


Figure 4.1: The effect of cluster mass (population) at intermediate β. The data are sampled from two normal distributions whose centers are marked by X. The computed representatives are marked by O. (a) Nonconstrained clustering. (b) Mass-constrained clustering.

It is therefore clear that the same representative locations will satisfy (4.12), 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 ($\beta \to \infty$), both the previous method and the mass-constrained method converge to the same process, namely, LBG [26] (or basic ISODATA [1] for the sum of squares distortion). This is so because the association probabilities in these 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 (Figure 4.1), and therefore yields a better method for hierarchical clustering, as well as a better chance of avoiding local minima.

4.3 The Travelling Salesman Problem

In the DA clustering algorithm, if we throw in enough representatives and let $\beta \to \infty$, 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 that would satisfy additional requirements. As an example, the EN approach to TSP [10][9][48][45] is considered here.

The problem statement is: Given a set of data points (usually called cities), find the shortest closed path which passes through all of them. In order to derive the EN 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 [10]. 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, and also 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 (4.6)

$$F' = F + \lambda \left(\sum_{k=1}^{N} |y_k - y_{k-1}|^2 - L \right), \qquad y_o = y_N. \tag{4.22}$$

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 [48]),

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

instead of D as defined in (2.6). Then, deriving the effective cost similarly to the derivation of (2.8) to (2.14), we would get F' of (4.22) instead of F of (2.14) (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

$$\frac{\partial}{\partial y_j} F' = 0 \qquad \forall j, \tag{4.23}$$

which by substituting (4.22) yields

$$\sum_x P(x \in C_j) \frac{\partial}{\partial y_j} d(x, y_j) + 2\lambda(2y_j - y_{j-1} - y_{j+1}) = 0. \tag{4.24}$$

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

$$\sum_x P(x \in C_j)(y_j - x) + \lambda(2y_j - y_{j-1} - y_{j+1}) = 0. \tag{4.25}$$

Note that this equation depends on β through the association probabilities (2.3). 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 [10] the formulation implies $\lambda \propto 1/\sqrt{\beta}$, while in [48] 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 is the tour length, normally. 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 particularly important at the vicinity of phase transitions where separating representatives should be ordered so as 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 [46]. 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 will be conveniently written as $h(Y) = L$. A necessary condition for an optimum is that the derivatives of the Lagrangian vanish; i.e.,

$$\frac{\partial F}{\partial y_j} + \lambda \frac{\partial h}{\partial y_j} = 0 \qquad \forall j. \tag{4.26}$$

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

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

It can be shown [31] that for such constrained optimization,

$$\lambda^* = -\frac{\partial F^*}{\partial L}. \tag{4.27}$$

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} is reached, which maintains some "tension" in the net. 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:

$$\frac{\partial \lambda^*}{\partial \beta} = -\frac{\partial}{\partial \beta}\left(\frac{\partial F^*}{\partial L}\right) = -\frac{\partial}{\partial L}\left(\frac{\partial F^*}{\partial \beta}\right), \tag{4.28}$$

where use was made of (4.27). Next we note that

$$\frac{\partial F^*}{\partial \beta} = \frac{\partial F}{\partial \beta}(Y^*, \beta) + \sum_k \frac{\partial F}{\partial y_k}(Y^*, \beta)\frac{\partial y_k^*}{\partial \beta}. \tag{4.29}$$

Differentiating the constraint we obtain

$$\sum_k \frac{\partial h}{\partial y_k}\frac{\partial y_k^*}{\partial \beta} = \frac{\partial L}{\partial \beta} = 0, \tag{4.30}$$

where 0 is obtained by the constant-L assumption. Adding λ^*·(4.30) to (4.29), we get

$$\frac{\partial F^*}{\partial \beta} = \frac{\partial F}{\partial \beta}(Y^*, \beta) + \sum_k (\frac{\partial F}{\partial y_k} + \lambda^*\frac{\partial h}{\partial y_k})\frac{\partial y_k^*}{\partial \beta}. \tag{4.31}$$

By (4.26) the second term equals 0 and so (4.31) reduces to

$$\frac{\partial F^*}{\partial \beta} = \frac{\partial F}{\partial \beta}(Y^*, \beta). \tag{4.32}$$

Let us now make the following observation,

$$\frac{\partial}{\partial \beta}(\beta F) = \sum_x \sum_j |y_j - x|^2 P(x \in C_j) = E. \tag{4.33}$$

So

$$\frac{\partial F}{\partial \beta} = \frac{E - F}{\beta}, \tag{4.34}$$

and by (4.32) we have

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

which when substituted into (4.28) yields

$$\frac{\partial \lambda^*}{\partial \beta} = -\frac{1}{\beta}(\frac{\partial E^*}{\partial L} - \frac{\partial F^*}{\partial L}) = -\frac{1}{\beta}(\frac{\partial E^*}{\partial L} + \lambda^*). \tag{4.35}$$

In practice this allows the use of the following approximation:

$$\Delta\lambda^*(\beta) \approx -\frac{\Delta\beta}{\beta}(\frac{\Delta E^*}{\Delta L} + \lambda^*), \tag{4.36}$$

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

Figure 4.2 shows our result for the ten-cities problem [9], which according to Durbin *et al.* is the optimal solution and 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.

Figure 4.3 shows our result for the first fifty-cities problem [10]. 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 the simulations β was increased exponentially, as had been done in [10]. 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 the 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 fifty-cities problem this required the rate $\Delta\beta/\beta = 0.0001$, as compared to $\Delta\beta/\beta = .01$ 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.

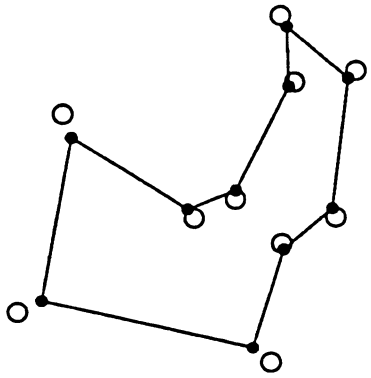


Figure 4.2: The ten-cities problem solved by deterministic annealing. This is the optimal tour for both the sum of distances, and the sum of squared distances.

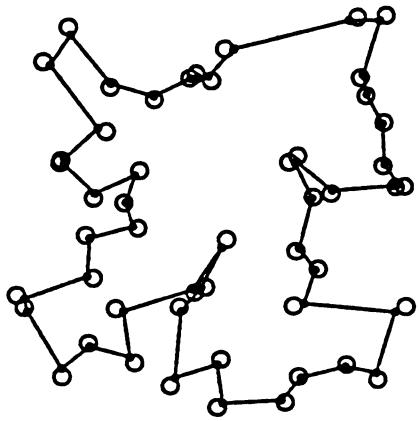


Figure 4.3: The DA result for the (first) fifty-cities problem.

4.4 Self-Organization

Kohonen [25] 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. This is also commonly called the dimensionality reduction problem.

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 reduced algorithms have been suggested and discussed earlier by Grossberg [15][16], and are based on competitive learning [39][17]. Chang and Gray [4] 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 empirically optimized, 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 close the circle by suggesting to extend the DA 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 (4.22) we now have

$$F' = F + \lambda \left(\sum_{k=2}^{N} |y_k - y_{k-1}|^2 - L \right). \tag{4.37}$$

If the data is one-dimensional and the number of representatives equals the number of data points, then what we get at the limit is a DA method for sorting, since the shortest open path is simply the ordered sequence of data points. Of course, DA 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, (4.37) is dealing with unsupervised learning of a linear network (linear topology), but can be easily extended to networks of higher topologies 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 (4.26-4.36) 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.

Figure 4.4 shows an example of the self-organization of a ten-unit linear network, given the same fifty-cities example we used for TSP. The results show a behavior similar to that of the stochastic method documented in [25] (the differences are discussed below). This demonstrates how a linear network tries to cope with two-dimensional input (dimensionality reduction). The biological plausibility of a stochastic version of such self-organization in cortical maps is discussed in [8]. In [48] 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, as the objective is to generalize from a training set.

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 keep in mind is that if KLP converges to a local minimum, it is *only at the limit* (equivalent to our $\beta \rightarrow \infty$). Intermediate results are stochastic and therefore should be taken with a grain of salt. DA, 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 a well-understood mathematical framework, where the process virtually always stays at the "statistical equilibrium" of its stochastic counterpart. I 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 (e.g., LBG) over sequential algorithms (e.g., k-means) in the field of clustering, and the inherent fuzzy nature of the problem, I conjecture that for self-organization, given a finite training set, DA should outperform KLP. Extensive experimentation is required to test this conjecture. Here, however, we were mainly interested in this subject as an example for constrained-clustering application. I intend to pursue this approach to unsupervised learning in a future study.

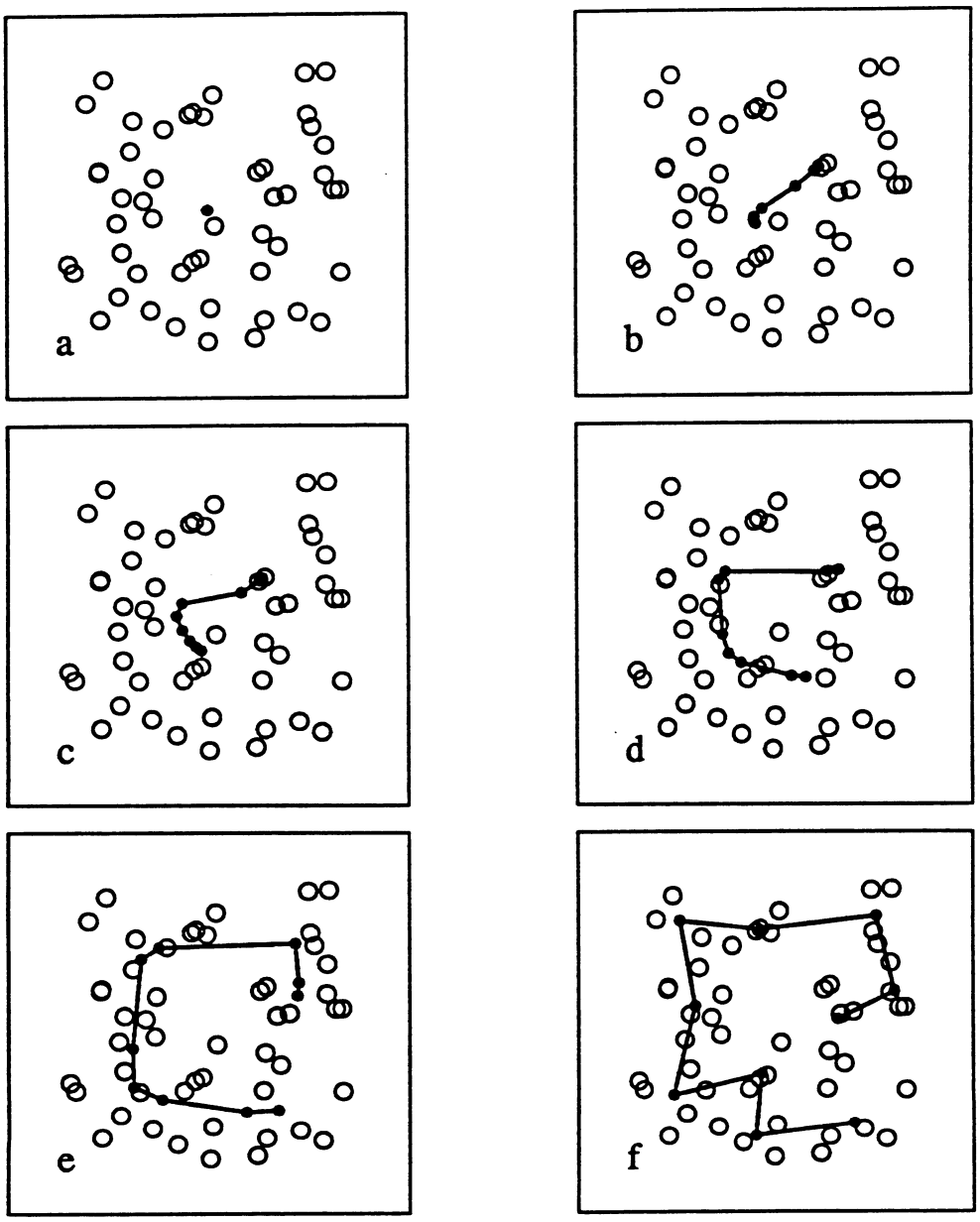


Figure 4.4: Self-organization of a linear network of ten units, given the fifty-cities problem data. (a) $\beta = 0$, (b) $\beta = 0.064$, (c) $\beta = 0.067$, (d) $\beta = 0.08$, (e) $\beta = 0.2$, (f) $\beta = 5$.

Chapter 5

Future Directions

In conclusion to the thesis, let us consider future directions for research. These belong "fuzzily" to two main categories, namely, applications, and generalizations. The first deals with the application of the ideas presented in this work to various related problems in different fields within science and engineering. The second deals with further generalization of the basic approach to cope with an even larger family of optimization problems.

One application mentioned in this work is entropy-constrained vector quantization. This is the right problem to solve if quantization values (the representatives) are to be encoded by a variable-length code. One should not be misled into thinking that it is constrained clustering as defined in Chapter 4. The constraint here is the representatives entropy, which is *not* a direct function of their values Y, but of their probabilities. The constraint is thus a function of the partition (set of associations) V. This therefore requires a more general formulation of constrained clustering than the one we have used.

Generalization to association or assignment problems that do not easily lend themselves to constrained-clustering formulation should be considered. An important class of such problems consists of problems where only a small subset of the fixed data is to be associated with the variables. One example is navigation,

where the solution consists of the shortest path, which is a small subset of the terrain data. Another example is tracking in the presence of clutter, where the estimated target returns to be assigned to the trajectory are a small subset of the detected data. Some preliminary work on multitarget tracking is summarized in [36] and [37].

A dramatic generalization would be to nonconvex optimization problems that are not necessarily association problems. This is maybe possible because the concept of deterministic annealing in its pure form is indeed general, but somewhat vague. However, this requires abandoning our probabilistic framework and the powerful annealing tool of fuzzy associations. It seems that instead of having a general DA method, appropriate probabilistic frameworks will have to be constructed for each family of optimization problems.

There are hosts of more or less immediate applications of DA as presented in this thesis. One application that has been treated here to some extent is self-organization and unsupervised learning in neural nets. Much more work is necessary to realize the potential of the DA approach. In particular, better understanding and quantification of the importance of fuzzy solutions and their relation to generalization are required. Other possible applications are in image processing and understanding, particularly image segmentation and image restoration. There is much interest in these applications in both science and engineering.

In summary, there seem to be many possible ways to continue this work. One main direction is further development and generalization of the basic approach. The other directions (many of which are not independent of the first) deal with a rich variety of applications.

Bibliography

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

[2] J. C. Bezdek. A convergence theorem for the fuzzy ISODATA clustering algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **2**:1–8, 1980.

[3] A. E. Cetin and V. Weerackody. Design vector quantizers using simulated annealing. *IEEE Transactions on Circuits and Systems*, **35**:1550, 1988.

[4] P.-C. Chang and R. M. Gray. Gradient algorithms for designing predictive vector quantizers. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **34**:679–690, 1986.

[5] P.A. Chou, T. Lookabaugh, and R. M. Gray. Entropy-constrained vector quantization. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **37**:31–42, 1989.

[6] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* Wiley-Interscience, New York, NY, 1974.

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

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

[9] R. Durbin, R. Szeliski, and A. Yuille. An analysis of the elastic net approach to the travelling salesman problem. *Neural Computation*, 1:348–358, 1989.

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

[11] S. Geman and D. Geman. Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[12] A. Gersho. On the structure of vector quantizers. *IEEE Transactions on Information Theory*, **28**:157–166, 1982.

[13] R. M. Gray. Vector quantization. *IEEE ASSP Magazine*, **1**:4–29, 1984.

[14] R. M. Gray and E. D. Karnin. Multiple local minima in vector quantizers. *IEEE Transactions on Information Theory*, **28**:256–261, 1982.

[15] S. Grossberg. Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. *Biological Cybernetics*, **23**:121–134, 1976.

[16] S. Grossberg. Adaptive pattern classification and universal recoding: II. Feedback, expectation, olfaction, illusions. *Biological Cybernetics*, **23**:187–202, 1976.

[17] S. Grossberg. Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science*, **11**:23–63, 1987.

[18] B. Hajek. Cooling schedules for optimal annealing. *Mathematics of Operations Research*, **13**:311–329, 1988.

[19] R. M. Haralick and L. G. Shapiro. Survey: Image segmentation techniques. *Computer Vision, Graphics, Image Processing*, **29**:100–132, 1985.

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

[21] E. T. Jaynes. Information theory and statistical mechanics. In R. D. Rosenkrantz, editor, *Papers on Probability, Statistics and Statistical Physics.* Kluwer Academic Publishers, Dordrecht, The Netherlands, 1989.

[22] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, **220**:671–680, 1983.

[23] J. Kittler and J. Illingworth. On threshold selection using clustering criterion. *IEEE Transactions on Systems, Man, Cybernetics*, **15**:652–655, 1985.

[24] R. Kohler. A segmentation system based on thresholding. *Computer Vision, Graphics, Image Processing*, **15**:319–338, 1981.

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

[26] Y. Linde, A. Buzo, and R. M. Gray. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, **28**:84–95, 1980.

[27] S. P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, **28**:129–137, 1982 (reprint of the 1957 paper).

[28] J. Max. Quantizing for minimum distortion. *IRE Transactions on Information Theory*, **6**:7–12, 1960.

[29] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys.*, **21**:1087–1091, 1953.

[30] N. M. Nasrabadi and R. A. King. Image coding using vector quantization: a review. *IEEE Transactions on Communications*, **36**:957–971, 1988.

[31] G. V. Reklaitis, A. Ravindran, and K. M. Ragsdell. *Engineering Optimization*. Wiley-Interscience, New York, NY, 1983.

[32] K. Rose, E. Gurewitz, and G. C. Fox. A deterministic annealing approach to clustering. *Pattern Recognition Letters*, **11**:589–594, 1990.

[33] K. Rose, E. Gurewitz, and G. C. Fox. Statistical mechanics and phase transitions in clustering. *Physical Review Letters*, **65**:945–948, 1990.

[34] K. Rose, E. Gurewitz, and G. C. Fox. Vector quantization by deterministic annealing. Technical Report C3P-895, California Institute of Technology, 1990 (submitted for publication).

[35] K. Rose, E. Gurewitz, and G. C. Fox. Constrained clustering as an optimization method. Technical Report C3P-919, California Institute of Technology, 1990 (submitted for publication).

[36] K. Rose, E. Gurewitz, and G. C. Fox. A nonconvex cost optimization approach to tracking multiple targets. In *IEEE International Workshop on Intelligent Robots and Systems IROS '90*, Japan, 1990.

[37] K. Rose, E. Gurewitz, and G. C. Fox. Multi-target tracking by graduated nonconvexity (in preparation).

[38] K. Rose, A. Heiman, and I. Dinstein. DCT/DST alternate-transform image coding. *IEEE Transactions on Communications*, **38**:94–101, 1990.

[39] D. E. Rumelhart and D. Zipser. Feature discovery by competitive learning. *Cognitive Science*, **9**:75–112, 1985.

[40] M. J. Sabin. Convergence and consistency of fuzzy c-means/ISODATA algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **9**:661–668, 1987.

[41] M. J. Sabin and R. M. Gray. Global convergence and empirical consistency of the generalized Lloyd algorithm. *IEEE Transactions on Information Theory*, **32**:148–165, 1986.

[42] P. K. Sahoo, S. Soltani, A. K. C. Wong, and Y. C. Chan. A survey of thresholding techniques. *Computer Vision, Graphics, Image Processing*, **4**:233–260, 1988.

[43] S. Z. Selim and M. A. Ismail. On the local optimality of the fuzzy ISODATA clustering algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **8**:284–288, 1989.

[44] J. E. Shore and R. W. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, **26**:26–37, 1980.

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

[46] A. G. Tsirukis, G. V. Reklaitis, and M. F. Tenorio. Nonlinear optimization using generalized Hopfield networks. *Neural Computation*, **1**:511–521, 1989.

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

[48] A. L. Yuille. Generalized deformable models, statistical physics, and matching problems. *Neural Computation*, **2**:1–24, 1990.

[49] K. Zeger and A. Gersho. Stochastic relaxation algorithm for improved vector quantizer design. *Electronics Letters*, **25**:896–898, 1989.