ARTICLES

Statistical mechanics of the maximum-likelihood density estimation

N. Barkai and H. Sompolinsky
Racah Institute of Physics and Center for Neural Computation, Hebrew University, Jerusalem 91904, Israel
(Received 11 February 1994)

Estimating the density of data generated by Gaussian mixtures, using the maximum-likelihood criterion, is investigated. Solving the statistical mechanics of this problem we evaluate the quality of the estimation as a function of the number of data points, $P = \alpha N$, N being the dimensionality of the points, in the limit of large N. Below a critical value of α, the estimated density consists of Gaussian centers that have zero overlap with the structure of the true mixture. We show numerically that estimating the centers by slowly reducing the estimated Gaussian width yields a good agreement with the theory even in the presence of many local minima.

PACS number(s): 05.20.−y, 02.50.−r, 02.70.Lq

This paper investigates the problem of estimating the parameters of a probability density function from a finite number of sampled data. This problem has diverse applications in science and technology, including statistics, pattern recognition, classification and clustering, and neural information processing [1]. One of the most widely used methods is the maximum-likelihood (ML) estimate [1,2], which chooses the parameter values that maximize the probability of the sampled data. This and related estimation methods encounter two major difficulties. One is concerned with the quality of the estimate if the sample size is not sufficiently big. Despite intense research, relatively little progress has been made in the theoretical understanding of the performance of the ML estimate except for the asymptotic limit of large sample size. The second is the presence of many local minima of the log-likelihood function, which often renders simple gradient descent methods for maximizing this function inadequate [1]. These problems are particularly severe in the case of high-dimensional data, which is of a major interest, as many applications of the ML method deal with dimensionalities ranging from several tens to a few thousands.

An important special case is the problem of estimating densities that are composed of several, relatively simple, component densities. This case is relevant when the data is naturally decomposed into several clusters [1,2]. A central problem in this case is how to estimate the number of different components in the underlying density. Recent studies [3–5] suggested the use of a smoothing parameter σ, analogous to "temperature," that controls the resolution in which the data is scrutinized. In addition, clustering algorithms have been recently proposed [3] that are based on estimating the loci of the cluster centers by local minimization of an effective energy function which depends on the smoothing parameter. First, the centers are found at high σ and then one iterates the solution to low σ by incremental reduction of σ. This process is known as *deterministic annealing* (DA). Similar DA algorithms have been found useful in spin-glass models [6] and a variety of optimization problems [7], but they still lack an adequate theoretical underpinning. It is thus important to understand how smoothing parameters affect the surface of optimization cost functions.

We consider a stochastic source consisting of a mixture of two Gaussians,

$$\mathcal{P}_0(\mathbf{S}) = \frac{1}{2}\mathcal{P}(\mathbf{S}|\mathbf{U}_1^0,\sigma_0) + \frac{1}{2}\mathcal{P}(\mathbf{S}|\mathbf{U}_2^0,\sigma_0), \tag{1}$$

where $\mathbf{S}$ is a vector in $\mathbb{R}^N$ and

$$\mathcal{P}(\mathbf{S}|\mathbf{U},\sigma) = \frac{1}{(2\pi\sigma)^{N/2}} \exp\left(-\frac{1}{2\sigma}|\mathbf{S}-\mathbf{U}|^2\right). \tag{2}$$

We assume for simplicity that the two Gaussian centers $\mathbf{U}_l^0$ are orthogonal and have equal magnitude, which we denote by $u_0^2 = |\mathbf{U}_l|^2/\sigma_0$. Note that u_0 also measures the normalized separation of the two Gaussians, since $\sqrt{2\sigma_0}u_0 = |\mathbf{U}_1^0 - \mathbf{U}_2^0|$. The ML estimator assumes that the data are generated by the mixture

$$\mathcal{P}(\mathbf{S}|\{\mathbf{U}_l\},\sigma) = \frac{1}{2}\mathcal{P}(\mathbf{S}|\mathbf{U}_1,\sigma) + \frac{1}{2}\mathcal{P}(\mathbf{S}|\mathbf{U}_2,\sigma) \ . \tag{3}$$

The parameters of this distribution are estimated using a set of observed data points $\mathbf{S}^\mu$, $\mu = 1,...,P$, generated at random according to Eq. (1). Note that the labels of $\mathbf{S}^\mu$, namely, the identity of the component distribution that generated them, are not provided. Hence the estimate is an instance of unsupervised learning. The estimate is performed by minimizing the log-likelihood energy function

$$E(\{\mathbf{U}_l\},\sigma) = -\sum_{\mu=1}^{P} \ln \mathcal{P}(\mathbf{S}^\mu|\{\mathbf{U}_l\},\sigma). \tag{4}$$

We study this problem in the thermodynamic limit, $N \to \infty$, keeping $\alpha = P/N$, σ_0 and u_0 finite. As we have shown previously [8], in this limit the overlapping volume between the density components is a finite fraction of the total effective volume. Note that in this limit, the width of each of the Gaussians is bigger by a factor of $\sqrt{N}$ than their separation. Examining E in the thermodynamic limit, we find the form

 ©1994 The American Physical Society

$$E = P\epsilon(\{\mathbf{U}_l\};\sigma) + \frac{NP}{2}[\ln(2\pi\sigma) + \sigma_0/\sigma], \tag{5}$$

where ϵ is of order 1. Since the last term is of the order of N^2, it follows that the ML estimation of σ is always $\sigma = \sigma_0$, independent of the estimation of $\mathbf{U}_l$. This feature is an immediate consequence of the thermodynamic limit. Nevertheless, we will consider here σ as a control parameter and evaluate the vectors $\mathbf{U}_l$ by minimizing E for a fixed value of σ. The reason for this is twofold. First, for many applications, it is the quality of the centers' estimation that is of primary interest and, as we will show, it is not necessarily optimal at $\sigma = \sigma_0$ when α is finite. Second, it may be useful to use σ as an annealing parameter, in a DA procedure discussed above.

In order to determine the Gaussian centers it is convenient to introduce the mean vector $\bar{\mathbf{U}} \equiv (\mathbf{U}_1+\mathbf{U}_2)/2$ and the splitting vector $\Delta\mathbf{U} \equiv (\mathbf{U}_1 - \mathbf{U}_2)/2$. It is straightforward to show that

$$\epsilon(\Delta\mathbf{U}, \bar{\mathbf{U}}) = \frac{1}{2\sigma}\bar{\mathbf{U}}\cdot(\bar{\mathbf{U}} - 2\bar{\mathbf{S}}) + \frac{1}{2\sigma}\Delta\mathbf{U}^2 - \frac{1}{P}\sum_{\mu=1}^{P}\ln\cosh(\Delta\mathbf{U}\cdot\mathbf{S}^\mu/\sigma), \tag{6}$$

where $\bar{\mathbf{S}} = P^{-1}\sum_\mu \mathbf{S}^\mu$, i.e., the center of mass of the P data points. Minimizing ϵ with respect to $\bar{\mathbf{U}}$ and $\Delta\mathbf{U}$ yields

$$\bar{\mathbf{U}} = \bar{\mathbf{S}} \tag{7}$$

for all σ and α. The splitting vector obeys

$$\Delta\mathbf{U} = \frac{1}{P}\sum_{\mu=1}^{P}\mathbf{S}^\mu\tanh(\Delta\mathbf{U}\cdot\mathbf{S}^\mu/\sigma) \ . \tag{8}$$

The vector $\Delta\mathbf{U}$ is particularly important because it determines the quality of classification of new inputs based on the estimated mixture density. In our case, a ML classifier classifies an arbitrary input $\mathbf{S}$ as $+1$ if $(\mathbf{S}-\mathbf{U}_1)^2 < (\mathbf{S}-\mathbf{U}_2)^2$ and vice versa; see Ref. [8]. The average classification error of this classifier is

$$\epsilon_C = H\left(\frac{u_0}{\sqrt{2}}\cos\theta\right), \tag{9}$$

where $H(x) = \int_x^\infty \frac{dt}{2\pi}e^{-\frac{1}{2}t^2}$. The angle θ is the angle between $\Delta\mathbf{U}$ and $\Delta\mathbf{U}^0 = (\mathbf{U}_1^0 - \mathbf{U}_2^0)/2$, i.e., $\theta = \cos^{-1}(\sqrt{2r^2/q})$, where

$$q = \sigma_0|\Delta\mathbf{U}|^2, \qquad r = u_0^{-1}\Delta\mathbf{U}\cdot\Delta\mathbf{U}^0 \ . \tag{10}$$

The order parameter q measures the separation between the two estimated centers, whereas r measures the overlap between the true and the estimated centers.

We have calculated analytically the order parameters r and q. This has been done by viewing the ML estimate as a zero-temperature limit of a statistical mechanical system with a Gibbs distribution

$$P_G(\{\mathbf{U}_l\}) \propto \exp[-\beta E(\{\mathbf{U}_l\})], \tag{11}$$

where the energy is given by Eq. (4) and β^{-1} is the temperature of the system. Note that in our formulation, β^{-1} is not related to the width parameter σ. Using the replica method [9] we have solved the mean-field, replica-symmetric theory of this system in the zero-temperature limit, thereby determining q and r for all σ and α. The results reveal three distinct phases, as shown in Fig. 1.

(i) Unsplit phase, $q = r = 0$. For large σ and large α the solution with maximum likelihood is one with a single center, i.e., $\mathbf{U}_1 = \mathbf{U}_2 = \bar{\mathbf{U}}$. The existence of this state is guaranteed by the symmetry of E under the transformation $\mathbf{U}_{1,2} \to \mathbf{U}_{2,1}$. As σ is lowered this state loses its stability and the single center splits into two distinct clusters. The occurrence of this bifurcation as σ decreases is in agreement with previous predictions. Here we find that the nature of the split phase depends crucially on the size of the sample, measured by α.

(ii) Split ordered phase, $q, r \neq 0$. For values of α above $\alpha_c = 4u_0^{-4}$ the single-center phase becomes unstable to splitting at $\sigma_1(\alpha) = (1 + u_0^2/2)(1 + 2\alpha^{-1}u_0^{-2})$. Here and in the following the values of σ are quoted in units of σ_0. The two clusters that appear below this line are split in a direction which has a finite projection on the direction of splitting of the true centers, as signaled by the nonzero value of r; see Eq. (10).

(iii) Split random phase, $q \neq 0$, $r = 0$. For $\alpha < \alpha_c$, the splitting into two centers appears at $\sigma_2(\alpha) = (1+\alpha^{-\frac{1}{2}})^2$. Below this line, the direction of the splitting of the two estimated Gaussians is determined almost exclusively by the random clumpiness of the data. This means that r is of the order of $1/N$ and vanishes in the thermodynamic limit.

We discuss the impact of this phase diagram on the quality of classification performed by the ML classifier, defined above. First, let us discuss the unsplit phase. Since $r = q = 0$, it would seem that in this regime the classifier will have a random performance. This, however, may not be the case. Equation (9) shows that, in our case, the classification error does not depend on the absolute magnitude of the splitting of the two clus-

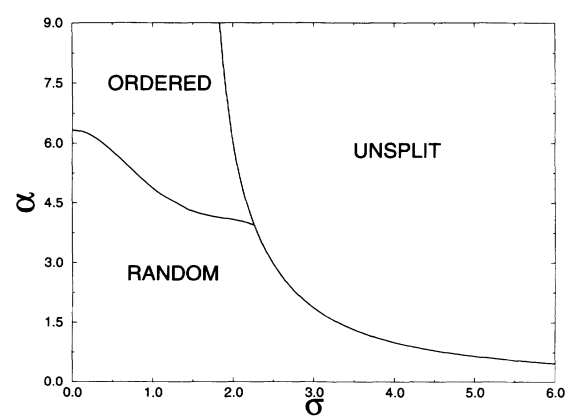

FIG. 1. Phase diagram for maximum-likelihood estimation of a mixture of two Gaussians for separation value $u_0 = 1$. The vertical axis denotes the number of data points per dimension, α. The horizontal axis denotes the width σ of the estimated Gaussians. The values of σ here and in the following figures are in units of the true width, σ_0.

ters but only on the angle between the splitting vector and $\Delta\mathbf{U}^0$, the value of which is ill defined in the unsplit phase, and is therefore dependent on the details of the dynamics. In the split ordered phase, $r, q > 0$; hence, $\epsilon_C < \frac{1}{2}$ as expected, as shown in Fig. 2. Note that as the transition to the unsplit phase is approached, ϵ_C remains below $\frac{1}{2}$. This is because, near the transition, $q \propto r^2$. On the other hand, $\epsilon_C = \frac{1}{2}$ throughout the random phase, since $\cos\theta = 0$. The results of Fig. 2 also show that the classification error is not minimal at the "correct" value of σ, namely, σ_0, but at a higher value. In fact, for these parameter values, the minimal error is achieved at the transition point $\sigma_1(\alpha)$. For other values of α or u_0 the minimum occurs at intermediate values of σ.

The above theoretical results do not reveal directly the existence of metastability in this system. To study this issue, we have performed computer simulations of the system by locally minimizing Eq. (4). The minimization was performed by iterating the map $\Delta\mathbf{U}(n+1) = \mathbf{F}(\Delta\mathbf{U}(n))$ where the function $\mathbf{F}$ is given by the right-hand side of Eq. (8). At each value of σ and α, a set of initial values of $\Delta\mathbf{U}$ was sampled at random from a Gaussian distribution. We have also studied the DA algorithm, in which σ was decreased in small steps from high values. At each value of σ, the initial condition for the iteration was the fixed point value of $\Delta\mathbf{U}$ at the previous value of σ (plus a small amount of noise). In Fig. 3, the results for the order parameter r are plotted as a function of σ for $P = N = 500$ and $u_0 = 2$. Note that for this value of u_0, $\alpha = 1$ is above α_c. For high values of σ, all initial conditions converge to a unique solution, which coincides with the DA solution. For $\sigma \leq 1.5$ many different local minima appear. As seen in Fig. 3, the value of r varies considerably across the spectrum of local minima, especially for small σ. Note that the onset of the metastability occurs well below the transition from the unsplit phase to the split ordered phase (at $\sigma = 4.5$). On the other hand, simulations at low α indicate that,

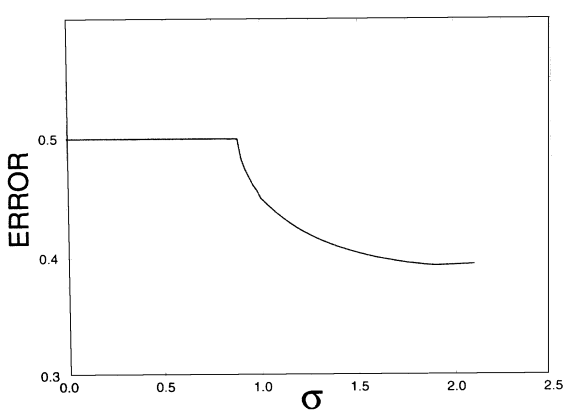

FIG. 2. Classification error as a function of σ for $\alpha = 5$, $u_0 = 1$. Width $\sigma = 2.1$ is the point of transition to the ordered phase (arrow). The error in the unsplit phase is ill defined (see text). Note that the error goes smoothly to 0.5 (random performance) at the crossing to the random phase ($\sigma = 0.90$).

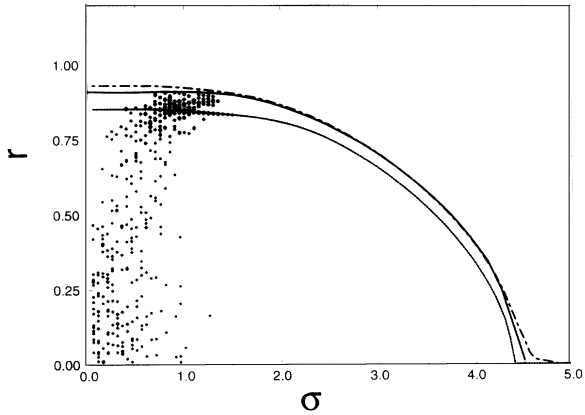

FIG. 3. The order parameter r as a function of σ for $\alpha = 1$ and $u_0 = 2$. The solid line is the result of the theory. The dashed line is the solution found by deterministic annealing from the local minimum at high values of σ; the dots are the solutions found for local minima generated numerically for $N = 500$ and random choices of initial conditions, at each σ. Note that for $\sigma \geq 1.5$ all initial conditions converge to the same solution.

for $\alpha < \alpha_c$, metastability appears, for large N, already for σ near the onset of the split random phase. The appearance of strong metastability suggests that at low σ and low α there appears a genuine thermodynamic spin-glass phase, marked by replica-symmetry breaking [9]. This spin-glass behavior occurs throughout the split random phase and in the low σ regime of the ordered phase. Testing this hypothesis requires evaluating the stability of the replica symmetric theory, which has not yet been done.

An interesting question is the relation between the DA solution and the theoretical results. In Fig. 3 we present the theoretical predictions for r together with the value obtained by averaging the DA results over 20 randomly sampled realizations of the sample points. The results reveal quite a reasonable agreement between the two. The deviations near $\sigma \approx 4.5$ are expected due to strong finite-size effects near the transition to the ordered phase. The deviations at low σ may reflect the effects of neglecting replica-symmetry breaking in the theory. Thus the results are consistent with the hypothesis that in this system the DA solution is in fact the global minimum of E or at least close to it. Further support for this hypothesis is gained from the fact that in our simulations of systems with $N \geq 500$ the DA solution was always lower in energy than the other local minima.

Further insight into the nature of the metastability is revealed by studying the evolution of individual local minima as σ is varied. This has been done by calculating local minima of E at some intermediate values of σ and following them by either decreasing or increasing σ incrementally. We have found that every existing local minimum varies smoothly with σ as σ decreases. On the other hand, upon increasing σ individual solutions disappear in a discontinuous fashion, and the system "jumps" to a different solution. An example is shown in Fig. 4,

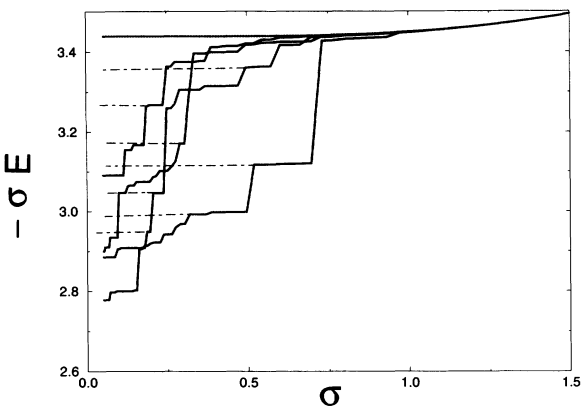

FIG. 4. Variation of the energies of individual local minima with σ, for $N = 500$. Solid (dashed) lines are the result of increasing (decreasing) σ. Only a few epochs of decreasing σ are shown. Upper line is the DA solution. Parameters are as in Fig. 3.

in which the evolution of the energies of several local minima is displayed. The solid lines are the result of increasing σ. Each time a jump occurs it represents a disappearance of a minimum and a transition to another existing minimum. This is demonstrated by dashed lines which show the evolution of the energies of several "new" minima upon reducing σ. Remarkably, we have found that, for $N \geq 500$, the energies of two existing minima never cross as σ decreases. This implies, in particular, that the DA solution remains the lowest energy solution for all σ, since when other local minima first appear they must be higher in energy than the DA state.

It is interesting to compare our results with the dependence of the local minima on temperature in spin glasses. Numerically, it has been found that following high temperature minima of the mean-field free energy of the infinite-range spin glass to low temperatures does not yield physical solutions at low temperature [10]. Furthermore, it has been argued on theoretical grounds that the correlations between spin-glass states at different temperatures or fields vanish in the thermodynamic limit [11]. On the other hand, our results are similar to those found for the *naive mean-field* equations for long- and short-range spin glasses, in the presence of a nonzero field [6]. Also, similar behavior has been found for a different, clustering cost function at low dimensions [4]. In our case, significant metastability begins to appear only for $N \geq 20$.

In conclusion, we have evaluated the phase diagram of a model of high-dimensional mixture density estimation by maximum likelihood. Below a critical (scaled) sample size α, the splitting of the estimated density into multiple different components is completely dominated by the random sampling and the components have negligible overlap with the centers in the underlying density. In addition, we find numerically that at low estimated width of the component densities (σ) or small α many local minima of the log-likelihood energy appear, which are reminiscent of a spin-glass phase. Our results indicate that, as σ is reduced, the surface of E roughens, thereby creating new minima, but the old minima and the order of their energies are not disrupted. This suggests that deterministic annealing may indeed yield the ground state of the system even in the presence of strong metastability.

Note added. After the completion of this work, we became aware of two recent papers which study unsupervised learning using statistical mechanics [12,13].

We are grateful to H. S. Seung for extremely helpful discussions on this work. Useful discussions with J. Buhmann, D. Hansel, G. Mato, and N. Tishby are also acknowleged. This research is partially supported by the Fund for Basic Research of the Israeli Academy of Science and Arts.

[1] R. A. Radner and H. F. Walker, SIAM Rev. **26**, 195 (1984), and references therein.
[2] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis* (Wiley, New York, 1973).
[3] K. Rose, E. Gurewitz, and G. Fox, Phys. Rev. Lett. **65**, 945 (1990).
[4] Y. Wong, Neural Comput. **5**, 89 (1993).
[5] J. Buhmann and H. Kuhnel, Neural Comput. **5**, 75 (1993).
[6] C. M. Soukoulis, K. Levin, and G. M. Grest, Phys. Rev. Lett. **48**, 1756 (1982).
[7] A. L. Yuille, Neural Comput. **2**, 1 (1990).
[8] N. Barkai, H. S. Seung, and H. Sompolinsky, Phys. Rev. Lett. **70**, 3167 (1993).
[9] M. Mezard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
[10] D. Ling *et al.*, Phys. Rev. B **23**, 262 (1983).
[11] K. Binder and A. P. Young, Rev. Mod. Phys. **58**, 801 (1986); H. Sompolinsky (unpublished).
[12] T. L. H. Watkin and J. P. Nadal, J. Phys. A **27**, 1899 (1994).
[13] M. Biehl and A. Mietzner, J. Phys. A **27**, 1885 (1994).