Towards EM-style Algorithms for *a posteriori* Optimization of Normal Mixtures

Eric S. Ristad Peter N. Yianilos

November 14, 1997*

Abstract

Expectation maximization (EM) provides a simple and elegant approach to the problem of optimizing the parameters of a normal mixture on an unlabeled dataset. To accomplish this, EM iteratively reweights the elements of the dataset until a locally optimal normal mixture is obtained. This paper explores the intriguing question of whether such an EM-style algorithm exists for the related and apparently more difficult problem of finding a normal mixture that maximizes the *a posteriori* class probabilities of a labeled dataset.

We expose a fundamental degeneracy in the relationship between a normal mixture and the *a posteriori* class probability functions that it induces, and use this degeneracy to prove that reweighting a dataset can almost always give rise to a normal mixture that exhibits any desired class function behavior. This establishes that EM-style approaches are sufficiently expressive for *a posteriori* optimization problems and opens the way to the design of new algorithms for them.

Keywords: *Expectation Maximization (EM), Normal Mixtures, Gaussian Mixtures, Supervised Learning, Maximize Mutual Information (MMI).*

*Eric Sven Ristad is with the Department of Electrical Engineering and Computer Science University of Illinois at Chicago (email: `ristad@eecs.uic.edu`), and is partially supported by Young Investigator Award IRI-0258517 from the National Science Foundation. Peter N. Yianilos is with the NEC Research Institute, 4 Independence Way, Princeton, NJ 08540 (email `pny@research.nj.nec.com`). An earlier version of the paper was distributed as a technical report entitled "On the Strange *a Posteriori* degeneracy of Normal Mixtures, and Related Reparameterization Theorems", and appeared in the second author's thesis [8].

1 Introduction

A finite *normal mixture* is a stochastic combination of normal densities. That is, a probability density function p on $\mathbb{R}^d$ of the form $p(x) = \sum_{i=1}^{k} m_i N_{\Sigma_i,\mu_i}(x)$ where $m_1, \ldots, m_k \geq 0$ with $\sum_{i=1}^{k} m_i = 1$, and $N_{\Sigma_i,\mu_i}(x)$ denotes the normal density with mean μ_i and covariance Σ_i. Each constituent normal density is referred to as a *component* of the mixture, and $m_1, \ldots, m_k$ are the *mixing coefficients*. Normal mixtures have proven useful in several areas including pattern recognition [4] and speech recognition [6, 2, 5] – along with vector quantization and many others.

The problem of finding a k-component normal mixture M that maximizes the likelihood $\prod_i p(s_i|M)$. of an unlabeled dataset $s_1, \ldots, s_n$ may be approached using the well-known expectation maximization (EM) algorithm [1, 3, 7]. EM iteratively re-weights each sample's membership in each of the k mixture components by the posteriori probability of the components given the sample.

Normal mixtures are also applied to pattern classification problems. For classification problems, each mixture component may be thought of as corresponding to a pattern class. Given a class label $\omega(s_i)$ for each element s_i in our dataset, we would like to maximize $\prod_i p(\omega(s_i)|s_i, M)$. Here the goal is to optimize the mixture's *a posteriori* performance, that is to predict the correct labels, not model the observation vectors themselves. This is sometimes called the *maximum mutual information* (MMI) criterion in the speech recognition literature [2] and may be viewed as probabilistic supervised learning.

A naive algorithm for this problem segregates the data elements by class, constructs the maximum-likelihood normal density for each class, and combines the resulting densities to form a mixture. This approach can be far from optimal, as illustrated by figure 1. Part (a) of the figure depicts the maximum-likelihood normal densities corresponding to the pair of points in each class. These densities are specified by the unweighted sample mean and unweighted sample covariance of each class in isolation. Although the resulting mixture maximizes the probability of the samples, it does not maximize the conditional probability of the classes given the samples. For example, the density shown in part (b) assigns greater conditional probability to the classes given the samples, although it assigns less probability to the samples themselves. To accomplish this, increased weight is assigned to the outermost two points in part (b). In this example the classes neatly divide and the optimal densities have zero-variance corresponding to placing all of the emphasis on a single element of each class. We view these weights as placing *emphasis* on the dataset, resulting in a modified collection of normal densities.

The question that motivates this paper is then:

> Does there exist an EM-style algorithm for the *a posteriori* maximization of normal mixtures?

By "EM-style" we mean an iterative algorithm that operates by varying the emphasis placed on a dataset rather than by viewing the model's natural parameters as variables in the optimization.

One might certainly imagine one based perhaps on the common machine learning heuristic where one increases the weight of a mistake in order to improve classification accuracy – a technique sometimes referred to as corrective training. But it is not at all clear that emphasis algorithms have the expressive power to discover the solution. The convex reweighting performed by EM cannot, for example, generate a mean vector outside of the dataset's convex hull.

This paper takes the first step towards such an algorithm by bringing the expressive power of emphasis schemes more clearly into focus. Section 2 defines two normal mixtures to be *class*

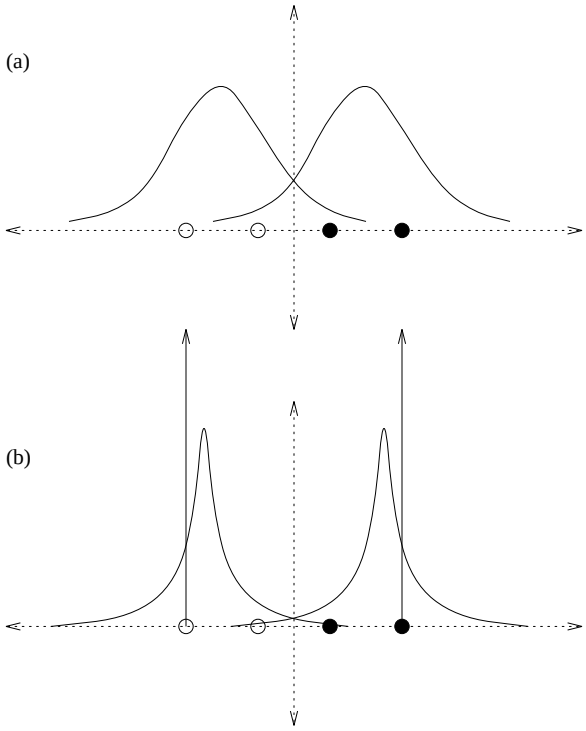

Figure 1: A simple one-dimensional classification problem. An even mixture formed with the two densities of part (a), each an optimal normal model of a single class, does not perform as well as a mixture formed with the densities of part (b) — where the performance measure is the probability with which the mixture predicts the correct label. The smaller variance densities perform better because their tails approach zero more rapidly so that less probability is ascribed to the incorrect class.

equivalent if they induce identical *a posteriori* class functions, i.e. perform identically as probabilistic classifiers. It is shown there that the relationship between mixtures and their class behavior is highly degenerate, and theorem 1 characterizes this degeneracy in a way that is useful for our objective. As a positive result of this degeneracy one can search the entire space of class functions without considering all possible mixtures. So to solve the *a posteriori* maximization problem above, it suffices to express *any* normal mixture that induces optimal class functions. We are then led to ask:

> When can emphasis of an observation set induce a mixture that is class equivalent to an arbitrary one?

Obviously the observation set must satisfy certain orthodoxy conditions such as being of sufficient size and possessing in some sense enough independence. But beyond these issues the fact that many mean vectors and covariance matrices are not expressible by emphasis would seem to present a much larger obstacle.

A positive answer to this question is given by theorem 2 of section 3. It states that one can almost always exactly match the *a posteriori* behavior of an arbitrary mixture by an emphasis method that slightly modifies the EM approach. Therefore any optimization problem that depends only on a mixture's induced class functions can be solved using this form of emphasis.

We conclude our introduction by reviewing EM and presenting our view of it as an emphasis reparameterization scheme.

Let $s_1, \ldots, s_n \in \mathbb{R}^d$ be a set of observations. Any set of positive weights $\gamma_1, \ldots, \gamma_n$, gives rise in a natural way to a normal density. That is, the normal density having mean:

$$\mu = \frac{1}{\sum_{i=1}^{n} \gamma_i} \sum_{i=1}^{n} \gamma_i s_i \tag{1}$$

and covariance:

$$\Sigma = \frac{1}{\sum_{i=1}^{n} \gamma_i} \sum_{i=1}^{n} \gamma_i (s_i - \mu)(s_i - \mu)^t \tag{2}$$

which may be thought of as the weighted maximum likelihood (ML) model corresponding to the observations. Given k sets of weights, as many normal densities arise. Adding k additional positive *mixing* weights $w_1, \ldots, w_k$ serves to specify a mixture of these k densities, with the probability of the ith component given by $w_i / \sum_{i=1}^{k} w_i$. Thus a total of $nk + k$ weights induce a normal mixture, and we loosely refer to them as *emphasis parameters*.

The well known expectation maximization (EM) method may be viewed as a process which accepts as input a normal mixture, and outputs an emphasis parameter set – from which an improved normal mixture arises. This cycle continues as the algorithm climbs towards a local maximum, which is a stationary point of the iteration. The global maximum is also stationary point.

More precisely, Given data values $s_1, \ldots, s_n$ and a normal mixture M, one begins by computing $p(\omega_i | s_j, M), \forall 1 \leq i \leq k, 1 \leq j \leq n$. Conceptually the result is a $k \times n$ table of nonnegative values. The rows of this table are then normalized so that their sum is one. Each row then corresponds to a convex combination of the samples. The entries along a row are thought of as *weights* attributed to the sample. Each improved mean vector μ_i is merely the corresponding weighted average (convex combination) of the sample vectors. Each improved covariance is the sample convex combination of outer products $(s_j - \mu_i)(s_j - \mu_i)^t$. The improved mixing parameters are obtained by normalizing the vector consisting of the table row weights prior to their normalization. This process is repeated in an elegant cycle, i.e. each table induces a mixture that gives rise to a new table, and so on.

From our viewpoint, what is interesting here is that it is possible to search for an optimal normal mixture, by searching over the space of values for the $nk + k$ emphasis parameters – rather than using the canonical parameterization consisting of the unknown mean vectors, covariance matrices, and mixing coefficients.

Only a locally optimal solution is found through EM but it is important to realize that the global optimum is a stationary point of the iteration. This means that it can be expressed via emphasis, i.e. that there exists a table which induces the globally optimal mixture. This is interesting because not all mixtures can be induced from the table. In particular, as noted earlier, the induced means must lie within the convex hull of the sample set, and the covariances are similarly constrained.

2 The Strange *a Posteriori* degeneracy of Normal Mixtures

This section illuminates certain fundamental aspects of the nature of normal mixtures. Thinking of each mixture component as a *class*, we focus on the corresponding *a posteriori* class probability functions. It is shown that the relationship between these functions and the mixture's parameters,

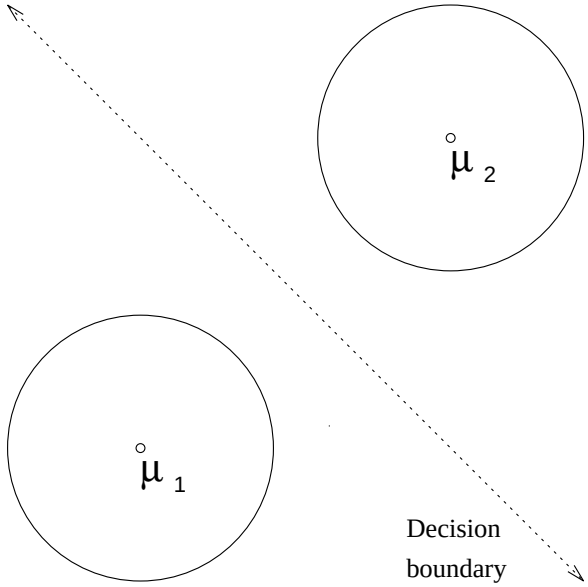

Figure 2: A depiction of an even mixture of two normal densities in $\mathbb{R}^2$, each with identity covariance. The two classes are separated by a linear decision boundary.

is highly degenerate – and that the precise nature of this degeneracy leads to somewhat unusual and counter-intuitive behavior. Even complete knowledge of a mixture's *a posteriori* class behavior, reveals essentially nothing of its absolute nature, i.e. mean locations and covariance norms. Consequently a mixture whose means are located in a small ball anywhere in space, can *project* arbitrary class structure everywhere in space.

A convenient visualization of class functions, focuses on *decision boundaries*, i.e. the surfaces along which classification is ambiguous. Imagery like ours in figure 2, and pages 28–31 of [4], suggest an intuitive relationship between mixture component locations, and the resulting *a posteriori* class structure and decision surfaces. One imagines each mean to be asserting ownership over some volume of space surrounding it. This view is however far from the truth and our theorem 1 reveals that the *a posteriori* class behavior of normal mixtures is far stranger and counter-intuitive than is generally understood. It shows that knowledge of a mixture's *a posteriori* class structure (which may be thought of as decision surfaces), tells us essentially nothing about the absolute location of the mixture's means – or the absolute nature of its covariances. One can easily imagine a normal mixture, and picture its corresponding class structure. Now if one is instead given only this class structure, the theorem says that infinitely many normal mixtures are consistent with it – and in particular that the means of such a mixture can be located within an arbitrarily small ball, located anywhere in space.

Despite the popularity and significant role of normal mixtures in statistical modeling and pattern recognition, the precise nature of this fascinating and highly degenerate relationship between class functions and mixtures has received little attention. Part of the reason may be that the simple view in which component means dominate some portion of the space around them is exactly the case when all covariances have identical determinant, and uniform mixing coefficients are used. The most common example of this setting consists of nearest-neighbor Euclidean clustering which corresponds to the identity covariance case. In practice, if the determinants are somewhat comparable, and the

mixing coefficients nearly uniform, the simple view is almost always valid. But in many cases, such as when mixture parameters are estimated using an optimization procedure, no such constraint is imposed, and the means, covariance matrices, and mixing coefficients are free to assume any values. Here the full range of strange behavior may be expected.

We now turn to the section's main mathematical development. Examples and discussion follow and the section ends with a technical convergence-rate result that is needed to establish the next sections' main result.

Definition 1 *A finite mixture is a probability function on a measure space $\mathcal{X}$, arising from k conditional probability functions (components) as follows:*

$$p(x) = \sum_{i=1}^{k} p(x, \omega_i) = \sum_{i=1}^{k} p(x|\omega_i)p(\omega_i)$$

where the constants $\{p(\omega_i)\}$ are referred to as the mixing coefficients. *Any such mixture induces k a* posteriori *class functions:*

$$p(\omega_i|x) = \frac{p(x|\omega_i)p(\omega_i)}{p(x)}$$

Two finite mixtures are class-equivalent if they induce the same class functions.

The class functions are not independent since they are constrained by: $\sum_i p(\omega_i|x) = 1$. We begin with a simple proposition, which allows us in what follows to focus on ratios of conditional probabilities, and ignore constant factors. This will be important since the nonconstant portion of the ratio of two normal densities is a rather simple object.

Proposition 1 *Let $p(x)$ be a k-component mixture, and $p'(x|\omega_1), \ldots, p'(x|\omega_k)$ be a collection of strictly positive conditional probability functions. If for some j, there exist constants $C_{i,j}$ such that $\forall i \neq j$ and $x \in \mathcal{X}$:*

$$\frac{p(x|\omega_i)}{p(x|\omega_j)} = C_{i,j} \cdot \frac{p'(x|\omega_i)}{p'(x|\omega_j)} \tag{3}$$

then there exist mixing coefficients $p'(\omega_1), \ldots, p'(\omega_k)$ such that the resulting mixture $p'(x)$ is class-equivalent to $p(x)$.

proof: We begin by giving a formula for mixing coefficients $p'(\omega_1), \ldots, p'(\omega_1)$, such that:

$$\underbrace{C_{i,j}\frac{p(\omega_i)}{p(\omega_j)}}_{k_{i,j}} = \frac{p'(\omega_i)}{p'(\omega_j)} \tag{4}$$

Relaxing for the moment the restriction $\sum_{\ell=1}^{k} p(\omega_\ell) = 1$, observe that if $p(\omega_j)$ were 1, then setting $p(\omega_i) = k_{i,j}, i \neq j$, would satisfy the equation. Normalization then yields a solution which does sum to 1:

$$
\begin{aligned}
p'(\omega_j) &= \frac{1}{1+\sum_{\ell\neq j} k_{\ell,j}} \\
p'(\omega_i) &= \frac{k_{i,j}}{1+\sum_{\ell\neq j} k_{\ell,j}} \quad i \neq j
\end{aligned}
$$

Multiplying each side of Eq. 4 by the corresponding side of Eq. 3 yields:

$$
\frac{p(x|\omega_i)p(\omega_i)}{p(x|\omega_j)p(\omega_j)} = \frac{p'(x|\omega_i)p'(\omega_i)}{p'(x|\omega_j)p'(\omega_j)} \quad \forall i \neq j, x \in \mathcal{X} \tag{5}
$$

The approach above works so long as $p(\omega_j) \neq 0$. If it is zero, then p may be treated as a $k-1$ component mixture, and the proposition follows by induction. Till now j has been fixed and Eq. 5 is established for $i \neq j$. This equation is however easily extended to any pair i, ℓ of indices. Let f_i denote $p(x|\omega_i)p(\omega_i)$ and f' denote $p'(x|\omega_i)p'(\omega_i)$. Then:

$$
\frac{f_i}{f_\ell} = \frac{f_i}{f_j} \cdot \frac{f_j}{f_\ell} = \frac{f'_i}{f'_j} \cdot \frac{f'_j}{f'_\ell} = \frac{f'_i}{f'_\ell}
$$

Now we may write:

$$
p(\omega_i|x) = \frac{f_i}{\sum_{\ell=1}^{k} f_\ell} = \frac{1}{1+\sum_{\ell\neq i} f_\ell/f_i}
$$

and a corresponding expression for $p'(\omega_i|x)$ in terms of $\{f'_\ell\}$. We have already seen that all ratios $f_i/f_\ell = f'_i/f'_\ell$, so $\sum_{\ell\neq i} f_\ell/f_i = \sum_{\ell\neq i} f'_\ell/f'_i$ whence $p(\omega_i|x) = p'(\omega_i|x)$ and we are done. □

We now specialize our discussion to mixtures whose components are normal densities.

Definition 2 *A d-dimensional k-component <u>normal mixture</u>, is a finite mixture of multivariate normal densities:*

$$
N_{\mu,\Sigma}(x) \triangleq \frac{1}{(2\pi)^{d/2}\Sigma^{1/2}} \cdot e^{-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)} \tag{6}
$$

The mixture's <u>parameters</u> are then $\Phi = \{\Sigma_1, \ldots, \Sigma_k, \mu_1, \ldots, \mu_k, p(\omega_1), \ldots, p(\omega_k)\}$.

The following theorem exposes the large degeneracy in the relationship between normal mixtures, and their induced class functions.

Theorem 1 *Let p be a d-dimensional normal mixture with k components. For any $x \in \mathbb{R}^d$ and $\epsilon > 0$, there exists a d-dimensional k-component normal mixture p', such that for $1 \leq i \leq k$:*

1. $\|\mu'_i - x\| < \epsilon$
2. $\|\Sigma'_i\|_2 < \epsilon$
3. *p' and p are class-equivalent*

where $\|\Sigma'_i\|_2$ refers to the Frobenius/Euclidean matrix norm.

proof: Begin by selecting some distinguished component j. By proposition 1 we have only to produce $\Sigma'_1, \ldots, \Sigma'_k$ and $\mu'_1, \ldots, \mu'_k$ such that the ratio conditions of the proposition are satisfied. Since constant factors do not matter, we ignore several portions of the definition of a normal density Eq. 6. Namely, the leading constant and the constant portion of the exponent once multiplied out. The result is that:

$$N_{\mu,\Sigma}(x) \propto e^{-\frac{1}{2}(x^t\Sigma^{-1}x - 2\mu^t\Sigma^{-1}x)}$$

The proportional ratio condition of proposition 1 then leads immediately to the following necessary and sufficient conditions:

$$\left.\begin{array}{rcl} \Sigma_i^{-1} - \Sigma_j^{-1} & = & \Sigma'^{-1}_i - \Sigma'^{-1}_j \\ \mu_i^t\Sigma_i^{-1} - \mu_j^t\Sigma_j^{-1} & = & \mu'^t_i\Sigma'^{-1}_i - \mu'^t_j\Sigma'^{-1}_j \end{array}\right\} \quad \forall i \neq j \tag{7}$$

We set $\mu'_j = x$ and begin by sketching the rest of the proof. The constraints are satisfied by choosing Σ'^{-1}_j so that each of its eigenvalues is *large*. Each resulting Σ'^{-1}_i must then be positive definite and will also have large eigenvalues. Both Σ'_j and Σ'_i then have small norm, satisfying the requirements of the theorem. In this process, it is only Σ'^{-1}_j we are free to choose. Each choice determines the Σ'_i, and μ'_i (since we have already set $\mu'_j = x$). In the limit, as we choose Σ'^{-1}_j with increasingly large eigenvalues, $\|\mu'_i - \mu'_j\| = \|\mu'_i - x\|$ approaches zero whence we can satisfy the theorem's other condition.

Denote by $\overline{\lambda}(A)$ the largest eigenvalue of positive definite matrix A, and by $\underline{\lambda}(A)$ the smallest. For matrices A, B, it then follows easily from the Cauchy-Schwartz inequality that $\underline{\lambda}(A+B) \geq \underline{\lambda}(A) - \overline{\lambda}(B)$, by writing $\|[(A+B)+(-B)]u\| \leq \|(A+B)u\| + \|-Bu\|$ where u denotes any unit length vector. Now the first constraint in Eq. 7 may be written:

$$\Sigma'^{-1}_i = \Sigma'^{-1}_j + (\Sigma_i^{-1} - \Sigma_j^{-1}) \tag{8}$$

and we then have:

$$\underline{\lambda}(\Sigma'^{-1}_i) \geq \underline{\lambda}(\Sigma'^{-1}_j) - \overline{\lambda}(\Sigma_i^{-1} - \Sigma_j^{-1})$$

The parenthesized term is constant since we are given both Σ_i^{-1} and Σ_j^{-1}, and by subadditivity of symmetric matrix spectral radii, does not exceed their sum $\overline{\lambda}(\Sigma_i^{-1}) + \overline{\lambda}(\Sigma_j^{-1})$. We can easily choose Σ'^{-1}_j such that $\underline{\lambda}(\Sigma'^{-1}_j)$ is arbitrarily large, e.g. $c \cdot I$ where c is a large positive constant. It follows then that $\underline{\lambda}(\Sigma'^{-1}_i)$ will also be arbitrarily large, ensuring that it is positive definite.

Next recall that the column norms of a matrix A cannot exceed its spectral radius (operate on each member of the canonical basis). In our positive definite setting each is then bounded above by $\overline{\lambda}(A)$, so that $\|A\|_2 \leq \sqrt{d}\overline{\lambda}(A)$. Choosing the smallest eigenvalue of Σ'^{-1}_j and Σ'^{-1}_i to be arbitrarily large, forces $\overline{\lambda}(\Sigma'_j)$ and $\overline{\lambda}(\Sigma'_i)$ to be arbitrarily small – satisfying the theorem's first condition.

We set μ'_j to be equal to x from the statement of the theorem, and the second constraint in Eq. 7 may be rearranged to yield:

$$\mu'_i = (\Sigma'_i\Sigma'^{-1}_j)x + [\Sigma'_i(\Sigma_i^{-1}\mu_i + \Sigma_j^{-1}\mu_j)] \tag{9}$$

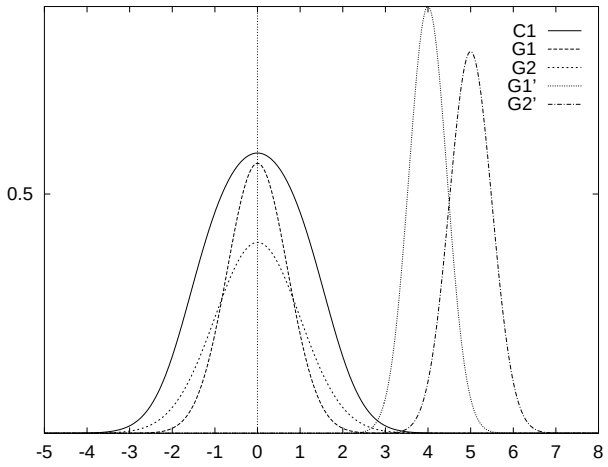

Figure 3: Two ways to induce a single class function. An illustration of the degeneracy in the relationship between normal mixtures, and their induced class functions. Zero mean normal densities G1 and G2 when mixed evenly induce C1, the *a posteriori* probability of G1. Normal densities G1' and G2' located in the right portion of the graph do not have zero means, but may be mixed (in a far from even way) so that C1 results.

By our earlier discussion, we may force Σ'_i to be arbitrarily close to zero – forcing the bracketed term in Eq. 9 to approach zero as well. We next show that the first parenthesized term $\Sigma'_i\Sigma'^{-1}_j$ tends to the identity matrix I as $\underline{\lambda}(\Sigma'^{-1}_j) \to \infty$. This then demonstrates that $\mu'_i \to x$ satisfying the theorem's second condition, and finishes the proof.

It is easy to see that $\Sigma'_i\Sigma'^{-1}_j \to I$ if and only if $(\Sigma'_i\Sigma'^{-1}_j)^{-1} = \Sigma'_j\Sigma'^{-1}_i \to I$. Using Eq. 8 this becomes:

$$I + \Sigma'_j(\Sigma_i^{-1} - \Sigma_j^{-1}) \to I$$

which is clearly the case since $\Sigma'_j \to 0$. □

To recapitulate, any location for μ'_j, and sufficiently large (in the $\underline{\lambda}$ sense) Σ'^{-1}_j, will give rise using the proof's constructions, to values for the other means, covariances, and mixture coefficients, such that a class-equivalent mixture results. The proof goes on to establish that in the limit, everything is confined as required within an ϵ-neighborhood.

Figure 3 provides a simple one dimensional example of the normal mixture to class function degeneracy. The left pair of Gaussians G1,G2 both have zero mean. The taller one, G1 corresponds to the first class and has variance 1/2, while G2 has variance 1 and corresponds to the second class. When mixed evenly, the *a posteriori* probability C1 results. Notice that C1 assumes value 1/2 where G1 crosses G2. The right pair of Gaussians G1',G2' have means at 4 and 5, and variances of 1/5 and 1/4 respectively. An even mixture of them would result in a class function very different from C1. But if very little weight ($\approx 5.74234 \times 10^{-5}$) is placed on G1', its mode drops under the graph of G2', and surprisingly, the induced class function is exactly C1. The parameters of this new mixture follow immediately from the calculations within the proof of theorem 1.

Figure 4 depicts another two class problem, this time in two dimensions. Here various elements of each class are arranged so that they may be perfectly separated by a circular decision boundary. It

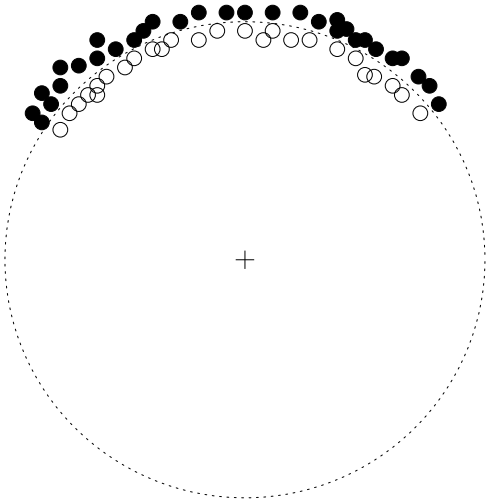

Figure 4: If data points of two colors are imagined to hug the inside and outside of a circular decision boundary, one might think that a Gaussian mixture model capable of separating the classes, would necessarily have component means at the center – but this is not the case.

is clear that we can create such a boundary by modeling each class with a normal density centered at the origin. We may evenly mix these components and set their covariance matrices to be multiples of the identity matrix; chosen so that the two density's intersection is the desired circle. This is in some sense the canonical solution but infinitely many others are possible. To begin with, it is not necessary that both covariance matrices be multiples of the identity. It suffices that their difference be such a multiple. It is not necessary that their means be located at the origin. Given a choice of covariance matrices, and the location of one mean, a location for the other may be computed which gives rise to the same boundary. Moreover, by choosing the covariances to be nearly zero, corresponding to highly peaked densities, the two means may be located within an arbitrary ϵ-neighborhood anywhere in $\mathbb{R}^2$. This illustrates the degeneracy by focusing only on the decision boundary, but theorem 1 shows that the family of class functions may be matched everywhere. The situation in figure 4 was first contrived in an attempt to disprove our *emphasis reparameterization theorem*, the subject to which we next turn. The data points are arranged along only part of the circle so that no convex combination of them, can possibly express the origin. That is, the origin is outside of the convex hull of the dataset. At that time we thought that this would prevent such combinations from leading to a mixture capable of inducing the desired circular decision boundary. The revelation of theorem 1 is that the component means of such a mixture can be located anywhere.

We have seen that all the Σ'_i converge to Σ'_j as the latter tends to zero. At the same time, the μ'_i approach μ'_j. The rate of convergence is our next topic. In particular we will show that $|\Sigma'_i - \Sigma'_j| \to 0$ quadratically as $\Sigma'_j \to 0$. This is formalized in the following proposition which is needed to establish the main result of the next section. We remark that the convergence of each μ'_i to μ'_j is only linear.

Proposition 2 $\|\Sigma'_i - \Sigma'_j\| = O(\overline{\lambda}^2(\Sigma'_j))$

proof: The difference $\Sigma_i^{-1} - \Sigma_j^{-1}$ is constant and is denoted by C. Then:

$$\Sigma'_i - \Sigma'_j = (\Sigma'^{-1}_j + C)^{-1} - (\Sigma'^{-1}_j)^{-1}$$

Denoting Σ'^{-1}_j as P for clarity this becomes:

$$\begin{aligned}(P+C)^{-1} - P^{-1} &= [(I + CP^{-1})P]^{-1} - P^{-1} \\ &= P^{-1}[I - (I + CP^{-1})^{-1}] \end{aligned} \tag{10}$$

Now clearly $I + CP^{-1} \to I$ with $\overline{\lambda}(P^{-1})$ and the eigenvalues of $I + CP^{-1}$ approach unity at this rate. So $(I + CP^{-1})^{-1} \to I$ as well – also with $\overline{\lambda}(P^{-1})$. Hence $\|I - (I + CP^{-1})^{-1}\| = O(\overline{\lambda}(P^{-1}))$ in Eq. 10. But $\|P^{-1}\| = O(\overline{\lambda}(P^{-1}))$ as well, so $\|(P+C)^{-1} - P^{-1}\| = O(\overline{\lambda}^2(P^{-1}))$. □

3 Emphasis Reparameterization

The question "Can emphasis induce a mixture class-equivalent to an arbitrary one?" posed at the beginning of this paper is reasonable since from theorem 1 we know that the means of such a mixture can lie within an arbitrarily small ball located anywhere in space. So confinement to the convex hull of the samples is not necessarily a restriction. Likewise the covariances can be chosen to have arbitrarily small Frobenius norm.

The main result of this section is a qualified positive answer to this question. It states that a modified form of emphasis can, for almost all sufficiently large observation sets, induce a mixture which is class equivalent to an arbitrary one.

Modifications are needed because even though the means and covariances may be chosen with a great deal of freedom, they are still highly constrained by the conditions of Eq. 7 – and it is sometimes impossible to generate a mixture which has satisfies these constraints by the form of emphasis employed by EM. The next section expands on the limits of strict EM-style emphasis.

The required modifications consist of a single change to Eq. 2 which becomes:

$$\Sigma = \sum_{i=1}^{n} \gamma_i (s_i - \mu)(s_i - \mu)^t \tag{11}$$

where the leading normalization has been removed. This allows the *scale* of the covariance to be adjusted without affecting the mean.

Before stating and proving our reparameterization result, a particular matrix is defined whose nonsingularity is a condition of the theorem.

Condition 1 *Denote by S_i the column vector formed by concatenating $s_i s_i^t$ and s_i, and let n (the number of samples) equal $d(d+3)/2$. Next form square matrix S by combining these columns. For example, if $d = 2$ we have:*

$$S \triangleq \begin{pmatrix} s_{1,1}s_{1,2} & s_{2,1}s_{2,2} & s_{3,1}s_{3,2} & s_{4,1}s_{4,2} & s_{5,1}s_{5,2} \\ s_{1,1}^2 & s_{2,1}^2 & s_{3,1}^2 & s_{4,1}^2 & s_{5,1}^2 \\ s_{1,2}^2 & s_{2,2}^2 & s_{3,2}^2 & s_{4,2}^2 & s_{5,2}^2 \\ s_{1,1} & s_{2,1} & s_{3,1} & s_{4,1} & s_{5,1} \\ s_{1,2} & s_{2,2} & s_{3,2} & s_{4,2} & s_{5,2} \end{pmatrix}$$

Our condition is that S be nonsingular.

It is not hard to show that this condition is almost always satisfied, a fact formalized by:

Proposition 3 *Regarding a sample of size $d(d+3)/2$ as a single random variable of dimension $d^2(d+3)/2$, the set of such vectors which give rise to singular S matrices, has measure zero.*

proof: Since the matrix consists of monomials of distinct structure, no non-zero linear combination (over the ring of polynomials) of its columns or rows vanishes. So its determinant, as a polynomial, is not identically zero. But the roots of a non-zero polynomial form a set of zero measure. □

Next we simplify our setting, adjust notation, and establish one more proposition before turning to the theorem. Without loss of generality, we may assume that our samples have mean zero. To see this, first let t denote some target mean, where we seek stochastic γ values such that $\sum \gamma_i s_i = t$. We may equivalently solve $\sum \gamma_i s'_i = t'$, where $s'_i \triangleq s_i - E[S]$ and $t' = t - E[S]$, since $\sum \gamma_i s'_i = (\sum \gamma_i s_i) - E[S]$. The problem of expressing a target covariance is unchanged by coordinate translation, since $(s_i - t) = (s'_i - t')$. Notice that this is true even if the γ are not stochastic. So in what follows, we will drop the *prime* notation and simply assume that $E[\{s_i\}] = 0$.

The mixture we generate will have all of its means located near zero, the mean of the sample set. The distinguished mixture component, identified with index j in the previous section, is located exactly at the origin. The location of the other mean vectors, are thought of as small displacements δ from the origin. We then write $s_i(\delta)$ to mean $s_i - \delta$, and the S matrix is then parameterized by δ in the obvious way, and denoted $S(\delta)$. Note that $S(0)$ is just S as originally defined.

Proposition 4 *Given nonsingular S, $\exists c > 0$ such that $\forall \delta$ satisfying $\|\delta\| < c$, $S(\delta)$ is nonsingular.*

proof: Immediate from the continuity of the δ parameterization of S. □

Each value of δ corresponds to a choice of mean vector, and gives rise to a linear mapping $S(\delta)$. That is, for each choice of δ, a linear map arises. The proposition tells us that so long as our mixture's means are kept within distance c of the origin, all associated linear mappings are invertible.

The utility of $S(\delta)$ is that it allows us to simultaneously express the objectives of obtaining by emphasis a specified mean vector, and covariance matrix. If Γ is a nonnegative emphasis vector, δ_t is the targeted mean vector, and Σ_t the desired covariance matrix, then we must solve the system $S(\delta_t)\Gamma = \Sigma_t : 0$. Here Σ_t is regarded as a vector which is concatenated (denoted ":") with the d-dimensional zero vector.

The Γ above is nonnegative, but not necessarily stochastic. This may seem to be inconsistent with our definition of the generation of a mean vector by emphasis, because of the normalization it includes. But it is not since $\sum \gamma_i(s_i - \delta) = 0$ is equivalent to $(1/\sum \gamma_i)\sum \gamma_i = \delta$, which is exactly our definition.

Theorem 2 *Given any d-dimensional normal mixture M with k components, and $s_1, \ldots, s_n \in \mathbb{R}^d$ such that $n \geq d(d+3)/2$ with some subset of size $d(d+3)/2$ satisfying condition 1, then there exists a $k \times n$ table of nonnegative values $\{\gamma_{i,j}\}$, and nonnegative values $m_1, \ldots, m_{k-1}$ with $\sum m_k \leq 1$, such that the normal mixture M' generated as described below, is class-equivalent to M.*

1. *The mixing parameters of M' are $m_1, \ldots, m_{k-1}, 1 - \sum_{j=1}^{k-1} m_j$.*

2. Each mean μ'_i within M' is given by:

$$\mu_i = \frac{1}{\sum_{j=1}^{n} \gamma_{i,j}} \sum_{j=1}^{n} \gamma_{i,j} s_j$$

3. Each covariance Σ'_i within M' is given by:

$$\Sigma_i = \sum_{j=1}^{n} \gamma_{i,j} (s_j - \mu_i)(s_j - \mu_i)^t$$

Also, the number of parameters may be reduced to $(k-1)d(d+3)/2 + d$, matching the complexity of the canonical parameterization, by replacing the top γ-table row, by a single new parameter α.

proof: Choose some subset of the $\{s_i\}$ of size exactly $d(d+3)/2$ that satisfies condition 1. We disregard the other elements entirely, and therefore simply assume that $n = d(d+3)/2$. As argued earlier, we may also assume without loss of generality that the $\{s_i\}$ have mean zero. Then the distinguished mixture component from theorem 1, with parameters (μ'_j, Σ'_j), is chosen as follows. Its mean μ'_j is the origin, and its covariance Σ'_j is proportional to the average of the element self-outer-products, i.e.:

$$\Sigma'_j = \alpha \frac{1}{n} \sum_{i=1}^{n} s_i s_i^t$$

This may be thought of as choosing the first row of the γ table to be $1/n$, and introducing a new scale parameter α. The table's top row elements are no longer parameters, so the total becomes $(k-1)d(d+3)/2$ table parameters, plus $d-1$ mixing parameters, plus α – matching the count in the statement of the theorem. As described in the proof of theorem 1, the choice of μ'_j is arbitrary, but Σ'_j may need to be scaled down to preserve the positive nature of all matrices. The number of resulting parameters in a direct parameterization (i.e. consisting of means, covariances, and mixing parameters), then matches our count.

As $\alpha \to 0$ we know that the $\mu'_i \to \mu'_j = 0$. Our first step is to choose α sufficiently small, so that the largest $\|\mu'_i\|$ is smaller than the c of proposition 4. Next, α is possibly reduced further until each covariance matrix Σ'_i arising from Eq. 4 is positive.

Each μ' corresponds to a δ value in our definition of $S(\delta)$. The reference mean, located at the origin, corresponds to $\delta = 0$, and by our construction, arises from weighting the s_ℓ by non-negative γ values – in this case uniform ones. Associated with each μ'_i there is also a Σ'_i. Recall from our earlier discussion, that our focus is on the equation $S(\mu_i)\Gamma = \Sigma_i : 0$. Since $S(\delta)$ is non-singular for each of the μ'_i, we know that some vector of γ values satisfies this equation. We have only to show that solutions consisting only of nonnegative values can be found.

We will say than a point is *representable* under some $S(\delta)$, if its inverse image under this mapping consists only of non-negative values. There exists a representable open neighborhood about $\alpha(\Sigma'_j : 0)$ under $S(0)$, since the inverse image of this point[1] is the constant vector α/n, and consists of strictly positive values.

[1] The inverse image is a unique point because $S(\delta)$ is invertible.

Within our bounds on δ, it uniform-continuously[2] parameterizes $S(\delta)$. Hence there exist values c' and ϵ, such that for all δ with $\|\delta\| \leq c'$, the open ball about $\alpha({\Sigma'}_j : 0)$, with radius ϵ, is representable under $S(\delta)$. To avoid introducing another symbol, let ϵ now denote the maximum such radius.[3]

Next notice that the space of representable points is convex and includes the origin. So representability of $\alpha({\Sigma'}_j : 0)$ implies representability for all α values. Now if necessary, reduce α further so that all the ${\mu'}$ are within c' of the origin.

Now let P denote the subspace of nonnegative Γ vectors. Let:

$$T \triangleq \bigcap_{\|\delta\| \leq c'} S(\delta)P$$

Subspace T is the portion of the range, representable under *any* $S(\delta)$ such that $\delta \leq c'$.

About $\alpha({\Sigma'}_j : 0)$ we have seen that there is a representable ball of radius ϵ which touches the boundary of T. We now denote this radius by $\epsilon(\alpha)$ since we are about to consider its behavior as $\alpha \to 0$. By our earlier comments, T includes $\alpha({\Sigma'}_j : 0)$, and all proportional vectors, in particular as $\alpha \to 0$.

The geometric observation key to our proof is that $\epsilon(\alpha)$ can shrink only as fast as α itself, since this value represents the shortest distance to some point set, in this case T^c. We imagine T to be a tunnel, leading to the origin while constricting, with the boundary of T^c forming the tunnel's wall.

Focus now on some ${\mu'}_i$. While there is a corresponding representable ball about $\alpha({\Sigma'}_j : 0)$ of radius $\epsilon(\alpha)$, there is no guarantee that ${\Sigma'}_i$ is within it. As $\alpha \to 0$, we have seen that the radius of this ball shrinks linearly with α. But the distance of ${\Sigma'}_i$ from ${\Sigma'}_j$ by proposition 2 shrinks quadratically with α, whence eventually, i.e. for small enough α, ${\Sigma'}_i$ becomes representable. Then α is reduced as necessary for for each component of the mixture, so that afterwards, every $\Sigma' : 0$ is representable. □

Other emphasis theorems are possible. In particular it is much easier to establish a similar result in which the weights may be either positive or negative because then the only requirement is that $S(\delta)$ be nonsingular.

4 The Limitations of Strict EM-style Emphasis

In the previous section we saw that using a particular form of emphasis reparameterization, the *a posteriori* behavior of any mixture could be matched. We say that such a sample set and emphasis method is *universal*. This section is by contrast essentially negative, and begins with an example demonstrating that even in one dimension, strict EM-style emphasis is not universal.

Our example is in $\mathbb{R}^1$, and includes two observation points located at 0 and 1. The leading normalization factors in Eq. 2,1 amount to enforcing convex combination, so there is only one degree of freedom given our two element observation set. If γ denotes the weight of the first point, then the second has weight $1-\gamma$. The induced mean is just γ and the induced variance $\gamma(1-\gamma)$. Our objective is to generate a two element mixture which is class equivalent to an arbitrary one, so two emphasis parameters γ_1 and γ_2 are needed. The two constant differences to be matched (from Eq. 7) are

[2] and in fact nearly linearly for small δ.

[3] A maximum exists because entries in S corresponding to covariance diagonal, are nonnegative, preventing many negative values from being representable.

denoted Δ_Σ and Δ_μ. That these are independent and unconstrained characteristics an equivalence class follows from the observation that we may, without loss of generality, set one of the target means to zero. The constraints then become:

$$\begin{aligned} \frac{1}{\gamma_1(1-\gamma_1)} - \frac{1}{\gamma_2(1-\gamma_2)} &= \Delta_\Sigma \\ \frac{1}{1-\gamma_1} - \frac{1}{1-\gamma_2} &= \Delta_\mu \end{aligned}$$

Choose $\Delta_\mu > 0$. From our second constraint we have $\gamma_1 = 1 - 1/[\Delta_m u + 1/(1-\gamma_2)]$ from which it follows that $\gamma_1 = \gamma_2$ only when they both are 1. But in this case $\Delta_\mu = 0$, contradicting our choice. So $\gamma_1 \neq \gamma_2$. Now when $\gamma_2 = 0$, $\gamma_1 > 0$, so here $\gamma_2 < \gamma_1$. Because they are never equal, and their relationship is continuous, this inequality must hold for any pair satisfying the constraints. But it is easily verified that $\Delta_\Sigma - \Delta_\mu = 1/\gamma_1 - 1/\gamma_2$ whence this quantity is negative. So no pair γ_1, γ_2 exists which satisfies the constraints given say $\Delta_\mu = 1$ and $\Delta_\Sigma = 2$.

We remark that if instead, one is interested in matching the *a posteriori* class behavior of a mixture at only the given sample points, then the example's argument does not apply. Indeed it may be verified that this less stringent requirement can be satisfied. This is not entirely surprising since given exactly $d(d+3)/2$ sample points, one has three free parameters, and must match only two. This is interesting, and we conjecture that it is true more generally for $d > 1$ and $k > 2$ – so long as $n = d(d+3)/2$. In any event, as a result of our arguments at the end of this section, it cannot be true for arbitrarily large n. Since we are interested in reparameterizing problems with arbitrarily many samples, we will not consider the matter further in this paper.

The example above applies to dimension 1 only, and a particular set of sample points. Moreover, its analytical approach does not generalize easily. We now show that in the general case strict EM-style emphasis is not universal.

Without loss of generality we will assume $\mu'_j = 0$ – since otherwise, the problem, and any solution thereto, can be translated appropriately. We may also assume the sample points are distinct. Eq. 7 becomes:

$$\Delta_{\mu_i} = \Sigma'^{-1}_i \mu'_i \tag{12}$$

where $\Delta_{\mu_i} \triangleq \Sigma_i^{-1}\mu_i - \Sigma_j^{-1}\mu_j$. Our focus is on the simple rearrangement of Eq. 12:

$$\Sigma'_i \Delta_{\mu_i} = \mu'_i \tag{13}$$

Independent of the $\Sigma'_1, \ldots, \Sigma'_k$ we are free to set $\mu'_1, \ldots, \mu'_k$ so that the Δ_{μ_i} have arbitrary values. In particular, we will set them so that $\Delta_{\mu_i} = (CC \ldots C)^t$ where $C > 1$ is a constant.

Now focus on target mixtures such that $\Delta_{\Sigma_i} \triangleq \Sigma_i^{-1} - \Sigma_j^{-1}$ has value αI where $\alpha \to \infty$. Then $\Sigma'^{-1}_i = \alpha I + \Sigma'^{-1}_j$. So independent of the choice of particular $\Sigma_1, \ldots, \Sigma_k$, and Σ'_j, each Σ'^{-1}_i, where $i \neq j$, will approach αI. There inverses therefore approach the diagonal matrix $1/\alpha I$.

As $\alpha \to \infty$, we adjust means as necessary so as to maintain $\Delta_{\mu_i} = (CC \ldots C)^t$. Then from Eq. 13 it is apparent that μ'_i is very nearly a scaled up copy of the diagonal of Σ'_i. Since the off diagonal elements approach zero, it must eventually be the case that $\|\Sigma'_i\| < \|\mu'_i\|$. Also, since the sample set is finite, it must eventually be that both are smaller than say 1/100th of the smallest

distance between sample points. When this happens, the distance from ${\mu'}_i$ to the nearest sample will be at least $\|{\mu'}_i\|$.

Now the covariance ${\Sigma'}_i$ is a convex combination of matrices of the form $(s_\ell - {\mu'}_i)(s_\ell - {\mu'}_i)^t$ – and the diagonals are nonnegative. The norm of ${\Sigma'}_i$ is at least equal to the norm of its diagonal, but the diagonal is just the distance from the sample point to ${\mu'}_i$. From this it follows that $\|{\Sigma'}_i\| \geq \|{\mu'}_i\|$ which presents a contradiction, whence strict EM-style emphasis is not universal. Our arguments above establish:

Theorem 3 *In the setting of theorem 2, altered after the fashion of EM so that each covariance* Σ'_i *is given by:*

$$\Sigma_i = \frac{1}{\sum_{j=1}^{n} \gamma_{i,j}} \sum_{j=1}^{n} \gamma_{i,j}(s_j - \mu_i)(s_j - \mu_i)^t$$

there exist target mixtures which may not be generated by emphasis.

We have seen that no number of sample points make strict EM-style emphasis universal. But given enough points, we can certainly identify a particular class equivalence-class, since all functions involved are analytic with a finite number of parameters. So given enough sample points, a reparameterization *must* be universal in order to match a target distribution at the sample points only. Now we are interested in reparameterizations which apply given an unlimited number of points. Therefore, unlike the modified emphasis reparameterization introduced in the previous section, strict EM-style reparameterization is simply not expressive enough to safely reparameterize *a posteriori* optimization problems.

5 Concluding Remarks

The *a posteriori* degeneracy we clarify in section 2 for normal mixtures must be added to the list of interesting characteristics of Gaussians. We have not considered analogues of theorem 1 for other densities but remark that proposition 1 applies more generally. Beta densities, for example, have the property that their ratio is again of the same functional form so that a similar degeneracy arises. It seems likely that results similar in flavor to theorem 1 might be established for this and other related densities.

Another interesting area for future work relates to densities for which there is essentially no degeneracy, i.e. for which the mixture may be identified given its class behavior. A simple example is provided by a mixture of two uniform univariate densities $U_{\theta_1}, U_{\theta_2}$ with support $[\theta_1, 1+\theta_1]$ and $[\theta_2, 1+\theta_2]$ respectively. Here, a form of degeneracy arises only when $\theta_1 = \theta_2$. The mixture is still identifiable but the mixing coefficients are not. This illustrates the manner in which mixtures of densities of finite support might be used to construct degeneracy-free mixtures. More complex and interesting possibilities exist.

The reparameterization of section 3 is of mathematical interest, and lends some credibility to approaches which attempt to maximize *a posteriori* probability by adjusting weights on the available samples – perhaps according to some error measure. An examination of reparameterization as a primary optimization technique, represents an interesting area for future work. We must however caution that while our proofs are constructive, we have not considered the matter of numerical

sensitivity. Indeed, if one attempts to emulate a mixture using means far displaced from their natural locations, the mixing parameters become quite small. In these cases floating point underflow is a virtual certainty.

One should not interpret section 4 to say that strict EM-style emphasis will not work in practice for many problems. It merely exposes its theoretical limitations.

Acknowledgments

We thank Leonid Gurvits for several helpful discussions – and in particular for pointing out the simple factorization in proof of proposition 2, and simplifying our proof of proposition 3. We also thank Joe Kilian for several helpful discussions which contributed to our understanding of degeneracy and class equivalence.

References

[1] L. E. Baum and J. E. Eagon, *An inequality with application to statistical estimation for probabalistic functions of a Markov process and to models for ecology*, Bull. AMS, 73 (1967), pp. 360–363.

[2] P. F. Brown, *Acoustic-phonetic modeling problem in automatic speech recognition*, PhD thesis, Carnegie-Mellon University, 1987.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum-likelihood from incomplete data via the EM algorithm*, J. Royal Statistical Society Ser. B (methodological), 39 (1977), pp. 1–38.

[4] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, Inc., 1973.

[5] X. D. Huang, Y. Ariki, and M. A. Jack, *Hidden Markov Models for Speech Recognition*, Edinburgh University Press, 1990.

[6] L. R. Rabiner, B. H. Juang, S. E. Levinson, and M. M. Sondhi, *Recognition of isolated digits using hidden markov models with continuous mixture densities*, AT&T Technical Journal, (1985).

[7] R. A. Redner and H. F. Walker, *Mixture densities, maximum likelihood, and the EM algorithm*, SIAM Review, 26 (1984), pp. 195–239.

[8] P. N. Yianilos, *Topics in Computational Hidden State Modeling*, PhD thesis, Princeton University, Computer Science Department, Princeton, NJ, June 1997.