Estimating a mixture of two product distributions

Yoav Freund[*] Yishay Mansour[†]

March 1, 1999

Abstract

We describe an efficient algorithm for estimating a mixture of two product distributions over binary vectors.

1 Introduction

There are two major lines in research in Machine learning, supervised learning and unsupervised learning. Supervised learning has been the more appealing to theoretical analysis, since the goal that it sets is very clear. We have some unknown target function that labels the examples, and we would like to build a good approximation for it. In unsupervised learning the goal is far less clear. We are given examples, but there is no labeling. The vague general goal is to find meaningful structure in the data.

One way to find a structure in the data is to approximate the distribution that generates it. This is the well studied problem of distribution estimation or density estimation. This problem has a well-defined and agreed upon goal and measure of performance. Given a sample generated from some target distribution, the goal is to output a distribution which is close to the target distribution. The most common measures of performance are the likelihood of the data given the model and the Kullback-Leibler (KL) divergence between the model and the true distribution.

Much of the research in distribution estimation is focused on mixture of simple distributions, e.g. mixture of Gaussians. The motivation is that each component in the mixture represents a "cluster" of similar instances. Given unlabeled data that is generated from this mixture we have two related tasks, one is to estimate the parameters of the mixture distribution, and the other, usually referred to as "clustering", is to identify each example with the cluster (or clusters) that is mostly to have generated it. Many algorithms for estimating mixture distributions, including our algorithm, work by alternating between the two tasks. Improving the model estimate is used to improve the accuracy of the association of each example with a cluster and improving the accuracy of these associations is used to improve the estimate of the parameters.

In this work our simple distributions are product distribution over binary n-bit vectors. The target distribution is a mixture of two such product distributions. Even this simple case turns our to be quite challenging, especially when we want our algorithm to be efficient for high dimensional data and the distance between the mixture components is small. We design an algorithm that given a large enough sample from the target distribution finds a mixture of two product distributions which are close to the target distribution in terms of the KL-divergence. Our algorithm is efficient both in its sample complexity and its computational complexity.

[*]AT&T Research Labs. `e-mail: yoav@research.att.com`

[†]AT&T Research Labs and Tel-Aviv University. `e-mail: mansour@research.att.com`

Our algorithm uses two novel ideas, which might also be useful in other cases. The first idea is to have a held-out set of *attributes*. We use one set of attributes to split in the sample into two parts, with the goal that the parts be very different from each other in terms of the ratio of instances originating from the two mixture components. If the components are sufficiently far from each other, than each part will contain almost exclusively instances from one component and the problem becomes trivial. However, even if the split is far from perfect we show how to estimate the distribution accurately. To do that we use the held-out attributes. The advantage of using the held-out attributes is that the marginal distribution of these attributes is a mixture of two product distributions for each part of the split sample. This is not the case with the original attributes that were used to split the sample. The second idea is that of a dimension reduction. We are given data from an n dimensional space and show how to reduce the n dimensional search to a one dimensional search over a line, which greatly reduces the computational complexity.

We do not attempt to give here a comprehensive review on the work on distribution estimation, but rather mention a few results whose main concern (as in this work) is the computational complexity. The work of Kearns et al. [KMR$^+$94] defines the computational setting of learning distributions, and also gives a few example of distributions which can be learn efficiently. The work of Freund and Ron [FR95] studies a related problem of identifying biased coins from sequences.

There is a vast literature in statistics on distribution estimation in general, and mixture models in particular. There is a necessary and sufficient condition for a mixture to be identifiable [YS68]. Identifiability implies that each mixture distribution has a unique representation. Note that identifiability is stronger than learnability, since for learnability it is enough to output some good model, and it might be that the true model is not unique.

We are interested here in the possibility of learning mixtures with similar components in a high dimensional space and with relatively small samples. One source of difficulty in this case is that the difference between the components might be undetectable when considering any small subset of the components. Thus any method that relies on low order statistics, such as the moment method for learning mixture distributions (described, for example, in [TSM85]) will necessarily require very large amounts of data. Another popular method for learning mixtures is EM, and our algorithm can be seen as a version of EM. However, while the analysis of EM usually guarantees only asymptotic convergence, we are interested here in designing an algorithm that can be proven to converge within a single iteration.

The paper is organized as follows. In Section 2 we define our notation and the learning model. In Section 3 we describe the algorithm. We prove its correctness in Section 4.

2 Model and Notation

2.1 Notation

Given a vector $\mathbf{x}$ we denote its ith coordinate by x_i. For a set of attributes $S \subset \{1, \cdots, n\}$, and a vector $\mathbf{q}$ of length n, we denote by $\mathbf{q}_S$ a vector that includes only the coordinates in S. The inner product of two n-dimensional vectors, $\mathbf{a}$ and $\mathbf{b}$, is $<\mathbf{a}, \mathbf{b}> = \sum_{i=1}^{n} a_i b_i$. For a set S of attributes then $<\mathbf{a}, \mathbf{b}>_S \doteq <\mathbf{a}_S, \mathbf{b}_S> = \sum_{i \in S} a_i b_i$. A *random set* $S \subset \{1, \ldots, n\}$ has $\Pr[i \in S] = 1/2$ independently for each i. The operation $\texttt{round}(\rho, \mathbf{q})$ receives a vector $\mathbf{q} \in [+1, -1]^n$ and returns a vector $\mathbf{q}'$ such that, if $q_i < \rho$ then $q'_i = \rho$, if $q_i > 1 - \rho$ then $q'_i = 1 - \rho$, otherwise $q'_i = q_i$.

2.2 Product distributions

A product distribution over binary vectors in $\{-1,+1\}^n$ is characterized by its expected value $\mathbf{q} \in [-1,1]^n$. A random vector $\mathbf{x}$ from the distribution $\mathbf{q}$ is generated by independently selecting each coordinate x_i to be 1 with probability $(1+q_i)/2$ or $x_i = -1$ with probability $(1-q_i)/2$. It is easy to verify that the expected value of vectors selected in this way is indeed $\mathbf{q}$, a fact which we express as

$$E_{\mathbf{x}\sim\mathbf{q}}[\mathbf{x}] = \mathbf{q}$$

A mixture of two product distributions is characterized by a triplet $(\mathbf{q}^{(1)}, \mathbf{q}^{(2)}, \gamma)$ where $\mathbf{q}^{(1)}, \mathbf{q}^{(2)} \in [-1,1]^n$ are the component *centers* and $\gamma \in [0, 1/2]$ is the mixture coefficient. To generate an example $\mathbf{x}$ from the mixture distribution, we select the index b to be 1 or 2 with probabilities γ or $1-\gamma$ respectively and then generate $\mathbf{x}$ according to the product distribution $\mathbf{q}^{(b)}$.

2.3 Learning Algorithm

A learning algorithm A samples examples from a mixture distribution $(\mathbf{q}^{(1)}, \mathbf{q}^{(2)}, \gamma)$. It outputs a mixture model $(\hat{\mathbf{q}}^{(1)}, \hat{\mathbf{q}}^{(2)}, g)$, such that the KL-divergence between the models is small, where The KL-divergence between two distributions P and Q is $KL(Q\|P) = E_Q[\log(Q/P)]$.

Formally, an algorithm A is a learning algorithm for mixture of two product distributions if, after sampling m examples, it outputs a model $(\hat{\mathbf{q}}^{(1)}, \hat{\mathbf{q}}^{(2)}, g)$, such that with probability $1-\delta$,

$$KL[(\mathbf{q}^{(1)}, \mathbf{q}^{(2)}, \gamma) \| (\hat{\mathbf{q}}^{(1)}, \hat{\mathbf{q}}^{(2)}, g)] \leq \epsilon.$$

Algorithm A is an efficient algorithm if it runs in time polynomial in $1/\epsilon$, $\log(1/\delta)$ and n.

3 The Algorithm

3.1 Overview

The basic idea of our algorithm is as follows. We split the coordinates into two sets S and $\bar{S}$. We use the coordinates in S to split the examples in the training set into two parts. We then consider the difference in the average of the two parts on the coordinates in $\bar{S}$ to reduce the problem into a one dimensional problem. We then perform a one dimensional search to find the best distribution model.

The splitting rule is a predicate $P(\mathbf{x}_S)$ that partitions the examples into two sets. We try to choose $P(\mathbf{x}_S)$ so that the distributions conditioned on this predicate contain very different ratios of the two components. In other words that the probability that $\mathbf{x}$ was generated from $\mathbf{q}^{(1)}$, given that $P(\mathbf{x}_S)$ is true, is either significantly smaller or significantly larger than γ.

Clearly, the distribution of the examples in each of the two parts is no longer a mixture of two product distributions. However, the marginal distribution of each part restricted to the coordinates in $\bar{S}$ *is* a mixture of two product distributions. Using this fact we can calculate unbiased approximations of the expected value of each of the two sample parts. By passing a line through these estimates we find an approximation of the line connecting the true centers with respect to the coordinates in $\bar{S}$.

To complete the algorithm we search this line for the optimal pair of centers to maximize the likelihood of the data. The main feature of this algorithm is that it reduced an n-dimensional search into a one dimensional search.

3.2 Description of the algorithm

The following is a high level sketch of the algorithm. (The detailed algorithm is described in Figure 1.) The main procedure (`MAIN`) receives as input a sample T of size m and outputs a model for the distribution. It works in the following way. First it computes the average of the entire sample and computes the likelihood it generated the sample T. Then it chooses a random set S (which is used to partition the coordinates) and calls the procedure `Estimate_SET` once with inputs S and T and once with inputs $\bar{S}$ and T. We would like a split S to be such, that the difference between the centers would be significant both in coordinates in S and in $\bar{S}$. (In Section4.1 we show that with probability 1/2 we have a good split.) The procedure `Estimate_SET` outputs a model for each call, and we combine the two models, for S and $\bar{S}$ to get a global model, and test its likelihood. The output of `MAIN` is the model with the highest likelihood.

The procedure `Estimate_SET` receives the sets S and T as input, and outputs a model for the distribution restricted to $\bar{S}$. Its operation consists of the following steps.

1. Choose a random vector $\mathbf{x}^*$ from the sample T of size m. (Each $\mathbf{x} \in T$ has probability $1/m$.) We like $\mathbf{x}^*$ to have the property, that when restricted to S, examples from different centers have different expected value of inner product with $\mathbf{x}^*$. (In Section 4.1 we show that the probability that $\mathbf{x}^*$ is good is proportional to γ times the distance between the centers.)

2. Given $\mathbf{x}^*$ we compute a threshold θ, and split T to $T^{(>)}$, which includes all the examples for which the inner product is more than θ and $T^{(\leq)}$ which includes the remaining examples. We set θ to the average value of the inner product of $\mathbf{x}$ with $\mathbf{y} \in T$, i.e., $(1/m)\sum_{\mathbf{y}\in T} < \mathbf{x}^*, \mathbf{y} >_S$. (In Section 4.3 we show that the fraction of examples in $T^{(\leq)}$ from any of the two centers is different than γ and $1-\gamma$.)

3. Compute the average of examples in T on $\bar{S}$ (denoted by $\hat{\mathbf{q}}^{(a)}_{\bar{S}}$) and the average of $T^{(\leq)}$ on $\bar{S}$ (denoted by $\hat{\mathbf{q}}^{(s)}_{\bar{S}}$). (In Section 4.4 we show that there is a non-negligible distance between the two vectors.)

4. We connect $\hat{\mathbf{q}}^{(a)}_{\bar{S}}$ and $\hat{\mathbf{q}}^{(s)}_{\bar{S}}$ to a line, and define on it points with a spacing of $\lambda\hat{\Delta}$, where $\hat{\Delta} = \hat{\mathbf{q}}^{(a)}_{\bar{S}} - \hat{\mathbf{q}}^{(s)}_{\bar{S}}$ and λ is a parameter. We also discretize the value of γ with a spacing parameter λ_γ. Given any pair of points $\mathbf{q}^{(1)}_{\bar{S}}$ and $\mathbf{q}^{(2)}_{\bar{S}}$ on the line and any value g for the mixture coefficient, we compute the likelihood that the model $(\mathbf{q}^{(1)}, \mathbf{q}^{(2)}, g)$ generated T. The model with the maximum likelihood is our estimator of the target mixture distribution on $\bar{S}$. The number of models we consider is at least $(\beta/\lambda)^2(1/\lambda_\gamma)$, where $\beta = 2/\max_j\{\hat{\Delta}_j\}$. (In Section 4.5 we show that there is a model which we consider and is near the target model. In Section 4.6 we show that it has a small KL divergence to the true model.)

4 Proof of learnability

In this section we show that our algorithms learns efficiently a mixture of two product distributions. As a part of the proofs we show how to determine the parameters of the algorithm.

4.1 Splitting the attributes

Given two vectors $\mathbf{q}^{(1)}, \mathbf{q}^{(2)} \in [-1,1]^n$, we assume that the distance between them is non-negligible, i.e. $||\mathbf{q}^{(1)} - \mathbf{q}^{(2)}||_2^2$ is large. we like to show that with non-negligible probability, if we choose a random set S, then both $\sum_{i\in S}(\mathbf{q}^{(1)}_i - \mathbf{q}^{(2)}_i)^2$ and $\sum_{i\notin S}(\mathbf{q}^{(1)}_i - \mathbf{q}^{(2)}_i)^2$ are not negligible.

Lemma 4.1 *Let $a_1 \cdots a_n$ be a set of non-negative real numbers whose maximal value is a_{max} and whose sum is L. Let S be a random subset of $\{1, \ldots, n\}$ which includes each element with probability $1/2$. Then with probability at least $1/2$, we have both $\sum_{i \in S} a_i \geq (L - a_{max})/4$ and $\sum_{i \notin S} a_i \geq (L - a_{max})/4$.*

Proof: There is a set $T \subset \{1, \ldots, n\}$ such that $|(\sum_{i \in T} a_i) - (\sum_{i \notin T} a_i)| \leq a_{\max}$. Let $\sum_{i \in T} a_i = L_1$. For any subset $S_1 \subset T$ either $\sum_{i \in S_1} a_i \geq L_1/2$ or $\sum_{i \in T - S_1} a_i \geq L_1/2$. Similarly, let $\sum_{i \in \bar{T}} a_i = L_2$. For any subset $S_2 \subset \bar{T}$ either $\sum_{i \in S_2} a_i \geq L_2/2$ or $\sum_{i \in \bar{T} - S_2} a_i \geq L_2/2$.

Choosing a random set of coordinates S by picking each coordinate independently at random is equivalent to choosing S_1 at random from T, S_2 from $\bar{T}$ and then choosing with probability $1/4$ one of the four combinations: $S_1 \cup S_2$, $S_1 \cup \bar{S}_2$, $\bar{S}_1 \cup S_2$, and $\bar{S}_1 \cup \bar{S}_2$. Therefore, the probability that the larger sum in S_1 is matched with the smaller sum in S_2 is $1/2$. This guarantees that the elements in S have a sum of at least $L_1/2$ and the elements in $\bar{S}$ have a sum of at least $L_2/2$. Since $L_1 + L_2 = L$ and $|L_1 - L_2| \leq a_{\max}$, the lemma follows. $\square$

To apply this lemma to our problem we define the distance between two centers, d, to be

$$d \doteq \frac{\|\mathbf{q}^{(1)} - \mathbf{q}^{(2)}\|_2^2 - \max\{(q_i^{(1)} - q_i^{(2)})^2\}}{4},$$

Setting $a_i = (q_i^{(1)} - q_i^{(2)})^2$ we get the following corollary directly from Lemma 4.1.

Corollary 4.2 *Let $\mathbf{q}^{(1)}, \mathbf{q}^{(2)}$ be any two vectors in $[+1, -1]^n$. With probability at least $1/2$, a random set of coordinates S has the property that both $\|\mathbf{q}_S^{(1)} - \mathbf{q}_S^{(2)}\|_2^2 \geq d$ and $\|\mathbf{q}_{\bar{S}}^{(1)} - \mathbf{q}_{\bar{S}}^{(2)}\|_2^2 \geq d$.*

4.2 Finding a Split

Let us review how we find a split: (1) we choose a random vector $\mathbf{x}^*$ from our sample, and a threshold θ. (2) For each $\mathbf{y} \in T$ we compute $< \mathbf{x}^*, \mathbf{y} >_S$. (3) If $< \mathbf{x}^*, \mathbf{y} >_S$ is larger than θ then $\mathbf{y}$ belongs to $T^{(>)}$ and otherwise $\mathbf{y}$ belongs to $T^{(\leq)}$. Our main aim will be to show that the fraction of elements from $\mathbf{q}^{(1)}$ in $T^{\leq}$ is significantly different from γ.

We start by showing that there exists some θ for which the behavior of the two centers is different. Assume that $\mathbf{x}$ is distributed according to $\mathbf{q}^{(1)}$ and $\mathbf{y}$ is distributed according to $\mathbf{q}^{(2)}$. Then,

$$E_{\mathbf{x} \sim \mathbf{q}^{(1)}, \mathbf{y} \sim \mathbf{q}^{(2)}}[< \mathbf{x}, \mathbf{y} >_S] =< \mathbf{q}^{(1)}, \mathbf{q}^{(2)} >_S$$

In the case that $\mathbf{y}$ is chosen from $\mathbf{q}^{(1)}$ we have

$$E_{\mathbf{x} \sim \mathbf{q}^{(1)}, \mathbf{y} \sim \mathbf{q}^{(1)}}[< \mathbf{x}, \mathbf{y} >_S] =< \mathbf{q}^{(1)}, \mathbf{q}^{(1)} >_S$$

We hope that there is a significant difference between the two expectations, so that it will influence the split. However, the difference in the expectation is $< \mathbf{q}^{(1)}, (\mathbf{q}^{(1)} - \mathbf{q}^{(2)}) >_S$, which might be zero. (Recall that we have no assumption on the norm of $\mathbf{q}^{(1)}$ and $\mathbf{q}^{(2)}$.)

We show that the difference can not be zero both in the case that $\mathbf{x}$ is chosen from $\mathbf{q}^{(1)}$ and $\mathbf{q}^{(2)}$. Let,

$$\mathtt{diff}(\mathbf{x}) \doteq E_{\mathbf{y}^{(1)} \sim \mathbf{q}^{(1)}}[< \mathbf{x}, \mathbf{y}^{(1)} >_S] - E_{\mathbf{y}^{(2)} \sim \mathbf{q}^{(2)}}[< \mathbf{x}, \mathbf{y}^{(2)} >_S]$$

We have that,

$$\begin{aligned} E_{\mathbf{x} \sim \mathbf{q}^{(1)}}[\mathtt{diff}(\mathbf{x})] + E_{\mathbf{x} \sim \mathbf{q}^{(2)}}[\mathtt{diff}(\mathbf{x})] &= < \mathbf{q}^{(1)}, \mathbf{q}^{(1)} - \mathbf{q}^{(2)} >_S + < \mathbf{q}^{(2)}, \mathbf{q}^{(2)} - \mathbf{q}^{(1)} >_S \\ &= \|\mathbf{q}_S^{(1)} - \mathbf{q}_S^{(2)}\|_2^2 \end{aligned}$$

This implies immediately the following lemma.

Lemma 4.3 *Either* $E_{\mathbf{x}\sim\mathbf{q}^{(1)}}[\mathtt{diff}(\mathbf{x})] \geq d/2$ *or* $E_{\mathbf{x}\sim\mathbf{q}^{(2)}}[\mathtt{diff}(\mathbf{x})] \geq d/2$.

We say that a vector $\mathbf{x} \in \{-1,+1\}^n$ is *good* if $\mathtt{diff}(\mathbf{x}) \geq d/4$. The next lemma gives a lower bound on the probability of choosing a good vector.

Lemma 4.4 *Let* $\mathbf{x}$ *be generated by* $(\mathbf{q}^{(1)}, \mathbf{q}^{(2)}, \gamma)$. *The probability that* $\mathbf{x}$ *is good is at least* $p \doteq \gamma \min\{d/32, 1/2\}$.

Proof: Without loss of generality assume that $E_{\mathbf{x}\sim\mathbf{q}^{(1)}}[\mathtt{diff}(\mathbf{x})] = \eta \geq d/2$. The probability that $\mathbf{x}$ is generated from $\mathbf{q}^{(1)}$ is at least γ. Assume that $\mathbf{x}$ is generated by $\mathbf{q}^{(1)}$.

By definition, $\mathtt{diff}(\mathbf{x}) = \sum_{j=1}^{n} X_j(\mathbf{q}_j^{(1)} - \mathbf{q}_j^{(2)})$, where X_j is $+1$ with probability $(\mathbf{q}_j^{(1)}+1)/2$ and otherwise -1. We can now apply a Chernoff bound on $\mathtt{diff}(\mathbf{x})$,

$$\Pr[|\mathtt{diff}(\mathbf{x}) - \eta| > \eta/2] \leq e^{-\eta/8} \leq e^{-d/16}$$

For $d \geq 16$ this probability is less than $1/2$. For $d < 16$ this probability is less than $1 - d/32$. This implies that the probability that $\mathtt{diff}(\mathbf{x}) \geq d/2$ is at least $\min\{1/2, d/32\}$. □

We expect that, with high probability, a large fraction of the examples in T would be good.

Corollary 4.5 *For* $m > (8/p)\log(1/\delta)$, *with probability* $1-\delta$, *at least* $pm/2$ *of the points in* T *are good.*

The above corollary guarantees that with probability of at least $p/2$ we chose a good $\mathbf{x}^*$.

4.3 Choosing the threshold $\hat{\theta}$

The next step is to choose the threshold parameter θ. Given $\mathbf{x}$ let $\theta(\mathbf{x}) = E[< \mathbf{x}, \mathbf{y} >_S]$, where $\mathbf{y}$ is distributed according to the mixture distribution $(\mathbf{q}^{(1)}, \mathbf{q}^{(2)}, \gamma)$. By definition,

$$\theta(\mathbf{x}) = \gamma E_{\mathbf{y}^{(1)}\sim\mathbf{q}^{(1)}}[< \mathbf{x}, \mathbf{y}^{(1)} >_S] + (1-\gamma) E_{\mathbf{y}^{(2)}\sim\mathbf{q}^{(2)}}[< \mathbf{x}, \mathbf{y}^{(2)} >_S].$$

Let

$$R(\mathbf{x}, \theta) = \min\{|E_{\mathbf{y}^{(1)}\sim\mathbf{q}^{(1)}}[< \mathbf{x}, \mathbf{y}^{(1)} >_S] - \theta|, |\theta - E_{\mathbf{y}^{(2)}\sim\mathbf{q}^{(2)}}[< \mathbf{x}, \mathbf{y}^{(2)} >_S]|\}.$$

Note that, $R(\mathbf{x}, \theta(\mathbf{x})) \geq \gamma \mathtt{diff}(\mathbf{x})$. This implies that if $\mathbf{x}$ is good then $R(\mathbf{x}, \theta(\mathbf{x})) \geq \gamma d/4$. Let,

$$\hat{\theta}(\mathbf{x}) = \frac{1}{m} \sum_{\mathbf{y} \in T} < \mathbf{x}, \mathbf{y} >_S$$

The following lemma (whose proof is in the Appendix) bounds the error in θ.

Lemma 4.6 *With probability* $1-\delta$ *we have that for each* $\mathbf{x} \in T$, $|\hat{\theta}(\mathbf{x}) - \theta(\mathbf{x})| \leq \gamma d/8$, *(and hence, if* $\mathbf{x}$ *is good then* $|R(\mathbf{x}, \hat{\theta}(\mathbf{x}))| \geq \gamma d/8$*) given that* $m > c(n/(\gamma^2 d^2))\log(1/\delta)$, *for some constant* c.

4.4 Properties of the split

Let $R = R(\mathbf{x}^*, \hat{\theta})$. We assume, without loss of generality, that $E_{\mathbf{y}\sim\mathbf{q}^{(1)}}[< \mathbf{x}^*, \mathbf{y} >_S] > \hat{\theta} + R$ and $E_{\mathbf{y}\sim\mathbf{q}^{(2)}}[< \mathbf{x}^*, \mathbf{y} >_S] < \hat{\theta} - R$. We are interested in computing the probabilities:

$$\mu^{(b)} = \Pr_{\mathbf{y}^b \sim q^{(b)}}[< \mathbf{x}^*, \mathbf{y}^b >_S \ > \hat{\theta}]$$

for $b \in \{1, 2\}$. The probability that a vector $\mathbf{x}$ was generated from the component $\mathbf{q}^{(1)}$, given that $< \mathbf{x}^*, \mathbf{y} >_S \ > \hat{\theta}$ is $\mu = \gamma\mu^{(1)}/(\gamma\mu^{(1)} + (1-\gamma)\mu^{(2)})$. We would like to show that $|\mu - \gamma|$ is large.

The simple case is if $d = \Omega(\sqrt{n \log n})$. This implies that $R = \Omega(\sqrt{n \log n})$. In such a case we have that $\mu^{(1)} \approx 1$, and thus we have a perfect split, and we are done. The more challenging case is when d is small, and thus R is small, and this is the case we address here. (The proof is given in the Appendix.)

Lemma 4.7 *Fix $\mathbf{x}$ and let $Q = (\mathbf{q}^{(1)}, \mathbf{q}^{(2)}, \gamma)$ be a model of mixture of two product distributions, such that $\sum_{j \in S} q_j^{(b)} \in [-1 + \alpha, 1 - \alpha]$, for some constant α, $b \in \{1, 2\}$. Let Q' be the distribution Q restricted to vectors $\mathbf{x}$ such that $< \mathbf{x}, \mathbf{y} >_S > \theta(\mathbf{x})$.*

Then $Q'_{\bar{S}}$ is a mixture distribution $(\mathbf{q}_{\bar{S}}^{(1)}, \mathbf{q}_{\bar{S}}^{(2)}, \mu)$ and $|\mu - \gamma| > c_3 R(\mathbf{x}, \hat{\theta}(\mathbf{x}))\gamma/\sqrt{n}$, for some constant c_3.

Now we like to say that if $\mathbf{x}^*$ is good, a similar claim holds for it. Recall that $T^{(\le)} = \{\mathbf{y} \in T :< \mathbf{x}, \mathbf{y} >_S \le \theta\}$. For the analysis we need to split the set T to a part generated by $\mathbf{q}^{(1)}$ and a part generated by $\mathbf{q}^{(2)}$. Clearly we do not have this split in the algorithm, but we use it only for the analysis. Let $T^{(b)}$ includes the examples in T generated using $\mathbf{q}^{(b)}$, for $b \in \{1, 2\}$. Let,

$$\hat{\mu}^{(b)} = \frac{|T^{(b)} \cap T^{(\le)}|}{|T^{(\le)}|}.$$

Lemma 4.8 *For $m > 16\lambda^{-2} \log(2/\delta)$, with probability $1 - \delta/2$, we have that $|\mu^{(b)} - \hat{\mu}^{(b)}| \le \lambda/4$.*

4.5 Choosing the candidate models

Up to now we have computed some split using part of the attributes, i.e. S. Now we will use the other part, $\bar{S}$, of the attributes to find candidate centers. We would like to show that we can find two centers which are very near the true centers, and thus would be a good approximation. From now on we will consider only the second part of the attributes, $\bar{S}$.

Let $\mathbf{q}_{\bar{S}}^{(a)} = \gamma\mathbf{q}_{\bar{S}}^{(1)} + (1-\gamma)\mathbf{q}_{\bar{S}}^{(2)}$. Then $\mathbf{q}_{\bar{S}}^{(1)} = \mathbf{q}_{\bar{S}}^{(a)} + (1-\gamma)\Delta$ and $\mathbf{q}_{\bar{S}}^{(2)} = \mathbf{q}^{(a)} - \gamma\Delta$, where $\Delta = \mathbf{q}_{\bar{S}}^{(1)} - \mathbf{q}_{\bar{S}}^{(2)}$. Using our sample we compute an estimate of $\mathbf{q}_{\bar{S}}^{(a)}$, which is $\hat{\mathbf{q}}_{\bar{S}}^{(a)} = (1/m)\sum_{\mathbf{y} \in T}(\mathbf{y})_{\bar{S}}$. Let $\epsilon_a = \hat{\mathbf{q}}_{\bar{S}}^{(a)} - \mathbf{q}_{\bar{S}}^{(a)}$.

Now assume that we are given another sample of the data distributed according to a different set of mixture coefficients. This is the effect that we get by using our split, we sample form $\mathbf{q}_{\bar{S}}^{(1)}$ with probability μ (rather than γ). Let $\mathbf{q}_{\bar{S}}^{(s)} = \mu\mathbf{q}_{\bar{S}}^{(1)} + (1-\mu)\mathbf{q}_{\bar{S}}^{(2)}$. Similarly $\hat{\mathbf{q}}_{\bar{S}}^{(s)}$ is the observed average, i.e. $(1/|T^{(\le)}|)\sum_{\mathbf{y} \in T^{(\le)}}(\mathbf{y})_{\bar{S}}$ and ϵ_s is the error. We have,

$$\Delta = \frac{\mathbf{q}_{\bar{S}}^{(a)} - \mathbf{q}_{\bar{S}}^{(s)}}{\gamma - \mu}.$$

We can estimate the direction of Δ using $\hat{\Delta} = \hat{\mathbf{q}}^{(a)} - \hat{\mathbf{q}}^{(s)}$, but since we do not know either γ or μ we can not approximate Δ directly. The following lemmas bound the errors in the estimations.

Lemma 4.9 *For $m > \lambda^{-2}\log(n/\delta)$, with probability $1-\delta$, we have that $\|\mathbf{q}_{\bar{S}}^{(a)} - \hat{\mathbf{q}}_{\bar{S}}^{(a)}\|_\infty \le \lambda$.*

Lemma 4.10 *For $m > 16\lambda^{-2}\log(3n/\delta)$, with probability $1-\delta$, we have that $\|\mathbf{q}_{\bar{S}}^{(s)} - \hat{\mathbf{q}}_{\bar{S}}^{(s)}\|_\infty \le \lambda$.*

Proof: By Lemma 4.8, $|\mu - \hat{\mu}| < \lambda/4$, with probability $1-\delta/2$. The average in $T^{(b)} \cap T^{(\le)}$ differs from $\mathbf{q}^{(b)}$ by at most $\lambda/4$, with probability at most $1-\delta/2$. The total error is at most $3\lambda/4$. □

We use $\hat{\mathbf{q}}_{\bar{S}}^{(a)}$ and $\hat{\mathbf{q}}_{\bar{S}}^{(s)}$ to create a line, and then discretize the line with a small spacing. The following lemma bound the distance of the "true" centers to the discrete points on this line.

Lemma 4.11 *With probability $1-\delta$, for $b \in \{1,2\}$, there exists an integer j_b such that,*

$$\|\mathbf{q}_{\bar{S}}^{(b)} - (\hat{\mathbf{q}}_{\bar{S}}^{(a)} + j_b\lambda(\hat{\mathbf{q}}_{\bar{S}}^{(a)} - \hat{\mathbf{q}}_{\bar{S}}^{(s)}))\|_\infty < \frac{6\lambda}{|\gamma-\mu|}$$

Proof: We do the proof for $b=1$, the case of $b=2$ is similar. Recall that $\mathbf{q}_{\bar{S}}^{(1)} = \mathbf{q}_{\bar{S}}^{(a)} + (1-\gamma)\Delta$. Let $\Delta' = (\hat{\mathbf{q}}_{\bar{S}}^{(a)} - \hat{\mathbf{q}}_{\bar{S}}^{(s)})/(\gamma-\mu)$. Note that, $\Delta' = \Delta + \frac{\epsilon_a - \epsilon_s}{\gamma-\mu}$. We can bound our error by,

$$\|\mathbf{q}_{\bar{S}}^{(a)} - \hat{\mathbf{q}}_{\bar{S}}^{(a)}\|_\infty + |j_1\lambda(\gamma-\mu) - (1-\gamma)|\max\{\|\Delta\|_\infty, \|\Delta'\|_\infty\} + \|\Delta - \Delta'\|_\infty$$

Lemma 4.9 bounds the first term by λ. The discretization error in $|j_1\lambda(\gamma-\mu)-(1-\gamma)|$ is bounded by λ. We can bound $\|\Delta\|_\infty$ by $2/|\gamma-\mu|$. We can bound $\|\Delta'\|_\infty$ by $\|\Delta\|_\infty$ plus $\|\Delta-\Delta'\|_\infty$, which is bounded by $3/|\gamma-\mu|$. Finally, $\|\Delta-\Delta'\|_\infty$, using Lemma 4.10 and 4.9 is bounded by $2\lambda/|\gamma-\mu|$. □

4.6 Maximum Likelihood

In this section we show that the maximum likelihood would give us a good approximation of the target distribution.

Lemma 4.12 *Let Q be any distribution. Let $P = (\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \gamma')$ be a model of mixture of two product distributions, such that $\mathbf{p}_j^{(i)} \in [\rho, 1-\rho]$. With probability $1-\delta$,we have,*

$$|E_{x\sim Q}[-\log P(x)] - \frac{1}{m}\sum_{y\in T} -logP(y)| \le n\log(1/\rho)\sqrt{\frac{\log(1/\delta)}{m}},$$

where $y \in T$ are i.i.d. sample from Q.

Proof: For any y we have $-\log P(y) < n\log 1/\rho$. The lemma follows from a Chernoff bound. □

The above lemma states that we have a good estimate of the log-loss of a distribution (and therefore of the KL divergence). We need to show that if we have a model which is near the true model then its log-loss is small. (The proof is found in the Appendix.)

Lemma 4.13 *Let $Q = (\mathbf{q}^{(1)}, \mathbf{q}^{(2)}, \gamma)$ be a model of mixture of two product distributions. Let $P = (\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \gamma')$ be a model of mixture of two product distributions. Also, $|\gamma-\gamma'| < \lambda_\gamma$ and $\mathbf{p}_j^{(b)} \in [\rho, 1-\rho]$. Then,*

$$KL(Q\|P) = E_Q[\log(Q(x)/P(x))] \le \gamma\frac{\|\mathbf{p}^{(1)} - \mathbf{q}^{(1)}\|_2^2}{2\rho(1-\rho)} + (1-\gamma)\frac{\|\mathbf{p}^{(2)} - \mathbf{q}^{(2)}\|_2^2}{2\rho(1-\rho)} + \lambda_\gamma n\log(1/\rho)$$

The following corollary would be later used to show that if d is small we can approximate the mixture by a single product distribution.

Corollary 4.14 *Let $Q = (\mathbf{q}^{(1)}, \mathbf{q}^{(2)}, \gamma)$ be a model of mixture of two product distributions and let $\mathbf{q}^{(a)} = \gamma\mathbf{q}^{(1)} + (1-\gamma)\mathbf{q}^{(2)}$, and $\mathbf{q}_j^{(b)} \in [\rho, 1-\rho]$, for $b \in \{1,2\}$. Then,*

$$KL(Q\|\mathbf{q}^{(a)}) = E_Q[\log(Q(x)/\mathbf{q}^{(a)}(x))] \leq \frac{\|\mathbf{q}^{(1)} - \mathbf{q}^{(2)}\|_2^2 - \max\{(q_j^{(1)} - q_j^{(2)})^2\}}{2\rho(1-\rho)}$$

4.7 Putting it all together

Theorem 4.15 *Our algorithm, given access to examples from $(\mathbf{q}^{(1)}, \mathbf{q}^{(2)}, \gamma)$, with probability $1-\delta$, outputs a model $(\hat{\mathbf{q}}^{(1)}, \hat{\mathbf{q}}^{(2)}, g)$ such that,*

$$KL[(\mathbf{q}^{(1)}, \mathbf{q}^{(2)}, \gamma)\|(\hat{\mathbf{q}}^{(1)}, \hat{\mathbf{q}}^{(2)}, g)] \leq \epsilon$$

and the running time and the sample size of our algorithm is bounded by $poly(n, \epsilon, \log(1/\delta)$ assuming that $q_j^{(b)} \in [\alpha, 1-\alpha]$ for $b \in \{1,2\}$ and for some constant α.

Proof: The proof will be done in the following steps. First we show (under some "optimistic" assumptions) that one of the models that we will consider is near to the true model, and thus our output model has a small KL divergence. Second we show that with high probability, in some iteration, all our "optimistic" assumptions hold. Third, we compute the sample size. Fourth, we compute the running time. Finally, we discuss the case where d is small.

We start by showing that, with high probability, one of the models that we will consider has a small KL divergence. Since we are approximating well the log-loss it implies that the model with the lowest log-loss has also the smallest KL divergence.

Let us make the assumption that S is good, $\mathbf{x}^*$ is good, and $R(\mathbf{x}^*, \hat{\theta}(\mathbf{x}^*)) \geq \gamma d/8$. By Lemma 4.11, with probability $1-\delta$, for $b \in \{1,2\}$, there is an integer j_b, such that

$$\|\mathbf{q}_{\bar{S}}^{(b)} - (\hat{\mathbf{q}}^{(a)} + j_b\lambda(\hat{\mathbf{q}}_{\bar{S}}^{(a)} - \hat{\mathbf{q}}_{\bar{S}}^{(s)}))\|_\infty < \frac{6\lambda}{|\gamma - \mu|} < \frac{c_3\sqrt{n}\lambda}{\gamma^2 d}$$

where the last inequality is by Lemma 4.7. Let $\hat{\mathbf{q}}^{(b,e)}$ be the vectors associated with j_b, $b \in \{1,2\}$.

Consider the model $(\hat{\mathbf{q}}_{\bar{S}}^{(1,e)}, \hat{\mathbf{q}}_{\bar{S}}^{(2,e)}, \gamma')$, where $\gamma' = \lambda_\gamma\lfloor\gamma/\lambda_\gamma\rfloor$. Now by Lemma 4.13 we have the the KL divergence between those two models is bounded by,

$$\lambda_\gamma n\log(1/\rho) + n\frac{n\lambda^2}{\rho(1-\rho)d^2\gamma^4}.$$

For

$$\lambda_\gamma < \frac{\epsilon}{n\log(1/\rho)} \quad \text{and} \quad \lambda < \frac{\epsilon d\gamma^2\sqrt{\rho(1-\rho)}}{n}$$

the model $(\hat{\mathbf{q}}_{\bar{S}}^{(1,e)}, \hat{\mathbf{q}}_{\bar{S}}^{(2,e)}, \gamma')$ has KL divergence is at most 2ϵ.

By Lemma 4.12 we have that the errors in our estimation of the log-loss is at most ϵ. This implies that the observer error of $(\hat{\mathbf{q}}_{\bar{S}}^{(1,e)}, \hat{\mathbf{q}}_{\bar{S}}^{(2,e)}, \gamma')$ is at most 3ϵ. Therefore the model with the lowest observed log-loss would have KL divergence of at most 4ϵ.

The next step is to show that with high probability all our assumptions hold. First we show the probability that S is good. Consider the procedure `MAIN`. The probability that we choose a good S is at least $1/2$ each time in the loop, by Corollary 4.2. Therefore with probability $1-\delta$ some set S we choose is good.

For each set S, in procedure `Estimate_SET` runs N loops. The probability that $\mathbf{x}^*$ is good is, by Lemma 4.4 and Corollary 4.5, at least $p/2 = \gamma \min\{d/64, 1/4\}$. Given that $\mathbf{x}^*$ is good, by Lemma 4.6, the probability that $R(\mathbf{x}^*, \hat{\theta}(\mathbf{x})) \geq \gamma d/8$ is at least $1-\delta$. Therefore, for $N = (\max\{128/d, 8\}\log(2/\delta)$ the probability of success is at least $1-\delta/2$. This completes the correctness.

The running time is bounded by,

$$O(\log(1/\delta) N m (\beta/\lambda)^2 (1/\lambda_\gamma) n) = poly(\log(1/\delta), n, \epsilon).$$

The only parameter we still need to bound is $\beta = 2/\max_i\{\hat{\Delta}_i\}$. For $\hat{\Delta} = \hat{\mathbf{q}}_{\bar{S}}^{(1)} - \hat{\mathbf{q}}_{\bar{S}}^{(2)}$, $\|\hat{\Delta}\|_\infty > d^{1.5}\gamma^2/n$. This implies that $\beta > 2n/d^{1.5}\gamma^2$. By our assumption that $\mathbf{q}_j^{(b)} \in [\alpha, 1-\alpha]$, we can set $\rho = \alpha$, which is a constant.

The sample size required is,

$$\max\{\frac{1}{p}\log(1/\delta), \frac{n}{(\gamma\ d)^2}\log(1/\gamma d\delta), \frac{n^2\gamma^4}{\epsilon^2 d^2 \rho(1-\rho)}, \frac{n^2}{\epsilon^2}(\log^2 1/\rho)(\log w/\delta))\} = poly(\log(1/\delta), n, \epsilon),$$

where w is the number of models we generate and test, which is at most $(\beta/\lambda)^2\lambda_\gamma$. Note that (ignoring logarithmic factors) the behavior is n^2/ϵ^2, which is a result of the accuracy requirement on the log-loss.

Note that both the sample size and the running time are inversely proportional to d, which is a problem if d is very small. We like to show that we can replace in the bounds d by $\max\{d, \epsilon\rho(1-\rho)/4\}$, which will solve the problem for small values of D. We show that if d is very small we can approximate the mixture distribution by a single product distribution. If $d < \epsilon\rho(1-\rho)/4$ then by Corollary 4.14,

$$KL[(\mathbf{q}^{(1)}, \mathbf{q}^{(2)}, \gamma)\|\mathbf{q}^{(a)}] \leq \frac{4d}{\rho(1-\rho)} \leq \epsilon$$

where $\mathbf{q}^{(a)} = \gamma\mathbf{q}^{(1)} + (1-\gamma)\mathbf{q}^{(2)}$. Since we test $\hat{\mathbf{q}}^{(a)}$ in `MAIN` the model we output has an observed KL divergence of at most 2ϵ, and thus a true KL divergence of at most 3ϵ. □

References

[CT91] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* Wiley, 1991.

[FR95] Yoav Freund and Dana Ron. Learning to model sequences generated by switching distribution s. In *Proceedings of the Eighth Annual Conference on Computational Learning Theory*, pages 41–50, 1995.

[Hoe56] W. Hoeffding. On the distribution of the number of successes in independent trials. *Annals of Mathematical Statistics*, 27:713–721, 1956.

[KMR+94] Michael Kearns, Yishay Mansour, Dana Ron, Ronitt Rubinfeld, Robert E. Schapire, and Linda Sellie. On the learnability of discrete distributions. In *Proceedings of the Twenty-Sixth Annual ACM Symposium on the Theory of Computing*, pages 273–282, 1994.

[TSM85] D.M. Titterington, A.F.M. Smith, and U.E. Makov. *Statistical Analysis of Finite Mixture Distributions.* John Wiley & Sons, 1985.

[YS68] S. J. Yakowitz and J. D. Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39:258–263, 1968.

A Proofs

Proof of Lemma 4.6: We can think of the sum $\sum_{\mathbf{y}\in T} < \mathbf{x}, \mathbf{y} >$ as being generated in the following fashion. In each of m independent trials we choose b to be 1 or 2 with probabilities γ or $1-\gamma$ respectively. Example $\mathbf{y}$ generated according to $\mathbf{q}^{(b)}$ are in $T^{(b)}$. Let m_b be the size of $T^{(b)}$. We rewrite $\hat{\theta}(\mathbf{x})$ as follows,

$$\hat{\theta}(\mathbf{x}) = \frac{m_1}{m}\left[\frac{1}{m_1}\sum_{y\in T^{(1)}} < \mathbf{x}, y >_S\right] + \frac{m_2}{m}\left[\frac{1}{m_2}\sum_{y\in T^{(2)}} < \mathbf{x}, y >_S\right]$$

Recall that $\theta(\mathbf{x})$ is,

$$\theta(\mathbf{x}) = \gamma E_{\mathbf{y}^{(1)}\sim\mathbf{q}^{(1)}}[< \mathbf{x}, \mathbf{y}^{(1)} >_S] + (1-\gamma)E_{\mathbf{y}^{(2)}\sim\mathbf{q}^{(2)}}[< \mathbf{x}, \mathbf{y}^{(2)} >_S].$$

With probability $1-\delta$ we have that $|\gamma - \frac{m_1}{m}| \leq \sqrt{\frac{\log(1/\delta)}{m}} = \epsilon_\gamma$. Now we bound the deviation in the estimation for each center.

$$\left|[\frac{1}{m_b}\sum_{y\in T^{(b)}} < \mathbf{x}, y >_S] - E_{\mathbf{y}^{(b)}\sim\mathbf{q}^{(b)}}[< \mathbf{x}, \mathbf{y}^{(b)} >_S]\right| \leq \sqrt{\frac{n\log(1/\delta)}{m_b}} = \epsilon_b$$

We can bound the total deviation by

$$\frac{m_1}{m}\epsilon_1 + \frac{m_2}{m}\epsilon_2 + \epsilon_\gamma \mathtt{diff}(\mathbf{x}) = \sqrt{\frac{m_1}{m}\frac{n\log(1/\delta)}{m}} + B\sqrt{\frac{m_2}{m}\frac{n\log(1/\delta)}{m}} + \sqrt{\frac{\log(1/\delta)}{m}}\mathtt{diff}(\mathbf{x}) < \frac{\gamma\mathtt{diff}(\mathbf{x})}{2}$$

for $m > c(n\log(1/\delta))/(\gamma^2 d^2)$. $\square$

Proof of Lemma 4.7: We are interested in bounding first $\mu^{(b)}$, for $b \in \{1,2\}$. Recall that $< \mathbf{x}^*, \mathbf{y}^b >_S = \sum_{j\in S} x^*_j y^b_j$. The problem is that the expectation of y^b_j is $q^{(b)}_j$, and is different for different js. The following lemma by Hoeffding [Hoe56] reduces this case to the case where all the y^b_j are i.i.d.

Lemma A.1 *Let X_i be n binary random variables which are independent. Let $\mu = E[\sum_{i=1}^n X_i]$. For an integer $k \leq \mu - 1$,*

$$Pr[\sum_{i=1}^n X_i \leq k] \quad \leq \quad Pr[\sum_{i=1}^n Y_i \leq k]$$

where Y_i are i.i.d binary random variables with $E[Y_i] = \mu/n$.

This enables us to reduce our case, with y^b_j of different distributions, to a simple case where we have z^b_j i.i.d., and the expectation of z^b_j is $\eta^{(b)} = \sum_{j\in S} q^{(b)}_j/n$, for $b \in \{1,2\}$. We now like to bound, $\Pr[\sum_{j\in S} x^*_j z^b_j > \hat{\theta}]$.

Since the z^b_j are i.i.d. random variables, the median is exactly (up to rounding) the expected value. Therefore,

$$\Pr[\sum_{j\in S} x^* z^1_j \quad \geq \hat{\theta} + R] \geq 1/2$$

Similarly,

$$\Pr[\sum_{j\in S} x^* z^2_j \quad \leq \hat{\theta} - R] \geq 1/2$$

Also, for some constant c_1, for $R < c_1\sqrt{n}$,

$$\Pr[\sum_{j\in S} x^* z_j^1 \ \in (\hat{\theta}, \hat{\theta}+R)] \geq c_2 \frac{R}{\sqrt{n}},$$

for some constant c_2. This implies,

$$\Pr[\sum_{j\in S} x^* z_j^1 \ \geq \hat{\theta}] \geq 1/2 + c_2 \frac{R}{\sqrt{n}}.$$

Therefore,

$$\mu^{(1)} = \Pr_{y\sim q^{(1)}}[< x, y >_S \ \geq \hat{\theta}] \geq \Pr[\sum_{j\in S} x^* z_j^1 \ \geq \hat{\theta}] \geq 1/2 + c_2 \frac{R}{\sqrt{n}}.$$

Similarly,

$$\mu^{(2)} = \Pr_{y\sim q^{(2)}}[< x, y >_S \ \geq \hat{\theta}] \leq \Pr[\sum_{j\in S} x^* z_j^2 \ \geq \hat{\theta}] = 1 - \Pr[\sum_{j\in S} x^* z_j^2 \ < \hat{\theta}] \leq 1/2.$$

This implies that,

$$\mu - \gamma = \frac{\gamma\mu^{(1)}}{\gamma\mu^{(1)} + (1-\gamma)\mu^{(2)}} - \gamma = \gamma(1-\gamma)\frac{\mu^{(1)} - \mu^{(2)}}{\gamma\mu^{(1)} + (1-\gamma)\mu^{(2)}} \geq c_3 \frac{\gamma R}{\sqrt{n}},$$

which completes the proof. □

Proof of Lemma 4.13: Using a general lemma (see [CT91]) we have that

$$KL(\gamma\mathbf{q}^{(1)} + (1-\gamma)\mathbf{q}^{(2)} \| \gamma\mathbf{p}^{(1)} + (1-\gamma)\mathbf{p}^{(2)}) \leq \gamma KL(\mathbf{q}^{(1)}\|\mathbf{p}^{(1)}) + (1-\gamma)KL(\mathbf{q}^{(2)}\|\mathbf{p}^{(2)}).$$

We are interested in computing,

$$KL(\gamma\mathbf{q}^{(1)} + (1-\gamma)\mathbf{q}^{(2)} \| \gamma'\mathbf{p}^{(1)} + (1-\gamma')\mathbf{p}^{(2)}).$$

Assume that $\gamma \leq \gamma'$. (The case $\gamma > \gamma'$ would be similar.) Rewrite the KL divergence as,

$$KL\left(\gamma\mathbf{q}^{(1)} + (1-\gamma)\mathbf{q}^{(2)} \| \gamma\mathbf{p}^{(1)} + (1-\gamma)[\frac{\gamma'-\gamma}{1-\gamma}\mathbf{p}^{(1)} + \frac{1-\gamma'}{1-\gamma}\mathbf{p}^{(2)}]\right)$$

We can bound this KL divergence by,

$$\gamma KL(\mathbf{q}^{(1)}\|\mathbf{p}^{(1)}) + (1-\gamma)KL(\mathbf{q}^{(2)}\|\mathbf{p}^{(2)}) + (\gamma'-\gamma)KL(\mathbf{q}^{(2)}\|\mathbf{p}^{(1)})$$

Computing $KL(\mathbf{q}^{(b)}\|\mathbf{p}^{(b)})$, $b \in \{1, 2\}$, is rather simple, since we have a product distribution.

$$KL(\mathbf{q}^{(b)}\|\mathbf{p}^{(b)}) = \sum_{i=1}^{n} KL(\mathbf{q}_i^{(b)}\|\mathbf{p}_i^{(b)}) \leq \frac{\|\mathbf{p}^{(b)} - \mathbf{q}^{(b)}\|_2^2}{2\rho(1-\rho)}$$

Finally, since $\mathbf{p}_i^{(b)} \in [\rho, 1-\rho]$, we can bound $(\gamma'-\gamma)KL(\mathbf{q}^{(2)}\|\mathbf{p}^{(1)})$ by $\lambda_\gamma n \log(1/\rho)$. □

Procedure MAIN

Input: A training set $T = \{x_1, \ldots, x_m\}$ where $x_i \in \{-1, +1\}^n$.

Compute the likelihood that the model $\hat{\mathbf{q}}^{(a)}$ generated T, where $\hat{\mathbf{q}}^{(a)}_{\bar{S}} = \frac{1}{m} \sum_{\mathbf{y}_j \in T} (\mathbf{y}_j)_{\bar{S}}$.

Repeat for $i = 1, 2, \ldots, \log(2/\delta)$

1. Choose a set S by including each index $1 \leq i \leq n$ in S with probability $1/2$. Let $\bar{S}$ be the complement of S.
2. Call `ESTIMATE_SET`(T, S) and `ESTIMATE_SET`$(T, \bar{S})$. Let $(\mathbf{q}^{(1,\bar{S})}, \mathbf{q}^{(2,\bar{S})}, g_1)$ and $(\mathbf{q}^{(1,S)}, \mathbf{q}^{(2,S)}, g_2)$ be the outputs.
3. Create eight models, $(\mathbf{q}^{(b_1,S)} \| \mathbf{q}^{(b_2,\bar{S})}, \mathbf{q}^{(\bar{b}_1,S)} \| \mathbf{q}^{(\bar{b}_2,S)}, g_{b_3})$, where $b_1, b_2, b_3 \in \{1, 2\}$, and $\bar{b} = 3 - b$.
4. Compute the likelihood that each model generated the sample T.

Output: The model with the maximum likelihood to generate T.

Figure 1: The algorithm for learning a mixture of two product distributions

Procedure ESTIMATE_SET

Input: A training set $T = \{x_1, \ldots, x_m\}$ where $x_i \in \{-1, +1\}^n$, and a set $S \subset \{1, \ldots, n\}$.

Repeat for $i = 1, 2, \ldots, N$

1. Select a vector $\mathbf{x}^*$ uniformly at random from the training set T.
2. Set a threshold $\hat{\theta} = \frac{1}{m} \sum_{\mathbf{y} \in T} < \mathbf{x}^*, \mathbf{y} >_S$.
3. Split the training set T into two parts:
$$T^{(\leq)} = \left\{ \mathbf{y} \in T :< \mathbf{y}, \mathbf{x} >_S \;\leq\; \hat{\theta} \right\}, \qquad T^{(>)} = \left\{ \mathbf{y} \in T :< \mathbf{y}, \mathbf{x} >_S \;>\; \hat{\theta} \right\}$$
4. Calculate the average of each subset on the remaining coordinates:
$$\hat{\mathbf{q}}_{\bar{S}}^{(s)} = \frac{1}{|T^{(\leq)}|} \sum_{\mathbf{y}_j \in T^{(\leq)}} (\mathbf{y}_j)_{\bar{S}}, \qquad \hat{\mathbf{q}}_{\bar{S}}^{(a)} = \frac{1}{m} \sum_{\mathbf{y}_j \in T} (\mathbf{y}_j)_{\bar{S}}.$$
5. Let $\hat{\Delta} = \hat{\mathbf{q}}_{\bar{S}}^{(s)} - \hat{\mathbf{q}}_{\bar{S}}^{(a)}$, and $\beta = 2/\max_j\{\hat{\Delta}_j\}$. Let F include the points $\hat{\mathbf{q}}^{(\leq)} + j\lambda\hat{\Delta}$ for integer j, $-\beta/\lambda \leq j \leq \beta/\lambda$. Let Γ include the numbers $i\lambda_\gamma$ for $0 \leq i \leq 1/\lambda_\gamma$.
6. For each pair $\mathbf{q}^{(1)}$ and $\mathbf{q}^{(2)}$ in F and g in Γ compute the likelihood that the model $(\texttt{round}(\rho, \mathbf{q}^{(1)}), \texttt{round}(\rho, \mathbf{q}^{(2)}), g)$ generated T.

Output: the model with the highest likelihood.

Figure 2: The procedure `Estimate_SET`. The parameters λ, λ_γ and N are determined in the proofs.