REPORT DOCUMENTATION PAGE

AFRL-SR-BL-TR-98-
0738

Public reporting burden for this collection of information is estimated to average 1 hour per response, incluc
and maintaining the data needed, and completing and reviewing the collection of information. Send con
information, including suggestions for reducing this burden, to Washington Headquarters Services, Director
1204, Arlington, VA 22202-4302, and to the Office of management and Budget, Paperwork Reduction Projec

s, gathering
collection of
hway, Suite

1. AGENCY USE ONLY (Leave Blank)

2. REPORT DATE
April, 1994

3. RE
Final

4. TITLE AND SUBTITLE
The Complexity of Learning from a Mixture of Labeled and Unlabeled Examples

5. FUNDING NUMBERS

6. AUTHORS
Joel E. Ratsaby

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Pennsylvania

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NI
4040 Fairfax Dr, Suite 500
Arlington, VA 22203-1613

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*
See Attachment

14. SUBJECT TERMS

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
Unclassified

18. SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

19. SECURITY CLASSIFICATION OF ABSTRACT
Unclassified

20. LIMITATION OF ABSTRACT
UL

THE COMPLEXITY OF LEARNING FROM A MIXTURE OF LABELED AND UNLABELED EXAMPLES

Joel E. Ratsaby
Advisor: Santosh S. Venkatesh

A DISSERTATION
in
ELECTRICAL ENGINEERING

Presented to the Faculties of the University of Pennsylvania in
Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy

April 1994

____

Supervisor of Dissertation

____

Graduate Group Chairperson

Dedication

To my wife Anat,

for her patience and understanding

To my parents Aron and Michele,

for their overall support

and to my son Nethanel,

for his gracefulness

Acknowledgement

I would like to express my sincere gratitude to my advisor Professor Santosh S. Venkatesh. He has provided me with guidance, support, and encouragement throughout my graduate studies. Both in his classes, and in personal consultations he has sparked my curiosity in numerous subjects and vigorously provided the opportunity to explore them. I thank him for his high mark of quality that served as an excellent reference for my research.

I greatly appreciate and acknowledge the fellowship from the Department of Defense, Air Force Office of Scientific Research, and for the partial support that was provided from AFOSR through research contracts.

Finally, I would like to thank my wife Anat who has borne the difficult life of being married to a graduate student.

ABSTRACT

THE COMPLEXITY OF LEARNING FROM A MIXTURE OF LABELED AND UNLABELED EXAMPLES

Joel E. Ratsaby

Santosh S. Venkatesh

The learning of a pattern classification rule rests on acquiring information to constitute a decision rule that is close to the optimal Bayes rule. Among the various ways of conveying information, showing the learner examples from the different classes is an obvious approach and ubiquitous in the pattern recognition field. Basically there are two types of examples: *labeled* in which the learner is provided with the correct classification of the example and *unlabeled* in which this classification is missing. Driven by the reality that often unlabeled examples are plentiful whereas labeled examples are difficult or expensive to acquire we explore the tradeoff between labeled and unlabeled sample complexities (the number of examples required to learn to within a specified error), specifically getting a quantitative measure of the reduction in the labeled sample complexity as a result of introducing unlabeled examples. This problem was posed in this form by T. M. Cover and may be succinctly, if inexactly, stated as follows: *How many unlabeled examples is one labeled example worth?*

The direction taken in this dissertation focuses on the archetypal problem of learning a classification problem with two pattern classes that are typified by feature vectors, i.e., examples drawn from class conditional Gaussian distributions and where the learning approaches are parametric and nonparametric. Denoting the dimensionality of the example-space as N, and the number of labeled and unlabeled examples as m and n respectively, then for specific algorithms, it is shown that under a nonparametric scenario the classification error probability decreases roughly as $O\left(\left(c_0 n^{-2/N}\right)^{\lg N}\right) + O\left(e^{-c_1 m}\right)$, and in the parametric scenario the error decreases roughly as $O\left(N^{3/5} n^{-1/5}\right) + O\left(e^{-c_1 m}\right)$, where $c_0, c_1 > 0$ are constants with respect to N, m, and n. This shows that in both the parametric and nonparametric cases it takes roughly exponentially more unlabeled examples than labeled examples for the same reduction in error. When considering the effect of the dimensionality N, roughly speaking, a labeled example is worth exponentially more in the nonparametric than in the parametric scenario.

The parametric approach uses the Maximum Likelihood technique with labeled and unlabeled samples to construct a decision rule estimate. In this scenario the learner knows the parametric form of the pattern class densities. Sufficient finite sample complexities are established by which the value of one labeled example in terms of the number of unlabeled examples is determined to be polynomial in the dimensionality N. The analysis may provide the details for broadening the results to other non Gaussian parametric based families of problems. An extension to the case of different *a priori* class probabilities is investigated under this parametric scenario, and for the non-unit covariance Gaussian problem it is conjectured that the value of a labeled example is still polynomial in N.

In the nonparametric scenario the primary focus is on an algorithm which is based on Kernel Density Estimation. It uses a mixed sample to construct a decision rule where now the learner has significantly less side information about the class densities. The finite sample complexities for learning the Gaussian based problem are established by which the value of one labeled example is determined to be exponential in the dimensionality N. An extension to a larger family of nonparametric classification problems is provided where the same tradeoff applies. A variant of this approach is investigated in which only a finite number of functional values of the underlying mixture density are estimated. This yields a smaller tradeoff but is still exponential in N. The mixed sample complexities for the classical k-means clustering procedure are also determined.

An experimental investigation using neural networks examines the value of a labeled example when learning a classification problem based on a Gaussian mixture. For other classification problems, the cost of learning measured by the labeled sample size as a function of the dimensionality N, is shown to be lower for a two-layer network than with the regular single layer Kohonen network. This is attributed to the better discrimination ability of the partition of the classifier.

Contents

List of Figures

Chapter 1

Introduction

The problem of learning a classification decision rule (cf. Duda & Hart [1], Fukunaga [2]) has been the subject of a large, diverse body of literature spanning at least the last 60 years. It has been approached by many methods in statistics and pattern recognition. In its basic form, we are given two classes, "1" and "2", of patterns that are represented by vectors whose elements represent various features about each of the two patterns. For instance, in the medical diagnosis of cancer, the two pattern classes are "malignant" and "benign" cells. A cell in a class has a variety of features such as size, color, shape, genetic code, etc., by which it is described. For our purposes, we assume the features are represented by a point x in N-dimensional Euclidean space. The objective is to find a decision rule which when presented with a pattern (i.e., a vector) that is drawn randomly either from class "1" (with probability p_1) or class "2" (with probability p_2), produces a label which identifies it as belonging to the true class of origin. Ideally we would desire a rule that never misclassifies a pattern. This however is only achievable if the pattern classes have non-overlapping probability one supports; in general the best achievable rule (the Bayes classifier) has a nonzero misclassification error, P_{Bayes}, determined by the class conditional probability density functions $f_1(x)$, $f_2(x)$ and the *a priori* class probabilities p_1 and p_2.

The Bayes decision rule is derived from the following: let $l(i, j)$, with $i, j \in 1, 2$, denote the loss incurred when the classifier decides "i" while the true class of the pattern is "j". We limit our discussion to the case of the symmetric 0–1 loss function: $l(1,1) = l(2,2) = 0$ and $l(1,2) = l(2,1) = 1$ for which the expected loss is identically the probability of misclassification, P_{error}. In this case we have

$$P_{error} = \mathbf{E}l(i,j) = \mathbf{E}_x\mathbf{E}\left(l(i,j)|x\right).$$

The inner expectation is

$$\begin{aligned} \mathbf{E}\left(l(i,j)|x\right) &= \mathbf{P}(i=1,j=2|x) \;+\; \mathbf{P}(i=2,j=1|x) \\ &= \mathbf{P}(i=1|j=2,x)\,p(j=2|x) \;+\; \mathbf{P}(i=2|j=1,x)\,p(j=1|x) \end{aligned}$$

where $p(j=1|x)$, $p(j=2|x)$ are the *a posterior* class probabilities. This expectation is a nonnegative quantity hence in order to minimize P_{error} it suffices to specify a classification rule which minimizes it.

A classifier can be considered as a mapping

$$C : \mathbb{R}^N \to \{1,2\}$$

or a partition of the feature space into disjoint regions R_1, R_2, where

$$C(x) = \begin{cases} 1 & \text{if } x \in R_1, \\ 2 & \text{if } x \in R_2. \end{cases}$$

This is a deterministic rule hence we have

$$\mathbf{P}(i=1|j=2,x) = 1_{R_1}(x)$$

and

$$\mathbf{P}(i=2|j=1,x) = 1_{R_2}(x)$$

where we use the notation $1_A(x)$ to denote the indicator function for the set A, i.e.,

$$1_A(x) = 1 \text{ if } x \in A \text{ and } 1_A(x) = 0 \text{ if } x \notin A.$$

The optimal (or Bayes) classifier is one which minimizes

$$p(j=1|x)1_{x\in R_2}+p(j=2|x)1_{x\in R_1}.$$

Only the decision regions R_1, R_2 are controllable and it is clear that the minimizing choice is

$$R_1=\{x:p(j=2|x)\leq p(j=1|x)\}\text{ and }R_2=\{x:p(j=1|x)<p(j=2|x)\}.$$

The decision border is

$$\{x:p(j=1|x)=p(j=2|x)\}=\{x:p_1f_1(x)=p_2f_2(x)\}, \tag{1.1}$$

where the last equality follows from Bayes' theorem, $f_1(x)p_1=p(j=1|x)f(x)$, with $f_i(x)$ being the class conditional densities, and p_i are the *a priori* class probabilities, $i=1,2$.

Hence $f_1(x), f_2(x)$ and p_1, p_2 determine the Bayes decision rule and the resulting (minimum) error of the Bayes classifier is zero if and only if the pattern classes have disjoint probability one supports.

If the class conditional densities and the priors were known. we can hence determine the Bayes optimal decision rule with $P_{error}\equiv P_{Bayes}$. However, realistically, this is a rare occurrence; as in the above medical diagnosis example, such detailed prior information is usually not available. We can at best hope for partial information about the classes, a typical scenario providing *randomly drawn* data according to the unknown probability distributions. This will be our focus here. Using a random sample, our goal is to determine a rule that achieves a given error probability which is not much larger than P_{Bayes}. More precisely, for $\epsilon>0$ chosen suitable small, we would like to obtain P_{error} bounded between P_{Bayes} and $P_{Bayes}(1+\epsilon)$.

Broadly speaking, the approach to classifier design is to use randomly drawn examples to estimate the class conditional densities and plug them into the above expression that relates the densities with the decision regions. The resulting rule has a classification error P_{error} which may differ from the optimal P_{Bayes}. This approach can be validated, at least asymptotically in the limit of large sample sizes. For example, it is shown in Glick [38] that sample-based density plug-in rules are asymptotically optimal i.e., minimize the classification error when the density estimates are themselves densities and are strongly consistent.

Classification methods vary according to the type and amount of additional side-information that is available. Direct information about the class densities leads to an estimate of the likelihood ratio and hence of the optimal decision border. More typically, only partial information is accessible; for instance: the parametric form of the distributions but not the parameter value; knowledge that the distributions are monotone decreasing; or that the mixture (i.e., weighted sum of the class conditionals) has k modes (peaks).

Traditionally, in the fields of statistics and pattern recognition, there are two main categories for density estimation: *parametric* and *non-parametric*. These are divided into various branches based upon the estimation method which depends on the information that is provided (or assumed) about the classes; for instance, if it is known that the class densities are of a given parametric form then the method of maximum-likelihood can be invoked. Once a density-estimation method is chosen, it remains to learn the constraints in the observed data and deduce the density that is closest (w.r.t. some quantitative measure) to the true underlying class densities.

If information regarding the densities is not available then one must resort to assumptions or heuristics based on some rules of thumb, in order to construct a decision border that hopefully has low P_{error}. For instance, observed data can be tested for clusters and a partition of the feature space is constructed such that each cluster is captured by one disjoint subset (a cell) of the partition. Then each cell gets associated with the class corresponding to the class of the majority of the observed examples. This induces a decision rule which may have low error. Neural network algorithms as applied to learning classification, are one of numerous *ad hoc* methods that fall under this category.

1.1 Classification Methodologies

The following is a nonexhaustive list of a few popular classification methodologies (cf. Fukunaga [2], Duda & Hart [1], Izenman [3]):

Parametric density estimation: We are restricted to a class of parametric density functions, $f(x|\theta)$, with known form and unknown true parameter θ_0.

Maximum Likelihood Estimation (MLE): The parameter is viewed as a deterministic variable, and one solves for the value of θ that achieves the global maximum of the likelihood function $L(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(x_i|\theta)$, where $\{x_i, 1 \leq i \leq n\}$ is a set of examples drawn independently and distributed according to $f(x|\theta_0)$. This value of θ is defined as the estimator $\hat{\theta}$. Defining the decision regions as in (1.1) using the estimates $f(x|\hat{\theta}_1)$ and $f(x|\hat{\theta}_2)$ for the two class densities yields an estimate of the Bayes classifier.

Maximum likelihood parameter estimates typically exhibit optimal properties (cf. Bickel & Doksum [39]). They are often asymptotically consistent, i.e., converge to the true unknown parameter as the sample size increases and are asymptotically efficient, i.e. the rate of decrease of their variance converges to the Cramér-Rao lower bound. Hence a decision rule based on the MLE-density estimates are, at least theoretically, attractive for solving a classification problem.

Bayesian Estimation: We seek the distribution of x given the random sample $x^n = (x_1, \ldots, x_n)$, i.e., $f(x|x^n)$. The parameter θ is viewed as a random variable with an *a priori* distribution $h(\theta)$. Any side information we might have about the unknown parameter is assumed to be contained in this distribution. We can write

$$f(x|x^n) = \int f(x|x^n, \theta) g(\theta|x^n)\, d\theta = \int f(x|\theta) g(\theta|x^n)\, d\theta$$

because x is independent of the sample x^n. By assumption, $f(x|\theta)$ is known, hence the desired density is the expected value of $f(x|\theta)$ w.r.t. the possible values of θ based on the random sample x^n.

Learning involves updating the *a posterior* distribution $f(\theta|x^n)$, whose variance decreases as the number of examples increases, whereby the integral on the right tends to $f(x|\theta_0)$. For instance in learning the mean of a Gaussian distributed random variable, the variance of the estimator $\hat{\theta}$ is asymptotic to σ^2/n as $n \to \infty$.

Moment Estimation: The parameter vector θ is composed of the moments m_i, $1 \leq i \leq k$, of the true distribution. These are estimated by the empirical (sample) moments $\hat{m}_i = \frac{1}{n} \sum_{j=1}^{n} x_j^i$. These estimates are consistent and yield consistent density estimates.

Non-parametric density estimation: Very little information is available. Neither the form of the class conditional distributions nor any of the parameters if such exist, are known.

Kernel Estimation: A function $K_{\sigma_n, x_i}(x)$, called the Kernel, is placed centered at each example x_i, i.e.,

$$K_{\sigma_n, x_i}(x) = K\left(\frac{x_i - x}{\sigma_n}\right)$$

where σ_n is a smoothing parameter. The smoothed average of the functions, each centered at one of the n examples, forms the density estimate

$$f_n(x) \equiv \frac{1}{n\sigma_n^N} \sum_{i=1}^{n} K\left(\frac{x_i - x}{\sigma_n}\right).$$

where N denotes the dimension. The bias of the estimate decreases as the smoothing parameter $\sigma_n \to 0$. However, the rate of decrease of the variance of the estimate as $n \to \infty$, becomes worse, i.e., slower as $\sigma \to 0$. By selecting σ_n to decrease to zero at the right rate, it is possible for $f_n(x)$ to be strongly consistent, uniformly for all $x \in \mathbb{R}^N$ (cf. Pollard [21]).

The shape of the kernel function can be designed to accelerate the decrease in the bias of the estimate as $n \to \infty$ (cf. Izenman [3], Silverman [40]). For learning classification, it may not be necessary for $f_n(x)$ to be a *bona fide* pdf as for instance in our investigation in Chapter 5. There the modes of the mixture density can directly determine the Bayes decision regions, and may be estimated by the modes of a kernel estimate which takes also negative values. With such a kernel, better rates of decrease for the bias are achievable.

Histogram Methods: The histogram method is an old and basic approach to density estimation. The feature space is divided into cells, $c_i \subset \mathbb{R}^N, 1 \leq i \leq M$, and the density function is approximated by the number of examples that fall in each cell. In a one dimensional sample space the estimate is

$$\hat{f}(x) = \frac{1}{n\sigma_n} \sum_{i=1}^{M} n_i 1_{x \in c_i},$$

where n_i is the number of examples in c_i, and σ_n is the cell width. The histogram density estimate is suboptimal and its defects include the discontinuity at cell boundaries and its strong sensitivity to the location of the origin, i.e., shifting the starting point of the first cell can result in very different looking histograms. For optimal error rates, the cell width needs to decrease slower than n^{-1} as $n \rightarrow \infty$. But even its optimal error rate is substantially slower than most other kinds of density estimators.

Direct classification approaches: These methods do not estimate the class conditional densities but instead directly construct a decision rule using the randomly drawn examples.

Nearest Neighbor Rule: A partition of the feature space is constructed by drawing m labeled examples, x_i, $1 \leq i \leq m$, and defining a voronoi cell of the partition to be

$$v_i \equiv \{x : |x - x_i| < |x - x_j|, j \neq i\}.$$

Each voronoi cell is assigned the label of the example x_i corresponding to it. The decision rule for classifying a given x is to assign to it the label of the voronoi cell in which it falls.

The nearest neighbor decision rule is suboptimal since even as the number of labeled examples tends to infinity its classification error need not tend to the Bayes optimal error. However its error is bounded (tightly) as

$$P_{Bayes} \leq P_{error,NN} \leq 2P_{Bayes}(1 - P_{Bayes}).$$

The simplicity of this rule, its near-optimal performance for small P_{Bayes} ($P_{error,NN}$ is upper bounded by twice the Bayes error for large sample sizes) and the fact that it is based on a Voronoi partition whose efficient implementation has been studied, are significant advantages.

Clustering Procedures: These methods aim to discover inherent clustering in the given sample of patterns, which possess strong feature-similarity. Each cluster constitutes one of the mutually disjoint decision regions of the classifier. The performance of the classifier depends on the metric that is used to measure similarity between examples. Different problems, with different class densities, may require different similarity measures for good classifier performance.

One way of learning the partition is to choose any one which extremizes a criterion function. For instance a simple criterion is

$$e(\{R_1, \ldots, R_c\}) = \sum_{i=1}^{c} \sum_{x \in R_i} |x - \mu_i|^2$$

where R_i, $1 \leq i \leq c$ are the clusters of a particular partition and μ_i is the average of the examples in the i^{th} cluster. Minimizing this function over the space of possible partitions may yield an optimal classifier, in particular for problems that have well separated pattern classes.

Gradient Descent Procedures: These procedures extremize some criterion function in order to obtain the desired classification rule. For instance, feedforward neural networks can implement general highly non linear mappings from the feature space $X = \mathbb{R}^N$ to the output space, Y, which for classification problems can be a finite set of integers whose elements represent the possible classes.

The neural net classifier is represented by a family of parametric functions $f_w(x)$, each indexed by a particular vector w of weights. There may be many neural nets, with different w that yield optimal classifiers. For a randomly drawn test vector $x \in X$, a criterion function $e(w)$, can be defined to measure the expected difference between the classifier output

$f_w(x)$ and the correct classification y of the vector x. Using a teaching set of examples, (x_i, y_i), $1 \leq i \leq m$, an algorithm, such as backpropagation (cf. Hinton et. al. [4]), can be used to search for any w that achieves a global minimum of the criterion function by stepping in small increments in the direction for which the gradient of $e(w)$ is minimum.

One of the main difficulties of such algorithms is choosing the starting point of the search. A bad starting point will yield a sequence of gradient descents leading to a local minimum associated with a nonoptimal classifier. Adding random noise to the learning rule may help increase the chance of reaching a global minimum.

1.2 Labeled and Unlabeled Examples

From our vantage point, we emphasize that, as in all scientific research where rules are to be discovered, observed data (or examples) is of primary necessity in all methods regardless of the *a priori* partial information. There are two fundamental types of examples, *labeled* and *unlabeled.* A labeled sample is a collection of m pairs (x_i, y_i), $1 \leq i \leq m$ where x_i is a feature vector and y_i is its corresponding class label; $y_i \in \{1,2\}$. The class label $y_i \in \{1,2\}$ is drawn at random according to the *a priori* probabilities p_1 and p_2; the feature vector x_i corresponding to y_i is then drawn at random according to the class–conditional density $f_{y_i}(x)$. An unlabeled sample consists only of the feature vectors x_i drawn according to the mixture density $f(x) = p_1 f_1(x) + p_2 f_2(x)$.

Labeled examples clearly contain more information than unlabeled examples and all things being equal would be the preferred form of data for the learner. However, as T. M. Cover [6] indicates, very often it is the case that unlabeled examples are more abundant and cheaper to acquire than labeled examples and for that reason mixed-sample learning is intuitively attractive.

Consider again the problem of medical cancer diagnosis. Here it is necessary to recognize malignant cells. Let us take as thesis that the process of generating unidentified malignant/benign pictures of cells from both cancer patients and healthy persons is substantially cheaper than having an expert determine whether a given cell is malignant: say that one needs $100/picture for an expert to identify the cells but only $10/picture for a technician who takes the pictures *. Ideally, one would want to take say 100 pictures and have the expert label only 10 of these as being malignant or not and then feed the whole information to a computer and with some clever algorithm, learn the classification to within a small error; this is preferred, costwise, over taking 25 pictures and having an expert label all of them †. As it stands today, the practitioner must resort to a variety of heuristics and knowledge from past experience in order to decide how many labeled and unlabeled examples need to be procured to obtain a good classification rule. As another example, consider the task of classifying trees in a forest by their names. Unlabeled examples are free as there are practically an endless number of trees that one can examine. The human expert charges by the hour for supplying the names of trees.

Let us denote explicitly by $P_{error}(m,n)$ the objective error (for a fixed algorithm) given a labeled sample of size m and an unlabeled sample of size n. Our interest here is to present a theoretical analysis that provides an insight into the tradeoff between the finite unlabeled and labeled sample sizes needed to learn (i.e., determine $P_{error}(m,n)$). The question may be succinctly put as follows: *How many unlabeled examples is one labeled example worth?*

In this thesis we present an approach which answers this question for some classification problems under different scenarios. Each scenario depends on the additional

*These figures may improve pending the fate of the new health care plan !

†Of course, what one really wants to do is to minimize the Bayes risk for appropriate loss functions

side information that is given to the learner, for instance the parametric form of the underlying mixture density, and also on the particular algorithm which approaches the estimation of the decision rule by a specific technique.

1.3 Organization

In Chapter 2 we present some additional motivation for being interested in the trade-off between the number of labeled and unlabeled examples for learning a classification rule. We refer to several established results that touch upon this area in the limiting sense—with infinitely many unlabeled examples and just one labeled example a decision rule having $P_{error} = 2P_{Bayes}(1 - P_{Bayes})$ can be achieved. Roughly speaking this means that the first labeled example contains one half of the classifying information (cf. Cover & Castelli [5]). Still with an infinity of unlabeled examples, as we increase the number, m, of labeled examples the classification error goes arbitrarily close to P_{Bayes} and exponentially fast with m. The question remains as to how fast does the error decrease with respect to increasing the unlabeled sample size. In Chapters 4, 5, 6 we determine the rates under different scenarios. We then describe our approach for learning with a mixed sample, which follows the Probably Approximately Correct (PAC) model of learning with examples. With the technical background on which PAC is based we can analyze learning with a mixed sample under different scenarios of side information given to the learner. We explore two such scenarios— a parametric, based on the MLE principle and a nonparametric one based on Kernel Density Estimation.

In Chapter 3 we present the necessary technical background in the form of established theorems, definitions and examples. The main technical results needed are various uniform strong laws of large numbers over classes of functions which arise from the pioneering work of V.N. Vapnik and A. Ya. Chervonenkis. The results themselves constitute powerful generalizations of the earliest and best known uniform strong law—the Glivenko-Cantelli theorem. The basic form of the uniform strong law that will be invoked asserts that, for various general function classes, the averages of functions (evaluated from a random sample) converge *uniformly* and *exponentially fast* to the corresponding expectations, this rates being governed by various covering numbers and a combinatorial parameter known as the Vapnik-Chervonenkis dimension.

The main contributions of the thesis are contained in Chapters 4, 5 and 6. In order to compare the tradeoff between labeled and unlabeled examples under the parametric and nonparametric scenario, we focus on learning the same problem, namely two N-dimensional Gaussian distributed pattern classes, but with the learner having different amounts of side information. Our overall approach is to obtain sample complexities, i.e., the sufficient number m of labeled examples and the number n of unlabeled examples for learning to classify under a prespecified accuracy ϵ and confidence $1 - \delta$. With the theory of Chapter 3 we obtain finite bounds on the sample complexities which enables us to quantify the tradeoff between m and n.

In Chapter 4 we investigate the parametric scenario where the learner knows the form of the underlying probability densities. We present theorems which state the finite sample complexities for two parametric algorithms, E and M. Algorithm E utilizes a purely labeled sample of size m which is polynomial in $\frac{1}{\epsilon}$ and in the dimensionality N. Algorithm M, which is based on maximizing the likelihood function, utilizes a mixed sample. The unlabeled examples are used for estimating the decision border, and the labeled examples are used to determine the optimal labeling of the partition. The proof of the sample complexities for algorithm M is intricate. We provide a preview of the proof and explain the basic concepts underlying it. As expected, algorithm M requires fewer labeled examples than algorithm E on account of using the unlabeled sample. We take the ratio of the number of unlabeled examples over this reduction as a representative of the value of one labeled example when learning under this parametric scenario. This is shown (in Chapter 6) to be polynomial in N, $\frac{1}{\epsilon}$, and $\log\frac{1}{\delta}$. We then investigate the sample complexities under the parametric scenarios of equal and different *a priori* class probabilities.

Chapter 5 focuses on the nonparametric scenario. The learner has no knowledge about the form of the underlying class densities and resorts to extracting all of the information solely from the n unlabeled examples and m labeled examples. Our approach is to use the modes of the Gaussian mixture to determine the Bayes decision border. Our algorithm utilizes Kernel Density Estimation to estimate the unknown mixture density f by $f_n(x)$. Using $f_n(x)$, it then constructs estimates of the modes of f which are shown to be consistent whence the decision rule can have P_{error} arbitrary close to P_{Bayes}. We provide a theorem which states the finite mixed sample complexity for this algorithm. The proof again utilizes the theory of Chapter 3, where the uniform strong law of large numbers plays a principle role in admitting a measure of complexity for this nonparametric approach. We then use the finite sample complexities to establish the tradeoff between m and n. This is shown (in Chapter 6) to be exponential in N, $\frac{1}{\epsilon}$. Requiring no parametric information suggests that the algorithm can be applied to other, non Gaussian, problems where the Bayes decision border can be identified via the modes of the class density mixture. We describe the family of problems for which algorithm K can be used and show that the decision rule is still close to optimal when the sample complexities are as for the Gaussian problem.

Also in Chapter 5, we discuss two other nonparametric approaches to learning classification— the Kohonen LVQ neural network, and the k-means procedure. The Kohonen LVQ neural network utilizes primarily the unlabeled sample to adapt a fixed number of vectors, called neurons, according to a sequential learning rule which performs gradient descent in the space of mean square error. In different fields, e.g., vector quantization, pattern classification, speech recognition, etc. practitioners have reported successful results using this type of learning rule. In our experimental investigation we focus on the tradeoff between the number of labeled and unlabeled examples that are necessary to achieve a specified classification accuracy. We consider experiments that aim to reduce the labeled sample complexity, m, by adding a second layer of neurons. This may be useful in situations where a labeled example is costly and unlabeled examples are abundant. We report on the family of classification problems for which a significant reduction in m w.r.t. the dimensionality N is evident.

The *ad hoc* clustering procedure known as the k-means method is based on a voronoi partition with center vectors y_i that adapt in a way to minimize the empirical mean square error (MSE) based on the randomly drawn unlabeled sample. The true MSE is defined as $\mathbf{E}\min_{1\leq i\leq k}|x-y_i|^2$ which measures the discrepancy between the input x and the output y where y is one of the k vectors y_i. In some problems the minimum MSE partitions achieve classification rates that are optimal or close to optimal when labeled correctly. We consider such a learning problem and using the uniform convergence laws of Chapter 3 we obtain the sufficient mixed sample complexities.

In Chapter 6 we accumulate the results of previous chapters and report the tradeoff between the unlabeled and labeled examples for learning a classification rule under the different scenarios. We discuss another possible approach based on algorithm K of Chapter 5 which uses fewer unlabeled examples by estimating the mixture density f only at a finite number of points. Here however the learner needs more side information than in algorithm K as the knowledge of a compact region which contains the modes of f is a necessary condition. We then discuss our ongoing work and a conjecture about the sample complexity for the Gaussian mixture problem with non-unit covariances. We also briefly discuss several extensions to learning with different types of examples and issues relating to side information.

The following table lists the notation that will be used.

$\mathbf{E}$	expectation
$\mathbf{P}$	probability measure
ϵ	positive accuracy parameter
$1-\delta$	positive confidence parameter
θ_0	the true unknown parameter vector
$B(\theta_0,\epsilon)$	a ball of radius ϵ centered at θ_0
$\partial B(\theta_0,\epsilon)$	the surface of the ball of radius ϵ centered at θ_0
$L(\theta)$	likelihood function based on the sample $(x_1,\ldots,x_n)$ evaluated at θ
θ_ϵ	point on the surface of the ball $B(\theta_0,\epsilon)$
$\lvert x\rvert = \sqrt{\sum_{i=1}^N x_i^2}$	Euclidean-norm of N-dimensional vector $x = [x_1, x_2, \ldots, x_N]$
$\mathrm{VC}(\mathcal{H})$	VC-dimension of class $\mathcal{H}$
d	VC-dimension
N	dimension of example-space
n	number of unlabeled examples
m	number of labeled examples
$\mathcal{N}(\epsilon,\mathcal{H},L_Q^1)$	covering number for class $\mathcal{H}$ under L^1-norm with probability measure Q
Θ	parameter space (compact set in Euclidean space)
$poly^r(x)$	r^{th}-degree polynomial in x
1_A	indicator function of the set A
$c_1, c_2, c_3, \ldots$	positive constants

Table 1.1: Notations

Chapter 2

The Problem

In the preceding section we argued that a mixed-sample approach to learning a classification rule has high appeal in practice. It is of interest to try to quantitatively describe the tradeoff in unlabeled versus labeled sample sizes required for learnability. To further our intuition, let us deduce some simple limiting results when either the number of labeled examples $m \to \infty$ or when the number of unlabeled examples $n \to \infty$.

Begin with the following observation: With an unlimited supply of independently drawn examples we can estimate any probability distribution function arbitrarily well. This is based on an extension of the Glivenko-Cantelli theorem, the oldest and best known uniform strong law of large numbers (cf. Pollard [21]).

Consider first the case $m = \infty$, $n = 0$, where there are an infinity of labeled examples and no unlabeled examples. An appeal to the Glivenko-Cantelli theorem shows that we can obtain exact estimates of each of the pattern-class probability distributions, and hence deduce the optimal Bayes rule. Consequently

$$P_{error}(\infty, 0) = P_{Bayes}.$$

Unlabeled examples alone are not sufficient to learn a decision rule; with infinitely many unlabeled examples the mixture density $f = p_1 f_1 + p_2 f_2$, can be learned exactly (via the Glivenko-Cantelli theorem) but even if the decision border can be uniquely identified from the mixture we still need labeled examples to associate every decision region with one of the two classes. Indeed suppose there are no labeled examples, $m = 0$, and an infinity of unlabeled examples $n = \infty$, and suppose the best case where the class-conditional densities f_1,f_2 and the priors p_1,p_2, can be extracted from the mixture (this notion of *identifiability* can be made mathematically precise). The Bayes decision border can thus be procured but we are still left with the problem of deciding the label "1" or "2" for the two regions R_1, R_2. With no recourse but random guessing one of the two labelings we hence obtain

$$P_{error}(0,\infty) = \frac{1}{2}P_{Bayes} + \frac{1}{2}(1 - P_{Bayes}) = \frac{1}{2},$$

a result no better than randomly guessing the label of an x in the first place! Nevertheless, it is clear that unlabeled examples *do* carry information, and with a small amount of additional information in the form of labeled examples we should be able to exploit this untapped source of information as we see in the sequel.

Let us now restrict ourselves to an identifiable mixture distribution $f(x)$ (cf. Teicher [29]). In particular, if $f(x) = \pi g(x) + (1-\pi)h(x)$, we can identify the exact value of π and the form of $g(x)$ and $h(x)$ given $f(x)$. Note, however, that it is not known which of $g(x)$ and $h(x)$ is $f_1(x)$ (the other will be $f_2(x)$) and whether $\pi = p_1$ or $\pi = p_2$. Nevertheless, given an infinity of unlabeled examples, the Bayes decision border $\{x : p_1 f_1(x) = p_2 f_2(x)\} = \{x : \pi g(x) = (1-\pi)h(x)\}$ can be identified. This is because the identifiable mixture $f(x)$ is obtained via the Glivenko-Cantelli theorem and the Bayes border is invariant to the labeling of g and h. As before, this decision border optimally partitions the feature space into two disjoint regions R_1 and R_2 and the difficulty is that we do not know which region should be labeled "1" and which "2".

Now suppose we have one labeled example (x, y), i.e., $m = 1$. Denote by E the event that a random x drawn according to $f(x)$ is misclassified. Draw one labeled example by first choosing a class according to the *a priori* probabilities p_1,p_2 and then drawing an x according to the density of the chosen class (which is either $f_1(x)$ or $f_2(x)$). Then label as class y the region that contains x, and the second region by the complement label. Clearly, there are two possible labelings: one has R_1 labeled "1", R_2 labeled "2"; this corresponds to the Bayes optimal labeling (denote it as L_{good}) which has (conditional) error probability $P_{error} = \mathbf{P}(E|L_{good}) = P_{Bayes}$; the other labeling has R_1 labeled "2", and R_2 labeled "1" (denoted by L_{bad}) and its conditional error probability is given by $P_{error} = \mathbf{P}(E|L_{bad}) = 1 - P_{Bayes}$. Any one of these two might be chosen. Consequently, the unconditional error probability of our decision rule is given by

$$\begin{aligned} P_{error}(1,\infty) \equiv \mathbf{P}(E) &= \mathbf{P}(E|L_{bad})\mathbf{P}(L_{bad}) + \mathbf{P}(E|L_{good})\mathbf{P}(L_{good}) \\ &= (1 - P_{Bayes})\mathbf{P}(L_{bad}) + P_{Bayes}\mathbf{P}(L_{good}) \end{aligned}$$

We have

$$\begin{aligned} \mathbf{P}(L_{bad}) &= \mathbf{P}(x \text{ has true label "1" and } x \text{ fell in } R_2) \\ &+ \mathbf{P}(x \text{ has true label "2" and } x \text{ fell in } R_1). \end{aligned}$$

This equals

$$p_1 \int_{R_2} f_1(x)\,dx + p_2 \int_{R_1} f_2(x)\,dx = P_{Bayes}.$$

Clearly $\mathbf{P}(L_{good}) = 1 - P_{Bayes}$ hence the total misclassification probability $\mathbf{P}(E)$ of the resulting classifier is

$$P_{error}(1,\infty) = 2P_{Bayes}(1 - P_{Bayes}) \leq 2P_{Bayes}.$$

Therefore for *any* problem with an identifiable class mixture and for *any* algorithm that produces a decision rule utilizing $n = \infty$ unlabeled examples and $m = 1$ labeled examples the classification performance is no worse than twice the best achievable error performance ! This result was demonstrated by T. M. Cover and V. Castelli [5] who also considered the case when there is more than one labeled example. We tackle this next.

With a few more labeled examples we can rapidly get as close as desired to Bayes performance. Suppose we have m labeled examples and $n = \infty$ unlabeled examples. Use the infinity of unlabeled examples to deduce the Bayes border $\{x : \pi g(x) = (1-\pi)h(x)\}$. Now w.l.o.g. suppose $R_1 = \{x : \pi g(x) > (1-\pi)h(x)\}$, R_2 is the complement of R_1, and $h(x) = f_2(x)$, $g(x) = f_1(x)$. We can determine exactly the quantities $\eta_1 = \mathbf{P}(2|x \in R_1)$, $\eta_2 = \mathbf{P}(1|x \in R_2)$, and $p = \int_{R_1} f(x)\,dx$. The quantities η_1 and η_2 are the probabilities that a randomly drawn test example x is misclassified given it is in R_1 or R_2, respectively. Also, $p = \mathbf{P}(R_1)$ and $1-p = \mathbf{P}(R_2)$. The procedure for labeling the regions is as follows: draw

$$m = \frac{1}{p_{min}\left(1 - 2\sqrt{\eta_{max}(1-\eta_{max}}\right)} \log \frac{3}{\delta}$$

labeled examples, where $p_{min} = \min(p, 1-p)$ and $\eta_{max} = \max(\eta_1, \eta_2)$ and $\delta > 0$ is arbitrarily small. Assign to each region R_1 and R_2, the label of the majority of the examples that fell in it. If no examples fell in R_i then label it "1" with probability $\frac{1}{2}$ and "2" with probability $\frac{1}{2}$. Then the resulting classifier has error probability

$$P_{error}(m, \infty) \leq P_{Bayes}(1-2\delta) + 4\delta.$$

(Cover & Castelli [5] have shown a similar bound.) We now briefly prove this result. Let E denote the event that a random x is misclassified. There are four possible labelings of the regions R_1, R_2, based on the labeled examples: L_{good} has R_1 labeled "1" and R_2 labeled "2"; $L_{bad,1}$ has R_1 labeled "2" and R_2 labeled "1"; $L_{bad,2}$ has R_1 and R_2 labeled "1"; $L_{bad,3}$ has R_1 and R_2 labeled "2". We have

$$\mathbf{P}(E) \quad = \quad \mathbf{P}(E|L_{good})\mathbf{P}(L_{good}) + \mathbf{P}(E|L_{bad,1})\mathbf{P}(L_{bad,1}) + \mathbf{P}(E|L_{bad,2})\mathbf{P}(L_{bad,2})$$

$$+ \quad \mathbf{P}(E|L_{bad,3})\mathbf{P}(L_{bad,3}).$$

We have $\mathbf{P}(E|L_{good}) = P_{Bayes}$ and clearly $P(L_{good}) \leq 1$. Also, $\mathbf{P}(E|L_{bad,1}) = 1-P_{Bayes}$ (see earlier), while $\mathbf{P}(E|L_{bad,2}) = p_2$ and $\mathbf{P}(E|L_{bad,3}) = p_1$.

Now, $\mathbf{P}(L_{bad,2})$ equals the probability that the majority of examples in R_1 are "1" and that the majority of examples in R_2 are "1", or that the majority of examples in R_1 are "1" and none fell in R_2 and "1" was chosen, or that the majority in R_2 are "1" and none fell in R_1 and "1" was chosen. Similarly $\mathbf{P}(L_{bad,3})$ equals the probability that the majority of examples in R_1 are "2" and that the majority of examples in R_2 are "2", or that the majority of examples in R_1 are "2" and none fell in R_2 and "2" was chosen, or that the majority in R_2 are "2" and none fell in R_1 and "2" was chosen. We have the probability that the majority of examples in R_2 are "1" given by

$$= \sum_{k=1}^{m} \binom{m}{k} (1-p)^k (p)^{m-k} \cdot \sum_{j=k/2}^{k} \binom{k}{j} \eta_2^j (1-\eta_2)^{k-j}.$$

Using Chernoff's bound for a binomial distributed random variable we can bound the inner sum by $(4\eta_2(1-\eta_2))^{k/2}$ whence obtain the upper bound

$$e^{-m(1-p)\left(1-2\sqrt{\eta_2(1-\eta_2)}\right)} \leq e^{-mp_{min}\left(1-2\sqrt{\eta_{max}(1-\eta_{max})}\right)}.$$

Similarly, the probability that the majority of examples in R_1 are "2" is given by

$$\sum_{k=1}^{m} \binom{m}{k} p^k (1-p)^{m-k} \cdot \sum_{j=k/2}^{k} \binom{k}{j} \eta_1^j (1-\eta_1)^{k-j} \leq e^{-mp_{min}\left(1-2\sqrt{\eta_{max}(1-\eta_{max})}\right)}.$$

We also have that the probability of the majority in R_1 are "1" and majority in R_2 are "1", is less than the probability that the majority of examples in R_2 are "1". Also, the probability that the majority in R_1 are "2" and the majority in R_2 are "2", is less than the probability that the majority in R_1 are "2". Similarly, the probability that the majority in R_1 are "1" and none fell in R_2 and "1" was chosen, is less than the probability that none fell in R_2 and "1" was chosen, which is bounded above by

$$(1-p)^m \frac{1}{2} \leq \frac{1}{2} e^{-mp} \leq \frac{1}{2} e^{-mp_{min}} \leq \frac{1}{2} e^{-mp_{min}\left(1-2\sqrt{\eta_{max}(1-\eta_{max})}\right)},$$

the last inequality following since $\left(1-2\sqrt{\eta_{max}(1-\eta_{max})}\right) \leq 1$. Define

$$\delta = e^{-mp_{min}\left(1-2\sqrt{\eta_{max}(1-\eta_{max})}\right)}.$$

Using the above we have

$$\mathbf{P}\left(L_{bad,2}\right) \leq \delta + \frac{1}{2}\delta + \frac{1}{2}\delta = 2\delta, \text{ and } \mathbf{P}\left(L_{bad,3}\right) \leq 2\delta.$$

We can similarly obtain that

$$\mathbf{P}\left(L_{bad,1}\right) \leq 2\delta.$$

Thus we have

$$\begin{aligned} \mathbf{P}(E) &\leq P_{Bayes} \cdot 1 + p_2 2\delta + p_1 2\delta + (1-P_{Bayes})2\delta \\ &= P_{Bayes}(1-2\delta) + 4\delta. \end{aligned}$$

The left side is by definition $P_{error}(m, \infty)$. This concludes the proof. As $\delta > 0$ can be arbitrarily small, we conclude that given an infinity of unlabeled examples, as the labeled sample size m increases, the classifier performance approaches P_{Bayes} exponentially fast in m.

Related to the limiting case of $P_{error}(\infty, 0)$ is a classical result of Cover & Hart [33] pertaining to the nearest neighbor (NN) classification rule (a nonparametric method). This classifier is based on a voronoi-partition of the feature space, each voronoi cell placed around one labeled example. They bound the error of the NN-classifier in the ∞-labeled sample limit as

$$P_{Bayes} \leq P_{error,NN}(\infty, 0) \leq 2P_{Bayes}(1-P_{Bayes}).$$

An improvement towards a more realistic case (having finite sample size instead of infinite) has been recently achieved by Psaltis, Snapp & Venkatesh [15], who showed that

$$P_{error,NN}(m,0) = P_{Bayes} + O(m^{-2/N})$$

for the nearest neighbor classifier in N-dimensional feature space.

Finally, for mixed-sample learning, the case of $P_{error}(m,n)$ is the most realistic since both sample sizes are finite. The form of the solution depends strongly on the algorithm used to learn the classification rule. Different methods may utilize different assumptions, or side information, regarding the class distributions and there may be various ways of learning from a mixture of labeled and unlabeled examples. For instance the learning of the decision border may be done solely with unlabeled examples while leaving the labeled sample only for labeling the regions. Another approach would use both labeled and unlabeled examples to learn the border. Our effort in this thesis is dedicated to tackle the $P_{error}(m,n)$ case. In this context, Cover & Castelli have suggested that for identifiable families, $P_{error}(m,n)$ may take the form $O(\frac{1}{n}) + O(e^{-\alpha m})$, i.e., that there is an exponential tradeoff between labeled and unlabeled examples.

Analyzing the size of finite labeled-samples as the basis of learning-complexity is the approach taken in the PAC (probably approximately correct) model of learning theory (Valiant [34], Blumer, Ehrenfeucht, Haussler & Warmuth [13],[14], Haussler [12]) and also in the analysis of the nearest-neighbor classifier of Psaltis, Snapp & Venkatesh [15]. (In contrast, another approach for representing learning-complexity is to get asymptotic ∞-sample limits as for instance, in Cover & Hart [33].) In these approaches, which utilize only labeled examples, the finite sample size may depend on a prespecified required accuracy, probabilistic-confidence of the result, dimensionality of the feature space and possibly more given parameters.

The theory on which PAC learning is based can be used to analyze learning with a mixed sample, i.e., with both unlabeled and labeled examples, and obtain finite bounds on m and n. These estimates can then be used as a measure of quantifying the tradeoff in unlabeled versus labeled examples in learning a classification decision rule. However, as was discussed in the previous section, there are really three forces at play here: the number of labeled examples m, the number of unlabeled examples n, and the amount of side information given *a priori* to the learner (for instance, in terms of assumptions on the class conditional densities). To see the tradeoff between any two of these three variables we need to fix the third. It is not clear how to quantify side information; there are still open issues to tackle here. Our approach will be to compare the tradeoff between labeled and unlabeled sample sizes under several (qualitative) scenarios of side information available to the learner.

Finite sample complexities results are more difficult to derive than asymptotic results, and typically require a case-by-case analysis — a fully general theory is still in abeyance. In this thesis we primarily investigate two scenarios: (1) the tradeoff between m and n, conditioned on the knowledge of the parametric form of the class-conditional densities; the analysis and results are presented for the specific parametric case of a multi-dimensional Gaussian mixture though the technique extends to other parametric families (2) The tradeoff between m and n conditioned on the knowledge that the modes of the mixture determine the Bayes optimal decision border (neither the parametric form of the mixture nor information about whether it is identifiable, are given to the learner). The function classes considered in the latter case are potentially much larger than the former parametric case.

We approached scenario (1) by choosing two parametric estimation methods, moment estimation and maximum likelihood parameter estimation (MLE). The former is easily applied to the case of a purely labeled sample since it involves estimating independently the moments of the two class conditional densities; the latter method can utilize unlabeled examples to estimate the mixture's parameters and labeled examples to choose the good labeling. Chapter 4 presents the analysis and results for this scenario. In scenario (2), we compared learning with only-labeled examples by the parametric moment-estimation method, to learning with a mixed sample with the Kernel Density Estimation method. The kernel based method invoked here can utilize unlabeled examples and requires no knowledge neither about the form of the class mixture nor whether it is identifiable but does utilize prior knowledge that the modes of the mixture $f(x)$ determine the Bayes border for the class under investigation. Results are presented in Chapter 5.

In both scenarios, the sample size of the purely labeled parametric approach represents a lower limit on the necessary number of labeled examples for learning classification when unlabeled examples are unavailable since the sufficient statistics are accessible and hence the method is efficient. One should expect that unlabeled examples are worth something and hence anticipate a reduction in the labeled sample size when learning with a mixed sample approach compared to the purely-labeled approach. As we will see, the relative amounts of side-information available to the learner in the two scenarios determines the tradeoff.

One common denominator between the MLE and the Kernel estimation technique used here, is that they both can be analyzed using the technical machinery of the uniform SLLN (reviewed in Chapter 3). This is also the fundamental principle behind the main branch of the field of computational learning theory. Using this theory, the complexity, or cost of learning general abstract problems, and also practical problems such as classification, regression, can be expressed quantitatively. A primary measure of cost is the number of examples that are sufficient (or even necessary) to learn the problem to within a prespecified accuracy and confidence. In subsequent chapters we base all of our algorithmic complexity measures on the sufficient sample sizes for learning a classification problem.

To compare the tradeoff between both scenarios, we restrict our discussion largely to learning one common classification problem, in which the two classes are distributed by

$$f_i(x) = \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}|x-\mu_i|^2}$$

for $i = 1, 2$, and $x, \mu_1, \mu_2 \in \mathbb{R}^N$, $|x| = \sqrt{x_1^2 + \ldots + x_N^2}$, and the classes have *a priori* probabilities p_1, p_2. In scenario (1) this problem belongs to a parametric family of classification problems since the form of the mixture is known to the learner. In scenario (2) this same problem belongs to a family of nonparametric classification problems where the learner knows very little about this mixture. The tradeoff between unlabeled and labeled examples is significantly different in both scenarios as will be shown in subsequent chapters.

In both scenarios, for the mixed-sample methods, we used unlabeled examples to estimate the mixture density (thereby learn the decision border) while labeled examples were used solely for labeling the decision regions. Had we chosen a different approach which also utilizes the labeled sample for learning the decision border, we might have needed fewer unlabeled examples. So in this respect, our results give an upper bound on the number of unlabeled examples required in a trade for every labeled example, when conditioned on fixed side information.

For scenario (1) we also considered the case where the class *a priori* probabilities are different and obtained the unlabeled versus labeled examples tradeoff. This is presented in Chapter 4. In Chapter 5 we investigated two additional nonparametric algorithms which use a mixed sample. We present computer simulation for a LVQ neural network classifier, and theoretical analysis for a related algorithm called k-means.

Before presenting our work, we first review several established theorems and examples of the theory which we use in succeeding sections.

Chapter 3

Technical Results

3.1 The Strong Law of Large Numbers

In this chapter we present several needed technical results to be used in calculating the finite values n and m which are sufficient for the two density estimation methods—maximum likelihood estimation (MLE) and kernel density estimation. Both the MLE and the kernel methods involve learning with randomly drawn unlabeled examples. It is possible to represent each as a problem of approximating the expectations of functions, h, in some large class $\mathcal{H}$ by the empirical means, $\frac{1}{n}\sum_{i=1}^{n} h(x_i)$ where $x_1, \ldots, x_n$ denotes a random sample. We would like to assert that the empirical means $\frac{1}{n}\sum_{i=1}^{n} h(x_i)$ converge *uniformly* over the class $\mathcal{H}$ to the corresponding expectations $\mathbf{E}h(x)$. This will be achieved by the principle of the uniform SLLN which is hence at the heart of both the MLE and kernel methods (as well as many other methods).

The SLLN is one of the fundamental laws in statistics and it arises whenever an empirical procedure which involves randomly drawn observations is believed to be governed by the laws of probability. In practice, we can only run empirical methods, e.g., MLE and kernel. The SLLN assists in bridging the gap between inference based on empirical measurements and that based on probability. We now present some fundamentals of the theory of uniform SLLN convergence for empirical means of functions to their expectations (these results were pioneered by Vapnik & Chervonenkis in [16], [17]).

The classical SLLN of Kolmogorov shows that we have arbitrarily small deviations between the empirical and true mean of a function h, with probability 1.

Theorem 3.1 (SLLN) *If x_i, $1 \leq i \leq n$, is an i.i.d. sequence of random variables with finite expectation $\mathbf{E}|x| < \infty$, then*

$$\frac{1}{n}\sum_{i=1}^{n} x_i \rightarrow \mathbf{E}x \ a.e.$$

Consequently for any measurable function $h(x)$ with $\mathbf{E}|h(x)| < \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} h(x_i) \rightarrow \mathbf{E}h(x) \text{ a.e.}$$

This guarantees *a.e.* convergence for any single function $h \in \mathcal{H}$ and consequently jointly for any *finite* collection of functions. When the class $\mathcal{H}$ of functions of interest is infinite, however, the strong law by itself will not suffice to guarantee uniform convergence over the whole class. There is a stronger version of this notion, however, called *uniform* SLLN which is the only convergence concept needed for our purposes and can be found in Pollard [21]. For the MLE and kernel methods it is necessary in Sections 4.3 and 5.2 to have convergence for a *whole* uncountable class $\mathcal{H}$ of functions. In technical terms the uniform SLLN is expressed as

$$\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n} h(x_i) - \mathbf{E}h\right| \rightarrow 0 \text{ a.e.}$$

Such uniform convergence cannot hold over all classes of functions — it is easy to construct instances of classes for which such convergence fails. The whole game is hence to identify classes $\mathcal{H}$ (as generally as possible) for which we can assert such uniform convergence. To facilitate the understanding of how such uniform convergence over a class of functions is proved and to introduce the basis for the finite-sample-size results that will be exhibited subsequently. we choose to express this notion with the following statement

$$\mathbf{P}\left(\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n}h(x_i)\ -\ \mathbf{E}h(x)\right|\ >\epsilon\right)\leq\ \delta \tag{3.1}$$

for any $\epsilon, \delta > 0$ and for all large enough n (which may be a function of δ and ϵ). (Note, since we may write $\delta = \delta(n, \epsilon)$, then if $\sum_{n=1}^{\infty}\delta(n,\epsilon)$ converges, then by the Borel-Cantelli Lemma, (cf. Chung [19]) the above definition of uniform *a.e.* convergence will follow; in the theorems to follow, this will be the case because $\delta(n,\epsilon) = e^{-\Omega(n)}$.)

If the class $\mathcal{H}$ is finite with cardinality K, then the regular SLLN can guarantee this convergence for every $h \in \mathcal{H}$ since then the sup becomes a max; in particular, a single application of Boole's inequality (the union bound) gives

$$\mathbf{P}\left(\max_{1\leq j\leq K}\ \left|\frac{1}{n}\sum_{i=1}^{n}h_j(x_i)\ -\ \mathbf{E}h(x)\right|>\epsilon\right)\ \leq\ K\,\delta$$

which can be made arbitrary small since $\delta > 0$ is arbitrary. However, if the class $\mathcal{H}$ has infinite cardinality (as is the case of the MLE and kernel methods), then the regular SLLN in conjunction with Boole's inequality do not suffice to ensure uniform convergence. It is necessary in such a case to approximate the class $\mathcal{H}$ by a finite collection of functions $\{h_1, h_2, \ldots, h_{cov(\mathcal{H})}\}$, called a finite *covering for* $\mathcal{H}$, such that every $h \in \mathcal{H}$ is "close" (in some metric-sense) to at least one function h_j in the covering. The integer $cov(\mathcal{H})$ is the minimum cardinality of such a covering and is called the *covering number* of $\mathcal{H}$. Then applying the SLLN uniformly over h_j, $1 \leq j \leq cov(\mathcal{H})$, (together with some technical details), results in uniform convergence for the whole class $\mathcal{H}$. In the MLE method we explicitly determine the covering number. In some situations, such as in the kernel method, it is easier to use a bound for the covering number. The covering number $cov(\mathcal{H})$ is bounded by a polynomial whose degree is the celebrated quantity called the Vapnik-Chervonenkis (VC) dimension denoted as VC($\mathcal{H}$). The VC-dimension VC($\mathcal{H}$) influences the upper bound (3.1) on the probability of ϵ-deviation from the mean and hence is an important quantity when determining the sufficient value of n for which this probability is at most δ. We now proceed with several definitions and theorems relating the VC-dimension to the covering number and apply them to get sample sizes that are sufficient for uniform SLLN convergence.

3.2 Uniform Strong Laws

The following theorems and definitions can be found in Pollard [21], Dudley [23], and Haussler [12]. Let X denote some universal set.

Definition 3.2 *Given a class $\mathcal{C}$ of sets in X, and a set $S \subset X$, denote by $\prod_{\mathcal{C}}(S)$ the set of all subsets of S that can be obtained by intersecting S with a set in $\mathcal{C}$, that is, $\prod_{\mathcal{C}}(S) = \{S \bigcap c \, : \, c \in \mathcal{C}\}$. The VC-dimension of $\mathcal{C}$, denoted by VC($\mathcal{C}$) is defined as the cardinality of the largest set $S \subset X$ such that $|\prod_{\mathcal{C}}(S)| = 2^{|S|}$. (Define VC(C) $= \infty$ if the property holds for S unboundedly large.)*

In words, the VC-dimension of $\mathcal{C}$ is the largest cardinality of a set S of points, all of whose subsets can be obtained by intersecting S with sets in $\mathcal{C}$.

EXAMPLE: Let $\mathcal{C}$ be the class of all finite intervals on the real line. When $|S| = 1$ then $|\prod_{\mathcal{C}}(S)| = 2$. When $|S| = 2$ it is 4. When $|S| > 2$, $|\prod_{\mathcal{C}}(S)| < 2^{|S|}$. Hence, VC($\mathcal{C}$) $=$ 2.

EXAMPLE: Let $\mathcal{C}$ be the class of all two-fold unions of intervals on the real line. W.l.o.g. take a set S of points $x_1 < x_2 < x_3 < x_4$, i.e. $|S| = 4$. From the previous example, it is clear that when $|S| = 4$ then $|\prod_{C}(S)| = 16$ because we can find intervals that achieve (by intersection with S) all 4 possible subsets of $\{x_1, x_2\}$ and intervals that achieve all 4 possible subsets of $\{x_3, x_4\}$. Taking these intervals in pairs gives us the 16 possible subsets of $\{x_1, x_2, x_3, x_4\}$. So VC($\mathcal{C}$) ≥ 4. Now try any set S with 5 points $\{x_1, x_2, x_3, x_4, x_5\}$ with $x_1 < x_2 < x_3 < x_4 < x_5$. There do not exist any pairs of intervals that can achieve the subset $\{\{x_1, x_3, x_5\}, \{x_2, x_4\}\}$. Thus VC($\mathcal{C}$) $= 4$.

EXAMPLE: A neuron (or a linear threshold element) with N inputs may be represented by an N-dimensional hyperplane. The number of dichotomies of a set of m points that such a hyperplane can separate is given by the quantity

$$2\sum_{j=0}^{N} \binom{m-1}{j}$$

which equals 2^m if and only if $m \leq N+1$. This is the celebrated result of L. Schlafli [46]. Hence the VC dimension of the class of all neurons with N inputs is equal to N.

Theorem 3.3 *Given any set S of cardinality $m \geq 0$ and a class $\mathcal{C}$ with VC($\mathcal{C}$) $=$ $d \geq 0$, then $\prod_{\mathcal{C}}(S) \leq \sum_{j=0}^{d} \binom{m}{j}$ if $m \geq d$ and $\prod_{\mathcal{C}}(S) = 2^m$ otherwise.*

For our purpose, we will use the fact that for $m \geq 2$ and $d \geq 2$, the sum $\sum_{j=0}^{d} \binom{m}{j} \leq m^d$ so that $\prod_{\mathcal{C}}(S) \leq m^d$, in consequence.

Definition 3.4 *The graph of a real-valued function $f(x)$ on a set X is defined as the subset $G_f = \{(x,y) : 0 \leq y \leq f(x) \text{ or } f(x) \leq y \leq 0\}$ of $X \times \mathbb{R}$.*

A figure of a graph of a function is displayed in Figure 3.1.

Definition 3.5 *The VC-dimension of a class $\mathcal{H}$ of real-valued functions $h(x)$ on X is the VC-dimension of the class of sets that are graphs of the functions in $\mathcal{H}$.*

Theorem 3.6 *Let the class $\mathcal{H}$ be a d-dimensional vector space of real valued functions from X to $\mathbb{R}^N$, i.e., the functions $h(x)$ are linear combinations of some basis set $\{\phi_1(x), \phi_2(x), \ldots, \phi_d(x)\}$. Then the class of sets of the form $\{x \in X : h(x) \geq 0\}$ has VC-dimension $= d$.*

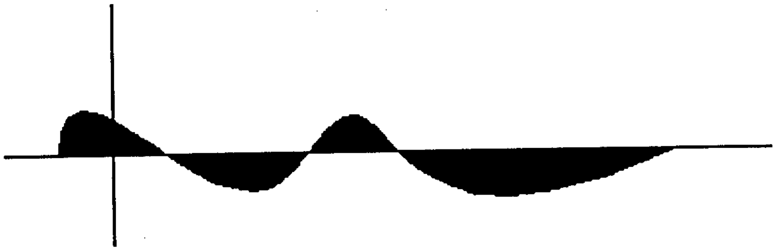


Figure 3.1:

EXAMPLE: Let $\mathcal{H}$ be all functions $h(x)$ of the form $h(x) = a_1 \phi_1 + a_2 \phi_2 + a_3 \phi_3$ where $\{\phi_1, \phi_2, \phi_3\} = \{ 1, x, x^2 \}$. Then VC($\mathcal{H}$) = 3.

EXAMPLE: Let $\mathcal{H}$ be a class of sets of the form $\{x : |x-\theta| \leq 1, x, \theta \in \mathrm{I\!R}^N\}$. Then we can express such sets by $\{x : g_\theta(x) \geq 0\}$ where $g_\theta(x) = -\sum_{i=1}^N x_i^2 - |\theta|^2 + 2\sum_{i=1}^N x_i\theta_i + 1$. Clearly g is a linear combination of the basis $\{1, x_1, x_1^2, \ldots, x_N, x_N^2\}$. Hence VC($\mathcal{H}$) $= 2N + 1$.

Definition 3.7 (Covering number) *Let Q be a probability measure on X and let $\mathcal{H}$ be a class of functions in $\mathcal{L}^1(Q)$, i.e., $\mathbf{E}_Q(|h|) < \infty$. For each $\epsilon > 0$, the covering number $\mathcal{N}(\epsilon, \mathcal{H}, L_Q^1)$ is defined as the smallest value of k for which there exist functions $g_1, g_2, \ldots, g_k$ (not necessarily in $\mathcal{H}$) such that $\min_j \mathbf{E}_Q|h - g_j| < \epsilon$ for each $h \in \mathcal{H}$. If no such k exists, then $\mathcal{N}(\epsilon, \mathcal{H}, L_Q^1) = \infty$.*

As mentioned earlier, uniform convergence is achieved by first approximating a class of functions by a finite covering. This introduces the covering number into the bounds on the deviation-probability. With the next theorem (from Pollard [21] and Haussler [12]) it is possible to replace the covering number in these bounds by a quantity that involves the VC-dimension (which may sometimes be easier to calculate). This was done to obtain Theorem 3.9 and Theorem 3.10. The definition of *permissibility* can be found in Pollard [21], and is a regularity condition guarding against some possible measurability difficulties; basically, if a class of functions can be shown to be indexed by some parameter that lives in a compact metric space, then it is possible to exhibit a finite covering for this class. In our applications, i.e., maximum-likelihood estimation and kernel-density estimation, we explicitly show the existence of a finite covering for the particular function class that is used.

Theorem 3.8 *Let $\mathcal{H}$ be a permissible class of functions from X to the interval [0,M] with VC($\mathcal{H}$) $= d$ for some $1 \leq d < \infty$. Let Q be any probability measure on X. Then for all $0 < \epsilon \leq M$,*

$$\mathcal{N}(\epsilon, \mathcal{H}, L^1_Q) \leq 2\left(\frac{2eM}{\epsilon} \log \frac{2eM}{\epsilon}\right)^d.$$

The following is Corollary 2 in Haussler [12].

Theorem 3.9 (the 1^{st} uniform convergence) *Let $\mathcal{H}$ be a permissible class of functions from X into a bounded interval [0,M] with VC($\mathcal{H}$) $= d$ for some $1 \leq d < \infty$. Assume $n \geq 1$, and draw a random n-sample independently according to any distribution Q on X. Then, for all $0 < \epsilon \leq M$,*

$$\begin{aligned}\mathbf{P}\left(\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n} h(x_i) - \mathbf{E}h(x)\right| > \epsilon\right) &\leq 4\mathcal{N}\left(\frac{\epsilon}{16}, \mathcal{H}, L^1_Q\right) e^{-\epsilon^2 n/64M^2} \\ &\leq 8\left(\frac{32eM}{\epsilon} \log \frac{32eM}{\epsilon}\right)^d e^{-\epsilon^2 n/64M^2}.\end{aligned}$$

where $\mathbf{P}$ denotes probability measure corresponding to independent sampling according to Q. Moreover, for

$$n \geq \frac{64M^2}{\epsilon^2}\left(2d\log\frac{16eM}{\epsilon} + \log\frac{8}{\delta}\right)$$

this probability is at most δ.

The following is based on Theorem 37 in Pollard [21]. The idea is that even if a class $\mathcal{H}$ depends on the sample size n (as will be the case when we deal with kernel density estimation) then it is still possible to have uniform convergence for the empirical means of functions $h \in \mathcal{H}$ to their true means. In Pollard [21] the condition is for functions in the class to have magnitude bounded by 1. We stated the result here under the condition $|h| \leq M$ for every $h \in \mathcal{H}$. Note that the result is distribution-free.

Theorem 3.10 (the 2^{nd} uniform convergence) *For each n, let $\mathcal{H}_n$ be a permissible class of functions whose covering number is bounded as in Theorem 3.8 and the constants M and d do not depend on n. Suppose $x_1, \ldots, x_n$ are obtained by independent sampling from an arbitrary probability distribution on X. If $|h| \leq M$ and* $\mathbf{E}(h^2) \leq \delta_n^2$ *for each $h \in \mathcal{H}_n$ where δ_n^2 satisfies $\frac{\log n/n}{\delta_n^2} \to 0$, then*

$$\mathbf{P}\left(\sup_{h \in \mathcal{H}_n} \left| \frac{1}{n} \sum_{i=1}^{n} h(x_i) - \mathbf{E}h(x) \right| > \epsilon \frac{\delta_n^2}{M}\right) \leq 24 \left(\frac{32eM^2}{\epsilon\, \delta_n^2} \log \frac{32eM^2}{\epsilon\, \delta_n^2} \right)^d e^{-n\, \epsilon^2 \delta_n^2 / 8192M^2}$$

and the RHS $\to 0$ faster than any power of n.

The fact that the bound goes to 0 faster than any power of n ensures a.e. convergence (via the Borel-Cantelli lemma) therefore achieving *uniform* convergence of empirical means to expectations over the whole class $\mathcal{H}_n$ for *arbitrary* sampling distributions.

Chapter 4

Parametric Scenario

In this chapter we discuss several variants of learning a classification rule with parametric underlying distributions. The overall theme is to determine the sample complexities of learning to classify when the form of the densities involved is known to the learner, and either a mixed or just a labeled sample is available. Following the approach of the Computational Learning Theory field (cf. [34], [12]), we measure the sample complexity of learning by the number of examples that are sufficient to achieve an accuracy $\epsilon > 0$ in learning the decision rule, with a certain level of confidence in excess of say $> 1 - \delta$. Hence all the statements of learnability that we make are probabilistic in nature, where the confidence parameter $\delta > 0$ can be arbitrarily chosen.

The investigated scenarios are limited to multi-dimensional Gaussian distributed pattern classes with unit covariances and the theorems are stated for this family of problems. It will be quite clear, however, that the analysis techniques pertains to other parametric families as well, albeit resulting in different constants and rates.

We start in Section 4.1 by determining the sample complexity of learning only with a labeled sample, where the classification problem has two equiprobable pattern classes. The learner uses algorithm E, based on moment estimation, to construct a decision rule. Tight bounds on the deviation of the moment estimates from the true values yield a tight sample complexity bound. This learning scenario represents a state in which the learner utilizes the labeled sample efficiently and has access to the sufficient statistics for estimating the class conditional densities. The cost of learning under such a scenario is therefore representative of the minimal cost in terms of exploitation of information under this scenario. This hence provides a good reference point for interpreting the cost of learning with a mixed sample. In particular, we determine the reduction in the labeled sample size due to introducing unlabeled examples. This establishes the tradeoff between labeled and unlabeled examples when the parametric form of the densities is available as side information.

In Section 4.2 the case of a mixed sample is considered. The problem is the same, i.e., two equiprobable pattern classes each distributed as a unit-covariance multi-dimensional Gaussian. The learner is given randomly drawn unlabeled and labeled examples and uses algorithm M, which is based on maximum likelihood estimation, to construct estimates of the two class conditional means using only unlabeled examples. These are then used to construct a linear decision border which approximates the Bayes partition. The labeled examples are used in this approach only for labeling the two regions of the hyperplane. As expected, the mixed sample approach uses fewer labeled examples. The reduction, compared to the purely labeled sample approach, is significant, being polynomial in the dimensionality of the feature space and in the accuracy and confidence parameters.

We will proceed as follows: in Section 4.1 we state Theorem 4.1 which pertains to learning with a purely-labeled sample, then preview the proof before providing the actual proof. In Section 4.2 we state Theorem 4.2 which pertains to the mixed sample learning, followed by a preview of the proof. The proof is given in Section 4.3. The referenced auxiliary lemmas are included in the proof. In Sections 4.4, 4.5 we analyze the same classification problem as the previous sections, except the two classes have

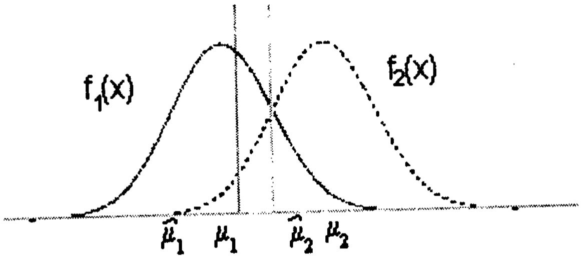


Figure 4.1:

different *a priori* probabilities. Discussions of the results are deferred to Chapter 6.

4.1 Purely Labeled Sample

Here the parametric form of the N-dimensional class conditional densities $f_1(x)$, $f_2(x)$, is known:

$$f_i(x) = \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}|x-\mu_{0i}|^2} \qquad i = 1, 2.$$

We denote this by writing $f_1(x) = g(x|\mu_{01})$ and $f_2(x) = g(x|\mu_{02})$. The only unknowns are the two mean vectors, μ_{01} and μ_{02}. The Bayesian decision border is a linear hyperplane orthogonal to $\mu_{02} - \mu_{01}$ (see Figure 4.1). A learner is given only labeled examples drawn according to the mixture $f(x) = \frac{1}{2}g(x|\mu_{01}) + \frac{1}{2}g(x|\mu_{02})$. Algorithm E, based on moment estimation, is used to determine close estimates of the means, with which a classifier is constructed.

We first state the algorithm then we state the theorem, provide a preview of the proof followed by the proof itself.

Algorithm E:

The setting: Two pattern classes with underlying Gaussian mixture density

$$f(x) = \frac{1}{2}g(x|\mu_{01}) + \frac{1}{2}g(x|\mu_{02}).$$

The teacher draws labeled examples according to $f(x)$ by choosing class "1" or class "2" with probability $\frac{1}{2}$ and then drawing according to the selected class conditional density $g(x|\mu_{0i})$, $i = 1, 2$.

Given: m_1 examples labeled as "1" and m_2 examples labeled as "2", where $m_1 + m_2 = m$.

Begin:

1) Let the mean estimates of μ_{0i}, $i = 1, 2$, be

$$\hat{\mu}_i \equiv \frac{1}{m_i} \sum_{k=1}^{m_i} x_k^i \qquad (i = 1, 2).$$

where we denote by x_k^i the k^{th} element of the i^{th} example x_i.

2) Let the decision border be the hyperplane that passes through the point $\frac{\hat{\mu}_1+\hat{\mu}_2}{2}$ and orthogonal to the vector $\hat{\mu}_2 - \hat{\mu}_1$.

3) Label the two decision regions across the hyperplane by the subscript of the mode estimate, $\hat{\mu}_i$, $i = 1, 2$, on that side, respectively.

End.

Theorem 4.1 *Suppose we are given two equiprobable classes which are distributed according to Gaussian probability densities $g(x|\mu_{01})$ and $g(x|\mu_{02})$, with means $\mu_{0i} \in \mathbb{R}^N$, $i = 1, 2$, and unit covariances. For small $\epsilon > 0$, arbitrary $\delta > 0$, given*

$$m = \frac{4N}{\epsilon} \log\left(\frac{8N}{\delta}\right)$$

labeled examples and $n = 0$ unlabeled examples, algorithm E results in a decision rule with a classification error

$$P_{Bayes} \leq P_{error}(m, 0) \leq P_{Bayes}(1 + c\epsilon)$$

with confidence at least $1-\delta$, where $c > 0$ is a constant depending only on the distance between the means.

We first preview the proof. The aim is to estimate the Bayes decision border (which is a hyperplane orthogonal to $\mu_{01} - \mu_{02}$) by a hyperplane that is orthogonal to the difference of the two sample averages, $\hat{\mu}_1 - \hat{\mu}_2$ and which passes through their midpoint.

The teacher draws labeled examples from the mixture $f(x)$. We first establish the sufficient sample size, m_1, of examples from class "1" for which $\hat{\mu}_1$ will be ϵ-close to μ_{01}. The estimate m_2 will be identical. Then we find an upper bound on m such that the number of "1"-labeled examples is at least m_1 and the number of "2"-labeled examples is at least m_2 with high confidence.

To obtain an exponentially small bound (w.r.t. sample size) on the deviation between each mean and its corresponding sample average, we utilize the Chernoff bound (cf. Papoulis [43]), which is a variant of Chebyshev's inequality. This bound uses the moment generating function and hence can be easily specialized to a normal random variable. It gives a high confidence for both sample averages to be ϵ-close to their respective means with m as above. Then we analyze the classification error of the resulting decision rule by finding the worst-case (error-wise) deviation of the hyperplane from the Bayes optimal hyperplane. As a consequence of having a linear decision boundary (hyperplane) the discriminant function which represents the decision region becomes a univariate-normally distributed random variable. This leads directly to a bound on P_{error} which depends on the above ϵ and hence is valid with the above confidence, given that m is as stated in the theorem. Note that the constants in the theorem are not the best possible and can doubtless be improved.

PROOF:

The aim is to show that $\mathbf{P}(\{|\hat{\mu}_1 - \mu_1| > \epsilon\} \cup \{|\hat{\mu}_2 - \mu_2| > \epsilon\})$ is at most δ when $m = m_1 + m_2$ is as specified in the theorem.

First we determine the sufficient sample size m_1 from class "1". We use μ_{1k} to denote the k^{th} element of the vector μ_1, and denote by x_k^i the k^{th} element of the i^{th} example x_i. To have $|\hat{\mu}_1 - \mu_1|^2 > \epsilon^2$ it is necessary that at least one of the N components $(\hat{\mu}_{1k} - \mu_{1k})^2 > \epsilon^2/N$, for $k = 1 \ldots N$. So, as a consequence of Boole's inequality,

$$\mathbf{P}\left(|\hat{\mu}_1 - \mu_1| > \epsilon\right) \leq \sum_{k=1}^{N} \mathbf{P}\left(|\hat{\mu}_{1k} - \mu_{1k}| > \epsilon/\sqrt{N}\right).$$

Now since x is a vector distributed as $\mathrm{N}(\mu_1, \mathrm{I})$, then each component x_k is distributed as $\mathrm{N}(\mu_{1k}, 1)$ hence the sample mean estimate $\hat{\mu}_{1k} = \frac{1}{m_1}\sum_{i=1}^{m_1} x_k^i$ is normal with mean μ_{1k} and variance $\frac{1}{m_1}$. Now note the useful elementary bound

$$\mathbf{P}(z > A) \leq \inf_{s \geq 0} e^{-sA}\, \mathbf{E}(e^{sz}).$$

Let

$$z = \frac{1}{m_1}\sum_{i=1}^{m_1} x_k^i - \mu_k$$

and $A = \epsilon/\sqrt{N}$. Simple algebra upper bounds $\mathbf{P}(z > A)$ by $e^{-\epsilon^2 m_1/2N}$. And hence

$$\mathbf{P}\left(\left|\frac{1}{m_1}\sum_{i=1}^{m_1} x_k^i - \mu_k\right| > \epsilon/\sqrt{N}\right) \leq 2e^{-\epsilon^2 m_1/2N}.$$

So from above,

$$\mathbf{P}(|\hat{\mu}_1 - \mu_1| > \epsilon) \leq 2\, N\, e^{-\epsilon^2 m_1/2N} \equiv \delta/2.$$

It follows that the sufficient sample size is

$$m_1 \geq (2\, N/\epsilon^2)\log(4\, N/\delta).$$

Repeating the same argument for class "2" we have that

$$m_2 \geq (2\, N/\epsilon^2)\log(4\, N/\delta)$$

is sufficient for

$$\mathbf{P}(|\hat{\mu}_2 - \mu_2| > \epsilon) \leq 2\, N\, e^{-\epsilon^2 m_2/2N} \equiv \delta/2.$$

Now we find the sufficient m so that the number of class "1" labeled examples is at least m_1 and similarly with class "2". Let $\hat{p}$ be the frequency of "1"-labeled examples, i.e., $\hat{p} = \frac{1}{m}\sum_{i=1}^{m} 1_{y_i=\text{``}1\text{''}}$ where y_i is the label of the i^{th} example. We have

$$m = \frac{m_1}{\hat{p}} \leq \frac{m_1}{p - \epsilon_1} = \frac{m_1}{\frac{1}{2} - \epsilon_1}$$

where ϵ_1 is a given constant representing the allowed deviation between the mean and the average of a binomial random variable. For this we have

$$\mathbf{P}\left(|p - \hat{p}| > \epsilon_1\right) \leq 2e^{-2m\epsilon_1^2} \equiv \delta_1$$

hence it suffices to draw

$$m \geq \frac{1}{2\epsilon_1^2} \log \frac{2}{\delta_1}$$

labeled examples to have

$$|p - \hat{p}| \leq \epsilon_1$$

with confidence $> 1 - \delta_1$. Therefore the overall m sufficient for obtaining m_1 "1"-examples with confidence $> 1 - \delta_1$ is

$$m \geq \max\left\{\frac{m_1}{\frac{1}{2} - \epsilon_1}, \frac{1}{2\epsilon_1^2} \log \frac{2}{\delta_1}\right\}$$

Repeating the above argument for "2"-examples (where $m_2 = m_1 = \frac{2N}{\epsilon^2} \log \frac{4N}{\delta}$) and combining we get that the sufficient m for obtaining m_1 "1"-examples and m_2 "2" examples with confidence $> 1 - 2\delta_1$ is

$$m \geq \max\left\{\frac{m_1}{\frac{1}{2} - \epsilon_1}, \frac{1}{2\epsilon_1^2} \log \frac{2}{\delta_1}\right\}.$$

To simplify the bound we select ϵ_1 so that the two terms inside the max are equal. Then substituting for ϵ_1, and replacing both δ and $2\delta_1$ by $\delta/2$ we obtain that with

$$m \geq \frac{4N}{\epsilon^2} \log\left(\frac{8N}{\delta}\right)$$

we have

$$\mathbf{P}\left(\{|\hat{\mu}_1 - \mu_1| > \epsilon\} \cup \{|\hat{\mu}_2 - \mu_2| > \epsilon\}\right) \leq \delta.$$

Now we aim to show that if there is an ϵ-deviation for each of the estimates from the true means then it results in the claimed error. We have two N-dimensional Gaussians, $f_1(x)$ with mean μ_1 and $f_2(x)$ with mean μ_2 and a hyperplane orthogonal to $\hat{\mu}_2 - \hat{\mu}_1$, passing through their midpoint. First translate the Gaussians so that the line between μ_1 and μ_2 is on first-dimension axis and the origin is equidistant from both. Let $\Delta \equiv (|\mu_1 - \mu_2|)/2$ and let $u^+ \equiv [\Delta, 0, \ldots, 0]^T$ and $u^- \equiv [-\Delta, 0, \ldots, 0]^T$. Now consider the decision rule that the hyperplane gives; denote it as $h(x)$. Clearly $h(x) = (\hat{\mu}_2 - \hat{\mu}_1)^T (x - \hat{\mu}_1) - \frac{1}{2} |\hat{\mu}_2 - \hat{\mu}_1|^2$ and the region where $h(x) > 0$ is classified as class "2", i.e., the decision point is at $h = 0$. The vector x is joint Gaussian and so $h(x)$, which is just a linear transformation of x is a one-dimensional Gaussian random variable conditioned on the high probability event that the estimates $\hat{\mu}_i$ are ϵ-close to μ_i. Letting $g(x) \equiv h(x)/|\hat{\mu}_2 - \hat{\mu}_1|$ yields $p_1(g) \sim N(\frac{(\hat{\mu}_2-\hat{\mu}_1)^T u^- + \frac{1}{2}(|\hat{\mu}_1|^2 - |\hat{\mu}_2|^2)}{|\hat{\mu}_2 - \hat{\mu}_1|}, 1)$ and $p_2(g) \sim N(\frac{(\hat{\mu}_2-\hat{\mu}_1)^T u^+ + \frac{1}{2}(|\hat{\mu}_1|^2 - |\hat{\mu}_2|^2)}{|\hat{\mu}_2 - \hat{\mu}_1|}, 1)$. The decision point is at $g = 0$ and it is away from the Bayes border of these two one-dimensional distributions by $\frac{\frac{1}{2}(|\hat{\mu}_1|^2 - |\hat{\mu}_2|^2)}{|\hat{\mu}_2 - \hat{\mu}_1|}$. The configuration of $\hat{\mu}_1$ and $\hat{\mu}_2$ that gives a good upper bound on the probability of error is achieved by minimizing the distance between the means and maximizing the distance from the border to 0. After some algebra we get an upper bound

$$P_{error} \leq \frac{1}{2}\Phi\left(-\frac{\epsilon\Delta}{\Delta - \epsilon} - \frac{\Delta^2}{\sqrt{\epsilon^2 + \Delta^2}}\right) + \frac{1}{2}\Phi\left(\frac{\epsilon\Delta}{\Delta - \epsilon} - \frac{\Delta^2}{\sqrt{\epsilon^2 + \Delta^2}}\right).$$

Approximating this expression for small $\epsilon > 0$ and using the fact that $\Phi(-\Delta) = P_{Bayes}$ we have $P_{error} \leq P_{Bayes} + c_1\epsilon^2 = P_{Bayes}(1 + c_2\epsilon^2)$, for some positive constants c_1, c_2. Replace ϵ^2 by ϵ both in this bound and in the bound for m, to get the claimed statement of the theorem. $\blacksquare$

We now proceed to mixed-sample learning, with side information of the parametric form of the mixture density of the classes.

4.2 Mixed Sample

In this section we consider the problem of classifying two pattern classes with equal *a priori* probabilities each distributed according to a multi-dimensional unit-covariance Gaussian density. Both unlabeled and labeled examples are drawn according to the mixture

$$f(x|\theta_0) = \frac{1}{2}f_1(x|\theta_{01}) + \frac{1}{2}f_2(x|\theta_{02})$$

where f_1, f_2 are Gaussians with means θ_{0i}, $i = 1, 2$, respectively and unit covariance matrix. (Here $\theta_0 = [\theta_{01}, \theta_{02}]$, and we use two functions f_1, f_2 since it will enable us to drop the parameters θ_{0i} for brevity.) The mixture is indexed by the unknown vector $\theta = [\theta_{01}, \theta_{02}]$ in a class of multi-dimensional Gaussian mixtures. This class is identifiable and hence if, using unlabeled examples, we estimate the unknown mixture $f(x|\theta)$ by some other function $f(x|\hat{\theta})$ in this same class, then it will uniquely identify two class conditional densities, $f_1(x|\hat{\theta}_1)$, $f_2(x|\hat{\theta}_2)$, $\hat{\theta} = [\hat{\theta}_1, \hat{\theta}_2]$, whose Bayes decision regions approximate the optimal unknown decision regions. The latter is a hyperplane orthogonal to $\theta_{01} - \theta_{02}$ and passing through their midpoint. If $\theta_{01} = \theta_{02}$ then any decision rule with regions R_1, R_2, in particular any hyperplane going through the point θ_{01}, yields $P_{error} = P_{Bayes} = \frac{1}{2}$. Thus in that case, the fact that the hyperplane cannot be identified is insignificant.

We use algorithm M which is based on Maximum Likelihood Estimation (MLE) with a mixed sample to construct a decision rule which has a P_{error} close to P_{Bayes}. The n unlabeled examples are used to find the point $\hat{\theta}$ which maximizes the likelihood

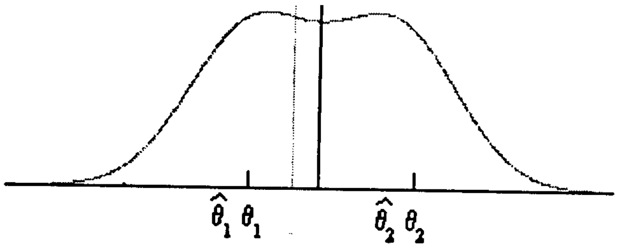


Figure 4.2:

function

$$L(\theta|x_1,\ldots,x_n) \equiv \frac{1}{n}\sum_{i=1}^{n}\log f(x_i|\theta).$$

By taking n sufficiently large, $\hat{\theta}$ is guaranteed with high confidence to be ϵ-close to the unknown θ_0. (The maximum likelihood estimation principle is discussed below). This implies that $\hat{\theta}_i$ is ϵ-close to θ_{0i}, $i = 1,2$ and a hyperplane is constructed as the decision border estimate (see Figure 4.2). The algorithm then uses the m labeled examples with the majority rule, assigning to each of the two regions the label of the majority of the examples that fell in it. The sample size m is taken sufficiently large so that with high confidence the labeling having the minimum error is picked. Because the majority rule results in an exponentially small bound (in m) on the probability of mislabeling the regions, the sample size m is very small, and in particular, is independent of N and ϵ.

The main reason that we chose the MLE for the unsupervised estimation procedure is for its direct coupling with the uniform SLLN principle. As mentioned in Chapter 2, this induces a clear notion of cost, through finite sample complexities, which is what we seek.

We now provide a brief review of the MLE principle.

4.2.1 Maximum Likelihood Estimation: A Review

The method of Maximum Likelihood Estimation (MLE) was first proposed by the German mathematician C.F. Gauss in 1821. However, the approach is usually credited to the English statistician R. A. Fisher who first investigated in 1922 the properties of this method (cf. Bickel & Doksum [39]). The intuition behind this method is based on the following: consider the frequency or density function $f(x|\theta)$ of the random variable X where θ is a parameter vector in a subset Θ of $\mathbb{R}^N$. Given n realizations $x_1, \ldots, x_n$, of X, drawn independently and identically distributed according to $f(x|\theta_0)$, the likelihood function $L(\theta|x^n)$ is defined as

$$L(\theta|x_1 \ldots, x_n) \equiv \frac{1}{n}\sum_{i=1}^{n} \log f(x_i|\theta).$$

(We will sometimes drop the dependence on the sample and write $L(\theta)$.) If the random variable X is discrete then for each θ the likelihood function represents the log of the probability of observing the sample $x_1, \ldots, x_n$. Thus $L(\theta|x_1, \ldots, x_n)$ represents a measure of how likely θ is to have produced the observed sample. The method of MLE aims at finding the parameter value $\hat{\theta} = \text{argsup}_{\theta\in\Theta} L(\theta|x_1, \ldots, x_n)$ which is the most likely to have produced the given sample.

To illustrate this method consider the following example. Let $x_1, \ldots, x_n$ be observations from a Gaussian $N(\mu_0, \sigma_0^2)$, s.t. $\theta_0 \equiv [\mu_0, \sigma_0^2] \in \Theta$ where the parameter space Θ is $-\infty < \mu < \infty$, $0 < \sigma^2 < \infty$. We seek an estimate $\hat{\theta}$ of θ_0. Simple algebra gives

$$L(\theta) = -n\log\sigma - \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}$$

and the maximum likelihood estimator is

$$\hat{\theta} = \left[\bar{x}, s^2\right]$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. By the law of large numbers this ML estimate is consistent, i.e., $\bar{x} \rightarrow \mu_0$ and $s^2 \rightarrow \sigma_0^2$ as $n \rightarrow \infty$. In general, as in this example, the MLE method yields asymptotically optimal parameter estimators, being asymptotically efficient, consistent and function of the minimal sufficient statistic.

However there are some possible difficulties with this method. The first is that $L(\theta)$ may be unbounded over Θ. An example of this (cf. Redner & Walker [7]) is when $f(x|\theta)$ is a mixture of two Gaussians having the same variance σ_0^2 and means μ_{01}, μ_{02}, i.e.

$$f(x|\theta_0) = \frac{1}{2}\left(\frac{1}{\sqrt{2\pi\sigma_0^2}}e^{-\frac{1}{2}\left(\frac{x-\mu_{01}}{\sigma_0}\right)^2} + \frac{1}{\sqrt{2\pi\sigma_0^2}}e^{-\frac{1}{2}\left(\frac{x-\mu_{02}}{\sigma_0}\right)^2}\right).$$

The likelihood function

$$L(\theta|x_1,\ldots,x_n) = \frac{1}{n}\sum_{i=1}^{n}\log\frac{1}{2}\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}(\frac{x_i-\mu_1}{\sigma})^2} + \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}(\frac{x_i-\mu_2}{\sigma})^2}\right)$$

can be unbounded for $\theta = [\mu_1, \mu_2, \sigma^2]$ in the limit as $\sigma \rightarrow 0$ and where one of the means μ_1, μ_2 coincides with an example x_i. Hence if Θ is $-\infty < \mu_i < \infty$, $i = 1,2$, and $0 < \sigma^2 < \infty$, and $\theta_0 \in \Theta$, then $\hat{\theta}$ will not yield a consistent estimate since the $\text{argsup}_{\theta\in\Theta}L(\theta)$ does not tend to θ_0 as $n \rightarrow \infty$. There are some ways to circumvent this difficulty, including various regularization techniques (cf. Grenander Ulf [47]) in which the parameter space Θ is allowed to change with n such that the singular points are contained only in the limit as $n \rightarrow \infty$.

We note that the function $L(\theta)$ may have several relative and global maxima. If the density $f(x|\theta)$ is identifiable, i.e., there do not exist two different parameters θ_a and θ_b which correspond to the same density, then under some weak conditions, it can be shown theoretically that the ML estimate is consistent. But if $f(x|\theta)$ is not identifiable then regardless of how large n is, $L(\theta)$ can possibly attain its maximum value at several different points and thus one is left with no clue as of which of these points should be chosen to be the estimate of the unknown parameter θ_0.

It should be mentioned that even if the density $f(x|\theta)$ is identifiable and $L(\theta)$ is well behaved then still finding the maximum of $L(\theta)$ can be costly in practice and involves some type of a global optimization technique (cf. Redner & Walker [7]).

The MLE method has a robust theoretical basis. There is a vast amount of literature about this method from both experimental and theoretical aspects. From the theoretical side, considerable work exists in proving convergence of the estimator $\hat{\theta}$ to the unknown θ_0 as the sample size $n \to \infty$, see Wald [8], Le Cam [9], Bahadur [10], Huber [11]. The experimental work regarding MLE is concerned with efficient algorithms of finding the global maximum of the likelihood function $L(\theta)$.

On the theoretical aspect of the MLE principle, a brief historic overview shows that initially in 1946, Cramér [35] established the consistency of a θ at which the likelihood function $L(\theta)$ has a relative maximum. This however is not strong enough since there may be several relative maxima and one cannot know which of these critical points to select as the estimator of θ_0. In 1949, Wald [8] established the consistency of the global maximum of $L(\theta)$. This means that the critical point at which $L(\theta)$ achieves its highest maximum, should be chosen as the estimator of θ_0, resulting in a consistent estimate of θ_0. Wald introduced an ingenious method utilizing the extremal properties of the Kullback-Leibler distance function with a uniform SLLN. His method is fundamental in the subject of MLE and it permits remarkable extensions to the case of infinite-dimensional abstract parameter spaces (cf. Grenander U. [47] Chapter 7). The main details of the ML principle will appear in the proof of Theorem 4.2. Much work has been done since then in weakening the conditions of Wald's proof. This includes the work of Huber [11].

4.2.2 Gaussian Mixture

In the previous section we mentioned that the MLE principle yields consistent parameter estimators. Consistency is an asymptotic notion, i.e., a property of the estimator as $n \to \infty$. Our interest is not in asymptotics but in a finite sample size n required for a prespecified estimation error when using the MLE method. Using that the tradeoff between the unlabeled and labeled sample sizes for learning classification can be calculated.

We now describe algorithm M and then proceed with the technical details.

Algorithm M:

`The setting:` Two pattern classes with underlying Gaussian mixture density

$$f(x|\theta_0) = \frac{1}{2} f_1(x|\theta_{01}) + \frac{1}{2} f_2(x|\theta_{02})$$

with $\theta_0 = [\theta_{01}, \theta_{02}]$ is in a compact set Θ of $\mathbb{R}^{2N}$. The teacher draws labeled and unlabeled examples independently according to $f(x|\theta_0)$ by choosing class "1" or class "2" with probability $\frac{1}{2}$ and then drawing according to the selected class conditional density $f_i(x|\theta_{0i})$, $i = 1, 2$.

`Given:` m labeled examples and n unlabeled examples.

`Begin:`

1) Find a point $\hat{\theta} \in \mathbb{R}^{2N}$ satisfying

$$\hat{\theta} = \text{argsup}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \log f(x_i|\theta).$$

2) Select as separating surface the hyperplane that passes through the point $\frac{\hat{\theta}_1 + \hat{\theta}_2}{2}$ and orthogonal to the vector $\hat{\theta}_2 - \hat{\theta}_1$.

3) Label each of the two decision regions separated by the hyperplane by the label of the majority of the labeled examples in the region.

`End.`

We state the following theorem, then preview its proof, followed by the proof itself.

Theorem 4.2 *Suppose the two pattern classes are distributed according to N-dimensional Gaussian probability densities $f_1(x|\theta_{01})$, $f_2(x|\theta_{02})$, both with unit covariance matrices and unknown means θ_{01}, θ_{02} where $\theta_0 \equiv [\theta_{01}, \theta_{02}] \in \Theta$ and Θ is a compact subset of $\mathbb{R}^{2N}$. Then for small $\epsilon > 0$, arbitrary $\delta > 0$, given*

$$n = c_1 \frac{N^2}{\epsilon^3 \delta} \left(N \log \frac{1}{\epsilon} + \log \frac{1}{\delta} \right)$$

unlabeled examples and

$$m \; = \; c_2 \log \frac{1}{\delta}$$

labeled examples, algorithm M determines a decision rule with classification error

$$P_{error}(m,n) \; \leq \; P_{Bayes}(1 + c_3 \epsilon)$$

with confidence at least $1 - \delta$. In the above, $c_1, c_3 > 0$ are constants which depend on θ_0, $c_2 > 0$ depends on P_{Bayes}. All constants may be replaced (with a slight worsening of the bounds) by absolute positive constants.

4.2.3 Preview of the proof of Theorem 4.2

We now outline the proof (for more details see Section 4.3). The proof can be divided into three sections: first, it is shown that with n as above, the maximum likelihood estimator, $\hat{\theta} \equiv [\hat{\theta}_1, \hat{\theta}_2]$, is ϵ-close to θ_0; consequently the hyperplane estimator is close to the Bayes hyperplane. Secondly, assuming that we picked the good labeling (as in Chapter 2, there are only two labelings) we determine the classification error of the resulting regions. This involves the same analysis as in Section 4.1. Thirdly, we determine a sufficient size for m that guarantees with some high confidence that the good labeling is picked by the majority-rule on each of the two regions. As this is a random labeling method, it influences the confidence parameter δ of producing the overall classification rule. This labeling method yields an exponentially small (w.r.t m) upper bound on the probability of choosing the bad labeling and therefore is superior.

The first section of the proof is more involved; it entails classical techniques in the field of probability and statistics. A preview of the first section of the proof is now provided.

It is known that Gaussian mixtures $f(x|\theta)$ are identifiable (see Section 4.3). Using this, the problem of estimating the true unknown distribution function, $f(x|\theta_0)$, becomes one of estimating only the true parameter θ_0 given that the learner has side information about the parametric form of $f(x|\theta)$. In the Gaussian mixture case, this is sufficient for estimating the optimal classifier since θ_0 alone identifies the Bayes border.

Given n random unlabeled examples drawn according to the true unknown mixture density $f(x|\theta_0)$, the aim is to show that any $\hat{\theta}$ that maximizes the function

$$L(\theta) \equiv \frac{1}{n}\sum_{i=1}^{n} \log f(x_i|\theta),$$

is ϵ-close to the unknown θ_0, i.e.

$$\left|\theta_0 - \hat{\theta}\right| \leq \epsilon$$

with high confidence provided n is large enough. The approach, which is based on the original proof of Wald [8] (see also Rao [22]), is to show that $L(\theta) < L(\theta_0)$ for all θ in the parameter space Θ such that $|\theta - \theta_0| > \epsilon$, and simultaneously that there exists a θ_a with $|\theta_a - \theta_0| \leq \epsilon$ such that $L(\theta_a) \geq L(\theta_0)$. So by calculating $\text{argsup}_\Theta L(\theta)$, the learner must obtain $\hat{\theta}$ which is ϵ-close to the true unknown parameter θ_0.

The first step is realizing that the Kullback-Leibler distance between two densities f and g defined by

$$\mathbf{E}_g \log \frac{f}{g}$$

achieves its maximum value of 0 uniquely when $f = g$. Then, because the Gaussian mixture is identifiable, it means that not only is there uniqueness in the the space of all Gaussian mixtures but also in the parameter space, i.e., there does not exist a θ differing from θ_0 at which

$$\Phi_{\theta_0}(\theta) \equiv \mathbf{E} \log \frac{f(x|\theta)}{f(x|\theta_0)}$$

attains the value 0. (In the proof, we will drop the subscript θ_0 and write only $\Phi(\theta)$, since θ_0 is fixed throught, and all expectations are w.r.t. $f(x|\theta_0)$.)

If we can guarantee that the empirical Kullback-Leibler function

$$\Phi_n(\theta) \equiv \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(x_i|\theta)}{f(x_i|\theta_0)} \tag{4.1}$$

which equals

$$L(\theta) - L(\theta_0)$$

is close enough to $\Phi(\theta)$ uniformly for all $\theta \in \Theta$ then it is not difficult to see that $\Phi_n(\theta)$ can be made < 0 for all θ s.t. $|\theta - \theta_0| > \epsilon$ for arbitrarily small $\epsilon > 0$. This in turn implies that $L(\theta) < L(\theta_0)$ for all θ such that $|\theta - \theta_0| > \epsilon$. It is also necessary to show that there exists at least one point θ_a with $L(\theta_a) \geq L(\theta_0)$ with $|\theta_a - \theta_0| \leq \epsilon$. Then it follows that $\text{argsup}_\Theta L(\theta)$ must be ϵ-close to θ_0—the needed result.

These two demands are satisfied with the help of the uniform SLLN (Theorem 3.9). It allows us to guarantee that

$$|\Phi_n(\theta) - \Phi(\theta)| \leq B_{\theta_0} \epsilon^2,$$

for θ s.t. $|\theta - \theta_0| = \epsilon$ where the constant $B_{\theta_0} > 0$ is s.t. $\Phi(\theta) \leq -B_{\theta_0}\epsilon^2$ for such θ and small $\epsilon > 0$. For all such θ we therefore have $L(\theta) < L(\theta_0)$. Using the continuity of $L(\theta)$ this implies that there exists θ_a, ϵ-close to θ_0, which is a maximum (not necessarily the global maximum) of $L(\theta)$. We then again use the same principle to show that

$$|\Phi_n(\theta) - \Phi(\theta)| \leq \alpha(\epsilon) \tag{4.2}$$

which makes $L(\theta) < L(\theta_0)$ for all θ s.t. $|\theta - \theta_0| > \epsilon$, where the deviation $\alpha(\epsilon)$ is within our control.

We can only use the uniform SLLN over a class of uniformly bounded functions (over $x \in \mathbb{R}^N$). The difficulty here is that the function $\log f(x|\theta)$ is unbounded over $x \in \mathbb{R}^N$. This difficulty is finessed by use of a truncation argument restricting the class of functions to which we apply the uniform SLLN to be a class

$$\{g(x|\theta) \equiv f(x|\theta)1_D(x) : \theta \in \Theta\}$$

where D is a properly selected set in $\mathbb{R}^N$. This is a class of bounded functions so we can get uniformly small deviations between the empirical and true means of such functions with high probability. To get such deviations over the complement, D^c, which is not compact, we must properly select D such that the tail of the Gaussian (i.e., the underlying probability measure) decreases fast enough over D^c so that the expectation of $\log f(x|\theta)1_{D^c}(x)$ is negligible.

It is crucial to find the necessary deviation needed to have $\Phi_n(\theta) < 0$ for all θ s.t. $|\theta - \theta_0| > \epsilon$ because from Theorem 3.9, it is clear that this deviation has a direct effect on the sample size n, i.e., appearing in the form of $\frac{1}{\alpha^2(\epsilon)}$ in the expression for n. If $\alpha(\epsilon)$ decreases exponentially fast as N increases or as ϵ decreases to 0 then the number of unlabeled examples will increase exponentially fast.

The last part of this analysis shows that the bound $\alpha(\epsilon)$ in (4.2) may be selected $\Theta(\epsilon^2)$ independent of N. Thus the sample size n stays polynomial in N and in $\frac{1}{\epsilon}$. The major part here is based on a technique that takes advantage of the low dimensional symmetry of the N-dimensional integrals that constitute the function $\Phi(\theta)$. Once such symmetry is identified, it follows that the set of values that $\Phi(\theta)$ takes for

$\theta \in \Theta \subset \mathbb{R}^{2N}$ remains the same for all $N \geq 3$. This establishes a sufficient deviation $\alpha(\epsilon)$ which is constant with N for $N \geq 3$.

4.3 Proof of Theorem 4.2

In the following, all expectations are taken w.r.t. $f(x|\theta_0)$. Initially we use the uniform SLLN to show that with n (as above) unlabeled examples it is possible to estimate the *true* parameter, θ_0, by $\hat{\theta}$ to within small deviation. This implies the decision rule is close to Bayes. And finally we calculate the m that guarantees with high confidence the labeling of the decision regions correctly.

The parameter space $\Theta \subset \mathbb{R}^{2N}$ since in our case θ consists of the two N-dimensional mean-vectors. Denote by θ_0 the unknown parameter which determines the optimal decision border. The likelihood function is defined as

$$L(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

where x_i are the unlabeled examples. The learner calculates the value of θ which achieves the global maximum of $L(\theta)$; call it $\hat{\theta}$. This $\hat{\theta}$ is then used for determining the decision rule as described above. Our aim is to show that $\hat{\theta}$ is ϵ-close to θ_0. First we find how large n suffices to guarantee that there exists a maximum (possibly a relative maximum) of $L(\theta)$, inside the closed ϵ-ball at θ_0 (denoted by $B(\theta_0, \epsilon)$), i.e., that there exists some $\theta_a \in B(\theta_0, \epsilon)$ such that $L(\theta_a) \geq L(\theta_0)$. Then we show that for all θ outside this ball, $L(\theta) < L(\theta_0)$. This will imply that by picking the global maximum of $L(\theta)$, the learner chooses a θ which is ϵ-close to θ_0. Theorem 2 and Proposition 1 of Teicher [29] together with Proposition 2 of Yakowitz [30] imply that mixtures of N-dimensional Gaussians are identifiable hence there can be only one unique true unknown parameter, θ_0, (we disregard the vector $[\theta_{02}, \theta_{01}]$ which differs only in the permutation since the decision border is the same in this case).

Because $L(\cdot)$ is continuous in θ, to have a maximum inside $B(\theta_0, \epsilon)$ it is sufficient to have $L(\theta_\epsilon) < L(\theta_0)$ where $\theta_\epsilon \in \partial B(\theta_0, \epsilon)$, the surface of the ball of radius ϵ at θ_0. We also use $\underline{\epsilon}$ to denote any $2N$-dimensional vector of magnitude ϵ. Hence we need $\sum_{i=1}^{n} \log \frac{f(x_i|\theta_\epsilon)}{f(x_i|\theta_0)} < 0$. For any two different distributions $g(x)$ and $h(x)$, it is easy to show that $\int g(x) \log \frac{h(x)}{g(x)} < 0$. Hence $\mathbf{E} \log \frac{f(x|\theta_\epsilon)}{f(x|\theta_0)} < 0$. (This is provided that both $\mathbf{E} \log f(x|\theta_0)$ and $\mathbf{E} \log f(x|\theta)$, exist, which is true in our case as is shown in Lemma 4.3.)

Lemma 4.3 *For a Gaussian mixture $f(x|\theta)$ with unit covariances and* a priori *probabilities $\frac{1}{2}$,*

$$\mathbf{E} \log \frac{f(x|\theta)}{f(x|\theta_0)} < \infty$$

for any fixed θ_0 and θ in $\mathbb{R}^{2N}$.

PROOF: It suffices to show that $\mathbf{E}_{\theta_0} \log f(x|\theta_0)$ is finite for any fixed θ and θ_0 in $\mathbb{R}^{2N}$. We have

$$\begin{aligned} f(x|\theta) &= \frac{1}{2} f_1(x|\theta_1) + \frac{1}{2} f_2(x|\theta_2) \\ &= \frac{1}{2} (2\pi)^{-N/2} \left(e^{-\frac{1}{2}|x-\theta_1|^2} + e^{-\frac{1}{2}|x-\theta_2|^2} \right) \\ &\leq (2\pi)^{-N/2} < 1 \end{aligned}$$

for all $x \in \mathbb{R}^N$. Consequently,

$$|\log f(x|\theta)| \leq \left| \log \frac{1}{2} f_1(x|\theta_1) \right| \leq \log 2 + \frac{|x-\theta_1|^2}{2}.$$

It follows that

$$\begin{aligned} \mathbf{E}_{\theta_0} \log f(x|\theta) &= \int f(x|\theta_0) \log f(x|\theta) \, dx \\ &\leq \int f(x|\theta_0) \left| \log f(x|\theta) \right| \, dx \\ &\leq \int f(x|\theta_0) \left(\log 2 + \frac{|x-\theta_1|^2}{2} \right) \, dx \end{aligned}$$

$$
\begin{aligned}
&= \log 2 + \frac{1}{4}\int |x-\theta_1|^2 f_1(x|\theta_{01})\,dx + \frac{1}{4}\int |x-\theta_2|^2 f_2(x|\theta_{02})\,dx \\
&< \infty
\end{aligned}
$$

as the two integrals on the right hand side just yield finite combinations of various second moments of the Gaussian. ▮

As mentioned, the class of N-dimensional Gaussian mixtures is identifiable up to permutation of the two parameters of the marginals. This means that if $f(x|\theta) = f(x|\theta_0)$ then either $\theta = [\theta_{01}, \theta_{02}]$ or $\theta = [\theta_{02}, \theta_{01}]$. Now, $\mathbf{E}\log\frac{f(x|\theta)}{f(x|\theta_0)}$ equals 0 if and only if $f(x|\theta) = f(x|\theta_0)$ (cf. Cover & Thomas [31]). Hence it follows that if $\theta \neq [\theta_{01}, \theta_{02}]$ and $\theta \neq [\theta_{02}, \theta_{01}]$ then $\mathbf{E}\log\frac{f(x|\theta)}{f(x|\theta_0)} < 0$. So our following argument will prove that there exists a maximum of $L(\theta)$ either ϵ-close to θ_0, i.e., $[\theta_{01}, \theta_{02}]$, or to the vector $[\theta_{02}, \theta_{01}]$. To be more clear we will only mention conditions which are sufficient that there exist a relative maximum, and later that there exist a global maximum, ϵ-close to θ_0; this will be apparent from the fact that the uniform SLLN is applied only over the ball $\{\theta \in B(\theta_0, \epsilon)\}$. However strictly speaking we should use the uniform SLLN over the region $\{\theta \in B(\theta_0, \epsilon) \bigcup B([\theta_{02}, \theta_{01}], \epsilon)\}$ which will yield the existence of a relative max of $L(\theta)$ either ϵ-close to $[\theta_{01}, \theta_{02}]$, or to the vector $[\theta_{02}, \theta_{01}]$. And similarly with the proof for the global maximum of $L(\theta)$. It turns out that the sample complexities are practically unaffected by this notational nuisance.

4.3.1 Local Maximum of the Likelihood Function

We would like to have with large enough n, and with high confidence, that

$$
\sup_{\theta \in B(\theta_0,\epsilon)} \left| \frac{1}{n}\sum_{i=1}^{n} \log f(x_i|\theta) - \mathbf{E}\log f(x|\theta) \right| \leq \alpha/2. \tag{4.3}
$$

Further ahead, we utilize the uniform SLLN (Theorem 3.9) to achieve this once we define an appropriate function class which satisfies the boundedness condition of the theorem. Once (4.3) is true, it implies that

$$\sup_{\theta_\epsilon}\left|\frac{1}{n}\sum_{i=1}^{n}\log f(x_i|\theta_\epsilon)-\mathbf{E}\log f(x|\theta_\epsilon)\right|\leq\alpha/2$$

and that

$$\left|\frac{1}{n}\sum_{i=1}^{n}\log f(x_i|\theta_0)-\mathbf{E}\log f(x|\theta_0)\right|\leq\alpha/2$$

since it is the special case when $\underline{\epsilon}=\underline{0}$. These imply

$$\sup_{\theta_\epsilon}\left|\frac{1}{n}\sum_{i=1}^{n}\log\frac{f(x_i|\theta_\epsilon)}{f(x_i|\theta_0)}-\mathbf{E}\log\frac{f(x|\theta_\epsilon)}{f(x|\theta_0)}\right|\leq\alpha.$$

Hence to get $\frac{1}{n}\sum_{i=1}^{n}\log\frac{f(x_i|\theta_\epsilon)}{f(x_i|\theta_0)}<0$ it is sufficient to have (4.3) true and choose

$$\alpha(\epsilon)=\inf_{\theta_\epsilon}\left|\mathbf{E}\log\frac{f(x|\theta_\epsilon)}{f(x|\theta_0)}\right|.$$

We need to estimate the dependence of $\alpha(\epsilon)$ on ϵ and show that $\alpha(\epsilon)$ is not arbitrary close to 0 for a fixed $\epsilon>0$. For small $\epsilon>0$ we expand $\mathbf{E}\log\frac{f(x|\theta_\epsilon)}{f(x|\theta_0)}$ in a $2N$-dimensional Taylor series around θ_0 as follows:

$$\mathbf{E}\log\frac{f(x|\theta_\epsilon)}{f(x|\theta_0)}=\int f(x|\theta_0)\log f(x|\theta_\epsilon)\,dx-\int f(x|\theta_0)\log f(x|\theta_0)\,dx$$

The first term becomes

$$\begin{aligned}\int f(x|\theta_0)\log f(x|\theta_\epsilon)\\ =\;&\int f(x|\theta_0)\log f(x|\theta_0)\,dx+\sum_{i=1}^{2N}\epsilon_i\int f(x|\theta_0)\frac{\partial}{\partial\theta_{0i}}\log f(x|\theta_0)\,dx\\ +\;&\frac{1}{2}\sum_{i,j=1}^{2N}\epsilon_i\epsilon_j\int f(x|\theta_0)\frac{\partial^2}{\partial\theta_{0i}\partial\theta_{0j}}\log f(x|\theta_0)\,dx\\ +\;&\frac{1}{6}\sum_{i,j,k=1}^{2N}\epsilon_i\epsilon_j\epsilon_k\int f(x|\theta_0)\frac{\partial^3}{\partial\theta_{0i}\partial\theta_{0j}\partial\theta_{0k}}\log f(x|\theta_0)|_{\theta_{\epsilon'}}\,dx\qquad(4.4)\end{aligned}$$

where $\underline{\epsilon}'$ is on the line between 0 and $\underline{\epsilon}$. Lemma 4.4 shows that the integral of the third order partials is bounded by some constant making the term bounded above by $c_0\epsilon^3$ for some positive constant c_0.

Lemma 4.4

$$\sum_{i,j,k=1}^{2N} \epsilon_i \epsilon_j \epsilon_k \int f(x|\theta_0) \frac{\partial^3}{\partial\theta_{0i}\partial\theta_{0j}\partial\theta_{0k}} \log f(x|\theta_{\epsilon'})\, dx \le c\epsilon^3 \tag{4.5}$$

for some constant $c > 0$.

PROOF: First work on the integral; denote $\frac{\partial}{\partial\theta_{0i}} f(x|\theta_0)|_{\theta_{\epsilon'}} \equiv f_i$. Then from page 61 we have

$$\frac{\partial^2}{\partial\theta_{0i}\partial\theta_{0j}} \log f(x|\theta_0)|_{\theta_{\epsilon'}} = \frac{f f_{ij} - f_i f_j}{f^2}$$

and hence (denote $f_0 \equiv f(x|\theta_0)$ and $f \equiv f(x|\theta_{\epsilon'})$)

$$\begin{aligned}
&\int f(x|\theta_0) \frac{\partial^3}{\partial\theta_{0i}\partial\theta_{0j}\partial\theta_{0k}} \log f(x|\theta_0)|_{\theta_{\epsilon'}}\, dx \\
&= \int f_0 \left(\frac{f f_{ij} - f_i f_j}{f^2} \right)_k dx = \int f_0 \frac{f^2(f_k f_{ij} + f f_{ijk} - f_{ik} f_j - f_i f_{jk}) - 2 f f_k (f f_{ij} - f_i f_j)}{f^4}\, dx \\
&= \int f_0 \left(\frac{f_k f_{ij} + f f_{ijk} - f_{ik} f_j - f_i f_{jk}}{f^2} - 2 \frac{f_k f f_{ij} - f_k f_i f_j}{f^3} \right) dx
\end{aligned}$$

Let $\hat f$ denote any one of the two class conditionals, i.e. $f(x|\theta_{\epsilon'1})$ or $f(x|\theta_{\epsilon'2})$ and $poly_r(x)$ denote any polynomial in $x_1, \ldots, x_N$ of degree $\le r$. Then $f_i = \hat f\, poly_1(x)$, $f_{ij} = \hat f\, poly_2(x)$, and $f_{ijk} = \hat f\, poly_3(x)$. Hence the above is bounded by

$$\begin{aligned}
\int f_0 \Bigg(\frac{\hat f^2 |poly_3(x)|}{f^2} + \frac{\hat f}{f} |poly_3(x)| + \frac{\hat f^2 \,|poly_3(x)|}{f^2} \quad &+ \quad \frac{\hat f^2 \,|poly_3(x)|}{f^2} \\
&+ \quad \frac{\hat f^2 \,|poly_3(x)|}{f^2} + \frac{\hat f^3 \,|poly_3(x)|}{f^3} \Bigg) dx
\end{aligned}$$

Recall that $f = f(x|\theta_{\epsilon'1})/2 + f(x|\theta_{\epsilon'2})/2$ hence $1/f \le 2/\hat f$ and so the above is bounded by $12 \int f_0\, |poly_3(x)|\, dx$. This is composed of a finite sum of products of terms such as

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}x_i^2} |x_i|^r \, dx_i \quad \text{for } 0 \le r \le 3,\ 1 \le i \le N$$

It is easy to show that such terms are finite. For instance, we can bound the 3^{rd} order term as follows (we use y to denote any one-dimensional component):

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}y^2}|y|^3\,dy \;=\; \frac{1}{\sqrt{2\pi}}\int_{|y|\le A} e^{-\frac{1}{2}y^2}|y|^3\,dy + \frac{1}{\sqrt{2\pi}}\int_{|y|>A} e^{-\frac{1}{2}y^2}|y|^3\,dy.$$

The first term is clearly finite since the integrand is continuous and over a closed set. The second term is bounded as follows:

$$\frac{1}{\sqrt{2\pi}}\int_{|y|>A} e^{-\frac{1}{2}y^2}|y|^3\,dy \;\le\; \frac{1}{\sqrt{2\pi}}\int_{|y|>A} e^{-y^2/2c}\,dy = \frac{1}{\sqrt{2\pi}}\int_{z^2>A^2/c} e^{-z^2/2}\,dz$$

where using the fact that $\forall y > A > e^{1/2}$, $\log y/y^2 \le \log A/A^2$ hence $e^{-y^2/2}y^3 \le e^{-y^2/2c}$ for $c \ge 1/(1-6\frac{\log A}{A^2})$; therefore it suffices to let $A \ge e^{1/2}$ and choose c accordingly. Finally, the last integral can be bounded by $2Nc^2/A^4$ using the variance of a chi-square and Chebyshev's inequality. So we have shown the integral in (4.5) is finite. Call it a_{ijk}. From (4.5) we have

$$\begin{aligned}\sum_{i,j,k}\epsilon_i\epsilon_j\epsilon_k a_{ijk} &\le \sum_{i,j}|\epsilon_i\epsilon_j|\sum_k|\epsilon_k a_{ijk}| \le \sum_{i,j}|\epsilon_i\epsilon_j|\sqrt{\sum_k \epsilon_k^2}\sqrt{\sum_k a_{ijk}^2} = \epsilon\sum_i|\epsilon_i|\sum_j|\epsilon_j b_{ij}|\\ &\le \epsilon^2\sum_i|\epsilon_i c_i| \le c\epsilon^3\end{aligned}$$

for some constant $c > 0$. ∎

For the other terms note that

$$f(x|\theta_0)\frac{\partial}{\partial\theta_{0i}}\log f(x|\theta_0) = \frac{\partial}{\partial\theta_{0i}}f(x|\theta_0).$$

Then the derivative can be represented as a limit of a bounded sequence (since $f(x|\theta_0)$ is differentiable) hence by the Lebesgue bounded convergence theorem we can take the limit, hence the derivative, outside. Doing this, the term with the first order derivatives becomes zero. Now consider the terms corresponding to the second order derivatives. We have:

$$
\begin{aligned}
&\int f(x|\theta_0)\frac{\partial}{\partial\theta_{0i}}\frac{\partial}{\partial\theta_{0j}}\log f(x|\theta_0)\,dx \\
&= \int f(x|\theta_0)\frac{\partial}{\partial\theta_{0i}}\left(\frac{\frac{\partial}{\partial\theta_{0j}}f(x|\theta_0)}{f(x|\theta_0)}\right) \\
&= \int f(x|\theta_0)\frac{f(x|\theta_0)\frac{\partial}{\partial\theta_{0i}}\frac{\partial}{\partial\theta_{0j}}f(x|\theta_0) - \frac{\partial}{\partial\theta_{0i}}f(x|\theta_0)\frac{\partial}{\partial\theta_{0j}}f(x|\theta_0)}{f^2(x|\theta_0)}\,dx \\
&= -\int f(x|\theta_0)\frac{\frac{\partial}{\partial\theta_{0i}}f(x|\theta_0)\frac{\partial}{\partial\theta_{0j}}f(x|\theta_0)}{f^2(x|\theta_0)}\,dx \\
&= -\mathbf{E}\left(\frac{\partial}{\partial\theta_{0i}}\log f(x|\theta_0)\frac{\partial}{\partial\theta_{0j}}\log f(x|\theta_0)\right) = -I_{ij}(\theta_0)
\end{aligned}
$$

where $I_{ij}(\theta_0)$ is the ij^{th} element in the Fisher information matrix evaluated at θ_0. Thus the above imply

$$\mathbf{E}\log\frac{f(x|\theta_\epsilon)}{f(x|\theta_0)} = -\frac{1}{2}\sum_{i,j}\epsilon_i\epsilon_j I_{ij}(\theta_0) + c_0\epsilon^3$$

The first term on the right is $-\frac{1}{2}B_{\theta_0}\epsilon^2$ where B_{θ_0} is a constant depending on θ_0 which is positive if $\theta_{01} \neq \theta_{02}$ since for such a θ_0, $I(\theta_0)$ is positive definite, as is shown next.

Lemma 4.5 *Given a mixture $f(x|\theta_0)$ of two Gaussians $f_1(x|\theta_{01})$ and $f_2(x|\theta_{02})$ with means $\theta_{01} \neq \theta_{02}$ and with unit covariance matrices, then the Fisher Information matrix $I(\theta_0)$ whose ij^{th} element is*

$$\mathbf{E}\left(\frac{\partial}{\partial\theta_{0i}}\log f(x|\theta_0)\frac{\partial}{\partial\theta_{0j}}\log f(x|\theta_0)\right)$$

is positive definite.

PROOF:

For any vector $u \neq 0$

$$
\begin{aligned}
u^T I u &= \sum_{i,j} u_i u_j \int f(x|\theta_0)\frac{\partial}{\partial\theta_{0i}}\log f(x|\theta_0)\frac{\partial}{\partial\theta_{0j}}\log f(x|\theta_0)\,dx \\
&= \int f(x|\theta_0)\left(\sum_i u_i\frac{\partial}{\partial\theta_{0i}}\log f(x|\theta_0)\right)^2 dx > 0
\end{aligned}
$$

which follows since

$$\left(\sum_i u_i \frac{\partial}{\partial \theta_{0i}} \log f(x|\theta_0)\right)^2$$

is not identically zero over the probability 1 support of $f(x|\theta_0)$ as we now show. The functions in the set

$$\left\{\frac{\partial}{\partial \theta_{0i}} \log f(x|\theta_0) : 1 \leq i \leq 2N\right\}$$

are linearly independent when $\theta_{01} \neq \theta_{02}$ since for any $u \neq 0$ we question whether there is a solution to

$$\sum_{i=1}^{2N} u_i \frac{\partial}{\partial \theta_{0i}} \log f(x|\theta_0) \quad = \quad \frac{1}{2}\sum_{i=1}^{N} u_i(x_i - \theta_{0i})\frac{f_1}{f} + \frac{1}{2}\sum_{i=N+1}^{2N} u_i(x_{i-N} - \theta_{0i})\frac{f_2}{f} = 0.$$

This is the same as asking if any function in the set

$$\{x_1 - \theta_{01}, x_2 - \theta_{02}, \ldots, x_N - \theta_{0N}, \frac{f_2}{f_1}(x_1 - \theta_{0,N+1}), \frac{f_2}{f_1}(x_2 - \theta_{0,N+2}), \ldots, \frac{f_2}{f_1}(x_N - \theta_{0,2N})\}$$

is a linear combination of the others. By inspection, as long as $\theta_{01} \neq \theta_{02}$, this is not possible because $\frac{f_2}{f_1}$ is a nonlinear function of the $x_i - \theta_{0i}$.

So for $\theta_0 = [y, z]$ where $y \neq z$ and $y, z \in \mathbb{R}^N$, we've shown that the Fisher matrix $I(\theta_0)$ is positive definite. ▮

Now from Rayleigh's quotient we have $\frac{\underline{\epsilon}^T I \underline{\epsilon}}{\epsilon^2} \geq \lambda_{min} > 0$ because all eigenvalues of a positive definite matrix are positive. This implies that $\underline{\epsilon}^T I \underline{\epsilon} = \Omega(\epsilon^2)$ and so for any $\underline{\epsilon}$

$$\mathbf{E} \log \frac{f(x|\theta_\epsilon)}{f(x|\theta_0)} \quad = \quad -\frac{1}{2}B_{\theta_0}\epsilon^2 + c_0\epsilon^3 < 0$$

therefore

$$\sup_{\theta_\epsilon} \mathbf{E} \log \frac{f(x|\theta_\epsilon)}{f(x|\theta_0)} = -\frac{1}{2}B_{\theta_0}\epsilon^2 + c_0\epsilon^3 < 0$$

for small enough ϵ, and $B_{\theta_0} > 0$.

(The case of $\theta_{01} = \theta_{02}$ is trivial as this results in $P_{error} = \frac{1}{2}$. Without loss of generality we henceforth assume $\theta_{01} \neq \theta_{02}$ whence $B_{\theta_0} > 0$ strictly.)

We did not consider the dependence of the right hand side on N but instead only as a function of ϵ and θ_0 since the left side is invariant for $N \geq 3$; we show later that this is true more generally. Hence we have

$$\alpha(\epsilon) = \inf_{\theta_\epsilon} \left| \mathbf{E} \log \frac{f(x|\theta_\epsilon)}{f(x|\theta_0)} \right| = B_{\theta_0}\epsilon^2 + c_0\epsilon^3 \tag{4.6}$$

and with this selection of $\alpha = \alpha(\epsilon)$, (4.3) will force $\sum_{i=1}^{n} \log \frac{f(x_i|\theta_\epsilon)}{f(x_i|\theta_0)} < 0$. (Henceforth we rename B_{θ_0} as c_1).

We now estimate the unlabeled sample size n needed to guarantee (4.3). The uniform SLLN holds for a class of functions that are uniformly bounded (see Theorem 3.9). Hence define a function class $\mathcal{G}$ as follows: let $D \subset \mathbb{R}^N$ be a compact subset of the probability one support of $f(x)$ and denote its complement as D^c. Let

$$\mathcal{G} = \{\log f(x|\theta) 1_D(x) : \theta \in B(\theta_0, \epsilon)\}.$$

The functions in $\mathcal{G}$ are bounded hence we can use the uniform SLLN over it. Then

$$\begin{aligned}
&\mathbf{P}\left(\sup_{\theta \in B(\theta_0,\epsilon)} \left| \mathbf{E} \log f(x|\theta) - \frac{1}{n}\sum_{i=1}^{n} \log f(x_i|\theta) \right| > \alpha/2 \right) \\
&= \quad \mathbf{P}\Bigg(\sup_{\theta \in B(\theta_0,\epsilon)} \Bigg| \int_D f(x|\theta_0) \log f(x|\theta)\, dx - \frac{1}{n}\sum_{i=1}^{n} \log f(x_i|\theta) 1_D(x_i) \\
&\quad + \int_{D^c} f(x|\theta_0) \log f(x|\theta)\, dx - \frac{1}{n}\sum_{i=1}^{n} \log f(x_i|\theta) 1_{D^c}(x_i) \Bigg| > \alpha/2 \Bigg).
\end{aligned}$$

Using Boole's inequality we upper bound this by

$$\begin{aligned}
&\mathbf{P}\left(\sup_{\theta \in B(\theta_0,\epsilon)} \left| \int_D f(x|\theta_0) \log f(x|\theta)\, dx - \frac{1}{n}\sum_{i=1}^{n} \log f(x_i|\theta) 1_D(x_i) \right| > \alpha/4 \right) \\
&+ \mathbf{P}\left(\sup_{\theta \in B(\theta_0,\epsilon)} \left| \int_{D^c} f(x|\theta_0) \log f(x|\theta)\, dx - \frac{1}{n}\sum_{i=1}^{n} \log f(x_i|\theta) 1_{D^c}(x_i) \right| > \alpha/4 \right).
\end{aligned} \tag{4.7}$$

The first term of the sum is the probability of uniform convergence for functions in $\mathcal{G}$. Hence Theorem 3.9 can be used here to make the first term arbitrary small. The theorem does not directly apply for the second term since $\log f(x|\theta)$ is unbounded over D^c. However this term can be made arbitrary small as we show next. Our aim is to choose the region D such that both the empirical and the true means over D^c are small (so their absolute difference must be small); that is, we do not need the SLLN to guarantee their difference is small. We achieve that by utilizing the rapid decay of the Gaussian outside a large enough sphere centered at the mean. We bound the second term of (4.7) by

$$\mathbf{P}\left(\sup_{\theta\in B(\theta_0,\epsilon)}\left|\int_{D^c} f(x|\theta_0)\log f(x|\theta)\,dx\right| + \sup_{\theta\in B(\theta_0,\epsilon)}\left|\frac{1}{n}\sum_{i=1}^{n}\log f(x_i|\theta)1_{D^c}(x_i)\right| > \alpha/4\right)$$

and

$$\sup_{\theta\in B(\theta_0,\epsilon)}\left|\int_{D^c} f(x|\theta_0)\log f(x|\theta)\,dx\right| \le \int_{D^c} f(x|\theta_0)\sup_{\theta\in B(\theta_0,\epsilon)}|\log f(x|\theta)|\,dx.$$

Without loss of generality suppose $\theta_{01}=0$ (rotational symmetry allows us to translate the coordinate system). We choose

$$D=\{x:|x|\le A \text{ or } |x-\theta_{02}|\le A\}$$

for some constant A. We then have

$$\begin{aligned}
\int_{x\in D^c} f(x|\theta_0)\sup_{\theta\in B(\theta_0,\epsilon)}|\log f(x|\theta)|\,dx &= \int_{|x|>A\,\cap\,|x-\theta_{02}|>A} f(x|\theta_0)\sup_{\theta\in B(\theta_0,\epsilon)}|\log f(x|\theta)|\,dx\\
&\le \frac{1}{2}\int_{|x|>A\,\cap\,|x-\theta_{02}|>A} f(x|\theta_{01})\sup_{\theta\in B(\theta_0,\epsilon)}\left|\log\frac{1}{2}f_1(x|\theta_1)\right|dx\\
&\quad+\frac{1}{2}\int_{|x|>A\,\cap\,|x-\theta_{02}|>A} f(x|\theta_{02})\sup_{\theta\in B(\theta_0,\epsilon)}\left|\log\frac{1}{2}f_2(x|\theta_2)\right|dx\\
&\le \frac{1}{2}\int_{|x|>A} f(x|\theta_{01})\sup_{\theta\in B(\theta_0,\epsilon)}\left|\log\frac{1}{2}f_1(x|\theta_1)\right|dx\\
&\quad+\frac{1}{2}\int_{|x-\theta_{02}|>A} f(x|\theta_{02})\sup_{\theta\in B(\theta_0,\epsilon)}\left|\frac{1}{2}\log f_2(x|\theta_2)\right|dx
\end{aligned}$$

$$\begin{aligned} &\le \quad \frac{1}{2}\int_{|x|>A} f(x|\theta_{01})\left|\log\frac{1}{2}f_\sigma(x)\right| dx \\ &\quad + \frac{1}{2}\int_{|x-\theta_{02}|>A} f(x|\theta_{02})\left|\log\frac{1}{2}f_\sigma(x-\theta_{02})\right| dx \\ &= \quad \frac{1}{2}\int_{|x|>A} \frac{1}{(2\pi)^{N/2}}e^{-|x|^2/2}\left|\log\frac{1}{2}f_\sigma(x)\right| dx \\ &\quad + \frac{1}{2}\int_{|x-\theta_{02}|>A} \frac{1}{(2\pi)^{N/2}}e^{-|x-\theta_{02}|^2/2}\left|\log\frac{1}{2}f_\sigma(x-\theta_{02})\right| dx, \end{aligned}$$

where the normal density

$$f_\sigma(x) = \frac{1}{(2\pi\sigma^2)^{N/2}}e^{-|x|^2/2\sigma^2}$$

is defined (by choosing an appropriate σ) such that

$$|\log f_1(x|\theta_1)| \le |\log f_\sigma(x)|$$

for $\theta \in B(\theta_0, \epsilon)$ and $\{x : |x| > A\}$. (Recall that θ_1 denotes the first N components of θ.) For this it is sufficient to choose σ that satisfies

$$\frac{1}{(2\pi)^{N/2}}e^{\{-(x_1+\epsilon)^2-x_2^2-\ldots-x_N^2\}/2}|_{x=[A,0,\ldots,0]} = \frac{1}{(2\pi\sigma^2)^{N/2}}e^{-|x|^2/2\sigma^2}|_{x=[A,0,\ldots,0]}.$$

Later, we discuss more specifically the choice of σ. The above is bounded by

$$\int_{|x|>A} \frac{1}{(2\pi)^{N/2}}e^{-|x|^2/2}\left(\log 2 + \frac{N}{2}\log\frac{1}{2\pi\sigma^2} + \frac{1}{2\sigma^2}|x|^2\right) dx \tag{4.8}$$

The terms of (4.8) can be bounded by the tail of a chi-square distribution (see Rao [22]). We outline how this is done. We use the fact that

$$e^{-z/2}\, z \;\le\; \frac{1}{e^{N/2}}e^{-z/2e}$$

for $z \ge 9N$ and for all $N \ge 1$ (since $\log 9N/N \le \log 9/9$ for all $N \ge 1$). So we have

$$\begin{aligned} &\frac{1}{(2\pi)^{N/2}}\int_{|x|^2>A^2} e^{-|x|^2/2}|x|^2\, dx \\ &\qquad \le \frac{1}{(2\pi)^{N/2}e^{N/2}}\int_{|x|^2>A^2} e^{-|x|^2/2e}\, dx = \frac{1}{(2\pi)^{N/2}}\int_{|x|^2>\frac{A^2}{e}} e^{-|x|^2/2}\, dx \end{aligned}$$

for $A^2 \geq 9N$ and $N \geq 1$. The last integral is equivalent to the probability that the sum of squares of N standard normal random variables, which is distributed as a chi-square with second moment $2N$, is $> A^2/e$. Hence using Chebyshev's inequality we can bound this integral by $2Ne^2/A^4$. The remaining terms of (4.8) are simple to bound by similar arguments. Finally, let us return to the choice of σ. The condition $A^2 \geq 9N$ together with the condition on σ (in the construction of $f_\sigma(x)$) yields:

$$\frac{e^{-(A+\epsilon)^2/2}}{(2\pi)^{N/2}} \geq \frac{e^{-A^2/2\sigma^2}}{(2\pi)^{N/2}\sigma^N} \quad \Rightarrow \quad \frac{(A+\epsilon)^2}{2} \leq \frac{A^2}{2\sigma^2} + N\log\sigma$$

Letting $y = 1/\sigma^2$ we have $\frac{1}{2}\log y = \log(1/\sigma)$. Solving for y by bootstrapping yields

$$1/\sigma^2 \geq 2(1 + b_0(\epsilon/\sqrt{N})) = b_1$$

(where b_0, b_1 are constants). Thus we can take $1/\sigma^2 = b_1$, i.e., treat it as a constant w.r.t. N in (4.8). So we get the bound

$$\sup_{\theta \in B(\theta_0,\epsilon)} \left| \int_{x \in D^c} f(x|\theta_0) \log f(x|\theta)\, dx \right| \leq c_2 \frac{N}{A^4}(N + c_3)$$

(where c_2, c_3 are positive constants) which is independent of θ_0. Denote this by Δ. We continue to bound the second term of (4.7). We have

$$\begin{aligned}
&\mathbf{P}\Big(\sup_{\theta \in B(\theta_0,\epsilon)} \left| \int_{x \in D^c} f(x|\theta_0) \log f(x|\theta)\, dx \right| + \sup_{\theta \in B(\theta_0,\epsilon)} \left| \frac{1}{n} \sum_{i=1}^{n} \log f(x_i|\theta) 1_{D^c}(x_i) \right| > \alpha/4 \Big) \\
&\leq \mathbf{P}\left(\Delta + \sup_{\theta \in B(\theta_0,\epsilon)} \left| \frac{1}{n} \sum_{i=1}^{n} \log f(x_i|\theta) 1_{D^c}(x_i) \right| > \alpha/4 \right) \\
&\leq \mathbf{P}\left(\sup_{\theta \in B(\theta_0,\epsilon)} \frac{1}{n} \sum_{i=1}^{n} |\log f(x_i|\theta)|\, 1_{D^c}(x_i) > \alpha/4 - \Delta \right) \\
&\leq \frac{\mathbf{E}\left| \sup_{\theta \in B(\theta_0,\epsilon)} \frac{1}{n} \sum_{i=1}^{n} |\log f(x_i|\theta)|\, 1_{D^c}(x_i) \right|}{|\alpha/4 - \Delta|}
\end{aligned}$$

the last step following from Markov's inequality. Now

$$\sup_{\theta \in B(\theta_0,\epsilon)} \frac{1}{n} \sum_{i=1}^{n} |\log f(x_i|\theta)|\, 1_{D^c}(x_i) \;\leq\; \frac{1}{n} \sum_{i=1}^{n} \sup_{\theta \in B(\theta_0,\epsilon)} |\log f(x_i|\theta)|\, 1_{D^c}(x_i).$$

Hence

$$\begin{aligned}\frac{\mathbf{E}\sup_{\theta\in B(\theta_0,\epsilon)}|\log f(x|\theta)|\,1_{D^c}(x)}{|\alpha/4-\Delta|} &= \frac{\int_{x\in D^c} f(x|\theta_0)\sup_{\theta\in B(\theta_0,\epsilon)}|\log f(x|\theta)|\,dx}{|\alpha/4-\Delta|}\\ &\leq \frac{\Delta}{\alpha/4-\Delta}\\ &\leq \frac{\delta}{4}\end{aligned}$$

the last bound achieved by selecting $\Delta \leq \frac{\alpha\delta}{4\delta+16}$ or equivalently $A^2 = c_4\frac{N}{\sqrt{\alpha\delta}}$. This is the choice for A such that the second term of (4.7) is at most $\delta/4$.

We now proceed to find a bound on the first term of (4.7). Here the functions are bounded since their domain D is bounded so that Theorem 3.9 is applicable. The procedure will be to directly find an upper bound on the covering number of this class instead of calculating its VC-dimension for bounding the covering number as mentioned just above Theorem 3.8. Then we can use the analysis in Haussler [12] and obtain a bound similar to Theorem 3.9. The class $\mathcal{G}$ is defined with the parameter θ restricted to within the closed ball $B(\theta_0,\epsilon)$. It is easy to calculate an upper bound for the covering number of this ball with respect to the Euclidean norm, denoted by $|\cdot|$. From Haussler [12], an ϵ'-cover for a set T is a finite set C (not necessarily in T) such that for all $x \in T$ there is a $y \in C$ with $|x-y| \leq \epsilon'$. The cardinality of the smallest ϵ'-cover for T is called the *covering number* and is denoted by $\mathcal{N}(\epsilon',\mathcal{T},|\cdot|)$. A set T is ϵ'-separated if for all distinct $x,y \in T$, $|x-y| > \epsilon'$. The size of the largest ϵ'- separated subset of T is called the *packing number* and is denoted by $\mathcal{M}(\epsilon',\mathcal{T},|\cdot|)$. It is easy to see that $\mathcal{N}(\epsilon',\mathcal{T},|\cdot|) \leq \mathcal{M}(\epsilon',\mathcal{T},|\cdot|)$. Consider an ϵ'-separated set with size $M(\epsilon',T,|\cdot|)$. Put around each point a sphere of radius ϵ' and let this be a covering. Suppose it is not an ϵ'-covering. Then there exists some point x whose distance from any of the points is $> \epsilon'$. But this would increment the size of the ϵ'-separated set by one, which contradicts the fact that it is the maximum ϵ'-separated set. Therefore there exists an ϵ' covering of size $M(\epsilon',T,|\cdot|)$ and clearly the smallest ϵ'-cover has cardinality smaller than it. That proves the inequality. Hence it suffices to find the packing number of our ball $B(\theta_0, \epsilon)$. The volume of a $2N$-dimensional ball is $k_{2N}r^{2N}$ where k_{2N} is a constant and r is its radius. The number of ϵ'- balls that can be contained without overlap in $B(\theta_0, \epsilon)$ is therefore at most $(k_{2N}\epsilon^{2N})/(k_{2N}\epsilon'^{2N}) = (\frac{\epsilon}{\epsilon'})^{2N}$. Clearly this is also the maximum number of $2\epsilon'$-separated set of points contained in $B(\theta_0, \epsilon)$ hence is equal to $\mathcal{M}(2\epsilon', B(\theta_0, \epsilon), |\cdot|)$. So therefore it follows that $\mathcal{N}(\epsilon', B(\theta_0, \epsilon), |\cdot|) \leq (\frac{2\epsilon}{\epsilon'})^{2N}$. We now proceed to find a bound on the covering number of the class $\mathcal{G}$ with respect to the L^1-norm (as in Definition 3.7).

Let the set $\{\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_{\mathcal{N}(\epsilon', B(\theta_0, \epsilon), |\cdot|)}\}$ be a covering of $B(\theta_0, \epsilon)$. We now construct a covering for $\mathcal{G}$. Fixing a particular $\tilde{\theta}_i$, we show that any function $g(x|\theta)$, with $|\theta - \tilde{\theta}_i| \leq \epsilon'$ is δ'-close to $g(x|\tilde{\theta}_i)$ in the $\sup_x$-norm. Using the notation for such a θ as $\theta_{\epsilon'}$ we therefore have $|\theta_{\epsilon'1} - \tilde{\theta}_{i1}| \leq \epsilon'$ and $|\theta_{\epsilon'2} - \tilde{\theta}_{i2}| \leq \epsilon'$ and so $\tilde{\theta}_{i1} = \theta_{\epsilon'1} + v_1$ and $\tilde{\theta}_{i2} = \theta_{\epsilon'2} + v_2$ where both v_1 and v_2 are of magnitude $\leq \epsilon'$. Then

$$\sup_{\theta_{\epsilon'}} \sup_{x \in \mathbb{R}^N} \left| g(x|\theta_{\epsilon'}) - g(x|\tilde{\theta}_i) \right| = \sup_{\theta_{\epsilon'}} \sup_{x \in \mathbb{R}^N} \left| \log f(x|\theta_{\epsilon'}) 1_D(x) - \log f(x|\tilde{\theta}_i) 1_D(x) \right|$$

$$\begin{aligned} &= \sup_{\theta_{\epsilon'}} \sup_{x \in \mathbb{R}^N} \Big| \log \left(e^{-\frac{1}{2}|x-\theta_{\epsilon'1}|^2} + e^{-\frac{1}{2}|x-\theta_{\epsilon'2}|^2} \right) 1_D(x) \\ &- \log \left(e^{-\frac{1}{2}|x-\theta_{\epsilon'1}-v_1|^2} + e^{-\frac{1}{2}|x-\theta_{\epsilon'2}-v_2|^2} \right) 1_D(x) \Big| \\ &= \sup_{\theta_{\epsilon'}} \sup_{x \in D} \left| \log \left(e^{-\frac{1}{2}|x-\theta_{\epsilon'1}|^2} + e^{-\frac{1}{2}|x-\theta_{\epsilon'2}|^2} \right) - \log \left(e^{-\frac{1}{2}|x-\theta_{\epsilon'1}-v_1|^2} + e^{-\frac{1}{2}|x-\theta_{\epsilon'2}-v_2|^2} \right) \right| \\ &= \sup_{y \in E_i} \left| \log \left(e^{-\frac{1}{2}|y_1|^2} + e^{-\frac{1}{2}|y_2|^2} \right) - \log \left(e^{-\frac{1}{2}|y_1-v_1|^2} + e^{-\frac{1}{2}|(y_2-v_2|^2} \right) \right| \end{aligned}$$

where $E_i = \{y : y = [x, x] - \theta_{\epsilon'}, x \in D, \theta_{\epsilon'} \in B(\tilde{\theta}_i, \epsilon')\}$ is a compact subset of $\mathbb{R}^{2N}$ as is now shown (the subscript i shows the dependence on $\tilde{\theta}_i$). A metric space X, e.g., $\mathbb{R}^{2N}$ in our case, is compact if every infinite sequence has a subsequence converging to a point in X (cf. Royden [32]). It suffices to consider a sequence $y_n \in E_i$ and prove that it has a convergent subsequence. Since $y_n \equiv [x_n, x_n] - \theta_{\epsilon'n}$ is in E_i then

$x_n \in D$. But D is compact hence $\exists$ a subsequence $x_{m(n)} \rightarrow \hat{x} \in D$. Corresponding to it we have the subsequence $y_{m(n)} = [x_{m(n)}, x_{m(n)}] - \theta_{\epsilon' m(n)}$. Now $\theta_{\epsilon' m(n)} \in B(\tilde{\theta}_i, \epsilon')$ hence $\exists$ a subsequence $\theta_{\epsilon' k(m(n))} \rightarrow \hat{\theta} \in B(\tilde{\theta}_i, \epsilon')$. Corresponding to this we have $y_{k(m(n))} = [x_{k(m(n))}, x_{k(m(n))}] - \theta_{\epsilon' k(m(n))}$. But clearly $x_{k(m(n))} \rightarrow \hat{x} \in D$. Hence the subsequence $y_{k(m(n))} \rightarrow \hat{y} \in E_i$ which proves the claim that E_i compact. Continuing, we have

$$\sup_{\theta_{\epsilon'}} \sup_{x \in \mathbb{R}^N} \left| g(x|\theta_{\epsilon'}) - g(x|\tilde{\theta}_i) \right| = \sup_{y \in E_i} |h(y) - h(y - v)|$$

where $h(y) = \log\left(e^{-\frac{1}{2}|y_1|^2} + e^{-\frac{1}{2}|y_2|^2}\right)$ is continuous and $|v| \leq \sqrt{2}\epsilon'$; note that for any $y \in E_i$, $y - v \in E_i$ since $\theta_{\epsilon'} + v \equiv \tilde{\theta}_i \in B(\tilde{\theta}_i, \epsilon')$. Clearly h is uniformly continuous over E_i so that, for any fixed $\epsilon' > 0$ we can find M_i such that

$$\sup_{y \in E_i} |h(y) - h(y - v)| \leq M_i \epsilon'.$$

It follows that

$$\sup_{x} |g(x|\theta) - g(x|\tilde{\theta}_i)| \leq M_i \epsilon' \leq M'_{\theta_0} \epsilon', \qquad \theta \in B(\tilde{\theta}_i, \epsilon')$$

where $M'_{\theta_0} \equiv \max_{1 \leq i \leq \mathcal{N}(\epsilon', \mathcal{B}(\theta_{\prime}, \epsilon), |\cdot|)} M_i$. Finally, take any function $g(x|\theta)$ where $\theta \in B(\theta_0, \epsilon)$. Then in the covering of $B(\theta_0, \epsilon)$ there exists $\tilde{\theta}_i$ such that $|\theta - \tilde{\theta}_i| \leq \epsilon'$. Corresponding to this $\tilde{\theta}_i$ there exists a function $g(x|\tilde{\theta}_i)$ such that $\sup_x |g(x|\theta) - g(x|\tilde{\theta}_i)| \leq M'_{\theta_0} \epsilon'$. This implies that for any function $g(x|\theta) \in \mathcal{G}$, there exists a $g(x|\tilde{\theta}_i)$ in the collection

$$\{g(x|\tilde{\theta_1}), g(x|\tilde{\theta_2}), \ldots, g(x|\tilde{\theta}_{\mathcal{N}(\epsilon', \mathcal{B}(\theta_0, \epsilon), |\cdot|)})\}$$

such that

$$\sup_{x \in \mathbb{R}^N} |g(x|\theta) - g(x|\tilde{\theta}_i)| \leq M'_{\theta_0} \epsilon'.$$

Hence

$$\mathbf{E}|g(x|\theta) - g(x|\tilde{\theta}_i)| \leq M'_{\theta_0} \epsilon' \equiv \delta'$$

where expectation is w.r.t. *any* probability measure on $\mathbb{R}^N$. So a subset of $\mathcal{G}$ with

$$\left(\frac{2\epsilon M'_{\theta_0}}{\delta'}\right)^{2N}$$

functions covers $\mathcal{G}$ to an accuracy of δ' w.r.t L^1_Q-norm hence $\mathcal{N}(\delta', \mathcal{G}, L^1_Q) \leq \left(\frac{2\epsilon M'_{\theta_0}}{\delta'}\right)^{2N}$ for any measure Q (see Definition 3.7). (Henceforth we rename M'_{θ_0} as c_5). Lastly, in order to use Theorem 3.9 we need to determine a bound on the range of the functions of $\mathcal{G}$. We have

$$\begin{aligned}
\sup_\theta \sup_{x\in D} |\log f| &= \sup_\theta \sup_{|x-\theta_{01}|\leq A \text{ or } |x-\theta_{02}|\leq A} \left|\log\left(\frac{1}{2}f_1 + \frac{1}{2}f_2\right)\right| \\
&\leq \sup_\theta \sup_{|x-\theta_{01}|\leq A} \left|\log\left(\frac{1}{2}f_1 + \frac{1}{2}f_2\right)\right| + \sup_\theta \sup_{|x-\theta_{02}|\leq A} \left|\log\left(\frac{1}{2}f_1 + \frac{1}{2}f_2\right)\right| \\
&\leq \sup_\theta \sup_{|x-\theta_{01}|\leq A} \left|\log \frac{1}{2}f_1\right| + \sup_\theta \sup_{|x-\theta_{02}|\leq A} \left|\log \frac{1}{2}f_2\right| \\
&= 2\sup_\theta \sup_{|x-\theta_{01}|\leq A} \left|\log \frac{1}{2}f_1\right|.
\end{aligned}$$

Define β as

$$\log\frac{1}{\beta} \equiv \sup_\theta \sup_{|x-\theta_{01}|\leq A} \left|\log \frac{1}{2}f_1\right|.$$

To find β we evaluate the maximum shifted Gaussian at the boundary of the region $\{x : |x-\theta_{01}| \leq A\}$ i.e., at $[A, 0, \ldots, 0]$ (since we take $\theta_{01} = 0$ as before). With

$$\beta = \frac{e^{-\frac{1}{2}(A+\epsilon)^2}}{2(2\pi)^{N/2}}$$

we have

$$\log\frac{1}{\beta} = \left|\log\frac{e^{-\frac{1}{2}(A+\epsilon)^2}}{2(2\pi)^{N/2}}\right| = \frac{(A+\epsilon)^2}{2} + \frac{N}{2}\log 2\pi + \log 2.$$

Now recall that $A^2 = c_4\frac{N}{\sqrt{\alpha\delta}}$ so $\log\frac{1}{\beta} \leq c_6\frac{N}{\sqrt{\alpha\delta}}$ for some positive constant c_6. We let this be M in Theorem 3.9 with $\alpha = c_1\epsilon^2 + c_0\epsilon^3$ and use $(\frac{2\epsilon c_7}{\alpha})^{2N}$ to bound the covering number there. This implies the probability that the first part of (4.7) is not true is at most $\frac{\delta}{4}$ when the number of unlabeled examples is

$$\begin{aligned} n &\geq \frac{64c_6N^2}{\alpha^3\delta}\left(N\log\frac{c_7\epsilon}{\alpha}+\log\frac{c_8}{\delta}\right) \\ &\geq \frac{c_9N^2}{\epsilon^6\delta}\left(N\log\frac{c_{10}}{\epsilon}+\log\frac{c_8}{\delta}\right) \end{aligned} \tag{4.9}$$

or simply when

$$n \geq \frac{c_{11}N^2}{\epsilon^6\delta}\left(N\log\frac{1}{\epsilon}+\log\frac{1}{\delta}\right).$$

With this many unlabeled examples, the probability of the first term of (4.7) is at most $\delta/4$. We have already bounded the second term of (4.7) by $\delta/4$ therefore this unlabeled sample complexity guarantees that (4.3) holds, and hence that there exists a maximum of $L(\theta)$ inside the ball $B(\theta_0, \epsilon)$ with probability $> 1 - \delta/2$.

4.3.2 Global Maximum of the Likelihood Function

Now we analyze the conditions needed to ensure that the global maximum of $L(\theta)$ is ϵ-close to θ_0. We have established above that there exists a maximum of $L(\theta)$ inside the closed ball $B(\theta_0, \epsilon)$ i.e., there exist some $\theta_a \in B(\theta_0, \epsilon)$ such that $L(\theta_a) \geq L(\theta_0)$. It remains to guarantee that for all $\theta \in \Theta\backslash B(\theta_0, \epsilon)$, $L(\theta) < L(\theta_0)$ where Θ denotes a compact region in $\mathbb{R}^{2N}$ which contains θ_0, and is the region where the learner searches for the argsup of $L(\theta)$.

There is a small notational nuisance here: in the preceding, we used the compactness of $B(\theta_0, \epsilon)$ when proving that the class $\mathcal{G}$ is finitely coverable. Now we use the notation $B(\theta_0, \epsilon)$ to mean an open ball; so that $\Theta\backslash B(\theta_0, \epsilon)$ is compact. Following the same steps as before, to guarantee to within some confidence that $L(\theta) < L(\theta_0)$ it is sufficient to have

$$\sup_{\theta\in\Theta\backslash B(\theta_0,\epsilon)}\left|\frac{1}{n}\sum\log\frac{f(x_i|\theta)}{f(x_i|\theta_0)} - \mathbf{E}\log\frac{f(x|\theta)}{f(x|\theta_0)}\right| \leq \alpha(\epsilon) \tag{4.10}$$

where

$$\alpha(\epsilon) \equiv \inf_{\theta \in \Theta \backslash B(\theta_0,\epsilon)} \left| \mathbf{E} \log \frac{f(x|\theta)}{f(x|\theta_0)} \right|.$$

We first redefine the class $\mathcal{G} = \{\log f(x|\theta) 1_D(x) : \theta \in \Theta\backslash B(\theta_0, \epsilon)\}$. Then we estimate the value for α, and in particular, analyze its dependence on ϵ and the dimension N. This is in analogy to our previous findings for α used for the sample complexity calculation for guaranteeing the consistency of a relative maximum of $L(\theta)$; we had earlier found the estimate $\alpha(\epsilon) = O(\epsilon^2)$ independent of N. Then it remains to find the covering number of the region $\Theta\backslash B(\theta_0, \epsilon)$ and use it to calculate the sample complexity needed such that (4.10) holds. As before, the covering number is still bounded by an exponential in N and the analysis leading to (4.9) still holds; we construct a $f_\sigma(x)$ and $\sigma = C$, C is constant w.r.t. N, ϵ. We only need to replace $\log \beta$ by $\log \beta_\Theta$ in the argument prior to (4.9), where β_Θ is the smallest value that any mixture $f(x|\theta)$ can take over the domain D when $\theta \in \Theta$; this is smaller than β but still, $\log \frac{1}{\beta_\Theta} = CN/\sqrt{\alpha\delta}$ for some constant $C > 0$.

We proceed to estimate $\alpha(\epsilon)$ anew once again for the new domain. Denote $\Phi(\theta) \equiv \mathbf{E} \log \frac{f(x|\theta)}{f(x|\theta_0)}$. For any given $\epsilon > 0$, using our new notation for the open ball, the region $\Theta\backslash B(\theta_0, \epsilon)$ is compact. The function Φ is continuous over the region hence Φ achieves its maximum value, which must differ from 0 (shown earlier) hence is strictly < 0 whence it follows that $\alpha(\epsilon) > 0$. To verify that $\Phi(\theta)$ is continuous over this region, write

$$\Phi(\theta) \;=\; \int f(x|\theta_0) \log f(x|\theta)\, dx - \int f(x|\theta_0) \log f(x|\theta_0)\, dx.$$

Let $\{\theta_n, n \geq 1\}$ be a sequence convergent to a point $\theta_a \in \Theta\backslash B(\theta_0, \epsilon)$ and such that for a constant $\rho > 0$, $|\theta_n - \theta_a| \leq \rho$ for each n. It suffices to show that the first term, denoted by $g(\theta)$ is continuous. It suffices to show

$$\lim_{n\to\infty} \int f(x|\theta_0) \log f(x|\theta_n)\, dx = \int f(x|\theta_0) \lim_{n\to\infty} \log f(x|\theta_n)\, dx = g(\theta_a). \qquad (4.11)$$

To justify the exchange of limit and integral, note that $\log f(x|\theta_n)$ is continuous and

$$\begin{aligned}
& f(x|\theta_0) \log f(x|\theta_n) \\
\leq \;& f(x|\theta_0)\, |\log f(x|\theta_n)| \leq f(x|\theta_0) \left|\log \frac{1}{2} f_1(x|\theta_{n1})\right| \leq C_1 f(x|\theta_0) + \frac{1}{2} f(x|\theta_0)|x - \theta_{n1}|^2 \\
\leq \;& C_1 f(x|\theta_0) + \frac{1}{2} f(x|\theta_0)|x|^2 + f(x|\theta_0)|x|C_2 + \frac{1}{2} f(x|\theta_0) C_2^2
\end{aligned}$$

where C_1, C_2 are constants and $|\theta_n| \leq C_2$ for all $\theta_n \in B(\theta_a, \rho)$. The integral of the right side is bounded by some finite constant w.r.t n (we partition the integral and then bound the noncompact part similar to page 64) hence the Lebesgue dominated convergence theorem permits the exchange of limit and integral in (4.11).

Now we show that $\alpha(\epsilon)$ does not depend on N. (In the following we will ignore the constant $\frac{1}{2}$ for clarity). We first split $\Phi(\theta)$ as follows:

$$\begin{aligned}
\Phi(\theta) \quad = \quad & \int \frac{1}{(2\pi)^{N/2}} (e^{-\frac{1}{2}|x-\theta_{01}|^2} + e^{-\frac{1}{2}|x-\theta_{02}|^2}) \log \frac{(e^{-\frac{1}{2}|x-\theta_1|^2} + e^{-\frac{1}{2}|x-\theta_2|^2})}{(e^{-\frac{1}{2}|x-\theta_{01}|^2} + e^{-\frac{1}{2}|x-\theta_{02}|^2})}\, dx \\
= \quad & \int \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}|x-\theta_{01}|^2} \log(e^{-\frac{1}{2}|x-\theta_1|^2} + e^{-\frac{1}{2}|x-\theta_2|^2})\, dx \\
+ \quad & \int \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}|x-\theta_{02}|^2} \log(e^{-\frac{1}{2}|x-\theta_1|^2} + e^{-\frac{1}{2}|x-\theta_2|^2})\, dx \\
- \quad & \int \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}|x-\theta_{01}|^2} \log(e^{-\frac{1}{2}|x-\theta_{01}|^2} + e^{-\frac{1}{2}|x-\theta_{02}|^2})\, dx \\
- \quad & \int \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}|x-\theta_{02}|^2} \log(e^{-\frac{1}{2}|x-\theta_{01}|^2} + e^{-\frac{1}{2}|x-\theta_{02}|^2})\, dx \\
= \quad & \int \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}|x-\theta_{01}|^2} \log e^{-\frac{1}{2}|x-\theta_1|^2}\, dx \\
+ \quad & \int \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}|x-\theta_{01}|^2} \log(1 + e^{\frac{1}{2}|x-\theta_1|^2 - \frac{1}{2}|x-\theta_2|^2})\, dx \\
+ \quad & \int \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}|x-\theta_{02}|^2} \log e^{-\frac{1}{2}|x-\theta_2|^2}\, dx \\
+ \quad & \int \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}|x-\theta_{02}|^2} \log(1 + e^{\frac{1}{2}|x-\theta_2|^2 - \frac{1}{2}|x-\theta_1|^2})\, dx \\
- \quad & \int \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}|x-\theta_{01}|^2} \log e^{-\frac{1}{2}|x-\theta_{01}|^2}\, dx \\
- \quad & \int \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}|x-\theta_{01}|^2} \log(1 + e^{\frac{1}{2}|x-\theta_{01}|^2 - \frac{1}{2}|x-\theta_{02}|^2})\, dx
\end{aligned}$$

$$- \int \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}|x-\theta_{02}|^2} \log e^{-\frac{1}{2}|x-\theta_{02}|^2}\, dx$$
$$- \int \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}|x-\theta_{02}|^2} \log(1 + e^{\frac{1}{2}|x-\theta_{02}|^2 - \frac{1}{2}|x-\theta_{01}|^2})\, dx$$

Four of the above terms sum to:

$$-\frac{1}{2}|\theta_{01} - \theta_1|^2 - \frac{N}{2} - \frac{1}{2}|\theta_{02} - \theta_2|^2 - \frac{N}{2} + \frac{N}{2} + \frac{N}{2} = -\frac{1}{2}|\theta_{01} - \theta_1|^2 - \frac{1}{2}|\theta_{02} - \theta_2|^2.$$

The other four terms (those with the $\log(1 + e^{(\cdot)})$ we denote by $I_1 + I_2 - I_3 - I_4$. We manipulate I_1 as follows:

$$I_1 = \int \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}|x-\theta_{01}|^2} \log(1 + e^{\frac{1}{2}|x-\theta_1|^2 - \frac{1}{2}|x-\theta_2|^2})\, dx$$

$$= \int \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}|y+\theta_1-\theta_{01}|^2} \log(1 + e^{\frac{1}{2}|y|^2 - \frac{1}{2}|y+\theta_1-\theta_2|^2}) dy$$

where we simply changed to the y-coordinate frame whose origin is at the point θ_1. Now rotate the coordinate frame to a new primed frame as $y' = Qy$ where Q is a unitary matrix chosen so that the y'_1-axis goes through the point θ_2 and the (y'_1, y'_2)-plane goes through θ_{01}. So we have $y = Q^T y'$ and the inverse Jacobian is just 1 since the determinant of Q is 1. Thus

$$I_1 = \int \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}|Q^T y' + \theta_1 - \theta_{01}|^2} \log(1 + e^{\frac{1}{2}|Q^T y'|^2 - \frac{1}{2}|Q^T y' + \theta_1 - \theta_2|^2})\, dy'$$

$$= \int \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}|Q^T y' + Q^T(\theta_1 - \theta_{01})'|^2} \log(1 + e^{\frac{1}{2}|Q^T y'|^2 - \frac{1}{2}|Q^T y' + Q^T(\theta_1 - \theta_2)'|^2})\, dy'$$

$$= \int \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}|Q^T(y' + (\theta_1 - \theta_{01})')|^2} \log(1 + e^{\frac{1}{2}|Q^T y'|^2 - \frac{1}{2}|Q^T(y' + (\theta_1 - \theta_2)')|^2})\, dy'$$

where $(\theta_{01} - \theta_1)'$ is a coordinate vector w.r.t the primed frame. Observe that

$$|Q^T y'|^2 = y'^T Q Q^T y' = (y', y) = |y'|^2$$

and

$$
\begin{aligned}
|Q^T(y' + (\theta_1 - \theta_{01})')|^2 &= (y' + (\theta_1 - \theta_{01})')^T Q Q^T (y' + (\theta_1 - \theta_{01})') \\
&= (y' + (\theta_1 - \theta_{01})', y' + (\theta_1 - \theta_{01})') = |y' + (\theta_1 - \theta_{01})'|^2.
\end{aligned}
$$

The integral can hence be written in the form

$$
\int \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}|y'+\theta_1'-\theta_{01}'|^2} \log(1 + e^{\frac{1}{2}|y'|^2 - \frac{1}{2}|y'+\theta_1'-\theta_2'|^2})\, dy'
$$

$$
\begin{aligned}
= \quad & \int \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}|y'|^2 - \frac{1}{2}|\theta_1'-\theta_{01}'|^2 + |\theta_1'-\theta_{01}'|(y', \frac{\theta_{01}'-\theta_1'}{|\theta_{01}'-\theta_1'|})} \\
\times \quad & \log\left(1 + e^{-\frac{1}{2}|\theta_1'-\theta_2'|^2 + |\theta_2'-\theta_1'|(y', \frac{\theta_2'-\theta_1'}{|\theta_2'-\theta_1'|})}\right)\, dy'.
\end{aligned}
$$

The inner-product $(y', \frac{\theta_2'-\theta_1'}{|\theta_2'-\theta_1'|})$ is the size of the projection of the N-dimensional vector y' on the y_1' axis because we chose the primed frame so that the vector $\theta_2' - \theta_1'$ is on the y_1' axis. Hence this equals the first component of y' i.e. y_1'. The inner-product $(y', \frac{\theta_{01}'-\theta_1'}{|\theta_{01}'-\theta_1'|})$ is the size of the projection of y' on the vector $\theta_{01}' - \theta_1'$ which is on the y_1', y_2' plane hence it must be a function only of the first two elements of the vectors y' and $\frac{\theta_{01}'-\theta_1'}{|\theta_{01}'-\theta_1'|}$, i.e., $y_1', y_2', < \frac{\theta_{01}'-\theta_1'}{|\theta_{01}'-\theta_1'|} >_1$ and $< \frac{\theta_{01}'-\theta_1'}{|\theta_{01}'-\theta_1'|} >_2$; denote it by $g(y_1', y_2')$. (Here the notation $< \cdot >_i$ means the i^{th} element of the vector.) For clarity we rename y', which is just a variable of integration, to x. Now we transform to cylindrical coordinate system:

$$
\begin{aligned}
&x_1 = x_1 \\
&x_2 = x_2 \\
&x_3 = r \cos\phi_1 \cos\phi_2 \ldots \cos\phi_{N-3} \\
&x_4 = r \sin\phi_1 \cos\phi_2 \ldots \cos\phi_{N-3} \\
&\ldots \\
&x_N = r \sin\phi_{N-3}.
\end{aligned}
$$

For example, for $N = 2, 3$ the Jacobian is 1. For $N = 4$ the inverse of the Jacobian is r. For $N = 5$ the inverse Jacobian is:

$$\begin{vmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -r\sin\phi_1\cos\phi_2 & -r\cos\phi_1\sin\phi_2 & \cos\phi_1\cos\phi_2 \\ 0 & 0 & r\cos\phi_1\cos\phi_2 & -r\sin\phi_1\sin\phi_2 & \sin\phi_1\cos\phi_2 \\ 0 & 0 & 0 & r\cos\phi_2 & \sin\phi_2 \end{vmatrix}$$
$$= \frac{1}{\sin\phi_1}\begin{vmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -r\sin\phi_1\cos\phi_2 & 0 & 0 \\ 0 & 0 & 0 & -r\sin\phi_2 & \cos\phi_2 \\ 0 & 0 & 0 & r\cos\phi_2 & \sin\phi_2 \end{vmatrix}$$
$$= -(r\cos\phi_2)(r).$$

For $N = 6$, the inverse Jacobian equals $-(r\cos\phi_2\cos\phi_3)(r\cos\phi_3)(r)$. In general, for $N \geq 4$ the Jacobian evaluates to $r^{N-3}\cos^{N-4}\phi_{N-3}\dots\cos\phi_2$. The variables range over values $0 \leq r \leq \infty$, $0 \leq \phi_1 \leq 2\pi$, $-\pi/2 \leq \phi_k \leq \pi/2$, $2 \leq k \leq N-3$. It is easy to see that the transformation is globally invertible. Also we have $|x|^2 = r^2 + {x_1}^2 + {x_2}^2$. So the integral becomes:

$$\begin{aligned} I_1 \quad = \quad & \int_{-\pi/2}^{\pi/2} \cos^{N-4}\phi_{N-3} \cdots \int_{-\pi/2}^{\pi/2} \cos\phi_2 \int_0^{2\pi} \int_0^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{N/2}} r^{N-3} \\ \times \quad & e^{-\frac{1}{2}(r^2+x_1^2+x_2^2)-\frac{1}{2}|\theta_1'-\theta_{01}'|^2+|\theta_1'-\theta_{01}'|g(x_1,x_2)} \\ \times \quad & \log(1+e^{-\frac{1}{2}|\theta_1'-\theta_2'|^2-|\theta_1'-\theta_2'|x_1})\,dx_1\,dx_2 dr d\phi_1 d\phi_2 \dots d\phi_{N-3}. \end{aligned}$$

The integrand does not depend on $\phi_1, \dots, \phi_{N-3}$ hence we can write $I_1 = G(N) \cdot I_1^*$ where

$$G(N) = \int_{-\pi/2}^{\pi/2} \cos^{N-4}\phi_{N-3} d\phi_{N-3} \cdots \int_{-\pi/2}^{\pi/2} \cos\phi_2 d\phi_2 \int_0^{2\pi} d\phi_1$$

and

$$\begin{aligned} I_1^* \quad = \quad & \int_0^{\infty} r^{N-3} \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}r^2}\,dr \quad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x_1^2+x_2^2)-\frac{1}{2}|\theta_{01}'-\theta_1'|^2+|\theta_{01}'-\theta_1'|g(x_1,x_2)} \\ \times \quad & \log\left(1+e^{-\frac{1}{2}|\theta_2'-\theta_1'|^2+|\theta_2'-\theta_1'|x_1}\right)\,dx_1\,dx_2. \end{aligned}$$

To evaluate $G(N)$, start with the identity

$$\int \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}|x|^2}\,dx = 1.$$

Passing to the cylindrical frame we obtain the identity

$$G(N)\int_0^\infty r^{N-3}\frac{1}{(2\pi)^{N/2}}e^{-\frac{1}{2}r^2}dr\int_{-\infty}^\infty\int_{-\infty}^\infty e^{-\frac{1}{2}(x_1^2+x_2^2)}\,dx_1\,dx_2 = 1,$$

whence

$$\begin{aligned} G(N) &= \left(\int_0^\infty r^{N-3}\frac{1}{(2\pi)^{N/2}}e^{-\frac{1}{2}r^2}dr\int_{-\infty}^\infty\int_{-\infty}^\infty e^{-\frac{1}{2}(x_1^2+x_2^2)}\,dx_1\,dx_2\right)^{-1} \\ &= \frac{\pi^{\frac{N}{2}-1}}{\Gamma(\frac{N}{2})} \sim \frac{1}{2\pi^{3/2}}\left(\frac{2\pi e}{N}\right)^{N/2} \qquad (N\to\infty). \end{aligned} \tag{4.12}$$

From (4.12) we obtain

$$\begin{aligned} I_1 &= G(N)I_1^* \\ &= \frac{\int_{-\infty}^\infty\int_{-\infty}^\infty e^{-\frac{1}{2}(x_1^2+x_2^2)-\frac{1}{2}|\theta'_{01}-\theta'_1|^2+|\theta'_{01}-\theta'_1|g(x_1,x_2)}\log(1+e^{-\frac{1}{2}|\theta'_2-\theta'_1|^2+|\theta'_2-\theta'_1|x_1})\,dx_1\,dx_2}{\int_{-\infty}^\infty\int_{-\infty}^\infty e^{-\frac{1}{2}(x_1^2+x_2^2)}\,dx_1\,dx_2}. \end{aligned}$$

From this we see that I_1 depends on $\theta_1, \theta_2, \theta_{01}$ only through the transformed-coordinate vector quantities: $|\theta'_{01}-\theta'_1|$, $|\theta'_2-\theta'_1|$, $< \frac{\theta'_{01}-\theta'_1}{|\theta'_{01}-\theta'_1|} >_1$, and $< \frac{\theta'_{01}-\theta'_1}{|\theta'_{01}-\theta'_1|} >_2$. Similarly, I_2 depends only on $|\theta'_{02}-\theta'_2|$, $|\theta'_1-\theta'_2|$, $< \frac{\theta'_{02}-\theta'_2}{|\theta'_{02}-\theta'_2|} >_1$, and $< \frac{\theta'_{02}-\theta'_2}{|\theta'_{02}-\theta'_2|} >_2$. For I_3 we have:

$$\begin{aligned} I_3 &= \int\frac{1}{(2\pi)^{N/2}}e^{-\frac{1}{2}|x-\theta_{01}|^2}\log(1+e^{\frac{1}{2}|x-\theta_{01}|^2-\frac{1}{2}|x-\theta_{02}|^2})\,dx \\ &= \int\frac{1}{(2\pi)^{N/2}}e^{-\frac{1}{2}|x|^2}\log(1+e^{\frac{1}{2}|x|^2-\frac{1}{2}|x+\theta'_{01}-\theta'_{02}|^2})\,dx \\ &= \int\frac{1}{(2\pi)^{N/2}}e^{-\frac{1}{2}|x|^2}\log(1+e^{-\frac{1}{2}|\theta'_{02}-\theta'_{01}|^2+2|\theta'_{02}-\theta'_{01}|(x,\frac{\theta'_{02}-\theta'_{01}}{|\theta'_{02}-\theta'_{01}|})})\,dx \end{aligned}$$

where we chose the primed frame such that θ_{01} is the origin and the x_1 axis through θ_{02} (we did not bother writing y' but kept x since it is only a variable of integration). Hence $(x, \frac{\theta'_{02}-\theta'_{01}}{|\theta'_{02}-\theta'_{01}|})$ is the projection of x on the x_1-axis which therefore equals x_1. Using the identical transformation again

$$\begin{aligned} I_3 &= G(N)\int_0^\infty r^{N-3}\frac{1}{(2\pi)^{N/2}}e^{-\frac{1}{2}r^2}dr \\ &\times \int_{-\infty}^\infty\int_{-\infty}^\infty e^{-\frac{1}{2}(x_1^2+x_2^2)}\log(1+e^{-\frac{1}{2}|\theta'_{02}-\theta'_{01}|^2+|\theta'_{02}-\theta'_{01}|x_1})\,dx_1\,dx_2. \end{aligned}$$

So I_3 depends on $|\theta'_{02} - \theta'_{01}|$. Similarly for I_4.

Now, given any four N-dimensional points $\theta_1, \theta_2, \theta_{01}, \theta_{02}$ we can apply the following procedure to form a vector v:

- Let θ_1 be the origin.
- Choose a primed-N-dimensional frame s.t. θ_2 lies on the y'_1-axis and θ_{01} on the y'_1, y'_2-plane.
- Record the values

$$v_1 \equiv |\theta'_{01} - \theta'_1|, v_2 \equiv |\theta'_2 - \theta'_1|, v_3 \equiv < \frac{\theta'_{01} - \theta'_1}{|\theta'_{01} - \theta'_1|} >_1, v_4 \equiv < \frac{\theta'_{01} - \theta'_1}{|\theta'_{01} - \theta'_1|} >_2 .$$

- Let θ_2 be the origin.
- Then choose another primed$N$-dimensional frame s.t. θ_1 lies on the y'_1-axis and θ_{02} on the y'_1, y'_2-plane.
- Record the values

$$v_5 \equiv |\theta'_{02} - \theta'_2|, v_6 \equiv < \frac{\theta'_{02} - \theta'_2}{|\theta'_{02} - \theta'_2|} >_1, v_7 \equiv < \frac{\theta'_{02} - \theta'_2}{|\theta'_{02} - \theta'_2|} >_2, v_8 \equiv |\theta'_{01} - \theta'_{02}|$$

Recall that the values of the other terms besides the $I_1, \ldots, I_4$, depended on $|\theta_{01} - \theta_{02}|$ (which equals v_8), hence it follows that $\Phi(\theta)$ is a function only of the vector v. But there exist four $3D$-vectors, $\theta_1, \theta_2, \theta_{01}, \theta_{02}$, with the same vector v. This follows since our four points in N-dimensional space lie on some $3D$-subspace. Let $N = 3$ and apply the above procedure to these points (which have $3D$ coordinate vectors w.r.t some frame in this $3D$-subspace). This must yield the same vector v because we did not disturb the points in any way. The vector v lives in some subset $\mathcal{V} \subseteq \mathbb{R}^8$ and the function Φ maps vectors $v \in \mathcal{V}$ to $\mathbb{R}^1$. In particular, take any 4 points $\theta_{01}, \theta_{02}, \theta_1, \theta_2$ in N-dimensions such that $|\theta - \theta_0| > \epsilon$ or equivalent by $|\theta_{01} - \theta_1|^2 + |\theta_{02} - \theta_2|^2 > \epsilon^2$. Then apply the procedure to get the vector v. As before, there exist 4 points, $\theta_{01}, \theta_{02}, \theta_1, \theta_2$ in 3D with the same v therefore with the same $|\theta_{01} - \theta_1|$ and $|\theta_{02} - \theta_2|$ therefore the same $|\theta - \theta_0| > \epsilon$ as the N-dimensional points. So for any $N \geq 3$, fixing $|\theta_{01} - \theta_{02}|$ results in fixing $\sup_{\theta \notin B(\theta_0, \epsilon)} \Phi(\theta)$. Recall that $\sup_{\theta \in \Theta \setminus B(\theta_0, \epsilon)} \Phi(\theta) = -\alpha(\epsilon)$. The above implies that, for a fixed $\epsilon > 0$ and $|\theta_{01} - \theta_{02}|$, $\alpha(\epsilon)$ is invariant for all $N \geq 3$. This validates the earlier claim, in the case of small $\epsilon > 0$, that $\alpha(\epsilon)$ is independent of N.

At this stage we've shown that given any $\epsilon > 0$, there exists $\alpha(\epsilon) > 0$ which when used in (4.10), yields a finite sample complexity that guarantees with some confidence, that the global maximum of $L(\theta)$ is ϵ-close to the true unknown parameter $\theta_0 \in \mathbb{R}^{2N}$. This $\alpha(\epsilon)$ is constant for all $N \geq 3$ hence its effect on the sample complexity cannot worsen (i.e., $\alpha(\epsilon)$ cannot decrease) with increasing N. But we are still left with the question of how fast $\alpha(\epsilon)$ can decrease with ϵ; if, for instance, it is decreasing as quickly as $e^{-1/\epsilon}$ then the sample complexity n will grow exponentially with the accuracy parameter ϵ.

Consider a ball $B(\theta_0, \epsilon')$ with $\epsilon' > 0$ such that (using our previous results) we have $\Phi(\theta_\epsilon) = -\frac{1}{2}c_1\epsilon^2 + c_0\epsilon^3$ for all θ_ϵ on the surface $\partial B(\theta_0, \epsilon)$, with $0 < \epsilon < \epsilon'$. As Φ is continuous it achieves its maximum value over the region $\Theta \setminus B(\theta_0, \epsilon')$ at a point θ_a (where $B(\theta_0, \epsilon')$ is an open ball). Now, let θ_b be the farthest point from θ_0 such that θ_b is in the closed ball $B(\theta_0, \epsilon')$ with $\Phi(\theta_a) \leq \Phi(\theta_b)$. ($\theta_b$ could be on $\partial B(\theta_0, \epsilon')$). By simple arguments it follows that for all θ_ϵ such that $|\theta_\epsilon - \theta_0| \leq |\theta_b - \theta_0|$, $\alpha(\epsilon) \geq |\Phi(\theta_\epsilon)|$ $= \frac{1}{2}c_1\epsilon^2 + c_0\epsilon^3$, and it is true for all $N \geq 3$. Hence for all sufficiently small $\epsilon > 0$ we can estimate $\alpha(\epsilon) = c_1\epsilon^2$ for all $N \geq 3$.

Hence we may proceed using $c_1\epsilon^2$ for $\alpha(\epsilon)$ in (4.10) and for all sufficiently small $\epsilon > 0$, the bound on the unlabeled sample complexity (4.9) not only guarantees that there exists a maximum of $L(\theta)$ ϵ-close to θ_0 but is also the global maximum, i.e. that $\hat{\theta}$ is ϵ-close to θ_0. Therefore we have $\mathbf{P}(\{|\hat{\theta}_1 - \theta_{01}| > \epsilon\} \bigcup \{|\hat{\theta}_2 - \theta_{02}| > \epsilon\}) \leq$

$\mathbf{P}(|\hat{\theta} - \theta_0| > \epsilon) \leq \frac{\delta}{2}$.

We've established above that the estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ are at most ϵ-away from θ_{01} and θ_{02} respectively. Using the analysis of Section 4.1, the classification error (under optimal labeling) can be written as $P_{error} = P_{Bayes}(1 + c_{12}\epsilon^2))$. We can replace ϵ^2 by ϵ here and in (4.9) to get that with

$$n = \frac{c_{13}N^2}{\epsilon^3\delta}\left(N \log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)$$

unlabeled examples, the P_{error} of the optimal labeling of the decision regions is $P_{Bayes}(1 + c_{12}\epsilon)$. It only remains to use labeled examples to guarantee that we pick this optimal labeling.

4.3.3 Labeling the Partition

We have two unlabeled regions separated by the hyperplane between $\hat{\theta}_1$ and $\hat{\theta}_2$ where both are ϵ-close to their respective true parameters. The good labeling has the above classification error. We use the labeled examples to control the confidence of picking the good labeling by the majority rule. We have the two regions R_1, R_2 on each side of the hyperplane. Draw m labeled examples. Assign to each region the label of the majority of the examples that fell in it. If no examples fell in R_i then label it "1" with probability $\frac{1}{2}$ and "2" with probability $\frac{1}{2}$. We now calculate m needed to guarantee we pick the good labeling. Denote $\eta_1 = \mathbf{P}(2|x \in R_1)$, $\eta_2 = \mathbf{P}(1|x \in R_2)$, and $p = \int_{R_1} f(x)\, dx$. The quantities η_1 and η_2 are the probabilities that a randomly drawn x is misclassified given it is in R_1 or R_2, respectively. Also, $p = \mathbf{P}(R_1)$ and $1 - p = \mathbf{P}(R_2)$. Let $p_{min} = \min(p, 1 - p)$ and $\eta_{max} = \max(\eta_1, \eta_2)$. Let the event $\{\text{a random } x \text{ is misclassified}\} \equiv E$. There are four possible labelings: L_{good} has R_1 labeled "1" and R_2 labeled "2"; $L_{bad,1}$ has R_1 labeled "2" and R_2 labeled "1"; $L_{bad,2}$ has R_1 and R_2 labeled "1"; $L_{bad,3}$ has R_1 and R_2 labeled "2". With the same analysis as in Section 2 with the replacement of

$$\delta \equiv 12e^{-mp_{min}\left(1-2\sqrt{\eta_{max}(1-\eta_{max})}\right)}$$

we obtain the probability of not choosing L_{good} is

$$\mathbf{P}(L_{bad,1}) + \mathbf{P}(L_{bad,2}) + \mathbf{P}(L_{bad,3}) \leq 3\left(\frac{\delta}{12} + \frac{\delta}{24} + \frac{\delta}{24}\right) = \frac{1}{2}\delta.$$

Using the analysis of the error of the decision rule in $\mathbb{R}^N$ of Section 4.1 we get that $p_{min} \geq \frac{1}{2} - c_{14}\epsilon$ and $\eta_{max} \leq P_{Bayes} + c_{15}\epsilon$. Plugging this into the exponential bound, for suitably small ϵ, we get $m = c_{16} \log \frac{1}{\delta}$ is sufficient to guarantee that we do not pick L_{good} with low probability, i.e.,

$$\mathbf{P}\left(P_{error} > P_{Bayes}(1 + c_{17}\epsilon) \middle| \{|\hat{\theta}_1 - \theta_{01}| \leq \epsilon\} \cap \{|\hat{\theta}_2 - \theta_{02}| \leq \epsilon\}\right) \leq \frac{\delta}{2}.$$

Combining this with the fact that

$$\mathbf{P}\left(\{|\hat{\theta}_1 - \theta_{01}| > \epsilon\} \cup \{|\hat{\theta}_2 - \theta_{02}| > \epsilon\}\right) \leq \mathbf{P}\left(|\hat{\theta} - \theta_0| > \epsilon\right) \leq \frac{\delta}{2}$$

completes the proof of the theorem. ▮

4.4 General *a priori* Class Probabilities

Here we extend the results of Section 4.1 to the case of general *a priori* class probabilities, p and $1-p$. The learner uses algorithm E_p with the randomly drawn labeled examples to construct the decision rule.

Algorithm E_p:

The setting: Two pattern classes with underlying Gaussian mixture density

$$f(x|[\mu_0, p]) = p\, g(x|\mu_{01}) + (1-p) g(x|\mu_{02}).$$

The teacher draws labeled examples randomly at least once according to $f(x)$ by choosing class "1" with probability p class "2" with probability $1-p$ and then drawing according to the selected class conditional density $g(x|\mu_{0i})$, $i = 1, 2$.

Given: m_1 examples labeled as "1" and m_2 examples labeled as "2", where

$$m_1 + m_2 = m > 0.$$

Begin:

1) If $m_1 = 0$ then label all of $\mathbb{R}^N$ as "2". If $m_2 = 0$ then label all of $\mathbb{R}^N$ as "1". Go to End.

2) Otherwise, continue with the following steps.

3) Let the mean estimates be

$$\hat{\mu}_i \equiv \frac{1}{m_i} \sum_{k=1}^{m_i} x_k^i \qquad (i = 1, 2)$$

where x_k^i denotes the k^{th} component of the example vector x_i.

4) Estimate p by

$$\hat{p} = \frac{1}{m} \sum_{i=1}^{m} 1_{y_i = \text{“}1\text{''}}$$

where y_i is the label of the i^{th} example.

5) Let the decision border be the hyperplane defined by

$$h(x) = (x - \hat{\mu}_1, \hat{\mu}_2 - \hat{\mu}_1) - \frac{1}{2}|\hat{\mu}_2 - \hat{\mu}_1|^2 - \log \frac{\hat{p}}{1-\hat{p}} = 0.$$

6) The classifier decides "1" for x if $h(x) < 0$ and decides "2" otherwise.

End.

The difference here, compared to the case of $p = \frac{1}{2}$, is that the Bayes decision border depends on p, i.e.,

$$h(x) = (x - \mu_{01}, \mu_{02} - \mu_{01}) - \frac{1}{2}|\mu_{02} - \mu_{01}|^2 - \log \frac{p}{1-p} = 0.$$

In this case

$$P_{Bayes} = (1-p)\Phi\left(-\frac{1}{2\Delta}\log\frac{1-p}{p} - \Delta\right) + p\,\Phi\left(\frac{1}{2\Delta}\log\frac{1-p}{p} - \Delta\right)$$

where $\Delta \equiv \frac{|\mu_{01}-\mu_{02}|}{2}$ and $\Phi()$ denotes the standard normal probability distribution. (We will use c_i to denote finite positive constants as before). So we need to estimate p by $\hat{p}$ in order to form an estimate of the Bayes decision border.

We now determine the labeled sample complexity m. W.l.o.g. we assume $p < 1-p$. Denote by A the event that the decision rule is as line (1) in the algorithm, and let A^c denote the complement. We have

$$P_{error} = \mathbf{P}\left(\text{error}|A\right)\mathbf{P}\left(A\right) + \mathbf{P}\left(\text{error}|A^c\right)\mathbf{P}\left(A^c\right).$$

Now

$$\mathbf{P}(A) = \mathbf{P}(\{m_1 = 0\} \text{ or } \{m_2 = 0\}) \leq 2(1-p)^m$$

and

$$\mathbf{P}(A^c) = \mathbf{P}(\{m_1 > 0\} \text{ and } \{m_2 > 0\}) \leq \mathbf{P}(\{m_1 > 0\}) = 1 - \mathbf{P}(m_1 = 0) = 1-(1-p)^m.$$

Now conditioned under event A, the probability of error is determined as follows: given an x labeled as "1" (with probability p) the algorithm misclassifies it only if $m_1 = 0$ which has probability $(1-p)^m$. Similarly, given an x labeled as "2" (with probability $1-p$) then the algorithm misclassifies it only if $m_2 = 0$ which has probability p^m. Using the inequality $(1-p)^m \leq e^{-mp}$ we have

$$\begin{aligned}\mathbf{P}(\text{error}|A) = (1-p)^m p + p^m(1-p) &= p\left((1-p)^m + p^{m-1}(1-p)\right) \\ &\leq p(1-p)\left((1-p)^{m-1} + p^{m-1}\right) \\ &\leq 2p(1-p)e^{-(m-1)p} \leq 2epe^{-mp}.\end{aligned}$$

So

$$\mathbf{P}\left(\text{error}|A\right)\mathbf{P}\left(A\right) \leq 4epe^{-2mp}.$$

Define the event

$$E = \{\ |\hat{\mu}_i - \mu_i| \leq \epsilon,\ i = 1, 2, \text{ and } |\hat{p} - p| \leq \epsilon\ \}.$$

We bound the term $\mathbf{P}\left(\text{error}|A^c\right)$. We have

$$\mathbf{P}\left(\text{error}|A^c\right) = \mathbf{P}(\text{error}|E)\mathbf{P}(E) + \mathbf{P}(\text{error}|E^c)\mathbf{P}(E^c) \leq \mathbf{P}(\text{error}|E) + \mathbf{P}(E^c).$$

In the rest of this section we estimate the two terms on the right.

First we determine $P(E^c)$. This is bounded above by the probability that $|\hat{p}-p| > \epsilon$ added to the probability that at least one of the mean estimates deviate by more than ϵ from the corresponding true mean.

For $\hat{p}$ we use the obvious estimate, i.e., $\hat{p} = \frac{1}{m}\sum_{i=1}^{m} 1_{y_i=\text{“}1\text{''}}$ where y_i is the label of the i^{th} example. For this we have

$$\mathbf{P}\left(|p - \hat{p}| > \epsilon_1\right) \leq 2e^{-2m\epsilon_1^2} \equiv \delta_1.$$

To estimate μ_{01} and μ_{02} we use the same ideas as in Section 4.1. This means the requirement is to have m_1 and $m_2 \geq \frac{2N}{\epsilon^2}\log\frac{4N}{\delta}$ examples from class "1" and the same number from class "2", respectively. The teacher draws labeled examples by using the mixture, i.e., first choosing class "1" according to the *a priori* probability p then drawing according to the selected class distribution.. If $\hat{p}$ is at most ϵ_1-far from p then we do not need m to be larger than $\frac{m_1}{p-\epsilon_1}$ in order to get m_1 class-"1" examples. The former has probability $> 1-\delta_1$. Similarly, m suffices to be $> \frac{m_2}{1-(p+\epsilon_1)}$ in order to get m_2 class-"2" examples with probability $> 1-\delta_1$. Equivalently, with probability $\geq 1-2\delta_1$, m needs to be at most $\frac{m_1}{p-\epsilon_1}$ to get m_1 examples of class "1" and at most $\frac{m_2}{1-(p+\epsilon_1)}$ to get m_2 examples of class "2". Clearly it suffices to take

$$m \geq \max\left\{\frac{2N}{(p-\epsilon_1)\epsilon^2}\log\frac{4N}{\delta}, \frac{2N}{(1-p-\epsilon_1)\epsilon^2}\log\frac{4N}{\delta}\right\}$$

to get m_1 class-"1" examples and m_2 class-"2" examples with probability $\geq 1-2\delta_1$. Using our assumption of $p < 1-p$, the above requirement is to have at least

$$m = \frac{2N}{(p-\epsilon_1)\epsilon^2}\log\frac{4N}{\delta}$$

which guarantees that with probability at least $1-2\delta_1$ we get the necessary m_1 and m_2 s.t. with probability at least $1-\delta$, $|\mu_{01}-\hat{\mu}_{01}| \leq \epsilon$ and $|\mu_{02}-\hat{\mu}_{02}| \leq \epsilon$ (for the last part see Section 4.1).

Therefore

$$\mathbf{P}(E^c) \leq (2\delta_1+\delta)+\delta_1 \leq 3\delta_1+\delta.$$

Now we determine $P(\text{error}|E)$ i.e., the P_{error} of the decision border based on the above ϵ-close estimates.

As in Section 4.1, $\hat{\mu}_1$ and $\hat{\mu}_2$, are ϵ-close to μ_{01} and μ_{02} respectively. The decision border is obtained by plugging the estimates into the functional form and solving for x in

$$\frac{e^{-\frac{1}{2}|x-\hat{\mu}_1|^2}}{e^{-\frac{1}{2}|x-\hat{\mu}_2|^2}} = \frac{1-\hat{p}}{\hat{p}}$$

which yields a decision border

$$h(x) = (x, \hat{\mu}_1 - \hat{\mu}_2) + \frac{|\hat{\mu}_2|^2 - |\hat{\mu}_1|^2}{2} - \log\frac{1-\hat{p}}{\hat{p}} = 0.$$

As in Section 4.1, conditioned on the high probability event that the estimates $\hat{\mu}_i$, $\hat{p}$ are ϵ-close to μ_i, p respectively, we have $h(x)$ is distributed as a one-dimensional Gaussian. For $g(x) \equiv h(x)/|\hat{\mu}_1 - \hat{\mu}_2|$ we have

$$p_1(g) \sim N\left(\frac{(\hat{\mu}_1 - \hat{\mu}_2)^T u^- \ + \ \frac{1}{2}(|\hat{\mu}_2|^2 - |\hat{\mu}_1|^2) - \log\frac{1-\hat{p}}{\hat{p}}}{|\hat{\mu}_2 - \hat{\mu}_1|}, 1\right)$$

and

$$p_2(g) \sim N\left(\frac{(\hat{\mu}_1 - \hat{\mu}_2)^T u^+ \ + \ \frac{1}{2}(|\hat{\mu}_2|^2 - |\hat{\mu}_1|^2) - \log\frac{1-\hat{p}}{\hat{p}}}{|\hat{\mu}_2 - \hat{\mu}_1|}, 1\right)$$

where u^-, u^+ are defined in Section 4.1. So we get a one dimensional decision problem which has the same P_{error} for the decision rule as the original N-dimensional problem. Considering the configuration of $\hat{\mu}_1$ and $\hat{\mu}_2$ that yields a good upper bound on P_{error}, we get

$$\begin{aligned} P_{error} \quad &\leq \quad (1-p)\Phi\left(-\Delta - \frac{1}{2\Delta}\log\frac{1-p}{p} + \Delta\epsilon + c_0\epsilon^2 + c_1\frac{\epsilon_1}{p}\right) \\ &+ \quad p\Phi\left(-\Delta + \frac{1}{2\Delta}\log\frac{1-p}{p} - \Delta\epsilon + c_2\epsilon^2 - c_3\frac{\epsilon_1}{p}\right). \end{aligned}$$

Breaking the $\Phi()$ into two parts and additional bounding gives

$$\begin{aligned} P_{error} \quad &\leq \quad (1-p)\Phi\left(-\Delta - \frac{1}{2\Delta}\log\frac{1-p}{p}\right) + p\Phi\left(-\Delta + \frac{1}{2\Delta}\log\frac{1-p}{p}\right) \\ &+ \quad c_4\left(\epsilon + \frac{\epsilon_1}{p^2}\right) \\ &= \quad P_{Bayes} + c_4\left(\epsilon + \frac{\epsilon_1}{p^2}\right). \end{aligned}$$

So

$$\mathbf{P}(\text{error}|A^c) \leq P_{Bayes} + c_4\left(\epsilon + \frac{\epsilon_1}{p^2}\right) + 3\delta_1 + \delta$$

and hence

$$P_{error} \;\leq\; \left(P_{Bayes} + c_4\left(\epsilon + \frac{\epsilon_1}{p^2}\right) + 3\delta_1 + \delta\right)(1 - (1-p)^m) + 4epe^{-2mp}.$$

So for arbitrary $0 < \alpha, \beta, \gamma$, in order to have an error

$$P_{error} \leq (P_{Bayes} + \alpha)\beta + \gamma$$

the sufficient sample complexity m is

$$m = c_5 \max\left\{\frac{N}{\alpha^2 p^4}\log\frac{N}{\alpha},\ \frac{\log\frac{1}{1-p}}{\log\frac{1}{1-\beta}},\ \frac{1}{p}\log\frac{p}{\gamma}\right\}.$$

For a prespecified P_{error}, m is polynomial in $\frac{1}{p}$, but as $p \to 0$ we can let $\alpha \to \infty$, $\beta, \gamma \to 0$, such that $m \to 1$ and $P_{error} \to 0$. (Note, we used the fact that algorithm E_p draws at least one labeled example). Now, for fixed $0 < p < \frac{1}{2}$, but increasing N or decreasing α, m grows like

$$c_6\frac{N}{\alpha^2 p^4}\log\frac{N}{\alpha}.$$

This is further discussed in Chapter 6.

4.5 Mixed Sample

In this section we use both labeled and unlabeled examples for learning the decision rule. As in Section 4.4, the Bayes decision border depends on the two means μ_{01}, μ_{02} and on the *a priori* class "1" probability p. So we need to estimate these by $\hat{\mu}_{01}$, $\hat{\mu}_{02}$ and $\hat{p}$ and use

$$h(x) = (x, \hat{\mu}_{01} - \hat{\mu}_{02}) + \frac{|\hat{\mu}_{02}|^2 - |\hat{\mu}_{01}|^2}{2} - \log\frac{1-\hat{p}}{\hat{p}} = 0.$$

as the decision border estimate.

We consider two approaches of utilizing the mixed sample: the first is based on algorithm M_1, which uses the labeled examples to estimate p by $\hat{p}$ to an accuracy ϵ_1

and confidence $> 1 - \delta_1$, and the unlabeled examples to estimate $\mu_0 = [\mu_{01}, \mu_{02}]$ to an accuracy ϵ and confidence $> 1 - \delta$ using the MLE procedure.

The second approach estimates the vector $\theta_0 = [\mu_{01}, \mu_{02}, p]$ using the MLE procedure with unlabeled examples and uses labeled examples only for labeling the decision regions. We assume w.l.o.g. that $p < 1 - p$. We start with the first method.

4.5.1 Learning using algorithm M_1

Algorithm M_1:

`The setting:` Two pattern classes with an underlying Gaussian mixture density

$$f(x|\mu_0, p) = p\, g(x|\mu_{01}) + (1-p) g(x|\mu_{02})$$

with $\mu_0 \in \mathcal{M}$ where $\mathcal{M}$ is a compact subset of $\mathbb{R}^{2N}$. The teacher draws labeled and unlabeled examples according to f by choosing class "1" with probability p, class "2" with probability $1-p$, and then drawing according to the selected class conditional density $g(x|\mu_{0i})$, $i = 1, 2$.

`Given:` First, the teacher draws $m = m_1 + m_2 > 0$ labeled examples.

`Begin:`

1) If $m_1 = 0$ then label all of $\mathbb{R}^N$ as "2". If $m_2 = 0$ then label all of $\mathbb{R}^N$ as "1". Go to End.

2) Otherwise request $n > 0$ unlabeled examples.

3) Estimate p by

$$\hat{p} = \frac{1}{m} \sum_{i=1}^{m} 1_{y_i = \text{"1"}}$$

where y_i is the label of the i^{th} example.

4) Using the unlabeled sample, estimate the mean vector $\mu_0 = [\mu_{01}, \mu_{02}]$ by the point $\hat{\mu}$,

$$\hat{\mu} = \text{argsup}_{\mu \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} \log f(x_i|\mu, \hat{p}).$$

5) Let the decision border be the hyperplane define by

$$h(x) = (x - \hat{\mu}_1, \hat{\mu}_2 - \hat{\mu}_1) - \frac{1}{2}|\hat{\mu}_2 - \hat{\mu}_1|^2 - \log\frac{\hat{p}}{1-\hat{p}} = 0.$$

6) The classifier decides "1" for x if $h(x) < 0$ and decides "2" otherwise.

End.

As in Section 4.4, we estimate $\hat{p}$ using the labeled sample to have that

$$\mathbf{P}\left(|p - \hat{p}| > \epsilon_1\right) \le 2e^{-2m\epsilon_1^2} \equiv \delta_1.$$

However now we use the unlabeled sample to estimate the means and get that the probability of either estimate being $> \epsilon$ from the true values is at most $\delta/2$ when

$$n = c_1 \frac{N^2 \log^2 \frac{1}{p}}{\epsilon^6 \delta}\left(N \log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)$$

and we select the bad labeling of the hyperplane partition with probability at most $\delta/2$ when m is at least

$$\frac{1}{c_2 p + c_3}\log\frac{1}{\delta}.$$

The analysis follows as in Section 4.4 to obtain

$$P_{error} \le \left(P_{Bayes} + c_4\left(\epsilon + \frac{\epsilon_1}{p^2}\right) + \delta_1 + \delta\right)\left(1 - (1-p)^m\right) + 4epe^{-2mp}.$$

So for arbitrary $0 < \alpha, \beta, \gamma$, in order to have an error

$$P_{error} \le (P_{Bayes} + \alpha)\beta + \gamma$$

the sufficient sample complexity m is

$$m = c_5 \max\left\{\frac{1}{\alpha^2 p^4}\log\frac{1}{\alpha},\ \frac{1}{c_6 p + 1}\log\frac{1}{\alpha},\ \frac{\log\frac{1}{1-p}}{\log\frac{1}{1-\beta}},\ \frac{1}{p}\log\frac{p}{\gamma}\right\}$$

and

$$n \ge c_7 \frac{N^3 \log^2 \frac{1}{p} \log\frac{1}{\alpha}}{\alpha^6}.$$

For a prespecified P_{error}, m is polynomial in $\frac{1}{p}$, and n is polynomial in $\log\frac{1}{p}$, For $p \to 0$, we can let $\alpha \to \infty$, and let both $\beta, \gamma \to 0$, such that $m \to 1$, $n \to 0$ and $P_{error} \to 0$ (Note that $P_{Bayes} \to 0$ as $p \to 0$, and we used the fact that the algorithm requires at least one labeled example.) For a fixed $0 < p < \frac{1}{2}$, m depends on α as

$$c_8 \frac{1}{\alpha^2} \log \frac{1}{\alpha}$$

and n depends on α, N, as

$$c_9 \frac{N^3 \log \frac{1}{\alpha}}{\alpha^6}.$$

Hence n is polynomial in N, and $\frac{1}{\alpha}$, while m only depends on α and is polynomial in $\frac{1}{\alpha}$.

This is further discussed in Chapter 6. Now consider the second approach.

4.5.2 Learning with algorithm M_2

Algorithm M_2:

`The setting:` Two pattern classes with an underlying Gaussian mixture density

$$f(x|\theta_0) = p g(x|\mu_{01}) + (1-p) g(x|\mu_{02})$$

where $\theta_0 = [\mu_{01}, \mu_{02}, p] \in \Theta$, and Θ is a compact subset of $\mathbb{R}^{2N+1}$, such that the teacher draws labeled and unlabeled examples according to f by choosing class "1" with probability p, class "2" with probability $1-p$, and then drawing according to the selected class conditional density $g(x|\mu_{0i})$, $i = 1, 2$.

`Given:` The teacher draws $m = m_1 + m_2 > 0$ labeled examples.

```
Begin:  1) If $m_1 = 0$ then label all of $\mathbb{R}^N$ as "2". If $m_2 = 0$ then label all of $\mathbb{R}^N$
           as "1". Go to End.
        2) Otherwise request $n > 0$ unlabeled examples.
        3) Estimate $\theta_0$ by
               $\hat{\theta} = \text{argsup}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \log f(x_i|\theta).$
        4) Let the decision border be the hyperplane defined by
               $h(x) = (x - \hat{\mu}_1, \hat{\mu}_2 - \hat{\mu}_1) - \frac{1}{2}|\hat{\mu}_2 - \hat{\mu}_1|^2 - \log \frac{\hat{p}}{1 - \hat{p}} = 0.$
        5) Label the two decision regions across the hyperplane by the label of
           the majority of the examples in each region.
End.
```

We proceed as in the previous two sections, except now the unlabeled examples are used to estimate the means and p, while the labeled examples are just used for picking the labeling. Then with algorithm M_2 it suffices to draw

$$n = \frac{c_1 N^2 \log^2 \frac{1}{p}}{\epsilon^6 \delta} \left(N \log \frac{1}{\epsilon} + \log \frac{1}{\delta} \right)$$

unlabeled examples to yield estimates $\hat{p}$, $\hat{\theta}_1$, $\hat{\theta}_2$, each ϵ-close to its corresponding unknown parameter. (c_1 is larger than c_1 in the expression for n in Section 4.5.1.) The decision rule based on these estimates has

$$P_{error} \leq P_{Bayes} + c_2 \epsilon \left(1 + \frac{1}{p^2} \right)$$

when labeled with the optimal labeling which is picked with confidence $> 1 - \delta$ when at least

$$m = \frac{c_3}{c_4 p + 1} \log \frac{1}{\delta}$$

labeled examples are drawn.

We obtain that for

$$P_{error} \leq (P_{Bayes} + \alpha)\beta + \gamma$$

the sufficient sample complexity m is

$$m = c_5 \max\left\{\frac{1}{c_4 p + 1}\log\frac{1}{\alpha}, \frac{\log\frac{1}{1-p}}{\log\frac{1}{1-\beta}}, \frac{1}{p}\log\frac{p}{\gamma}\right\}$$

and the sufficient n is

$$n = c_6 \frac{N^3 \log^2 \frac{1}{p} \log \frac{1}{\alpha}}{\alpha^7 p^{14}}$$

for arbitrary $0 < \alpha, \beta, \gamma$. For a prespecified P_{error}, m is practically uneffected by p, and n is a polynomial in $\frac{1}{p}$. However, as in the preceding sections, for $p \rightarrow 0$ we can let $\alpha \rightarrow \infty$, and let both $\beta, \gamma \rightarrow 0$, such that $m \rightarrow 1$, $n \rightarrow 0$, and $P_{error} \rightarrow 0$ (where we used the fact that algorithm M_2 draws at least one labeled example). For fixed $0 < p < \frac{1}{2}$, but variable N and α, m is constant, while

$$n = c_7 \frac{N^3 \log \frac{1}{\alpha}}{\alpha^7}.$$

So n is effected by N, growing polynomial in N, and $\frac{1}{\alpha}$.

This is further discussed in Chapter 6.

Chapter 5

Nonparametric Scenario

In this chapter we study another approach to learning classification of two pattern classes. The approach is called Kernel estimation—a non parametric approach which assumes very little knowledge about the form of the distribution. This method can be used with both a purely labeled or a mixed scenario. When learning with a purely labeled sample one would estimate each of the nonparametric class conditional densities and then use the estimates for defining an estimate decision rule using the Bayesian approach (Chapter 1). This falls under the field of Nonparametric Discriminant Analysis (cf. Silverman [40]). Theoretical studies of nonparametric discrimination techniques indicate that they yield asymptotically optimal decision rules (cf. Silverman [40]).

We are interested in using the kernel technique with the mixed sample scenario. In the last chapter, we found the unlabeled and labeled sample complexities of learning a Gaussian mixture. It is interesting to ask what is the complexity of learning the problem, based on the same underlying mixture, with the nonparametric kernel estimation. In this scenario the learner does not utilize the information that he had in Chapter 4, namely that f can be indexed by a finite real vector and that the decision rule can be determined from the class conditional densities which can be inferred uniquely from the mixture estimate.

So investigating the complexity of a nonparametric method for such a family shows us how much more complexity, in terms of sample sizes, is needed when, for example, parametric side information is not available. We then show how to extend this approach to a large nonparametric class of distributions which includes the Gaussian mixtures.

One way to try to use the Kernel method with unlabeled examples is by limiting consideration to a family of problems whose underlying mixture $f(x)$ is identifiable (not necessarily parametric) (cf. Cover & Castelli [5]). The family is chosen so that given $f(x)$, it is possible to uniquely determine its components, $f_1(x)$, $f_2(x)$, and the *a priori* probabilities p_1, p_2. The mixed sample can be used by some nonparametric method, say kernel estimation, to estimate $f(x)$ by $f_n(x)$ to within accuracy ϵ uniformly over x. Then, there exists an identifiable function $\hat{f}(x)$, such that $|\hat{f}(x) - f(x)| \leq \epsilon$. By careful selection of the functions that are elements of this family, it may be possible to have the latter imply that the two corresponding components are close, i.e., $|\hat{p}_i \hat{f}_i - p_i f_i| \leq 2\epsilon$, $i = 1, 2$. Then construct the decision rule based on $\hat{p}_1$, $\hat{p}_2$, $\hat{f}_1$, $\hat{f}_2$ which has P_{error} close to P_{Bayes}. The goal would be to try to match the richness of this family with the power of the estimation technique, e.g., the kernel method has only a few restrictions on the types of f that can be estimated and therefore it can handle a very rich family of functions. This way the large sample complexities (which we expect for a powerful estimation technique) will be justifiable for learning the family of functions that we defined above.

One difficulty in this approach is in finding $\hat{f}$ from f_n, especially if the identifiable family of mixtures is nonparametric. This translates into difficulty in finding the decision rule estimate. While in principle (with an appeal to the continuum hypothesis) it may be possible to order all functions in this family and then search for any function that is ϵ close to f_n (we know that there exist at least one, namely the

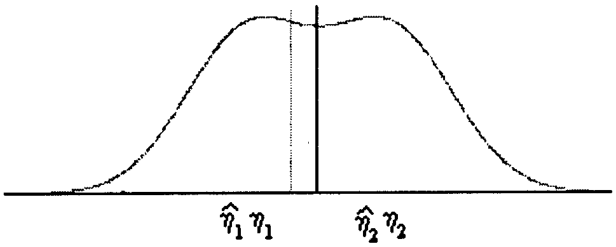


Figure 5.1:

unknown true mixture f), it is not clear that a practical polynomial-time algorithm can be constructed to find such an $\hat{f}$.

Consider for a moment a Gaussian mixture (Figure 5.1). It appears that the modes (i.e., the two global maxima) of this one-dimensional mixture may determine the Bayes border. In Lemma 5.5 we show that this is true for the N-dimensional mixture, while the condition that the means of the class conditional densities are a certain distance apart. This suggests that there may be another approach using the kernel technique for constructing a classifier. There may be an algorithm that first estimates the mixture f by $f_n(x)$ (where the subscript n shows the dependence on the sample) to within ϵ-accuracy, then determines consistent estimates, $\hat{\eta}_1$, $\hat{\eta}_2$, of the modes η_1, η_2, of f using $f_n(x)$. This method would be categorized as pseudo-direct because it skips the estimation of the class conditional densities, however it still uses density estimation for the mixture. Of course, this approach can be used for a large generic class of nonparametric problems whose Bayes border is determined by modes. We pursue this approach here.

To begin with, we will consider an algorithm, called *algorithm K*, which follows the above intuition, and learns the classification problem with an underlying Gaussian mixture. We then present a description of a rich nonparametric family of mixtures that contains the Gaussian mixture, in addition to other types of mixtures which algorithm K can handle. For every mixture f in this family, the Bayes border is a linear hyperplane and is identified by the modes of f.

In this nonparametric scenario, if we view the problem of finding the best estimate $f_n(x)$ for $f(x)$ as a learning problem in the generalized framework of PAC (cf. [12]) then we expect that a larger sample is needed to learn $f(x)$ (compared to the parametric scenario of Chapter 4) since the class of functions of which $f_n(x)$, $f(x)$ are members is significantly richer. Our analysis of kernel estimation uses the principle of the uniform SLLN differently from the PAC framework, i.e., there is no overall empirical loss which is minimized in order to learn $f(x)$. Nonetheless a quantity called the VC-dimension (see Chapter 3) which bounds the covering number through the expression of (3.8), will emerge as a clear indication of the complexity that is involved in learning $f(x)$ when no parametric knowledge is assumed about its form. In Chapter 4 we saw that the complexity of the problem of learning a Gaussian mixture as an element of the parametric class of Gaussian mixtures was reflected in the covering number of a related class (the class $\mathcal{G}$ on page 63). So the covering number plays an important role in realizing the difference in complexities of learning a particular function $f(x)$ both in the parametric and nonparametric scenarios.

The results Chapter 5 indicate that for the nonparametric scenario a sufficient unlabeled sample for learning the Gaussian mixture problem is significantly larger than in the parametric scenario of Chapter 4. The labeled sample size is the same, hence the value of a labeled example is significantly higher when less side information is at hand. This is discussed further in Chapter 6 where we discourse on the measurement of this value.

The material of this chapter is organized as follows: In Section 5.1 we provide a review of kernel estimation, which contains the important points that we need in later sections. In Section 5.2 we investigate learning the classification problem based on a Gaussian mixture, using algorithm K. The main result of this section is embodied in Theorem 5.1 which gives the mixed sample complexity for learning with algorithm K. Before proving the theorem in Section 5.4 we state several lemmas in Section 5.3.

In Section 5.5 we consider learning using algorithm K (with the same sample complexities stated in Theorem 5.1) a classification problem which is based on pattern class densities whose mixture f is of a more general type. We describe the nonparametric family, and then prove the consistency of the mode estimates which are constructed by algorithm K.

In Section 5.6 we present an alternate nonparametric technique— learning vector quantization in neural networks, which can use unlabeled examples and labeled examples in learning a classification rule by exploiting clustering. In Section 5.7 we analyze a learning-classification approach, called *k-means*, which is based on minimizing the empirical MSE of the voronoi partition on which the classifier is based. We find the labeled and unlabeled sample complexities for learning a general decision problem having well-separated pattern classes.

We now discuss the principle ideas behind the Kernel technique (cf. Silverman [40], Duda & Hart [1]).

5.1 Kernel Density Estimation: A Review

The naive one-dimensional kernel density estimate is based on the idea of placing a σ-scaled version of the window function

$$w(x) = \begin{cases} 1 & |x| \leq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

around the test point x then counting the number of examples x_i that fall inside the window, and normalizing by the total number n of examples and the window width

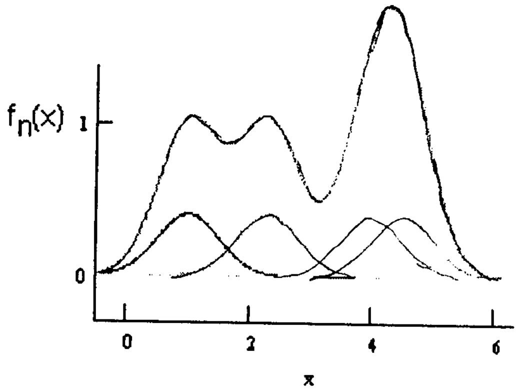


Figure 5.2:

σ. The value of the estimate $f_n(x)$ at x is therefore expressed as

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sigma}w\left(\frac{x-x_i}{\sigma}\right).$$

This estimate, however, does not possess some important properties, such as smoothness, that will be discussed later. We hence introduce a more general estimate which depends on a function $K(x)$ called the kernel, that satisfies

$$\int_{-\infty}^{\infty} K(x)\,dx = 1.$$

That is,

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sigma}K\left(\frac{x-x_i}{\sigma}\right).$$

The estimate $f_n(x)$ is still defined using the same equation as for the window function. However now it can be viewed as a sum of "humps" centered at the examples x_i (see Figure 5.2). The kernel function K determines the shape of the humps and the parameter σ determines their effective width.

As σ goes to 0 the estimate is a sum of delta functions at the examples. Such an estimate does not interpolate the data at all. At the other extreme, as σ tends to infinity, all the humps overlap and sum up to a very smooth function that hides all the high frequency detail of the underlying density.

There are several ways to measure the goodness of the kernel estimator. Viewing $f_n(x)$ as a random variable because of its dependence on the random sample, admits the mean square error (MSE) as a measure of error, i.e.,

$$\mathbf{E}\left(f_n(x) - f(x)\right)^2$$

with expectation w.r.t. the joint density of the examples $x_1, \ldots, x_n$. (Here x is nonrandom and all the randomness resides in the examples.) The MSE can be written as a sum of two terms:

$$\left(\mathbf{E}f_n(x) - f(x)\right)^2 + \operatorname{var}\left(f_n(x)\right) = \operatorname{bias}^2\left(f_n(x)\right) + \operatorname{var}\left(f_n(x)\right).$$

We could also view $f_n(x)$ as a regular function and hence use the sup norm as a measure of discrepancy,

$$\sup_x |f_n(x) - f(x)|$$

which itself is a random variable (as f_n is random). So one may define the error measure as

$$\mathbf{P}\left(\sup_x |f_n(x) - f(x)| > \epsilon\right).$$

The event

$$\left\{\sup_x |f_n(x) - f(x)| > \epsilon\right\}$$

implies that

$$\left\{\sup_x \left|f_n(x) - \bar{f}(x)\right| > \frac{\epsilon}{2}\right\} \text{ or } \left\{\sup_x \left|\bar{f}(x) - f(x)\right| > \frac{\epsilon}{2}\right\}$$

where

$$\bar{f}(x) = \mathbf{E}f_n(x).$$

The first term is analogous to the variance and corresponds to a random event. The second term is the magnitude of the bias of $f_n(x)$ and is a deterministic event. For a good estimate we demand that the probability of the event

$$\sup_x |f_n(x) - f(x)| > \epsilon$$

be less than δ where $\delta > 0$ is chosen suitably small. This amounts to demanding that the bias term

$$\sup_x \left|\bar{f}(x) - f(x)\right| \leq \frac{\epsilon}{2}$$

with probability 1 (i.e., it is a deterministic event) and that

$$\mathbf{P}\left(\sup_x \left|f_n(x) - \bar{f}(x)\right| > \frac{\epsilon}{2}\right) \leq \delta.$$

We will refer to these two components as the random and the bias parts of the error. This error measurement will be used in this chapter because it fits nicely in the framework of uniform SLLN convergence which was introduced in Chapter 3.

We now review some well known properties of these two components of the error. We limit the discussion to the univariate case. Consider the bias part. Since the examples are identically and independently distributed, we have

$$\bar{f}(x) = \mathbf{E} f_n(x) = \frac{1}{\sigma}\int f(y) K\left(\frac{y-x}{\sigma}\right) dy.$$

Using the fact that by choice $K(x)$ integrates to 1, we have

$$\begin{aligned} \bar{f}(x) - f(x) &= \frac{1}{\sigma}\int f(y) K\left(\frac{y-x}{\sigma}\right) dy - \int f(x) K(z)\, dz \\ &= \int f(x+\sigma z) K(z)\, dz - \int f(x) K(z)\, dz \\ &= \int K(z) \left(f(x+\sigma z) - f(x)\right) dz. \end{aligned}$$

Now, expand in Taylor series around the point $\sigma = 0$ to obtain

$$f(x+\sigma z) = f(x) + \sigma z f'(x) + \frac{1}{2}\sigma^2 z^2 f''(x) + \cdots,$$

whence

$$\bar{f}(x) - f(x) = \sigma f'(x) \int zK(z)\,dz + \frac{\sigma^2}{2} f''(x) \int z^2 K(z)\,dz + \cdots.$$

If we let σ decrease to 0 with increasing n then the bias goes to zero. The rate of decrease can be made faster by choosing $K(x)$ such that its first r moments are identically 0, i.e.,

$$\int K(x)x^i\,dx = 0, \qquad 1 \le i \le r.$$

Such a choice of kernel guarantees that the first r terms of the bias are 0 so that the expression for the bias becomes

$$\frac{\sigma^{r+1}}{(r+1)!} f^{(r+1)}(x) \int K(z) z^{r+1}\,dz + O(\sigma^{r+1}), \qquad \sigma \to 0.$$

However, if $K(x)$ is to have zero moments of order ≥ 2 then it must take negative as well as positive values. The estimate $f_n(x)$ may itself be negative at some points. This is not acceptable if the estimate is to be a density. In our case we will use $f_n(x)$ to estimate only the modes of $f(x)$ allowing $f_n(x)$ to be negative at places at need.

Now let us consider the random part of the error, namely

$$\mathbf{P}\left(\sup_x \left| f_n(x) - \bar{f}(x) \right| > \frac{\epsilon}{2}\right).$$

By the definition of $f_n(x)$ this can be written as

$$\mathbf{P}\left(\sup_{K_{\sigma,x} \in \mathcal{K}_\sigma} \left| \frac{1}{n} \sum_{i=1}^{n} K_{\sigma,x}(x_i) - \mathbf{E}K_{\sigma,x}(x_1) \right| > \epsilon\sigma\right)$$

where $\mathcal{K}_\sigma$ is a class of functions indexed by the scalar x, i.e.,

$$\mathcal{K}_\sigma = \left\{ K_{\sigma,x}(y) \equiv K\left(\frac{x-y}{\sigma}\right) : x \in \mathbb{R} \right\}.$$

Theorem 3.10 bounds this probability by a quantity of the form

$$\left(\frac{1}{\epsilon\sigma} \log \frac{1}{\epsilon\sigma}\right)^{VC(\mathcal{K}_\sigma)} e^{-n\epsilon^2\sigma} \equiv \delta$$

where we ignored some of the constants which are irrelevant to our discussion here and we used the fact that δ_n^2 in the theorem is proportional to σ if the underlying distribution can be uniformly bounded and K is square integrable. Thus the confidence of getting an ϵ accurate estimate $f_n(x)$ is at least $1 - \delta$.

The above discussion identifies two critical variables that influence whether this bound is small, and how fast it decreases with n. First, the choice of σ. The bound on the probability of the random error term is reduced as σ increases. That means we can make the accuracy parameter ϵ smaller while keeping the confidence δ the same, i.e., the random error term is reduced as σ increases. The intuition behind it is that as the "window size" σ gets larger, the variance of the estimate $f_n(x)$ decreases. However, as we noted before, the bias increases as σ increases. This conflict of interest demonstrates one of the fundamental problems of density estimation.

The second conflict of interest appears when trying to shape the kernel $K(x)$ so as to have a bias that depends on high order terms of σ (as noted above). As we saw, the bias can be made to decrease faster if we choose $K(x)$ which is orthogonal to x^i, $1 \leq i \leq r$. However, as will be shown in the proof of Theorem 5.1, one class $\mathcal{K}_\sigma$ of such kernels, exhibits a VC-dimension which increases as r increases. As a result, the bound on the random part of the error increases, unless we increase the error deviation ϵ of the estimator in order to keep the same confidence $1 - \delta$. This is obviously not desirable. So as we try to exhibit a shape for the kernel that has the first r moments identically zero in order to reduce the bias at a faster rate w.r.t n, there is an adverse effect, coming from the random part of the error through the increase in $VC(\mathcal{K}_\sigma)$. The intuition behind this is that a kernel function having more zero moments is likely to have more relative maxima and heavier negative parts over its fixed support in which case $\sup_{x,y:|x-y|\leq\alpha} |K(x) - K(y)| \equiv M$ increases. It can be shown by an argument as in Lemma 5.6 that the covering number of the class of functions increases as M increases. A higher covering number implies a higher VC-dimension (by Theorem 3.8) which agrees with the above.

With some algebra it can be shown that the fastest possible decrease of σ w.r.t. n, while keeping a tight bound on the random component of the error, is such that $\frac{\sigma}{\log n/n} \to \infty$ as $n \to \infty$ (cf. Silverman, B.W. [41], Stute, W. [48]). In fact, this yields a bound that decreases fast enough to achieve uniform a.e. convergence of $\sup_x |f_n(x) - \bar{f}(x)|$ to 0. In choosing the kernel one minimizes the bound on the random part w.r.t. r.

5.2 Gaussian Mixture

In this section we consider the mixed sample complexities for learning the classification decision rule for a problem whose pattern classes are distributed as N-dimensional Gaussians with unit covariances where the learner, in the absence of specific parametric side-information about the class, opts for a kernel estimation approach.

The modes of the Gaussian mixture determine the Bayes decision whenever the mixture has two modes, which holds if the means satisfy $|\theta_{01} - \theta_{02}| > 2$. This together with the fact that the Bayes decision border is the hyperplane equidistant from the modes and perpendicular to the line $\eta_2 - \eta_1$, is shown in Lemma 5.5. (Note, the modes of the mixture do not equal the means of the class conditional).

Algorithm K (shown below) is used to construct the decision border by first using the unlabeled sample to estimate the mixture $f(x)$ by the kernel estimate $f_n(x)$. Then two modes $\hat{\eta}_1$, $\hat{\eta}_2$ of $f_n(x)$ are obtained such that they are consistent estimates of the two modes of the mixture f. The intuition here is that for sufficiently small accuracy $\epsilon > 0$ the main humps of f_n capture the modes η_1, η_2 of f. Hence the value $\hat{\eta}_i$ at which f_n is maximized over the i^{th} hump is a consistent estimate of the η_i.

Using the same analysis as for the MLE in Chapter 4, the decision regions separated by the hyperplane between the mode estimates yield close-to-Bayes misclassification error when labeled with the good labeling (as in the MLE we also have L_{good}, L_{bad} and use the same number, m, of labeled examples to choose L_{good} with some confidence).

We first state the algorithm, then we state the theorem, followed by the proof. (Algorithm K can be used to learn the Gaussian-mixture based problem as well as the larger class of mixtures of general form which is discussed in Section 5.5. In order not to break the flow, we only state the algorithm here while its discussion is delayed to Section 5.5. For the Gaussian mixture problem we will show the construction of the mode estimates in the proof of Theorem 4.2 hence for the moment it suffices to focus only on the main part of the algorithm without procedure P which describes the construction of the mode estimates for the more general case.)

Algorithm K:

```
The setting:  $f(x)$ has global modes, $\eta_i$, $1 \leq i \leq k$, where $k \geq 2$.

Given:        $n$ unlabeled examples, $m$ labeled examples drawn randomly according to
              the unknown $f(x)$.

Begin:
    1) Use kernel estimation to obtain $f_n(x)$.
    2) Use procedure P to determine the mode estimates, $\hat{\eta}_i$, $1 \leq i \leq k$.
    3) Use the mode estimates to construct a decision border as the hyperplane
       which passes through the point $\bar{\eta}$,
           $\bar{\eta} = \frac{1}{k}\sum_{i=1}^{k} \hat{\eta}_i$
       and which is perpendicular to the straight line which is the closest
       (in the mean-squared-error sense) to $\hat{\eta}_i$, $1 \leq i \leq k$. (Note, in the
       Gaussian mixture case $k = 2$ hence this step produces the hyperplane
       which is perpendicular to the line through $\hat{\eta}_1$ and $\hat{\eta}_2$.)
    4) Label the two decision regions across the hyperplane by the label of
       the majority of the examples in each region.
End.
```

```
Procedure P:  Definitions:
      • $M \equiv f(\eta_i)$, $1 \le i \le k$.
      • $D = \{x : f'(x; u_{\eta_i,x}) = 0, x \neq \eta_j, f(\eta_i) = M, f(\eta_j) = M, 1 \le i,j \le k\}$
        $L \equiv \sup_{x \in D} f(x)$
        where $f'(x; u_{\eta_i,x})$ is the directional derivative of $f$ at $x$ in the direction
        of the unit vector $u_{\eta_i,x}$ whose direction is the same as the ray starting
        at $\eta_i$ going through $x$.
      • Choose $\epsilon < \frac{M-L}{8}$.
      • $\hat{\eta}_1 \equiv \mathrm{argsup}_{x \in \mathbb{R}^N} f_n(x)$.
      • $B_\epsilon = \{x : f_n(x) > f_n(\hat{\eta}_1) - 4\epsilon\}$.
      • $A_{i,\epsilon} = \{y : |y - \hat{\eta}_i| \le \inf_x \inf_{z \in r_{\hat{\eta}_i,x}} |z - \hat{\eta}_i|, f_n(z) < f_n(\hat{\eta}_1) - 6\epsilon\} \cup \{\hat{\eta}_i\}$.
        for $1 \le i \le k$, where $r_{\hat{\eta}_i,x}$ is a ray from $\hat{\eta}_i$ going through $x$.
      • $Y = B - A_1$.
      • $i = 2$.

Do While:    $Y \neq \emptyset$
      1) $\hat{\eta}_i \equiv \mathrm{argsup}_{x \in Y} f_n(x)$.
      2) $Y = Y - A_i$.
      3) $i = i + 1$.
End Do.

End Procedure.
```

We now state the theorem, followed by its preview and proof.

Theorem 5.1 *Suppose we are given two classes which are distributed according to Gaussian probability densities $f_1(x)$, $f_2(x)$, with means θ_{01}, and θ_{02} respectively, and with unit covariance matrices. Suppose further that $0 < P_{Bayes} \leq 0.16$ (or, equivalently $|\theta_{01} - \theta_{02}| > 2$). Then there exists a positive constant b (determined by $|\theta_{01} - \theta_{02}|$) such that for $0 < \epsilon < b$ and arbitrary $\delta > 0$, given*

$$n = c_1 \frac{13^{N \log(5+\log N)}}{\epsilon^{\frac{N}{2 \log N}}} \log \frac{1}{\epsilon\delta}$$

unlabeled examples and

$$m = c_2 \log \frac{1}{\delta}$$

labeled examples, algorithm K determines a decision rule with a classification error

$$P_{error}(m, n) \leq P_{Bayes}(1 + c_3 \epsilon)$$

with confidence at least $1 - \delta$, where $c_1 > 0$ an absolute constant, $c_2 > 0$ is a constant depending on P_{Bayes}, and $c_3 > 0$ depends on θ_0.

REMARK: The restriction on P_{Bayes} is a consequence of the constraint on the two means of the class conditional densities to be sufficiently distant in order for the mixture to have two modes and thereby identify the Bayes border.

We now provide a preview of the proof.

5.2.1 Preview of the proof of Theorem 5.1

As mentioned at the start of Chapter 5, our nonparametric approach here is kernel estimation which utilizes the unlabeled sample to estimate the mixture $f(x)$ by the estimate

$$f_n(x) \equiv \frac{1}{n} \sum_{i=1}^{n} \sigma^{-N} K\left(\frac{\zeta_i - x}{\sigma}\right). \tag{5.1}$$

where we use here the notation $\zeta_1, \ldots, \zeta_n$, to represent the randomly drawn unlabeled sample of size n. We denote by x, a vector in $\mathbb{R}^N$, and x_i, $1 \leq i \leq N$ denotes its components. We will use the particular kernel function defined by the polynomial over the interval $[-1, 1]$ as

$$K_1(x) = \begin{cases} \sum_{i=0}^{r-1} a_i x_1^i & |x_1| \leq 1, \\ 0 & \text{otherwise}, \end{cases}$$

where x_1^i denotes the variable x_1 raised to the i^{th} power with r an even integer and with coefficients a_i selected so that $K_1(x_1)$ is orthogonal to x_1^i for $1 \leq i \leq r-1$ and such that $\int K_1(x_1)\, dx_1 = 1$. (The subscript in K_1 denotes it is a one dimensional kernel.) The N-dimensional kernel used in (5.1) is defined as the product

$$K(y) = K_1(y_1) K_1(y_2) \cdots K_1(y_N), \qquad y = [y_1, \ldots, y_N].$$

As mentioned before, having $r-1$ zero moments helps in the reduction of the bias.

The estimation discrepancy will be defined as the worst case (over all x) deviation of $f_n(x)$ from $f(x)$ divided by the absolute value of $f(x)$. This allows us to compare the performance as a function of dimensionality N, i.e., fix an ϵ-accuracy between P_{error} and P_{Bayes} to hold for all dimensions, and compare the sample complexities for different N with this fixed error criterion. As before, the total error is split into the bias and the random components

$$\frac{\sup_x |f_n(x) - f(x)|}{\sup_x f(x)} \leq \frac{\sup_x |\bar{f}(x) - f(x)|}{\sup_x f(x)} + \frac{\sup_x |f_n(x) - \bar{f}(x)|}{\sup_x f(x)}.$$

Because we are interested in estimating the unknown distribution $f(x)$ *uniformly* for all $x \in \mathbb{R}^N$ by the estimate $f_n(x)$, the aim will be to transform the random component of the error partly into a uniform SLLN convergence over a class $\mathcal{K}_\sigma$ of bounded functions using the same truncation ideas as before. Each function in $\mathcal{K}_\sigma$ is indexed by a point $x \in D \subset \mathbb{R}^N$, where D is a suitably chosen compact set, and hence uniformly approximating this class of functions gives the uniform approximation of $f(x)$ over D. The uniform SLLN cannot be applied over D^c as it is a non compact region. However, we can choose D so that the magnitude of $f(x)$, $f_n(x)$, and hence their difference, is sufficiently small, to obtain a good estimate, $f_n(x)$, uniformly for $x \in \mathbb{R}^N$.

We first proceed to determine a bound on the bias component. We show that there exists a coefficient vector a which satisfies the above orthogonality conditions on the kernel K_1. Lemma 5.3 is used to bound the magnitude $|K_1(x_1)|$ from above. This bound is needed for both the bias and the random parts.

The bias is then expanded in a Taylor series with remainder around $\sigma = 0$ which yields a polynomial in σ whose coefficients depend on the first $r + 1$ derivatives of the unknown mixture $f(x)$, which is a Gaussian mixture in this case, and also on the first r moments of the one dimensional kernel. Using the orthogonality of K_1 we are left with one term which depends on a bound of $f^{(r)}(x)$ and on the r^{th} moment of K_1. The former is bounded using the theory of Hermite polynomials, see Lemma 5.4, and the latter is bounded using the results of Lemma 5.3.

The random part of the error is bounded using the uniform SLLN (Theorem 3.10). Adapting an approach from Pollard [21], we view the estimate $f_n(x)$ as the empirical mean of a class of N-dimensional functions $K_{\sigma,x}(y) = K\left(\frac{y-x}{\sigma}\right) \in \mathcal{K}_\sigma$, each indexed by a "parameter" $x \in D \subset \mathbb{R}^N$. The uniformity that is sought for the estimate $f_n(x)$ over all $x \in \mathbb{R}^N$ is achieved by invoking the uniform SLLN theorem over the class $\mathcal{K}_\sigma$. The reason that Theorem 3.10 and not Theorem 3.9 was used here is due to the $1/\sigma^N$ factor which is a part of $f_n(x)$ and is allowed to increase as the sample size n is increased.

We cannot let $K_{\sigma,x}(y)$ be defined simply as $\frac{1}{\sigma^N}K(\frac{y-x}{\sigma})$ because this would make the magnitude of functions $K_{\sigma,x}$ in the class, $\mathcal{K}_\sigma$, depend on n; in particular, the magnitude of $K_{\sigma,x}$ would increase with n.

The bounds on the random part of the error involve the quantity $VC(\mathcal{K}_\sigma)$. This is evaluated easily by Theorem 3.6 together with Definitions 3.4, 3.5. We have chosen the one dimensional Kernel to be an $(r-1)^{th}$ degree polynomial, i.e., a linear combination of basis functions, specifically in order to be able to apply Theorem 3.6 and yield a bound for $VC(\mathcal{K}_\sigma)$.

Finally, we combine the bounds on the two error parts and deduce the finite unlabeled sample complexity.

With that accomplished, we then describe the learning procedure (which is based on algorithm K that is described in details in Section 5.5) and show how the modes of $f_n(x)$ yield consistent estimates of the modes of $f(x)$. Using the same analysis as in the parametric cases of Chapter 4 we use the small labeled sample with the majority rule to confidently pick the good labeling of the partition.

We first collect auxiliary lemmas needed in the proof of the theorem to avoid breaking up the flow subsequently.

5.3 Auxiliary Lemmas

The following lemma can be found in Szego & Polya [25, page 89].

Lemma 5.2 *For any arbitrary polynomial $P(x) = \sum_{i=0}^{r} a_i x^i$ with real coefficients a_i such that $\int_{-1}^{1}(P(x))^2\,dx = 1$ then for $-1 \le x \le 1$ we have $|P(x)| \le \frac{r+1}{\sqrt{2}}$.*

PROOF: $P(x)$ is a an arbitrary polynomial of degree r hence can be expanded using the Legendre basis as

$$P(x) = \sum_{k=0}^{r} a_k \sqrt{\frac{2k+1}{2}} P_k(x).$$

Then by the condition on $P(x)$ we have

$$\int_{-1}^{1} P^2(x)\,dx = a_0^2 + a_1^2 + \ldots + a_r^2 = 1.$$

From Holder's Inequality we have

$$\sum_i a_i b_i \le \sum_i |a_i b_i| \le \sqrt{\sum_i a_i^2}\sqrt{\sum_i b_i^2}$$

hence $(\sum_i a_i b_i)^2 \le \sum_i a_i^2 \sum_i b_i^2$. Regarding $\sqrt{\frac{2k+1}{2}}P_k(x)$ as a sequence in k we have

$$P^2(x) = \left(\sum_k a_k \sqrt{\frac{2k+1}{2}} P_k(x)\right)^2 \le (\sum_k a_k^2) \sum_k \frac{2k+1}{2} P_k^2(x) \le \sum_k \frac{2k+1}{2}$$

because $|P_k(x)| \le 1$ for $|x| \le 1$ (the reason for that is provided below). Finally,

$$\frac{1}{2}\sum_{k=0}^{r} 2k+1 = \sum_{k=0}^{r} k + \frac{1}{2}(r+1) = \frac{(r+1)^2}{2}$$

which proves the theorem.

Now we show that for the Legendre polynomial $P_n(x)$ we have $|P_n(x)| \le 1$ over $|x| \le 1$. We first prove that $P_n(x)$ satisfies a recursion equation, then from this we find the generating function for the sequence $P_n(x)$, in n. Denote the coefficient of x^n of $P_n(x)$ as k_n.

$$P_n(x) - \frac{k_n}{k_{n-1}} x P_{n-1}(x)$$

is an $(n-1)$-polynomial hence can be expanded as linear combination of Legendre basis,

$$P_n(x) - \frac{k_n}{k_{n-1}} x P_{n-1}(x) = \sum_{k=0}^{n-1} c_k P_k(x).$$

Multiply both sides by $(P_i(x), \cdot)$ where $(\cdot,\cdot) \equiv \int_{-1}^{1}(\cdot)(\cdot)\,dx$. Note that

$$(xP_{n-1}, P_i) = (xP_i, P_{n-1})$$

and $xP_i(x)$ is an $(i+1)$-polynomial therefore if $i+1 < n-1$ then $(xP_i, P_{n-1}) = 0$. Also, $(P_n, P_i) = 0$ if $i < n-2$. So the LHS is zero for $i < n-2$. The RHS is

$$c_0(P_0, P_i) + \dots c_{n-3}(P_{n-3}, P_i) + c_{n-2}(P_{n-2}, P_i) + c_{n-1}(P_{n-1}, P_i).$$

If $i = 0$ then the RHS is $c_0(P_0, P_0)$ hence $c_0 = 0$. Similarly with $i = 1, \ldots, n-3$. So $c_0 = c_1 = c_{n-3} = 0$. So

$$P_n(x) = (\frac{k_n}{k_{n-1}}x + c_{n-1})P_{n-1}(x) + c_{n-2}P_{n-2}(x).$$

Now we show that

$$P_n(x) = P_n(-x)(-1)^n.$$

We have

$$\int_{-1}^{1} P_n(-x)P_m(-x)\,dx = \int_{-1}^{1} P_n(x)P_m(x)\,dx = \delta_{nm}.$$

So the polynomials $\{P_n(-x)\}$ form an orthogonal basis and hence we can expand

$$P_n(x) = \sum_{k=0}^{n} \alpha_k P_k(-x)$$

since $P_n(x)$ is an n-polynomial. Now, $(P_n(x), P_i(-x)) = 0$ when $i < n$ so $\alpha_i = 0$ for $i < n$. So $P_n(x) = \alpha_n P_n(-x)$. Equate coefficients of x^n and find that $\alpha_n = (-1)^n$. Hence $P_n(x) = (-1)^n P_n(-x)$.

With this we can find the value for c_{n-1}. Clearly, $(-1)^n P_n(-x)$ satisfies the recurrence equation hence

$$(-1)^n P_n(-x) = (\frac{k_n}{k_{n-1}}x + c_{n-1})(-1)^{n-1}P_{n-1}(-x) + c_{n-2}(-1)^{n-2}P_{n-2}(-x).$$

Subtract it from the original recursion and we get $c_{n-1} = 0$. Now to find c_{n-2} we use the fact that $P_n(1) = 1$ which can be seen from the general formula (cf. Szego & Polya [25])

$$P_n(x) = \left(\frac{x+1}{2}\right)^n \sum_{k=0}^{n} \binom{n}{k}^2 \left(\frac{x-1}{x+1}\right)^k$$

(where we use the notation $0^0 = 1$). We have,

$$P_n(1) = \frac{k_n}{k_{n-1}} \cdot 1 \cdot P_{n-1}(1) + c_{n-2}P_{n-2}(1)$$

which implies $c_{n-2} = 1 - \frac{k_n}{k_{n-1}}$. Finally, it is easy to show that $k_n = \frac{(2n)!}{2^n(n!)^2}$ so $\frac{k_n}{k_{n-1}} = \frac{2n-1}{n}$ and $c_{n-2} = -\frac{n-1}{n}$.

Hence we proved the recurrence

$$P_n(x) = \frac{2n-1}{n} x P_{n-1}(x) - \frac{n-1}{n} P_{n-2}(x).$$

Now from the recurrence we can get the generating function of $P_n(x)$,

$$f(w) = \sum_{n\geq 0} P_n(x) w^n.$$

We have

$$nP_n = (2n-1)xP_{n-1} - (n-1)P_{n-2}.$$

We multiply both sides by $\sum_{n\geq 0} w^{n-1}$ and after some manipulation get

$$f'(w)(1 - 2xw + w^2) = (x - w)f(w).$$

This yields

$$f(w) = \frac{1}{\sqrt{1 - 2xw + w^2}}.$$

Let $x = \cos\theta$. Plug into the g.f. and get

$$\frac{1}{\sqrt{1 - 2\cos(\theta)w + w^2}} = \frac{1}{\sqrt{1 - e^{i\theta}w}} \frac{1}{\sqrt{1 - e^{-i\theta}w}}.$$

Using the identity

$$\frac{1}{\sqrt{1-4x}} = \sum_{k\geq 0} \binom{2k}{k} x^k$$

the right hand side becomes $\sum_{k\geq 0} \binom{2k}{k} 4^{-k} e^{i\theta k} w^k \sum_{j\geq 0} \binom{2j}{j} 4^{-j} e^{-i\theta j} w^j$. Taking the coefficient of w^n on both sides we have

$$\begin{aligned} P_n(\cos\theta) &= \sum_{j+k=n} \tbinom{2j}{j} \tbinom{2k}{k} 4^{-n} e^{i\theta(k-j)} \\ &= \tbinom{0}{0} \tbinom{2n}{n} 4^{-n} 2\cos(n\theta) + \tbinom{2}{1} \tbinom{2(n-1)}{n-1} 4^{-n} 2\cos((n-1)\theta) \\ &+ \ldots + \tbinom{n}{n/2} \tbinom{n}{n/2} 4^{-n}. \end{aligned}$$

All the terms multiplying the cosines are positive and $\cos(\cdot) \leq 1$ so therefore $|P_n(\cos\theta)| \leq P_n(1) = 1$ where the last equality we already established above. This proves that $|P_n(x)| \leq 1$. ▌

Lemma 5.3 *Suppose the polynomial*

$$K_1(x_1) = \sum_{i=0}^{r-1} a_i x_1^i 1_{\{|x_1| \leq 1\}}$$

satisfies

$$\int_{-1}^{1} K_1(x_1) x_1^j \, dx_1 = \delta_{j0}$$

for $0 \leq j \leq r-1$, *where* δ_{j0} *is the Kronecker delta. Then for an even integer* r,

$$a_0 = \frac{r^2}{2} \left(2^{-r} \binom{r}{r/2} \right)^2.$$

In particular, as $r \to \infty$ *through the even integers,*

$$a_0 \sim \frac{r}{\pi}.$$

PROOF: By Cramér's rule we have

$$a_0 = \frac{2^{r-1} \begin{vmatrix} \frac{1}{3} & 0 & \frac{1}{5} & \cdots & \frac{1}{r-1} & 0 & \frac{1}{r+1} \\ 0 & \frac{1}{5} & \cdots & \frac{1}{r-1} & 0 & \frac{1}{r+1} & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \frac{1}{r+1} & \cdots & 0 & 0 & \frac{1}{2r-3} & 0 \\ \frac{1}{r+1} & 0 & \cdots & 0 & \frac{1}{2r-3} & 0 & \frac{1}{2r-1} \end{vmatrix}}{2^{r} \begin{vmatrix} 1 & 0 & \frac{1}{3} & 0 & \frac{1}{5} & \cdots & \frac{1}{r-1} & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{5} & \cdots & \frac{1}{r-1} & 0 & \frac{1}{r+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{1}{r-1} & 0 & \frac{1}{r+1} & 0 & \cdots & 0 & \frac{1}{2r-3} & 0 \\ 0 & \frac{1}{r+1} & 0 & \cdots & \cdots & \cdots & 0 & \frac{1}{2r-1} \end{vmatrix}} \tag{5.2}$$

We will now reduce the determinant of the denominator as follows: number the rows and columns from 0 to $r-1$ (with r even). Consider the matrix

$$\begin{vmatrix} 1 & 0 & \frac{1}{3} & 0 & \frac{1}{5} & \cdots & \frac{1}{r-1} & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{5} & \cdots & \frac{1}{r-1} & 0 & \frac{1}{r+1} \\ \frac{1}{3} & 0 & \frac{1}{5} & \cdots & \frac{1}{r-1} & 0 & \frac{1}{r+1} & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{1}{r-1} & 0 & \frac{1}{r+1} & 0 & \cdots & 0 & \frac{1}{2r-3} & 0 \\ 0 & \frac{1}{r+1} & 0 & \cdots & \cdots & \cdots & 0 & \frac{1}{2r-1} \end{vmatrix}.$$

Note that row $2j$ can be written as

$$\left[\frac{1}{2j+1}\,0\,\frac{1}{2j+3}\,0\cdots 0\,\frac{1}{2j+r+1}\,0\right]$$

while row $2j+1$ can be written as

$$\left[0\,\frac{1}{2j+3}\,0\,\frac{1}{2j+5}\,0\cdots 0\,\frac{1}{2j+r+1}\right].$$

Now multiply all even numbered rows $2j$ for $j=1,2,3,\ldots,(r-2)/2$ by -1, whence the determinant becomes

$$(-1)^{(r-2)/2}\begin{vmatrix} 1 & 0 & \frac{1}{3} & 0 & \frac{1}{5} & \cdots & \frac{1}{r-1} & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{5} & \cdots & \frac{1}{r-1} & 0 & \frac{1}{r+1} \\ \frac{-1}{3} & 0 & \frac{-1}{5} & \cdots & \frac{-1}{r-1} & 0 & \frac{-1}{r+1} & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{-1}{r-1} & 0 & \frac{-1}{r+1} & 0 & \cdots & 0 & \frac{-1}{2r-3} & 0 \\ 0 & \frac{1}{r+1} & 0 & \cdots & \cdots & \cdots & 0 & \frac{1}{2r-1} \end{vmatrix}$$

Now add the top row to all the rows that start with a non-zero element yielding

$$(-1)^{(r-2)/2}\begin{vmatrix} 1 & 0 & \frac{1}{3} & 0 & \frac{1}{5} & \cdots & \frac{1}{r-1} & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{5} & \cdots & \frac{1}{r-1} & 0 & \frac{1}{r+1} \\ \frac{2}{3\cdot 1} & 0 & \frac{2}{5\cdot 3} & \cdots & \frac{2}{(r-1)(r-3)} & 0 & \frac{2}{(r+1)(r-1)} & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{r-2}{(r-1)1} & 0 & \frac{r-2}{(r+1)3} & 0 & \cdots & 0 & \frac{r-2}{(2r-3)(r-1)} & 0 \\ 0 & \frac{1}{r+1} & 0 & \cdots & \cdots & \cdots & 0 & \frac{1}{2r-1} \end{vmatrix}.$$

First factor out the numerators of alternating rows and get

$$(-1)^{(r-2)/2}(r-2)(r-4)\ldots(4)(2)\begin{vmatrix} 1 & 0 & \frac{1}{3} & 0 & \frac{1}{5} & \cdots & \frac{1}{r-1} & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{5} & \cdots & \frac{1}{r-1} & 0 & \frac{1}{r+1} \\ \frac{1}{3\cdot 1} & 0 & \frac{1}{5\cdot 3} & \cdots & \frac{1}{(r-1)(r-3)} & 0 & \frac{1}{(r+1)(r-1)} & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{1}{(r-1)1} & 0 & \frac{1}{(r+1)3} & 0 & \cdots & 0 & \frac{1}{(2r-3)(r-1)} & 0 \\ 0 & \frac{1}{r+1} & 0 & \cdots & \cdots & \cdots & 0 & \frac{1}{2r-1} \end{vmatrix}$$

Then factor the denominators of alternating columns to get

$$(-1)^{(r-2)/2}\frac{(r-2)(r-4)\ldots(4)(2)}{(r-1)(r-3)\cdots(3)}\begin{vmatrix} 1 & 0 & 1 & 0 & 1 & \cdots & 1 & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{5} & \cdots & \frac{1}{r-1} & 0 & \frac{1}{r+1} \\ \frac{1}{3} & 0 & \frac{1}{5} & \cdots & \frac{1}{(r-1)} & 0 & \frac{1}{(r+1)} & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{1}{(r-1)} & 0 & \frac{1}{(r+1)} & 0 & \cdots & 0 & \frac{1}{(2r-3)} & 0 \\ 0 & \frac{1}{r+1} & 0 & \cdots & \cdots & \cdots & 0 & \frac{1}{2r-1} \end{vmatrix}$$

Now repeat the operation on the columns: start with columns whose top element is 1 (excluding the first column), and multiply them by (-1). The determinant now becomes

$$(-1)^{r-2}\frac{(r-2)(r-4)\ldots(4)(2)}{(r-1)(r-3)\cdots(3)}\begin{vmatrix} 1 & 0 & -1 & 0 & -1 & \cdots & -1 & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{5} & \cdots & \frac{1}{r-1} & 0 & \frac{1}{r+1} \\ \frac{1}{3} & 0 & \frac{-1}{5} & \cdots & \frac{-1}{(r-1)} & 0 & \frac{-1}{(r+1)} & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{1}{(r-1)} & 0 & \frac{-1}{(r+1)} & 0 & \cdots & 0 & \frac{-1}{(2r-3)} & 0 \\ 0 & \frac{1}{r+1} & 0 & \cdots & \cdots & \cdots & 0 & \frac{1}{2r-1} \end{vmatrix}$$

Since r is even, then $(-1)^{r-2} = 1$. Now add the first column to all the columns that start with a -1 to get

$$\frac{(r-2)(r-4)\ldots(4)(2)}{(r-1)(r-3)\cdots(3)}\begin{vmatrix} 1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{5} & \cdots & \frac{1}{r-1} & 0 & \frac{1}{r+1} \\ \frac{1}{3} & 0 & \frac{2}{5\cdot 3} & \cdots & \frac{r-4}{(r-1)3} & 0 & \frac{r-2}{(r+1)3} & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{1}{(r-1)} & 0 & \frac{2}{(r+1)(r-1)} & 0 & \cdots & 0 & \frac{r-2}{(2r-3)(r-1)} & 0 \\ 0 & \frac{1}{r+1} & 0 & \cdots & \cdots & \cdots & 0 & \frac{1}{2r-1} \end{vmatrix};$$

factor out the numerators from the alternating columns to get

$$\frac{((r-2)(r-4)\ldots(4)(2))^2}{(r-1)(r-3)\cdots(3)}\begin{vmatrix} 1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{5} & \cdots & \frac{1}{r-1} & 0 & \frac{1}{r+1} \\ \frac{1}{3} & 0 & \frac{1}{5\cdot 3} & \cdots & \frac{1}{(r-1)3} & 0 & \frac{1}{(r+1)3} & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{1}{(r-1)} & 0 & \frac{1}{(r+1)(r-1)} & 0 & \cdots & 0 & \frac{1}{(2r-3)(r-1)} & 0 \\ 0 & \frac{1}{r+1} & 0 & \cdots & \cdots & \cdots & 0 & \frac{1}{2r-1} \end{vmatrix}.$$

And finally factor the denominators from the alternating rows to get

$$\left(\frac{(r-2)(r-4)\ldots(4)(2)}{(r-1)(r-3)\cdots(3)}\right)^2 \begin{vmatrix} 1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{5} & \cdots & \frac{1}{r-1} & 0 & \frac{1}{r+1} \\ 1 & 0 & \frac{1}{5} & \cdots & \frac{1}{(r-1)} & 0 & \frac{1}{(r+1)} & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & 0 & \frac{1}{(r+1)} & 0 & \cdots & 0 & \frac{1}{(2r-3)} & 0 \\ 0 & \frac{1}{r+1} & 0 & \cdots & \cdots & \cdots & 0 & \frac{1}{2r-1} \end{vmatrix}.$$

The determinant in the above expression equals the determinant on the numerator of (5.2). So the denominator of (5.2) is

$$\left(\frac{(r-2)(r-4)\cdots(4)(2)}{(r-1)(r-3)\cdots(3)}\right)^2 \cdot 2^r \begin{vmatrix} \frac{1}{3} & 0 & \frac{1}{5} & \cdots & \frac{1}{r-1} & 0 & \frac{1}{r+1} \\ 0 & \frac{1}{5} & \cdots & \frac{1}{r-1} & 0 & \frac{1}{r+1} & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \frac{1}{r+1} & \cdots & 0 & 0 & \frac{1}{2r-3} & 0 \\ \frac{1}{r+1} & 0 & \cdots & 0 & \frac{1}{2r-3} & 0 & \frac{1}{2r-1} \end{vmatrix}$$

and so $a_0 = \frac{1}{2}\left(\frac{(r-1)(r-3)\cdots(3)}{(r-2)(r-4)\cdots(4)(2)}\right)^2$. Simple manipulations results in the alternative form

$$a_0 = \frac{r^2}{2}\left(2^{-r}\binom{r}{r/2}\right)^2.$$

A simple application of Stirling's formula to the central term of the binomial gives

$$2^{-r}\binom{r}{r/2} \sim \frac{\sqrt{2}}{\sqrt{\pi r}}$$

as $r \to \infty$. This completes the proof.

Lemma 5.4 *Let $f(x)$ be a Gaussian N-dimensional density, then*

$$\sup_x \left| f^{(r)}_{x_{i_1},x_{i_2},\ldots,x_{i_r}}(x) \right| \leq C(2\pi)^{-N/2} r^{\frac{r}{2}} e^r$$

where $i_1, i_2, \ldots, i_r \in \{1, 2, \ldots, N\}$, and C is some positive constant.

PROOF: We need to bound

$$\left| \frac{\partial^{t_1}}{\partial x_1^{t_1}} \frac{\partial^{t_2}}{\partial x_2^{t_2}} \cdots \frac{\partial^{t_r}}{\partial x_N^{t_N}} e^{-\frac{1}{2}|x|^2} / (2\pi)^{N/2} \right|$$

where $0 \le t_i \le r$, and $\sum_{i=1}^N t_i = r$, and use the convention that $d^0/dx^0 f(x) = f(x)$. Without loss of generality suppose the derivatives are taken w.r.t. $x_1, \ldots, x_l$, i.e.,

$$\left|\frac{\partial^{t_1}}{\partial x_1^{t_1}}\frac{\partial^{t_2}}{\partial x_2^{t_2}}\cdots\frac{\partial^{t_l}}{\partial x_l^{t_l}}e^{-\frac{1}{2}|x|^2}/(2\pi)^{N/2}\right|$$

where $1 \le t_i \le r$, $\sum_{i=1}^l t_i = r$. Clearly $1 \le l \le r$. We will suppress the (2π) factor for brevity. This can be written as

$$\begin{aligned}&\left|\frac{\partial^{t_1}}{\partial x_1^{t_1}}e^{-\frac{1}{2}x_1^2}\right|\left|\frac{\partial^{t_2}}{\partial x_2^{t_2}}e^{-\frac{1}{2}x_2^2}\right|\cdots\left|\frac{\partial^{t_l}}{\partial x_l^{t_l}}e^{-\frac{1}{2}x_l^2}\right|\left|e^{-\frac{1}{2}x_{l+1}^2}\right|\cdots\left|e^{-\frac{1}{2}x_N^2}\right|\\ &\le \left|\frac{\partial^{t_1}}{\partial x_1^{t_1}}e^{-\frac{1}{2}x_1^2}\right|\left|\frac{\partial^{t_2}}{\partial x_2^{t_2}}e^{-\frac{1}{2}x_2^2}\right|\cdots\left|\frac{\partial^{t_r}}{\partial x_l^{t_l}}e^{-\frac{1}{2}x_l^2}\right|\end{aligned}$$

We first bound any one of the one dimensional factors, i.e., denoted by

$$\left|\frac{\partial^n}{\partial x^n}e^{-\frac{1}{2}x^2}\right|.$$

From the theory of Hermite polynomial we have

$$e^{-x^2}H_n(x) = (-1)^n\frac{\partial^n}{\partial x^n}e^{-x^2}.$$

So

$$(-1)^n\frac{\partial^n}{\partial x^n}e^{-\frac{1}{2}x^2} = (\frac{1}{\sqrt{2}})^n e^{-\frac{1}{2}x^2}H_n(\frac{1}{\sqrt{2}}x)$$

and therefore

$$\left|\frac{\partial^n}{\partial x^n}e^{-\frac{1}{2}x^2}\right| = \left|(\frac{1}{\sqrt{2}})^n e^{-\frac{1}{2}x^2}H_n(\frac{1}{\sqrt{2}}x)\right|.$$

We now bound the right side. Using the identity (cf. [26, page 102]),

$$H_n(x) = \sum_{k=0}^{\lfloor n/2\rfloor}(-1)^k\frac{n!}{k!(n-2k)!}(2x)^{n-2k}$$

the above becomes

$$\begin{aligned}\left|\frac{\partial^n}{\partial x^n}e^{-\frac{1}{2}x^2}\right| &= \left|(\frac{1}{\sqrt{2}})^n\sum_{k=0}^{n/2}(-1)^k\frac{n!}{k!(n-2k)!}(\sqrt{2}x)^{n-2k}e^{-\frac{1}{2}x^2}\right|\\ &= \left|\sum_{k=0}^{n/2}(-1)^k\frac{n!}{k!(n-2k)!}2^{-k}x^{n-2k}e^{-\frac{1}{2}x^2}\right|\end{aligned}$$

where we used $n/2$ instead of $\lfloor n/2 \rfloor$ for simplicity. This is further bounded by

$$\sum_{k=0}^{n/2} \frac{n!}{k!(n-2k)!} 2^{-k} \left| x^{n-2k} e^{-\frac{1}{2}x^2} \right| .$$

Now

$$\frac{n!}{k!(n-2k)!} 2^{-k} \le 2^{-k} \frac{n^{n-k}}{(n-2k)!} .$$

Also, simple differentiation shows

$$x^a e^{-\frac{1}{2}x^2} \le \left(\frac{a}{e} \right)^{a/2} , \quad a = 1, 2, 3, \ldots$$

Thus

$$\begin{aligned} \left| \frac{\partial^n}{\partial x^n} e^{-\frac{1}{2}x^2} \right| &< \sum_{k=0}^{n/2} 2^{-k} \frac{n^{n-k}}{(n-2k)!} \left(\frac{n-2k}{e} \right)^{\frac{n-2k}{2}} = \sum_{i=n,n-2,\ldots,0} 2^{\frac{n-i}{2}} \frac{n^{(n+i)/2}}{i!} \left(\frac{i}{e} \right)^{\frac{i}{2}} \\ &= \left(\frac{n}{2} \right)^{n/2} \sum_{i=n,n-2,\ldots,0} \frac{1}{i!} \left(\frac{2ni}{e} \right)^{i/2} . \end{aligned}$$

By Stirling's formula,

$$i! > \sqrt{2\pi} i^{i+1/2} e^{-i} e^{1/(12i+1)} \ge \sqrt{2\pi} \left(\frac{i}{e} \right)^i .$$

So

$$\left| \frac{\partial^n}{\partial x^n} e^{-\frac{1}{2}x^2} \right| < c_1 \left(\frac{n}{2} \right)^{n/2} \sum_{i=n,n-2,\ldots,0} \left(\frac{2ne}{i} \right)^{i/2}$$

where $c_1 = 1/\sqrt{2\pi}$, and using the convention that $(1/0)^0 = 1$. The function $\left(\frac{2ne}{i} \right)^{i/2}$ increases monotonically in the range $0 < y \le n$ since its derivative is positive there. Thus the sum is bounded by $\frac{n}{2}(2e)^{n/2}$ and we have

$$\left| \frac{\partial^n}{\partial x^n} e^{-\frac{1}{2}x^2} \right| < c_1 \left(\frac{n}{2} \right)^{n/2} \frac{n}{2} (2e)^{n/2} = c_2 n^{n/2+1} e^{n/2}$$

where c_2 is an absolute positive constant. Finally, we have

$$\left| \frac{\partial^{t_1}}{\partial x_1^{t_1}} \frac{\partial^{t_2}}{\partial x_2^{t_2}} \cdots \frac{\partial^{t_l}}{\partial x_l^{t_l}} e^{-\frac{1}{2}|x|^2} \right| \le c_3 t_1^{t_1/2+1} e^{t_1/2} t_2{}^{t_2/2+1} e^{t_2/2} \cdots t_l{}^{t_l/2+1} e^{t_l/2} \le c_3 \prod_{i=1}^{l} t_i r^{r/2} e^{r/2} .$$

Now, recall that $1 \le t_i \le r$, and $\sum_{i=1}^{l} t_i = r$. We can use the following bound (cf. Marcus & Minc [42, page 106])

$$\prod_{i=1}^{l} t_i \le \left(\frac{1}{l}\sum_{i=1}^{l} t_i\right)^{l} = \left(\frac{r}{l}\right)^{l}.$$

Simple differentiation gives

$$\left(\frac{r}{l}\right)^{l} \le \left(\frac{r}{r/e}\right)^{r/e} = e^{r/e}$$

hence

$$\left|\frac{\partial^{t_1}}{\partial x_1^{t_1}}\frac{\partial^{t_2}}{\partial x_2^{t_2}}\cdots\frac{\partial^{t_l}}{\partial x_l^{t_l}}e^{-\frac{1}{2}|x|^2}/(2\pi)^{N/2}\right| \le \frac{c_3 r^{r/2} e^{r(1/2+1/e)}}{(2\pi)^{N/2}} \le c_3\frac{r^{r/2}e^{r}}{(2\pi)^{N/2}}.$$

Lemma 5.5 *The N-dimensional Gaussian mixture with unit covariance matrices and equal* a priori *class probabilities has two modes whenever the means, θ_1, θ_2 of the class conditional densities satisfy $|\theta_1 - \theta_2| > 2$. In this case, the modes determine the Bayes border (hyperplane).*

PROOF: The modes of the mixture are denoted by η_1, η_2. We have

$$f(x) = \frac{1}{2(2\pi)^{N/2}}e^{-\frac{1}{2}|x-\theta_1|^2} + \frac{1}{2(2\pi)^{N/2}}e^{-\frac{1}{2}|x-\theta_2|^2}$$

First, translate the frame so that the point whose coordinate vector is θ_1 is at the origin. Then transform to a new primed-coordinate system, $x' = Qx$ s.t. the coordinates of the means, θ_1' and θ_2' are on the x_1'-axis. (We will still denote the point by θ_1' although it is the origin). This is simply a rotation hence Q is unitary and the Jacobian equals 1 yielding

$$f(x') = \frac{1}{2(2\pi)^{N/2}}e^{-\frac{1}{2}|Q^T x' - Q^T\theta_1'|^2} + \frac{1}{2(2\pi)^{N/2}}e^{-\frac{1}{2}|Q^T x' - Q^T\theta_2'|^2}$$

$$\begin{aligned} &= \frac{1}{2(2\pi)^{N/2}} e^{-\frac{1}{2}|x'-\theta'_1|^2} + \frac{1}{2(2\pi)^{N/2}} e^{-\frac{1}{2}|x'-\theta'_2|^2} \\ &= \frac{1}{2(2\pi)^{N/2}} \left(e^{-\frac{1}{2}(x'_1-\theta'_{11})^2} + e^{-\frac{1}{2}(x'_1-\theta'_{21})^2} \right) e^{-\frac{1}{2}{x'_2}^2} e^{-\frac{1}{2}{x'_3}^2} \ldots e^{-\frac{1}{2}{x'_N}^2} \end{aligned}$$

Clearly the x' which maximizes $f(x')$ has $x'_2 = \ldots = x'_N = 0$. So the modes must be on the x'_1-axis and hence on the line through the means. We differentiate $f(x')$ w.r.t. x'_1 and equate to zero getting

$$\frac{e^{-\frac{1}{2}(x'_1-\theta'_{11})^2}}{e^{-\frac{1}{2}(x'_1-\theta'_{21})^2}} = \frac{\theta'_{21} - x'_1}{x'_1 - \theta'_{11}}$$

The left side is positive hence the solutions for x'_1 must be between θ'_{11} and θ'_{21}. Additional manipulation yields

$$x'_1 = \frac{1}{\theta'_{11} - \theta'_{21}} \log \frac{\theta'_{21} - x'_1}{x'_1 - \theta'_{11}} + \frac{\theta'_{21} + \theta'_{11}}{2}$$

and substituting $y = x'_1 - \frac{\theta'_{21}+\theta'_{11}}{2}$, $a = \frac{\theta'_{21}-\theta'_{11}}{2}$ gives

$$y = \frac{1}{2a} \log \frac{y+a}{a-y} \tag{5.3}$$

(Note, $a = |\theta_1 - \theta_2|/2$). The right side is an odd function around zero. At $y = -a$ and $y = a$ it equals $-\infty$ and ∞ respectively. Taking its derivative w.r.t. y yields $\frac{1}{a^2-y^2}$ which never equals zero hence it has no critical points. There are two cases for the derivative at $y = 0$: (1) < 1 which happens when $a^2 > 1$, (2) ≥ 1, occurring for $a^2 \leq 1$. Case (1) implies that the right side of (5.3) intersects the line (the function on the left side is a line y) only at the two points y_a and y_b (besides 0) which are equidistant from 0. This implies $f(x')$ has three critical points: at $x'_1 = \frac{\theta'_{11}+\theta'_{21}}{2}$ (corresponding to $y = 0$), and at two points which are equidistant from $\frac{\theta'_{11}+\theta'_{21}}{2}$. The first is a relative minimum and the other two are the modes hence the mixture has two modes. So in case (1) we showed that the modes are equidistant from the average of the means (which is where the Bayes hyperplane passes) hence the hyperplane passes through the average of the mode, and the line through the modes is perpendicular to the Bayes hyperplane. Therefore the modes determine the Bayes hyperplane under the condition that $|\theta_1 - \theta_2| > 2$. In case (2), there is only one extrema, and it is a maximum at $x'_1 = \frac{\theta'_{11}+\theta'_{21}}{2}$; the mixture has only one mode. The Bayes hyperplane goes through this point, however it is not possible to determine which of the infinitely many possible hyperplanes is the Bayes border since the line perpendicular to the Bayes hyperplane cannot be determined.

Lemma 5.6 *The class $\mathcal{K}_\sigma$ of kernels can be finitely covered.*

(We prove this lemma since finite coverability is a necessary condition for Theorem 5.1, in which it is stated as a permissibility condition.)

PROOF: In denoting the class $\mathcal{K}_\sigma = \{K_{\sigma,x}(y) : x \in D\}$ and N-dimensional kernel $K_{\sigma,x}$, we will omit the σ since it is the same for all functions in $\mathcal{K}_\sigma$, and D is compact subset of $\mathbb{R}^N$. In the following, all vectors, such as x, y are in $\mathbb{R}^N$. We first show that for a fixed $\tilde{x}$ (a center of a sphere in the covering of D), and fixed ϵ, $\sup_{x_\epsilon} \mathbf{E}\,|K_{\tilde{x}}(y) - K_{x_\epsilon}(y)| \leq c\epsilon$, where $x_\epsilon \in \mathbb{R}^N$ is s.t. $|\tilde{x} - x_\epsilon| \leq \epsilon$, and $c > 0$ is a constant. We assume that the distribution of y is absolutely continuous and denote the pdf by $f(y)$; we also need that $f(y)$ has a probability-1 support containing the region $\{z : |z - y| \leq 1, y \in D\}$. In what follows we denote the one dimensional kernel by $K_1(y_i - x_i) = poly(y_i - x_i)1_{|y_i - x_i| \leq 1}$, $1 \leq i \leq N$. We start with:

$$\begin{aligned}
&\sup_{x_\epsilon} \mathbf{E}\,|K_{\tilde{x}}(y) - K_{x_\epsilon}(y)| \\
&\quad = \sup_{x_\epsilon} \int \Big| poly(y_1 - x_{\epsilon 1})1_{|y_1 - x_{\epsilon 1}| \leq 1} poly(y_2 - x_{\epsilon 2})1_{|y_2 - x_{\epsilon 2}| \leq 1} \cdots poly(y_N - x_{\epsilon N})1_{|y_N - x_{\epsilon N}| \leq 1} \\
&\quad - poly(y_1 - \tilde{x}_1)1_{|y_1 - \tilde{x}_1| \leq 1} poly(y_2 - \tilde{x}_2)1_{|y_2 - \tilde{x}_2| \leq 1} \cdots poly(y_N - \tilde{x}_N)1_{|y_N - \tilde{x}_N| \leq 1} \Big| f(y)\, dy
\end{aligned}$$

In the above, let $A_1(y, x_\epsilon, \tilde{x})$ denote the quantity inside the absolute value. We need to define the following:

$$\begin{aligned} L^i_{1,x_\epsilon,\tilde{x}} &\equiv \{y_i : |y_i - x_{\epsilon i}| > 1, |y_i - \tilde{x}_i| \leq 1\} \\ L^i_{2,x_\epsilon,\tilde{x}} &\equiv \{y_i : |y_i - x_{\epsilon i}| \leq 1, |y_i - \tilde{x}_i| \leq 1\} \\ L^i_{3,x_\epsilon,\tilde{x}} &\equiv \{y_i : |y_i - x_{\epsilon i}| \leq 1, |y_i - \tilde{x}_i| > 1\} \end{aligned}$$

where subscript i denotes i^{th} component of a vector. Continuing we have

$$\begin{aligned} &\sup_{x_\epsilon} \int |A_1(y, x_\epsilon, \tilde{x})| \, f(y) \, dy \\ &= \sup_{x_\epsilon} \int |A_1(y, x_\epsilon, \tilde{x})| \, 1_{\{y_1 \in L^1_{1,x_\epsilon,\tilde{x}}\}} f(y) \, dy + \sup_{x_\epsilon} \int |A_1(y, x_\epsilon, \tilde{x})| \, 1_{\{y_1 \in L^1_{2,x_\epsilon,\tilde{x}}\}} f(y) \, dy \\ &+ \sup_{x_\epsilon} \int |A_1(y, x_\epsilon, \tilde{x})| \, 1_{\{y_1 \in L^1_{3,x_\epsilon,\tilde{x}}\}} f(y) \, dy \qquad (5.4) \end{aligned}$$

The first term equals

$$\begin{aligned} &\sup_{x_\epsilon} \int \Big| poly(y_1 - x_{\epsilon 1}) 1_{|y_1 - x_{\epsilon 1}| \leq 1} 1_{\{y_1 \in L^1_{1,x_\epsilon,\tilde{x}}\}} \prod_{j=2}^{N} poly(y_j - x_{\epsilon j}) 1_{|y_j - x_{\epsilon j}| \leq 1} \\ &\quad - \; poly(y_1 - \tilde{x}_1) 1_{|y_1 - \tilde{x}_1| \leq 1} 1_{\{y_1 \in L^1_{1,x_\epsilon,\tilde{x}}\}} \prod_{j=2}^{N} poly(y_j - \tilde{x}_j) 1_{|y_j - \tilde{x}_j| \leq 1} \Big| f(y) \, dy \\ &= \; \sup_{x_\epsilon} \int \Big| poly(y_1 - \tilde{x}_1) 1_{\{y_1 \in L^1_{1,x_\epsilon,\tilde{x}}\}} \prod_{j=2}^{N} poly(y_j - \tilde{x}_j) 1_{|y_j - \tilde{x}_j| \leq 1} \Big| f(y) \, dy \end{aligned}$$

since $1_{\{y_1 \in L^1_{1,x_\epsilon,\tilde{x}}\}} 1_{|y_1 - x_{\epsilon 1}| \leq 1} = 0$ and $1_{\{y_1 \in L^1_{1,x_\epsilon,\tilde{x}}\}} 1_{|y_1 - \tilde{x}_1| \leq 1} = 1_{\{y_1 \in L^1_{1,x_\epsilon,\tilde{x}}\}}$. The above is

$$\leq \; C^{N-1} \sup_{x_\epsilon} \int_{|y_2 - \tilde{x}_2| \leq 1, \ldots, |y_N - \tilde{x}_N| \leq 1} f(y_2, \ldots, y_N) \int_{L^1_{1,x_\epsilon,\tilde{x}}} f(y_1 | y_2, \ldots, y_N) dy$$

where C bounds the one-dimensional kernel over its support. Now the only factor depending on x_ϵ is $L^1_{1,x_\epsilon,\tilde{x}}$ which is $\subset L^1_{\tilde{x},\epsilon}$ where

$$L^1_{\tilde{x},\epsilon} = L^1_{1,x_\epsilon,\tilde{x}} : |x_{\epsilon 1} - \tilde{x}_1| = \epsilon$$

(recall that $|\tilde{x} - x_\epsilon| \leq \epsilon$ hence $|x_{\epsilon 1} - \tilde{x}_1| \leq \epsilon$). That is, $L^1_{\tilde{x},\epsilon}$ is the one interval that corresponds to $L^1_{1,x_\epsilon,\tilde{x}}$ with the specific x_ϵ that is ϵ away from $\tilde{x}_1$. Hence the above is

$$\begin{aligned} &\leq \; C^{N-1} \int_{|y_2-\tilde{x}_2|\leq 1,\ldots,|y_N-\tilde{x}_N|\leq 1} \int_{L^1_{\tilde{x},\epsilon}} f(y_1|y_2,\ldots,y_N)\,dy_1 f(y_2,\ldots,y_N)\,dy_2\ldots dy_N \\ &\leq \; C^{N-1} M_1 \epsilon \end{aligned}$$

where we assume $f(y_1|y_2,\ldots,y_N) \leq M_1$ over the support of the integral, for some positive constant M_1. Note, because of the initial assumption on the probability-1 support of $f(y)$ it follows that $L^1_{\tilde{x},\epsilon}$ cannot contain all the probability mass and hence the above follows.

We can similarly bound the third term of (5.4) by $C^{N-1}M_3\epsilon$, for some positive constant M_3, using the fact that $1_{\{y_1 \in L^1_{3,x_\epsilon,\tilde{x}}\}} 1_{|y_1-\tilde{x}_1|\leq 1} = 0$ and $1_{\{y_1 \in L^1_{3,x_\epsilon,\tilde{x}}\}} 1_{|y_1-x_{\epsilon 1}|\leq 1} = 1_{\{y_1 \in L^1_{3,x_\epsilon,\tilde{x}}\}}$. We use the fact that $1_{\{y_1 \in L^1_{2,x_\epsilon,\tilde{x}}\}} 1_{|y_1-\tilde{x}_1|\leq 1} = 1_{\{y_1 \in L^1_{2,x_\epsilon,\tilde{x}}\}}$ and $1_{\{y_1 \in L^1_{2,x_\epsilon,\tilde{x}}\}} 1_{|y_1-x_{\epsilon 1}|\leq 1}$ $= 1_{\{y_1 \in L^1_{2,x_\epsilon,\tilde{x}}\}}$ to get the second term of (5.4)

$$\begin{aligned} \sup_{x_\epsilon} \int \Big| & poly(y_1 - x_{\epsilon 1}) 1_{\{y_1 \in L^1_{2,x_\epsilon,\tilde{x}}\}} \prod_{j=2}^{N} poly(y_j - x_{\epsilon j}) 1_{|y_j - x_{\epsilon j}|\leq 1} \\ - \;& poly(y_1 - \tilde{x}_1) 1_{\{y_1 \in L^1_{2,x_\epsilon,\tilde{x}}\}} \prod_{j=2}^{N} poly(y_j - \tilde{x}_j) 1_{|y_j - \tilde{x}_j|\leq 1} \Big| f(y)\,dy \end{aligned}$$

Denote the quantity in the absolute value as $A_2(y, x_\epsilon, \tilde{x})$. We break it as

$$\begin{aligned} & \sup_{x_\epsilon} \int |A_2(y,x_\epsilon,\tilde{x})|\, 1_{\{y_2 \in L^2_{1,x_\epsilon,\tilde{x}}\}} f(y)\,dy \\ & \quad + \int |A_2(y,x_\epsilon,\tilde{x})|\, 1_{\{y_2 \in L^2_{2,x_\epsilon,\tilde{x}}\}} f(y)\,dy + \int |A_2(y,x_\epsilon,\tilde{x})|\, 1_{\{y_2 \in L^2_{3,x_\epsilon,\tilde{x}}\}} f(y)\,dy \end{aligned}$$

Using the same ideas as before, we find the first and third terms are bounded from above by some positive constant multiple of ϵ. Then continue to break the second term into

$$\begin{aligned} \sup_{x_\epsilon} \int \Big| & poly(y_1 - x_{\epsilon 1}) 1_{\{y_1 \in L^1_{2,x_\epsilon,\tilde{x}}\}} poly(y_2 - x_{\epsilon 2}) 1_{\{y_2 \in L^2_{2,x_\epsilon,\tilde{x}}\}} \prod_{j=3}^{N} poly(y_j - x_{\epsilon j}) 1_{|y_j - x_{\epsilon j}|\leq 1} \\ - \;& poly(y_1 - \tilde{x}_1) 1_{\{y_1 \in L^1_{2,x_\epsilon,\tilde{x}}\}} poly(y_2 - \tilde{x}_2) 1_{\{y_2 \in L^2_{2,x_\epsilon,\tilde{x}}\}} \prod_{j=3}^{N} poly(y_j - \tilde{x}_j) 1_{|y_j - \tilde{x}_j|\leq 1} \Big| f(y)\,dy \end{aligned}$$

Denote the quantity in absolute value by $A_3(y, x_\epsilon, \tilde{x})$ and continue to break as before until we end up with all terms which are bounded from above by some positive constants-multiple of ϵ, and one term which is as follows

$$\begin{aligned}
&\sup_{x_\epsilon} \int_{y_1 \in L^1_{2,x_\epsilon,\tilde{x}}, \ldots, y_N \in L^N_{2,x_\epsilon,\tilde{x}}} \left| \prod_{i=1}^N poly(y_i - x_{\epsilon i}) - \prod_{i=1}^N poly(y_i - \tilde{x}_i) \right| f(y)\, dy \\
\leq \; & \sup_{x_\epsilon} \sup_{y_1 \in L^1_{2,x_\epsilon,\tilde{x}}, \ldots, y_N \in L^N_{2,x_\epsilon,\tilde{x}}} \left| \prod_{i=1}^N poly(y_i - x_{\epsilon i}) - \prod_{i=1}^N poly(y_i - \tilde{x}_i) \right| \\
\leq \; & \left| \prod_{i=1}^N poly(y_i^* - x_{\epsilon i}^*) - \prod_{i=1}^N poly(y_i^* - \tilde{x}_i) \right| \qquad (5.5)
\end{aligned}$$

where x_ϵ^*, y^* are where the maximum is achieved (it is achieved since each of the one-dimensional kernel functions, $K_{x_{\epsilon i}}(y_i)$, is bounded over the set on which (x_ϵ, y) vary). But each of the one-dimensional factors,

$$|poly(y_i^* - x_{\epsilon i}^*) - poly(y_i^* - \tilde{x}_i)| \leq M_i' \epsilon$$

where M_i' is some finite constant, because

$$\sup_{x_{\epsilon i}} \sup_{y_i \in L^i_{2,x_{\epsilon i},\tilde{x}_i}} |poly(y_i - x_{\epsilon i}) - poly(y_i - \tilde{x}_i)| \leq \sup_{x_{\epsilon i}} \sup_{y_i : |y_i - \tilde{x}_i| \leq 1} |poly(y_i - x_{\epsilon i}) - poly(y_i - \tilde{x}_i)|$$

since $L^i_{2,x_{\epsilon i},\tilde{x}_i} \subseteq \{y_i : |y_i - \tilde{x}_i| \leq 1\}$. And therefore the above is bounded by

$$\sup_{s,t : \tilde{x}_i - 1 - \epsilon \leq s,\, t \leq \tilde{x}_i + 1 + \epsilon, |s-t| \leq \epsilon} |poly(s) - poly(t)| \leq M_i' \epsilon$$

because $|x_{\epsilon i} - \tilde{x}_i| \leq \epsilon$, and $poly()$ is continuous over the compact set

$$\{s, t : \tilde{x}_i - 1 - \epsilon \leq s,\, t \leq \tilde{x}_i + 1 + \epsilon\}.$$

Hence in (5.5), $poly(y_i^* - x_{\epsilon i}^*) = poly(y_i^* - \tilde{x}_i) + c_0 \epsilon$ for some constant $c_0 > 0$; by inspection it is clear that (5.5) becomes a constant multiple of ϵ for some positive constant.

Hence $\sup_{x_\epsilon} \mathbf{E} |K_{\tilde{x}}(y) - K_{x_\epsilon}(y)| \leq c\epsilon$, for some positive constant c. Now, for any $K_x(y) \in \mathcal{K}_\sigma$, (hence $x \in D$), there is a $\tilde{x}_k$, a center of a sphere in the covering of D, such that $|x - \tilde{x}_k| \leq \epsilon$. Corresponding to this $\tilde{x}_k$, $\exists\ K_{\tilde{x}_k}(y)\ \ni\ \mathbf{E}\,|K_{\tilde{x}_k}(y) - K_x(y)| \leq c\epsilon$ because we showed it is true for any x_ϵ with $|x_\epsilon - \tilde{x}_k| \leq \epsilon$. This implies the existence of a finite collection of functions $\{K_{\tilde{x}_1}(y), K_{\tilde{x}_2}(y), \ldots, K_{\tilde{x}_{cov(D)}}(y)\}$ which covers $\mathcal{K}_\sigma$ in the L_1 norm to an accuracy $c\epsilon$. ▮

5.4 Proof of Theorem 5.1

We will use here the notation $\zeta_1, \ldots, \zeta_n$, to represent the randomly drawn unlabeled sample of size n, where n is stated in Theorem 5.1. We denote by x, a vector in $\mathbb{R}^N$, and x_i, $1 \leq i \leq N$ denotes its components.

Initially we show that with this n-sample it is possible to estimate $f(x)$ by $f_n(x)$ to within small deviation where the goodness of fit is measured by

$$\frac{\sup_x |f_n(x) - f(x)|}{\sup_x f(x)}$$

where the sup is over $\mathbb{R}^N$. This implies that the mode-estimates are good and hence the decision rule $h(x)$ is close to the Bayes rule. The reason for this measure of fit is to enable a comparison of performance, i.e. error versus sample size, across different dimensions N. This choice will ensure that the $O(\epsilon)$ term of P_{error} in the theorem is independent of N (it may depend on quantities such as the distance between the true modes) and the range allowed for ϵ holds for all N; we still say *for small* $\epsilon > 0$ in order for some approximations to hold, but the choice will be independent of N, and in particular, does not decrease with N.

We define a function $K_{\sigma,x} \in \mathcal{K}_\sigma$ as follows: let

$$K_{\sigma,x}(y) \;\equiv\; K\left(\frac{y-x}{\sigma}\right)$$

where $y \in \mathbb{R}^N$, and x is in a compact set D in $\mathbb{R}^N$ (to be specified later), and K is a real valued function chosen as

$$K(y) \equiv K_1(y_1)K_1(y_2)\cdots K_1(y_N)$$

where $K_1(y_1)$ is an $(r-1)^{th}$ degree polynomial which is orthogonal to $y_1, {y_1}^2, \ldots, {y_1}^{r-1}$ and such that

$$\int_{-1}^{1} K_1(y_1)\, dy_1 = 1.$$

(The subscript in K_1 indicates that it is a function on $\mathbb{R}^1$.) We later describe the reason for this choice and its construction in detail. Define the estimate $f_n(x)$ as the empirical mean of the function

$$\sigma^{-N} K_{x,\sigma}(\cdot)$$

i.e.,

$$f_n(x) \;\equiv\; \frac{1}{n}\sum_{i=1}^{n} \sigma^{-N} K\left(\frac{\zeta_i - x}{\sigma}\right)$$

where

$$K_{\sigma,x}(y) \equiv K\left(\frac{y-x}{\sigma}\right), \qquad y, x \in \mathbb{R}^N, \qquad \sigma \in \mathbb{R}.$$

(Note, we use a double subscript for $K_{\sigma,x}$ which indicates it is not the same function as the one dimensional kernel K_1.) We treat x as a constant, acting as the index of the function in the class $\mathcal{K}_\sigma$, while the only randomness is in the sample $\zeta_1, \zeta_2, \ldots, \zeta_n$. Clearly $f_n(x)$ is a random variable with expected value $\bar{f}(x) = \mathbf{E}(\sigma^{-N} K_{x,\sigma})$. The bias of the estimate is then

$$\frac{\sup_x |\bar{f}(x) - f(x)|}{\sup_x f(x)}.$$

We can express the error of the estimate in terms of the bias, i.e.,

$$\frac{\sup_x |f_n(x) - f(x)|}{\sup_x f(x)} \leq \frac{\sup_x |\bar{f}(x) - f(x)|}{\sup_x f(x)} + \frac{\sup_x |f_n(x) - \bar{f}(x)|}{\sup_x f(x)}.$$

In the current context, $f(x)$ is Gaussian, hence $\sup_x f(x) = 1/(2\pi)^{N/2}$. The bias is nonrandom and, as we later show, decreases to zero as $\sigma \rightarrow 0$. The learner aims at reducing the kernel-window but not too fast (w.r.t n) because the probability of the second error component decreases at a rate which becomes worse (i.e., slower) if $\sigma \rightarrow 0$ too fast with n. The second component is random and is the deviation of the empirical mean from the true mean of $K_{x,\sigma}$ which we can bound (after we partition the domain of the indexing-parameter into D and D^c) using the uniform SLLN over the class $\mathcal{K}_\sigma$ of functions $K_{x,\sigma}$, $x \in D$. We start with the bias component.

5.4.1 The bias component

As will be seen shortly, the bias term can be made smaller by constructing a kernel $K_1(x_1)$ taking both negative and positive values, and which is orthogonal to $x_1, {x_1}^2, \ldots, {x_1}^{r-1}$. We first define the one-dimensional kernel. Let r be an even integer and

$$K_1(x_1) = \begin{cases} \sum_{i=0}^{r-1} a_i x_1^i & |x_1| \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

The a_i are chosen as the solution of the r equations,

$$\begin{array}{lllllll} a_0(x_1^0, x_1^0) & + & a_1(x_1^0, x_1^1) & + & \cdots & + & a_{r-1}(x_1^0, x_1^{r-1}) & = & 1 \\ a_0(x_1^1, x_1^0) & + & a_1(x_1^1, x_1^1) & + & \cdots & + & a_{r-1}(x_1^1, x_1^{r-1}) & = & 0 \\ & & & & \cdots & & & & \\ a_0(x_1^{r-1}, x_1^0) & + & a_1(x_1^{r-1}, x_1^1) & + & \cdots & + & a_{r-1}(x_1^{r-1}, x_1^{r-1}) & = & 0 \end{array}$$

where

$$(f, g) \equiv \int_{-1}^{1} f(x_1) g(x_1)\, dx_1.$$

REMARK: $(x_1^i, x_1^j) = 0$ if $i + j$ is odd, and $(x_1^i, x_1^j) = \frac{2}{(i+j+1)}$ if $i + j$ is even.

Denote the solution vector as a, and the matrix of the dot products as A. There always exist a solution vector a for any chosen r since the matrix A has a nonvanishing determinant. That is because the quadratic form

$$(b, Ab) = \int_{-1}^{1} \sum_{i,j=0}^{r-1} b_i x_1^i b_j x_1^j \, dx_1 = \int_{-1}^{1} \Bigl(\sum_{ij=0}^{r-1} b_i x_1^i \Bigr)^2 > 0$$

whenever $b \neq 0$. Our one-dimensional kernel function $K_1(x_1)$ satisfies $(K_1(x_1), x_1^i) = 0$ for $1 \leq i \leq r-1$, and $(K_1(x_1), 1) = 1$. Before we show the effect of r on the bias, we prove an upper bound on the magnitude of this kernel; this will be used later when we require uniform SLLN convergence for this class of kernel functions. We look for an upper bound on $|K_1(x_1)|$. From Szego & Polya [25, page 89], for any arbitrary polynomial $P(x_1)$ of r^{th} degree with real coefficients such that

$$\int_{-1}^{1} (P(x_1))^2 \, dx_1 = 1,$$

we have

$$|P(x_1)| \leq \frac{r+1}{\sqrt{2}},$$

uniformly for all $-1 \leq x_1 \leq 1$. For the proof see Lemma 5.2 in Section 5.3. In our case the polynomial $K_1(x_1)$ is of degree $r-1$ and satisfies

$$\int_{-1}^{1} (K_1(x_1))^2 \, dx_1 = (a, Aa) = (a, [10\ldots 0]^t) = a_0.$$

Hence

$$\int_{-1}^{1} \left(\frac{K_1(x_1)}{\sqrt{a_0}} \right)^2 dx_1 = 1$$

so that

$$|K_1(x_1)| \leq r\sqrt{\frac{a_0}{2}}$$

for $|x_1| \leq 1$. Now we calculate a_0. Without loss of generality, take r to be even. By Cramér's rule we have

$$a_0 = \frac{2^{r-1} \begin{vmatrix} \frac{1}{3} & 0 & \frac{1}{5} & \cdots & \frac{1}{r-1} & 0 & \frac{1}{r+1} \\ 0 & \frac{1}{5} & \cdots & \frac{1}{r-1} & 0 & \frac{1}{r+1} & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \frac{1}{r+1} & \cdots & 0 & 0 & \frac{1}{2r-3} & 0 \\ \frac{1}{r+1} & 0 & \cdots & 0 & \frac{1}{2r-3} & 0 & \frac{1}{2r-1} \end{vmatrix}}{2^{r} \begin{vmatrix} 1 & 0 & \frac{1}{3} & 0 & \frac{1}{5} & \cdots & \frac{1}{r-1} & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{5} & \cdots & \frac{1}{r-1} & 0 & \frac{1}{r+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{1}{r-1} & 0 & \frac{1}{r+1} & 0 & \cdots & 0 & \frac{1}{2r-3} & 0 \\ 0 & \frac{1}{r+1} & 0 & \cdots & \cdots & \cdots & 0 & \frac{1}{2r-1} \end{vmatrix}}. \tag{5.6}$$

With some manipulations (see Lemma 5.3) and for even r, we obtain

$$a_0 = \frac{1}{2}\left(\frac{(r-1)(r-3)\cdots(3)}{(r-2)(r-4)\cdots(4)(2)}\right)^2 = \frac{r^2}{2}\left(2^{-r}\binom{r}{r/2}\right)^2$$

whence from Stirling's formula applied to the central term of the binomial, we obtain $a_0 \sim \frac{r}{\pi}$ as $r \to \infty$ through the even integers. Consequently,

$$|K_1(x_1)| \leq r\sqrt{\frac{a_0}{2}} \sim \frac{r^{3/2}}{\sqrt{2\pi}}, \qquad (r \to \infty).$$

With $x = [x_1, x_2, \ldots, x_N] \in \mathbb{R}^N$, from before we have the N-dimensional kernel as

$$K(x) = K_1(x_1)K_1(x_2)\cdots K_1(x_N).$$

We now show the effect of r on the bias by expressing the bias as follows (all integrals are over $\mathbb{R}^N$ unless explicitly specified):

$$\bar{f}(x) - f(x) \ = \ \int \sigma^{-N} K\left(\frac{y-x}{\sigma}\right) f(y)dy - f(x)\int K(y)dy$$

since

$$\int K(y)dy = \int_{-1}^{1} K_1(y_1)dy_1 \int_{-1}^{1} K_1(y_2)dy_2 \cdots \int_{-1}^{1} K_1(y_N)dy_N = 1.$$

Changing the variable of integration to $z = \frac{y-x}{\sigma}$ we obtain

$$\bar{f}(x) - f(x) = \int K(z)(f(x+\sigma z) - f(x))dz.$$

We expand $f(x+\sigma z)$ in a Taylor series around $\sigma = 0$. The bias becomes

$$\begin{aligned} \sup_x \left|\bar{f}(x) - f(x)\right| (2\pi)^{N/2} = (2\pi)^{N/2} \sup_x \bigg| \int K(z) \bigg(f(x) + \sigma \sum_{i_1=1}^{N} z_{i_1} f_{i_1}^{(1)}(x) \\ + \quad \frac{\sigma^2}{2} \sum_{i_1,i_2=1}^{N} z_{i_1} z_{i_2} f_{i_1,i_2}^{(2)}(x) + \cdots \\ + \quad \frac{\sigma^r}{r!} \sum_{i_1,i_2,\ldots,i_r=1}^{N} z_{i_1} z_{i_2} \cdots z_{i_r} f_{i_1,i_2,\ldots,i_r}^{(r)}(x+cz) - f(x)\bigg) dz \bigg| \end{aligned}$$

$$
\begin{aligned}
\leq \quad & (2\pi)^{N/2} \sup_x \left| \sigma \sum_{i_1=1}^{N} f^{(1)}_{i_1}(x) \int K(z) z_{i_1} dz \right| \\
+ \quad & (2\pi)^{N/2} \sup_x \left| \frac{\sigma^2}{2} \sum_{i_1,i_2=1}^{N} f^{(2)}_{i_1,i_2}(x) \int K(z) z_{i_1} z_{i_2} dz \right| + \cdots \\
+ \quad & (2\pi)^{N/2} \sup_x \left| \frac{\sigma^r}{r!} \sum_{i_1,i_2,\ldots,i_r=1}^{N} \int f^{(r)}_{i_1,i_2,\ldots,i_r}(x+cz) K(z) z_{i_1} z_{i_2} \cdots z_{i_r} dz \right|
\end{aligned}
$$

where $0 \leq c \leq \sigma$. Now, by Lemma 5.4 in Section 5.3

$$
|f^{(k)}_{u_{i_1} u_{i_2} \ldots u_{i_k}}| \leq M_k
$$

uniformly over $i_1, i_2, \ldots i_k \in \{1, 2, \ldots N\}$ for $k \leq r+1$ where

$$
M_k = C(2\pi)^{-N/2} k^{\frac{k}{2}} e^k \tag{5.7}
$$

with C an absolute positive constant. Then by the mean value theorem we have

$$
|f^{(r)}(y) - f^{(r)}(x)| \leq M_{r+1} |y - x|
$$

implying uniform continuity of $f^{(r)}$ hence

$$
\forall \delta > 0 \;\; \exists\, c(\delta) \ni |f^{(r)}(x + c(\delta) z) - f^{(r)}(x)| \leq \delta_1
$$

where we view $f^{(r)}$ as a function of x, while $c(\delta_1)z$ as a small deviation. It suffices that $|zc(\delta_1)| = \delta_1 / M_{r+1}$. Now $|z| \leq 2\sqrt{N}$ as z ranges over $[-1, +1]^N$ only. To achieve δ_1 deviation the lowest necessary c is $\delta_1 / 2\sqrt{N} M_{r+1}$. With this choice we can write

$$
f^{(r)}(x + cz) \leq f^{(r)}(x) + 2\sqrt{N} M_{r+1} \sigma \leq M_r + \delta_1
$$

as $0 \leq c \leq \sigma$, for a suitably small choice of $\sigma < \delta_1 / 2\sqrt{N} M_{r+1}$ (from (5.7) we have $M_{r+1} < \infty$). To avoid carrying the nuisance factor of δ_1 around, increase each M_r

slightly to include an extra δ_1. The bias is now bounded above by

$$\begin{aligned}\sup_x \left|\bar{f}(x) - f(x)\right| (2\pi)^{N/2} &\leq (2\pi)^{N/2} M_1 \sigma \sum_{i_1=1}^{N} \left|\int K(z) z_{i_1} dz\right| \\ &+ (2\pi)^{N/2} M_2 \frac{\sigma^2}{2} \sum_{i_1,i_2=1}^{N} \left|\int K(z) z_{i_1} z_{i_2} dz\right| + \cdots + \\ &+ (2\pi)^{N/2} \frac{\sigma^r}{r!} M_r \sum_{i_1,i_2,\ldots,i_r=1}^{N} \left|\int K(z) z_{i_1} z_{i_2} \cdots z_{i_r} dz\right|.\end{aligned}$$

Using the orthogonality of $K_1(z_1)$ to the first $r-1$ powers of z_1, only the last term survives so that

$$\begin{aligned}&\sup_x |\bar{f}(x) - f(x)|(2\pi)^{N/2} \\ &= (2\pi)^{N/2} M_r \frac{\sigma^r}{r!} \sum_{i_1,i_2,\ldots,i_r=1}^{N} \left|\int K(z) z_{i_1} z_{i_2} \cdots z_{i_r} dz\right| \\ &= (2\pi)^{N/2} M_r \frac{\sigma^r}{r!} \left(\left|\int K_1(z_1) z_1^r dz_1\right| + \left|\int K_1(z_2) z_2^r dz_2\right| + \cdots + \left|\int K_1(z_N) z_N^r dz_N\right| \right) \\ &\leq (2\pi)^{N/2} N c_2 M_r \frac{\sigma^r}{r!}\end{aligned}$$

where, for any $1 \leq i \leq N$, $\left|\int_{-1}^{1} K_1(z_i) z_i^r \, dz_i\right| = c_2$ is an absolute positive constant as

$$\begin{aligned}\left|\int_{-1}^{1} K_1(z_i) z_i^r \, dz_i\right| &\leq \int_{-1}^{1} |K_1(z_i) z_i^r| \, dz_i \\ &\leq \sqrt{\int_{-1}^{1} K_1^2(z_i) \, dz_i} \sqrt{\int_{-1}^{1} z_i^{2r} \, dz_i} \\ &= \sqrt{a_0} \sqrt{\frac{2}{2r+1}} \leq c_2\end{aligned}$$

from Lemma 5.3 for a_0. Now, Lemma 5.4 shows that

$$M_r \leq C(2\pi)^{-N/2} r^{\frac{r}{2}} e^r. \qquad (5.8)$$

The bias is hence bounded above by

$$c_3 r^{\frac{r}{2}} e^r \frac{\sigma^r}{r!}$$

for some positive constant c_3. Using Stirling's formula for $r!$, we see that the latter goes to 0 as $r \to \infty$ given that $\sigma < 1$. (Note that the condition on σ from before translates into the requirement $\sigma \leq c_4(2\pi)^{N/2}N^{-1/2}r^{-r/2}e^{-r}$.)

5.4.2 The random part of the error

We now treat the second component of the error, i.e.

$$\frac{\sup_x |f_n(x) - \bar{f}(x)|}{\sup_x f}.$$

We partition the domain of x, the index-parameter, and apply the uniform SLLN over the compact part of the partition. Let D be a compact subset of $\mathbb{R}^N$ to be specified. We have

$$\begin{aligned}
&\mathbf{P}\left(\frac{\sup_x |f_n(x) - \bar{f}(x)|}{\sup_x f} > \epsilon\right) \\
\leq\ & \mathbf{P}\left(\frac{\sup_{x\in D} |f_n(x) - \bar{f}(x)|}{\sup_x f} + \frac{\sup_{x\in D^c} |f_n(x) - \bar{f}(x)|}{\sup_x f} > \epsilon\right) \\
\leq\ & \mathbf{P}\left(\frac{\sup_{x\in D} |f_n(x) - \bar{f}(x)|}{\sup_x f} > \epsilon/2\right) + \mathbf{P}\left(\frac{\sup_{x\in D^c} |f_n(x) - \bar{f}(x)|}{\sup_x f} > \epsilon/2\right)
\end{aligned} \tag{5.9}$$

where the first term on the right is ready for application of the uniform SLLN. We first show that the second term is subdominant. We have

$$\begin{aligned}
\mathbf{P}\left(\frac{\sup_{x\in D^c} |f_n(x) - \bar{f}(x)|}{\sup_x f} > \epsilon/2\right) &\leq \mathbf{P}\left(\frac{\sup_{x\in D^c} |f_n(x)|}{\sup_x f} + \frac{\sup_{x\in D^c} \left|\bar{f}(x)\right|}{\sup_x f} > \epsilon/2\right) \\
= \ \mathbf{P}&\left(\frac{\sup_{x\in D^c} \left|\frac{1}{n}\sum_{i=1}^n \frac{1}{\sigma^N} K_{\sigma,x}(\zeta_i)\right|}{\sup_x f} + \frac{\sup_{x\in D^c} \left|\int \frac{1}{\sigma^N} K_{\sigma,x}(y) f(y)\, dy\right|}{\sup_x f} > \epsilon/2\right).
\end{aligned} \tag{5.10}$$

Now suppress the σ to simplify notation, and choose

$$D \equiv \{x : |x - \theta_{01}| \leq A, \text{ or } |x - \theta_{02}| \leq A\}$$

where A will be specified shortly. (We denote by $poly(x)$ an r^{th} degree polynomial in $x_1, x_2, \ldots, x_N$.) We have

$$\begin{aligned}
\sup_{x \in D^c} \left| \int K_{\sigma,x}(y) f(y)\, dy \right| &= \sup_{|x-\theta_{01}|>A \cap |x-\theta_{02}|>A} \left| \int poly(y-x) 1_{|y-x|\le 1} f(y)\, dy \right| \\
&\le \sup_{|x-\theta_{01}|>A \cap |x-\theta_{02}|>A} \left| \int poly(y-x) 1_{|y-x|\le 1} f_1(y|\theta_{01})\, dy \right| \\
&\quad + \sup_{|x-\theta_{01}|>A \cap |x-\theta_{02}|>A} \left| \int poly(y-x) 1_{|y-x|\le 1} f_2(y|\theta_{02})\, dy \right| \\
&\le \sup_{|x-\theta_{01}|>A} \int |poly(y-x)|\, 1_{|y-x|\le 1} f_1(y|\theta_{01})\, dy \\
&\quad + \sup_{|x-\theta_{02}|>A} \int |poly(y-x)|\, 1_{|y-x|\le 1} f_2(y|\theta_{02})\, dy \\
&= \sup_{|x-\theta_{01}|>A} \int_{|y-\theta_{01}|>A-1} |poly(y-x)|\, 1_{|y-x|\le 1} \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}|y-\theta_{01}|^2}\, dy \\
&\quad + \sup_{|x-\theta_{02}|>A} \int_{|y-\theta_{02}|>A-1} |poly(y-x)|\, 1_{|y-x|\le 1} \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}|y-\theta_{02}|^2}\, dy \\
&\le c_4 \frac{r^{\frac{3N}{2}}}{(2\pi)^{N/2}} \int_{|y-\theta_{01}|>A-1} \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}|y-\theta_{01}|^2}\, dy \\
&\quad + c_4 \frac{r^{\frac{3N}{2}}}{(2\pi)^{N/2}} \int_{|y-\theta_{02}|>A-1} \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}|y-\theta_{02}|^2}\, dy.
\end{aligned}$$

From page 66 these two integrals are bounded above by $c2Ne^2/A^4$ for some positive constant c. Therefore

$$\frac{\sup_{x \in D^c} \left| \int \frac{1}{\sigma^N} K_{\sigma,x}(y) f(y)\, dy \right|}{\sup_x f} = \frac{c_5 r^{3N/2} N}{\sigma^N A^4} \equiv \Delta.$$

Applying Markov's inequality to (5.10) we now have

$$\begin{aligned}
&\mathbf{P}\left(\frac{\sup_{x \in D^c} \left| \frac{1}{n} \sum_{i=1}^n \frac{1}{\sigma^N} K_{\sigma,x}(\zeta_i) \right|}{1/(2\pi)^{N/2}} > \frac{\epsilon}{2} - \Delta \right) \\
&\le \mathbf{P}\left(\frac{\sup_{x \in D^c} \frac{1}{n} \sum_{i=1}^n \left| \frac{1}{\sigma^N} K_{\sigma,x}(\zeta_i) \right|}{1/(2\pi)^{N/2}} > \frac{\epsilon}{2} - \Delta \right) \le \frac{\mathbf{E} \frac{1}{n} \sum_{i=1}^n \sup_{x \in D^c} \left| \frac{1}{\sigma^N} K_{\sigma,x}(\zeta_i) \right|}{\left| (\frac{\epsilon}{2} - \Delta)/(2\pi)^{N/2} \right|} \\
&= \frac{\frac{1}{\sigma^N} \int_{|y|>A-1} \sup_{x \in D^c} \left| poly(y-x) 1_{|y-x|\le 1} f(y)\, dy \right|}{\left| (\frac{\epsilon}{2} - \Delta)/(2\pi)^{N/2} \right|} \le \frac{\Delta}{\left| \frac{\epsilon}{2} - \Delta \right|} = \frac{\Delta}{\frac{\epsilon}{2} - \Delta}
\end{aligned}$$

since for any ϵ, σ, and N we can choose A large enough to make Δ suitably small compared to ϵ. We hence obtain

$$\mathbf{P}\left(\frac{\sup_{x\in D^c}\left|f_n(x)-\bar{f}(x)\right|}{\sup_x f} > \frac{\epsilon}{2}\right) \leq c_6\Delta = \frac{c_7 N r^{3N/2}}{A^4} \equiv \delta/2$$

where A is chosen accordingly. Hence the second term on the right of (5.9) is $\leq \delta/2$. Now we estimate the size of n needed to make the first term of (5.9) less than $\delta/2$.

The stage is set for an application of Theorem 3.10. The class $\mathcal{K}_\sigma$ of functions is the set $\{K_{\sigma,x}, x \in D\}$. These functions are uniformly bounded by choice of compact D (note that σ is permitted to decrease with n). Lemma 5.6 shows that this class is finitely coverable which is a condition for its permissibility (needed by the theorem). We begin by bounding $\mathbf{E}K^2_{\sigma,x}$. We have

$$\mathbf{E}\,K^2_{x,\sigma}(y) = \int f(y)K^2_{x,\sigma}(y)dy = \sigma^N \int f(x+\sigma z)K^2(z)dz$$

with the change of variable $z = (y-x)/\sigma$. In our case f is a multi-multivariate Gaussian mixture and

$$f \leq 1/(2\pi)^{N/2}. \tag{5.11}$$

The bound hence becomes

$$\begin{aligned}
\mathbf{E}\,K^2_{x,\sigma} &\leq (\sigma^N/(2\pi)^{N/2})\int K^2(z)dz \\
&= (\sigma^N/(2\pi)^{N/2})\int_{-1}^{1} K_1^2(z_1)dz_1 \int_{-1}^{1} K_1^2(z_2)dz_2 \cdots \int_{-1}^{1} K_1^2(z_N)dz_N \\
&\leq (\sigma^N/(2\pi)^{N/2})\left(\frac{r^2}{2}\left(2^{-r}\,\tbinom{r}{r/2}\,\right)^2\right)^N \\
&\leq (\sigma^N/(2\pi)^{N/2})\left(\frac{r}{\pi}\right)^N \\
&= \left(\frac{\sigma r}{\sqrt{2\pi^3}}\right)^N.
\end{aligned}$$

We let this be δ_n^2 in the theorem. In order to satisfy the conditions of the theorem, we need to select σ as functions of n such that

$$\frac{\log n/n}{{\delta_n}^2} \to 0 \qquad (n \to \infty).$$

This is satisfied if

$$n \geq \left(\frac{\sqrt{2\pi^3}}{\sigma r}\right)^N \log^2 \left(\frac{\sqrt{2\pi^3}}{\sigma r}\right)^N .$$

We will see that this condition is trivially met in our subsequent choice of n. Now for M in the theorem, we bound the functions in the class as

$$|K_{\sigma,x}(z)| = \prod_{i=1}^{N} \left|K_1\left(\frac{x_i - z_i}{\sigma}\right)\right| \leq \left(r\sqrt{\frac{a_0}{2}}\right)^N \leq \frac{r^{3N/2}}{(2\pi)^{N/2}} \equiv M$$

by again using Lemma 5.3. We thus have

$$\begin{aligned}
&\mathbf{P}\left(\sup_{x\in D}\left|\frac{1}{n}\sum_{i=1}^{n} K\left(\frac{\zeta_i - x}{\sigma}\right) - \mathbf{E}\, K\left(\frac{\zeta_1 - x}{\sigma}\right)\right| > \epsilon\frac{\delta_n^2}{M}\right) \\
&= \mathbf{P}\left(\sup_{x\in D}\left|\frac{1}{n}\sum_{i=1}^{n} K\left(\frac{\zeta_i - x}{\sigma}\right) - \mathbf{E}\, K\left(\frac{\zeta_1 - x}{\sigma}\right)\right| > \epsilon\sigma^N r^{-N/2}\pi^{-N}\right) \\
&= \mathbf{P}\left(\sup_{K_{\sigma,x}\in\mathcal{K}_\sigma}\left|\frac{1}{n}\sum_{i=1}^{n} K_{\sigma,x}(\zeta_i) - \mathbf{E}\, K_{\sigma,x}(\zeta_1)\right| > \epsilon\sigma^N r^{-N/2}\pi^{-N}\right) \\
&\leq 24\left(\frac{32e}{\epsilon\left(\sqrt{\frac{2}{\pi}}\frac{\sigma}{r^2}\right)^N} \log \frac{32e}{\epsilon\left(\sqrt{\frac{2}{\pi}}\frac{\sigma}{r^2}\right)^N}\right)^d e^{-n\,\epsilon^2\left(\sqrt{\frac{2}{\pi}}\frac{\sigma}{r^2}\right)^N/8192} \qquad (5.12)
\end{aligned}$$

The left hand side is equivalent to

$$\begin{aligned}
\mathbf{P}\left(\sup_{x\in D}\left|f_n(x) - \bar{f}(x)\right| > \epsilon r^{-N/2}\pi^{-N}\right) &= \mathbf{P}\left(\frac{\sup_{x\in D}|f_n - \bar{f}|}{\sup_x f} > \epsilon r^{-N/2}\pi^{-N}(2\pi)^{N/2}\right) \\
&= \mathbf{P}\left(\frac{\sup_{x\in D}|f_n - \bar{f}|}{\sup_x f} > \epsilon\left(\frac{2}{r\pi}\right)^{N/2}\right).
\end{aligned}$$

Redefining ϵ as $\epsilon\left(\frac{2}{r\pi}\right)^{N/2}$ and calling the right hand side of (5.12) $\delta/2$ we obtain

$$\mathbf{P}\left(\frac{\sup_{x\in D}|f_n(x) - \bar{f}(x)|}{\sup_x f(x)} > \epsilon\right) \leq \delta/2$$

and hence combining with the previous result we have

$$\mathbf{P}\left(\frac{\sup_x |f_n(x) - \bar{f}(x)|}{\sup_x f(x)} > \epsilon\right) \leq \delta,$$

when

$$n \geq \frac{c_8 r^N (2/\pi)^{N/2}}{\sigma^N \epsilon^2} \left(d \log \frac{r^{3N/2}}{\epsilon \sigma^N} + \log \frac{1}{\delta} \right).$$

We can use this bound once we calculate $d = VC(\mathcal{K}_\sigma)$ which is obtained (see Definition 3.5) by noticing that $\mathcal{K}_\sigma$ is class of functions that are linear combinations of a finite basis of functions (as in Theorem 3.6). It will transpire that we can easily calculate the *VC-dimension* of the graphs of such functions (Definition 3.4) which by Definition 3.5 is VC($\mathcal{K}_\sigma$).

Define $\mathcal{F}$ as the class of graphs of the N-dimensional kernels $K_{\sigma,x} \in \mathcal{K}_\sigma$ where from before we have $K_{\sigma,x}(y) \equiv K(\frac{y-x}{\sigma})$. Recall that, by definition, K has a compact support $[-1,1]^N$. Each function in $\mathcal{K}_\sigma$, and hence each graph in $\mathcal{F}$, has the same fixed σ but has a different N-dimensional vector x which indexes it in the class. The graphs are sets in $\mathbb{R}^N \times \mathbb{R}$ since a function $K_{\sigma,x}(y)$ is a mapping from $\mathbb{R}^N$to $\mathbb{R}$. By Definition 3.5, VC($\mathcal{K}_\sigma$) = VC($\mathcal{F}$) so our aim is to find VC($\mathcal{F}$). Replace x with θ to indicate the parameter indexing a function and let x now denote the domain of the function, i.e., $K_{\sigma,\theta}(x)$. The one-dimensional kernel equals

$$\sum_{i=0}^{r-1} a_i (x_1 - \theta_1)^i 1_{|x_1 - \theta_1| \leq 1}, \quad x_1, \theta_1 \in \mathbb{R}.$$

This is an $(r-1)^{th}$-degree polynomial in x_1 whose coefficients are comprised of the θ_1^k and a_k. Hence we write the N-dimensional kernel as a product

$$K_{\sigma,\theta}(x) = poly^{r-1}(x_1) poly^{r-1}(x_2) \cdots poly^{r-1}(x_N) 1_{|x_1-\theta_1|\leq 1} 1_{|x_2-\theta_2|\leq 1} \cdots 1_{|x_N-\theta_N|\leq 1}$$

where $poly^{r-1}(\cdot)$ denotes a polynomial of degree $r-1$ in a single variable x_i. Denote by

$$p(x) \equiv poly^{r-1}(x_1) poly^{r-1}(x_2) \cdots poly^{r-1}(x_N).$$

Write

$$B \equiv r\sqrt{\frac{a_0}{2}} \geq |K_1(x_1)|$$

for the bound on the magnitude of the one dimensional kernel. Use the notation $|x - \theta| \leq \underline{1}$ to denote the set based on the *vector* inequality

$$\{|x_i - \theta_i| \leq 1, \ \ 1 \leq i \leq N\} ;$$

the function $1_{|x-\theta|\leq\underline{1}}$ then connotes the indicator for the cube (in $\mathbb{R}^N$) of side 2 at θ. Let $\mathcal{G}_\pm$ denote the graphs of the functions $\pm B 1_{|x-\theta|\leq\underline{1}}$ respectively. Also, let

$$\mathcal{P}_+ = \{(x,y) : 0 \leq y \leq p(x)\}$$

$$\mathcal{P}_- = \{(x,y) : p(x) \leq y < 0\} .$$

The graph of $K_{\sigma,\theta}(x)$ is then represented simply by $(\mathcal{P}_+ \cap \mathcal{G}_+) \cup (\mathcal{P}_- \cap \mathcal{G}_-)$. Our aim now is to express this set by intersection/unions of sets of the form

$$\left\{(x,y) : \sum_i a_i \phi_i(x,y) > 0\right\}$$

where the sum is finite. Then we can directly apply Theorem 3.6 to find the VC-dimension of $K_{\sigma,\theta}(x)$. We first construct a function

$$h_{p,a}(x,y) \equiv p(x) - ay$$

where a is a real scalar. We have

$$\{(x,y) : h_{p,1}(x,y) \geq 0\} \bigcap \{(x,y) : y \geq 0\} = \{(x,y) : 0 \leq y \leq p(x)\}$$

and

$$\{(x,y) : h_{-p,-1}(x,y) \geq 0\} \bigcap \{(x,y) : y < 0\} = \{(x,y) : p(x) \leq y < 0\}$$

We can index any function of the form

$$p(x) = poly^{r-1}(x_1) poly^{r-1}(x_2) \cdots poly^{r-1}(x_N)$$

using the basis

$$\{\{1, x_1, x_1^2, \ldots, x_1^{r-1}\} \times \{1, x_2, x_2^2, \ldots, x_2^{r-1}\} \times \cdots \times \{1, x_N, x_N^2, \ldots, x_N^{r-1}\}\}.$$

This basis has cardinality r^N. Hence the function $h_{p,a}(x, y)$ can be expressed as a linear combination of $r^N + 1$ terms. By Theorem 3.6 it follows that the VC dimension, d, of the class, $\mathcal{H}$, of sets $\{(x, y) : h_{p,a}(x, y) > 0\}$ is at most $r^N + 1$. The class, $\mathcal{H}'$, of intersections of such sets with $\{(x, y) : y \geq 0\}$ can pick out at most the same number of dichotomies of an m-sample as $\mathcal{H}$ does. To see this, note that if (x^i, y^i), $1 \leq i \leq m$ is any m-sample shattered by $\mathcal{H}'$ then *necessarily* we must have $y^i \geq 0$ for each i. Now, take any dichotomy of this sample, say the one achieved by a set $A' \equiv A \cap \{(x, y) : y \geq 0\}$ which is an element of the class of sets $\mathcal{H}'$. (Note, the set A is in $\mathcal{H}$). Clearly, the set A achieves the same dichotomy of the sample. Hence $\mathcal{H}$ must shatter this m-sample. Hence $VC(\mathcal{H}) \geq VC(\mathcal{H}')$.

So therefore the VC dimension of the class of sets

$$\{(x, y) : h_{p,a}(x, y) > 0\} \bigcap \{(x, y) : y \geq 0\} = \{(x, y) : 0 \leq y \leq p(x)\}$$

is at most $r^N + 1$ and likewise the VC dimension of the class of sets

$$\{(x, y) : h_{-p,a}(x, y) > 0\} \bigcap \{(x, y) : y < 0\} = \{(x, y) : p(x) \leq y < 0\}$$

is at most $r^N + 1$. Continuing, we have by definition,

$$\mathcal{G}_+ = \bigcap_i \left\{(x, y) \in \mathbb{R}^N \times \mathbb{R} : 0 \leq y \leq B \cdot 1_{|x_i - \theta_i| \leq 1}\right\}.$$

It suffices hence to estimate the VC dimension of the class of sets

$$\{(x, y) : 0 \leq y \leq B \cdot 1_{|x_1 - \theta_1| \leq 1}\}.$$

Define

$$\alpha(y) = \begin{cases} 1 & \text{if } 0 \leq y \leq B \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to see that

$$\{(x,y) : 0 \le y \le B \cdot 1_{|x_1-\theta_1|}\} = \{(x,y) : \alpha(y) - (x_1-\theta_1)^2 > 0\}.$$

The function $\alpha(y) - (x_1-\theta_1)^2$ is a linear combination of the function basis $\{x_1^2, x_1, 1, \alpha(y)\}$. Hence by Theorem 3.6 the class of sets $\{(x,y) : \alpha(y) - (x_1-\theta_1)^2 > 0\}$ and therefore the class of sets $\{(x,y) : 0 \le y \le B \cdot 1_{|x_1-\theta_1|}\}$ has VC dimension ≤ 4. This is true for every one of the N classes, $\{(x,y) : 0 \le y \le B \cdot 1_{|x_i-\theta_i|}\}, 1 \le i \le N$. By Theorem 3.3, the number of dichotomies of an m-sample that is picked out by any such class is $\le m^4$. Hence the class of sets $\mathcal{G}_+$ can pick out at most m^{4N} dichotomies of any m-sample. It follows that the family of graphs $\mathcal{P}_+ \cap \mathcal{G}_+$ picks out at most m^{r^N+4N+1} dichotomies of any m-sample. An analogous treatment shows that the number of dichotomies of an m-sample picked out by the class of sets $\mathcal{P}_- \cap \mathcal{G}_-$ is at most m^{r^N+4N+1}. It follows that the class of sets

$$\mathcal{F} = (\mathcal{P}_+ \cap \mathcal{G}_+) \cup (\mathcal{P}_- \cap \mathcal{G}_-)$$

achieves no more than

$$m^{2r^N+8N+2}$$

subsets out of any collection of m points. Denote the exponent by $c = 2r^N + 8N + 2$. By definition, the VC-dimension of $\mathcal{F}$ is bounded above by the largest value of m for which

$$m^c \ge 2^m$$

whence direct computation shows

$$VC(\mathcal{F}) \le 1.37c \log_2 c < 2c \log c = (4r^N + 16N + 4)\log(2r^N + 8N + 2).$$

Note that $r^N \ge 4N+1$ when $r \ge 5 + \log N$ for every $N \ge 1$. For this range of r then,

$$VC(\mathcal{F}) \le 8r^N \log\left(4r^N\right) < 32r^{2N},$$

as $\log x < x$ for all x. We complete the bound on n by using the following bounds on d and M,

$$d \leq 32r^{2N},$$

$$M \leq \frac{r^{3N/2}}{(2\pi)^{N/2}},$$

and obtain that

$$n \geq \frac{c_{10}r^N (2/\pi)^{N/2}}{\sigma^N \epsilon^2} 32r^{2N} \log \frac{r^{3N/2}}{\epsilon\sigma^N} + \frac{c_{10}r^N (2/\pi)^{N/2}}{\sigma^N \epsilon^2} \log \frac{1}{\delta} \tag{5.13}$$

is sufficient in order to have

$$\mathbf{P}\left(\frac{\sup_x |f_n(x) - \bar{f}(x)|}{\sup_x f(x)} > \epsilon\right) \leq \delta/4$$

for δ, ϵ arbitrary positive, and σ sufficiently small as before.

Together with the bias we have that with the same n,

$$\frac{\sup_x |f_n(x) - f(x)|}{\sup_x f(x)} > \epsilon + c_3 r^{\frac{r}{2}} e^r \frac{\sigma^r}{r!} N$$

with probability $\leq \delta/4$. To simplify this we replace ϵ with $\sqrt{\epsilon}/2$ and replace $c_3 r^{\frac{r}{2}} e^r \frac{\sigma^r}{r!} N$ by $\sqrt{\epsilon}/2$ (to find σ) yielding total error of $\sqrt{\epsilon}$. First look at the first term (ignoring the less significant log() part) in the bound for n. We have

$$\frac{r^N (2/\pi)^{N/2} 2^{N/r} e^{r\frac{N}{r}} r^{(\frac{r}{2})(N/r)} \left(\frac{N}{r!}\right)^{N/r} 32r^{2N}}{\epsilon^{1+N/2r}}. \tag{5.14}$$

Now we are free to choose r, so let $r = 5 + \log N$. We have

$$\left(\frac{N}{(5+\log N)!}\right)^{N/(5+\log N)} \leq \left(\frac{N}{2^{5+\log N}}\right)^{N/(5+\log N)} \leq 1^{N/(5+\log N)} = 1.$$

So (5.14) is bounded by

$$\frac{c_{11}(2^{\frac{7}{2}})^{N\log(5+\log N)} \left(\frac{2e^2}{\pi}\right)^{N/2} 2^{N/(5+\log N)}}{\epsilon^{1+N/2\log N}}. \tag{5.15}$$

Now $2^{\frac{7}{2}} < 12$, and there exists a c_{12} such that for all $N \geq 1$,

$$\left(\frac{2e^2}{\pi}\right)^{N/2} \leq c_{12}\left(\frac{13}{12}\right)^{\frac{1}{3}N\log(5+\log N)}.$$

Use similar arguments for the $2^{N/(5+\log N)}$ to get that (5.14) is bounded by

$$\frac{c_{12}(12)^{N\log(5+\log N)}\left(\frac{13}{12}\right)^{\frac{2}{3}N\log(5+\log N)}}{\epsilon^{N/2\log N}}.$$

Now the less significant log() part in the first term of (5.13) is bounded by

$$c_{13}\frac{13}{12}^{N\log(5+\log N)/3}\log\frac{1}{\epsilon}.$$

It follows that (5.13) is bounded by

$$c_{14}\frac{13^{N\log(5+\log N)}}{\epsilon^{N/2\log N}}\log\frac{1}{\epsilon\delta}.$$

Hence

$$\frac{\sup_x |f_n(x) - f(x)|}{\sup_x f(x)} > \sqrt{\epsilon} \tag{5.16}$$

when

$$n \geq c_{14}\frac{13^{N\log(5+\log N)}}{\epsilon^{N/2\log N}}\log\frac{1}{\epsilon\delta}$$

with probability $\leq \delta/4$. We now need to see how this translates into the decision rule error by showing that the mode-estimates will also be close.

Since $f_n(x)$ may have many relative maxima we need to choose two such modes that can be used to estimate closely the modes of $f(x)$. This is achieved by the following procedure which is based on the more general algorithm K (see Section 5.5). The requirements that are necessary for this procedure is that $0 < \epsilon < b$ where b is proportional to $|\theta_{01} - \theta_{02}|$, and that $f(x)$ must have two modes (not just one); for this reason we need the requirement on the means, θ_{01} and θ_{02}, of $f(x)$ to satisfy $|\theta_{01} - \theta_{02}| > 2$ (see Lemma 5.5).

Consider first the case of $x \in \mathbb{R}$. Denote the true modes as η_1 and η_2 and without loss of generality let $\eta_2 \geq \eta_1$. We first describe how the mode estimates are calculated. For simplicity we proceed with the assumption that $\sup_x |f_n(x) - f(x)| \leq \epsilon/2$ and later replace that by $\frac{\sup_x |f_n(x)-f(x)|}{\sup_x f(x)} \leq \sqrt{\epsilon}$ as in (5.16). For small enough ϵ, the learner determines the maximum of $f_n(x)$ to be, say, closer to η_1 which puts it under the first hump of $f(x)$; denote it by $\hat{\eta}_1$ and this will be the estimate for η_1. Then the learner determines the x-coordinates, x_{1a} and x_{1b}, of the two points closest to $\hat{\eta}_1$, where a horizontal line through the point $(\hat{\eta}_1, f_n(\hat{\eta}_1) - 8\epsilon)$ cuts $f_n(x)$. Note, because $|f_n(x) - f(x)| \leq \epsilon$ it follows that regardless of where $\hat{\eta}_1$ is, we have $\eta_1 \in (x_{1a}, x_{1b})$; in fact that is the case even if the line is defined with 3ϵ instead of 8ϵ. This guarantees that $f(x)$ is decreasing, when moving away from x_{1a} or from x_{1b} by a small amount (again under the main assumption of small enough ϵ). Using this, together with the fact that $|f_n(x) - f(x)| \leq \epsilon$ we have that for all x s.t. $x \leq x_{1a}$ or $x \geq x_{1b}$ and such that x is closer to η_1 than η_2, then $f_n(x) \leq f_n(\hat{\eta}_1) - 6\epsilon$. We claim that there exists a point whose x-coordinate, x_2, is closer to η_2 than η_1 and $f_n(x_2) > f_n(\hat{\eta}_1) - 6\epsilon$. This can be seen by drawing a line through $(\hat{\eta}_1, f(\hat{\eta}_1) - 4\epsilon)$ which intersects $f(x)$ on the second hump at a point $(x_2, f(x_2))$. The reason it intersects is because $|f_n(\hat{\eta}_1) - f(\eta_1)| \leq 3\epsilon$ which implies $f_n(\hat{\eta}_1) - 4\epsilon \leq f(\eta_1) - \epsilon < f(\eta_1) = f(\eta_2)$. The former follows from the fact that

$$|f_n(\hat{\eta}_1) - f(\hat{\eta}_1)| \leq \epsilon$$

and

$$|f(\hat{\eta}_1) - f(\eta_1)| \leq 2\epsilon.$$

The former is trivial. The latter is true since assuming the contrary implies that $f_n(\hat{\eta}_1) < f_n(\eta_1)$ and this is clearly false since it contradicts the definition of $\hat{\eta}_1$, as being the $\sup_x f_n(x)$. Moreover, $f_n(x_2) \geq f(x_2) - \epsilon = f_n(\hat{\eta}_1) - 5\epsilon > f_n(\hat{\eta}_1) - 6\epsilon$, which proves the above claim.

So the learner needs only to use the horizontal line through $(\hat{\eta}_1, f_n(\hat{\eta}_1) - 8\epsilon)$, determine x_{1a}, x_{1b}, look for the maximum of $f_n(x)$ over $x \ni x \leq x_{1a}$ or $x \geq x_{1b}$ and let it be $\hat{\eta}_2$. This guarantees $\hat{\eta}_2$ is closer to η_2 than η_1, and hence the consistency of $\hat{\eta}_i$ to η_i as $\epsilon \to 0$. For $f(x)$ with $x \in \mathbb{R}^N$ a similar procedure can be applied to find the mode estimates.

Now we find the possible deviation for $\hat{\eta}_i$ from η_i where $i = 1, 2$. Without loss of generality let the means be $\mu_1 = 0$ and $\mu_2 = \mu$. The modes are η_1 and η_2. We look for the largest deviation, y, from η_1 such that it is possible for $f_n(\eta_1 + y) \geq f_n(\eta_1)$. This is the largest deviation for $\hat{\eta}_1$ from η_1 and by symmetry also for $\hat{\eta}_2$ from η_2 . When $\frac{\sup|f_n - f|}{\sup f} \leq \epsilon/2$ then at the point, $y + \eta_1$, we have $f(\eta_1) - f(y + \eta_1) = \epsilon/(2\pi)^{N/2}$. Therefore we look for the y such that $f(\eta_1) - f(y + \eta_1) > \epsilon/(2\pi)^{N/2}$ which implies that $\frac{\sup|f_n - f|}{\sup f} > \epsilon/2$. After some algebra we determine $y = c_{15}\epsilon$ for some positive constant c_{15}. Then, adhering to the statement of (5.16) we replace $\epsilon/2$ by $\sqrt{\epsilon}$ for the deviation, and get that $|f_n(\cdot) - f(\cdot)| \geq c_{16}\sqrt{\epsilon}$ if the maximum deviation from $\hat{\eta}_i$ to η_i, $i = 1, 2$, is $c_{17}\sqrt{\epsilon}$. So from last paragraph, it follows that $\{|\hat{\eta}_1 - \eta_1| > c_{17}\sqrt{\epsilon})$ or $|\hat{\eta}_2 - \eta_2| > c_{17}\sqrt{\epsilon}\}$ has probability at most $\delta/2$.

Lastly, the learner outputs the hyperplane orthogonal to the line between $\hat{\eta}_1$ and $\hat{\eta}_2$ as a decision rule. Since the Bayes border is between η_1 and η_2 then from Section 4.1, it follows that the above deviation yields a $P_{error} \leq P_{Bayes} + c_{18}\sqrt{\epsilon}^2 = P_{Bayes} + c_{18}\epsilon$ *if* the regions across the hyperplane are labeled correctly. Lastly, using the same analysis for the labeled sample complexity as in Section 4.3 but using $\sqrt{\epsilon}$ there instead of ϵ for the deviation of the mode-estimates we get that with $m = c_{19} \log \frac{1}{\delta}$ labeled examples and given that both mode-estimates are $c_{17}\sqrt{\epsilon}$-close to the true modes, then the probability of choosing a labeling (by the majority rule) with $P_{error} > P_{Bayes} + c_{20}\epsilon$ is at most $\delta/2$. Hence the probability is at most δ that the learner outputs a decision rule whose $P_{error} > P_{Bayes} + c_{20}\epsilon = P_{Bayes}(1 + c_{21}\epsilon)$. This completes the proof of the theorem. ∎

5.5 Mixture of a General Form

In the previous section we established the complexity of learning a decision rule for a problem based on a Gaussian mixture f using a procedure that estimates the modes of f. The requirements of this procedure, together with the resulting exponentially large n stated in Theorem 5.1 suggest that the procedure is powerful to handle a richer variety of mixtures. That is, we already saw in Chapter 4 that the Gaussian mixture can be learned with polynomial sized samples given that parametric side information is available. So having an exponential sized sample suggests that the technique may be powerful enough to learn problems based on a richer variety of mixtures.

In this section, we extend the intuition that modes can determine the Bayes border for a large nonparametric class (containing Gaussian mixtures) where the mixtures are not necessary identifiable. We define a family, $\mathcal{P}$, of classification problems, each specified by a pair of class conditional densities and having a Bayes border that is identified by the modes of the mixture corresponding to the specific problem. We denote by $\mathcal{F}$ the class of mixtures which is induced by $\mathcal{P}$.

First we will prescribe the general form for the pair of densities that a problem in $\mathcal{P}$ may have, through several conditions. One of the consequence of these conditions is that a sufficient mixed sample complexity for learning a decision rule for any problem in $\mathcal{P}$ is the same as that of Theorem 5.1. It is of no concern whether there are several different problems in $\mathcal{P}$ that are associated with the *same* mixture (in which case the mixture is not identifiable) because the hyperplane identified by this mixture is the Bayes border for all these problems. We will impose the conditions on the class $\mathcal{P}$ in the course of the discussion, rather than all at once, for better comprehension.

Secondly, we will describe the algorithm K and prove its consistent estimation of the modes of an $f \in \mathcal{F}$.

We now proceed with the description of the problem class $\mathcal{P}$. Let a classification problem in $\mathcal{P}$ be defined as having the following class conditional densities

$$f_{g,\theta_1}(x) = \frac{1}{c^N} g(|x - \theta_1|^2),$$

$$f_{g,\theta_2}(x) = \frac{1}{c^N} g(|x - \theta_2|^2),$$

for $\theta_i \in \mathbb{R}^N$ $(i = 1, 2)$ and where g is smooth, decreasing on $[0, \infty)$, bounded above by 1, satisfying

$$\sup_x \left| g^{(r)}_{x_{i_1}, x_{i_2}, \ldots, x_{i_r}}(|x|^2) \right| \leq c_1 c_2^r r^{\frac{r}{2}}$$

where $i_1, i_2, \ldots, i_r \in \{1, 2, \ldots, N\}$, c_1, c_2 are positive constants, and that the absolute value of the lower partial derivatives of order less than r is bounded by some positive constants uniformly over x. (If we use a bound of the form $c_3 r^r$ then the unlabeled sample complexity will differ from the one in Theorem 5.1 only in the constant raised to $N \log \log N$). An additional condition on g is that its induced mixture $f_{g,\theta} \in \mathcal{F}$, where $\theta = [\theta_1, \theta_2]$, must have at least two modes. (In our discussion below we show that the last condition is easily satisfied by many types of g.) For brevity, when there is no danger of misinterpretation, we will drop the subscript g, θ and refer to a mixture just as f.

The pattern classes in $\mathcal{P}$ have a mixture of the following form

$$f_{g,\theta}(x) = \frac{1}{c^N} \left(\frac{1}{2} g(|x - \theta_1|^2) + \frac{1}{2} g(|x - \theta_2|^2) \right)$$

which may have multiple modes (a mode is a local maximum) although $g(|x - \theta|^2)$ has a single mode at $x = \theta$. Now consider the region

$$\{x : g(|x - \theta_1|^2) = g(|x - \theta_2|^2)\}$$

which is the Bayes decision border of the problem. This region is equivalent to $\{x : |x-\theta_1| = |x-\theta_2|\}$, ($g(y)$ has an inverse as it is decreasing), which is a hyperplane passing through the midpoint $\frac{\theta_1+\theta_2}{2}$ and perpendicular to the line through θ_1 and θ_2. (As a consequence of the succeeding discussion it follows that the points θ_1 and θ_2 are *not* modes of the mixture f.) We now prove that the global modes of $f(x)$ are on a line through θ_1 and θ_2, and that they identify this hyperplane and therefore the Bayes border.

We will show that the set of modes on the line through θ_1 and θ_2 (which includes the global modes) has an average which is exactly the point $\frac{\theta_1+\theta_2}{2}$. Moreover, if we take only the global modes (assuming that there are more than one) their average is also this quantity because all the modes on the line appear in symmetric pairs. Hence the global modes of $f(x)$ identify a point (i.e., their average) and a line (going through them). The hyperplane which goes through this point and perpendicular to this line yields the Bayes decision border.

In the following, we ignore the normalizing constant and the *a priori* probabilities which are $\frac{1}{2}$. Translate the coordinate frame so that the origin is at the point θ_1. Then transform to a new primed-coordinate system, $x' = Qx$ s.t. the coordinates of θ_1' and θ_2' are on the x_1'-axis (the first point is the origin however we will refer to it by the name θ_1'). This is simply a rotation hence Q is unitary and the Jacobian equals 1 yielding

$$\begin{aligned} f(x') &= g(|Q^T x' - Q^T\theta_1'|^2) + g(|Q^T x' - Q^T\theta_2'|^2) \\ &= g(|x' - \theta_1'|^2) + g(|x' - \theta_2'|^2) \\ &= g((x_1' - \theta_{11}')^2 + x_2'^2 + \ldots x_N'^2) + g((x_1' - \theta_{21}')^2 + x_2'^2 + \ldots x_N'^2). \end{aligned}$$

Note that the global modes of $f(x)$ are on the x_1' axis since the x' which maximizes $f(x')$ has $x_2'^2 = \ldots = x_N'^2 = 0$ because these are all non negative quantities while $g(y)$

decreases as y increases.

We now show that the midpoint between the modes of $f(x')$ equals $\frac{\theta'_1+\theta'_2}{2}$. Since the global modes are on the x'_1-axis, and since at these points, all the partial derivatives are 0, in particular the partial w.r.t. x_1, then the solution set of

$$\frac{\partial}{\partial x'_1}\left(g((x'_1-\theta'_{11})^2)+g((x'_1-\theta'_{21})^2)\right)\equiv 0$$

must contain the first elements of the global mode vectors (which is the only nonzero elements, w.r.t. the primed frame). We get

$$\frac{g'((x'_1-\theta'_{11})^2)}{g'((x'_1-\theta'_{21})^2)}=\frac{\theta'_{21}-x'_1}{x'_1-\theta'_{11}}$$

where $g'(\cdot)$ denotes the derivative of $g(\cdot)$. For convenience, let $y\equiv x'_1-\frac{\theta'_{11}+\theta'_{21}}{2}$ and $a\equiv\frac{\theta'_{21}-\theta'_{11}}{2}$. The above equation becomes

$$\frac{g'((y+a)^2)}{g'((y-a)^2)}=\frac{a-y}{y+a}. \tag{5.17}$$

Now since g is decreasing then g' is negative hence the left side is positive which implies that the solution, y, to the equation satisfies $-a\leq y\leq a$. Clearly $y=0$ is a solution. Also, suppose y_0 is a solution, then it follows that $-y_0$ is also a solution since

$$\frac{g'((y+a)^2)}{g'((y-a)^2)}=\frac{a-y}{y+a}$$

implies that

$$\frac{g'((a-y_0)^2)}{g'((a+y_0)^2)}=\frac{a+y_0}{a-y_0}$$

and hence

$$\frac{g'((-y_0+a)^2)}{g'((-y_0-a)^2)}=\frac{a-(-y_0)}{-y_0+a}.$$

So the solutions that differ from 0 appear in symmetric pairs. Regardless of the number of solutions, clearly their average must be $y=0$. Moreover, considering only the global modes (which must be on this line), then their average is also at $y = 0$ again due to the symmetry. That is, given that z is a global mode then $-z$ is also a global mode since the function f achieves the same value at both of these points, i.e.,

$$\begin{aligned} f([z0\ldots 0]) &= g\left((z+a)^2\right) + g\left((z-a)^2\right) \\ &= g\left((-z+a)^2\right) + g\left((-z-a)^2\right) \\ &= f([-z0\ldots 0]). \end{aligned}$$

So taking the average of the global modes yields the point $x' = \frac{\theta'_{11}+\theta'_{21}}{2}$ which is precisely the point we needed to show.

As was shown above, the line through the modes is the line through θ_1 and θ_2. Hence given that there are at least one solution pair (i.e., at least two modes), we can identify this line and choose the hyperplane perpendicular to it that goes through the point which is the average of the modes; this yields the Bayes border.

Suppose there exist other two-pattern classes in the same family of problems with class conditional distributions say $\hat{g}(|x-\hat{\theta}_1|^2)$ and $\hat{g}(|x-\hat{\theta}_2|^2)$ with $\hat{g} \neq g$, and $\hat{\theta} \neq \theta$, $\hat{\theta} \neq [\theta_2, \theta_1]^T$ (the last condition ensuring we are not considering the simple permutation which trivially would yield the same decision regions) such that the mixture is the same, i.e.,

$$\hat{g}(|x-\hat{\theta}_1|^2) + \hat{g}(|x-\hat{\theta}_2|^2) = g(|x-\theta_1|^2) + g(|x-\hat{\theta}_2|^2) \equiv f.$$

Then, arguing as above, we get that the average of the global modes of f are located on the line through θ_1, θ_2 and their average is $\frac{\theta_1+\theta_2}{2}$ and similarly that the global modes are also on the line through $\hat{\theta}_1, \hat{\theta}_2$ and their average is $\frac{\hat{\theta}_1+\hat{\theta}_2}{2}$. Clearly, because the modes of f are fixed, the two averages must be the same point. Now, if we assume that f has two global modes, then the above two lines must coincide. So although the mixture f does not identify a unique class conditional pair it still identifies a unique border which is the Bayes border of both different problems, i.e.,

$$\begin{aligned}\{x : \hat{g}(|x-\hat{\theta}_1|^2) = \hat{g}(|x-\hat{\theta}_2|^2)\} &= \{x : |x-\hat{\theta}_1|^2 = |x-\hat{\theta}_2|^2\} \\ &= \{x : |x-\theta_1|^2 = |x-\theta_2|^2\} \\ &= \{x : g(|x-\theta_1|^2) = g(|x-\theta_2|^2)\}.\end{aligned}$$

Hence this algorithm will yield the Bayes border just from knowledge of the f although it is not possible to uniquely identify the class conditional densities from f. As seen above, the two global mode requirement is also necessary for determining a line to which the optimal hyperplane should be perpendicular. So in order for the algorithm to work properly the mixture needs to have at least two global modes.

We now show that there exist functions g which give rise to mixtures f that have at least two modes. We specify g' by construction, constraining the graph of $g'(y)$ to go through two points (y_i, A) and (y_j, B) satisfying

$$0 < y_i < y_j < a,\ \frac{a-y_i}{y_i+a} < A,\ \frac{a-y_j}{y_j+a} > B \tag{5.18}$$

where y_i, y_j, a, are any fixed scalars and $A, B > 0$.

This guarantees that the first point lies above the curve $\frac{a-y}{a+y}$ and the second point below it. Now we choose any continuous, negative function g' such that

$$\frac{g'((y_i+a)^2)}{g'((y_i-a)^2)} = A \text{ and } \frac{g'((y_j+a)^2)}{g'((y_j-a)^2)} = B \tag{5.19}$$

This guarantees that the curve of $\frac{g'((y+a)^2)}{g'((y-a)^2)}$ intersects the curve of $\frac{a-y}{y+a}$ at least at one point (y_k, C) where $y_i \le y_k \le y_j$, i.e., $y_k \ne 0$ is a solution of (5.17). By the above, it follows that $-y_k$ is also a solution. Let us just show that there exist functions g' that satisfies (5.19). Fix some values for y_i, y_j, a, A, B that satisfies (5.18). Arbitrarily pick negative values for $g'((y_i+a)^2)$ and $g'((y_j+a)^2)$ (negative since g is required to be decreasing). Then determine the necessary values of $g'((y_i-a)^2)$ and $g'((y_j-a)^2)$.

This completes the specification of g', i.e., specifying the value that g' takes at the points $y_i - a, y_i + a, y_j - a, y_j + a$. Clearly there are infinitely many such g' hence functions g that satisfy this, and they need not be indexed by a finite parameter vector. Similarly we can show that there are functions g with more than two modes.

So there are infinitely many functions, g, which are decreasing on $[0, \infty)$ and have continuous derivatives, g', such that the corresponding mixtures,

$$f_{g,\theta}(x) = \frac{1}{c^N}\left(\frac{1}{2}g(|x-\theta_1|^2) + \frac{1}{2}g(|x-\theta_2|^2)\right)$$

have at least two modes. Taking the hyperplane perpendicular to the line through these modes and which passes through the point of their average, yields the Bayes optimal decision border.

We now explain the conditions on f (on page 146) which result in the same sample complexity as for learning the Gaussian mixture of Section 5.4. The only terms depending on $f(x)$ (as opposed to the kernel function $K(x)$) which influence the sample complexity are the bound on $f(x)$ (used in (5.11)) and on its r^{th} derivative (5.8). We demand that the first $r-1$ derivatives be bounded by some finite constants uniformly over x. The bound need only be specified for the r^{th} derivative since by the orthogonality of $K(x)$ (see page 109), only the r^{th} term survives.

In the Gaussian case, $M_r = c_1(2\pi)^{-N/2}r^{\frac{r}{2}}e^r$ for some positive constant c_1 which results in the term $c_2 e^N r^{N/2}$ in (5.14). It follows that if g satisfies

$$\sup_x \left| g^{(r)}_{x_{i_1},x_{i_2},\ldots,x_{i_r}}(|x|^2)\right| \leq c_1 r^{\frac{r}{2}} c_2^r$$

where $i_1, i_2, \ldots, i_r \in \{1, 2, \ldots, N\}$, c_1, c_2 are positive constants, and the absolute values of the lower order partial derivatives of g is each bounded by some finite positive constant, then the sample complexity n is the same as in Theorem 5.1.

The other part where f has any bearing is in the bound of $\mathbf{E}K^2_{\sigma,x}$ in (5.11). We just use c^N instead of $(2\pi)^{N/2}$. This does not change the sample complexity order of growth since it only introduces a different constant in (5.15).

We now discuss algorithm K (stated on page 105) which is used to determine consistent estimates of the modes of $f \in \mathcal{F}$ and identify a decision rule that has a P_{error} close to P_{Bayes}.

After this subsection, there will be one more condition enforced on the class $\mathcal{P}$, more precisely, on its induced mixture class $\mathcal{F}$.

5.5.1 Discussion of Algorithm K

Above, we described the type of mixtures $f \in \mathcal{F}$ for which algorithm K can construct a decision rule. The intuition of the algorithm is based on the fact that for sufficiently small $\epsilon > 0$, knowing $f_n(x)$ allows to determine regions $A_{i,\epsilon}$ which contain the global modes η_i of f and the mode estimates $\hat{\eta}_i$. These regions are closely related to the global humps of the true mixture f (a hump is a region where one global mode is the only extrema). We now prove the consistency of this algorithm.

We assume that f is of the form stated in the preceding section. In general, although g is decreasing, f may have multiple global maxima and relative extrema, also on the x_1'-axis. Denote $M \equiv \sup_x f(x)$. We assume that f has at least two global modes η_1, η_2, such that $f(\eta_i) = M$. (We showed before that there exist g that have at least two modes; using the same argument we can show there exist g that have more than two modes.) Moreover let us assume that there are a finite number of such global modes, η_i, $1 \leq i \leq k$, at which $f(\eta_i) = M$. (This further restricts the type of g that may be used above but it is easy to show by the same construction that there exist such functions g.) Let

$$L \equiv \sup_{x \in D} f(x), \quad D = \{x : f'(x; u_{\eta_i,x}) = 0, x \neq \eta_j, f(\eta_i) = M, f(\eta_j) = M, 1 \leq i,j \leq k\}$$

where $f'(x; u_{\eta_i,x})$ is the directional derivative of f at x in the direction of the unit vector $u_{\eta_i,x}$ whose direction is the same as the ray starting at η_i going through x. We now show additional constraints on g s.t. $M - L > 0$ and s.t. f decreases monotonically along any ray, starting from a mode η_i, in some small ball around each mode. These will be crucial for our algorithm.

We show that there exists a $\delta > 0$ s.t. f is decreasing along any ray in the direction of $u_{\eta,x}$ where x is in the ball $B(\eta, \delta)$ around a mode η (we do not use subscripts for indexing one of k modes and instead reserve the subscripts to specify elements of a vector, except for θ_1 and θ_2). We will use the primed frame (page 147) which has the points corresponding to θ_1 and θ_2 on the x_1'-axis however here we drop the prime $'$ from the notations of all vectors. To prove this it suffices to show that $f'(x; u) < 0$ for any $x \in B(\eta, \delta)$ s.t. $x \neq \eta$ and where $u = \frac{x-\eta}{|x-\eta|}$. By definition,

$$\begin{aligned}
f'(x;u) &= \frac{\partial f(x)}{\partial x_1}\frac{x_1-\eta_1}{|x-\eta|} + \frac{\partial f(x)}{\partial x_2}\frac{x_2-\eta_2}{|x-\eta|} + \cdots + \frac{\partial f(x)}{\partial x_N}\frac{x_N-\eta_N}{|x-\eta|} \\
&= \left(g'(|x-\theta_1|^2)2(x_1-\theta_{11}) + g'(|x-\theta_2|^2)2(x_1-\theta_{21})\right)\frac{x_1-\eta_1}{|x-\eta|} \\
&+ \left(g'(|x-\theta_1|^2)2x_2 + g'(|x-\theta_2|^2)2x_2\right)\frac{x_2}{|x-\eta|} \\
&+ \cdots + \left(g'(|x-\theta_1|^2)2x_N + g'(|x-\theta_2|^2)2x_N\right)\frac{x_N}{|x-\eta|}
\end{aligned}$$

where we used the fact that θ_1, θ_2 and η are on the x_1-axis. We inquire whether the above is < 0. Denote by $a \equiv g'(|x-\theta_1|^2)$ and $b \equiv g'(|x-\theta_2|^2)$. Then it is the same as asking if

$$(a(x_1-\theta_{11}) + b(x_1-\theta_{21}))(x_1-\eta_1) + (a+b)(x_2^2 + \cdots + x_N^2) < 0.$$

But the second term is negative because $g' < 0$ as g is decreasing on $[0, \infty)$. So it suffices to check for the value of x_1 for which

$$(a(x_1-\theta_{11}) + b(x_1-\theta_{21}))(x_1-\eta_1) < 0. \qquad (5.20)$$

(We still have $|x - \eta| \leq \delta$, i.e., $|x_1 - \eta_1| \leq \delta$, as a constraint on x_1.) Without loss of generality take $\eta_1 < \eta_2$, so we have $\theta_{11} < \eta_1 < \theta_{21}$. There are two cases:

- $x_1 \leq \eta_1$, in which case we need $|b(x_1 - \theta_{21})| > |a(x_1 - \theta_{11}|$ for (5.20) to hold.
- $x_1 > \eta_1$, where we need $|a(x_1 - \theta_{21})| > |b(x_1 - \theta_{11}|$ for (5.20) to hold.

It suffices to consider the two following cases:

- $x_1 \leq \eta_1$, need $\frac{g'((x_1-\theta_{21})^2)}{g'((x_1-\theta_{11})^2)} > \frac{x_1-\theta_{11}}{\theta_{21}-x_1}$,
- $x_1 > \eta_1$, need $\frac{g'((x_1-\theta_{21})^2)}{g'((x_1-\theta_{11})^2)} < \frac{x_1-\theta_{11}}{\theta_{21}-x_1}$.

We know from previous analysis that the roots of $\frac{g'((x_1-\theta_{21})^2)}{g'((x_1-\theta_{11})^2)} = \frac{x_1-\theta_{11}}{\theta_{21}-x_1}$ are modes. But by definition, $\eta = [\eta_1 0 \cdots 0]$ is a mode. So using the same notation as in Section 5.5, and generalizing for all the k global modes, the above two requirements are satisfied if for $1 \leq i \leq k$, g is chosen s.t. for $\eta_i - \delta \leq y \leq \eta_i$ the function $\frac{g'((y+a)^2)}{g'((y-a)^2)}$ is above the graph of $\frac{a-y}{y+a}$ and for $\eta_i \leq y \leq \eta_i + \delta$ it is below this graph. We use the same construction of g as in (5.18) in order to satisfy these conditions and hence there are infinitely many such functions. The existence of $\delta > 0$ follows from having a finite number of modes. So we showed that there are infinitely many functions g s.t. f is decreasing along any ray $r_{\eta_i,x}$ for any $x \in B(\eta_i, \delta)$, $1 \leq i \leq k$, for some $\delta > 0$. We now use this to construct an algorithm for estimating the modes.

We estimate these modes by $\hat{\eta}_i$, $1 \leq i \leq k$ by using $f_n(x)$, the estimate of $f(x)$ where $\sup_x |f(x) - f_n(x)| \leq \epsilon$. (Note, we do not need $f_n(x)$ to be continuous in this algorithm; this is crucial since the kernel estimate gives a discontinuous $f_n(x)$ as the window functions, i.e., the polynomials, are truncated at ±1.) For the algorithm to work we need the error accuracy of the kernel estimate $\epsilon < \frac{M-L}{8}$.

First, find the $\text{argsup}_{x \in X} f_n(x) \equiv \hat{\eta}_1$ and suppose, w.l.o.g. $|\hat{\eta}_1 - \eta_1| < |\hat{\eta}_1 - \eta_i|$, $i \neq 1$. (If there is more than one such point, then choose any one.) Define H_i, to be the region

$H_i \equiv \{x : |x - \eta_i| < |x - \eta_j|, j \neq i, f$ decreases on line $l_{\eta_i,x}$ in direction of $x, x \in \mathbb{R}^N\} \cup \{\eta_i\}$

and let $H \equiv H_1 \cup H_2 \cdots \cup H_k$ where $H_i \cap H_j = \emptyset$.

We have

$$f(x) > L \Rightarrow x \in H.$$

To see this, suppose $x \notin H$. That means as we walk along a ray from some η_i towards x we encounter a point z at which $f'(z; u_{\eta_i,z}) = 0$. Moreover, there exist such a z satisfying $f(z) \geq f(x)$. Now $z \in D$ (where D is the region in the definition of L). Hence $f(z) \leq L$. And therefore $f(x) \leq L$ which proves it.

Now, we have, $\hat{\eta}_1 \in H_1$. This follows since

$$f_n(\hat{\eta}_1) \geq f_n(\eta_1) \geq f(\eta_1) - \epsilon = M - \epsilon$$

where the first inequality is because $\hat{\eta}_1$ is the argsup$f_n(\cdot)$ over $X' \ni \eta_1$. The second inequality follows from the fact that for any x, $|f_n(x) - f(x)| \leq \epsilon$. So,

$$f(\hat{\eta}_1) \geq f_n(\hat{\eta}_1) - \epsilon \geq M - 2\epsilon > L$$

where the last inequality follows from the restriction on ϵ, i.e., $\epsilon < \frac{M-L}{8}$.

Now define

$$B_\epsilon = \{x : f_n(x) > f_n(\hat{\eta}_1) - 4\epsilon\}.$$

(We will not carry the subscript ϵ for brevity.) Clearly $x \in B \Rightarrow x \in H$ i.e.,

$$B \subset H$$

since

$$x \in B \Rightarrow f_n(x) > f_n(\hat{\eta}_1) - 4\epsilon \geq M - \epsilon - 4\epsilon = M - 5\epsilon,$$

$$f(x) \geq f_n(x) - \epsilon \geq M - 6\epsilon > L,$$

and from above, $f(x) > L \Rightarrow x \in H$. We will concentrate only on subsets of B hence throught the discussion below, we have f with the ray-decreasing property of H.

We next define a region A_i which depends on the estimate $\hat{\eta}_i$. Up to now we only defined $\hat{\eta}_1$, so at this stage only A_1 can be defined. However, later when we define the rest of the estimates $\hat{\eta}_i$, $i = 2, \dots, k$, we will use the following definition for A_i:

$$A_{i,\epsilon} = \{y : |y - \hat{\eta}_i| \leq \inf_{x} \inf_{z \in r_{\hat{\eta}_i, x}} |z - \hat{\eta}_i|, f_n(z) < f_n(\hat{\eta}_1) - 6\epsilon\} \cup \{\hat{\eta}_i\} \text{ for } 1 \leq i \leq k,$$

where $r_{\hat{\eta}_i, x}$ is a ray from $\hat{\eta}_i$ going through x. (We omit the ϵ subscript for brevity.) So A_i is simply a ball around $\hat{\eta}_i$ with the above-specified radius.

We claim that the region

$$(B - A_1) \cap H_1 = \emptyset.$$

We proceed by showing that all points x in H_1 that have $f_n(x) > f_n(\hat{\eta}_1) - 4\epsilon$ must be also in A_1. First, we prove that

$$\eta_1 \in A_1.$$

Suppose the contrary. Then there exists a $z \in \partial A_1$ on the ray $r_{\eta_1, \hat{\eta}_1}$ and by definition of A_1, $f_n(z) \leq f_n(\hat{\eta}_1) - 6\epsilon$. Also, f is decreasing between η_1 and $\hat{\eta}_1$, (since we showed that $\hat{\eta}_1 \in H_1$, and by definition $\eta_1 \in H_1$) hence

$$f(\eta_1) \geq f(z) \geq f(\hat{\eta}_1) \Rightarrow f_n(z) \geq f(z) - \epsilon \geq f(\hat{\eta}_1) - \epsilon \geq f_n(\hat{\eta}_1) - 2\epsilon.$$

This is a contradiction. Now we prove the claim making use of the fact that $\eta_1 \in A_1$. Suppose the contrary. Then there exists an x satisfying $x \in B$, $x \notin A_1$, and $x \in H_1$. This implies there exist some $z \in \partial A_1$, i.e., on the border of A_1 as in the definition of A_1, such that z lies in between η_1 and x, i.e. on the ray $r_{\eta_1, x}$. Now, $|z - \eta_1| \leq |\eta_1 - x|$. Since $x, \eta_1 \in H_1$, then $f(z) \geq f(x)$. Also, since $z \in \partial A_1$ then $f_n(z) < f_n(\hat{\eta}_1) - 6\epsilon$.

And $f_n(x) \leq f_n(z) + 2\epsilon$ because at any point, f_n can jump by at most 2ϵ from its current value (due to $|f_n(\cdot) - f(\cdot)| \leq \epsilon$). Hence $f_n(x) \leq f_n(\hat{\eta}_1) - 6\epsilon + 2\epsilon = f_n(\hat{\eta}_1) - 4\epsilon$. So x does not satisfy $f_n(x) > f_n(\hat{\eta}_1) - 4\epsilon$. So $x \notin B$. This is a contradiction. Hence after removing A_1 from B, we are left with no points from H_1, i.e. $(B - A_1) \cap H_1 = \emptyset$.

Now we show that H_i, $i \neq 1$, has points that are in B, i.e. after the removal of A_1, we still have

$$(B - A_1) \cap H_i \neq \emptyset.$$

First we show that

$$\eta_i \notin A_1, \text{ for } 2 \leq i \leq k.$$

We have three key points, $\hat{\eta}_1$, η_1 and η_i. Consider the line l_{η_1,η_i}. Clearly there exist a $y \in l_{\eta_1,\eta_i}$ s.t. $f(y) \leq L$. That is because, in general, on a ray through any two modes, η_i, η_j of f, there must be a point y at which the directional derivative $f'(y; u_{\eta_i,\eta_j}) = 0$, and also recall the definition of L. (Note that $y \neq \eta_i$ since the two modes differ, i.e., $\eta_1 \neq \eta_i$.) So therefore

$$f_n(y) \leq L + \epsilon < M - 7\epsilon = f(\eta_1) - 7\epsilon \leq f_n(\eta_1) - 6\epsilon \leq f_n(\hat{\eta}_1) - 6\epsilon \Rightarrow f_n(y) < f_n(\hat{\eta}_1) - 6\epsilon$$

where the first inequality from the left follows from the condition on ϵ. Hence either $y \notin A_1$ or $y \in \partial A_1$ Therefore the radius of A_1 is $\leq |\hat{\eta}_1 - y|$ by definition of A_1. Also $|\hat{\eta}_1 - \eta_1| < |\hat{\eta}_1 - \eta_i|$ by definition of $\hat{\eta}_1$. So we have a circle centered at $\hat{\eta}_1$ with η_i on the circle and η_1, y, both inside the circle, both lying on a line through η_i. It is now simple to see that $|\hat{\eta}_1 - y| < |\hat{\eta}_1 - \eta_i|$ by taking a radius of size $|\hat{\eta}_1 - \eta_i|$ and rotating it until it goes through the point y. So the radius of A_1 is $< |\hat{\eta}_1 - \eta_i|$. Hence $\eta_i \notin A_1$.

Now we prove that there exist at least one point in $(B - A_1) \cap H_i$, namely η_i itself, for $2 \leq i \leq k$. It suffices to check if $f_n(\eta_i) > f_n(\hat{\eta}_1) - 4\epsilon$, i.e. the definition of B, since we've already proved that $\eta_i \notin A_1$ and by definition of H_i, $\eta_i \in H_i$. But this follows trivially as

$$f_n(\eta_i) \geq f(\eta_i) - \epsilon = f(\eta_1) - \epsilon \geq f(\hat{\eta}_1) - \epsilon \geq f_n(\hat{\eta}_1) - 2\epsilon > f_n(\hat{\eta}_1) - 4\epsilon.$$

We now define the rest of the estimates, namely $\hat{\eta}_2, \ldots, \hat{\eta}_k$. Find the point that equals $\text{argsup}_{x \in B - A_1} f_n(x)$. This must yield a point which is not in H_1 since we've shown above that $(B - A_1) \cap H_1 = \emptyset$. This point must be in $B \cap H_i$ for some $2 \leq i \leq k$ since we proved that there exists at least one point, namely η_i, in $B \cap H_i$ for $2 \leq i \leq k$, and since $B \subset H$. W.l.o.g. suppose this point falls closer to η_2 than to any other η_i, $3 \leq i \leq k$. We define this point as $\hat{\eta}_2$. From these last statements and from the definition of H_2, we have $\hat{\eta}_2 \in H_2$. We can then define A_2 as was done above for the general A_i. There is a slight asymmetry in the way we defined the estimates since we used $\hat{\eta}_1$ as the pilot, in the definition of all A_i, $1 \leq i \leq k$ and for the definition of the region B. Hence we will go through the proofs once more, to show that they still work when for getting $\hat{\eta}_i$, $2 \leq i \leq k$.

As was the case for A_1, here too we claim

$$((B - A_1) - A_2) \cap H_2 = \emptyset$$

as we now show. First we prove that

$$\eta_2 \in A_2.$$

Suppose the contrary. Then there exists a $z \in \partial A_2$ on the ray $r_{\eta_2, \hat{\eta}_2}$ and by definition of A_2, $f_n(z) \leq f_n(\hat{\eta}_1) - 6\epsilon$. Also, f is decreasing between η_2 and $\hat{\eta}_2$, (since we showed that $\hat{\eta}_2 \in H_2$, and by definition $\eta_2 \in H_2$) hence

$$f(\eta_2) \geq f(z) \geq f(\hat{\eta}_2) \Rightarrow f_n(z) \geq f(z) - \epsilon \geq f(\hat{\eta}_2) - \epsilon \geq f_n(\hat{\eta}_2) - 2\epsilon \geq f_n(\hat{\eta}_1) - 4\epsilon > f_n(\hat{\eta}_1) - 6\epsilon$$

This is a contradiction. (Note, the inequality before last follows since

$$|f_n(\hat{\eta}_1 - f_n(\hat{\eta}_2))| \leq 2\epsilon.$$

In fact, for any $1 \leq i \neq j \leq k$, $|f_n(\hat{\eta}_i) - f_n(\hat{\eta}_j)| \leq 2\epsilon$ because for any $1 \leq i \leq k$,

$$f_n(\hat{\eta}_i) \geq f_n(\eta_i) \geq f(\eta_i) - \epsilon = M - \epsilon$$

and $f_n(\hat{\eta}_i) \leq M + \epsilon$.) Now we prove the claim making use of the fact that $\eta_2 \in A_2$. Suppose the contrary, i.e., that $x \in B$, and $x \in H_2$ but $x \notin A_2$. This implies there exist some $z \in \partial A_2$, i.e., on the border of A_2 as in the definition of A_2, such that z lies in between η_2 and x, i.e., on the ray $r_{\eta_2,x}$. Now, $|z - \eta_2| \leq |\eta_2 - x|$. Also, since $z \in \partial A_2$ then $f_n(z) < f_n(\hat{\eta}_1) - 6\epsilon$. Since $x, \eta_2 \in H_2$, then $f(z) \geq f(x)$. And $f_n(x) \leq f_n(z) + 2\epsilon$ because at any point, f_n can jump by at most 2ϵ from its current value (due to $|f_n(\cdot) - f(\cdot)| \leq \epsilon$). Hence $f_n(x) \leq f_n(\hat{\eta}_1) - 6\epsilon + 2\epsilon = f_n(\hat{\eta}_1) - 4\epsilon$. So x does not satisfy $f_n(x) > f_n(\hat{\eta}_1) - 4\epsilon$. So $x \notin B$. This is a contradiction. Hence after removing A_2 from B, we are left with no points from H_2, i.e. $((B - A_1) - A_2) \cap H_2 = \emptyset$.

As before, after removal of A_2 from $B - A_1$ we still have points in H_i, $3 \leq i \leq k$, which are in B, i.e.

$$((B - A_1) - A_2) \cap H_i \neq \emptyset \qquad \text{for } 3 \leq i \leq k.$$

First we show that $\eta_i \notin A_2$, for $3 \leq i \leq k$. We have three key points, $\hat{\eta}_2$, η_2 and η_i. Consider the line l_{η_2,η_i}. Clearly there exist a $y \in l_{\eta_2,\eta_i}$ s.t. $f(y) \leq L$. That is because, in general, on a ray through any two modes, η_i, η_j of f, there must be a point y at which the directional derivative $f'(y; u_{\eta_i,\eta_j}) = 0$, and also recall the definition of L. (Note that $y \neq \eta_i$ since the two modes differ, i.e., $\eta_2 \neq \eta_i$.) So therefore

$$f_n(y) \leq L + \epsilon < M - 7\epsilon = f(\eta_1) - 7\epsilon \leq f_n(\eta_1) - 6\epsilon \leq f_n(\hat{\eta}_1) - 6\epsilon \Rightarrow f_n(y) < f_n(\hat{\eta}_1) - 6\epsilon$$

where the first inequality from the left follows from the condition on ϵ. Hence either $y \notin A_2$ or $y \in \partial A_2$. But therefore the radius of A_2 is $\leq |\hat{\eta}_2 - y|$ by definition of A_2. Also $|\hat{\eta}_2 - \eta_2| < |\hat{\eta}_2 - \eta_i|$ by definition of $\hat{\eta}_2$. So we have a circle centered at $\hat{\eta}_2$ with

η_i on the circle and η_2, y inside the circle, both lying on a line through η_i. It is now simple to see that $|\hat{\eta}_2 - y| < |\hat{\eta}_2 - \eta_i|$ by taking a radius, of size $|\hat{\eta}_2 - \eta_i|$, and rotating it until it goes thorough the point y. So the radius of A_2 is $< |\hat{\eta}_2 - \eta_i|$. Hence $\eta_i \notin A_2$.

Now we prove that there exist at least one point in $((B - A_1) - A_2) \cap H_i$, namely η_i itself, for $3 \leq i \leq k$. It suffices to check if $f_n(\eta_i) > f_n(\hat{\eta}_1) - 4\epsilon$, i.e. the definition of being in B, since we already proved that $\eta_i \notin A_2$ and by definition of H_i, $\eta_i \in H_i$. We have

$$f_n(\eta_i) \geq f(\eta_i) - \epsilon = f(\eta_1) - \epsilon \geq f(\hat{\eta}_1) - \epsilon \geq f_n(\hat{\eta}_1) - 2\epsilon > f_n(\hat{\eta}_1) - 4\epsilon.$$

Using the above if we take $\text{argsup}_{x \in B - A_1 - A_2} f_n(x)$, this must yield a point which is not in H_1, nor in H_2. This point must lie in $B \cap H_i$ for some $3 \leq i \leq k$. W.l.o.g. suppose the point falls closer to η_3 than to any other η_i, $4 \leq i \leq k$. We define this point to be $\hat{\eta}_3$. From these last statements and from the definition of H_3, we have $\hat{\eta}_3 \in H_3$.

So it is clear that our procedure for finding $\hat{\eta}_i$, $1 \leq i \leq k$ continues as above until all $\hat{\eta}_i$ have been found, the last one being $\hat{\eta}_k = \text{argsup}_{x \in B - A_1 - \cdots - A_{k-1}} f_n(x)$. After that stage, we have removed A_i, $1 \leq i \leq k$ from B, i.e., we are left with the region

$$B \cap A_1^c \cap \cdots \cap A_k^c$$

(where intersection by the complement is the same as subtracting a region) whose intersection with any of H_i, $1 \leq i \leq k$, is empty. But recall that $B \subset H_1 \cup H_2 \ldots \cup H_k$. This means the region that the learner is left with does not have any point x s.t. $f_n(x) > f_n(\hat{\eta}_1) - 4\epsilon$ and at that stage he stops the algorithm since no point is returned for the argsup of f_n, i.e.,

$$\text{argsup}_{B \cap A_1^c \cap \cdots \cap A_k^c} f_n(x) = \text{argsup}_{\emptyset} f_n(x) = \emptyset.$$

Note that the learner does not need to know the number of modes, k, of f since all he needs to do is keep finding the argsup of f_n over a region which is totally defined by the first estimate, $\hat{\eta}_1$, and which becomes smaller and smaller until it becomes empty exactly when there is no need to estimate anymore modes.

Finally we show that the above estimates are consistent as $\epsilon \rightarrow 0$. From above, both $\hat{\eta}_i, \eta_i \in A_{i,\epsilon}$. Considering the terms inside the definition of $A_{i,\epsilon}$ we have

$$\hat{\eta}_1 \rightarrow \eta_1 \text{ as } \epsilon \rightarrow 0$$

since

$$\hat{\eta}_1 = \text{argsup}_{x \in X} f_n(x) \rightarrow \text{argsup}_{x \in X} f(x) = \eta_1$$

since $|f_n(x) - f(x)| \rightarrow 0$ for any $x \in X$. Also,

$$f_n(\hat{\eta}_1) \rightarrow f(\hat{\eta}_1) = f(\text{argsup}_{x \in X} f_n(x)) \rightarrow f(\text{argsup}_{x \in X} f(x)) = f(\eta_1)$$

and $f_n(z) \rightarrow f(z)$. Hence

$$\begin{aligned} A_{i,\epsilon} &= \{y : |y - \hat{\eta}_i| \leq \inf_x \inf_{z \in r_{\hat{\eta}_i,x}} |z - \hat{\eta}_i|, f_n(z) < f_n(\hat{\eta}_1) - 6\epsilon\} \cup \{\hat{\eta}_i\} \\ &\rightarrow \{y : |y - \eta_i| \leq \inf_x \inf_{z \in r_{\eta_i,x}} |z - \eta_i|, f(z) < f(\eta_1)\} \cup \{\eta_i\} \\ &= \{y : |y - \eta_i| \leq 0\} \cup \{\eta_i\} = \eta_i \end{aligned}$$

As both η_i and $\hat{\eta}_i$ are in $A_{\epsilon,i}$, the above implies that $\hat{\eta}_i \rightarrow \eta_i$ as $\epsilon \rightarrow 0$, for all $1 \leq i \leq k$.

5.5.2 The resulting P_{error}

In the previous subsection, it was only necessary to know $M - L$ in order to specify the allowed range for the accuracy parameter ϵ, without needing to know the number of modes of f. Thus far the algorithm yields consistent estimates for the modes of f, where $f \in \mathcal{F}$.

Using the modes estimates we can form a hyperplane estimate of the optimal Bayes hyperplane as follows: Given the k mode estimates $\hat{\eta}_i$, $1 \leq i \leq k$, we minimize the function

$$e = \sum_{i=1}^{k} \sum_{j=1}^{N-1} \left(\frac{([\hat{\eta}_i, 1], w_j)}{|w_j|} \right)^2$$

w.r.t. the $(N-1)$ unknown $(N+1)$-dimensional unit vectors $w_1, \ldots, w_{N-1}$, under the constraints

$$(w_j, w_k) = 0, \qquad \text{for } 1 \leq k \leq N-1 \text{ and } k \neq j.$$

(Nonlinear programming is one approach to solve this.) This will find a line l in $\mathbb{R}^N$ which is least-square-close to the k mode estimates. We used here the fact that a line in $\mathbb{R}^N$ can be represented as intersection of affined hyperplanes in $\mathbb{R}^N$, i.e., as the set of all points that are orthogonal to a specific set of vectors w_j, $1 \leq j \leq N-1$. The term inside the double summation represents the distance of the i^{th} point to the j^{th} hyperplane. Thus e represents the total distance squared of the k points from the line in $\mathbb{R}^N$, and minimizing e obtains the least square line l.

We then form the average

$$\overline{\eta} \equiv \frac{1}{k} \sum_{i=1}^{k} \hat{\eta}_i$$

and define the hyperplane estimator to be the unique hyperplane which is orthogonal to the line l and which goes through the point $\overline{\eta}$ (which is not necessarily on the line l).

As $n \to \infty$, the mode estimates converge to the true modes and the hyperplane estimate converges to the optimal Bayes hyperplane. So there exists some function $h(\epsilon)$ such that the classification error of the decision rule based on this hyperplane is

$$P_{error} = P_{Bayes}(1 + h(\epsilon))$$

where $h(\epsilon) \to 0$ as $\epsilon \to 0$ when the regions are labeled optimally. As in the Gaussian case, we can use the majority rule with $c_1 \log \frac{1}{\delta}$ labeled examples, where $c_1 > 0$ is some constant, to guarantee with confidence $> 1 - \delta$ that this is true.

Now, the function $h(\epsilon)$ is the accuracy parameter of the probability of error of the classifier. The function h depends on the type of g's that are permitted in the definition of the problem family $\mathcal{P}$ since it is directly related to the amount of deviation possible by the mode estimates $\hat{\eta}_i$ when the kernel estimate $f_n(x)$ deviates by no more than ϵ from $f(x)$ uniformly over $x \in \mathbb{R}^N$. The flatter the main humps of f, the more such deviation is possible and the more P_{error} can deviate from P_{Bayes}.

Therefore in order to be able to claim an accuracy $h(\epsilon)$ uniformly for all problems in $\mathcal{P}$, given the sample complexities of Theorem 5.1 we need to ensure that we define $\mathcal{P}$ with dependence on h. One way to create a $\mathcal{P}$, is to consider a union of families of classification problems, the i^{th} family $\mathcal{P}_i$ being composed of density functions

$$f_{g_i,\theta_1}(x) = g_i(|x-\theta_1|^2), \qquad f_{g_i,\theta_2}(x) = g_i(|x-\theta_2|^2), \qquad \theta_1, \theta_2 \in \mathbb{R}^N$$

and such that for the class $\mathcal{P}_i$, the misclassification error accuracy is $h_i(\epsilon)$. (It is not difficult to approximate $h_i(\epsilon)$ since it suffices to consider one type of function g_i.) Then define

$$\mathcal{P}_h = \cup_{i=1}^{l} \mathcal{P}_i$$

where $l < \infty$, and h is an envelope function for all the h_i, $i = 1, \ldots, l$, i.e.

$$h(x) \equiv \sup_x h_i(x).$$

Then any classification problem in the family $\mathcal{P}_h$ can be learned to an accuracy $h(\epsilon)$ using algorithm K with the sample complexities of Theorem 5.1.

5.6 Neural Network Clustering

Here we describe simulation results of a neural network (Kohonen [24]) based on the Kohonen self-organizing maps, which can learn using unlabeled examples. Our results are qualitative, giving the intuition for comparing two extreme scenarios: an only-labeled sample with side information about the class densities, versus a mixed sample (few labeled examples) without side information; the latter is implemented via a neural network.

The Kohonen neural network is a popular algorithm that has found numerous applications among a wide range of fields, e.g., statistical pattern recognition, robot control, adaptive communication schemes, and speech recognition. It is biologically inspired by the cortical maps in the brain that are topologically ordered and organized with high dependency on their input features. Its algorithm is very similar to the k-means algorithm which is an *ad hoc* procedure used to partition multivariate data into cells that resemble the clustering of the underlying distribution. We first describe the algorithm and then show how it can be used for learning classification.

The neural network consists of k neurons with real weight vectors $w_i \in X$, $i = 1, \ldots, k$ where X is the space over which examples x are drawn according to some distribution. The neurons are arranged in a two-dimensional array which defines their spatial neighboring. It is then possible to define the influence of a neuron on the adaptation of other neurons in its vicinity. The notion of vicinity is *not* in X space but is measured by the array-index according to which the neurons are ordered. The weight vectors are adapted by the following iterative procedure

$$w_i(t+1) = \begin{cases} w_i(t) + \alpha(t)\,(x(t) - w_i(t)) & \text{if } i \in N_c(t), \\ w_i(t) & \text{if } i \notin N_c(t). \end{cases} \tag{5.21}$$

where t represents discrete time, $\alpha(t)$ is an adaptation-gain, and $N_c(t)$ is an index set of the neurons around the winner neuron whose index is c. We define the winner as the one neuron whose weight vector w_c is the closest to x w.r.t. the Euclidean norm, i.e., $|x - w_c| = \min_{1 \leq i \leq k} |x - w_i|$. One can view the quantity $|x - w_i|$ as a real-valued output of the i^{th} neuron. Hence in effect this algorithm is a model of a collection of neurons all seeing the same input vector x and adapting their sensitivity to x (i.e., the weight vectors) according to both the input x and the activity of other neurons. As time evolves, the activity of only the nearer neighbors influence the adaptation of a neuron's weight vector. The parameters $\alpha(t)$ and $N_c(t)$ start at some initial value and decrease at the rate of $O(1/t)$. This choice is *ad hoc*, however with it, the vectors w get ordered in a way which resembles the natural clustering of the examples which are drawn according to the unknown underlying distribution. This is the fame of the Kohonen self-organization phenomenon; it is based on the intuition that the density of weight vectors w_i in X space tends to imitate the probability density of the examples x. In this regard it is similar to some non-parametric density estimation techniques. It is possible to use this neural network for learning a two-class classification problem as we shall see below.

We define a nearest neighbor partition of the weight vectors w with the i^{th} vector w_i corresponding uniquely to a voronoi cell

$$v_i = \{x : |x - w_i| \leq |x - w_j|\} \tag{5.22}$$

for $j \neq i$. Clearly, if this partition is labeled, i.e., each cell gets a label 1 or 2, then we have a decision rule: given an x, classify it by the label of the cell in which it falls. Therefore the following learning procedure emerges: pick randomly k weight vectors then show n unlabeled examples x, while adapting the w vectors according to the Kohonen rule. Define a nearest neighbor partition using the w vectors, then show m labeled examples and use the majority rule per cell, to label each cell, and the resulting labeled partition is the classification decision rule. This is the basis for our neural networks learning classification experiments. We now describe our results.

A 4 Neuron Net trained on a 2 dimensional mixture of two gaussians with different variances. After training only on unlabeled example, the neuron weight vectors define a voronoi partition that is then labeled using labeled examples.

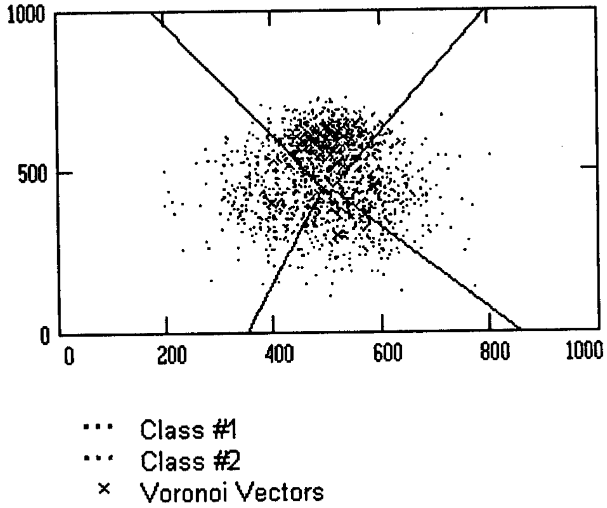


Figure 5.3:

5.6.1 The value of a labeled example versus P_{error}

We investigated the mixed sample sizes, n and m with dimensionality $N = 2$, for achieving a specified error; as in previous sections, the labeled examples are used only for labeling the decision regions. We simulated a 4-neuron network with weight vectors in $\mathbb{R}^2$. Unlabeled examples $x \in \mathbb{R}^2$ were drawn according to a mixture of two Gaussians each with a different covariance matrix (both were diagonal matrices). Figure 5.3 shows the actual data drawn from this mixture; the lines represent the voronoi cell borders of the partition. We ran six experiments, each differing in the number of randomly drawn unlabeled examples ranging from $n = 20$ up to $n = 10,000$. In each experiment we then showed m randomly drawn labeled examples, where m ranged from the minimum number necessary in order not to leave out any cell unlabeled, up to $m = 100$. We measured the classification error as a function of

n and m; the learning curves are shown below in Figure 5.4. The curves of the neural network corresponding to the lower unlabeled sample sizes do not reach the low error rate as m increases because we did not utilize labeled examples to further adjust the decision border as in the different variants of the LVQ algorithm of Kohonen [24]. For each experiment we averaged the learning curves of 50 different neural networks. As a reference we then conducted an experiment, using the knowledge of the parametric form of the class densities to estimate the sufficient statistics, i.e., the means and covariances for each of the Gaussians, by using only labeled examples with which the Bayes optimal decision border was estimated. With respect to the neural net, this experiment is different since significantly more information (parametric/identifiability knowledge of the class conditional distributions) is provided to the algorithm.

Examining the intersection of the (dotted) curve of the purely labeled experiment with the (solid) curves of the neural network, gives approximately the number of unlabeled examples necessary for the labeled sample size of the neural network to differ by one example from the labeled sample size of the parametric algorithm. This intuitively represents an upper bound on the value of one labeled example in terms of unlabeled examples because it says that for a fixed error, with no side information and with minimal usage of labeled examples we need this many unlabeled examples and one fewer labeled examples than the case which has maximum side information and uses labeled examples efficiently. The points where the parametric algorithm curve intersects the neural net curves are plotted in Figure 5.5. There we see that the value of a labeled example increases sharply as the objective P_{error} is reduced.

5.6.2 m versus the dimensionality N

In this section we describe the effect on the labeled sample size m when increasing the dimensionality N. A partition can be labeled by any one of the 2^k labelings, where k

The set of solid curves corresponds to learning with both unlabeled and labeled examples. n = # unlabeled examples.
The dotted curve represents learning with only labeled examples.

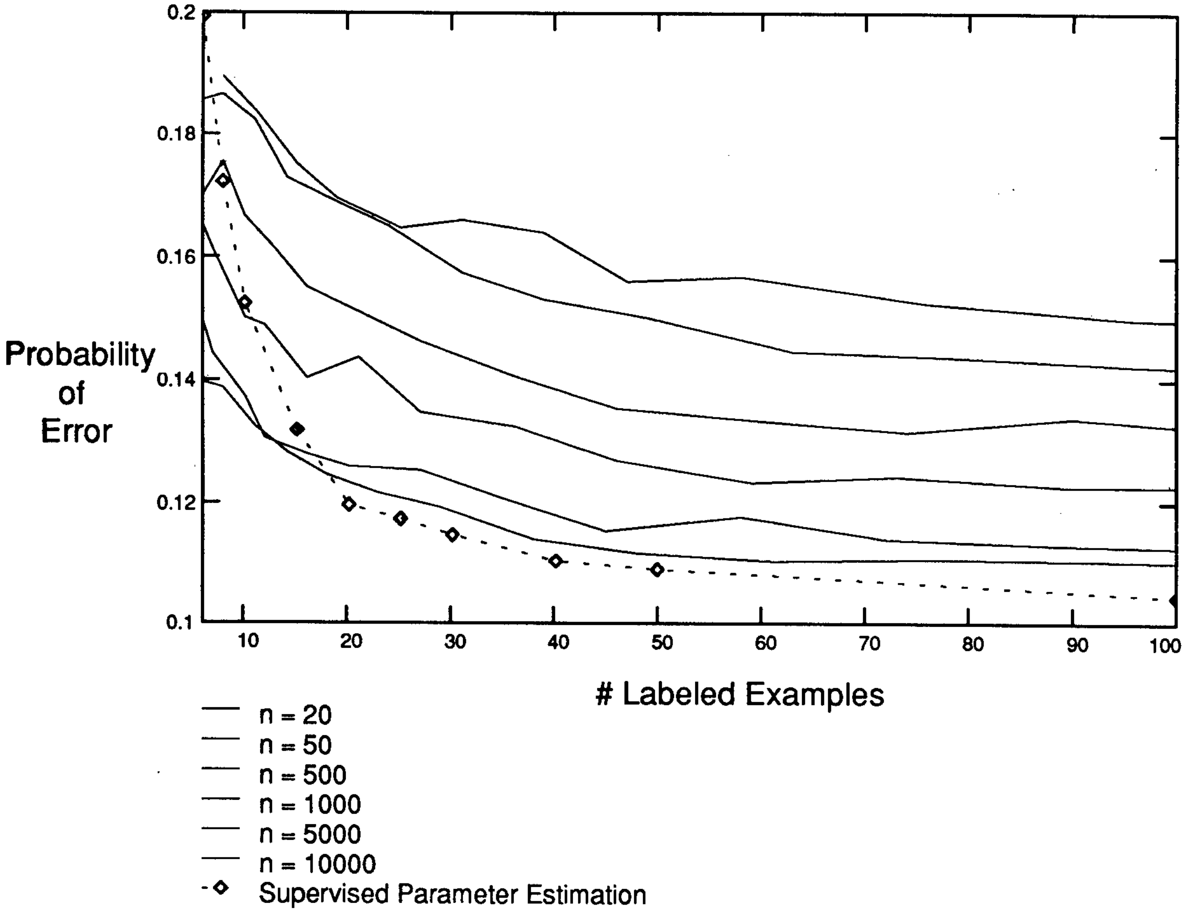


Figure 5.4:

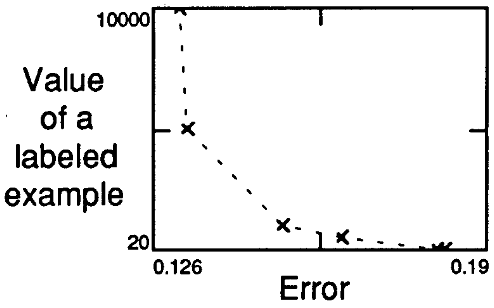


Figure 5.5:

is the number of cells. Theoretically, we expect that as the dimensionality increases, k needs to increase in order to construct a partition which achieves a given fixed error criteria (under the optimal labeling). This in turn requires more labeled examples in order to pick the optimal labeling with a fixed confidence. This effect of N on m is clearly undesirable and since many realistic problems have high dimensionality we sought an approach that can reduce this effect.

As unlabeled examples are taken to be abundant, we did not attempt to limit their supply when choosing the algorithm. Our main focus was to limit the labeled sample size. The Kohonen neural network, by principle, fits this criteria as it can utilize primarily unlabeled examples for learning the decision regions. We therefore considered a variant of its architecture.

The limitation of the voronoi partition, produced by the Kohonen network, arises from the piecewise-linearity of the cells (from (5.22) a voronoi cell has hyperplane borders with its surrounding cells). When the classification problems consist of pattern classes that are not linearly separable, it takes many voronoi cells to establish a reasonable decision border. This raises the labeled sample size required to optimally label the partition. However, if the cells have nonlinear borders then in many problem situations one can do with fewer cells and hence considerably reduce the labeled sample size.

The question therefore was whether a neural network utilizing primarily unlabeled examples in a self organizing unsupervised learning can produce non linear cells and adapt them to achieve a good partition, in particular, a partition that does not have too many cells and which needs fewer labeled examples to be optimally labeled.

Our simulation results (described below) show that a two-layer Kohonen network, featuring self organization in both layers, performed with the above desired characteristics for a variety of problems. On some problems this network had a labeled sample complexity superior (w.r.t N) to the voronoi-partition classifier (based on the regular single layer Kohonen net), for instance, there were problems in which the labeled sample size was a constant w.r.t. N. We now describe the simulations.

The architecture that we considered has two self organizing layers. The first layer is a Kohonen network, i.e., a collection of neurons each of whose inputs is a vector x which represents the pattern-class feature vector. The i^{th} neuron is associated with a weight vector w_i. The output of the i^{th} neuron is a real scalar g_i which measures the Euclidean distance from x, i.e. $g_i = |x - m_i|$. These neurons adapt their weight vector according to the Kohonen adaptation rule of (5.21).

The second layer consists of neurons each having a weight vector y_i. Their input is the vector g of outputs of the first layer neurons. The neurons of this layer also adapt their y_i according to the Kohonen rule.

Using unlabeled examples, we first train the first layer producing the adapted w_i vectors. Then using the same examples we train the second layer neurons. This results in a partition of the transformed feature space G i.e. X is transformed to G by the mapping $g = [|x - m_1|, |x - m_2|, \ldots, |x - m_k|]$. Each of the second layer neurons is associated with a voronoi cell, i.e.,

$$v_i = \{g : |g - y_i| \leq |g - y_j|\}.$$

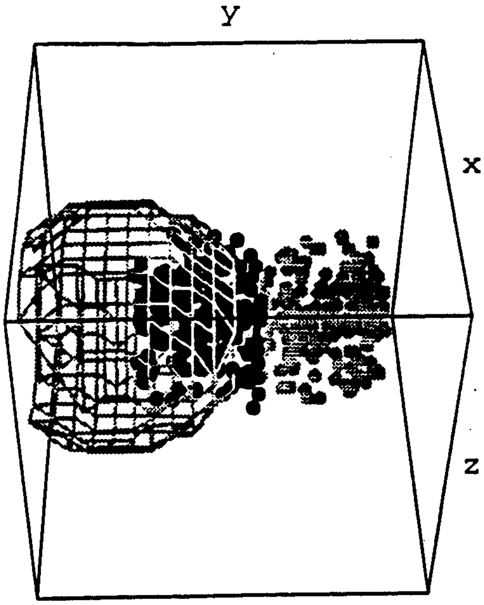
Using labeled examples, we label the partition, assigning each cell with the label of the majority of the examples that fell in it (or drawing with probability $\frac{1}{2}$ its label if none fell).

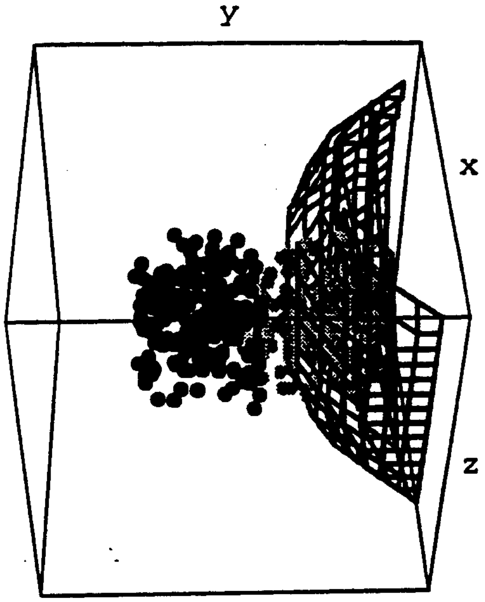
Given a test vector x, the network classifies it with the label of the cell in which it falls. This establishes the classifier which we denote as the 2-layer network.

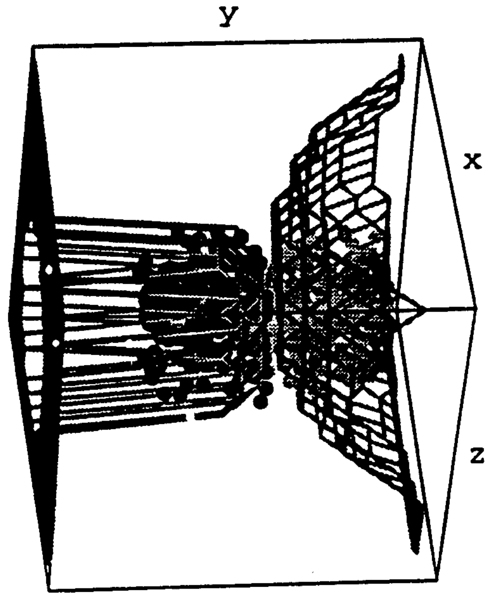
We simulated this 2-layer network and compared its m v.s. N performance with that of the 1-layer network (i.e., the usual Kohonen network). Some examples of the types of activation regions, i.e., cells, that the 2-layer network exhibited are displayed in Figures 5.6 and 5.7. In these the input space X dimensionality is $N = 3$. The region between the two mesh surfaces is a cell corresponding to one of the second layer neurons. The first pattern class is represented by the black dots and the second class by the gray dots. The nonlinearity of the surfaces is apparent.

We trained both the 1-layer and the 2-layer networks on a problem consisting of four N-dimensional cubes, mutually contained as $Cube_1 \subset Cube_2 \subset Cube_3 \subset Cube_4$ and where the first class is defined as $Cube_1 \bigcup Cube_3$ and the second class as $Cube_2 \bigcup Cube_4$. We first drew unlabeled examples distributed uniformly and then labeled examples distributed uniformly over each class. The case of $N = 2$ is displayed in Figure 5.8. We measured the sample complexity m w.r.t. N for both networks, which is needed to achieve a constant error rate across the range of N. Figure 5.9 shows the labeled sample complexity versus dimension N which, for the 1-layer-network, increases with increasing N. The 2-layer network needed only a constant number of labeled examples.

We then considered random classification problems, picking 30 clusters per class randomly positioned over a two ring region. Figure 5.10 shows an instance of the problem with $N = 2$. The vertices of the mesh indicate the position of the first layer neurons and the black and gray clusters are class 1 and 2 respectively.


y
x
z


y
x
z


y
x
z

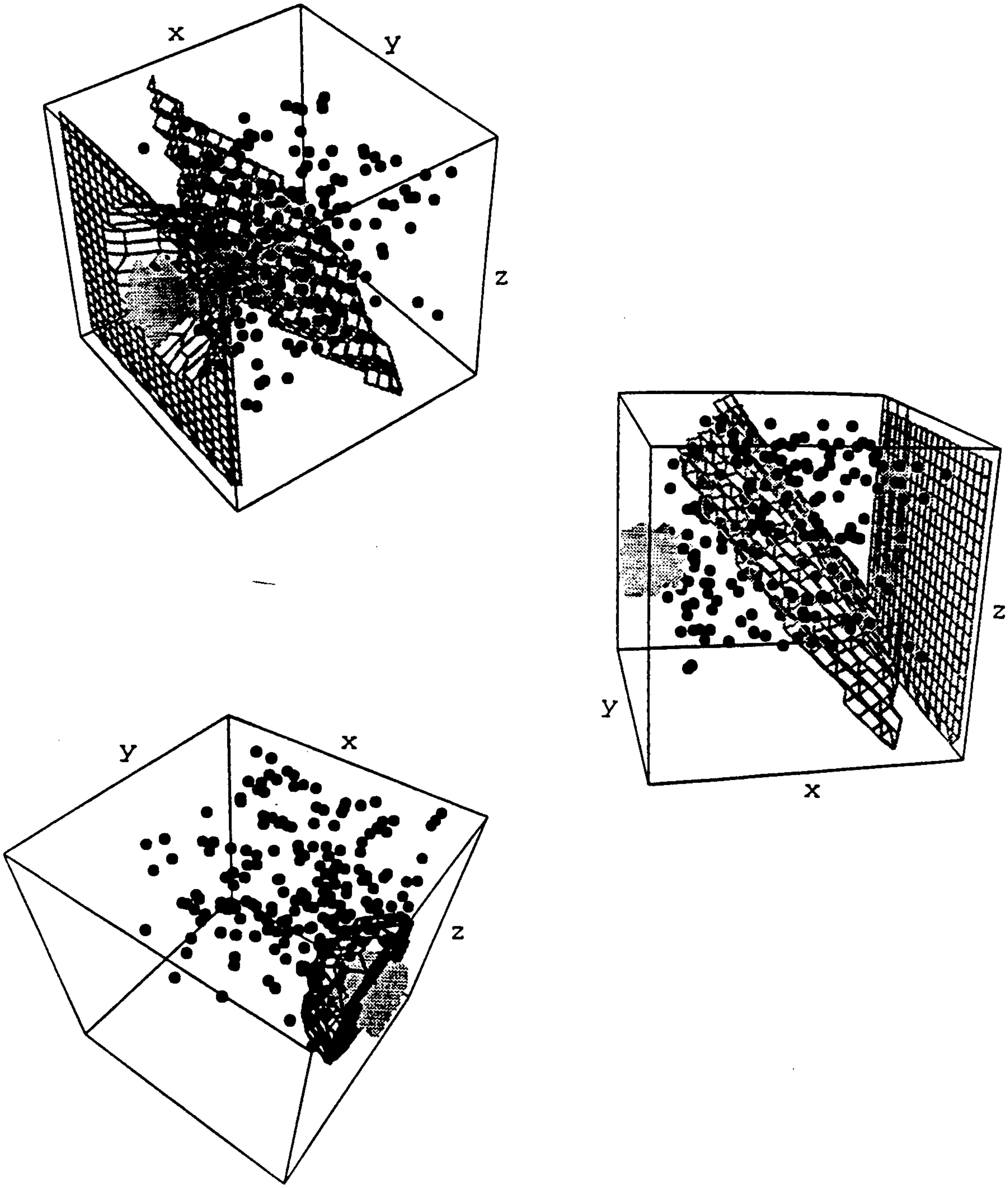


Figure 5.7:

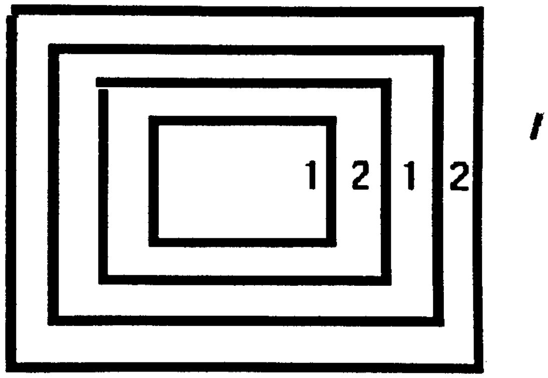


Figure 5.8:

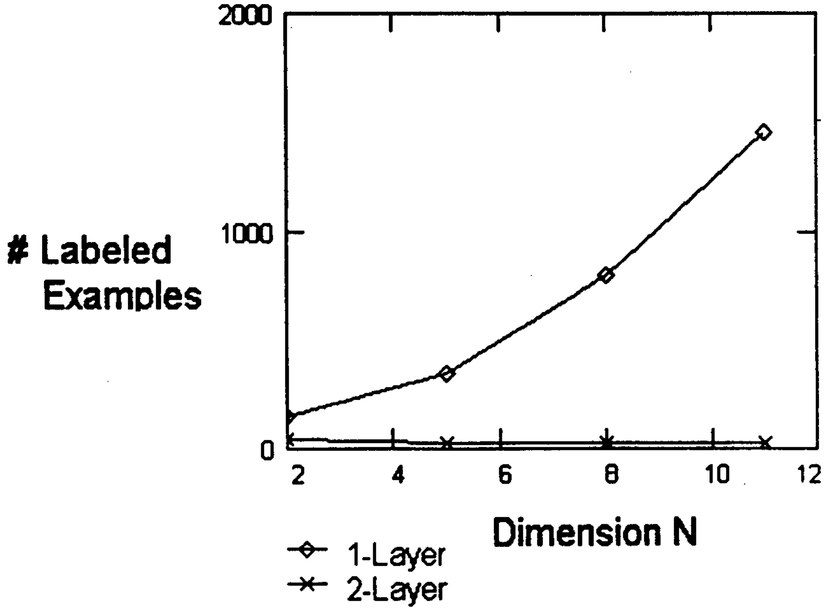


Figure 5.9:

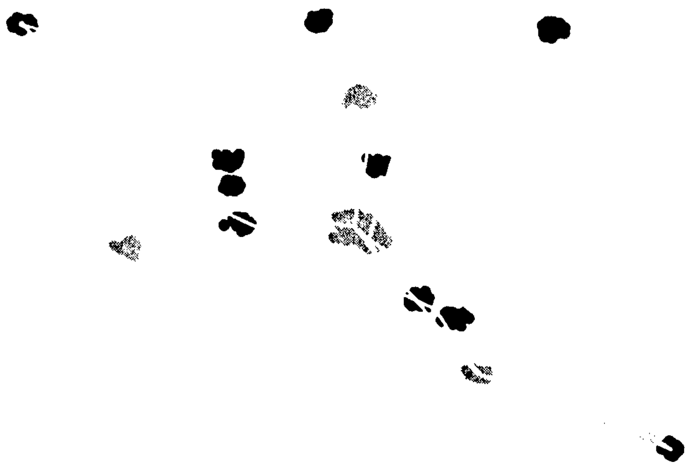


Figure 5.10:

We measured the error versus labeled sample complexity of 20 randomly chosen 1-layer networks and 20 randomly chosen 2-layer networks per dimension N and plotted the results over a range $2 \leq N \leq 12$. As a measure of comparison between the 1-layer and 2-layer net we used the following ratio

$$\frac{m_{1\text{-layer}}/m_{2\text{-layer}}}{P_{error,2}/P_{error,1}} \equiv \frac{R_m}{R_p} \equiv R$$

which measures the number, R_m, of 1-layer examples needed for every one 2-layer example in order to achieve a ratio of R_p 2-layer misclassifications per one 1-layer misclassification. The higher it is the worse the 1-layer performance w.r.t the 2-layer performance. As seen in Figure 5.11, this ratio increases as N increases. This suggests that on average, the 2-layer network requires fewer labeled examples than the 1-layer network, for the random-clusters-on-rings problems, and the saving in labeled examples increases with the input dimensionality.

The ring problems are particularly difficult for a linear partition since class 1 encloses class 2 and the difficulty becomes worse with multiple rings. There are a host of easier problems, e.g., a cluster in the shape of a "C" next to another cluster,

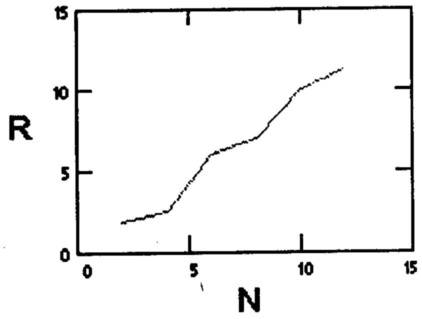


Figure 5.11:

but not containing it. Here the 1-layer network performs better (i.e., fewer labeled examples are needed for same error) then in the multiple ring case, but the 2-layer net still does better than the 1-layer.

We also ran both types of networks on problems with classes that consist of randomly positioned non overlapping clusters but not on ring contours as before. Again, performance represented the labeled sample complexity versus dimensionality N. Here, the 1-layer performed as well as the 2-layer net. Trying to decrease the number of cells in the second layer resulted in poor performance. This is due to the limitation of the nonlinearity of the cells. The decision borders that can be achieved with partitions based on the 2-layer architecture are not better than those that the usual voronoi partition achieves, when considering the average performance over these types of randomly generated problems.

The above ideas fit under a common strategy of sample reduction, namely fitting the family of possible classification decision regions to the family of classification problems. The 2-layer network reduced the sample complexity for problems where the class clusters are distributed over N-dimensional spherical contours.

5.7 The Method of k-Means

The Kohonen neural network uses an adaptation rule for the neural weights that is similar to the *k-means* algorithm (cf. MacQueen [36]). This procedure adapts k vectors, $y_1, \ldots, y_k$, so to minimize the empirical mean square error (MSE)

$$e_n(Y) = \frac{1}{n} \sum_{i=1}^{n} \min_{1 \leq j \leq k} |x_i - y_j|^2$$

where $Y \equiv [y_1, \ldots, y_k]$ with y_i being N-dimensional column vectors and the x_i are the N-dimensional example vectors. The matrix Y defines a voronoi partition with k cells. The true MSE is denoted by

$$e(Y) = \mathbf{E} \min_{1 \leq j \leq k} |x - y_j|^2$$

where expectation is w.r.t. the underlying pattern mixture density $f(x)$. A necessary condition for minima of $e(Y)$ is that the partition must have its k vectors as the centroids of the corresponding cells (cf. Gersho A. [37]).

Using the theory of uniform SLLN we can find the sufficient number of examples x_i, such that for any partition Y, $|e(Y) - e_n(Y)| \leq \epsilon$ (cf. Pollard [21]). It follows that an algorithm which finds a partition Y' that minimizes $e_n(Y)$ in effect finds a partition which minimizes $e(Y)$ to within ϵ accuracy. This idea can be used for classification problems. Suppose one of the r minima of $e(Y)$, denoted by Y^*, is a partition for which $P_{error} = P_{Bayes}$. We can then use unlabeled examples together with an algorithm that locates the r local minima of $e_n(Y)$ to estimate the local minima of $e(Y)$ and use labeled examples to choose the best performing partition out of the r possible ones. This would require an algorithm which uses unlabeled examples to discover consistent estimates of the r global minima of $e(Y)$ and hence would be similar to the mode estimation of Section 5.5.

We look at the problem in which $r = 1$, i.e., $e(Y)$ has one global minimum and hence the labeled examples are not needed for choosing one of several possible partitions, but just to label one partition estimate, $\hat{Y}$, of Y^*.

We first present the algorithm and then derive its sample complexities.

Algorithm S:

`The setting:` $C \subset \mathbb{R}^{2N}$ is a compact parameter space, $Y = [y_1, y_2] \in C$, where Y indexes a two-vector voronoi partition.
The MSE is defined as

$$e(Y) = \mathbf{E} \min_{1 \leq j \leq 2} |x - y_i|^2.$$

The two pattern classes are such that there exists a two-vector voronoi partition, $Y^* \in C$, being the partition of the Bayes classifier, and such that $e(Y^*)$ is the unique global minimum of the function $e(Y)$.

`Given:` m labeled examples and n unlabeled examples drawn according to an unknown mixture $f(x)$.

`Begin:`

1) Using the unlabeled examples find the point $\hat{Y}^*$ which minimizes the empirical MSE, i.e.,

$$\hat{Y}^* = \text{arginf}_{Y \in C} \frac{1}{n} \sum_{i=1}^{n} \min_{1 \leq j \leq 2} |x_i - y_j|^2.$$

2) Form the hyperplane perpendicular to the line through $\hat{y}_i^*$, $i = 1, 2$, and which passes through their midpoint.

3) Label the two decision regions across the hyperplane by the label of the majority of the examples on either side.

`End.`

As an instance of such a problem, consider for simplicity, the classification problem which consists of two pattern classes, each in a cluster sufficiently separated so that the partition, Y^*, defined by the hyperplane perpendicular to the line through the two cluster-centroids, separates the two clusters and achieves the global minimum of $e(Y)$. We now find the sufficient number of unlabeled examples, n, and labeled examples, m, to learn the Bayes rule to within small error and high confidence.

Unlabeled examples are used to compute the empirical MSE $e_n(Y)$. In order for minimization of the empirical MSE to yield a good partition we need

$$\mathbf{P}\left(\sup_{[y_1,y_2]\in C}\left|\mathbf{E}\min\{|x-y_1|^2,|x-y_2|^2\}-\frac{1}{n}\sum_{i=1}^{n}\min\{|x_i-y_1|^2,|x_i-y_2|^2\}\right|>\epsilon\right)\leq\delta/2 \tag{5.23}$$

where δ,ϵ are two arbitrarily small positive constants, and y_1,y_2 are the two vectors defining a partition Y whose two regions are

$$R_1=\{x:|x-y_1|<|x-y_2|\},\quad R_2=\{x:|x-y_1|\geq|x-y_2|\}$$

and C is a compact subset of $\mathbb{R}^N\times\mathbb{R}^N$ which contains the optimal $[y_1^*,y_2^*]$-based partition which achieves the minimum MSE.

In order to use Theorem 3.9 we need to define a class of bounded functions. Let $g_{y_1,y_2}(x)=\min\{|x-y_1|^2,|x-y_2|^2\}$. Then

$$\begin{aligned}
&\mathbf{P}\left(\sup_{[y_1,y_2]\in C}\left|\mathbf{E}g_{y_1,y_2}(x)-\frac{1}{n}\sum_{i=1}^{n}g_{y_1,y_2}(x_i)\right|>\epsilon\right)\\
&\leq\ \mathbf{P}\left(\sup_{(y_1,y_2)\in C}\left|\int_D g_{y_1,y_2}(x)dP-\frac{1}{n}\sum_{i=1}^{n}g_{y_1,y_2}(x_i)1_{x_i\in D}\right|>\epsilon/2\right)\\
&\ +\ \mathbf{P}\left(\sup_{[y_1,y_2]\in C}\left|\int_{D^c} g_{y_1,y_2}(x)dP-\frac{1}{n}\sum_{i=1}^{n}g_{y_1,y_2}(x_i)1_{x_i\in D^c}\right|>\epsilon/2\right)
\end{aligned} \tag{5.24}$$

where D is a compact subset of $\mathbb{R}^N$. Define the class

$$\mathcal{H}\equiv\{g_{y_1,y_2}(x)1_{x\in D}:[y_1,y_2]\in C\}.$$

The functions in $\mathcal{H}$ are bounded since

$$\min\{|x-y_1|^2,|x-y_2|^2\}\ \leq\ 2|x|^2+2|x|(|y_1|+|y_2|)+|y_1|^2+|y_2|^2\leq M$$

as $|x|^2$ is bounded over the compact region D and $|y_1|^2,|y_2|^2$ are bounded over the compact set C. Theorem 3.9 can be used to bound the first term of (5.24) by $\delta/4$.

The second term is bounded by

$$\mathbf{P}\left(\sup_{[y_1,y_2]\in C}\left|\int_{D^c} g_{y_1,y_2}(x)dP\right| + \sup_{[y_1,y_2]\in C}\left|\frac{1}{n}\sum_{i=1}^{n} g_{y_1,y_2}(x_i)1_{x_i\in D^c}\right| > \epsilon/2\right)$$

and

$$\sup_{[y_1,y_2]\in C}\left|\int_{D^c} g_{y_1,y_2}(x)dP\right| \le \int_{D^c}\left|\sup_{[y_1,y_2]\in C} g_{y_1,y_2}(x)\right| dP \le \int_{D^c}|x|^2dP + M_1\int_{D^c}|x|dP + M_2$$

where $M_1, M_2 < \infty$ and we used the fact that $|y_1|^2, |y_2|^2$ are bounded over C. We assume that the class mixture $f(x)$ is such that there exists a compact D that makes these last integrals arbitrarily small; in particular, for the case in which the class clusters are bounded we can let D be a ball which contains both clusters which makes these integrals equal zero since the probability measure (corresponding to the mixture distribution) is zero outside the ball. Hence we can bound the the right side by an arbitrarily small quantity $\Delta \ge 0$. As on page 67, the second term of (5.24) is then bounded by $\delta/4$ where δ is a given arbitrary confidence parameter.

Proceeding to find a bound on the first term of (5.24) by using Theorem 3.9, we only need to calculate the VC dimension of the class $\mathcal{H}$ of functions. First we find the VC of the class $\mathcal{A}$ of functions $\{|x-a|^2 : a \in \mathbb{R}^N\}$. A function here can be expressed as a linear combination

$$|x-a|^2 = \sum_{i=1}^{N} x_i^2 - 2\sum_{i=1}^{N} a_i x_i + |a|^2 = \sum_{i=1}^{2N+1} \alpha_i \phi_i(x)$$

where α_i are constants and the $\phi_i(x)$ are basis functions. The class of graphs of these functions has VC $= 2N+1$ by Theorem 3.6 and by Definition 3.5 it follows that $VC(\mathcal{A}) = 2N+1$. The graphs of functions of $\mathcal{H}$ are intersections of graphs of functions of $\mathcal{A}$ intersected with a fixed set $\{(x,z) : 0 \le z \le 1_{x\in D}\}$, i.e.,

$$\{(x,z) : 0 \le z \le |x-y_1|^2\} \cap \{(x,z) : 0 \le z \le |x-y_2|^2\} \cap \{(x,z) : 0 \le z \le 1_{x\in D}\}.$$

Given an m-sample, the number of dichotomies achieved by such class of graphs is at most $m^{VC(\mathcal{A})}m^{VC(\mathcal{A})}$. Finding the largest m such that it equals 2^m yields

$$VC(\mathcal{H}) \leq 2(4N+2)\log(4N+2).$$

Plugging this into Theorem 3.9 we have the sufficient number of unlabeled examples n in order for (5.23) to be satisfied is

$$n \geq \frac{64M^2}{\epsilon^2}\left((16N+8)\log(4N+2)\log\frac{16eM}{\epsilon} + \log\frac{16}{\delta}\right).$$

With ϵ-accuracy when estimating $e(Y)$ by $e_n(Y)$, implies we can estimate $Y^* \equiv [y_1^*, y_2^*]$ by $\hat{Y}^* \equiv [\hat{y}_1^*, \hat{y}_2^*]$ s.t. $|y_i^* - \hat{y}_i^*| \leq \alpha$, $i = 1,2$ where $\alpha > 0$ is arbitrary small depending on ϵ, Y^*, and $\hat{Y}^*$. Developing on this theme, we have

$$e(\hat{Y}^*) - \epsilon \leq e_n(\hat{Y}^*) < e_n(Y^*) \leq e(Y^*) + \epsilon$$

where the middle inequality follows from $\hat{Y}^* \equiv \text{argmin}_{Y \in C} e_n(Y)$. So

$$\left|e(Y^*) - e(\hat{Y}^*)\right| \leq 2\epsilon.$$

Now assume that $e(\cdot)$ is continuous and 1-1 at least in some ball around Y^* (for conditions cf. Apostol [28] p. 370). Then (cf. Rudin [27] p. 90) its inverse is continuous there and therefore

$$\left|e(Y^*) - e(\hat{Y}^*)\right| \leq 2\epsilon \Rightarrow \left|Y^* - \hat{Y}^*\right| \leq \alpha$$

for small enough $\epsilon > 0$, where α is arbitrarily small and depends on ϵ, $\hat{Y}^*$, Y^*. Hence the learner draws n (as above) unlabeled examples then locates the argmin of $e_n(Y)$. This vector, $\hat{Y}^*$, is α-close to the vector Y^* at which the minimum of the true MSE, $e(\cdot)$, occurs. That implies, $|\hat{y}_i^* - y_i^*| \leq \alpha$, $i = 1,2$. Recall that by assumption, the partition based on the hyperplane which is associated with this Y^*

achieves the minimum P_{error}. From Section 4.1 we have $P_{error} = P_{Bayes}(1 + O(\alpha^2(\epsilon)))$ when labeled optimally. From Section 4.3 we need $m = A \log \frac{1}{\delta}$, for some constant A, labeled examples to guarantee that the probability of not labeling optimally is at most $\delta/2$.

Combining all of the above it follows that with confidence $> 1 - \delta$, and

$$n \geq \frac{64M^2}{\epsilon^2}\left((16N+8)\log(4N+2)\log\frac{16eM}{\epsilon} + \log\frac{16}{\delta}\right)$$

unlabeled examples and

$$m = A \log \frac{1}{\delta}$$

labeled examples, algorithm S which minimizes the empirical MSE finds a classification rule which has

$$P_{error} = P_{Bayes}(1 + c\alpha^2(\epsilon))$$

for some positive constant c and where $\alpha(\cdot)$ depends on the problem by depending on the MSE function $e(Y)$.

Chapter 6

Conclusions

Based on the finite sample complexity results we now discuss their implications on the worth of a label example under several different scenarios. All constants c_i, $i = 0, 1, 2 \ldots$ are finite and positive.

First we compile the results concerning the learning of a classification problem with an underlying Gaussian mixture. In the following discussion, f will denote the unknown underlying Gaussian mixture of two equiprobable pattern classes, unless stated otherwise.

We showed in Section 4.2 that with knowledge of both the parametric form of f and that f is a member of an identifiable family, algorithm M suffices with

$$n_M = c_1 \frac{N^2}{\epsilon^3 \delta} \left(N \log \frac{1}{\epsilon} + \log \frac{1}{\delta} \right)$$

unlabeled examples

$$m_M \ = \ c_2 \log \frac{1}{\delta}$$

labeled examples. We compare this to the purely-labeled sample scenario of Section 4.1 where a learner using algorithm E required

$$m_E = \frac{4N}{\epsilon} \log \left(\frac{8N}{\delta} \right)$$

labeled examples to learn the same problem. Clearly there is a reduction by introducing n_M unlabeled examples. Dividing n_M by the difference $m_E - m_M$ yields a rough estimate of the worth of a label example, namely

$$\frac{c_3 N^2}{\epsilon^3 \delta \log N} \tag{6.1}$$

unlabeled examples. This is polynomial in N, $\frac{1}{\epsilon}$, and $\frac{1}{\delta}$.

In Section 5.2 we showed that when learning the same classification problem but not knowing the parametric form of f nor the fact that the class of Gaussian mixtures is identifiable, algorithm K required

$$n_K = c_4 \frac{13^{N \log(5+\log N)}}{\epsilon^{\frac{N}{2\log N}}} \log \frac{1}{\epsilon\delta}$$

unlabeled examples and

$$m_K = c_5 \log \frac{1}{\delta}$$

labeled examples. As before, comparing this to m_E we have that a labeled example is worth roughly

$$\frac{c_6 13^{N \log(5+\log N)}}{N \log N \epsilon^{\frac{N}{2\log N}}} \tag{6.2}$$

unlabeled example.

This is roughly exponential in N and $\frac{1}{\epsilon}$ and is therefore considerably more than the previous polynomial expression. As discussed in Chapter 5, the same n_K and m_K apply also for algorithm K when learning a function with a more general form than the Gaussian mixture f. So the reason that it takes significantly more unlabeled examples to learn f with algorithm K than with algorithm M is due to the larger complexity of the family of functions of which f is a member under algorithm K.

Therefore when learning the same Gaussian mixture f, under the nonparametric scenario, a labeled example is worth an exponentially more unlabeled examples than in the parametric scenario.

We can equivalently express the above results by showing $P_{error}(m, n)$ as a function of m and n, i.e., $P_{error}(m, n) = g(m) + h(n)$. For the mixed sample learning, in both the parametric and the nonparametric scenarios the function $g(m)$ is $O(e^{-b_0 m})$, $m \to \infty$, for some constant $b_0 > 0$ independent of N. For the parametric case the function $h(n)$ is $O\left(\left(\frac{N}{n}\right)^{b_1}\right)$, $n \to \infty$, and in the nonparametric case the function $h(n)$ is $O\left(\left(\frac{b_2}{n^{2/N}}\right)^{\log N}\right)$, $n \to \infty$, where $b_1, b_2 > 0$ are independent of N. Cover & Castelli [5] report a similar expression for the P_{error}, namely a polynomial decrease in n and exponential decrease in m however our results specifically point out the dependency on the dimensionality N. For fixed N, it is clear that in both scenarios the number of unlabeled examples is exponentially more than the number of labeled examples. With variable N we conclude that the value of a labeled example is exponentially more when in the nonparametric scenario. In the purely labeled sample case, $P_{error}(m)$ is $O\left(\left(\frac{N}{m}\right)^{b_4}\right)$, $m \to \infty$, for $b_4 > 0$ is constant with N. Hence the rate of decrease of the error w.r.t. m is exponential when unlabeled examples are introduced to the learning. Without unlabeled examples, it is only polynomially fast.

In Sections 4.4, 4.5, we considered the same Gaussian mixture problem, but with general *a priori* class probabilities, p and $1-p$. The algorithms E_p, M_1, and M_2 assume that f is a member of a parametric family of identifiable mixtures. The parameter vector indexing an f in this family is $[\theta, p]$, where p is the class "1" *a priori* probability and θ is the two-means vector. W.l.o.g we assume $p < 1-p$. We now determine the tradeoff between the number of labeled and unlabeled examples, under the case of general $0 < p < \frac{1}{2}$ in the following three scenarios: (1) both θ and p are learned using a purely labeled sample (2) θ is learned using an unlabeled sample, p is learned using a labeled sample (3) both θ and p are learned using an unlabeled sample. (The last two are the mixed-sample cases.)

The sample complexities sufficient to achieve $P_{error} \leq P_{Bayes}(1 + c_7\epsilon)$ with confidence $> 1-\delta$ are as follows: in (1), for constant p, the sufficient labeled sample is

$$m_{E_p} = c_8 \frac{N}{\epsilon^2} \log \frac{N}{\epsilon}.$$

When p varies it effects m through a factor polynomial in $\frac{1}{p}$, however as $p \to 0, m \to 1$ while $P_{error} \to 0$.

In (2), for fixed p, the labeled sample complexity decreases to

$$m_{M_1} = c_9 \frac{1}{\epsilon^2} \log \frac{1}{\epsilon}$$

which is independent of N, on account of introducing

$$n_{M_1} = c_{10} \frac{N^3 \log \frac{1}{\epsilon}}{\epsilon^6}$$

unlabeled examples. The parameter p effects n through a factor of $\log^2 \left(\frac{1}{p}\right)$ and m through a factor polynomial in $\frac{1}{p}$ but $n \to 0$ and $m \to 1$ as $p \to 0$.

In case (3), the labeled sample m_{M_2} is practically a constant, while for a fixed p, the unlabeled sample size is

$$n_{M_2} = c_{11} \frac{N^3 \log \frac{1}{\epsilon}}{\epsilon^6}$$

For variable p, n depends on p through a factor of $\frac{\log^2(\frac{1}{p})}{p^{14}}$, however as $p \to 0, n \to 0$, and $m \to 1$.

From the above it follows that n,m, are effected by p in the worst-case through a factor polynomial in $\frac{1}{p}$. When p is estimated using the unlabeled sample, n grows w.r.t. p, faster than the rate at which m grows w.r.t. p under the scenario where the labeled sample is used to estimate p. So there is a tradeoff between cost and amount of examples— unlabeled examples are cheaper but more of them are needed. The same situation also applies for the estimation of the means—when unlabeled examples are used, more of them are needed than when labeled examples are used.

For $p \to 0$, the labeled sample goes to 1, and the unlabeled sample goes to 0. This is anticipated since small p implies that one of the pattern classes has a negligible effect on the decision rule and hence can be ignored, letting all of $\mathbb{R}^N$ have the complement label while suffering a small misclassification error. The algorithms that are used, all need at least one labeled example.

Observing the number of unlabeled examples needed to reduce from m_{E_p} to m_{M_2}, we have as a rough estimate that one labeled example is worth

$$c_{12}\frac{N^2 \log \frac{1}{\epsilon}}{\epsilon^4 p^{11} \log N}$$

unlabeled examples, which is a polynomial in $\frac{1}{p}$.

We also partially analyzed algorithm M which is based on the MLE technique, for the two classes having the same non unit covariance matrix. We encountered some difficulty in the part of the proof (of Theorem 4.2) where it is necessary to show the independence of the function $\Phi(\theta)$ for all N greater than some constant. However in all other parts there was no difficulty (the work is more involved than for the unit covariance case). In particular, we let the parameter $\theta = [\mu_1, \mu_2, \Sigma^{-1}]$ and under the condition that the unknown mixture $f(x|\theta_0)$ has a nonsingular covariance matrix Σ_0, we can define a compact set containing θ_0 and not containing any points with singular covariance matrices. This enables us to define a *bona fide* class of bounded functions, indexed by the parameter θ, and thereby be able to apply the uniform SLLN theorems. We conjecture that for the part of the proof which was not yet completed, there exists a way to find the necessary symmetry in the integrals that define $\Phi(\theta)$ such that for $N \geq N_0$, $\sup_{\theta \notin B(\theta_0,\epsilon)} \Phi(\theta)$ is constant, where $3 < N_0 < \infty$. Based on this, the unlabeled sample complexity will still remain polynomial in N, $\frac{1}{\epsilon}$, $\frac{1}{\delta}$.

In Section 5.7 we considered the k-means method—a nonparametric method to learning classification which is based on an *ad hoc* clustering approach. As before, the classifier here consists of a partition and a labeling of each of the decision regions.

The partition comprises of voronoi cells, each associated with a vector y_i, hence the partition can be indexed by a matrix Y of k vectors $y_i \in \mathbb{R}^N$, $1 \leq i \leq k$. The classical vector-quantization techniques (cf. Gersho [37]) uses such a partition as a mapping from the input space containing the vector x to the output space which is the finite set of vectors y_i, $1 \leq i \leq k$. A common measure of discrepancy between the random input x and the output y is the means squared error (MSE). Thus a good partition, for vector quantization, is one which minimizes the MSE. A classifier can be constructed based on the partition provided that each voronoi cell is assigned a label of either class. There are several labelings possible and the learner is to pick the one which minimizes the probability P_{error} of misclassification. In general, a partition which has a minimum MSE does not necessarily has a minimum P_{error} under optimal labeling but for some problems the MSE partition does yield a good classifier.

We considered the problem for which there exists a unique minimum MSE partition with $P_{error} = P_{Bayes}$. Using algorithm S we showed that it is sufficient to have n_{MS} be polynomial in $\frac{1}{\epsilon}$, $\log\frac{1}{\delta}$, and N, to achieve P_{error} which is $c_{13}\alpha^2(\epsilon)$ where α is a function depending on the smoothness of the MSE function $e(Y) = \mathbf{E}\min\{|x-y_1|^2, |x-y_2|^2\}$ over the region $C \subset \mathbb{R}^{2N}$. Here α is analogous to the function h of Section 5.5 in that they both represent the worst misclassification error deviation when the uniform deviation between the empirical and the true means over a class of functions is at most ϵ. The labeled sample size m_{MS} is practically an absolute constant.

When compared to the results of algorithm K, it may first seem surprising that this nonparametric k-means classification scheme requires only a polynomial number of unlabeled examples. However we note that algorithm S is really parametric since it searches for an optimal partition, or its associated parameter Y^*, in a Euclidean parameter space, (we still call this a nonparametric approach to classification since knowledge of the parametric form of the class densities is not required.) So we should expect its sample complexities to be of the same order of magnitude as for algorithm M which is also based on a search for a function in class of parametric functions. Consequently we do not anticipate such a parametric approach to perform satisfactorily on a rich nonparametric family of density mixtures since the complexity of such a family of functions mismatches the complexity of the parametric function class on which algorithm S is based. However with heuristics, as for instance in the different variants of the LVQ algorithm (cf. Kohonen [24]) where the partition is adjusted using labeled examples, it may be possible to improve the classification error *ad hoc*.

We now discuss another possible approach to estimate the modes of f by using the kernel technique. In Chapter 5, algorithm K estimated $f(x)$ using the kernel technique *uniformly* for all $x \in \mathbb{R}^N$, i.e. we used $\sup_{x \in \mathbb{R}^N} \frac{|f_n(x)-f(x)|}{\sup_x f(x)}$ as the estimation discrepancy. However only the modes of f were needed by the algorithm for constructing the decision border. This suggests that perhaps it suffices to estimate the functional values of f at its k modes η_i, $1 \leq i \leq k$. The difficulty is that the modes η_i are not known hence one cannot even specify what is there to be estimated. However suppose that the learner does know that the modes of f are restricted to be in a ball B of radius ρ centered at some point x_0 of $\mathbb{R}^N$. He can then consider an ϵ-cover, S, (w.r.t. the Euclidean norm) of B having a cardinality which is bounded above by $s_\epsilon \equiv (\frac{2\rho}{\epsilon})^N$. (For brevity we also denote it by s). By definition, for any point $x \in B$ there exists a $y_j \in S$ such that $|y_j - x| \leq \epsilon$. For continuous f and small enough $\epsilon > 0$ we can guarantee that there exist $y_j \in S$, each ϵ-close to its corresponding η_j such that $|f(y_j) - f(\eta_j)| \leq \alpha$ for arbitrary $\alpha > 0$, and $1 \leq j \leq k$. (Note, the learner knows the points y_j, $1 \leq j \leq s$.)

The learner may then use the kernel technique to generate the s values, $f_n(y_j)$, as α-estimates of the points $f(y_j)$, i.e.,

$$f_n(y_j) = \frac{1}{n}\sum_{i=1}^{n} \sigma^{-N} K\left(\frac{x_i - y_j}{\sigma}\right), \quad 1 \leq j \leq s,$$

where x_i, $1 \leq i \leq n$ are the unlabeled examples, such that $|f_n(y_j) - f(y_j)| \leq \alpha$. It follows that for each η_i, $1 \leq i \leq k$, there is a subset A_i of S which contains points y_j such that $|f_n(y_j) - f(\eta_i)| \leq 2\alpha$, and from above, there exists such a point which is ϵ-close to η_i. Thus using a variant of algorithm K the learner can obtain ϵ-close estimates of the modes of f.

It only remains to show the sample complexities for this approach. The analysis follows identically as the one in Section 5.2, except now the class $\mathcal{K}_\sigma$ is defined as

$$\mathcal{K}_\sigma = \left\{K_{y_j,\sigma}(x) : y_j \in S\right\}$$

instead of

$$\mathcal{K}_\sigma = \left\{K_{y,\sigma}(x) : y \in D \subset \mathbb{R}^N\right\}$$

which was the case in Chapter 5 where $f_n(y)$ estimated $f(y)$ uniformly for $y \in \mathbb{R}^N$. We can use the cardinality $|\mathcal{K}_\sigma|$ instead of a covering number (and hence not requiring VC($\mathcal{K}_\sigma$)). From above we have $|\mathcal{K}_\sigma| = s_\epsilon$. Following (5.12) and using the cardinality w.r.t. the appropriate accuracy yields a sample complexity

$$n \;=\; c_{14}\frac{4^{N\log(5+\log N)}}{\epsilon^{\frac{N}{2\log N}}}\log\frac{\rho}{\epsilon\delta}$$

which differs from the n of Theorem 5.1 in the 4 instead of the 13. So this other approach contributes to an exponential reduction in the number of unlabeled examples from our algorithm K approach however the resulting unlabeled sample size is still exponential in N and $\frac{1}{\epsilon}$, and an additional constraint is for the learner to know *a priori* the region B which contains the true modes of f.

As a summary, in this thesis we considered an interesting simple question raised by T. M. Cover [5], which asks what is the tradeoff between labeled and unlabeled examples for learning a classification rule. We used theory from statistics and mathematical analysis to formalize this question in terms of a probabilistic setting in which sample sizes not only influence the misclassification accuracy but also the confidence of the classification decision. It became clear as we began to get expressions for the value of a labeled example, that side information is a crucial assumption in learnability. That is, learning by examples is very dependent on what is assumed to be known by the learner *a priori.*

It seems that there should be a way to represent both examples and side information in the same model of learning, so that for instance, one would be able to tell how many more examples are needed if less side information is at hand, or vice versa. This is an interesting theoretical question. If an elegant and intuitive model could be contrived to represent it then its consequences will have strong theoretical, if not also practical, implications on our understanding of information in conjunction with learning by examples.

There has been some related work on this question, cf. Abu-Mostafa [45] represents this in terms of giving the learner hints and uses VC dimension arguments to formalize it. Haussler, Kearns & Schapire [44] consider a Bayesian-information theoretic model in which the unknown function is drawn according to some prior density from a function class, and different degrees of side information are represented by using different types of priors.

From our work we saw that in both the parametric and nonparametric scenarios, there was a primary function class defined, which was indirectly related to the class of mixtures that contained the unknown underlying mixture. This primary function class in essence represents the "engine" of the so called uniform SLLN—the mathematical machinery that produced for us all the sample complexity results. For instance, going from the nonparametric kernel technique to the parametric MLE takes us from the function class

$$\mathcal{K}_\sigma = \left\{K_{\sigma,x}(y) : x \in E \subset \mathbb{R}^N\right\}$$

to

$$\mathcal{G} = \left\{g(x|\theta) = \log f(x|\theta)1_D(x) : \theta \in \Theta \subset \mathbb{R}^{2N}\right\}.$$

Both engine classes have a complexity measure, namely the covering number or the VC-dimension, as we showed in preceding chapters. The complexity of the class $\mathcal{K}_\sigma$ is exponential in N while the complexity of $\mathcal{G}$ is linear in N. So it seems intuitive that whatever side information is available to the learner in the parametric scenario, and which is abeyant in the nonparametric kernel scenario, could possibly be represented by some function of the difference between the complexities of this two function classes. It is as though the teacher points his finger at a lower dimensional area in some high dimensional space, and thereby communicates to the learner this side information, i.e., restricting the learner to search for the unknown function over a a less complex set of functions which contains the unknown desired function. This results in requiring fewer examples, for instance, as we saw in the significant reduction of the unlabeled sample complexity when going from the kernel to the MLE scenario. In ongoing research, we are investigating possible models that represent these ideas in a formal framework.

We have used the majority rule algorithm with the labeled examples for all the mixed-sample learning algorithms when choosing the labeling of the partition of the classifier. It is worth noting that the same rule can be used in the operation of the classifier, i.e., when one needs to test a hypothesis of having class "1" or "2". One can take r test examples and let the classifier assign a labeling to each example. Then choose the hypothesis which corresponds to the label of the majority of the r examples. As in the classical hypothesis testing, the larger r the better performance, i.e., the lower the probability of making a bad hypothesis decision. In fact, the probability of error decreases to 0 exponentially fast with r.

Extensions to the case of more than two pattern classes and to other parametric forms can be tackled using the same approach as we have taken here. In the parametric approach one would need to have identifiable mixtures and the MLE analysis will be very similar to the Gaussian case. The function class over which the uniform SLLN is to be applied will possibly have a different complexity but in most cases we expect a polynomial growth for the number of unlabeled examples w.r.t. the important variables $\frac{1}{\epsilon}$, $\frac{1}{\delta}$, and the dimensionality N. In the nonparametric scenario, it would be interesting to extend and determine other classes of mixtures where the modes of a mixture f can determine the Bayes partition, in particular for the cases of more than two pattern classes and for nonlinear decision borders.

An interesting extension for learning classification is using examples that can take on a label of a fuzzy nature. For instance one type of such examples is denoted by (x_i, y_i), $1 \leq i \leq l$, where $x_i \in \mathbb{R}^N$ is a feature vector and $0 \leq y_i \leq 1$ represents the probability that x_i is of class 1. If $y_i = 1$ then it corresponds to having a labeled example from class 1, while if $y_i = 0$ then it is a labeled example from class 2. When $y_i = \frac{1}{2}$ the example is considered as unlabeled. This kind of examples therefore allow for a full spectrum of confidence in the label (i.e., from 0 to 1) and may be useful in situations where the teacher can only provide likelihoods or confidences about the true origin of a particular feature vector x. Referring to this type of examples as the fuzzy kind it is interesting to ask what is the sample complexity, l, for learning a decision rule using examples of this kind. One approach is to use the l-sample to estimate the *a posterior* functions $p(1|x)$ and $p(2|x)$ which directly identify the Bayes rule as in (1.1). It suffices to estimate $p(1|x)$ as $p(2|x) = 1 - p(1|x)$. Let us assume that the mixture density $f(x)$ and the class conditional densities are parametric. In this case the function $p(1|x)$ is a member of a parametric class of functions, indexed by a finite real vector ϕ and denote it by $p_\phi(1|x)$, where $\phi = [p_1, p_2, \theta_1, \theta_2]$ also indexes the mixture $f(x|\phi)$ in the class of mixtures. We can let the teacher draw (x_i, y_i) according to some arbitrary probability density $g(x,y)$ and define a discrepancy measure as $\mathbf{E}\left(p_\phi(1|x) - y\right)^2$ where y is a function of x and represents the true unknown $p_{\phi_0}(1|x)$, and $\phi, \phi_0 \in A \subset \mathbb{R}^k$ where k does not necessarily equal the dimension N of the feature vectors x (the expectation is w.r.t. $g(x,y)$). We can apply the uniform SLLN over the class of functions $\mathcal{H} = \{h_\phi(x,y) \equiv \left(p_\phi(1|x) - y\right)^2 : \phi \in A \subset \mathbb{R}^k\}$ to get $\sup_{\phi\in A}\left|\mathbf{E}h_\phi(x,y) - \frac{1}{l}\sum_{i=1}^{l} h_\phi(x_i,y_i)\right| \leq \epsilon$ with confidence $> 1-\delta$.

The learner then finds a function h_{ϕ^*} which minimizes the empirical mean $\frac{1}{l}\sum_{i=1}^{l}$ $h_\phi(x_i, y_i)$ over all $\phi \in A$, and for sufficiently small $\epsilon > 0$ it follows that $p_{\phi^*}(1|x)$ is a close estimate of the true unknown *a posterior* $p_{\phi_0}(1|x)$ in the MSE sense w.r.t. any probability density $g(x,y)$ (due to the distribution independent results of the uniform SLLN theorems). In order for this to hold a sufficient sample complexity can be calculated by determining a bound on the covering number of $\mathcal{H}$. Roughly speaking since the parameter set A is in $\mathbb{R}^k$ which is also the parameter space of the mixture density $f(x|\phi)$ then the size l of the fuzzy sample will differ by a constant factor (w.r.t. the dimensionality k of the parameter space) from the unlabeled sample size n that is sufficient to estimate the parametric mixture $f(x|\phi)$. In all the mixed sample cases investigated earlier we saw that the number of labeled examples m is only $c\log\frac{1}{\delta}$ which is significantly smaller than n so therefore l is of the same order as total $m+n$.

Hence when learning a classification problem one can either use unlabeled examples to estimate the class conditional densities and subsequently the decision regions and then label them using the labeled sample or use a fuzzy sample to estimate the *a posterior* functions $p(1|x)$ and $p(2|x)$ which directly result in an estimate of the Bayes decision rule. In both cases the total number of examples is roughly the same. There are many other types of samples that one may investigate, for instance, introduce a noise component to the labels by making the label be a random variable z which takes on the value y (which is the true label) with probability $1 - \alpha$ for small $\alpha > 0$ and takes on the complement label y^c with probability α. This type of examples are useful when representing the possibility of miscommunication between the teacher and the learner. Clearly an investigation of learning with such examples and other variants of this type are interesting and will require more work.

In the mixed sample approaches we considered both the unlabeled and labeled examples as being drawn according to the underlying true unknown densities $f_1(x)$ and $f_2(x)$. This represents a natural setting in which there is no teacher which controls the learning process but instead a passive "nature" presents the examples. It is a suitable representation when learning is primarily done using unlabeled examples as was true in the cases we investigated and as we mentioned above side information is related to the complexity of the engine-class of functions over which the uniform SLLN is applied.

However when dealing with a purely labeled sample or a fuzzy sample as above there is an obvious role for a teacher to represent side information— having an active teacher which is free to choose a particular probability density $g(x, y)$ for randomly drawing the examples. In the previous discussion regarding the fuzzy sample the sample size l was distribution independent, hence the freedom to choose $g(x, y)$ was not exploited. When the teacher uses a particular distribution to draw the examples he restricts the learner to search for the unknown function in an effectively simpler class of functions, i.e., one whose complexity is the covering number w.r.t. probability distribution $g(x, y)$, which may be smaller than the upper bound based on the VC-dimension of (3.8). This complexity measure has a direct bearing on the sample complexity (see (3.9)) hence it may be possible to reduce the sufficient sample sizes (both for the purely labeled and the fuzzy sample) by selecting particular probability densities $g(x, y)$ which is effectively showing side information. When the sample space is discrete and the functions are indicator of sets (cf. Benedek & Itai [49]) it is clearly seen that the sample size l is directly related to the distribution $g(x, y)$. Using these ideas it is possible to choose distributions that place mass only on "interesting" or relevant sets of the class and therefore in effect reduce the complexity of the class resulting in a reduction of the sample size. For more related work see Benedek & Itai [50], Barlett & Williamson [51].

The subject of animal learning is related to learning with labeled and unlabeled examples. In real life, an animal gets penalized when making a wrong choice. The penalty can be viewed as the negative label. In this respect, it is reasonable to expect that an animal tends to minimize the number of labeled examples that it needs for learning basic primitive tasks as the cost of negative labeled examples is high (for instance, a negative label may mean the animal looses a limb or perhaps its life). Young animals in the wild need to learn very quickly (relative to humans) in order to achieve the stage in which they no longer rely on their parents hence considering that labeled examples are rare or costly (especially for animals who do not have a language of exact communication) it is rational to suppose that animals rely on learning with unlabeled examples which are abundant in their natural habitat and less on labeled examples (of course the genetic factor is also very important here since it may result in many fewer things necessary to learn). The biological nervous system in particular the brain of the animal perhaps uses mechanisms or neural architectures which put more weight on learning with unlabeled examples and minimize the need for labeled examples.

The self organizing neural networks that we considered in Chapter 5 are similar to some biological neural networks (cf. Kohonen [24]). It is known that topological maps similar to those which arise in self organizing Kohonen neural networks (which use primarily unlabeled examples) are common in the real brain. In our simulations we saw that specific neural architectures need fewer labeled examples therefore it is conceivable that biological neural networks possess architecture that need fewer labeled examples. This may be achieved by specialization, i.e., networks that represent decision rules which are based on partitions having separating surfaces that are fit to a particular class of pattern classification problems.

Bibliography

[1] Duda R. O. and Hart P. E., *Pattern Classification and Scene Analysis* (1973) John Wiley & Sons, New York

[2] Fukunaga K., *Introduction to Statistical Pattern Recognition*, (1972) Academic Press, New York

[3] Izenman A. J., Recent Developments in Nonparametric Density Estimation, *Journal of the American Statistical Association* (1991), Vol. 86, No. 413.

[4] McClelland J. L., Rumelhart D. E. and the PDP Research Group, *Parallel Distributed Processing*, MIT Press, 1986.

[5] Cover T. M., Castelli V., in *Computational Learning Theory*, Valiant L. G., Warmuth M. K., ACM SIGACT/SIGART August 5 1991, Morgan Kaufmann Publishers Inc., San Mateo CA.

[6] Cover T. M., Castelli V., in *Advances in Neural Information Processing System* (1991), Moody J. E., Hanson S. J., Lippmann R. P., Morgan Kauffmann Publishers Inc., San Mateo CA.

[7] Redner R. A. and Walker H. F., Mixture Densities, Maximum Likelihood and the EM Algorithm, *SIAM Review* (1984) Vol. 26, No. 2

[8] Wald A., Note on the consistency of the Maximum Likelihood Estimate *Annals of Mathematical Statistics* (1949) Vol 20 p.595-601

[9] Le Cam L., On the asymptotic theory of estimation and testing hypotheses, *Proc. Third Berkeley Symp. Math. Statist. Prob.* (1955) Vol 1 p.129-156.

[10] Bahadur R.R., Rates of convergence of estimates and test statistics, *Annals of Mathematical Statistics* (1967) Vol. 38, p.303-324

[11] Huber P.J., The behavior of Maximum Likelihood Estimates under Nonstandard Conditions, *Fifth Berkeley Symp. Math. Statist. Prob* p.221- 233.

[12] Haussler D., Decision Theoretic Generalizations of the PAC Model for Neural Net and Other Learning Applications, *Technical Report* September (1989) UCSC-CRL-91-02.

[13] Blumer A.,Ehrenfeucht A., Haussler D. and Warmuth M, Classifying Learnable Geometric Concepts with the Vapnik-Chervonenkis Dimension, (1986) ACM

[14] Blumer A., Ehrenfeucht A., Haussler D. and Warmuth M, Learnability and the Vapnik-Chervonenkis Dimension, *Journal of the ACM* (1989) Vol 36, No. 4., p.929-965.

[15] Psaltis D., Snapp R. R., Venkatesh S. S., On the finite sample performance of the nearest neighbor classifier, *Advances in Neural Information Processing System* (1991), Morgan Kauffmann, also to appear in *IEEE Transactions on Information Theory.*

[16] Vapnik V. N and Chervonenkis A. Ya., On the uniform convergence of relative frequencies of events to their probabilities. *Theoret. Probl. and Its Appl.* (1971) Vol. 16 , 2, p.264-280.

[17] Vapnik V. N and Chervonenkis A. Ya., Necessary and sufficient conditions for the uniform convergence of means to their expectations. *Theoret. Probl. and Its Appl.* (1981) Vol. 26, 3, p.532-553.

[18] Rissanen J., *Stochastic Complexity in Statistical Inquiry*, (1989) Series in Computer science–Vol. 15.

[19] Chung K. L., *A course in Probability Theory* (1974) Vol 21 in Probability and Mathematical Statistics.

[20] Feller W., *An Introduction to Probability Theory and Its Applications* (1959).Vol. 1, 2.

[21] Pollard D., *Convergence of Stochastic Processes*, Springer Series in Statistics, (1984).

[22] Rao E. R., *Linear Statistical Inference and Its Applications* (1973) 2^{nd} Ed. John Wiley & Sons, New York

[23] Wencour R.S. and Dudley R.M., Some special Vapnik-Chervonenkis classes, *Discrete Mathematics* (1981) Vol. 33, p.313-318.

[24] Kohonen T., The Self-Organizing Map, *Proceedings of the IEEE* (1990) Vol. 78, No. 9.

[25] Szego G. and Polya G., *Problems and theorems in Analysis* (1976) Volumes I & II, Springer Verlag.

[26] Szego G. *Orthogonal Polynomials* (1939) American Mathematical Society, N.Y.C.

[27] Rudin W., *Principles of Mathematical Analysis* (1976) 3^{rd} Ed. McGraw Hill, New York.

[28] Apostol T., *Mathematical Analysis* (1975) 2^{nd} Ed. Addison-Wesley Publishing, Menlo Park, CA.

[29] Teicher H., Identifiability of finite mixtures, *Annals of Mathematical Statistics* (1963) Vol. 34, p.1265-1269

[30] Yakowitz S.J. and Spragins J.D., On identifiability of finite mixtures, *Annals of Mathematical Statistics* (1968) Vol. 39, p.209-214

[31] Cover T. M., Thomas J., *Elements of Information Theory*(1991) John Wiley and Sons, New York.

[32] Royden H.L., *Real Analysis* (1968) 2^{nd} Ed., The Macmillan Company, London.

[33] Cover T. M. and Hart P. E., Nearest Neighbor Pattern Classification, *IEEE Transaction on Information Theory* (1967) Vol IT-13, No. 1

[34] Valiant L. G., A Theory of the learnable, *Comm. ACM* (1984) 27:11, p. 1134-1142.

[35] Cramér H., *Mathematical Methods of Statistics*, (1946) Princeton University Press, Princeton, N.J.

[36] MacQueen J., Some methods for classification and analysis of multivariate observations, *Proc. Fifth Berkeley Symposium on Math. Stat. and Prob., I*, p.281-297.

[37] Gersho A., On the structure of vector quantizers, *IEEE Transactions on Information Theory*, Vol. IT-28, No. 2, March 1982.

[38] Glick N., Sample-Based Classification Procedures Derived from Density Estimators, *Journal of the American Statistical Association*, March 1972 Vol. 67, Num. 337.

[39] Bickel P. J. and Doksum K. A. , *Mathematical Statistics: Basic Ideas and Selected Topics*, (1977) Holden-Day, Inc., San Francisco, CA.

[40] Silverman B.W. , *Density Estimation*, (1986) Chapman and Hall, N.Y.

[41] Silverman B.W., Weak and strong uniform consistency of the kernel estimate of a density and its derivatives, *Annals of Statistics*, Vol 6, p.177-184.

[42] Marcus M., Minc H., *A Survey of Matrix Theory and Matrix Inequalities*, (1964) Allyn and Bacon, Inc., Boston

[43] Papoulis A., *Probability, Random Variables, and Stochastic Processes*, (1984) McGraw-Hill, Inc.

[44] Haussler D., Kearns M., Schapire R., *Bounds on the Sample Complexity of Bayesian Learning Using Information Theory and the VC Dimension*, Technical Report, UCSC-CRL-91-44.

[45] Abu-Mostafa Y. S. *Learning from Hints in Neural Networks*, Journal of Complexity, 6, p.192-198 (1990).

[46] Cover T. M., Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition, *IEEE Trans. Electron. Comput.* Vol. EC-14, p. 326-334, 1965.

[47] Grenander U., *Abstract Inference*, (1981) Wiley, New York.

[48] Stute, W., A law of the logarithm for kernel density estimators, em Annals of Probability, Vol 10, p.86-107.

[49] Benedek G., Itai A., Learnability by fixed distributions, in *Computational Learning Theory*, Haussler D., Pitt L., ACM SIGACT/SIGART August 3 1988, Morgan Kaufmann Publishers Inc., San Mateo CA.

[50] Benedek G., Itai A., Dominating Distributions and Learnability, in *Computational Learning Theory*, ACM SIGACT/SIGART July 27 1992, Morgan Kaufmann Publishers Inc., San Mateo CA.

[51] Barlett P. L., Williamson R. C., Investigating the Distribution Assumptions in the PAC Learning Model *Computational Learning Theory*, Valiant L. G., Warmuth M. K., ACM SIGACT/SIGART August 5 1991, Morgan Kaufmann Publishers Inc., San Mateo CA.