Frugal Hypothesis Testing and Classification

by

Kush R. Varshney

B.S., Electrical and Computer Engineering,
Cornell University, 2004

S.M., Electrical Engineering and Computer Science,
Massachusetts Institute of Technology, 2006

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in Electrical Engineering and Computer Science
at the Massachusetts Institute of Technology

June 2010

Author: _____

Department of Electrical Engineering and Computer Science
February 11, 2010

Certified by: _____

Alan S. Willsky
Edwin Sibley Webster Professor of Electrical Engineering
Thesis Supervisor

Accepted by: _____

Terry P. Orlando
Chair, Department Committee on Graduate Students

Frugal Hypothesis Testing and Classification

by Kush R. Varshney

Submitted to the Department of Electrical Engineering and Computer Science
on February 11, 2010, in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

The design and analysis of decision rules using detection theory and statistical learning theory is important because decision making under uncertainty is pervasive. Three perspectives on limiting the complexity of decision rules are considered in this thesis: geometric regularization, dimensionality reduction, and quantization or clustering. Controlling complexity often reduces resource usage in decision making and improves generalization when learning decision rules from noisy samples. A new margin-based classifier with decision boundary surface area regularization and optimization via variational level set methods is developed. This novel classifier is termed the geometric level set (GLS) classifier. A method for joint dimensionality reduction and margin-based classification with optimization on the Stiefel manifold is developed. This dimensionality reduction approach is extended for information fusion in sensor networks. A new distortion is proposed for the quantization or clustering of prior probabilities appearing in the thresholds of likelihood ratio tests. This distortion is given the name mean Bayes risk error (MBRE). The quantization framework is extended to model human decision making and discrimination in segregated populations.

Thesis Supervisor: Alan S. Willsky
Title: Edwin Sibley Webster Professor of Electrical Engineering

Acknowledgments

The drinker leaves his home to find the House of Wine, but does not know
The way, and fears achievement must be but for an instructed few;
And each from whom he asks the way has something new and strange to say.
In fact, you reach the House of Wine by any path you may pursue.
Harbans Rai Bachchan

Ithaka gave you the marvelous journey.
Without her you would not have set out.
C. P. Cavafy

I would like to thank my thesis supervisor Alan Willsky for all of the guidance and support he has provided during my time at MIT. He has really shaped my outlook on the science of information and decision systems, and has been a constant source of inspiration.

I appreciate the efforts of the other members of my doctoral committee: John Fisher, Polina Golland, and Josh Tenenbaum. John, in particular, has always asked the apropos question. I express my gratitude to Müjdat Çetin, Justin Dauwels and Sujay Sanghavi for listening to my research ideas and providing their thoughts, to Nikos Paragios for helping me get going with level set methods, and to Barry Chen and Ryan Prenger for showing me different perspectives on pattern recognition and statistical learning. I appreciate the assistance of Praneetha Mukhatira and Rachel Cohen.

I am grateful to Anima Anandkumar, Biz Bose, Jason Chang, Lei Chen, Michael Chen, Ayres Fan, Emily Fox, Matt Johnson, Junmo Kim, Pat Kreidl, Dahua Lin, Andrew Mastin, Mike Siracusa, Walter Sun, Vincent Tan, and Jason Williams for having things to say in grouplet when I didn't. Thanks also to Venkat Chandrasekaran, Jin Choi, Alex Ihler, Jason Johnson, Ying Liu, Dmitry Malioutov, James Saunderson, and Erik Sudderth for making the group everything that it has been. I would like to thank Jason, Pat, Matt, Anima, James, and especially Emily for being such great officemates.

My inquiry into frugal hypothesis testing and classification began as a result of a conversation with Lav Varshney, and some of the results in Chapter 5 are joint work with him. Besides this intellectual impetus, I thank Lav for all of the other impetuses he has provided.

I thank my family for their encouragement. Above all, I would like to thank my parents for everything.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

DECISION making under uncertainty is a task faced in a variety of domains. One specific decision-making task is to determine which of two or more alternatives known as *hypotheses* or *classes* is exhibited in measurements. Basketball referees decide whether a player is committing a foul [159], geologists and geophysicists decide whether a layer of the earth's crust is composed of sandstone or shale, communication system receivers decide what symbol was transmitted, art connoisseurs decide who painted a painting [100], missile defense systems decide whether a target is a warhead, and physicians decide whether a patient has breast cancer [127]. In all of these cases and many other such examples, the decision is based on imperfect measurements, but is also based on a utility or loss function, and on prior information or models of the world.

The typical formulation for modeling and designing optimal decision rules appeals to the maximization of expected utility [20, 67, 136, 162, 174], equivalently the minimization of expected loss. The simplest loss function is the error rate. A unit cost is accrued if the decision is for one hypothesis, but a different hypothesis is true. In practice, loss functions may take all sorts of other factors into account. The form of decision rules is typically a partition of the measurement domain using decision boundaries. According to Weirich [215], maximization of expected utility is the guiding principle for decision making even when precise probabilities or utilities are unavailable.

Proponents of frugal engineering and constraint-based innovation favor the view that limitless material resources are not usually available, and even if they are, they might actually be a hindrance; frugalness can often be a blessing [80]. Reporting on cognition research, Gladwell [84] writes, "in good decision making, frugality matters. ... [E]ven the most complicated of relationships and problems ... have an identifiable underlying pattern. ... [I]n picking up these sorts of patterns, less is more. Overloading the decision makers with information ... makes picking up that signature harder, not easier."

In general, decision-making systems are limited by a scarcity of resources, whether manifested as finite training data, time constraints, processing constraints, or memory constraints. With finite training data, the structural risk minimization principle from statistical learning theory formalizes the idea that the complexity of decision rules should be controlled to find an optimal complexity level that neither overfits nor underfits [206]. Sensor networks are often used for decision making and are severely

power-limited which means that they are also limited in computation and especially communication [36]. Human decision makers have information processing limits as well, known in the economics literature as bounded rationality [46, 168].

The topic covered in the thesis is frugal decision making, specifically the decision-making tasks of hypothesis testing and classification. Three modes of frugality in decision making are studied: frugality with the surface area of decision boundaries, frugality with the dimensionality input into decision rules, and frugality with the precision of prior probabilities used in constructing decision rules. All three modes can be viewed as limits on the decision boundary. Frugalness with surface area is directly expressed in terms of the decision boundary. Frugalness with dimensionality amounts to limiting decision boundaries to be cylinders or generalized cylinders in the domain of measurements. Frugalness with prior probability precision limits the set of possible decision boundaries to a discrete set.

Limiting the surface area of decision boundaries is applied to control complexity and thereby improve performance in the context of the supervised learning of classifiers. This leads to a novel margin-based classifier whose regularization term is the surface area of the decision boundary; the new regularization term implies a new inductive bias for learning [130]. The proposed classifier is given the name *geometric level set classifier*. It is competitive in performance to classifiers from the literature and gives the best performance on certain datasets. The training of this classifier is approached using variational level set methods [141].

Mapping measurements to lower-dimensional features is generally termed *dimensionality reduction*. The mapping may be either linear or nonlinear. Many approaches to dimensionality reduction, such as principal component analysis [148, 177], do not contain decision making in their formulation. In this thesis, dimensionality reduction mappings are chosen specifically for supervised margin-based classification. Dimensionality reduction is a way to improve classification performance, and also to conserve resource usage when taken to sensor network settings. The approach to learning both a classifier and a dimensionality reduction mapping jointly includes optimization on the Stiefel manifold [60, 188].

The prior probability factors into the optimal decision rule of Bayesian hypothesis testing [221]. Quantizing or clustering this prior probability diminishes its precision. Quantization and clustering are not usually examined in problems of decision making. The specific problem considered in the thesis leads to a new distortion criterion, the *mean Bayes risk error*. Quantization or clustering the prior probability reduces resource usage, and in certain setups, clustering improves decision-making performance. This mode of frugality provides a model for human decision making, and an extension predicts social discrimination in segregated human populations.

Examples throughout the thesis illustrate the 'less is more' paradigm of decision making when decision rules are determined on the basis of a finite set of noisy data samples. With complexity measured through a decision boundary surface area constraint, through reduced-dimensionality, or through the number of clusters, decision-

making quality first improves with an increase in complexity but then gets worse with too much complexity, implying that throwing resources at the problem is not always advantageous.

1.1 Overview of Thesis Contributions and Methods

This section walks through the remaining chapters of the thesis, previewing the contributions and methods of the thesis as well as giving a glimpse of its organization.

1.1.1 Chapter 2: Background

Preliminary material related to five mathematical topics is presented in Chapter 2. These five topics are:

- detection theory with a focus on Bayesian hypothesis testing,
- statistical learning theory with a focus on supervised margin-based classification,
- variational level set methods,
- linear and nonlinear dimensionality reduction, and
- quantization theory and k-means clustering.

The first two of these set the stage of decision making under uncertainty, the problem considered in the thesis. The remaining three topics enter into the thesis to support the three modes of frugality mentioned. Variational level set methods are used to learn classifiers with limited surface area. Dimensionality reduction is used to limit the dimensionality of the input space for classifiers. Quantization and clustering are used to limit the precision of prior probabilities.

The detection theory section states the hypothesis testing problem with two classes, formulates the Bayesian objective to hypothesis testing known as the Bayes risk, and derives the optimal decision rule that minimizes Bayes risk—the likelihood ratio test. The section also derives properties of the complementary receiver operating characteristic and discusses the Bayesian hypothesis testing problem with more than two classes.

The statistical learning theory section states the supervised classification problem and discusses the structural risk minimization principle, including generalization bounds based on Vapnik–Chervonenkis (VC) dimension and Rademacher complexity [13, 99, 206]. It also details a particular formulation for supervised classification known as margin-based classification which includes the popular support vector machine (SVM) [99, 176].

The section on variational level set methods describes the types of partitioning problems that can be solved using those methods: energy minimization problems with region-based and boundary-based terms [141]. It also discusses the gradient descent optimization framework known as curve evolution or contour evolution, and implementation using level set functions [143].

The dimensionality reduction section gives a general formulation for linear dimensionality reduction as a constrained optimization problem on the Stiefel manifold of matrices and discusses gradient descent-based optimization that respects the Stiefel manifold constraint [60, 188]. Zonotopes, types of polytopes that are intimately related to linear dimensionality reduction matrices, and their properties are discussed [37, 65]. Nonlinear dimensionality reduction, especially the Isomap method and its formulation through a data-dependent kernel function, is also described [19, 193].

The final section of the chapter focuses on quantization theory [79]. The quantization problem is stated first; k-means clustering is a version of quantization based on samples from a distribution rather than on the distribution itself. Optimality conditions for a minimum distortion quantizer are given along with the Lloyd–Max and k-means algorithms, which are used to find minimum distortion quantizers and clusterings respectively [118, 125]. Finally, it is discussed that the distortion in quantization is monotonically decreasing in the number of quantization levels.

■ 1.1.2 Chapter 3: Surface Area of the Decision Boundary

The primary contribution of Chapter 3 is the introduction of the geometric level set (GLS) classifier to statistical learning. Geometric approaches and analysis are closely tied to statistical learning [92, 111, 185]. The GLS classifier takes geometric partial differential equation ideas [142, 173] and applies them to the supervised classification problem. A key feature of the GLS classifier is regularization through the penalization of decision boundaries with large surface area.

The first section of the chapter formulates the optimization problem of the GLS classifier with two classes. The objective to be minimized is composed of an empirical risk term containing a margin-based loss function applied to the training data and a term that is the surface area of the decision boundary weighted by a regularization parameter. The section also details how this geometrically-regularized margin-based loss objective can be minimized using the Euler–Lagrange descent approach known as curve evolution or contour evolution, implemented using level set methods [141]. Examples of classifier learning are presented on illustrative datasets.

The second section extends the GLS classifier from binary classification to multicategory classification. Margin-based classifiers are typically extended for multicategory classification using the one-against-all construction [163]. The proposal in this section is a new alternative to one-against-all that requires many fewer decision functions: logarithmic rather than linear in the number of classes. It is based on a binary encoding scheme enabled by the representation of decision boundaries with signed distance functions, a type of level set function [214]. An illustrative example is presented in this section as well.

Level set methods are nearly always implemented with a pixel or voxel grid-based representation in their incarnations as methods of solution for problems in physics and image processing. This implementation strategy is only tractable in domains with dimensionality up to three or four. The third section describes a level set implementation

amenable to higher-dimensional spaces, which are routinely encountered in decision making, through radial basis functions [78, 216]. The decision-making performance of the GLS classifier with radial basis function implementation is shown as a function of the weight given to the surface area regularization term on benchmark datasets, illustrating the structural risk minimization principle. The performance of the GLS classifier is also compared to several popular classifiers on several benchmark datasets. The GLS classifier is competitive with other classifiers and performs the best on one dataset.

Statistical learning theory analysis of the new GLS classifier is presented in the fourth section. Specifically, the VC dimension is measured empirically as a function of the weight on the surface area regularization term [205]. Additionally, the Rademacher complexity is characterized analytically [13, 121]. The GLS classifier is also shown to be consistent [116]. The Rademacher complexity analysis and the consistency analysis rely on finding the ϵ-entropy [106] of GLS classifiers, which is done in the section.

Variational level set methods are known to be fairly flexible in the types of objectives that can be included. The inductive bias of the GLS classifier, caused by the preference for small decision boundary surface area, is included in a straightforward manner. Another geometric inductive bias that may be included for classification is proposed in the fifth section of the chapter. An additional regularization term that promotes feature subset selection is proposed. This term is small when the decision boundary is an axis-aligned cylinder in a few dimensions of the input measurement space. This additional regularization term is just one example of geometric preferences that can be encoded in the variational level set framework.

■ 1.1.3 Chapter 4: Dimensionality of the Classifier Input

The main contribution of Chapter 4 is the development of a method for joint dimensionality reduction and margin-based classification. Dimensionality reduction of data may be performed without having an eye toward what the final use of the reduced-dimensional data is. However, if it is known beforehand that dimensionality reduction is to be followed by classification, then the reduced-dimensional space should be optimized for classification. The specific classifiers considered in the chapter are non-parametric margin-based classifiers that do not rely on strong assumptions about the data likelihood functions, and include the GLS classifier proposed in Chapter 3 and the SVM. The formulation is extended to sensor networks, a setting in which dimensionality reduction has the dual advantages of improving classification performance and reducing the amount of communication.

Decision functions of margin-based classifiers are typically defined in the input measurement space of full dimensionality. The first section in this chapter formulates a classifier in which the decision function is defined in a reduced-dimensional linear subspace of the input space. The classifier decision function and the linear subspace are learned jointly based on the training data in order to minimize a margin-based loss objective. The linear subspace is specified through a matrix on the Stiefel manifold, the set of

matrices that have orthonormal columns [188]. Therefore, the training objective is a functional of both the decision function and the dimensionality reduction matrix. The learning is approached through coordinate descent minimization: alternating minimizations for the matrix with the decision function fixed, and for the decision function with the matrix fixed. Some illustrative examples are shown with both the GLS classifier and the SVM. Classification performance as a function of the reduced dimensionality is given for several benchmark datasets with the SVM. The best classification performance for a ten thousand-dimensional dataset is achieved by reducing to twenty dimensions because dimensionality reduction helps prevent overfitting.

The second section considers nonlinear dimensionality reduction rather than linear dimensionality reduction. The interest is in finding low-dimensional nonlinear manifolds rather than linear subspaces on which the classifier is defined. A formulation is proposed that represents the nonlinear manifold using a data-dependent kernel function arising from Isomap, a manifold learning technique that does not have supervised classification as its objective [19, 193]. With the kernel function representation, optimization of the manifold for margin-based classification amounts to finding a matrix on the Stiefel manifold as in the linear case. Therefore, the coordinate descent for joint linear dimensionality reduction and margin-based classification also applies to nonlinear dimensionality reduction. An illustrative example is shown in which nonlinear dimensionality reduction is superior to linear dimensionality reduction. A comparison of classification performance as a function of the reduced dimension between linear and nonlinear dimensionality reduction is also given for several benchmark datasets.

Rademacher complexity and consistency are studied in the third section of the chapter [13, 116]. As in Chapter 3, the analysis builds upon the ϵ-entropy of the set of classifiers [106], but here with the additional factor of dimensionality reduction. This additional factor is accounted for through an analysis of zonotopes, which are polytopes that are convex, centrally-symmetric, and whose faces are also centrally-symmetric in all lower dimensions [37, 65]. The mapping of a high-dimensional hypercube to a reduced-dimensional space by a matrix on the Stiefel manifold is a zonotope.

The fourth section concentrates on the application of sensor networks. Prior work on dimensionality reduction for classification has not studied sensor network applications [117, 150, 151, 202]. Also, prior work on sensor networks has not focused on supervised classification, particularly in combination with dimensionality reduction [38, 194, 203, 213]. The joint dimensionality reduction and classification formulation has a clear interpretation in the sensor network setting. Sensor nodes take measurements and perform dimensionality reduction, whereas a fusion center makes the decision. Communication emanating from sensors is of the reduced dimension rather than the measured dimension, resulting in savings of resources. With a single sensor and fusion center, the coordinate descent minimization procedure developed without the sensor network application in mind is unchanged and training can be performed with communication related to the reduced dimension rather than the full dimensionality of the measurements. An extension is described for tree-structured multisensor fusion net-

works in which the training involves the passing of simple messages between nodes, with the amount of communication related to the reduced dimension rather than the input dimension. A simple physical model of sensor networks is studied and classification performance as a function of power consumption due to communication is examined. The structural risk minimization principle appears in this setting as well, again indicating that more resource usage does not necessarily mean better performance.

■ 1.1.4 Chapter 5: Precision of the Prior Probabilities

The main contribution of Chapter 5 is the proposition of minimum mean Bayes risk error (MBRE) quantization/clustering of prior probabilities for hypothesis testing. The decision rule that minimizes Bayes risk is the likelihood ratio test with a threshold incorporating the costs of different types of errors and incorporating the prior probabilities of the hypotheses [221]. Over a population of many objects with different prior probabilities, keeping track of the prior probabilities of each individual object consumes resources. Moreover, only an estimate of these prior probabilities may be observable. Quantization or clustering the prior probabilities over the population serves to reduce resource usage. When only an estimate of the priors is available, clustering also serves to improve decision-making performance. Studies in psychology suggest that humans categorize objects in populations [129]; quantization and clustering are one way to model categorization.

The first section states the problem of quantizing the prior probabilities appearing in the threshold of the likelihood ratio test when these prior probabilities come from a probability density function describing a population. A new distortion function for quantization is proposed, given the name Bayes risk error. This distortion function measures the difference in Bayes risk between a decision rule utilizing a quantized prior probability in the threshold and one utilizing the true prior probability in the threshold. Convexity and quasiconvexity properties, among others, are derived for the new distortion function. Conditions satisfied by locally minimum MBRE quantizers are derived and shown to be sufficient conditions [200]. Examples of MBRE-optimal quantizers are shown under different population distributions and Bayes costs, with Gaussian likelihood functions. Minimum MBRE quantizers are also compared to minimum mean absolute error quantizers in the quantization of prior probabilities for hypothesis testing.

High-rate or high-resolution analysis asymptotically characterizes quantization when the number of quantization cells is large [89]. The second section of the chapter looks at the high-rate quantization regime for minimum MBRE quantizers. The Bayes risk error, unlike absolute error and squared error for example, may not be expressed as a function of the difference between the quantized and unquantized values. Many distortion functions that model human perception are nondifference distortion functions as well. The high-resolution quantization analysis in the section is based on a locally quadratic approximation that draws from analysis used with perceptual distortions [113]. The examples from the first section are analyzed in the high-rate regime.

In contrast to having direct access to the population distribution, the third section looks at learning about the population from a finite number of noisy measurements per object for a finite number of objects. Imperfect observations of the hypothesis are used to estimate the prior probabilities of an object. These estimates across the population are then clustered using k-means with the Bayes risk error distortion. The cluster center assigned to an object is used in the likelihood ratio test threshold. This empirical, data-based model shows elements of overfitting and the structural risk minimization. Beyond a certain number of clusters, decision-making performance gets worse. This is in contrast to quantization with known population distribution, where increasing the number of quantization levels always improves decision-making performance. Additionally, the more observations per object, the better the decision-making performance and the greater the optimal number of clusters.

Decision making by humans on populations of humans is studied in the fourth section. Arrests by police officers, verdicts by jurors, and foul calls by referees are examples of the type of situation under consideration. Human decision makers have information processing limits and tend to categorize [129], a phenomenon captured by quantization or clustering. Regardless of resource or processing limitations, clustering is the right thing to do when noisy estimates of the prior probabilities of population members are available to the decision maker. Humans tend not only to categorize members of a population into groups related to how likely they are to commit a crime or foul, but also to categorize based on social indicators such as race [122]. Separate categorization based on race is modeled in this section by separate quantizers for different racial populations. The model predicts higher Bayes risk when the decision maker and population member are of different races than when they are of the same race. This is due to more quantization levels being allocated to the same race population because of differences in the amount of inter-race and intra-race societal interaction [59]. High Bayes risk can result from either high missed detection probability, high false alarm probability, or both. A high false alarm probability on population members of a different race than the decision maker occurs when the cost of a missed detection is much higher than the cost of a false alarm. Many econometric studies indicate that human decision makers do have a higher false alarm probability on members of a different race than members of the same race, rather than a significantly higher missed detection probability [8, 53, 159, 190]. The quantization model proposed in this section explains these studies if the decision maker is precautionary, i.e. the cost of a missed detection is higher than the cost of a false alarm for the decision maker.

■ 1.1.5 Chapter 6: Conclusion

The final chapter provides a summary of the thesis contributions, in particular that the three discussed notions of frugalness in decision making: surface area regularization, dimensionality reduction, and quantization/clustering share the commonality that some complexity reduction improves generalization when learning from samples and decreases the usage of resources.

Several directions of future research are laid out in the chapter as well. One such direction is model selection, including determining when the GLS classifier should be used. Another direction is extension to learning scenarios such as online learning, semisupervised learning, and Neyman–Pearson learning. Also discussed are further possible developments within the MBRE quantization and clustering framework, the nonlinear dimensionality reduction framework, and in the sensor network application. It is suggested that Markov chain Monte Carlo sampling be used to characterize the posterior probability of GLS classifier decision rules and get a sense of classifier uncertainty. Some ideas regarding broader application of methods based on variational and geometric partial differential equations are presented.

■ 1.2 Notes

Two-dimensional grid-based examples shown in Chapter 3 and Chapter 4 are implemented using the Level Set Methods Toolbox for Matlab by Barış Sümengen, available at `http://barissumengen.com/level_set_methods`. Mutual information estimates used to initialize dimensionality reduction matrices in Chapter 4 are calculated using the Kernel Density Estimation Toolbox for Matlab by Alex Ihler and Mike Mandel, available at `http://www.ics.uci.edu/~ihler/code`.

Portions of the material in this thesis have been previously presented at the 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing [207], the 2008 IEEE Workshop on Machine Learning for Signal Processing [209], and the 2009 International Conference on Information Fusion [210]. Additionally, portions of the material appear in or have been submitted to the IEEE Transactions on Signal Processing [208, 212] and the Journal of Machine Learning Research [211].

This work was supported in part by a National Science Foundation Graduate Research Fellowship, by Shell International Exploration and Production, Inc., by a MURI funded through AFOSR Grant FA9550-06-1-0324, and by a MURI funded through ARO Grant W911NF-06-1-0076.

Chapter 2

Background

THIS background chapter describes several topics in applied mathematics, probability, and statistics that form the theoretical foundation for the thesis. Five topics are covered: detection theory, statistical learning theory, variational level set methods, dimensionality reduction, and quantization theory. The theories of detection [221], statistical learning [206], and quantization [79] grew out of the fields of electrical engineering and computer science, but are fairly abstract. Variational level set methods were first developed to solve differential equations for applications in physics [143], and were further developed when transferred to applications in image processing and computer vision [141, 142]. Dimensionality reduction has been a part of data analysis and statistics since their beginnings [69, 95, 96, 148], but is now becoming essential in a world awash in data and possible in a world with abundant computational resources. Entire courses may be devoted to each of these five topics individually; the treatment in this chapter is limited in scope to that required for the remainder of the thesis.

Typeface is used throughout the thesis to distinguish random variables from samples, and also to distinguish scalars, vectors, matrices, and sets. Random variables have sans-serif typeface whereas samples have serif typeface. Vectors and matrices have bold typeface whereas scalars do not. Vectors are in lowercase and matrices are in uppercase. Sets are indicated through calligraphic typeface.

■ 2.1 Detection Theory

Detection theory—the theory of distinguishing classes of objects or states of the world based on noisy measurements—forms the central thread linking the entire thesis together. The focus of the thesis is on constraints introduced into decision making (for good or for bad); the background material on detection gives the limits on the best possible performance in decision making if there are no constraints, as well as the structure of the decision rule that achieves this best possible performance.

The optimal Bayes risk defined in Section 2.1.2 is used to define classifier consistency later. Decision rules proposed in the thesis are shown to be consistent in Section 3.4 and Section 4.3. The optimal Bayes risk is also used to define a novel distortion function for quantization in Chapter 5. The likelihood ratio test, the optimal decision rule, and its associated notion of a sufficient statistic described in Section 2.1.3 motivates the

development of dimensionality reduction techniques in Chapter 4. The structure of the likelihood ratio test is retained, but its inputs are constrained in the decision rules of Chapter 5. The complementary receiver operating characteristic and its properties discussed in Section 2.1.4 are used to develop properties of the distortion function for quantization in Chapter 5. Detection with more than two classes, mentioned in Section 2.1.5, forms a baseline for comparison to a novel multicategory classification method proposed in Section 3.2.

■ 2.1.1 Binary Hypothesis Testing Problem Statement

Consider an object that is in one of two states, the hypotheses $y = -1$ and $y = +1$, having prior probabilities $p_- = \Pr[y = -1]$ and $p_+ = \Pr[y = +1] = 1 - p_-$. In the hypothesis testing problem, the task is to determine the state of the object using an imperfect observation $\mathbf{x} \in \Omega \subset \mathbb{R}^D$, which is also known as the measurement or input data. The measurements and the hypotheses are related by the likelihood functions $f_{\mathbf{x}|y}(\mathbf{x}|y = -1)$ and $f_{\mathbf{x}|y}(\mathbf{x}|y = +1)$. The prior probabilities and the likelihood functions together specify the joint probability density function of the measurements and hypotheses $f_{\mathbf{x},y}(\mathbf{x}, y)$.

A function $\hat{y}(\mathbf{x})$, known as the decision rule, is designed to determine the hypothesis from the measurement.[1] It uniquely maps every possible $\mathbf{x}$ to either -1 or $+1$. Equivalently, $\hat{y}$ partitions the measurement space Ω into two regions: one corresponding to $\hat{y} = -1$ and the other corresponding to $\hat{y} = +1$.

There are two types of errors, with the following probabilities:

$$p_F = \Pr[\hat{y}(\mathbf{x}) = +1|y = -1],$$
$$p_M = \Pr[\hat{y}(\mathbf{x}) = -1|y = +1],$$

where p_F is the probability of false alarm and p_M the probability of missed detection. The complement of the probability of missed detection is the probability of detection $p_D = 1 - p_M$.

■ 2.1.2 Bayes Risk

In the Bayesian formulation to the hypothesis testing problem, the decision rule $\hat{y}$ is chosen to minimize the Bayes risk $\mathrm{R} = \mathrm{E}[c(y = i, y' = j)]$, an expectation over a nonnegative cost function $c(i, j)$. This gives the following specification for the Bayes optimal decision rule $\hat{y}^*(\mathbf{x})$:

$$\hat{y}^*(\cdot) = \arg\min_{f(\cdot)} \mathrm{E}[c(y, f(\mathbf{x}))], \tag{2.1}$$

where the expectation is over both y and $\mathbf{x}$. The optimal decision rule has Bayes risk $\mathrm{R}(\hat{y}^*)$. The cost function values are denoted as follows: $c_{--} = c(-1, -1)$ (the cost of a

[1] Note that in this thesis, only deterministic, nonrandomized decision rules are considered.

true alarm), $c_{-+} = c(-1,+1)$ (the cost of a false alarm), $c_{+-} = c(+1,-1)$ (the cost of a missed detection), and $c_{++} = c(+1,+1)$ (the cost of a detection).

The Bayes risk, which is the decision-making performance, may be expressed in terms of the error probabilities as:

$$\mathrm{R} = [c_{+-} - c_{--}]p_- p_F + [c_{-+} - c_{++}]p_+ p_M + c_{--}p_- + c_{++}p_+. \tag{2.2}$$

Often, no cost is assigned to correct decisions, i.e. $c_{--} = c_{++} = 0$, which is assumed in the remainder of this thesis. In this case, the Bayes risk simplifies to:

$$\mathrm{R}(p_-) = c_{+-}p_- p_F(p_-) + c_{-+}[1 - p_-]p_M(p_-). \tag{2.3}$$

In (2.3), the dependence of the Bayes risk R and error probabilities p_F and p_M on p_- has been explicitly noted. The error probabilities depend on p_- through the decision rule $\hat{y}^*$. The function $\mathrm{R}(p_-)$ is zero at the points $p_- = 0$ and $p_- = 1$ and is positive-valued, strictly concave, and continuous in the interval $(0,1)$ [49, 218, 221].

When measuring performance by probability of decision error, p_E, both types of errors take equal weight and $c_{-+} = c_{+-} = 1$; the Bayes risk further simplifies to:

$$\mathrm{R} = p_E = p_- p_F + p_+ p_M. \tag{2.4}$$

■ 2.1.3 Likelihood Ratio Test

The decision rule that minimizes the Bayes risk, the solution to (2.1), is now derived following the derivation of [221].

Consider an arbitrary, fixed decision rule $\hat{y}$ and its Bayes risk $\mathrm{R} = \mathrm{E}[c(y, \hat{y}(\mathbf{x}))]$. Using iterated expectation, this is equivalent to

$$\mathrm{R} = \mathrm{E}\left[\mathrm{E}[c(y, \hat{y}(\mathbf{x}))|\mathbf{x} = \mathbf{x}]\right] \tag{2.5}$$

$$= \int_\Omega \check{\mathrm{R}}(\hat{y}(\mathbf{x}), \mathbf{x}) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}, \tag{2.6}$$

where $\check{\mathrm{R}}(y, \mathbf{x}) = \mathrm{E}[c(y, y)|\mathbf{x} = \mathbf{x}]$. Because $f_{\mathbf{x}}(\mathbf{x})$ is nonnegative, the minimum of (2.5) occurs when $\check{\mathrm{R}}(\hat{y}(\mathbf{x}), \mathbf{x})$ is minimum at each point $\mathbf{x} \in \Omega$.

Looking at a particular point $\mathbf{x} = \check{\mathbf{x}}$, if the decision rule is such that $\hat{y}(\check{\mathbf{x}}) = -1$, then

$$\check{\mathrm{R}}(-1, \check{\mathbf{x}}) = c_{-+} \Pr[y = +1|\mathbf{x} = \check{\mathbf{x}}]. \tag{2.7}$$

If the decision rule is such that $\hat{y}(\check{\mathbf{x}}) = +1$, then

$$\check{\mathrm{R}}(+1, \check{\mathbf{x}}) = c_{+-} \Pr[y = -1|\mathbf{x} = \check{\mathbf{x}}]. \tag{2.8}$$

Between the two possible decision rules at $\check{\mathbf{x}}$, the one with smaller $\check{\mathrm{R}}$ is optimal. The overall optimal decision rule is then a comparison between (2.7) and (2.8) for all $\mathbf{x} \in \Omega$, which can be written:

$$c_{-+} \Pr[y = +1|\mathbf{x} = \mathbf{x}] \underset{\hat{y}^*(\mathbf{x})=-1}{\overset{\hat{y}^*(\mathbf{x})=+1}{\gtrless}} c_{+-} \Pr[y = -1|\mathbf{x} = \mathbf{x}]. \tag{2.9}$$

The decision rule (2.9) may be rearranged as

$$\frac{\Pr[y=+1|\mathbf{x}=\mathbf{x}]}{\Pr[y=-1|\mathbf{x}=\mathbf{x}]} \underset{\hat{y}^*(\mathbf{x})=-1}{\overset{\hat{y}^*(\mathbf{x})=+1}{\lessgtr}} \frac{c_{+-}}{c_{-+}}. \tag{2.10}$$

Applying the Bayes theorem yields the likelihood ratio test

$$\frac{f_{\mathbf{x}|y}(\mathbf{x}|y=+1)}{f_{\mathbf{x}|y}(\mathbf{x}|y=-1)} \underset{\hat{y}^*(\mathbf{x})=-1}{\overset{\hat{y}^*(\mathbf{x})=+1}{\lessgtr}} \frac{p_- c_{+-}}{(1-p_-)c_{-+}}. \tag{2.11}$$

The function on the left side of (2.11) is known as the likelihood ratio:

$$\Lambda(\mathbf{x}) = \frac{f_{\mathbf{x}|y}(\mathbf{x}|y=+1)}{f_{\mathbf{x}|y}(\mathbf{x}|y=-1)}.$$

The right side of (2.11) is the threshold:

$$\eta = \frac{p_- c_{+-}}{(1-p_-)c_{-+}}.$$

The likelihood ratio test may also be expressed as follows:

$$\hat{y}^*(\mathbf{x}) = \text{sign}(\Lambda(\mathbf{x}) - \eta). \tag{2.12}$$

This form reveals that there is a decision function $\varphi(\mathbf{x}) = \Lambda(\mathbf{x}) - \eta$ whose zero level set is a decision boundary. Whenever $\varphi(\mathbf{x})$ is below zero, the decision is -1, and whenever $\varphi(\mathbf{x})$ is above zero, the decision is $+1$. The likelihood ratio function is a mapping from D dimensions to one dimension. It is a scalar sufficient statistic for detection. Regardless of D, the dimension of the observations, applying the likelihood ratio results in dimensionality reduction that is lossless, i.e. $\text{R}(\hat{y}^*(\mathbf{x})) = \text{R}(\hat{y}^*(\Lambda(\mathbf{x})))$ where $\hat{y}^*$ takes an appropriate-dimensional argument.

■ 2.1.4 Complementary Receiver Operating Characteristic

As seen in (2.11), the optimal decision rule, and consequently the two types of error probabilities, depends on the ratio of the costs c_{-+} and c_{+-}. Different ratios of costs correspond to different values of the threshold $\eta \in [0, \infty)$. The threshold parameterizes different operating points of the decision rule with different error probability pairs (p_F, p_M). The curve traced out on the p_F–p_M plane as η is varied from zero to infinity is the complement of what is known as the receiver operating characteristic, which is the curve traced out on the p_F–p_D plane. This complementary receiver operating characteristic (CROC) takes values $(p_F = 0, p_M = 1)$ when $\eta \to \infty$ and $(p_F = 1, p_M = 0)$ when $\eta \to 0$. Several important properties of the CROC arising from the likelihood ratio test are derived, again following the derivations of [221].

The expression for the derivative $\frac{dp_M(\eta)}{d\eta}$ is now obtained. First, define the regions $\mathcal{R}_+$ and $\mathcal{R}_-$, which are subsets of Ω, as follows:

$$\mathcal{R}_-(\eta) = \{\mathbf{x}|\Lambda(\mathbf{x}) - \eta < 0\},$$
$$\mathcal{R}_+(\eta) = \{\mathbf{x}|\Lambda(\mathbf{x}) - \eta > 0\}.$$

As an integral over the region $\mathcal{R}_-$, the probability of missed detection is:

$$p_M(\eta) = \int_{\mathcal{R}_-(\eta)} f_{\mathbf{x}|y}(\mathbf{x}|y=+1)d\mathbf{x} \tag{2.13}$$
$$= \int_{\mathcal{R}_-(\eta)} \Lambda(\mathbf{x}) f_{\mathbf{x}|y}(\mathbf{x}|y=-1)d\mathbf{x} \tag{2.14}$$
$$= \int_{\Omega} \text{step}(-\Lambda(\mathbf{x}) + \eta)\Lambda(\mathbf{x}) f_{\mathbf{x}|y}(\mathbf{x}|y=-1)d\mathbf{x} \tag{2.15}$$
$$= \text{E}\left[\text{step}(-\Lambda(\mathbf{x}) + \eta)\Lambda(\mathbf{x})|y=-1\right], \tag{2.16}$$

where (2.14) is obtained using the definition of the likelihood ratio, (2.15) is obtained by introducing a unit step function allowing the integration to be over the entire space, and (2.16) is obtained via the definition of expectation. Treating the likelihood ratio as a random variable,

$$p_M(\eta) = \text{E}\left[\text{step}(-\Lambda + \eta)\Lambda|y=-1\right] \tag{2.17}$$
$$= -\int_{\eta}^{\infty} \Lambda f_{\Lambda|y}(\Lambda|y=-1)d\Lambda. \tag{2.18}$$

Then, taking the derivative of (2.18) using Leibniz' rule, the derivative of the probability of missed detection with respect to the threshold is:

$$\frac{dp_M}{d\eta} = \eta f_{\Lambda|y}(\eta|y=-1). \tag{2.19}$$

Note that $\frac{dp_M}{d\eta}$ is nonnegative since η and the conditional density are both nonnegative. Thus $p_M(\eta)$ is a nondecreasing function.

A similar procedure may be followed to obtain the derivative $\frac{dp_F}{d\eta}$.

$$p_F(\eta) = \int_{\mathcal{R}_+(\eta)} f_{\mathbf{x}|y}(\mathbf{x}|y=-1)d\mathbf{x} \tag{2.20}$$
$$= \int_{\Omega} \text{step}(\Lambda(\mathbf{x}) - \eta) f_{\mathbf{x}|y}(\mathbf{x}|y=-1)d\mathbf{x} \tag{2.21}$$
$$= \text{E}\left[\text{step}(\Lambda(\mathbf{x}) - \eta)|y=-1\right]. \tag{2.22}$$

Again taking the likelihood ratio to be a random variable,

$$p_F(\eta) = \text{E}\left[\text{step}(\Lambda - \eta)|y=-1\right] \tag{2.23}$$
$$= \int_{\eta}^{\infty} f_{\Lambda|y}(\Lambda|y=-1)d\Lambda, \tag{2.24}$$

and differentiating yields

$$\frac{dp_F}{d\eta} = -f_{\Lambda|y}(\eta|y=-1). \tag{2.25}$$

Here note that $\frac{dp_M}{d\eta}$ is nonpositive and that $p_F(\eta)$ is nonincreasing.

Based on the derivatives of the two error probabilities with respect to the threshold, the derivative of the CROC is simply

$$\frac{dp_M}{dp_F} = \frac{dp_M}{d\eta}\frac{d\eta}{dp_F} = \frac{\eta f_{\Lambda|y}(\eta|y=-1)}{-f_{\Lambda|y}(\eta|y=-1)} = -\eta. \tag{2.26}$$

Since η is always nonnegative, the derivative of the CROC is always nonpositive and therefore the CROC is a nonincreasing function. It may additionally be shown through an argument relying on randomized decision rules that the CROC is a convex function [221]. In summary, the key properties discussed in this section are that the CROC is a convex, nonincreasing function from $(p_F = 0, p_M = 1)$ to $(p_F = 1, p_M = 0)$; that $p_M(\eta)$ is nondecreasing; and that $p_F(\eta)$ is nonincreasing.

■ 2.1.5 M-ary Hypothesis Testing

The discussion thus far has focused on the case in which there are two hypotheses. This section considers hypothesis testing problems with $M > 2$ hypotheses and $y \in \{1, 2, \ldots, M\}$. There are M prior probabilities $p_i = \Pr[y = i]$ and likelihood functions $f_{\mathbf{x},y}(\mathbf{x}|y = i)$ for $i = 1, 2, \ldots, M$. The decision rule $\hat{y}(\mathbf{x})$ is a mapping from $\Omega \subset \mathbb{R}^D$ to $\{1, 2, \ldots, M\}$ and partitions Ω into M regions.

The optimization criterion for the Bayes risk optimal decision rule is the same for M-ary hypothesis testing as for binary hypothesis testing: the expression (2.1); however with general M, there are M^2 costs c_{ij}, with $i = 1, 2, \ldots, M$ and $j = 1, 2, \ldots, M$.

The M-ary Bayes optimal decision rule relies on the comparison of likelihood ratio functions like the binary decision rule, but there are $M - 1$ such likelihood ratio functions. These likelihood ratios are:

$$\Lambda_i(\mathbf{x}) = \frac{f_{\mathbf{x},y}(\mathbf{x}|y=i)}{f_{\mathbf{x},y}(\mathbf{x}|y=1)}, \quad i = 2, \ldots, M \tag{2.27}$$

The decision rule, derived in the same way as the binary rule (2.11), is:

$$\hat{y}(\mathbf{x}) = \arg \min_{i \in \{1,2,\ldots,M\}} \left\{ p_1 c_{1i} + \sum_{j=2}^{M} p_j c_{ij} \Lambda_j(\mathbf{x}) \right\}. \tag{2.28}$$

■ 2.2 Statistical Learning Theory

Statistical learning has two main branches: generative learning and discriminative learning. Generative learning builds probabilistic models of objects or variables starting from a parametric or nonparametric prior model that is updated using a finite data sample.

Discriminative learning encompasses all of the methods and theory associated with determining decision rules based on a finite data sample. Much of the theoretical work is concerned with developing performance bounds, whereas the work of a more applied nature aims to develop methods that yield small error empirically. One of the central problems in statistical learning is supervised classification, which is described in Section 2.2.1.

A key theoretical point from discriminative learning that recurs throughout the thesis is the structural risk minimization principle, discussed in Section 2.2.2. When learning from finite data, inserting constraints into decision making may be beneficial because the constraints provide regularization and prevent overfitting. One of the ways to measure the complexity of a classifier function class, the Rademacher complexity, described also in Section 2.2.2, is characterized for the classifiers proposed in the thesis in Section 3.4 and Section 4.3. The support vector machine described in Section 2.2.3 is one example from the class of margin-based classifiers, which is described in general in Section 2.2.4. The geometric level set classifier proposed in Chapter 3 is a new margin-based classifier. The dimensionality reduction techniques proposed in Chapter 4 have margin-based classification as their objective. Examples in that chapter use the support vector machine as well as the new geometric level set classifier.

■ 2.2.1 Supervised Classification Problem Statement

Supervised classification is the archetypal problem of discriminative learning. Supervised classification is like hypothesis testing (discussed in Section 2.1) in many respects, but with one key difference. In hypothesis testing, the joint probability density function of the measurements and hypotheses is available when constructing the decision rule, but it is not available in supervised classification. Instead, a finite number of samples from the probability distribution are given.

Specifically, the training dataset $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ drawn according to $f_{\mathbf{x},y}(\mathbf{x}, y)$ is given, with measurement vectors $\mathbf{x}_j \in \Omega \subset \mathbb{R}^D$ and class labels $y_j \in \{-1, +1\}$. The goal is to find the decision rule or classifier $\hat{y} : \Omega \to \{-1, +1\}$ that minimizes the probability of error $\Pr[y \neq \hat{y}(\mathbf{x})]$, known as the generalization error. Note that in learning $\hat{y}$, the true objective to be minimized is the generalization error, but a direct minimization is not possible since the joint distribution of $\mathbf{x}$ and y is not available. Also note that in the typical formulation, although not necessary, the two types of errors have equal costs and the generalization error is equivalent to p_E given in (2.4). The generalization error of a classifier learned from n samples is denoted $\mathrm{R}(\hat{y}^{(n)})$.

In practice, the classifier $\hat{y}$ is selected from a function class $\mathcal{F}$ to minimize a loss function of the training data.[2] The general form of the supervised classification problem is an expression like the Bayes risk expression (2.1), but based on samples rather than

[2] In Bayesian hypothesis testing, no restriction is imposed on $\hat{y}$; it can be any function.

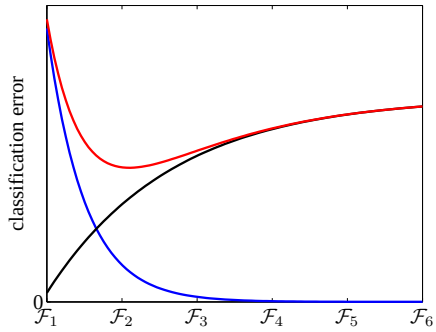


Figure 2.1. The structural risk minimization principle states that for nested function spaces $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$, the training error (blue line) tends to decrease and the complexity (black line) increases. The generalization error (red line) is the sum of the training error and complexity term and has an intermediate function space complexity at which it is minimum.

distributions:

$$\hat{y}(\cdot) = \arg \min_{f(\cdot) \in \mathcal{F}} \frac{1}{n} \sum_{j=1}^{n} \ell(y_j, f(\mathbf{x}_j)). \tag{2.29}$$

Loss functions ℓ in classification may, and in practice do, take into account factors besides the label output provided by the decision rule.

As the cardinality of the training dataset grows, the generalization error of a *consistent* classifier converges to $\mathrm{R}(\hat{y}^*)$, the Bayes risk of the likelihood ratio test (2.11). Specifically, the sequence $\mathrm{R}(\hat{y}^{(n)})$ converges in probability to $\mathrm{R}(\hat{y}^*)$ and equivalently, the sequence $\mathrm{R}(\hat{y}^{(n)}) - \mathrm{R}(\hat{y}^*)$ converges to zero in probability.

■ 2.2.2 Structural Risk Minimization

Since the decision rule $\hat{y}$ is learned based on finite training data, but is applied to and evaluated on new unseen samples $\mathbf{x} \sim \mathbf{x}$, it is critical to consider the phenomenon of overfitting. A decision rule that yields perfect classifications on the training data may or may not have small generalization error. This will depend on whether the classification algorithm has locked on to the vagaries of the samples in the training set or on to the regularities that reoccur in a different sample. The structural risk minimization principle states that a classifier with good generalizability balances training error and complexity [206]. Classifiers with too much complexity overfit the training data. Classifier complexity is a property of the function class $\mathcal{F}$. The structural risk minimization principle is illustrated schematically in Figure 2.1.

The generalization error can be bounded by the sum of the error of $\hat{y}$ on the training set, and a penalty that is larger for more complex $\mathcal{F}$. One such penalty is the Vapnik–Chervonenkis (VC) dimension $C^{\mathrm{VC}}(\mathcal{F})$ [206] and another is the Rademacher complexity $C_n^{\mathrm{Rad}}(\mathcal{F})$ [13, 107]. The definition of VC dimension is based on the concept

of shattering. A set $\mathcal{F}$ can shatter n points if it contains decision rules producing all different combinations of labelings of the points. For example, the set of linear decision boundaries in $\mathbb{R}^2$ can shatter $n = 3$ points but cannot shatter $n = 4$ points. The VC dimension $C^{\mathrm{VC}}(\mathcal{F})$ is the maximum number of points $\mathcal{F}$ can shatter [99, 206].

The definition of Rademacher complexity is based on independent random variables s_j taking the values -1 and $+1$ with equal probability; these random variables are known as Rademacher random variables. Then, with

$$\hat{C}_n^{\mathrm{Rad}}(\mathcal{F}) = \frac{2}{n} \mathrm{E}\left[\sup_{f \in \mathcal{F}} \left| \sum_{j=1}^{n} \mathsf{s}_j f(\mathbf{x}_j) \right| \right], \tag{2.30}$$

where the expectation is over the s_j, the Rademacher complexity $C_n^{\mathrm{Rad}}(\mathcal{F})$ is $\mathrm{E}[\hat{C}_n^{\mathrm{Rad}}(\mathcal{F})]$ with the expectation over $\mathbf{x}_j$.

With probability greater than or equal to $1 - \delta$, a bound on the generalization error utilizing the VC dimension for a specified decision rule $\hat{y}$ is [99, 206]:

$$\mathrm{R}(\hat{y}) \leq \frac{1}{n} \sum_{j=1}^{n} \mathrm{I}(y_j \neq \hat{y}(\mathbf{x}_j)) + \sqrt{\frac{C^{\mathrm{VC}}(\mathcal{F}) \left(\ln(2n/C^{\mathrm{VC}}(\mathcal{F})) + 1 \right) - \ln(4\delta)}{n}}, \tag{2.31}$$

where I is an indicator function. The first term on the right side of (2.31) is the training error and the second term is complexity. With probability greater than or equal to $1-\delta$, Bartlett and Mendelson [13] give a similar bound on the generalization error based on Rademacher complexity:

$$\mathrm{R}(\hat{y}) \leq \frac{1}{n} \sum_{j=1}^{n} \mathrm{I}(y_j \neq \hat{y}(\mathbf{x}_j)) + \frac{C_n^{\mathrm{Rad}}(\mathcal{F})}{2} + \sqrt{\frac{-\ln(\delta)}{2n}}. \tag{2.32}$$

Considering nested function spaces $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$, those with larger index are larger sets and have higher complexity. As the training objective in (2.29) does not change, but the constraint set gets larger with higher classifier complexity, the training error tends to decrease with an increase in classifier complexity. Both bounds reflect the structural risk minimization principle. As a function of increasing classifier complexity, the training error decreases and the complexity term increases; the generalization error is the sum of the two, and thus there exists an optimal intermediate classifier complexity that balances the two terms and minimizes generalization error [206].

■ 2.2.3 Support Vector Machine

One of the most popular classification methods used today is the support vector machine (SVM). The specification of the SVM classifier is based on the concept of margin maximization. The SVM is derived in this section following the derivations of [99, 176] to a large degree, starting with the SVM with linear decision boundary.

To begin, consider a classifier with a linear decision boundary that passes through the origin: $\hat{y}(\mathbf{x}) = \text{sign}(\boldsymbol{\theta}^T\mathbf{x})$. The D-dimensional parameter vector $\boldsymbol{\theta}$ specifies the decision boundary hyperplane $\boldsymbol{\theta}^T\mathbf{x} = 0$. Due to the special encoding $y_j \in \{-1, +1\}$, $y_j\boldsymbol{\theta}^T\mathbf{x}_j > 0$ if a classification is correct, and $y_j\boldsymbol{\theta}^T\mathbf{x}_j < 0$ if incorrect. If the classifier is correct on all n samples in the training set,[3] then there exists a $\gamma > 0$ such that $y_j\boldsymbol{\theta}^T\mathbf{x}_j \geq \gamma$, for all $j = 1, \ldots, n$.

The value $y_j\boldsymbol{\theta}^T\mathbf{x}_j/\|\boldsymbol{\theta}\|$ is equal to the distance from the decision boundary to the sample $\mathbf{x}_j$. This distance value is known as the *margin* of sample $\mathbf{x}_j$. The SVM is based on the idea that a classifier with good generalization has large margin. The SVM objective is to maximize the minimum margin among the training samples while preserving the constraint that they are all correctly classified. Mathematically, the SVM optimization problem is:

$$\begin{aligned} &\text{maximize} && \gamma/\|\boldsymbol{\theta}\| \\ &\text{such that} && y_j\boldsymbol{\theta}^T\mathbf{x}_j \geq \gamma,\ j = 1, \ldots, n. \end{aligned} \tag{2.33}$$

Maximizing $\gamma/\|\boldsymbol{\theta}\|$ is equivalent to minimizing $\frac{1}{2}\|\boldsymbol{\theta}\|^2/\gamma^2$, yielding:

$$\begin{aligned} &\text{minimize} && \tfrac{1}{2}\|\boldsymbol{\theta}\|^2/\gamma^2 \\ &\text{such that} && y_j\boldsymbol{\theta}^T\mathbf{x}_j \geq \gamma,\ j = 1, \ldots, n, \end{aligned} \tag{2.34}$$

which can also be written

$$\begin{aligned} &\text{minimize} && \tfrac{1}{2}\|\boldsymbol{\theta}/\gamma\|^2 \\ &\text{such that} && y_j\left(\boldsymbol{\theta}/\gamma\right)^T\mathbf{x}_j \geq 1,\ j = 1, \ldots, n. \end{aligned} \tag{2.35}$$

The optimization problem (2.35) contains only the ratio $\boldsymbol{\theta}/\gamma$, and the decision rules $\hat{y}$ based on $\boldsymbol{\theta}/\gamma$ and on $\boldsymbol{\theta}$ are equivalent. Thus, γ can be set to one without loss of generality, resulting in:

$$\begin{aligned} &\text{minimize} && \tfrac{1}{2}\|\boldsymbol{\theta}\|^2 \\ &\text{such that} && y_j\boldsymbol{\theta}^T\mathbf{x}_j \geq 1,\ j = 1, \ldots, n. \end{aligned} \tag{2.36}$$

The problem (2.36) is a quadratic program that may be solved efficiently to give the linear SVM classifier $\hat{y}(\mathbf{x}) = \text{sign}(\boldsymbol{\theta}^T\mathbf{x})$.

The classifier function class $\mathcal{F}$ may be enlarged slightly to also include linear decision boundaries that do not pass through the origin via an offset parameter θ_0. The optimization problem in this case is:

$$\begin{aligned} &\text{minimize} && \tfrac{1}{2}\|\boldsymbol{\theta}\|^2 \\ &\text{such that} && y_j\left(\boldsymbol{\theta}^T\mathbf{x}_j + \theta_0\right) \geq 1,\ j = 1, \ldots, n, \end{aligned} \tag{2.37}$$

[3] The classifier can only be correct on all training data if the set is linearly separable. The requirement that the classifier be correct on all training samples is relaxed later in this section.

giving the decision rule $\hat{y}(\mathbf{x}) = \text{sign}(\boldsymbol{\theta}^T\mathbf{x} + \theta_0)$.

The training dataset is not usually linearly separable in real-world problems, rendering the formulation (2.37) infeasible, and prompting a formulation with slack variables. With slack variables ξ_j, the relaxation to (2.37) is:

$$\begin{aligned} \text{minimize} \quad & \tfrac{1}{2}\|\boldsymbol{\theta}\|^2 + \tfrac{1}{\lambda}\sum_{j=1}^{n}\xi_j \\ \text{such that} \quad & y_j\left(\boldsymbol{\theta}^T\mathbf{x}_j + \theta_0\right) \geq 1 - \xi_j,\ j = 1, \ldots, n \\ & \xi_j \geq 0,\ j = 1, \ldots, n. \end{aligned} \tag{2.38}$$

The parameter λ represents the tradeoff between violating the margin constraint and maximizing the minimum margin. A training sample $\mathbf{x}_j$ is incorrectly classified if its corresponding slack variable ξ_j is greater than one. If the slack variable is between zero and one, then the sample is within the margin, though still correctly classified. Notably,

$$\xi_j = \begin{cases} 1 - y_j\left(\boldsymbol{\theta}^T\mathbf{x}_j + \theta_0\right), & y_j\left(\boldsymbol{\theta}^T\mathbf{x}_j + \theta_0\right) \leq 1 \\ 0, & \text{otherwise} \end{cases} \tag{2.39}$$

$$= \max\left\{0, 1 - y_j\left(\boldsymbol{\theta}^T\mathbf{x}_j + \theta_0\right)\right\}. \tag{2.40}$$

Again, once the optimization is performed, the SVM classifier is $\hat{y}(\mathbf{x}) = \text{sign}(\boldsymbol{\theta}^T\mathbf{x} + \theta_0)$.

The problem (2.38) with slack variables can be dualized with variables $\alpha_1, \ldots, \alpha_n$ to yield the dual optimization problem [99, 176]:

$$\begin{aligned} \text{maximize} \quad & \sum_{j=1}^{n}\alpha_j - \tfrac{1}{2}\sum_{j=1}^{n}\sum_{j'=1}^{n}\alpha_j\alpha_{j'}y_jy_{j'}\mathbf{x}_j^T\mathbf{x}_{j'} \\ \text{such that} \quad & \alpha_j \geq 0,\ j = 1, \ldots, n \\ & \alpha_j \leq \tfrac{1}{\lambda},\ j = 1, \ldots, n \\ & \sum_{j=1}^{n}\alpha_j y_j = 0. \end{aligned} \tag{2.41}$$

The dual formulation is also a quadratic program that can be solved efficiently. The primal variables in terms of the dual variables are:

$$\boldsymbol{\theta} = \sum_{j=1}^{n}\alpha_j y_j \mathbf{x}_j, \tag{2.42}$$

and consequently the SVM classifier is

$$\hat{y}(\mathbf{x}) = \text{sign}\left(\sum_{j=1}^{n}\alpha_j y_j \mathbf{x}_j^T\mathbf{x} + \theta_0\right). \tag{2.43}$$

Many datasets encountered in practice are not well-classified by linear classifiers, and thus an extension of the SVM to nonlinear classifiers is desirable. Consider a vector of D_{nonlin} nonlinear functions of the input data $\mathbf{x}$ denoted $\boldsymbol{\phi}(\mathbf{x})$. Treat that D_{nonlin}-dimensional space linearly, replacing $\mathbf{x}$ by $\boldsymbol{\phi}(\mathbf{x})$ in (2.41), which results in:

$$\begin{aligned} \text{maximize} \quad & \sum_{j=1}^{n} \alpha_j - \tfrac{1}{2} \sum_{j=1}^{n} \sum_{j'=1}^{n} \alpha_j \alpha_{j'} y_j y_{j'} \boldsymbol{\phi}(\mathbf{x}_j)^T \boldsymbol{\phi}(\mathbf{x}_{j'}) \\ \text{such that} \quad & \alpha_j \geq 0, \ j = 1, \ldots, n \\ & \alpha_j \leq \tfrac{1}{\lambda}, \ j = 1, \ldots, n \\ & \sum_{j=1}^{n} \alpha_j y_j = 0. \end{aligned} \tag{2.44}$$

Only the inner product $\boldsymbol{\phi}(\mathbf{x}_j)^T \boldsymbol{\phi}(\mathbf{x}_{j'})$, and not $\boldsymbol{\phi}(\mathbf{x}_j)$ separately, appears in (2.44). Kernel functions $\mathrm{K}(\mathbf{x}_j, \mathbf{x}_{j'})$ are a means to compute the inner product without having to compute the full vector of nonlinear functions, thus providing computational savings. The equivalent optimization problem with a kernel function is:

$$\begin{aligned} \text{maximize} \quad & \sum_{j=1}^{n} \alpha_j - \tfrac{1}{2} \sum_{j=1}^{n} \sum_{j'=1}^{n} \alpha_j \alpha_{j'} y_j y_{j'} \, \mathrm{K}(\mathbf{x}_j, \mathbf{x}_{j'}) \\ \text{such that} \quad & \alpha_j \geq 0, \ j = 1, \ldots, n \\ & \alpha_j \leq \tfrac{1}{\lambda}, \ j = 1, \ldots, n \\ & \sum_{j=1}^{n} \alpha_j y_j = 0. \end{aligned} \tag{2.45}$$

The classifier can also be expressed in terms of the kernel function in a manner like (2.43) as:

$$\hat{y}(\mathbf{x}) = \text{sign} \left(\sum_{j=1}^{n} \alpha_j y_j \, \mathrm{K}(\mathbf{x}, \mathbf{x}_j) + \theta_0 \right). \tag{2.46}$$

Note that because of the kernel function, the dimensionality of the nonlinear space D_{nonlin} does not play a role in (2.45) and (2.46). Consequently, even very high-dimensional spaces of nonlinear functions may be considered. Also, kernel functions may be specified directly without explicitly specifying the nonlinear functions $\boldsymbol{\phi}$. Certain directly-specified kernel functions correspond to implicitly infinite-dimensional vectors of nonlinear functions. An example of such a kernel is the radial basis function (RBF) kernel:

$$\mathrm{K}(\mathbf{x}_j, \mathbf{x}_{j'}) = \exp \left(-\frac{\|\mathbf{x}_j - \mathbf{x}_{j'}\|^2}{2\sigma^2} \right), \tag{2.47}$$

with scale parameter σ.

2.2.4 Margin-Based Classification

The SVM is an example of a larger category of classifiers known as margin-based classifiers. When a classifier has the form $\hat{y}(\mathbf{x}) = \text{sign}(\varphi(\mathbf{x}))$ and the decision function φ is chosen to minimize the functional:

$$\text{L}(\varphi) = \sum_{j=1}^{n} \ell(y_j \varphi(\mathbf{x}_j)) + \lambda \, \text{J}(\varphi), \tag{2.48}$$

it is known as a margin-based classifier. The value $y\varphi(\mathbf{x})$ is the margin; it represents the distance that $\mathbf{x}$ is from the classifier decision boundary $\varphi(\mathbf{x}) = 0$. The function ℓ is known as a margin-based loss function. Examples of such functions are the logistic loss function:

$$\ell_{\text{logistic}}(z) = \log\left(1 + e^{-z}\right)$$

and the hinge loss function:

$$\ell_{\text{hinge}}(z) = \max\{0, 1 - z\}.$$

Also J, the second term on the right side of (2.48) with nonnegative weight λ, represents a regularization term that penalizes the complexity of the decision function [12, 116].

To see how the SVM fits the form of margin-based classifiers, write the constrained SVM optimization problem (2.38) in unconstrained form making use of the slack variable expression (2.40), giving:

$$\text{minimize} \quad \tfrac{1}{2}\|\boldsymbol{\theta}\|^2 + \tfrac{1}{\lambda}\sum_{j=1}^{n} \max\left\{0, 1 - y_j\left(\boldsymbol{\theta}^T\mathbf{x}_j + \theta_0\right)\right\}, \tag{2.49}$$

and also

$$\text{minimize} \quad \sum_{j=1}^{n} \max\left\{0, 1 - y_j\left(\boldsymbol{\theta}^T\mathbf{x}_j + \theta_0\right)\right\} + \tfrac{\lambda}{2}\|\boldsymbol{\theta}\|^2. \tag{2.50}$$

With $\varphi(\mathbf{x}) = \boldsymbol{\theta}^T\mathbf{x} + \theta_0$, ℓ being the hinge loss function, and $\text{J}(\varphi) = \frac{1}{2}\|\boldsymbol{\theta}\|^2$, the linear SVM is in the form of a general margin-based classifier. The nonlinear SVM has $\varphi(\mathbf{x}) = \boldsymbol{\theta}^T\boldsymbol{\phi}(\mathbf{x}) + \theta_0 = \sum_{j=1}^{n} \alpha_j y_j K(\mathbf{x}, \mathbf{x}_j) + \theta_0$. Another example of margin-based classification is logistic regression, which uses the logistic loss function and the same regularization term as the SVM.

2.3 Variational Level Set Methods

Variational methods are predicated on the belief that the state and dynamics of a physical system are the optimum of some energy function or energy functional. Variational level set methods are concerned with finding (through optimization) the state and dynamics of interfaces such as the boundary between a flame and the air, or soap bubbles.

They are not restricted, however, to only being used with physical systems and physical energies. One may define a system and appropriate energy to be optimized in any application domain.

Section 2.3.1 details the type of energy functionals appropriate for optimization through variational level set methods. A margin-based classifier with a new regularization term, the surface area of the decision boundary, is proposed in Chapter 3 whose objective fits the requirements of Section 2.3.1. Section 2.3.2 describes gradient descent flow minimization of energy functionals, a procedure known as contour evolution, and Section 2.3.3 discusses implementation of contour evolution using the level set representation. Contour evolution implemented with level sets is derived for the new margin-based classifier in Chapter 3.

■ 2.3.1 Region-Based and Boundary-Based Functionals

Consider a bounded domain $\Omega \subset \mathbb{R}^D$ containing the region $\mathcal{R}$. The rest of the domain is $\Omega \setminus \bar{\mathcal{R}}$, where $\bar{\mathcal{R}}$ is the closure of $\mathcal{R}$. The boundary between the two regions is the contour $\mathcal{C} = \partial\mathcal{R}$. The region $\mathcal{R}$ may be simply connected, multiply connected, or composed of several components. The contour $\mathcal{C}$ is a curve when $D = 2$, a surface when $D = 3$, and a hypersurface in higher dimensions. Points in Ω are $\mathbf{x} = \begin{bmatrix} x_1 & \cdots & x_D \end{bmatrix}^T$. This section describes typical energy functionals $\mathrm{L}(\mathcal{C})$ considered in variational level set methods: region-based functionals and boundary-based functionals. The goal is to minimize $\mathrm{L}(\mathcal{C})$.

Region-based functionals are integrals over either the region $\mathcal{R}$ or the region $\Omega \setminus \bar{\mathcal{R}}$ of a function $g_r(\mathbf{x})$. Since $\mathcal{R}$ depends on $\mathcal{C}$, these integrals are functions of $\mathcal{C}$. The functionals are:

$$\mathrm{L}(\mathcal{C}) = \int_{\mathcal{R}} g_r(\mathbf{x}) d\mathbf{x} \tag{2.51}$$

and

$$\mathrm{L}(\mathcal{C}) = \int_{\Omega \setminus \bar{\mathcal{R}}} g_r(\mathbf{x}) d\mathbf{x}, \tag{2.52}$$

respectively. The function $g_r(\mathbf{x})$ depends on the application at hand. For example in the image processing application of segmentation, it may depend on both the observed image values and the expected statistics of the foreground and background of the image.

Boundary-based functionals are contour integrals over the boundary $\mathcal{C}$ of a function $g_b(\mathcal{C}(\mathbf{s}))$ where the variable $\mathbf{s}$ parameterizes $\mathcal{C}$. The boundary-based functional is:

$$\mathrm{L}(\mathcal{C}) = \oint_{\mathcal{C}} g_b(\mathcal{C}(\mathbf{s})) d\mathbf{s}. \tag{2.53}$$

An example of a function $g_b(\mathcal{C}(\mathbf{s}))$ in image segmentation is one that is small at strong image edges, and large at smooth, nonedge locations in the image. Overall energy functional objectives for an application may contain linear combinations of (2.51), (2.52), and (2.53).

■ 2.3.2 Contour Evolution

The objective is to minimize $\mathrm{L}(\mathcal{C})$; the approach taken in variational level set methods is to start with some initial contour and to follow a gradient descent flow. This procedure is known as curve evolution or contour evolution. The direction of the gradient descent is given by the negative of the first variation of the energy functional obtained using the calculus of variations [70], denoted $-\frac{\delta \mathrm{L}}{\delta \mathcal{C}}$. The first variations of region-based and boundary-based functionals are derived in this section; the derivations do not follow any particular reference per se, but are an amalgam of the expositions found in [34, 50, 105, 141, 142, 146, 201, 225].

First, the first variation of the region-based functional (2.51) is found. Let $g_r(\mathbf{x}) = \nabla \cdot \mathbf{v}(\mathbf{x})$ for some vector field $\mathbf{v}(\mathbf{x}) = \begin{bmatrix} v_1(\mathbf{x}) & \cdots & v_D(\mathbf{x}) \end{bmatrix}^T$. The divergence in expanded form is:

$$\nabla \cdot \mathbf{v} = \frac{\partial v_1}{\partial x_1} + \cdots + \frac{\partial v_D}{\partial x_D}, \tag{2.54}$$

and the energy functional is:

$$\mathrm{L}(\mathcal{C}) = \int_{\mathcal{R}} \nabla \cdot \mathbf{v}(\mathbf{x}) d\mathbf{x}. \tag{2.55}$$

By the (Gauss–Ostrogradsky) divergence theorem [68]:

$$\mathrm{L}(\mathcal{C}) = \oint_{\mathcal{C}} \mathbf{v}(\mathbf{x}(\mathbf{s})) \cdot \mathbf{n} d\mathbf{s}, \tag{2.56}$$

where $\mathbf{n}$ is the outward unit normal to the contour $\mathcal{C}$.

For simplicity of exposition, consider $D = 2$, in which case the normal vector to the curve $\mathcal{C}$ is:

$$\mathbf{n} = \begin{bmatrix} \frac{dx_2}{ds} \\ -\frac{dx_1}{ds} \end{bmatrix} = \begin{bmatrix} \dot{x}_2 \\ -\dot{x}_1 \end{bmatrix}. \tag{2.57}$$

The notation $\dot{x}_1 = \frac{dx_1}{ds}$ and $\dot{x}_2 = \frac{dx_2}{ds}$ is introduced for simplicity in later manipulations. Let $\mathrm{I}(\mathbf{x}, \dot{\mathbf{x}}) = v_1(\mathbf{x}(s))\dot{x}_2 - v_2(\mathbf{x}(s))\dot{x}_1$. Then,

$$\mathrm{L}(\mathcal{C}) = \oint_{\mathcal{C}} \mathrm{I}(\mathbf{x}(s), \dot{\mathbf{x}}) ds. \tag{2.58}$$

The first result of the calculus of variations is the Euler–Lagrange equation stating that if $\mathrm{L}(\mathcal{C})$ is minimum, then the first variation $\frac{\delta \mathrm{L}}{\delta \mathcal{C}} = 0$ [70]. Moreover, the first variation of L is:

$$\begin{bmatrix} \frac{\delta \mathrm{L}}{\delta x_1} \\ \frac{\delta \mathrm{L}}{\delta x_2} \end{bmatrix} = \begin{bmatrix} \frac{\partial \mathrm{I}}{\partial x_1} - \frac{d}{ds}\frac{\partial \mathrm{I}}{\partial \dot{x}_1} \\ \frac{\partial \mathrm{I}}{\partial x_2} - \frac{d}{ds}\frac{\partial \mathrm{I}}{\partial \dot{x}_2} \end{bmatrix}. \tag{2.59}$$

The first variation expression may be manipulated as follows:

$$\begin{bmatrix} \frac{\delta \mathrm{L}}{\delta x_1} \\ \frac{\delta \mathrm{L}}{\delta x_2} \end{bmatrix} = \begin{bmatrix} \frac{\partial v_1}{\partial x_1}\dot{x}_2 - \frac{\partial v_2}{\partial x_1}\dot{x}_1 - \frac{d}{ds}\left\{-v_2(\mathbf{x}(s))\right\} \\ \frac{\partial v_1}{\partial x_2}\dot{x}_2 - \frac{\partial v_2}{\partial x_2}\dot{x}_1 - \frac{d}{ds}\left\{v_1(\mathbf{x}(s))\right\} \end{bmatrix} \tag{2.60}$$

$$= \begin{bmatrix} \frac{\partial v_1}{\partial x_1}\dot{x}_2 - \frac{\partial v_2}{\partial x_1}\dot{x}_1 + \frac{\partial v_2}{\partial x_1}\dot{x}_1 + \frac{\partial v_2}{\partial x_2}\dot{x}_2 \\ \frac{\partial v_1}{\partial x_2}\dot{x}_2 - \frac{\partial v_2}{\partial x_2}\dot{x}_1 - \frac{\partial v_1}{\partial x_1}\dot{x}_1 - \frac{\partial v_1}{\partial x_2}\dot{x}_2 \end{bmatrix} \tag{2.61}$$

$$= \left(\frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2}\right) \begin{bmatrix} \dot{x}_2 \\ -\dot{x}_1 \end{bmatrix} \tag{2.62}$$

which can be recognized as $g_r(\mathbf{x})\mathbf{n}$, cf. (2.54) and (2.57). The result $\frac{\delta \mathrm{L}}{\delta \mathcal{C}} = g_r(\mathbf{x})\mathbf{n}$ is true for $D > 2$ as well, which can be shown by following the same steps as for $D = 2$. Note that the first variation does not depend on the specific choice of the vector field $\mathbf{v}(\mathbf{x})$, as that choice is arbitrary as long as $\nabla \cdot \mathbf{v}(\mathbf{x}) = g_r(\mathbf{x})$. The outward normal of the region $\Omega \setminus \bar{\mathcal{R}}$ is negative of the outward normal of the region $\mathcal{R}$. Thus, for integrals over $\Omega \setminus \bar{\mathcal{R}}$ (2.52), the first variation is $\frac{\delta \mathrm{L}}{\delta \mathcal{C}} = -g_r(\mathbf{x})\mathbf{n}$.

Now the first variation of the boundary-based functional (2.53) is derived. The approach followed is to first express the boundary-based $\mathrm{L}(\mathcal{C})$ in the same form as (2.56). Then, the first variation of the boundary-based functional is simply stated based on the first variation of functionals of the form (2.56), which has already been derived in the region-based context. Expand $\mathrm{L}(\mathcal{C}) = \oint_{\mathcal{C}} g_b(\mathcal{C}(\mathbf{s}))d\mathbf{s}$, as follows:

$$\mathrm{L}(\mathcal{C}) = \oint_{\mathcal{C}} g_b(\mathcal{C}(\mathbf{s}))\mathbf{n} \cdot \mathbf{n} d\mathbf{s}, \tag{2.63}$$

which is equivalent since $\mathbf{n} \cdot \mathbf{n} = 1$. Using the findings from the region-based functional, $\frac{\delta \mathrm{L}}{\delta \mathcal{C}} = (\nabla \cdot \mathbf{v})\,\mathbf{n}$, where here, $\mathbf{v} = g_b(\mathcal{C}(\mathbf{s}))\mathbf{n}$. That is:

$$\frac{\delta \mathrm{L}}{\delta \mathcal{C}} = (\nabla \cdot \mathbf{v})\mathbf{n} = (\nabla \cdot (g_b\mathbf{n}))\,\mathbf{n}. \tag{2.64}$$

The divergence of a vector field multiplied by a scalar field has the following equivalent expression known as the product rule of divergence:

$$\frac{\delta \mathrm{L}}{\delta \mathcal{C}} = ((\nabla g_b) \cdot \mathbf{n} + g_b \nabla \cdot \mathbf{n})\,\mathbf{n}. \tag{2.65}$$

The divergence of the normal is related to the mean curvature by $\kappa = -\nabla \cdot \mathbf{n}$ [141], so the first variation for boundary-based functionals is $((\nabla g_b) \cdot \mathbf{n} - g_b\kappa)\,\mathbf{n}$.

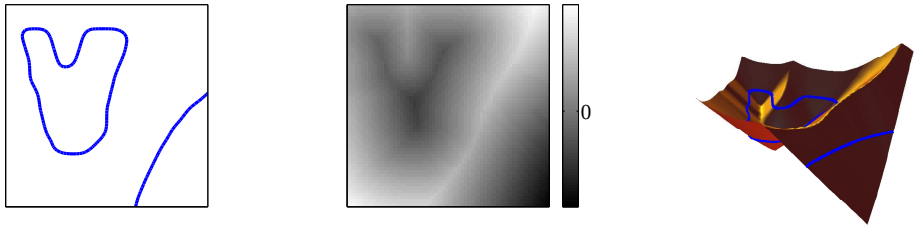
The contour evolution, i.e. the gradient descent flow, is parameterized by a time t, so that the contour evolution is $\mathcal{C}(t)$. With the change in the contour being in the negative first variation direction, the partial derivatives of the contour with respect to t have the following expressions:

$$\mathcal{C}_t = -g_r\mathbf{n} \tag{2.66}$$

for the region-based functional (2.51), and

$$\mathcal{C}_t = (g_b\kappa - (\nabla g_b) \cdot \mathbf{n})\,\mathbf{n} \tag{2.67}$$

for the boundary-based functional (2.53).



(a) (b) (c)

Figure 2.2. An illustration of the level set function representation of a contour with $D = 2$. The contour is shown in (a), its level set function is shown by shading in (b), and as a surface plot marked with the zero level set in (c).

■ 2.3.3 Level Set Representation

In order to numerically implement contour evolution, some representation for $\mathcal{C}(t)$ is required. One possible representation is through a collection of control points that lie on the contour. Another possible representation is through a level set function $\varphi(\mathbf{x}; t)$. The contour is represented as the zero level set of φ. Given a particular level set function φ, the contour $\mathcal{C}$ it describes is unique.

To help understand the level set concept, an analogy with the Hawai'ian islands is useful. The elevation of the land, both underwater and above water can be considered as the level set function $\varphi(\mathbf{x})$, and sea level as the zero level. The coastlines are the contour $\mathcal{C}$. Over time, the elevation can increase or decrease locally, changing the coastline. If the land rises up very high, two islands can merge and if there is sinking, one island may split into more than one island. Also, 'almost' islands which are peaks but not above sea level, can rise and create new islands. Level set methods naturally handle changes in topology to the contour $\mathcal{C}$, unlike the control point representation; this property among several others makes the level set representation the preferred representation for contour evolution.

The level set function is a smooth, Lipschitz continuous function that satisfies the following properties:

$$\varphi(\mathbf{x}; t) < 0, \quad \mathbf{x} \in \mathcal{R}(t), \tag{2.68}$$

$$\varphi(\mathbf{x}; t) = 0, \quad \mathbf{x} \in \mathcal{C}(t), \tag{2.69}$$

$$\varphi(\mathbf{x}; t) > 0, \quad \mathbf{x} \notin \bar{\mathcal{R}}(t). \tag{2.70}$$

An example of a contour and its corresponding level set function are shown in Figure 2.2. Another appealing property of the level set representation is the ease with which geometric quantities such as the normal vector $\mathbf{n}$ and curvature κ may be calculated. The outward unit normal to $\mathcal{C}$ is the normalized gradient of the level set function

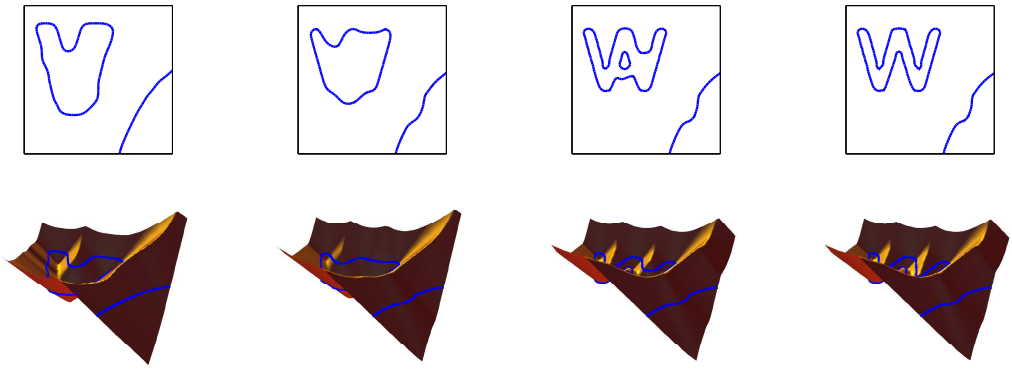


Figure 2.3. Iterations of an illustrative contour evolution proceeding from left to right. The top row shows the contour and the bottom row shows the corresponding signed distance function.

(evaluated on $\mathcal{C}$):

$$\mathbf{n} = \frac{\nabla\varphi}{\|\nabla\varphi\|}. \tag{2.71}$$

The curvature can also be expressed in terms of the level set function:

$$\kappa = -\nabla \cdot \left(\frac{\nabla\varphi}{\|\nabla\varphi\|} \right). \tag{2.72}$$

Oftentimes, the particular level set function known as the signed distance function is employed. The magnitude of the signed distance function at a point equals its distance to $\mathcal{C}$, and its sign indicates whether it is in $\mathcal{R}$ or not. The signed distance function satisfies the additional constraint that $\|\nabla\varphi(\mathbf{x})\| = 1$ and has Lipschitz constant equal to one. With a signed distance function, the normal simplifies to $\mathbf{n} = \nabla\varphi$ and the curvature to $\kappa = -\nabla^2\varphi$. Given a level set function that is not a signed distance function, an equivalent signed distance function with the same zero level set may be obtained through the following Eikonal partial differential equation [192]:

$$\varphi_t(\mathbf{x}) = \text{sign}(\varphi(\mathbf{x}))(1 - \|\nabla\varphi(\mathbf{x})\|), \tag{2.73}$$

where φ_t is the partial derivative of φ with respect to time parameter t.

Contour evolutions correspond to evolutions of their level set functions. An example of a contour evolution through the evolution of its level set function is shown in Figure 2.3. Based on the normal and curvature expressed in terms of the signed distance function and the region-based and boundary-based gradient descent flows (2.66) and (2.67), the gradient descent flow of the following energy functional

$$\text{L}(\mathcal{C}) = \int_{\mathcal{R}} g_r(\mathbf{x})d\mathbf{x} + \lambda \oint_{\mathcal{C}} g_b(\mathcal{C}(\mathbf{s}))d\mathbf{s} \tag{2.74}$$

expressed in terms of the signed distance function is:

$$\varphi_t(\mathbf{x}) = -g_r(\mathbf{x})\nabla\varphi(\mathbf{x}) - \lambda\left[g_b(\mathbf{x})\nabla^2\varphi(\mathbf{x}) + (\nabla g_b(\mathbf{x}))\cdot(\nabla\varphi(\mathbf{x}))\right]\nabla\varphi(\mathbf{x}). \tag{2.75}$$

The update (2.75) does not preserve the signed distance property of the level set function in general; the level set function must be periodically reinitialized as a signed distance function using (2.73).

■ 2.4 Dimensionality Reduction

Dimensionality reduction, the mapping of high-dimensional data to lower-dimensional features, is an important procedure in many data analysis applications. It aids in visualization and human interpretation of data, allows the identification of important data components, reduces the computational and memory requirements of further analysis, and can provide noise suppression.

In Section 2.4.1, the linear dimensionality reduction problem is introduced as an optimization problem on the Stiefel manifold of matrices. The optimization framework is used to propose a method for joint linear dimensionality reduction and margin-based classification in Section 4.1, with an extension to sensor networks given in Section 4.4. Section 2.4.2 describes zonotopes and their relationship to linear dimensionality reduction. The content of zonotopes is an important ingredient in the consistency and complexity analysis of Section 4.3. Nonlinear dimensionality reduction and manifold learning techniques are discussed in Section 2.4.3 and extended for margin-based classification in Section 4.2.

■ 2.4.1 Linear Dimensionality Reduction and the Stiefel Manifold

Linear dimensionality reduction is the mapping of D-dimensional data to $d \leq D$ dimensions by a linear function. It can be represented by a matrix $\mathbf{A} \in \mathbb{R}^{D\times d}$ with elements a_{ij}. With a data vector $\mathbf{x} \in \mathbb{R}^D$, $\tilde{\mathbf{x}} = \mathbf{A}^T\mathbf{x}$ is in d dimensions. Typically, scalings of the reduced-dimensional data are not of interest, so the set of possible matrices is limited to those which involve orthogonal projection, i.e. to the *Stiefel manifold* of $D \times d$ matrices:

$$\mathcal{V}(D,d) = \{\mathbf{A} \in \mathbb{R}^{D\times d}, d \leq D | \mathbf{A}^T\mathbf{A} = \mathbf{I}\}. \tag{2.76}$$

The Stiefel manifold is the set of all linear subspaces with basis specified. Linear dimensionality reduction involves finding a mapping $\mathbf{A}$ that minimizes an objective $\mathrm{L}(\mathbf{A})$ [188]:

$$\min \mathrm{L}(\mathbf{A}) \quad \text{such that } \mathbf{A} \in \mathcal{V}(D,d), \tag{2.77}$$

where L is a scalar-valued function and the optimization is constrained to the Stiefel manifold. For some specific choices of $\mathrm{L}(\mathbf{A})$, e.g. those corresponding to the popular linear dimensionality reduction methods principal component analysis (PCA) [95, 96, 148] and Fisher's linear discriminant analysis (FDA) [69], this optimization problem can be solved through eigendecomposition.

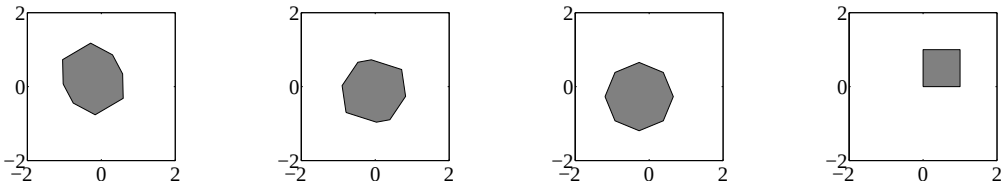


Figure 2.4. Several zonotopes in $\mathcal{Z}(4,2)$. The zonotopes in (a) and (b) are generated by random Stiefel manifold matrices, the zonotope in (c) is content-maximizing, and the zonotope in (d) is content-minimizing.

Several iterative gradient-based minimization algorithms exist for differentiable functions $\mathrm{L}(\mathbf{A})$ [60, 124, 140]. The expression for gradient descent along geodesics of the Stiefel manifold given by Edelman et al. [60] is as follows. Let $\mathbf{L_A}$ denote the $D \times d$ matrix with elements $\partial \mathrm{L} / \partial a_{ij}$. The gradient is:

$$\mathbf{G} = \mathbf{L_A} - \mathbf{A}\,\mathbf{L_A}^T\,\mathbf{A}. \tag{2.78}$$

Starting at an initial $\mathbf{A}(0)$, a step of length τ in the direction $-\mathbf{G}$ to $\mathbf{A}(\tau)$ is:

$$\mathbf{A}(\tau) = \mathbf{A}(0)\mathbf{M}(\tau) + \mathbf{Q}\mathbf{N}(\tau), \tag{2.79}$$

where $\mathbf{QR}$ is the QR decomposition of $(\mathbf{A}\mathbf{A}^T\mathbf{G} - \mathbf{G})$, and

$$\begin{bmatrix}\mathbf{M}(\tau)\\ \mathbf{N}(\tau)\end{bmatrix} = \exp\left\{\tau \begin{bmatrix}-\mathbf{A}^T\mathbf{G} & -\mathbf{R}^T\\ \mathbf{R} & 0\end{bmatrix}\right\}\begin{bmatrix}\mathbf{I}\\ 0\end{bmatrix}.$$

The step size τ may be optimized by a line search.

■ 2.4.2 Zonotopes

Consider the D-dimensional unit hypercube, denoted $\Omega = [0,1]^D$, and a matrix $\mathbf{A} \in \mathcal{V}(D,d)$. The set $Z = \mathbf{A}^T\Omega \subset \mathbb{R}^d$, the orthogonal shadow cast by Ω due to the projection $\mathbf{A}$, is a zonotope, a particular type of polytope that is convex, centrally-symmetric, and whose faces are also centrally-symmetric in all lower dimensions [37, 65]. For reference, Figure 2.4 shows several zonotopes for $D = 4$ and $d = 2$. The matrix $\mathbf{A}^T$ is known as the generator of the zonotope Z; the notation $Z(\mathbf{A})$ is used to denote the zonotope generated by $\mathbf{A}^T$. Also, let

$$\mathcal{Z}(D,d) = \{Z(\mathbf{A}) | \mathbf{A} \in \mathcal{V}(D,d)\}. \tag{2.80}$$

Although the relationship between zonotopes and their generators is not bijective, zonotopes provide a good means of visualizing Stiefel manifold matrices, especially when $d = 2$.

As seen in Figure 2.4, the area (equivalently the content[4] for $d > 2$) of zonotopes is variable. For $d = 2$, the area is maximized by a regular polygon with $2D$ sides having area $\cot(\pi/2D)$, and minimized by a square having area 1 [37]. Figure 2.4(a) and Figure 2.4(b) show zonotopes generated by random Stiefel manifold matrices [43], Figure 2.4(c) shows an area-maximizing zonotope generated by

$$\mathbf{A}^T = \frac{1}{\sqrt{2\alpha^2+2}} \begin{bmatrix} \alpha & -\alpha & -1 & -1 \\ -1 & -1 & \alpha & -\alpha \end{bmatrix}$$

where $\alpha = \sqrt{2}+1$, and Figure 2.4(d) shows an area-minimizing zonotope generated by

$$\mathbf{A}^T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

As already discussed, for $Z \in \mathcal{Z}(D,2)$,

$$1 \leq V(Z) \leq \cot\left(\tfrac{\pi}{2D}\right). \tag{2.81}$$

For general $\mathcal{Z}(D,d)$, with $d \geq 2$, the same lower bound is achieved when the zonotope is a d-dimensional unit hypercube. There is no tight closed-form upper bound for $V(Z)$, but an upper bound is developed in [37] that is asymptotically of the correct order of magnitude for fixed d as D goes to infinity. Specifically, for $Z \in \mathcal{Z}(D,d)$

$$1 \leq V(Z) \leq \omega_d \left(\frac{\omega_{d-1}}{\omega_d} \sqrt{\frac{D}{d}} \right)^d, \tag{2.82}$$

where $\omega_d = \sqrt{\pi}^d / \Gamma(1+d/2)$ is the content of the d-dimensional unit hypersphere.

■ 2.4.3 Nonlinear Dimensionality Reduction

Section 2.4.1 is concerned with finding a d-dimensional linear subspace within a D-dimensional space. However, it may be the case that data samples are better represented by a d-dimensional nonlinear manifold—a space that is locally but not globally equivalent to a Euclidean space. This section deals with dimensionality reduction that finds nonlinear manifolds [17, 85, 166, 193]. Methods for nonlinear dimensionality reduction, also known as manifold learning, have samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$, with $\mathbf{x}_j \in \mathbb{R}^D$, as input and produce an embedding of those points on a manifold $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n$, with $\tilde{\mathbf{x}}_j \in \mathbb{R}^d$. Manifolds are locally smooth, and thus the embedding should preserve local geometric quantities such as the distances between neighboring points in the high-dimensional space.

Multidimensional scaling is a technique that, given all of the pairwise distances or dissimilarities between n samples, produces the low-dimensional Euclidean embedding of those n samples that minimizes the error between the distances in the embedding

[4] The content of a polytope is also known as its volume or hypervolume.

and those input [47]. Specifically, the input is an $n \times n$ matrix $\mathbf{B}$ with elements $b_{ij} = \rho(\mathbf{x}_i, \mathbf{x}_j)^2$, where $\rho(\cdot, \cdot)$ is a symmetric distance or dissimilarity function, and the output is the low-dimensional embedding $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n$ which may be expressed as a $d \times n$ matrix $\tilde{\mathbf{X}}$ with each column being one of the embedded samples. The criterion that is minimized in finding the embedding is the Frobenius norm $\| -\frac{1}{2}\mathbf{H}(\mathbf{B} - \tilde{\mathbf{X}}^T\tilde{\mathbf{X}})\mathbf{H}\|_F$, where $\mathbf{H} = \mathbf{I} - 1/n$ is known as the centering matrix. Left multiplication by the centering matrix makes the mean of each column of a matrix zero. Let $\mathbf{M} = -\frac{1}{2}\mathbf{HBH}$ with elements

$$m_{ij} = -\frac{1}{2}\left(\rho(\mathbf{x}_i, \mathbf{x}_j)^2 - \frac{1}{n}\sum_{i'=1}^{n}\rho(\mathbf{x}_i, \mathbf{x}_{i'})^2 - \frac{1}{n}\sum_{j'=1}^{n}\rho(\mathbf{x}_{j'}, \mathbf{x}_j)^2 + \frac{1}{n^2}\sum_{i'=1}^{n}\sum_{j'=1}^{n}\rho(\mathbf{x}_{i'}, \mathbf{x}_{j'})^2\right).$$

The multidimensional scaling solution is

$$\tilde{\mathbf{x}}_j = \begin{bmatrix} \sqrt{l_1}v_{1,j} \\ \vdots \\ \sqrt{l_d}v_{d,j} \end{bmatrix}, \tag{2.83}$$

where $\mathbf{v}_k = \begin{bmatrix} v_{k,1} & \cdots & v_{k,n} \end{bmatrix}^T$ is the eigenvector of $\mathbf{M}$ corresponding to the kth largest eigenvalue l_k.

The method of multidimensional scaling is agnostic to the distance or dissimilarity ρ; the key idea of the nonlinear dimensionality reduction technique Isomap is to use an approximation to geodesic distance along the manifold for ρ, computed in the high-dimensional space [193]. As mentioned previously in the section, manifolds may be approximated as Euclidean locally, but not globally. Thus geodesic distance on the manifold from a sample may be approximated by Euclidean distance to its neighbors, but not to samples outside its neighborhood. However, distances to samples far away may be approximated by taking local hops from sample to neighboring sample until reaching the far away point. The distances ρ and matrix $\mathbf{B}$ employed by Isomap are computed by first constructing a neighborhood graph with the samples as the vertices. Edges exist only between those samples that are in the same neighborhood; one way of defining the neighborhood is through the k-nearest neighbors by Euclidean distance in $\mathbb{R}^D$. Then $\rho(\mathbf{x}_i, \mathbf{x}_j)$ is the length of the shortest path in the graph between the vertices corresponding to $\mathbf{x}_i$ and $\mathbf{x}_j$, where the weight of an edge is the Euclidean distance between the two samples it connects. With this definition of ρ, the Isomap embedding is (2.83).

Isomap and other nonlinear dimensionality techniques only produce low-dimensional embeddings of the given samples, but do not provide a function that can be applied to a new sample $\mathbf{x}$ not included in the original set of samples. Bengio et al. [19] provide a function mapping $\mathbf{g}(\mathbf{x}) \in \mathbb{R}^d$ that can be applied to vectors $\mathbf{x}$ that are not part of the input set, but which is equivalent to the nonlinear dimensionality reduction technique at the input samples; they also discuss the functional minimized to yield $\mathbf{g}(\mathbf{x})$. For

Isomap, the function mapping is:

$$\mathbf{g}(\mathbf{x}) = \begin{bmatrix} \sum_{j=1}^{n} a_{j,1} \, \mathrm{K}(\mathbf{x}, \mathbf{x}_j) \\ \vdots \\ \sum_{j=1}^{n} a_{j,d} \, \mathrm{K}(\mathbf{x}, \mathbf{x}_j) \end{bmatrix}, \tag{2.84}$$

where

$$\mathrm{K}(\mathbf{w}, \mathbf{z}) = -\frac{1}{2} \left(\rho(\mathbf{w}, \mathbf{z})^2 - \frac{1}{n} \sum_{j=1}^{n} \rho(\mathbf{w}, \mathbf{x}_j)^2 - \frac{1}{n} \sum_{j=1}^{n} \rho(\mathbf{x}_j, \mathbf{z})^2 + \frac{1}{n^2} \sum_{j=1}^{n} \sum_{j'=1}^{n} \rho(\mathbf{x}_j, \mathbf{x}_{j'})^2 \right), \tag{2.85}$$

and

$$a_{jk} = \frac{1}{\sqrt{l_k}} v_{k,j}, \tag{2.86}$$

with the eigenvalues and eigenvectors as before.

Defining the length n vector $\tilde{\mathbf{K}}(\mathbf{x}) = \begin{bmatrix} \mathrm{K}(\mathbf{x}, \mathbf{x}_1) & \cdots & \mathrm{K}(\mathbf{x}, \mathbf{x}_n) \end{bmatrix}^T$ and the $n \times d$ matrix $\mathbf{A}$ with elements a_{jk} from (2.86), the dimensionality-reduced image of the sample $\mathbf{x}$ is:

$$\tilde{\mathbf{x}} = \mathbf{A}^T \tilde{\mathbf{K}}(\mathbf{x}). \tag{2.87}$$

Since the columns of $\mathbf{A}$ are proportional to eigenvectors of a symmetric matrix (the centered matrix of distances $\mathbf{M}$ is symmetric), they are orthogonal. The different dimensions in the low-dimensional space have different scale factors due to the eigenvalues $l_1, \ldots, l_d$. In certain applications, it may be the case that this scaling of the dimensions is not important; in those cases, $\mathbf{A} \in \mathcal{V}(n, d)$ if the dimensions are normalized to be on the same scale.

■ 2.5 Quantization Theory

Quantization arises when analog signals are represented digitally, and in other similar scenarios. Whenever there is quantization, there is necessarily some distortion. The simplest quantizer takes a number and maps it to the nearest value from a preselected set of allowed numbers, e.g. rounding real numbers to the nearest integer. Quantization theory considers the design of mappings that minimize the expected distortion that is incurred over a probability distribution.

Section 2.5.1 posits the quantization problem for a generic setting with a generic distortion function. The general framework is used in a specific manner for the quantization of prior probabilities for hypothesis testing in Chapter 5. The chapter proposes a new distortion function for that problem, the Bayes risk error. Section 2.5.1 also briefly discusses high-rate quantization; a high-rate characterization of the minimum mean Bayes risk error (MBRE) quantizer is presented in Section 5.2. Section 2.5.2 gives the conditions for optimality of a generic quantizer. These conditions are specialized

to the minimum MBRE quantizer in Section 5.1. The algorithm used to design quantizers, both based on probability distributions and based on samples, is discussed in Section 2.5.3 and applied in Section 5.1 and Section 5.3. Section 2.5.4 discusses the convergence of quantization error as a function of quantizer complexity, which is useful in drawing conclusions in Section 5.4.

■ 2.5.1 Quantization Problem Statement

Consider a scalar random variable $\mathsf{p} \in \mathbb{R}$ whose samples are to be represented using a small number of discrete values. In the quantization problem, the task is find a function known as the quantizer that partitions the real line $\mathbb{R}$ into cells $\mathcal{Q}_1, \ldots, \mathcal{Q}_k$ and has representation points $a_1, \ldots, a_k$ such that sample $p \in \mathcal{Q}_i$ is represented by a_i in a manner that minimizes an expected distortion ϱ. A quantizer is a nonlinear function $q_k(\cdot)$ such that $q_k(p) = a_i$ for $p \in \mathcal{Q}_i$, $i = 1, \ldots, k$.

A k-point regular quantizer partitions the set of real numbers into k intervals $\mathcal{Q}_1 = (b_0, b_1]$, $\mathcal{Q}_2 = (b_1, b_2]$, $\mathcal{Q}_3 = (b_2, b_3]$, $\ldots$, $\mathcal{Q}_k = (b_{k-1}, b_k)$, with $b_0 = -\infty$ and $b_k = +\infty$, and $b_{i-1} < a_i \leq b_i$. The performance of a quantizer is measured with respect to a given distortion criterion $\rho(p, a)$ between the sample p and the representation point a. A common distortion is squared error: $\rho_2(p, a) = |p - a|^2$ [75, 79]. Another common criterion is absolute error: $\rho_1(p, a) = |p - a|$ [72, 79, 103]. A distortion function must satisfy $\rho(p, a) \geq 0$.

For a given value of k, the objective is to find the quantizer that minimizes the expected value of the distortion over the distribution for p:

$$\varrho = \mathrm{E}[\rho(\mathsf{p}, q_k(\mathsf{p}))] = \int \rho(p, q_k(p)) f_{\mathsf{p}}(p) dp. \tag{2.88}$$

The performance of the quantizer is the expected distortion ϱ, which depends on k. High-rate quantization theory is the study of the distortion-rate function $\varrho(k)$ for large k, especially asymptotically as k grows [79, 89]. At the limit $k \to \infty$, the notion of representation points $p = a_i$, $i = 1, \ldots, k$ is replaced by the point density function $\lambda(p)$. Integrating the point density over an interval yields the fraction of the representation points in that interval.

■ 2.5.2 Optimality Conditions

For a given k, the goal is to find the quantizer that minimizes the average distortion ϱ given in (2.88). This problem does not have a closed-form solution in general. However, there are three conditions that an optimal quantizer must satisfy. These three necessary conditions are known as the nearest neighbor condition, the centroid condition, and the zero probability of boundary condition, and are discussed in this section.

First consider finding the optimal quantization cells $\mathcal{Q}_1, \ldots, \mathcal{Q}_k$ for given representation points $a_1, \ldots, a_k$. There is no partition better than the one that maps points to the closest representation point. Formally, this observation is the nearest neighbor condition.

Condition 2.5.1 (Nearest Neighbor). *For a given set of representation points $a_1, \ldots, a_k$, the quantization cells satisfy*

$$\mathcal{Q}_i \subset \{p | \rho(p, a_i) \leq \rho(p, a_{i'}) \text{ for all } i' \neq i\}.$$

For ρ_1 and ρ_2, the absolute error and squared error respectively, the nearest neighbor condition implies that the quantization cell boundaries are the midpoints between two adjacent representation points: $b_i = (a_i + a_{i+1})/2$, $i = 1, \ldots, k-1$.

Now consider finding the optimal representation points $a_1, \ldots, a_k$ for a given set of quantizer cells $\mathcal{Q}_1, \ldots, \mathcal{Q}_k$. The optimal representation point for a given quantization cell is found by minimizing the conditional expected distortion. Defining the centroid $\text{cent}(\mathcal{Q})$ of a random variable p in a cell $\mathcal{Q}$ with respect to a distortion function $\rho(\cdot, \cdot)$ as

$$\text{cent}(\mathcal{Q}) = \arg\min_a \mathrm{E}[\rho(\mathsf{p}, a) | \mathsf{p} \in \mathcal{Q}], \tag{2.89}$$

the centroid condition that an optimal quantizer must satisfy is the following.

Condition 2.5.2 (Centroid). *For a given set of quantization cells $\mathcal{Q}_1, \ldots, \mathcal{Q}_k$, the representation points satisfy*

$$a_i = \text{cent}(\mathcal{Q}_i).$$

The third necessary condition for quantizer optimality arises when dealing with an $f_{\mathsf{p}}(p)$ having a discrete component. Consider a quantizer that satisfies the nearest neighbor and centroid conditions, and has a cell boundary b_i between two adjacent representation points $a_i < a_{i+1}$, with $b_i \in \mathcal{Q}_i$ (and $b_i \notin \mathcal{Q}_{i+1}$). Also $f_{\mathsf{p}}(p = b_i)$ is an impulse so that b_i has positive probability mass. If the quantizer is modified so that $b_i \in \mathcal{Q}_{i+1}$, the total expected distortion overall does not change, but the centroid of $\mathcal{Q}_i$ is changed. Due to this phenomenon, the zero probability of boundary condition is required.

Condition 2.5.3 (Zero Probability of Boundary). *The random variable to be quantized has zero probability of occurring at a boundary between quantization cells.*

When $f_{\mathsf{p}}(p)$ is absolutely continuous with respect to a Lebesgue measure, the zero probability of boundary condition is always satisfied.

It is shown in [200] that the conditions necessary for optimality of the quantizer are also sufficient conditions for local optimality if the following hold. The first condition is that $f_{\mathsf{p}}(p)$ must be positive and continuous. The second condition is that $\int \rho(p, a) f_{\mathsf{p}}(p) dp$ must be finite for all a. The third condition is that the distortion function $\rho(p, a)$ must satisfy some properties. It must be zero only for $p = a$, continuous in p for all a, and convex in a. Further conditions on ρ and $f_{\mathsf{p}}(p)$ are given in [200] for there to be a unique locally optimal quantizer, i.e. the global optimum.

■ 2.5.3 Lloyd–Max Algorithm and k-Means Clustering

The necessary conditions given in the previous section suggest an iterative algorithm for the design of locally optimal quantizers known as the Lloyd–Max algorithm. It alternates between application of the nearest neighbor condition and the centroid condition.

Algorithm 2.5.1 (Lloyd–Max). *Given the number of quantization cells k, the probability density $f_{\mathsf{p}}(p)$, and the distortion ρ:*

1. *Choose arbitrary initial representation points $a_1, \ldots, a_k$,*
2. *For all $i = 1, \ldots, k$ set $\mathcal{Q}_i = \{p | \rho(p, a_i) \leq \rho(p, a_{i'})$ for all $i' \neq i\}$,*
3. *For all $i = 1, \ldots, k$ set $a_i = \text{cent}(\mathcal{Q}_i)$,*
4. *Repeat steps 2 and 3 until change in average distortion ϱ is negligible.*

The average distortion decreases or remains the same after each execution of steps 2 and 3. The algorithm is widely used due to its simplicity, effectiveness, and convergence properties [89]. If the three sufficient conditions of [200] are satisfied, then the Lloyd–Max algorithm is guaranteed to converge to a local optimum. The algorithm may be run many times with different initializations to find the global optimum. If the further conditions for unique local optimality given in [200] hold, then the Lloyd–Max algorithm is guaranteed to find the globally minimum quantizer.

As discussed previously in this chapter, hypothesis testing deals with the case when the density function $f_{\mathsf{x},y}(\mathbf{x}, y)$ is given and supervised classification deals with the case when n samples of $(\mathbf{x}, y)$ are given instead. The quantization objective, optimality conditions, and Lloyd–Max algorithm all rely on the probability density function $f_{\mathsf{p}}(p)$. Like hypothesis testing, quantization deals with the case when the distribution is given; the case of finding data partitions when given n samples $p_1, \ldots, p_n$ is a learning problem (analogous to supervised classification) known as unsupervised clustering. In particular, the sample-based version of the quantization problem is known as k-means clustering. The objective is to partition the $n > k$ samples into k clusters $\mathcal{S}_i$ and produce a representation point for each cluster a_i, $i = 1, \ldots, k$, to minimize an average distortion.

Finding the globally optimal clustering is computationally difficult (NP-hard) [5, 55, 76], but a variation of the Lloyd–Max algorithm for a collection of samples, known as the k-means algorithm, finds locally optimal clusterings efficiently. For the sample-based version, the centroid is defined as:

$$\text{cent}(\mathcal{S}) = \arg\min_{a} \frac{1}{|\mathcal{S}|} \sum_{p_j \in \mathcal{S}} \rho(p_j, a). \tag{2.90}$$

The algorithm is as follows.

Algorithm 2.5.2 (k-Means). *Given the number of clusters k, samples $p_1, \ldots, p_n$ with $n > k$, and the distortion ρ:*

1. *Choose arbitrary initial representation points* $a_1, \ldots, a_k$,
2. *For all* $i = 1, \ldots, k$ *set* $\mathcal{S}_i = \{p_j | \rho(p_j, a_i) \leq \rho(p_j, a_{i'})$ *for all* $i' \neq i\}$,
3. *For all* $i = 1, \ldots, k$ *set* $a_i = \text{cent}(\mathcal{S}_i)$,
4. *Repeat steps 2 and 3 until there is no change in the* $\mathcal{S}_i$.

As the number of samples n increases, the sequence of clusterings learned from data converges to the quantizer designed from $f_{\mathsf{p}}(p)$ [88, 169].

■ 2.5.4 Monotonic Convergence

Let $\varrho^*(k)$ denote the expected distortion of an optimal k-point quantizer. This section shows that $\varrho^*(k)$ monotonically converges as k increases. The argument follows the same logic as the argument for training error in supervised classification tending to decrease with increasing classifier complexity. The value of $\varrho^*(k)$ is:

$$\varrho^*(k) = \sum_{i=1}^{k} \int_{\mathcal{Q}_i^*} \rho(p, a_i^*) f_{\mathsf{p}}(p) dp.$$

The optimal k-point quantizer is the solution to the following problem:

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{k} \int_{b_{i-1}}^{b_i} \rho(p, a_i) f_{\mathsf{p}}(p) dp \\ \text{such that} \quad & b_0 = -\infty \\ & b_k = +\infty \\ & b_{i-1} < a_i, \ i = 1, \ldots, k \\ & a_i \leq b_i, \ i = 1, \ldots, k. \end{aligned} \tag{2.91}$$

Add the additional constraint $b_{k-1} = +\infty$ to (2.91), forcing $a_k = +\infty$ and degeneracy of the kth quantization cell. The optimization problem for the k-point quantizer (2.91) with the additional constraint is equivalent to the optimization problem for the $(k-1)$-point quantizer. Thus, the $(k-1)$-point design problem and the k-point design problem have the same objective function, but the $(k-1)$-point problem has an additional constraint. Therefore, $\varrho^*(k-1) \geq \varrho^*(k)$ and $\varrho^*(k)$ is a nonincreasing sequence.

Since the distortion function $\rho(p, q_k(p))$ is greater than or equal to zero, the expected distortion ϱ is also greater than or equal to zero. Since the sequence $\varrho^*(k)$ is nonincreasing and bounded from below, it converges. Expected distortion cannot get worse when more quantization cells are employed. In typical settings, performance always improves with an increase in the number of quantization cells.

Chapter 3

Surface Area of the Decision Boundary

CLASSIFIERS partition the input space of measurements using decision boundaries. A geometric property of a decision boundary in a bounded input space that can be measured is its surface area. The frugality pursued in this chapter is to limit the surface area of classifier decision boundaries. Classifier decision boundaries are points in one-dimensional input spaces and the surface area is a count of the number of points. They are curves in two-dimensional input spaces and the surface area is the curve length. Decision boundaries are surfaces in three-dimensional input spaces and hypersurfaces in higher-dimensional input spaces, with appropriate notions of surface area. Being frugal with decision boundary surface area prevents overfitting. The mathematical approach followed is to set up a margin-based classification problem and solve it using variational level set methods [209, 211].

Variational level set methods, pioneered by Osher and Sethian [143], have found application in fluid mechanics, computational geometry, image processing, computer vision, computer graphics, materials science, and numerous other fields, but have heretofore found little application in statistical learning. This chapter introduces a level set approach to the problem of supervised classification. An implicit level set representation for classifier decision boundaries, a margin-based objective regularized by a surface area penalty, and an Euler–Lagrange contour evolution algorithm for training are proposed.

Several well-developed techniques for supervised discriminative learning exist in the literature, including the perceptron algorithm [164], logistic regression [61], and SVMs [206]. All of these approaches, in their basic form, produce linear decision boundaries. Nonlinear boundaries in the input space can be obtained by mapping the input space to a feature space of higher (possibly infinite) dimension by taking nonlinear functions of the input variables. As discussed in Section 2.2, learning algorithms are then applied to the new higher-dimensional feature space by treating each dimension linearly and they retain the efficiency of the input lower-dimensional space through the use of kernels [176].

As an alternative to kernel methods for generalizing linear methods, the proposal in this chapter is to find nonlinear decision boundary contours directly in the input space.

An energy functional for classification is proposed that is composed of an empirical risk term that uses a margin-based loss function and a complexity term that is the the surface area of the decision boundary. The empirical risk term is standard in many classification methods. What is new in this work is the measurement of decision boundary complexity by surface area and the idea of using variational level set methods for optimization in discriminative learning.

The connection between level set methods (particularly for image segmentation) and classification has been noticed before, but there has been little prior work in this area. Boczko et al. [27] only hint at the idea of using variational level set methods for classification. Tomczyk and Szczepaniak [197] do not consider fully general input spaces. Specifically, samples in the training and test sets must be pixels in an image with the measurement vector containing the spatial index of the pixel along with other variables. Cai and Sowmya [31] do consider general input spaces, but have a very different energy functional than the proposed margin-based loss functional. Theirs is based on counts of training samples in grid cells and is similar to the region-based functional used in image segmentation that separates the mean values of the image foreground and background. Their learning is also based on one-class classification rather than standard discriminative classification, which is the framework followed in this thesis. Yip et al. [223] use variational level set methods for density-based clustering in general input spaces, rather than for learning classifiers.

Cremers et al. [48] dichotomize image segmentation approaches into those that use spatially continuous representations and those that use spatially discrete representations, with level set methods being the main spatially continuous approaches. There have been methods using discrete representations that bear some ties to the methods introduced in this chapter. An example of a spatially discrete approach uses normalized graph cuts [182], a technique that has been used extensively in unsupervised learning for general features unrelated to images as well. Normalized decision boundary surface area is implicitly penalized in this discrete setting. Geometric notions of complexity in supervised classification tied to decision boundary surface area have been suggested by Ho and Basu [94], but also defined in a discrete way related to graph cuts. In contrast, the continuous formulation employed here using level sets involves very different mathematical foundations, including explicit minimization of a criterion involving surface area. Moreover, the continuous framework—and in particular the natural way in which level set functions enter into the criterion—lead to new gradient descent algorithms to determine optimal decision boundaries. By embedding the criterion in a continuous setting, the surface area complexity term is defined intrinsically rather than being defined in terms of the graph of available training samples.

There are some other methods in the literature for finding nonlinear decision boundaries directly in the input space related to image segmentation, but these methods use neither contour evolution for optimization, nor the surface area of the decision boundary as a complexity term, as in the level set classification method proposed in this chapter. A connection is drawn between classification and level set image segmenta-

tion in [180, 219], but the formulation is through decision trees, not contour evolution. Tomczyk et al. [198] present a simulated annealing formulation given the name adaptive potential active hypercontours for finding nonlinear decision boundaries in both the classification and clustering problems. Pölzlbauer et al. [154] construct nonlinear decision boundaries in the input space from connected linear segments. In some ways, their approach is similar to active contours methods in image segmentation such as snakes that do not use the level set representation: changes in topology of the decision boundary in the optimization are difficult to handle. (As mentioned in Section 2.3, the implicit level set representation takes care of topology changes naturally.)

The theory of classification with Lipschitz functions is discussed by von Luxburg and Bousquet [121]. As mentioned in Section 2.3, level set functions are Lipschitz functions and the signed distance function specifically has a unit Lipschitz constant. The Lipschitz constant is minimized in [121], whereas the Lipschitz constant is fixed in the formulation proposed in this chapter. The von Luxburg and Bousquet [121] formulation requires the specification of a subspace of Lipschitz functions over which to optimize in order to prevent overfitting, but does not resolve the question of how to select this subspace. Being frugal with the surface area of the decision boundary provides a natural specification for subspaces of signed distance functions. The maximum allowable surface area parameterizes nested subspaces.

The chapter is organized as follows. Section 3.1 details geometric level set classification in the binary case, describing the objective to be minimized, the contour evolution to perform the minimization, as well as illustrative examples. Section 3.2 goes over multicategory level set classification. A level set implementation using radial basis functions is described in Section 3.3; that implementation is used to compare the classification test performance of geometric level set classification to the performance of several other classifiers. Theoretical analysis of the level set classifier is provided in Section 3.4, including characterizations of consistency and complexity. A variational level set method for both margin-based classification and feature subset selection is described in Section 3.5. A summary of the chapter is provided in Section 3.6.

■ 3.1 Binary Classification Using Geometric Level Sets

The margin-based approach to supervised classification is theoretically well-founded and has excellent empirical performance on a variety of datasets [176]. A margin-based classifier with a new geometric regularization term is proposed in this section. The new classifier is termed the geometric level set (GLS) classifier.

■ 3.1.1 Classification Functional with Surface Area Regularization

Recall the margin-based classification objective (2.48) from Section 2.2:

$$\mathrm{L}(\varphi) = \sum_{j=1}^{n} \ell(y_j \varphi(\mathbf{x}_j)) + \lambda \, \mathrm{J}(\varphi), \tag{3.1}$$

where the training dataset $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ has measurements $\mathbf{x}_j \in \Omega \subset \mathbb{R}^D$ and class labels $y_j \in \{-1, +1\}$. The margin-based classifier is $\hat{y}(\mathbf{x}) = \text{sign}(\varphi(\mathbf{x}))$, where φ is a decision function. The functional J is a regularization term and ℓ is a margin-based loss function.

Also recall the generic form of energy functionals minimized using variational level set methods (2.74) given in Section 2.3:

$$\text{L}(\mathcal{C}) = \int_{\mathcal{R}_-} g_r(\mathbf{x}) d\mathbf{x} + \lambda \oint_{\mathcal{C}} g_b(\mathcal{C}(\mathbf{s})) d\mathbf{s}, \tag{3.2}$$

where the contour $\mathcal{C}$ is the set of $\mathbf{x} \in \Omega \subset \mathbb{R}^D$ where the signed distance function $\varphi(\mathbf{x}) = 0$. The region $\mathcal{R}_-$ is the set of $\mathbf{x}$ where the signed distance function is negative. Energy functionals may also be integrals over the region $\mathcal{R}_+$, the subset of Ω where the signed distance function is positive.

The main idea of the GLS classifier is to combine margin-based classification with variational level set methods. Toward that end, the decision function $\varphi(\mathbf{x})$ is taken to be a signed distance function defined over Ω, and $\mathcal{C}$ is the decision boundary. With linear margin-based classifiers, including the original primal formulation of the SVM (2.38), the concept of margin is proportional to Euclidean distance from the decision boundary in the input space Ω. With kernel methods, this relationship to distance is in the implicit feature space, but the relationship in the input space is lost; the quantity referred to as the margin, $y\varphi(\mathbf{x})$, is not the same as distance from $\mathbf{x}$ to the decision boundary in Ω. As discussed by Akaho [2], oftentimes it is of interest that the definition of margin truly be distance to the decision boundary in the input space. With the signed distance function representation, the margin $y\varphi(\mathbf{x})$ is equivalent to Euclidean distance from $\mathcal{C}$ and hence is a satisfying nonlinear generalization to linear margin-based methods.

Furthermore, the regularization term $\text{J}(\varphi)$ in (3.1) is taken to be a boundary-based energy functional. Specifically, it is proposed that the regularization term be the surface area of the decision boundary, that is:

$$\text{J}(\varphi) = \oint_{\varphi=0} d\mathbf{s}. \tag{3.3}$$

The training objective to be minimized for the GLS classifier is then:

$$\text{L}(\varphi) = \sum_{j=1}^{n} \ell(y_j \varphi(\mathbf{x}_j)) + \lambda \oint_{\mathcal{C}} d\mathbf{s}, \tag{3.4}$$

with $\varphi(\mathbf{x})$ a signed distance function.

The expression (3.4) takes the form of a variational level set energy functional. In particular, the surface area regularization is a boundary-based functional with $g_b = 1$, and the margin-based loss term can be expressed as the sum of region-based functionals over $\mathcal{R}_-$ and $\mathcal{R}_+$ with $g_r(\mathbf{x})$ incorporating $\ell(y_j \varphi(\mathbf{x}_j))$. Therefore, the GLS classifier may be learned from training data by contour evolution.

■ 3.1.2 Contour Evolution for Classifier Learning

As just noted, (3.4) can be minimized using contour evolution. Recall the contour evolution level set function update equation (2.75) from Section 2.3 to minimize the generic energy (3.2):

$$\varphi_t(\mathbf{x}) = -g_r(\mathbf{x})\nabla\varphi(\mathbf{x}) - \lambda\left[g_b(\mathbf{x})\nabla^2\varphi(\mathbf{x}) + (\nabla g_b(\mathbf{x}))\cdot(\nabla\varphi(\mathbf{x}))\right]\nabla\varphi(\mathbf{x}), \tag{3.5}$$

Also recall that updating signed distance functions using (3.5) generally results in a level set function that is not a signed distance function; to recover the signed distance function with the same zero level set, the following update is applied:

$$\varphi_t(\mathbf{x}) = \text{sign}(\varphi(\mathbf{x}))(1 - \|\nabla\varphi(\mathbf{x})\|). \tag{3.6}$$

Applying (3.5) to the margin-based classification objective with decision boundary surface area regularization, the gradient descent flow is found to be:

$$\varphi_t(\mathbf{x})|_{\mathbf{x}=\mathbf{x}_j} = \begin{cases} \ell(y_j\varphi(\mathbf{x}_j))\nabla\varphi(\mathbf{x}_j) - \lambda\nabla^2\varphi(\mathbf{x}_j)\nabla\varphi(\mathbf{x}_j), & \varphi(\mathbf{x}_j) < 0 \\ -\ell(y_j\varphi(\mathbf{x}_j))\nabla\varphi(\mathbf{x}_j) - \lambda\nabla^2\varphi(\mathbf{x}_j)\nabla\varphi(\mathbf{x}_j), & \varphi(\mathbf{x}_j) > 0 \end{cases}. \tag{3.7}$$

The two cases in (3.7) arise because $\sum_{j=1}^{n}\ell(y_j\varphi(\mathbf{x}_j))$ is defined over both $\mathcal{R}_-$ and $\mathcal{R}_+$. As discussed in Section 2.3, the gradient direction has opposite signs when the region-based functional is defined over $\mathcal{R}_-$ and $\mathcal{R}_+$.

Maintaining the signed distance property of the level set function using (3.6) is more important here than with functionals employed in other level set applications such as image segmentation because (3.4) uses the magnitude of $\varphi(\mathbf{x})$, not just its sign. Note that the surface area of the decision boundary is never computed in doing the contour evolution. Computing the value of the surface area is oftentimes intractable and only its gradient descent flow is required.

■ 3.1.3 Examples

Two synthetic examples are now presented to illustrate the GLS classifier. In both examples, there are $n = 1000$ samples in the training set with $D = 2$. The first example has 502 samples with label $y_j = -1$ and 498 samples with label $y_j = +1$ and is separable by an elliptical decision boundary. The second example has 400 samples with label $y_j = -1$ and 600 samples with label $y_j = +1$ and is not separable by a simple shape, but has the -1 labeled samples in a strip.

In these two examples, in the other examples in the rest of the chapter, and in the performance results of Section 3.3, the logistic loss function is used for ℓ in the margin-based classification objective. In these two examples, the surface area penalty has weight $\lambda = 0.5$; the value $\lambda = 0.5$ is a default parameter value that gives good performance with a variety of datasets regardless of their dimensionality D and can be used if one does not wish to optimize λ using cross-validation. Classification error and classifier complexity as a function of λ are shown later in the chapter.

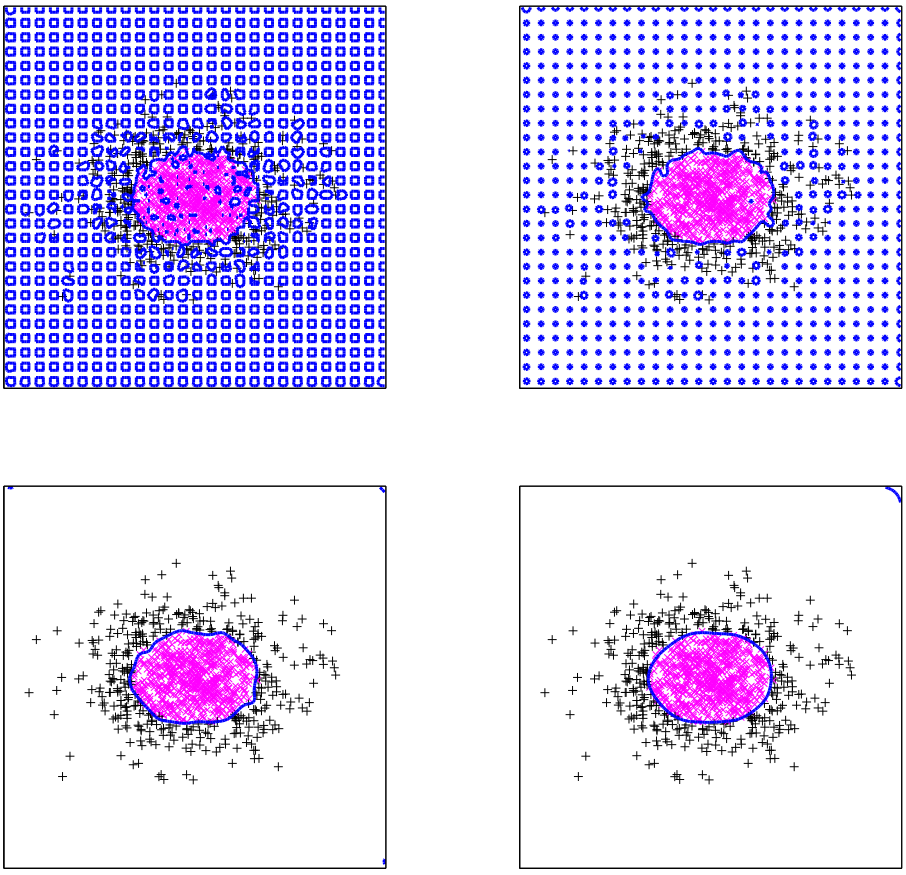


Figure 3.1. Contour evolution iterations for an example training set with $\lambda = 0.5$ proceeding in raster scan order from top left to bottom right. The magenta $\times$ markers indicate class label -1 and the black $+$ markers indicate class label $+1$. The blue line is the decision boundary.

Contour evolution minimization requires an initial decision boundary. In the portion of Ω where there are no training samples, the initialization used here sets the decision boundary to be a uniform grid of small components; this small seed initialization is common in level set methods. In the part of Ω where there are training samples, the locations and labels of the training samples are used to set the initial decision boundary. A positive value is assigned to the initial signed distance function in locations of positively labeled samples and a negative value in locations of negatively labeled samples. The initial decision boundaries for the two examples are shown in the top left panels of Figure 3.1 and Figure 3.2.

Two intermediate iterations and the final decision boundary are also shown in Fig-

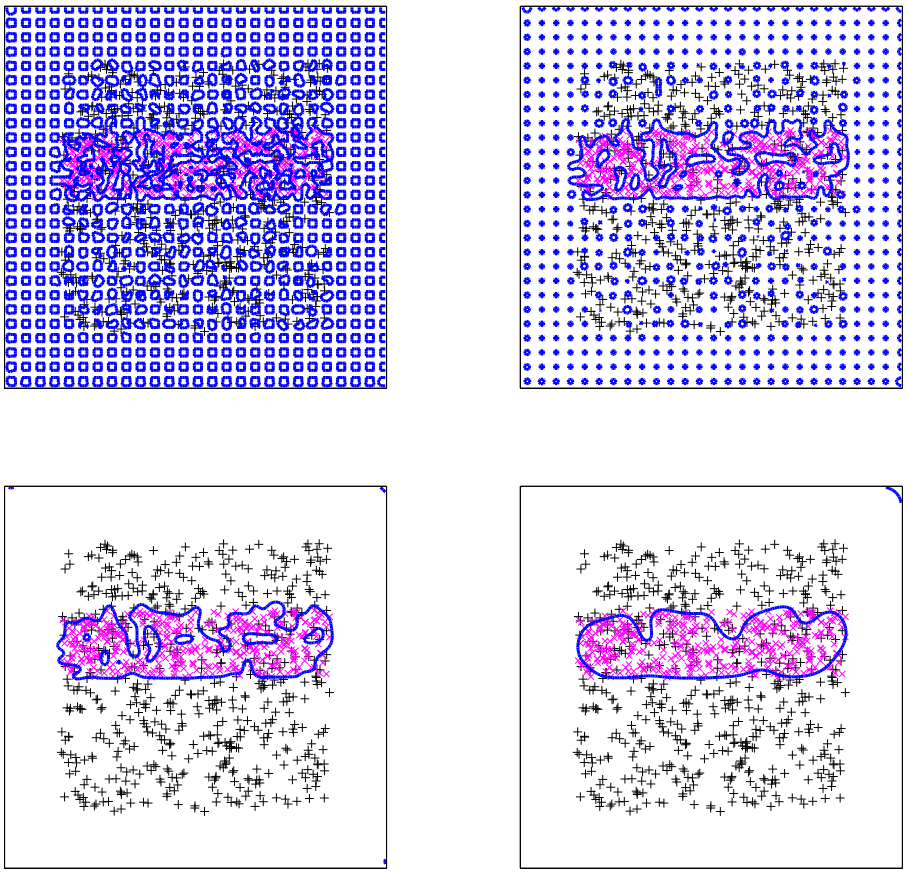
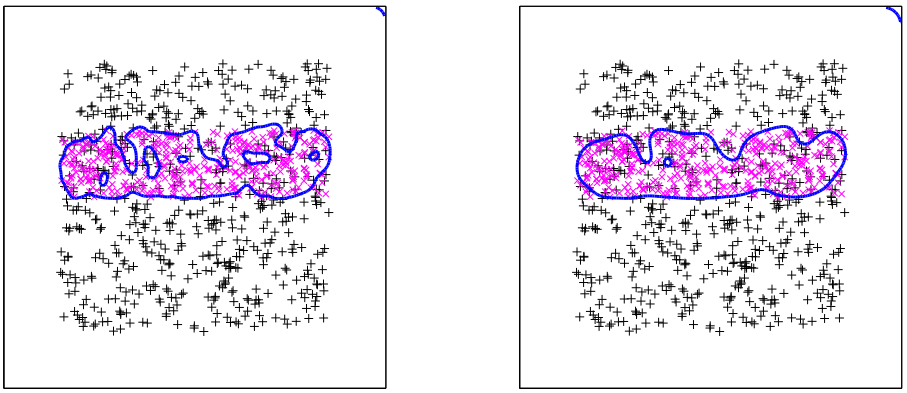


Figure 3.2. Contour evolution iterations for an example training set with $\lambda = 0.5$ proceeding in raster scan order from top left to bottom right. The magenta $\times$ markers indicate class label -1 and the black $+$ markers indicate class label $+1$. The blue line is the decision boundary.

ure 3.1 and Figure 3.2. Solutions are as expected: an elliptical decision boundary and a strip-like decision boundary have been recovered. In the final decision boundaries of both examples, there is a small curved piece of the decision boundary in the top right corner of Ω where there are no training samples. This piece is an artifact of the initialization and the regularization term, and does not affect classifier performance. (The corner piece of the decision boundary is a minimal surface, a surface of zero mean curvature, which is a critical point of the surface area regularization functional (3.3), but not the global minimum. It is not important, assuming that the training set is representative.)

For a visual comparison of the effect of the surface area penalty weight, in Figure 3.3



(a) (b)

Figure 3.3. Solution decision boundaries for an example training set with (a) $\lambda = 0.005$ and (b) $\lambda = 0.05$. The magenta $\times$ markers indicate class label -1 and the black $+$ markers indicate class label $+1$. The blue line is the decision boundary.

the solution decision boundaries of the GLS classifier are shown for two other values of λ, 0.005 and 0.05, with the dataset used in the example of Figure 3.2. As can be seen in comparing this figure with the bottom right panel of Figure 3.2, the smaller the value of λ, the longer and more tortuous the decision boundary. Small values of λ, which correspond to large decision boundary surface areas, may lead to overfitting.

This section has described the basic method for nonlinear margin-based binary classification based on variational level set methods and illustrated its operation on two synthetic datasets. The following sections build upon this core binary GLS classifier in several directions, including multicategory classification, theoretical analysis, and joint feature subset selection. Classification performance on several benchmark datasets along with comparison to other methods is also given.

■ 3.2 Multicategory Classification Using Geometric Level Sets

Many interesting applications of classification contain more than two classes. For the multicategory classification problem with $M > 2$ classes, binary margin-based classification methods are typically extended using the *one-against-all* construction [97]. The one-against-all scheme represents the classifier with M decision functions that each distinguish one class from all of the other classes. In this section, a more frugal representation of multicategory margin-based classification is proposed that uses $\lceil \log_2 M \rceil$ decision functions. A collection of $\log_2 M$ level set functions can implicitly specify M regions using a binary encoding akin to a Venn diagram [214]. This proposed logarithmic multicategory classification is new, as there does not seem to be any M-ary classifier

representation in the statistical learning literature utilizing as few as $\lceil \log_2 M \rceil$ decision functions. Methods that combine binary classifier outputs using error-correcting codes make use of a logarithmic number of binary classifiers with a larger multiplicative constant, such as $\lceil 10 \log M \rceil$ or $\lceil 15 \log M \rceil$ [4, 163]. The M-ary Bayesian hypothesis testing decision rule (2.28) discussed in Section 2.1.5 employs $M-1$ decision functions. The GLS classifier is extended for multicategory problems in this section.[1]

■ 3.2.1 Multicategory Margin-Based Functional and Contour Evolutions

The classification problem considered in Section 3.1 has binary-valued class labels, whereas the class labels in this section take one of M values. As in the binary case, n training samples are given with measurement vectors $\mathbf{x}_j \in \Omega \subset \mathbb{R}^D$. For the multicategory case, the class labels are $y_j \in \{1, \ldots, M\}$. The classifier $\hat{y}(\mathbf{x})$ is a mapping from Ω to $\{1, \ldots, M\}$.

As in Section 3.1, a margin-based objective regularized by the surface area of the decision boundaries is proposed, with training through variational level set methods. The decision boundaries are represented using $m = \lceil \log_2 M \rceil$ signed distance functions $\varphi^{(1)}(\mathbf{x}), \ldots, \varphi^{(m)}(\mathbf{x})$. These signed distance functions can represent 2^m regions $\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_{2^m}$ through a binary encoding scheme [214]. The regions are defined as follows:

$$\begin{aligned}
&\varphi^{(1)}(\mathbf{x}) < 0,\ \ldots,\ \varphi^{(m-1)}(\mathbf{x}) < 0,\ \varphi^{(m)} < 0, \quad \mathbf{x} \in \mathcal{R}_1, \\
&\varphi^{(1)}(\mathbf{x}) < 0,\ \ldots,\ \varphi^{(m-1)}(\mathbf{x}) < 0,\ \varphi^{(m)} > 0, \quad \mathbf{x} \in \mathcal{R}_2, \\
&\qquad\qquad\qquad \vdots \\
&\varphi^{(1)}(\mathbf{x}) > 0,\ \ldots,\ \varphi^{(m-1)}(\mathbf{x}) > 0,\ \varphi^{(m)} > 0, \quad \mathbf{x} \in \mathcal{R}_{2^m}.
\end{aligned}$$

As discussed in [228, 229], the same margin-based loss functions used in the binary case, such as the hinge loss and logistic loss, may be used in defining multicategory margin-based classification objectives. In binary classification, the special encoding $y \in \{-1, +1\}$ allows $y\varphi(\mathbf{x})$ to be the argument to the margin-based loss function, because multiplication by the class label value makes $y\varphi(\mathbf{x})$ positive for correct classifications and negative for incorrect classifications, and preserves the magnitude as the distance to the decision boundary. The argument to the margin-based loss function for multicategory classification, with the proposed representation using a logarithmic number of signed distance functions, must be specified with care. It is proposed that the argument to the margin-based loss function be through functions $\psi_y(\mathbf{x})$, which are

[1] It is certainly possible to use one-against-all with the binary GLS classifier. In fact, there are M-ary level set methods that use M level set functions [144, 171], but they are less frugal than the approach followed in this section.

also specified through a binary encoding:

$$\begin{aligned}
\psi_1(\mathbf{x}) &= \max\left\{+\varphi^{(1)}(\mathbf{x}), \ldots, +\varphi^{(m-1)}(\mathbf{x}), +\varphi^{(m)}(\mathbf{x})\right\}, \\
\psi_2(\mathbf{x}) &= \max\left\{+\varphi^{(1)}(\mathbf{x}), \ldots, +\varphi^{(m-1)}(\mathbf{x}), -\varphi^{(m)}(\mathbf{x})\right\}, \\
&\vdots \\
\psi_{2^m}(\mathbf{x}) &= \max\left\{-\varphi^{(1)}(\mathbf{x}), \ldots, -\varphi^{(m-1)}(\mathbf{x}), -\varphi^{(m)}(\mathbf{x})\right\}.
\end{aligned}$$

To extend the GLS classifier to the multicategory case, the surface area regularization term must also be specified. The full set of decision boundaries is the union of the zero level sets of $\varphi^{(1)}(\mathbf{x}), \ldots, \varphi^{(m)}(\mathbf{x})$. By the inclusion-exclusion principle, the surface area of the full set of decision boundaries can be approximated by the sum of the surface areas of the zero level sets of the m signed distance functions. This approximation is quite good in practice because the zero level sets of different signed distance functions rarely intersect.

Combining the multicategory margin and surface area regularization, the M-ary GLS classification energy functional that is proposed is:

$$\mathrm{L}(\varphi^{(1)}, \ldots, \varphi^{(m)}) = \sum_{j=1}^{n} \ell(\psi_{y_j}(\mathbf{x}_j)) + \frac{\lambda}{m} \sum_{k=1}^{m} \oint_{\varphi^{(k)}=0} d\mathbf{s}. \tag{3.8}$$

The gradient descent flows for the m signed distance functions are

$$\begin{aligned}
\varphi_t^{(1)}(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}_j} &= \begin{cases} \ell(\psi_{y_j}(\mathbf{x}_j))\nabla\varphi^{(1)}(\mathbf{x}_j) - \frac{\lambda}{m}\nabla^2\varphi^{(1)}(\mathbf{x}_j)\nabla\varphi^{(1)}(\mathbf{x}_j), & \varphi^{(1)}(\mathbf{x}_j) < 0 \\ -\ell(\psi_{y_j}(\mathbf{x}_j))\nabla\varphi^{(1)}(\mathbf{x}_j) - \frac{\lambda}{m}\nabla^2\varphi^{(1)}(\mathbf{x}_j)\nabla\varphi^{(1)}(\mathbf{x}_j), & \varphi^{(1)}(\mathbf{x}_j) > 0 \end{cases} \\
&\vdots \\
\varphi_t^{(m)}(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}_j} &= \begin{cases} \ell(\psi_{y_j}(\mathbf{x}_j))\nabla\varphi^{(m)}(\mathbf{x}_j) - \frac{\lambda}{m}\nabla^2\varphi^{(m)}(\mathbf{x}_j)\nabla\varphi^{(m)}(\mathbf{x}_j), & \varphi^{(m)}(\mathbf{x}_j) < 0 \\ -\ell(\psi_{y_j}(\mathbf{x}_j))\nabla\varphi^{(m)}(\mathbf{x}_j) - \frac{\lambda}{m}\nabla^2\varphi^{(m)}(\mathbf{x}_j)\nabla\varphi^{(m)}(\mathbf{x}_j), & \varphi^{(m)}(\mathbf{x}_j) > 0 \end{cases}.
\end{aligned}$$

In the case $M = 2$ and $m = 1$, the energy functional and gradient flow revert back to binary level set classification described in Section 3.1. The proposed multicategory classifier is different from one-against-all both because it treats all M classes simultaneously in the objective and because the decision regions are represented by a logarithmic rather than linear number of decision functions. Zou et al. [228] also treat all M classes simultaneously in the objective, but their multicategory kernel machines use M decision functions.

■ 3.2.2 Example

An example showing multicategory level set classification with $M = 4$ and $D = 2$ is now given. The dataset has 250 samples for each of the four class labels $y_j = 1$, $y_j = 2$,

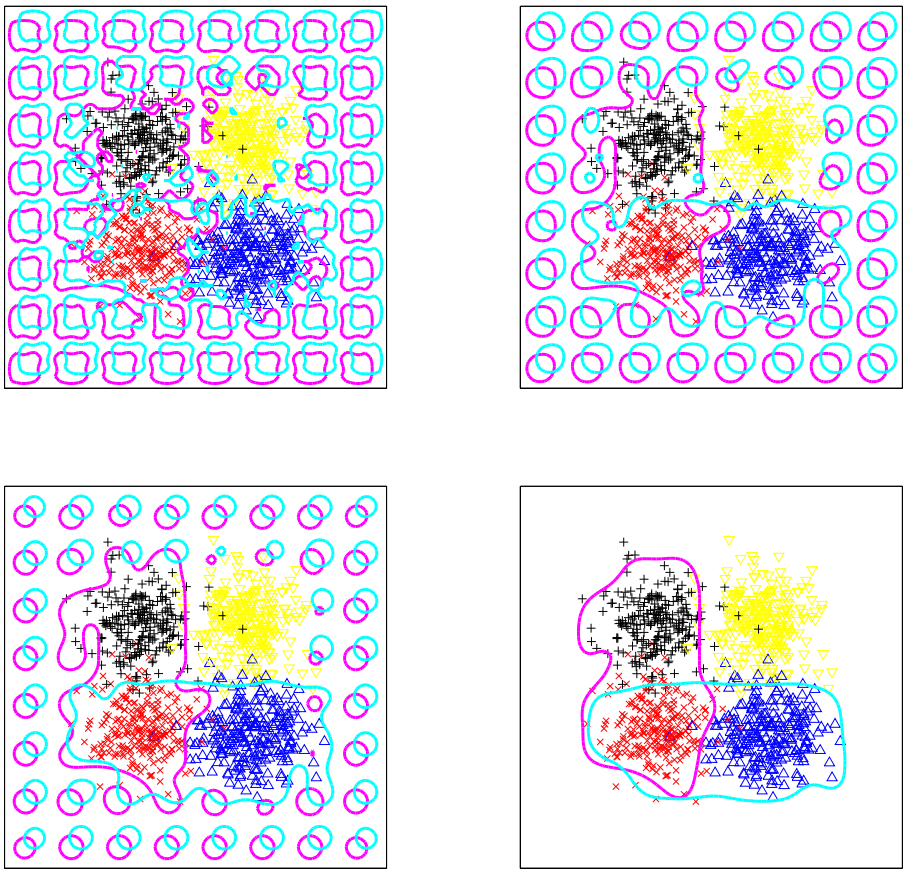


Figure 3.4. Contour evolution iterations for multicategory classification with $\lambda = 0.5$ proceeding in raster scan order from top left to bottom right. The red $\times$ markers indicate class label 1, the black $+$ markers indicate class label 2, the blue $\triangle$ markers indicate class label 3, and the yellow $\triangledown$ markers indicate class label 4. The magenta and cyan lines are the zero level sets of the $m = 2$ signed distance functions and together make up the decision boundary.

$y_j = 3$, and $y_j = 4$. The classes are not perfectly separable by simple boundaries. With four classes, there are $m = 2$ signed distance functions.

Figure 3.4 shows the evolution of the two contours, the magenta and cyan curves. The same type of initialization described in Section 3.1.3 is employed; here the small seeds in the part of Ω not containing samples are offset from each other for the two different signed distance functions. The final decision region for class $y = 1$, $\mathcal{R}_1$, is the portion of Ω inside both the magenta and cyan curves, and coincides with the training samples having class label 1. The final $\mathcal{R}_2$ is the region inside the magenta curve but outside the cyan curve; the final $\mathcal{R}_3$ is the region inside the cyan curve but outside the

magenta curve, and the final $\mathcal{R}_4$ is outside both curves. The final decision boundaries are fairly smooth and partition the space with small training error.

■ 3.3 Geometric Level Set Classifier Implementation and Performance

The level set representation of classifier decision boundaries, the surface area regularization term, and the logarithmic multicategory classification scheme are not only interesting theoretically, but also practically. In this section, the classification performance of the GLS classifier is compared with many classifiers used in practice on several binary and multicategory datasets from the UCI Repository [10], and found to be competitive.

Level set methods are usually implemented on a discretized grid, i.e., the values of the level set function are maintained and updated on a grid. In physics and image processing applications, it nearly always suffices to work in two- or three-dimensional spaces. In classification problems, however, the input data space can be high-dimensional. Implementation of level set methods for large input space dimension becomes cumbersome due to the need to store and update a grid of that large dimension. One way to address this practical limitation is to represent the level set function by a superposition of RBFs instead of on a grid [35, 78, 186]. This implementation strategy is followed in obtaining classification results.

■ 3.3.1 Radial Basis Function Level Set Method

There have been many developments in level set methods since the original work of Osher and Sethian [143]. One development in particular is to represent the level set function by a superposition of RBFs instead of on a grid [35, 78, 186]. Grid-based representation of the level set function is not amenable to classification in high-dimensional input spaces because the memory and computational requirements are exponential in the dimension of the input space. A nonparametric RBF representation, however, is tractable for classification. An RBF level set method is used in this section to minimize the energy functionals (3.4) and (3.8) for binary and multicategory margin-based classification. The method is most similar to that described by Gelas et al. [78] for image processing.

The starting point of the RBF level set approach is describing the level set function $\varphi(\mathbf{x})$ via a strictly positive definite[2] RBF $\mathrm{K}(\cdot)$ as follows:

$$\varphi(\mathbf{x}) = \sum_{j=1}^{n} \alpha_j \, \mathrm{K}\left(\|\mathbf{x} - \mathbf{x}_j\|\right). \tag{3.9}$$

The zero level set of $\varphi(\mathbf{x})$ defined in this way is the contour $\mathcal{C}$. For the classification

[2] A more complete discussion including conditionally positive definite RBFs would add a polynomial term to (3.9), to span the null space of the RBF [216].

problem, the centers $\mathbf{x}_j$ are taken to be the data vectors of the training set.[3] Then, constructing an $n \times n$ matrix $\mathbf{H}$ with elements $h_{jk} = \mathrm{K}(\|\mathbf{x}_j - \mathbf{x}_k\|)$, and letting $\boldsymbol{\alpha}$ be the vector of coefficients in (3.9):

$$\begin{bmatrix} \varphi(\mathbf{x}_1) \\ \vdots \\ \varphi(\mathbf{x}_n) \end{bmatrix} = \mathbf{H}\boldsymbol{\alpha}.$$

To minimize an energy functional of $\mathcal{C}$, the level set optimization is over the coefficients $\boldsymbol{\alpha}$ with $\mathbf{H}$ fixed. In order to perform contour evolution with the RBF representation, a time parameter t is introduced like in Section 2.3, giving:

$$\mathbf{H}\frac{d\boldsymbol{\alpha}}{dt} = \begin{bmatrix} \varphi_t(\mathbf{x})|_{\mathbf{x}=\mathbf{x}_1} \\ \vdots \\ \varphi_t(\mathbf{x})|_{\mathbf{x}=\mathbf{x}_n} \end{bmatrix}. \tag{3.10}$$

For the binary margin-based classification problem with surface area regularization, the gradient flow (3.7) is substituted into the right side of (3.10). For the multicategory classification problem, there are m level set functions as discussed in Section 3.2 and each one has a gradient flow to be substituted into an expression like (3.10).

The iteration for the contour evolution indexed by i is then:

$$\boldsymbol{\alpha}(i+1) = \boldsymbol{\alpha}(i) - \tau\mathbf{H}^{-1} \begin{bmatrix} \varphi_t(\mathbf{x};i)|_{\mathbf{x}=\mathbf{x}_1} \\ \vdots \\ \varphi_t(\mathbf{x};i)|_{\mathbf{x}=\mathbf{x}_n} \end{bmatrix}, \tag{3.11}$$

where τ is a small step size and $\varphi(\mathbf{x};i)$ comes from $\boldsymbol{\alpha}(i)$. The vector $\boldsymbol{\alpha}$ is normalized according to the ℓ_1-norm after every iteration. The RBF-represented level set function is not a signed distance function. However, as discussed by Gelas et al. [78], normalizing the coefficient vector $\boldsymbol{\alpha}$ with respect to the ℓ_1-norm after every iteration of (3.11) has a similar effect as reinitializing the level set function as a signed distance function. The Lipschitz constant of the level set function is constrained by this normalization, which is important because the magnitude of the level set function appears in the argument of the margin-based loss function.

The RBF level set approach is similar to the SVM with RBF kernel discussed in Section 2.2.3 in the sense that the decision function is represented by a linear combination of RBFs. However, the SVM and other kernel methods in the literature minimize a reproducing kernel Hilbert space squared norm for regularization, whereas the GLS classifier minimizes decision boundary surface area for regularization. The regularization term and consequently inductive bias of the GLS classifier is new and different

[3] It is not required that the RBFs be collocated with the training samples, or even that the number of RBFs be as many as the number of training samples. An extension would consider optimizing the number and placement of the RBFs to further reduce complexity. This direction is considered in [186].

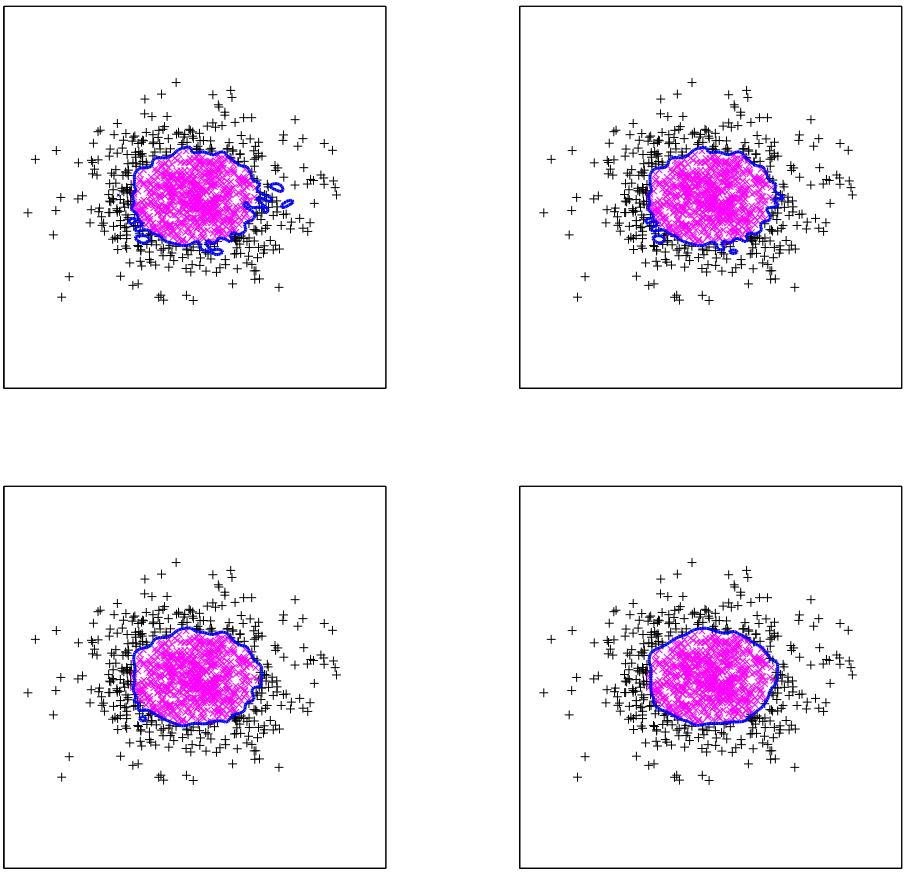


Figure 3.5. Contour evolution iterations with RBF implementation and $\lambda = 0.5$ for example training set proceeding in raster scan order from top left to bottom right. The magenta $\times$ markers indicate class label -1 and the black $+$ markers indicate class label $+1$. The blue line is the decision boundary.

compared to existing kernel methods. The solution decision boundary is the zero level set of a function of the form given in (3.9). Of course this representation does not capture all possible functions, but, given that a number of RBFs equal to the number of training samples is used, the granularity of this representation is well-matched to the data. This is similar to the situation found in other contexts such as kernel machines using RBFs.

The initialization for the decision boundary used here has $\boldsymbol{\alpha} = n(\mathbf{H}^{-1}\mathbf{y})/\|\mathbf{H}^{-1}\mathbf{y}\|_1$, where $\mathbf{y}$ is a vector of the n class labels in the training set. Figure 3.5 shows this initialization and following RBF-implemented contour evolution on the elliptically-separable dataset presented in Section 3.1.3. The initial decision boundary is tortuous. It is

smoothed out by the surface area penalty during the course of the contour evolution, thereby improving the generalization of the learned classifier as desired. To initialize the m vectors $\boldsymbol{\alpha}$ in M-ary classification, m vectors of length n containing positive and negative ones constructed from the binary encoding are used instead of $\mathbf{y}$.

■ 3.3.2 Classification Results

Classification performance results on benchmark datasets from the UCI Machine Learning Repository [10] are given for the GLS classifier and compared to the performance of several other classifiers, with the conclusion that GLS classification is a competitive technique. Tenfold cross-validation classification error performance with RBF level set implementation is presented on four binary datasets: Pima Indians Diabetes ($n = 768$, $D = 8$), Wisconsin Diagnostic Breast Cancer ($n = 569$, $D = 30$), BUPA Liver Disorders ($n = 345$, $D = 6$) and Johns Hopkins University Ionosphere ($n = 351$, $D = 34$), and four multicategory datasets: Wine Recognition ($n = 178$, $M = 3$, $D = 13$), Iris ($n = 150$, $M = 3$, $D = 4$), Glass Identification ($n = 214$, $M = 6$, $D = 9$), and Image Segmentation ($n = 2310$, $M = 7$, $D = 19$). For the binary datasets, there is $m = 1$ level set function, for the wine and iris datasets $m = 2$ level set functions, and for the glass and segmentation datasets $m = 3$ level set functions.

Before training the classifier, the data is scaled and shifted so that each of the input dimensions has zero mean and unit variance. The RBF:

$$\mathrm{K}(\|\mathbf{x} - \mathbf{x}_j\|) = e^{-\|\mathbf{x}-\mathbf{x}_j\|^2}$$

is used along with the logistic loss function, $\tau = 1/m$, and the initialization $\boldsymbol{\alpha} = n(\mathbf{H}^{-1}\mathbf{y})/\|\mathbf{H}^{-1}\mathbf{y}\|_1$. The step size τ is scaled by the number of level set functions for reasons of stability.

First, classification error is examined as a function of the regularization weight of the surface area penalty λ. Figure 3.6 shows the tenfold cross-validation training and test errors for the Pima and WDBC datasets; other datasets yield similar plots. The plots show evidence of the structural risk minimization principle described in Section 2.2.2.[4] For small λ (corresponding to large surface area), the model class is too complex and although the training error is zero, the test error is not minimal due to overfitting. For large λ, the model class is not complex enough; the training error is large and the test error is not minimal due to underfitting. There is an intermediate value of λ that achieves the minimal test error. However, the test error is fairly insensitive to the value of λ. The test error does not change much over the plotted range.

Table 3.1 and Figure 3.7 report the tenfold cross-validation test error (as a percentage) on the eight datasets and compare the performance to nine other classifiers.[5] On

[4]The horizontal axis in Figure 3.6 is shown with increasing λ from left to right and thus decreasing complexity, which is opposite of the horizontal axis in Figure 2.1 with increasing complexity from left to right.

[5]For lower-dimensional datasets (up to about $D = 12$), it is possible to use optimal dyadic decision trees [24, 180]. The results using such trees are not significantly better than those obtained using

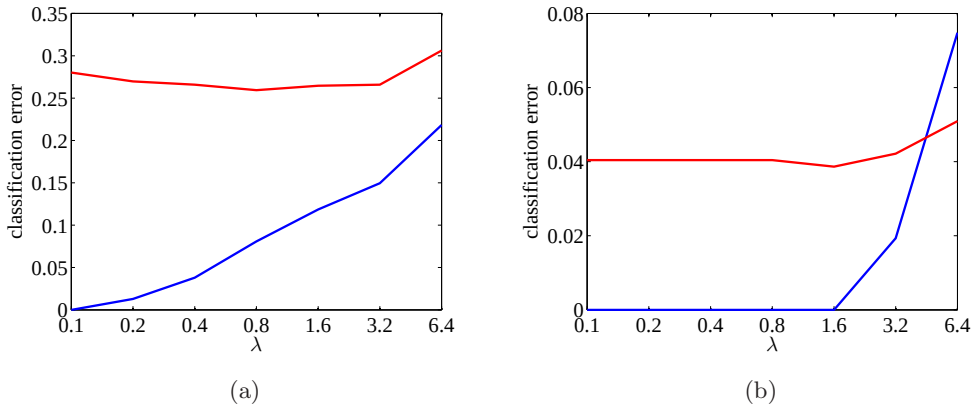


Figure 3.6. Tenfold cross-validation training error (blue line) and test error (red line) for the (a) Pima, and (b) WDBC datasets as a function of the regularization parameter λ on a logarithmic scale.

Dataset (M, D)	NB	BN	kNN	C4.4	C4.5	NBT	SVM	RBN	LLS	GLS
Pima $(2, 8)$	23.69	25.64	27.86	27.33	26.17	25.64	22.66	24.60	29.94	25.94
WDBC $(2, 30)$	7.02	4.92	3.68	7.20	6.85	7.21	2.28	5.79	6.50	4.04
Liver $(2, 6)$	44.61	43.75	41.75	31.01	31.29	33.87	41.72	35.65	37.39	37.61
Ionos. $(2, 34)$	17.38	10.54	17.38	8.54	8.54	10.27	11.40	7.38	13.11	13.67
Wine $(3, 13)$	3.37	1.11	5.00	6.14	6.14	3.37	1.67	1.70	5.03	3.92
Iris $(3, 4)$	4.00	7.33	4.67	4.00	4.00	6.00	4.00	4.67	3.33	6.00
Glass $(6, 9)$	50.52	25.24	29.89	33.68	34.13	24.78	42.49	34.50	38.77	36.95
Segm. $(7, 19)$	18.93	9.60	5.20	4.27	4.27	5.67	8.07	13.07	14.40	4.03

Table 3.1. Tenfold cross-validation error percentage of GLS classifier with RBF level set implementation on several datasets compared to error percentages of various other classifiers reported in [31]. The other classifiers are: naïve Bayes classifier (NB), Bayes net classifier (BN), k-nearest neighbor with inverse distance weighting (kNN), C4.4 decision tree (C4.4), C4.5 decision tree (C4.5), naïve Bayes tree classifier (NBT), SVM with polynomial kernel (SVM), radial basis function network (RBN), and learning level set classifier (LLS) of Cai and Sowmya [31].

each of the ten folds, λ is set using cross-validation. Specifically, fivefold cross-validation is performed on the nine tenths of the full dataset that is the training data for that fold. The value of λ is selected from the set of values $\{0.2, 0.4, 0.8, 1.6, 3.2\}$ to minimize the fivefold cross-validation test error. The performance results of the nine other classifiers are as given by Cai and Sowmya [31], who report the same tenfold cross-validation test error as that given for the GLS classifier. Details about parameter settings for the other nine classifiers may be found in [31].

The GLS classifier outperforms each of the other classifiers at least once among the

the C4.4 and C4.5 decision trees (which could be applied to all of the datasets without concern for dimensionality).

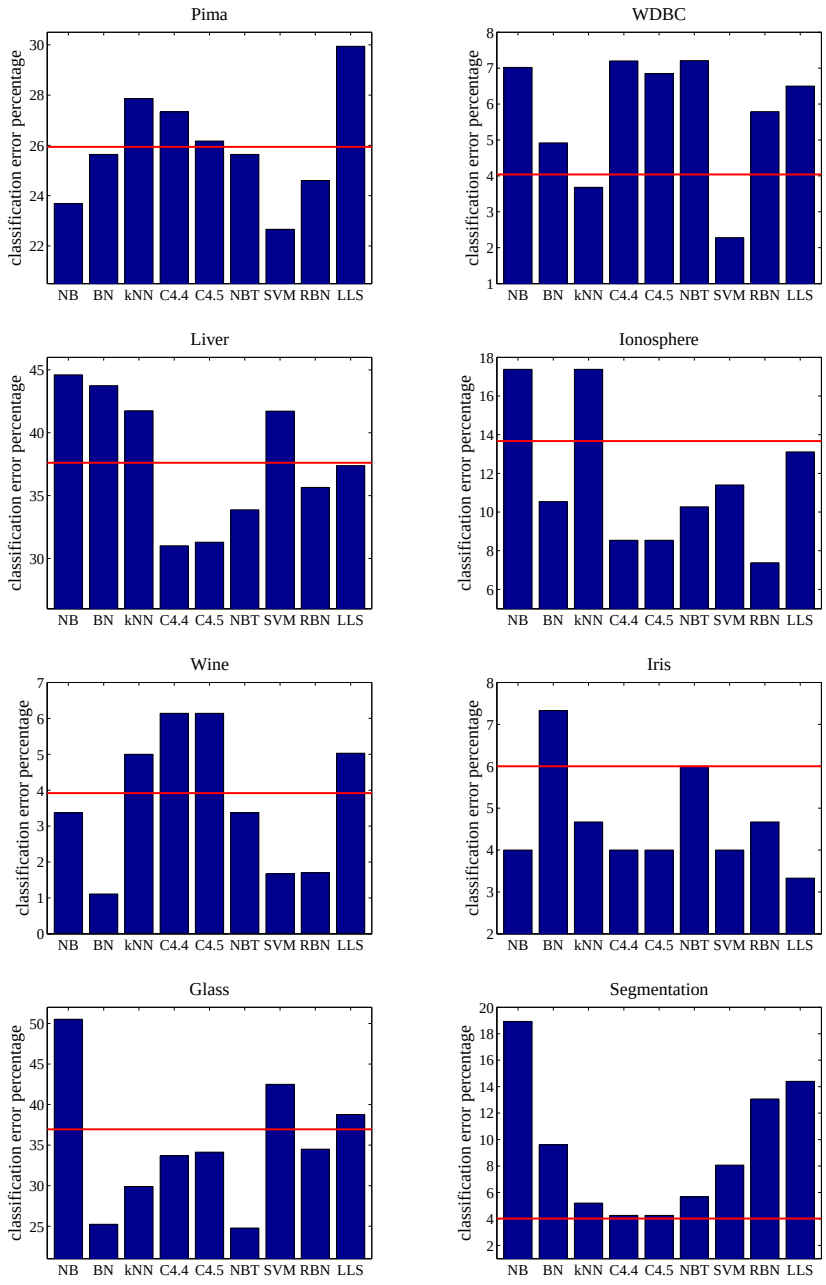


Figure 3.7. Tenfold cross-validation error percentage of GLS classifier with RBF level set implementation (red line) on several datasets compared to error percentages of various other classifiers (blue bars) reported in [31]. The other classifiers are: naïve Bayes classifier (NB), Bayes net classifier (BN), k-nearest neighbor with inverse distance weighting (kNN), C4.4 decision tree (C4.4), C4.5 decision tree (C4.5), naïve Bayes tree classifier (NBT), SVM with polynomial kernel (SVM), radial basis function network (RBN), and learning level set classifier (LLS) of Cai and Sowmya [31].

four binary datasets, and is generally competitive overall. The GLS classifier is also competitive on the multicategory datasets. In fact, it gives the smallest error among all of the classifiers on the segmentation dataset. The proposed classifier is competitive for datasets of both small and large dimensionality D; there is no apparent relationship between D and the performance of the GLS classifier in comparison to other methods.

■ 3.4 Complexity and Consistency Analysis

In this section, the GLS classifier proposed in this chapter is characterized in terms of complexity, both VC dimension [206] and Rademacher complexity [13, 107], as well as consistency. VC dimension is examined through the empirical procedure described in [205]. The main tool used in the characterization of Rademacher complexity and consistency is ϵ-entropy [106]. An expression for the ϵ-entropy of the set of geometric level set classifiers is derived and then Rademacher consistency and complexity results from learning theory that are based on it are applied. The main findings are that GLS classifiers are consistent, and that complexity is monotonically related to decision boundary surface area frugality, and thus the surface area regularization term can be used to control underfitting and overfitting.

■ 3.4.1 Empirical Vapnik–Chervonenkis Dimension

The VC dimension $C^{\text{VC}}(\mathcal{F})$, the maximum number of points that $\mathcal{F}$ can shatter as described in Section 2.2.2, is difficult to specify analytically except in special cases such as classifiers with linear decision boundaries. However, Vapnik et al. [205] have outlined a procedure for empirically measuring the VC dimension of a classifier based on classification error on uniformly drawn samples $\mathbf{x}_j$ with class labels $y_j \in \{-1, +1\}$ assigned randomly with equal probability. The empirically measured VC dimension $\hat{C}^{\text{VC}}$ is given below for the GLS classifier as a function of the surface area regularization parameter λ.

The GLS classifier that is considered is defined in $D = 2$ dimensions. Starting from the same small seed decision boundary initialialization as in Figure 3.1 and Figure 3.2, contour evolution is performed to minimize the margin-based energy functional (3.4) on twenty-five randomly generated training sets with 3000 samples having label $y_j = -1$ and 3000 samples having label $y_j = +1$. Carrying this out for fifty different values of the regularization weight λ, estimating the VC dimension using the calculation of [205], and averaging over the twenty-five trials gives a plot of estimated VC dimension $\hat{C}^{\text{VC}}$ as a function of λ. The training sets are uniformly random, as that provides the worst case for shattering; other application-specific data distributions could be considered for complexity analysis as well.

The relationship between $\hat{C}^{\text{VC}}$ and λ, shown in Figure 3.8 is essentially monotonic. Figure 3.9 shows the decision boundaries for different values of λ corresponding to one instance of the random training set. The smoother, less tortuous contours corresponding

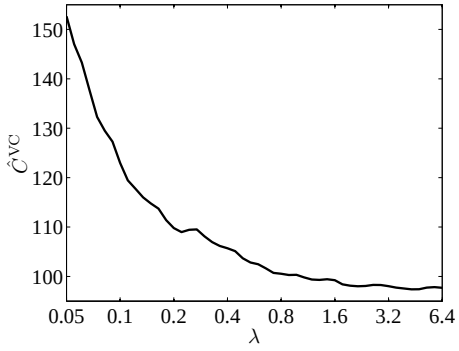


Figure 3.8. Estimated VC dimension as a function of the surface area regularization weight on a logarithmic scale.

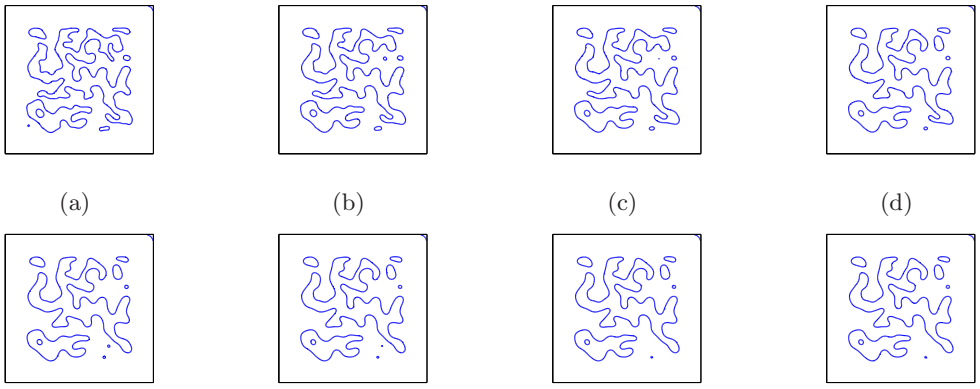


Figure 3.9. Classifiers learned from one instance of a random training set for different values of λ used to estimate VC dimension by the procedure of [205]: (a) $\lambda = 0.05$, (b) $\lambda = 0.1$, (c) $\lambda = 0.2$, (d) $\lambda = 0.4$, (e) $\lambda = 0.8$, (f) $\lambda = 1.6$, (g) $\lambda = 3.2$, and (h) $\lambda = 6.4$. Note that these decision boundaries are not iterations of a contour evolution, but final boundaries for different values of λ.

to the larger values of λ can shatter fewer points.[6] Figure 3.8 shows empirically that the complexity of the GLS classifier measured by VC dimension can be directly controlled using the weight on the surface area regularization term. The VC dimension behavior as a function of λ combined with the VC generalization bound (2.31) given in Section 2.2.2 matches well with the behavior of test error as a function of λ seen in Figure 3.6.

■ 3.4.2 Rademacher Complexity

As discussed in Section 2.2.2, the VC dimension is one way to quantify the complexity of a classifier, and another is through the Rademacher complexity $\hat{C}_n^{\text{Rad}}(\mathcal{F})$ defined in (2.30). An analytical bound for $\hat{C}_n^{\text{Rad}}$ is developed in this section. Like $\hat{C}^{\text{VC}}$, it is found that the bound for $\hat{C}_n^{\text{Rad}}$ decreases with λ and correspondingly increases with a constraint on the decision boundary surface area.

In characterizing $\hat{C}_n^{\text{Rad}}(\mathcal{F})$, $\mathcal{F}$ is taken to be the set of signed distance functions on Ω and $\mathcal{F}_s$ to be the subset of signed distance functions whose zero level sets have surface area less than s, that is $\oint_{\mathcal{C}} d\mathbf{s} < s$. Such a constraint is related to the regularized margin-based loss expression $\text{L}(\varphi)$ given in (3.4) through the method of Lagrange multipliers, with λ inversely related to s. In classification, it is always possible to scale and shift the data and this is often done in practice. Without losing much generality and dispensing with some bookkeeping, consider signed distance functions defined on the unit hypercube: $\Omega = [0,1]^D$.

It is shown in [121] that the Rademacher average of an arbitrary function class $\mathcal{F}_s$ satisfies:

$$\hat{C}_n^{\text{Rad}}(\mathcal{F}_s) \leq 2\epsilon + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\frac{\epsilon}{4}}^{\infty} \sqrt{H_{\rho_\infty,\epsilon'}(\mathcal{F}_s)}\, d\epsilon', \tag{3.12}$$

where $H_{\rho_\infty,\epsilon}(\mathcal{F}_s)$ is the ϵ-entropy of $\mathcal{F}_s$ with respect to the metric

$$\rho_\infty(\varphi_1, \varphi_2) = \sup |\varphi_1(\mathbf{x}) - \varphi_2(\mathbf{x})|.$$

The ϵ-covering number $N_{\rho_\infty,\epsilon}(\mathcal{F}_s)$ of a metric space is the minimal number of sets with radius not exceeding ϵ required to cover that space; the ϵ-entropy is the base-two logarithm of the ϵ-covering number [106].

Given the Rademacher generalization bound (2.32) presented in Section 2.2.2 and the Rademacher complexity term (3.12), an expression for $H_{\rho_\infty,\epsilon}(\mathcal{F}_s)$ must be found to characterize the prevention of overfitting by being frugal with the decision boundary surface area. The ϵ-covering number and ϵ-entropy are useful values in characterizing learning [12, 108, 116, 121, 189, 220]. Calculations of ϵ-entropy for various classes of functions and various classes of sets are provided in [57, 58, 106] and other works, but the ϵ-entropy of the class of signed distance functions with a constraint on the surface area of the zero level set does not appear in the literature. The second and third examples in Section 2 of [106] are related, and the general approach taken here for obtaining the ϵ-entropy of GLS classifiers is similar to those two examples.

[6] See Figure 3.3 as well for decision boundaries corresponding to different values of λ.

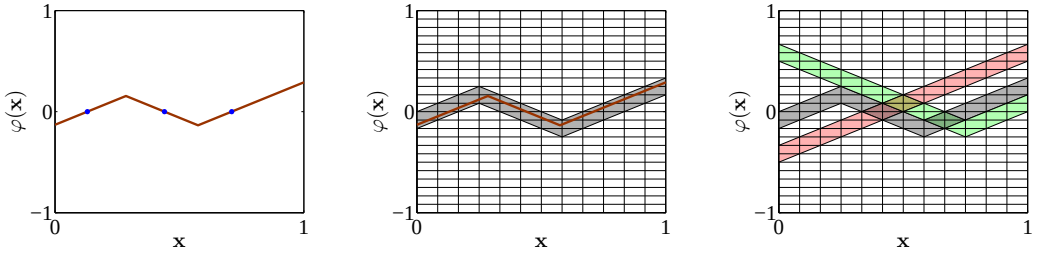


Figure 3.10. Illustration of ϵ-corridors with $D = 1$. A one-dimensional signed distance function in $\Omega = [0, 1]$ is shown in (a), marked with its zero level set. The ϵ-corridor with $\epsilon = 1/12$ that contains the signed distance function is shown in (b), shaded in gray. The ϵ-corridor of (b), whose center line has three zero crossings is shown in (c), again shaded in gray, along with an ϵ-corridor whose center line has two zero crossings, shaded in green, and an ϵ-corridor whose center line has one zero crossing, shaded in red.

The exposition begins with the $D = 1$ case and then comes to general D. Figure 3.10(a) shows a signed distance function over the unit interval. Due to the $\|\nabla\varphi\| = 1$ constraint, its slope is either -1 or $+1$ almost everywhere. The slope changes sign exactly once between two consecutive points in the zero level set. The signed distance function takes values in the range between positive and negative one.[7] As mentioned at the beginning of the chapter, the surface area is the number of points in the zero level set in the $D = 1$ context, for example three in Figure 3.10(a).

Sets known as ϵ-*corridors* are used in finding $H_{\rho_\infty,\epsilon}(\mathcal{F}_s)$. They are particular balls of radius ϵ measured using ρ_∞ in the space of signed distance functions. This corridor terminology is the same as in [106], but the definition here is slightly different. An ϵ-corridor is a strip of height 2ϵ for all $\mathbf{x}$. Define $\nu = \lceil 1/\epsilon \rceil$. At $\mathbf{x} = 0$, the bottom and top of a corridor are at $2i\epsilon$ and $2(i + 1)\epsilon$ respectively for some integer i, where $-\nu \leq 2i < \nu$. The slope of the corridor is either -1 or $+1$ for all $\mathbf{x}$ and the slope can only change at values of $\mathbf{x}$ that are multiples of ϵ. Additionally, the center line of the ϵ-corridor is a signed distance function, changing slope halfway between consecutive points in its zero level set and only there. The ϵ-corridor in which the signed distance function of Figure 3.10(a) falls is indicated in Figure 3.10(b). Other ϵ-corridors are shown in Figure 3.10(c).

By construction, each signed distance function is a member of exactly one ϵ-corridor. This is because since at $\mathbf{x} = 0$ the bottom and top of ϵ-corridors are at consecutive integer multiples of 2ϵ and since the center line of the corridor is a signed distance function, each signed distance function starts in one ϵ-corridor at $\mathbf{x} = 0$ and does not

[7] There are several ways to define the signed distance function in the two degenerate cases $\mathcal{R}_- = \Omega$ and $\mathcal{R}_- = \emptyset$, including the assignments $-\infty$ and $+\infty$, or -1 and $+1$ [50]. For the purposes of this section, it suffices to say that a unique function for the $\mathcal{R}_- = \Omega$ case and a unique function for the $\mathcal{R}_- = \emptyset$ case has been chosen.

escape from it in the interval $(0, 1]$. Also, an ϵ-corridor whose center line has s points in its zero level set contains only signed distance functions with at least s points in their zero level sets.

Theorem 3.4.1. *The ϵ-entropy of the set of signed distance functions defined over $\Omega = [0, 1]$ with zero level set having less than s points is:*

$$H_{\rho_\infty,\epsilon}(\mathcal{F}_s) = \log_2 \left(\sum_{k=1}^{s} \binom{\nu - 1}{k - 1} \right) + 1.$$

Proof. Since ϵ-corridors only change slope at multiples of ϵ, the abscissa can be divided into ν pieces. (Each piece has width ϵ except the last one if $1/\epsilon$ is not an integer.) In each of the ν subintervals, the center line of a corridor is either wholly positive or wholly negative. Enumerating the full set of ϵ-corridors is equivalent to enumerating binary strings of length ν. Thus, without a constraint s, there are 2^ν ϵ-corridors. Since, by construction, ϵ-corridors tile the space of signed distance functions, $N_{\rho_\infty,\epsilon}(\mathcal{F}) = 2^\nu$.

With the s constraint on ϵ-corridors, the enumeration is equivalent to twice the number of compositions of the positive integer ν by a sum of s or less positive integers. Twice because for every composition, there is one version in which the first subinterval of the corridor center is positive and one version in which it is negative. As an example, the red corridor in Figure 3.10(c) can be composed with two positive integers $(5 + 7)$, the green corridor by three $(7 + 4 + 1)$, and the gray corridor by four $(1 + 4 + 4 + 3)$. The number of compositions of ν by k positive integers is $\binom{\nu-1}{k-1}$. Note that the zero-crossings are unordered for this enumeration and that the set $\mathcal{F}_s$ includes all of the signed distance functions with surface area smaller than s as well. Therefore:

$$N_{\rho_\infty,\epsilon}(\mathcal{F}_s) = 2 \sum_{k=1}^{s} \binom{\nu - 1}{k - 1}.$$

The result then follows because $H_{\rho_\infty,\epsilon}(\mathcal{F}_s) = \log_2 N_{\rho_\infty,\epsilon}(\mathcal{F}_s)$. ■

The combinatorial formula in Theorem 3.4.1 is difficult to work with, so a highly accurate approximation is given as Theorem 3.4.2.

Theorem 3.4.2. *The ϵ-entropy of the set of signed distance functions defined over $\Omega = [0, 1]$ with zero level set having less than s points is:*

$$H_{\rho_\infty,\epsilon}(\mathcal{F}_s) \approx \nu + \log_2 \Phi \left(\frac{2s - \nu}{\sqrt{\nu - 1}} \right),$$

where Φ is the standard Gaussian cumulative distribution function.

Proof. The result follows from the de Moivre–Laplace theorem and continuity correction, which are used to approximate the binomial distribution with the Gaussian distribution. ■

Since Φ is a cumulative distribution function taking values in the range zero to one, $\log_2 \Phi$ is nonpositive. The surface area constraint only serves to reduce the ϵ-entropy.

The ϵ-entropy calculation has been for the $D = 1$ case thus far. Moving to the case with general D, recall that $\Omega = [0,1]^D$. Once again, ϵ-corridors are constructed that tile the space of signed distance functions. In the one-dimensional case, the ultimate object of interest for enumeration is a string of length ν with binary labels. In the two-dimensional case, the corresponding object is a ν-by-ν grid of ϵ-by-ϵ squares with binary labels, and in general a D-dimensional Cartesian grid of hypercubes with content ϵ^D, ν on each side. The surface area of the zero level set is the number of interior faces in the Cartesian grid whose adjoining hypercubes have different binary labels.

Theorem 3.4.3. *The ϵ-entropy of the set of signed distance functions defined over $\Omega = [0,1]^D$ with zero level set having surface area less than s is:*

$$H_{\rho_\infty,\epsilon}(\mathcal{F}_s) \approx \nu^D + \log_2 \Phi \left(\frac{2s - D\,(\nu-1)\,\nu^{D-1} - 1}{\sqrt{D\,(\nu-1)\,\nu^{D-1}}} \right),$$

where Φ is the standard Gaussian cumulative distribution function.

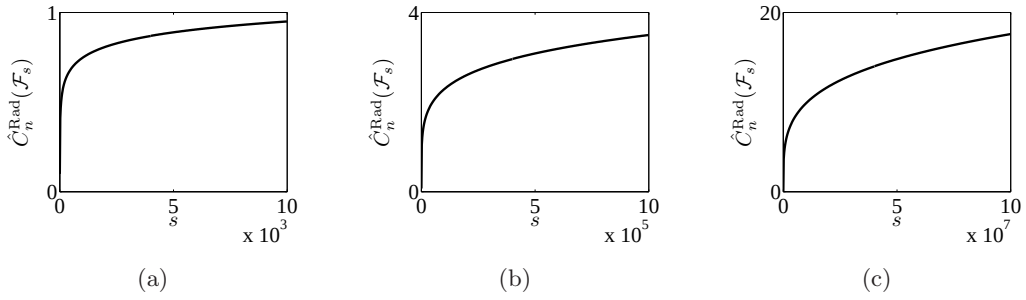
Proof. In the one-dimensional case, it is easy to see that the number of segments is ν and the number of interior faces is $\nu - 1$. For a general D-dimensional Cartesian grid with ν hypercubes on each side, the number of hypercubes is ν^D and the number of interior faces is $D(\nu-1)\nu^{D-1}$. The result follows by substituting ν^D for ν and $D(\nu-1)\nu^{D-1}$ for $\nu - 1$ in the appropriate places in Theorem 3.4.2. ■

Theorem 3.4.2 is a special case of Theorem 3.4.3 with $D = 1$. It is common to find the dimension of the space D in the exponent of ϵ^{-1} in ϵ-entropy calculations as is found here.

The expression for the ϵ-entropy of signed distance functions with surface area constraint can be used along with the Rademacher complexity expression (3.12) to characterize GLS classifier complexity. With $\Omega = [0,1]^D$, the upper limit of the integral in (3.12) is one rather than infinity because ϵ cannot be greater than one. The right side of (3.12) is plotted as a function of the surface area constraint s in Figure 3.11 for three values of D, and fixed ϵ and n. As the value of s increases, decision boundaries with more surface area are available. Decision boundaries with large surface area are more complex than smoother decision boundaries with small surface area. Hence the complexity term increases as a function of s. Consequently, the surface area penalty can be used to control the complexity of the classifier, and prevent underfitting and overfitting. The same relationship between the empirical VC dimension and the surface area penalty appears in 3.4.1.

■ 3.4.3 Consistency

As mentioned in Section 2.2.1, $\mathrm{R}(\hat{y}^{(n)})$, the generalization error of a classifier learned using a training set of size n drawn from $f_{\mathbf{x},y}(\mathbf{x}, y)$, converges in the limit as n goes



(a) (b) (c)

Figure 3.11. Rademacher complexity as a function of the surface area constraint s for signed distance functions on $\Omega = [0,1]^D$ with (a) $D = 2$, (b) $D = 3$, and (c) $D = 4$. The values of ϵ and n are fixed at 0.01 and 1000 respectively.

to infinity to the Bayes optimal probability of error $\mathrm{R}(\hat{y}^*)$ if the classifier is consistent. This convergence for classifier consistency is in probability, as $\mathrm{R}(\hat{y}^{(n)})$ is a random variable due to the stochastic nature of the training set. The learned GLS classifier $\hat{y}^{(n)}(\mathbf{x}) = \mathrm{sign}(\varphi^{(n)}(\mathbf{x}))$ minimizes an energy functional composed of both margin-based loss and surface area regularization, and consequently the properties of $\mathrm{R}(\hat{y}^{(n)})$ are affected by both the margin-based loss function ℓ and by the surface area regularization term or surface area constrained function class $\mathcal{F}_s$. Lin [116], Steinwart [189], and Bartlett et al. [12] have given conditions on the loss function necessary for a margin-based classifier to be consistent. Lin [116] calls a loss function that meets the necessary conditions *Fisher-consistent*. Common margin-based loss functions including the logistic loss function are Fisher-consistent.[8] Fisher consistency of the loss function is not enough, however, to imply consistency of the classifier overall. The regularization term must also be analyzed; since the regularization term based on decision boundary surface area introduced in this chapter is new, so is the following analysis.

Theorem 4.1 of [116], which is based on ϵ-entropy, is adapted to show consistency of the GLS classifier. The analysis is based on the method of sieves, where sieves $\mathcal{F}_n$ are an increasing sequence of subspaces of a function space $\mathcal{F}$. For the case considered here, $\mathcal{F}$ is the set of signed distance functions on Ω and the sieves, $\mathcal{F}_{s(n)}$, are subsets of signed distance functions whose zero level sets have surface area less than $s(n)$. In the following, the function $s(n)$ is increasing in n and thus the conclusions of the theorem provide asymptotic results on consistency as the strength of the regularization term decreases as more training samples are made available. The sieve estimate is:

$$\varphi^{(n)} = \arg \min_{\varphi \in \mathcal{F}_{s(n)}} \sum_{j=1}^{n} \ell(y_j \varphi(\mathbf{x}_j)). \tag{3.13}$$

[8] The conditions on ℓ for Fisher consistency are mainly related to incorrect classifications incurring more loss than correct classifications.

Theorem 3.4.4. *Let ℓ be a Fisher-consistent loss function in* (3.13)*; let*

$$\tilde{\varphi} = \arg\min_{\varphi \in \mathcal{F}} \mathrm{E}[\ell(y\varphi(\mathbf{x}))],$$

where $\mathcal{F}$ is the space of signed distance functions on $[0,1]^D$; and let $\mathcal{F}_{s(n)}$ be a sequence of sieves. Then[9] *for sieve estimate $\varphi^{(n)}$:*

$$\mathrm{R}(\hat{y}^{(n)}) - \mathrm{R}(\hat{y}^*) = \mathrm{O}_P\left(\max\left\{n^{-\tau}, \inf_{\varphi \in \mathcal{F}_{s(n)}} \int (\varphi(\mathbf{x}) - \tilde{\varphi}(\mathbf{x}))^2 f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}\right\}\right),$$

where

$$\tau = \begin{cases} \frac{1}{3}, & D = 1 \\ \frac{1}{4} - \frac{\log\log n}{2\log n}, & D = 2 \\ \frac{1}{2D}, & D \geq 3 \end{cases}.$$

Proof. The result is a direct application of Theorem 4.1 of [116], which is in turn an application of Theorem 1 of [181]. In order to apply this theorem, two things must be noted. First, that signed distance functions on $[0,1]^D$ are bounded in the L_∞ norm. As noted previously in the section, signed distance functions take values between -1 and $+1$ for $D = 1$. In general, they take values between $-\sqrt{D}$ and $+\sqrt{D}$, and thus are bounded. Second, that there exists a B such that $H_{\rho_\infty,\epsilon}(\mathcal{F}_s) \leq B\epsilon^{-D}$. Based on Theorem 3.4.3, $H_{\rho_\infty,\epsilon}(\mathcal{F}_s) \leq \nu^D$ because the logarithm of the cumulative distribution function is nonpositive. Since $\nu = \lceil 1/\epsilon \rceil$, if $1/\epsilon$ is an integer, then $H_{\rho_\infty,\epsilon}(\mathcal{F}_s) \leq \epsilon^{-D}$ and otherwise there exists a B such that $H_{\rho_\infty,\epsilon}(\mathcal{F}_s) \leq B\epsilon^{-D}$. ■

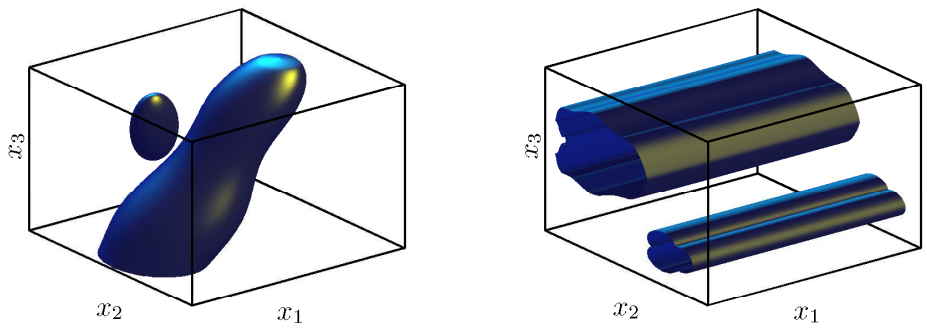
Clearly $n^{-\tau}$ goes to zero as n goes to infinity. Also, $\inf_{\varphi \in \mathcal{F}_{s(n)}} \int (\varphi(\mathbf{x}) - \tilde{\varphi}(\mathbf{x}))^2 f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}$ goes to zero when $s(n)$ is large enough so that the surface area constraint is no longer applicable.[10] Thus, level set classifiers are consistent.

■ 3.5 Feature Subset Selection Using Geometric Level Sets

An advantage of using variational level set methods for supervised classification is that they allow for the inclusion of geometric preferences and priors for the decision boundary in the input space, which are more difficult to include in kernel methods for example. Being frugal with the decision boundary surface area is one such geometric preference. In this section, an additional such preference is considered which incorporates local feature relevance [52] and promotes feature subset selection in a manner similar to the ℓ_1 feature selection of [137]. The decision boundary adapts the surface area frugality to the relevance of different input space dimensions for classification.

[9] The notation $\mathsf{a}_n = \mathrm{O}_P(b_n)$ means that the random variable $\mathsf{a}_n = b_n \mathsf{c}_n$, where c_n is a random variable bounded in probability [204]. Thus, if b_n converges to zero, then a_n converges to zero in probability.

[10] For a given ϵ, there is a maximum possible surface area; the constraint is no longer applicable when the constraint is larger than this maximum possible surface area.



(a) (b)

Figure 3.12. Decision boundary in three-dimensional space that (a) uses all input variables, and (b) selects the two variables x_2 and x_3 for classification. Note that the decision boundary contours are shown, not the signed distance function.

■ 3.5.1 Variational Level Set Formulation

In input spaces where some of the dimensions are irrelevant or not informative for classification, feature subset selection is important to prevent overfitting [137]. The idea is to learn classifiers which only make use of the relevant dimensions. As described in Section 2.2.3, margin-based classifiers with linear decision boundaries in a D-dimensional space have decision function $\varphi(\mathbf{x}) = \boldsymbol{\theta}^T\mathbf{x} + \theta_0$, where $\boldsymbol{\theta}$ is a length D vector of coefficients. Feature subset selection can be formulated through the preference that $\boldsymbol{\theta}$ be sparse, that is have few nonzero elements. An ℓ_1-norm penalty is well known for producing sparse solutions as well as being tractable [40, 196]. The idea of ℓ_1-based feature subset selection for linear decision boundaries is extended here to decision boundaries represented by level set functions.

First, note that a classifier that ignores a particular measurement dimension has a decision boundary that is constant and does not change as a function of that unused dimension. As seen in Figure 3.12, such a decision boundary is a generalized cylinder parallel to the unused dimension axis. The partial derivative of the level set function with respect to the unused variable is zero for all $\mathbf{x} \in \Omega$. The magnitude of the partial derivative is used by Domeniconi et al. [52] to locally indicate feature relevance.

If the partial derivative $\varphi_{x_k}(\mathbf{x})$ is zero for all $\mathbf{x} \in \Omega$, then the scalar quantity $\int_\Omega |\varphi_{x_k}(\mathbf{x})|\, d\mathbf{x}$ equals zero. Consequently, a length D vector:

$$\begin{bmatrix} \int_\Omega |\varphi_{x_1}(\mathbf{x})|\, d\mathbf{x} \\ \vdots \\ \int_\Omega |\varphi_{x_D}(\mathbf{x})|\, d\mathbf{x} \end{bmatrix}$$

may be constructed, which should be sparse for feature subset selection.

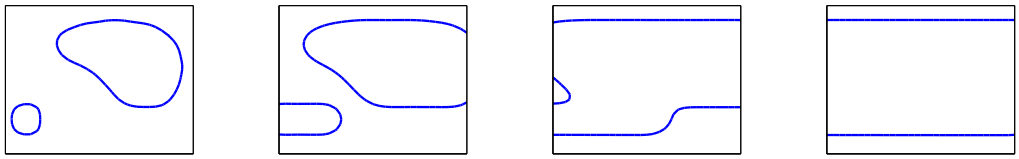


Figure 3.13. Contour evolution with the surface area penalty and one of the partial derivative terms for feature subset selection. The evolution from the initial contour to the final contour is shown from left to right. For this illustration, the energy functional contains no empirical risk term. The final contour is a cylinder.

Applying the ℓ_1-norm to this vector and appending it to (3.4), gives the following energy functional:

$$\mathrm{L}(\varphi) = \sum_{j=1}^{n} \ell(y_j \varphi(\mathbf{x}_j)) + \lambda_1 \oint_{\mathcal{C}} d\mathbf{s} + \lambda_2 \sum_{k=1}^{D} \int_{\Omega} |\varphi_{x_k}(\mathbf{x})| \, d\mathbf{x}, \tag{3.14}$$

where λ_2 is the weight given to the feature subset selection term. The ℓ_1 regularization further restricts the model class $\mathcal{F}_s$. The gradient descent flow for the ℓ_1 term is

$$\varphi_t(\mathbf{x}) = -\lambda_2 \left(\sum_{k=1}^{D} |\varphi_{x_k}(\mathbf{x})| \right) \nabla \varphi(\mathbf{x}). \tag{3.15}$$

Thus, contour evolution to minimize (3.14) may be used for feature subset selection integrated with classifier training in the same way as, for example, ℓ_1-regularized logistic regression for linear decision boundaries [137]. Figure 3.13 shows contour evolution from an initial contour with the energy functional containing one of the D partial derivative feature subset selection terms and containing no empirical risk term. The final contour is a cylinder, as in Figure 3.12(b). In practice, the empirical risk term guides which dimension or dimensions are selected.

■ 3.5.2 Example

GLS classifiers trained with the additional feature subset selection term are now shown on the same dataset given in the second example of Section 3.1.3. With $\lambda_1 = 0.5$, the final decision boundaries for four different values of λ_2 are shown in Figure 3.14. The final decision boundary in Figure 3.14(a) is with $\lambda_2 = 0$ and is thus the same as the final decision boundary in the second example of Section 3.1.3. For larger values of λ_2, the decision boundaries increasingly make the horizontal dimension irrelevant for classification. The data guides which of the two dimensions to make irrelevant, as regularization terms for both the horizontal and vertical dimensions are included in the objective functional (3.14).

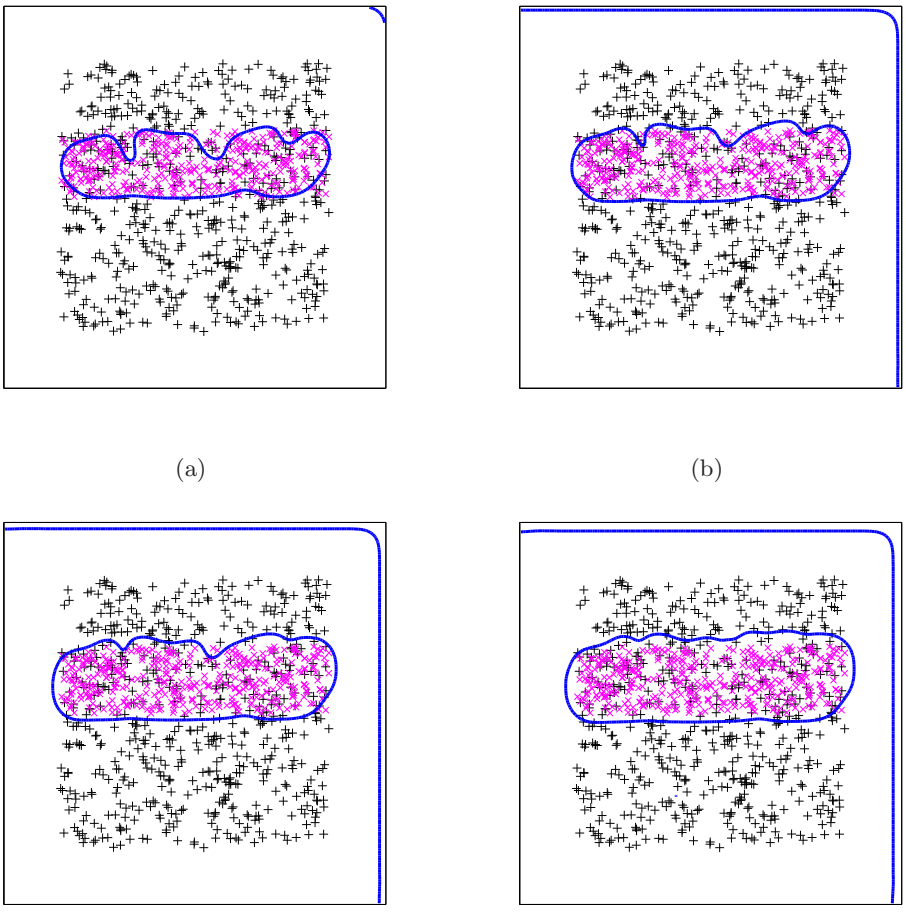


Figure 3.14. Final decision boundaries with feature subset selection term weighted by (a) $\lambda_2 = 0$, (b) $\lambda_2 = 3$, (c) $\lambda_2 = 5$, and (d) $\lambda_2 = 7$. The magenta $\times$ markers indicate class label -1 and the black $+$ markers indicate class label $+1$. The blue line is the decision boundary.

The variational level set formulation is flexible in allowing the inclusion of various geometric priors defined in the input space. The energy functional of feature relevance measured using the partial derivative of the signed distance function is one example. Others may be included as desired.

3.6 Chapter Summary

Level set methods are powerful computational techniques that have not yet been widely adopted in statistical learning. This chapter contributes to opening a conduit between the application area of learning and the computational technique of level set methods. Toward that end, a nonlinear, nonparametric classifier based on variational level set methods has been developed that minimizes margin-based empirical risk in both the binary and multicategory cases, and is regularized by a geometric complexity penalty novel to classification. This approach with decision boundary surface area frugality is an alternative to kernel machines for learning nonlinear decision boundaries in the input space and is in some ways a more natural generalization of linear methods.

A multicategory level set classification procedure has been described with a logarithmic number of decision functions, rather than the linear number that is typical in classification and decision making, through a binary encoding made possible by the level set representation. A characterization of Vapnik–Chervonenkis and Rademacher complexities, and consistency results have been provided. The variational level set formulation is flexible in allowing the inclusion of various geometric priors defined in the input space. One example is the energy functional of feature relevance measured using the partial derivative of the signed distance function proposed for ℓ_1-regularized feature subset selection.

It is a known fact that with finite training data, no one classification method is best for all datasets. Performance of classifiers may vary quite a bit depending on the data characteristics because of differing inductive biases. The classifier presented in this chapter provides a new option when choosing a classifier. The results on standard datasets indicate that the GLS classifier is competitive with other state-of-the-art classifiers.

Chapter 4

Dimensionality of the Classifier Input

DECISION rules in hypothesis testing are simplified through sufficient statistics such as the likelihood ratio. Calculation of a sufficient statistic losslessly reduces the dimensionality of high-dimensional measurements before applying a decision rule defined in the reduced-dimensional space. Decision rules that are learned in most supervised classification methods, in contrast, are defined in the full high-dimensional input space rather than in a reduced-dimensional space. The frugality pursued in this chapter is to limit the dimensionality of the space in which classification decision boundaries are defined. A method for simultaneously learning both a dimensionality reduction mapping, represented by a matrix on the Stiefel manifold, and a margin-based classifier defined in the reduced-dimensional space is proposed [210, 212]. Not only does dimensionality reduction simplify decision rules, but it also decreases generalization error by preventing overfitting [25, 131, 206, 230].

As mentioned in Section 2.4, many methods for linear dimensionality reduction can be posed as optimization problems on the Stiefel or Grassmann manifold with different objectives [188]. In a similar manner to how the SVM with linear decision boundary can be extended to a classifier with nonlinear decision boundary through the use of kernels, linear dimensionality reduction methods that are optimization problems on the Stiefel or Grassmann manifold can be extended as nonlinear dimensionality reduction methods using kernels [15, 128, 176, 177]. In this chapter, an optimization problem on the Stiefel manifold is proposed whose objective is that of margin-based classification and an iterative coordinate descent algorithm for its solution is developed. The proposed objective and coordinate descent are extended for distributed dimensionality reduction in sensor networks. In that resource-constrained setting, not only does dimensionality reduction improve classifier generalization, but also reduces the amount of communication by sensor nodes.

The most popular method of dimensionality reduction for data analysis is PCA [95, 96, 148]. PCA (and its nonlinear extension using kernels [177]) only makes use of the measurement vectors, not the class labels, in finding a dimensionality reduction mapping. Popular nonlinear dimensionality reduction methods such as Isomap [193],

locally linear embedding [166], and Laplacian eigenmaps [17] also do not make use of class labels.[1] If the dimensionality reduction is to be done in the context of supervised classification, the class labels should also be used. Several *supervised* dimensionality reduction methods exist in the literature. These methods can be grouped into three broad categories: those that separate likelihood functions according to some distance or divergence, those that try to match the probability of the labels given the measurements with the probability of the labels given the dimensionality-reduced measurements, and those that attempt to minimize a specific classification or regression objective.

FDA is a supervised linear dimensionality reduction method that assumes that the likelihood functions are Gaussian with the same covariance and different means [69]. It (and its nonlinear extension using kernels [15, 128]) returns a matrix on the Stiefel manifold that maximally separates (in Euclidean distance) the clusters of the different labels [188]. The method of Lotlikar and Kothari [119] also assumes Gaussian likelihoods with the same covariance and different means, but with an even stronger assumption that the covariance matrix is a scalar multiple of the identity. The probability of error is explicitly minimized using gradient descent; the gradient updates do not enforce the Stiefel manifold constraint, but the Gram-Schmidt orthonormalization procedure is performed after every step to obtain a matrix that does meet the constraint. With a weaker assumption only that the likelihood functions are Gaussian, but without restriction on the covariances, other methods maximize Bhattacharyya divergence or Chernoff divergence, which are surrogates for minimizing the probability of error [195].

The method of Patrick and Fischer [147], like FDA, maximally separates the clusters of the different labels but does not make the strong Gaussian assumption. Instead, it performs kernel density estimation of the likelihoods and separates those estimates. The optimization is gradient ascent and orthonormalization is performed after every step. Similarly, information preserving component analysis also performs kernel density estimation and maximizes Hellinger distance, another surrogate for minimizing the probability of error, with optimization through gradient ascent and the Stiefel manifold constraint maintained in the gradient steps [33]. Other approaches with information-theoretic criteria include [135, 160, 199].

Like [33, 147], the method of Sajama and Orlitsky [170] also estimates probability density functions for use in the criterion for dimensionality reduction. The particular criterion, however, is based on the idea that the dimensionality reduction mapping should be such that the conditional probability of the class labels in the supervised classification problem given the input high-dimensional measurements equals the conditional probability of the class labels given the reduced-dimensional features. The same criterion appears in [73, 74, 114, 115] and many references given in [42]. These papers describe various methods of finding dimensionality reduction mappings to optimize the

[1] Incidentally, the nonlinear dimensionality reduction methods Isomap, locally linear embedding, and Laplacian eigenmaps can be expressed in terms of kernels in a manner similar to kernel PCA, as shown by Bengio et al. [19]. The expression of Isomap in terms of a kernel function is given at the end of Section 2.4.3.

criterion with different assumptions.

Some supervised dimensionality reduction methods explicitly optimize a classification or regression objective. The support vector singular value decomposition machine has a joint objective for dimensionality reduction and classification with the hinge loss function [150]. However, the matrix it produces is not guaranteed to be on the Stiefel manifold, and the space in which the classifier is defined is not exactly the dimensionality-reduced image of the high-dimensional space. It also changes the regularization term from what is used in the standard SVM formulation. Maximum margin discriminant analysis is another method based on the SVM; it finds the reduced-dimensional features one by one instead of giving the mappings for all of the reduced dimensions at once and it does not simultaneously give a classifier [202]. The method of [117, 188] is based on the nearest neighbor classifier. A linear regression objective and a regression parameter/Stiefel manifold coordinate descent algorithm is developed in [151].

The objective function and optimization procedure proposed in this chapter have some similarities to many of the methods discussed, but also some key differences. First of all, no explicit assumption is made on the statistics of likelihood functions, and indeed no assumptions are explicitly used, e.g., no assumption of Gaussianity is employed.[2] Moreover, the method proposed in the chapter does not require nor involve estimation of the probability density functions under the two hypotheses nor of the likelihood ratio. The direct interest is only in learning decision boundaries and using margin-based loss functions to guide both this learning *and* the optimization over the Stiefel manifold to determine the reduced-dimensional space in which decision making is to be performed. Density estimation is a harder problem than finding classifier decision boundaries and it is well known that when learning from finite data, it is best to only solve the problem of interest and nothing more. Similarly, the desideratum that the conditional distributions of the class labels given the high-dimensional and reduced-dimensional measurements match is more stringent than wanting good classification performance in the reduced-dimensional space.

Rather than nearest neighbor classification or linear regression, the objective in the proposed method is margin-based classification. The proposed method finds all reduced-dimensional features in a joint manner, and gives both the dimensionality reduction mapping and the classifier as output. Unlike in [150], the classifier is defined exactly without approximation in the reduced-dimensional space that is found. Additionally, the regularization term and consequently inductive bias of the classifier is left unchanged.

The preceding represent the major conceptual differences between the framework developed in this chapter and that considered in previous work. Coordinate descent optimization procedures are used in this work, which are also employed in other works, e.g. [150, 151], but the setting in which these are used is new. The proposed frame-

[2]Note that there is an intimate relationship between margin-based loss functions used in this chapter and statistical divergences [139].

work also allows the development of some new theoretical results on consistency and Rademacher complexity. Moreover, the framework allows a natural generalization to distributed dimensionality reduction for classification in sensor networks, a problem that has not been considered previously.

This work fits into the general category of the supervised learning of frugal data representations. Examples from this category include supervised learning of undirected graphical models [172], sparse signal representations [98, 123], directed topic models [26, 109], quantizer codebooks [110], and dimensionality reduction mappings, which is the topic of this chapter.

The chapter is organized as follows. Section 4.1 combines the ideas of margin-based classification and optimization on the Stiefel manifold to give a joint linear dimensionality reduction and classification objective as well as an iterative algorithm. Illustrative examples and results on several datasets are also given in this section. Section 4.2 extends the linear approach of Section 4.1 to nonlinear dimensionality reduction for margin-based classification through the use of kernel functions. In Section 4.3, an analysis of Rademacher complexity and consistency is provided. Section 4.4 shows how the formulation can be extended to multisensor data fusion networks, discusses a physical model of wireless sensor networks, and gives experimental results of classification performance as a function of transmission power consumed in the network. Section 4.5 is a brief summary of the chapter.

■ 4.1 Linear Dimensionality Reduction for Margin-Based Classification

Decision functions of margin-based classifiers, including the SVM and the GLS classifier proposed in Chapter 3, are defined on the full-dimensional input measurement space. In this section, the input measurement vectors are linearly mapped to a reduced-dimensional space over which the decision function is defined. The linear mapping, represented by a matrix on the Stiefel manifold, and the decision function are both optimized to obtain a classifier with small regularized margin-based training loss.

■ 4.1.1 Joint Objective Functional

Recall the binary margin-based objective functional

$$\mathrm{L}(\varphi) = \sum_{j=1}^{n} \ell(y_j \varphi(\mathbf{x}_j)) + \lambda \, \mathrm{J}(\varphi), \tag{4.1}$$

presented in Section 2.2, where $\mathbf{x}_j \in \Omega \subset \mathbb{R}^D$. The proposal in this section is to formulate a joint linear dimensionality reduction and classification minimization problem by extending the margin-based functional (4.1).

The decision function φ is defined in the reduced-dimensional space $Z = \mathbf{A}^T \Omega \subset \mathbb{R}^d$, where $d \leq D$ and $\mathbf{A}$ is a linear dimensionality reduction matrix on the Stiefel manifold. Aside from including $\mathbf{A}$ in the argument of the decision function, the classification objective is left unchanged. In particular, the regularization term J is not altered, thereby

allowing any regularized margin-based classifier to be extended for dimensionality reduction. The margin-based classification objective is extended to include a $D \times d$ matrix $\mathbf{A}$ as follows:

$$\mathrm{L}(\varphi, \mathbf{A}) = \sum_{j=1}^{n} \ell(y_j \varphi(\mathbf{A}^T \mathbf{x}_j)) + \lambda \, \mathrm{J}(\varphi), \tag{4.2}$$

with the constraint $\mathbf{A} \in \mathcal{V}(D, d)$. With the definition $\tilde{\mathbf{x}} = \mathbf{A}^T \mathbf{x}$, the objective function is also

$$\mathrm{L}(\varphi, \mathbf{A}) = \sum_{j=1}^{n} \ell(y_j \varphi(\tilde{\mathbf{x}}_j)) + \lambda \, \mathrm{J}(\varphi). \tag{4.3}$$

With different loss functions ℓ and different regularization terms J, various margin-based classifiers can thus be extended for joint linear dimensionality reduction. Once φ and $\mathbf{A}$ are found, the classifier is $\hat{y}(\mathbf{x}) = \operatorname{sign}\left(\varphi\left(\mathbf{A}^T \mathbf{x}\right)\right)$.

The joint linear dimensionality reduction and classification framework is applicable to the multicategory GLS classifier proposed in Section 3.2 as well. The objective functional in the multicategory case is

$$\mathrm{L}(\varphi^{(1)}, \ldots, \varphi^{(m)}, \mathbf{A}) = \sum_{j=1}^{n} \ell(\psi_{y_j}(\mathbf{A}^T \mathbf{x}_j)) + \frac{\lambda}{m} \sum_{k=1}^{m} \oint_{\varphi^{(k)}=0} d\mathbf{s}. \tag{4.4}$$

■ 4.1.2 Coordinate Descent Minimization

The option pursued for performing the minimization of $\mathrm{L}(\varphi, \mathbf{A})$ given in (4.2) is coordinate descent: alternating minimizations with fixed $\mathbf{A}$ and with fixed φ. The problem is conceptually similar to level set image segmentation along with pose estimation for a shape prior [165], especially when the classifier is the GLS classifier. When $\mathbf{A}$ is fixed, optimization for φ is a standard margin-based classification problem in the reduced-dimensional space. If the margin-based classifier is the SVM, then this optimization step may be performed by quadratic programming techniques, as mentioned in Section 2.2.3. If the margin-based classifier is the GLS classifier, then this optimization step may be performed by contour evolution, as detailed in Chapter 3.

When φ is fixed, the problem at hand is to minimize a function of $\mathbf{A}$ lying on the Stiefel manifold, a problem touched on in Section 2.4.1. The function

$$\mathrm{L}(\mathbf{A}) = \sum_{j=1}^{n} \ell(y_j \varphi(\mathbf{A}^T \mathbf{x}_j))$$

is differentiable with respect to $\mathbf{A}$ for differentiable loss functions. The first derivative is:

$$\mathbf{L}_{\mathbf{A}} = \sum_{j=1}^{n} y_j \, \ell'(y_j \varphi(\mathbf{A}^T \mathbf{x}_j)) \mathbf{x}_j \left[\varphi_{\tilde{x}_1}(\mathbf{A}^T \mathbf{x}_j) \quad \cdots \quad \varphi_{\tilde{x}_d}(\mathbf{A}^T \mathbf{x}_j)\right]. \tag{4.5}$$

Note that $\mathbf{x}_j$ is a $D \times 1$ vector and that $\begin{bmatrix}\varphi_{\tilde{x}_1}(\mathbf{A}^T\mathbf{x}_j) & \cdots & \varphi_{\tilde{x}_d}(\mathbf{A}^T\mathbf{x}_j)\end{bmatrix}$ is a $1 \times d$ vector, where $\varphi_{\tilde{x}_k}(\cdot)$ is the partial derivative of the decision function with respect to dimension $\tilde{x}_k$ in the reduced-dimensional space. For the logistic loss function:

$$\ell'_{\text{logistic}}(z) = -\frac{e^{-z}}{1+e^{-z}}$$

and for the hinge loss function:

$$\ell'_{\text{hinge}}(z) = -\operatorname{step}(1-z),$$

where $\operatorname{step}(\cdot)$ is the Heaviside step function. In order to minimize the margin-based loss with respect to $\mathbf{A}$, gradient descent along Stiefel manifold geodesics is performed, which involves applying (2.78) and (2.79) given in Section 2.4.1 with the matrix derivative (4.5).

Similarly for the multicategory GLS classifier, the matrix derivative is:

$$\mathbf{L}_\mathbf{A} = \sum_{j=1}^{n} \ell'(\psi_{y_j}(\mathbf{A}^T\mathbf{x}_j))\mathbf{x}_j \begin{bmatrix}\psi_{y_j,\tilde{x}_1}(\mathbf{A}^T\mathbf{x}_j) & \cdots & \psi_{y_j,\tilde{x}_d}(\mathbf{A}^T\mathbf{x}_j)\end{bmatrix}, \tag{4.6}$$

where $\psi_{y_j,\tilde{x}_k}$ is the partial derivative of ψ_{y_j} with respect to dimension $\tilde{x}_k$ in the reduced d-dimensional space. The matrix derivative (4.6) is substituted into (2.78) and (2.79) in this case as well to perform gradient descent along Stiefel manifold geodesics.

■ 4.1.3 Examples

A binary example is now presented in which the input dimensionality is $D = 8$. The first two dimensions of the data, x_1 and x_2, are informative for classification and the remaining six are completely irrelevant. The first two dimensions of the data are the same as in the first example of Section 3.1.3, separable by an ellipse. The values in the other six dimensions are independent samples from an identical Gaussian distribution without regard for class label. Linear dimensionality reduction to $d = 2$ dimensions is sought.

The desiderata for this example are that the correct two-dimensional projection is identified and, assuming that it is, that the decision boundary is essentially elliptical. If the correct projection is identified, then the last six rows of the $\mathbf{A}$ matrix will be small compared to the first two rows, and the corresponding zonotope will be nearly square. Since rotations and reflections of the reduced-dimensional space are inconsequential, it is not necessary that the first two rows of $\mathbf{A}$ be the identity matrix, nor that the orientation of the zonotope $Z(\mathbf{A})$ line up with the coordinate axes. The two likelihood functions have the same mean and are not Gaussian, and thus not very amenable to FDA. The matrices obtained using PCA and FDA are given in Table 4.1 and visualized in Figure 4.1 using $Z(\mathbf{A})$. Neither PCA nor FDA is successful at recovering the informative subspace: the x_1–x_2 plane.

PCA	FDA
$\begin{bmatrix} 0.9942 & -0.0027 \\ -0.0055 & 0.0836 \\ 0.0006 & -0.1572 \\ -0.0905 & 0.3921 \\ -0.0188 & -0.5629 \\ 0.0346 & 0.1785 \\ -0.0343 & -0.1310 \\ -0.0260 & -0.6699 \end{bmatrix}$	$\begin{bmatrix} -0.0366 & 0.4280 \\ -0.0285 & -0.1829 \\ -0.0639 & 0.1498 \\ 0.0171 & -0.0587 \\ -0.2199 & 0.7443 \\ -0.0219 & 0.1451 \\ 0.8840 & 0.0175 \\ 0.4040 & 0.4268 \end{bmatrix}$

Table 4.1. Dimensionality reduction matrices produced by PCA and FDA.

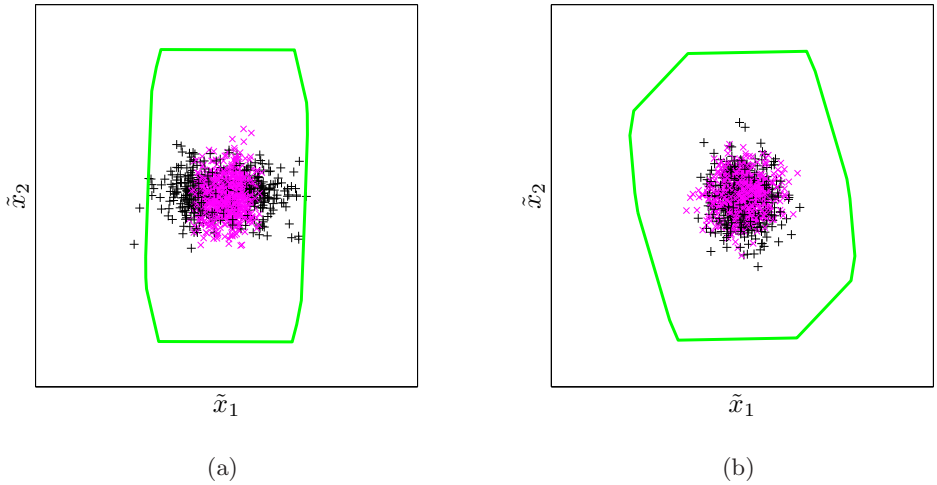


Figure 4.1. PCA and FDA projections. Magenta × markers indicate label −1. Black + markers indicate label +1. The green line outlines the zonotope $Z(\mathbf{A})$ from (a) PCA, and (b) FDA solutions.

The classification–linear dimensionality reduction coordinate descent is run on the example dataset as well. The expression (4.2) is minimized to find both an $\mathbf{A}$ matrix and decision boundary using two different margin-based classifiers: the SVM with RBF kernel and the GLS classifier of Chapter 3 with the logistic loss function. In order to show the robustness of the coordinate descent, a poor random initialization is used for $\mathbf{A}$.

The top left panel of Figure 4.2 shows the decision boundary resulting from the first optimization for φ using the GLS classifier with the random initialization for $\mathbf{A}$, before the first gradient descent step on the Stiefel manifold. As the coordinate descent progresses, the zonotope becomes more like a square, i.e., $\mathbf{A}$ aligns with the x_1–x_2 plane, and the decision boundary becomes more like an ellipse. The bottom right panel of Figure 4.2 shows the final learned classifier and linear dimensionality reduction matrix. Figure 4.3 shows the coordinate descent with the SVM. Here also, the zonotope becomes

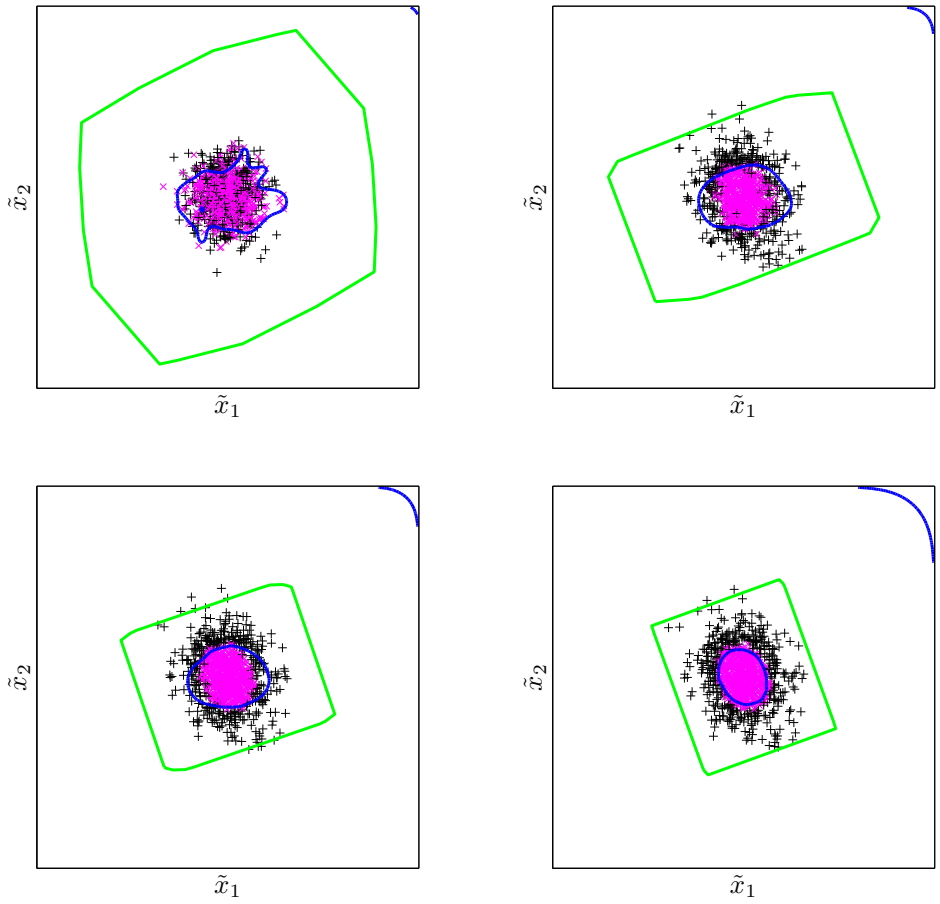


Figure 4.2. Joint linear dimensionality reduction and margin-based classification coordinate descent with the GLS classifier proceeding in raster scan order from top left to bottom right. The magenta $\times$ markers indicate class label -1 and the black $+$ markers indicate class label $+1$. The blue line is the decision boundary. The green line outlines the zonotope $Z(\mathbf{A})$.

more like a square and the decision boundary becomes more like an ellipse throughout the minimization.

The random initial $\mathbf{A}$ matrix and the final $\mathbf{A}$ matrix solutions for the GLS classifier and the SVM are given in Table 4.2. Conforming to the expected behavior, the final decision boundary is almost an ellipse and the final $\mathbf{A}$ has very little energy in the bottom six rows with both margin-based classifiers. As this example indicates, the procedure is capable of making quite large changes to $\mathbf{A}$.

As a second example, the dataset used in Section 3.1 and Section 3.5 is considered, in which the samples labeled -1 are in a strip parallel to the x_1 axis. Here the dimen-

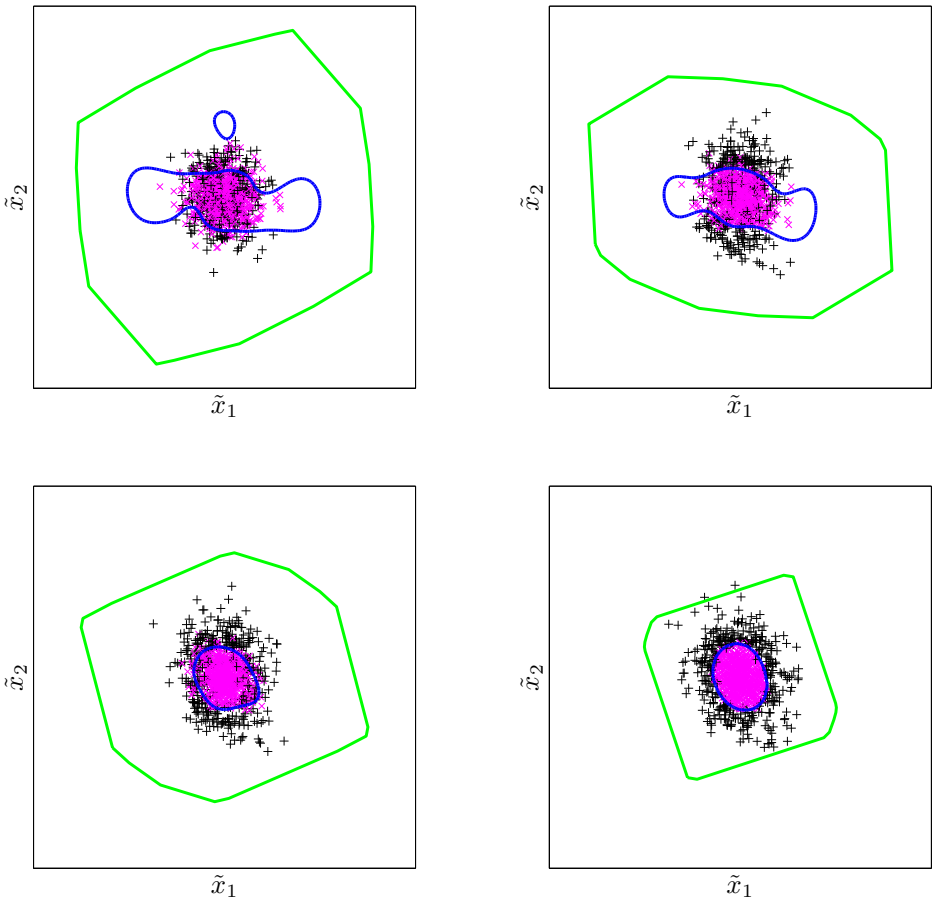


Figure 4.3. Joint linear dimensionality reduction and margin-based classification coordinate descent with the SVM classifier proceeding in raster scan order from top left to bottom right. The magenta $\times$ markers indicate class label -1 and the black $+$ markers indicate class label $+1$. The blue line is the decision boundary. The green line outlines the zonotope $Z(\mathbf{A})$.

sionality reduction is from $D = 2$ to $d = 1$ dimensions. As noted in Section 3.5, the vertical dimension x_2 is useful for classifying this dataset and the horizontal dimension is not. Therefore, the expected solution for $\mathbf{A}$ is either $\begin{bmatrix} 0 & 1 \end{bmatrix}^T$ or $\begin{bmatrix} 0 & -1 \end{bmatrix}^T$.

The margin-based classifier used in this example is the GLS classifier. The initial and final $\mathbf{A}$ matrices are given in Table 4.3 and the initial and final signed distance functions are shown in Figure 4.4. The final recovered $\mathbf{A}$ matrix is as expected with nearly all of the energy in the second input variable. The final signed distance function in Figure 4.4(b) partitions the $\tilde{x}_1 = \mathbf{A}^T\mathbf{x}$ axis as expected: it is negative in the strip where there are negatively labeled samples and positive elsewhere. (Remnants of the

Random Initialization	Coordinate Descent with GLS	Coordinate Descent with SVM
$\begin{bmatrix} 0.0274 & -0.4639 \\ 0.4275 & 0.2572 \\ 0.4848 & 0.1231 \\ -0.0644 & 0.4170 \\ 0.0138 & 0.3373 \\ 0.5523 & 0.2793 \\ 0.1333 & 0.0283 \\ 0.5043 & -0.5805 \end{bmatrix}$	$\begin{bmatrix} 0.3386 & -0.9355 \\ 0.9401 & 0.3406 \\ 0.0118 & -0.0110 \\ 0.0103 & -0.0196 \\ 0.0246 & -0.0675 \\ -0.0172 & 0.0181 \\ 0.0186 & -0.0580 \\ -0.0108 & -0.0027 \end{bmatrix}$	$\begin{bmatrix} 0.3155 & -0.9425 \\ 0.9446 & 0.3098 \\ 0.0334 & 0.0936 \\ 0.0037 & 0.0356 \\ 0.0061 & -0.0318 \\ -0.0716 & 0.0121 \\ -0.0411 & -0.0410 \\ -0.0151 & -0.0537 \end{bmatrix}$

Table 4.2. Dimensionality reduction matrices in coordinate descent minimization.

Initial	Final
$\begin{bmatrix} 0.2425 \\ 0.9701 \end{bmatrix}$	$\begin{bmatrix} -0.0180 \\ 0.9998 \end{bmatrix}$

Table 4.3. Initial and final dimensionality reduction matrices in coordinate descent minimization of joint dimensionality reduction and GLS classification.

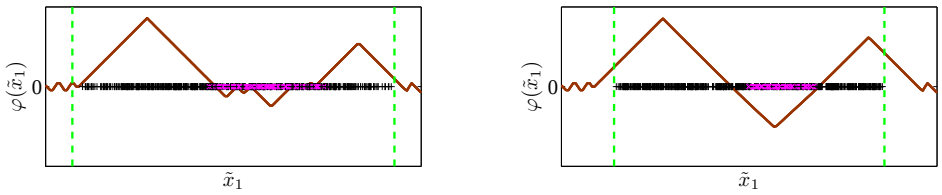
initialization remain at the edges of the domain where there are no data samples.) In the two examples of this section, it happens that the expected solutions for $\mathbf{A}$ are zero in a subset of the input variables for illustratory reasons. However, the real power of matrix optimization on the Stiefel manifold is that other linear combinations of input variables can be obtained.

■ 4.1.4 Classification Error for Different Reduced Dimensions

Experimental classification results are presented on several datasets from the UCI Machine Learning Repository [10]. The joint linear dimensionality reduction and margin-based classification method proposed earlier in this section is run for different values of the reduced dimension d, showing that performing dimensionality reduction does in fact improve classification performance in comparison to not performing dimensionality reduction. The margin-based classifier that is used is the SVM with RBF kernel and default parameter settings from the Matlab bioinformatics toolbox.

Training error and test error as a function of the reduced dimension are investigated for five different datasets from varied application domains: Wisconsin diagnostic breast cancer ($D = 30$), ionosphere ($D = 34$), sonar ($D = 60$), arrhythmia ($D = 274$ after preprocessing to remove dimensions containing missing values), and arcene ($D = 10000$). On the first four datasets, tenfold cross-validation training and test errors are examined. The arcene dataset has separate training and validation sets, which are used.

For the initialization of $\mathbf{A}$, estimates of the mutual informations between the label y and individual data dimensions x_k, $k = 1, \ldots, D$, are used. The first column of $\mathbf{A}$ is taken to be the canonical unit vector corresponding to the dimension with the largest



(a) (b)

Figure 4.4. Joint linear dimensionality reduction and GLS classification. The magenta $\times$ markers indicate class label -1 and the black $+$ markers indicate class label $+1$. The maroon line is the signed distance function whose zero level set forms the decision boundary. The dashed green line delimits the zonotope generated by $\mathbf{A}^T$. The initial and final dimensionality reduction projection and signed distance function are in (a) and (b) respectively.

mutual information. The second column of $\mathbf{A}$ is taken to be the canonical unit vector corresponding to the dimension with the second largest mutual information, and so on. The last, i.e. dth, column of $\mathbf{A}$ is taken to be zero in the rows already containing ones in the first $(d-1)$ columns, and nonzero in the remaining rows with values proportional to the mutual informations of the remaining dimensions. Kernel density estimation is used in estimating mutual information.

The tenfold cross-validation training error is shown with a blue line and the tenfold cross-validation test error is shown with a red line for the first four datasets in Figure 4.5. Figure 4.6 gives the training and test performance for the arcene dataset. For the Wisconsin diagnostic breast cancer, ionosphere, and sonar datasets, classification performance is shown for all possible reduced dimensions. For the arrhythmia and arcene datasets, reduced dimensions up to $d = 50$ and $d = 100$ are shown, respectively.

The first thing to notice in the plots is that the training error quickly converges to zero with an increase in the reduced dimension d. The margin-based classifier with linear dimensionality reduction perfectly separates the training set when the reduced dimension is sufficiently large. However, this perfect separation does not carry over to the test error—the error of most interest. In all of the datasets, the test error first decreases as the reduced dimension is increased, but then starts increasing. There is an intermediate optimal value for the reduced dimension. For the five datasets, these values are $d = 3$, $d = 9$, $d = 16$, $d = 10$, and $d = 20$, respectively. This test error behavior is evidence of overfitting if d is too large. Dimensionality reduction improves classification performance on unseen samples by preventing overfitting. Remarkably, even the ten thousand-dimensional measurements in the arcene dataset can be linearly reduced to twenty dimensions.

The classification error as a function of d using the proposed joint linear dimen-

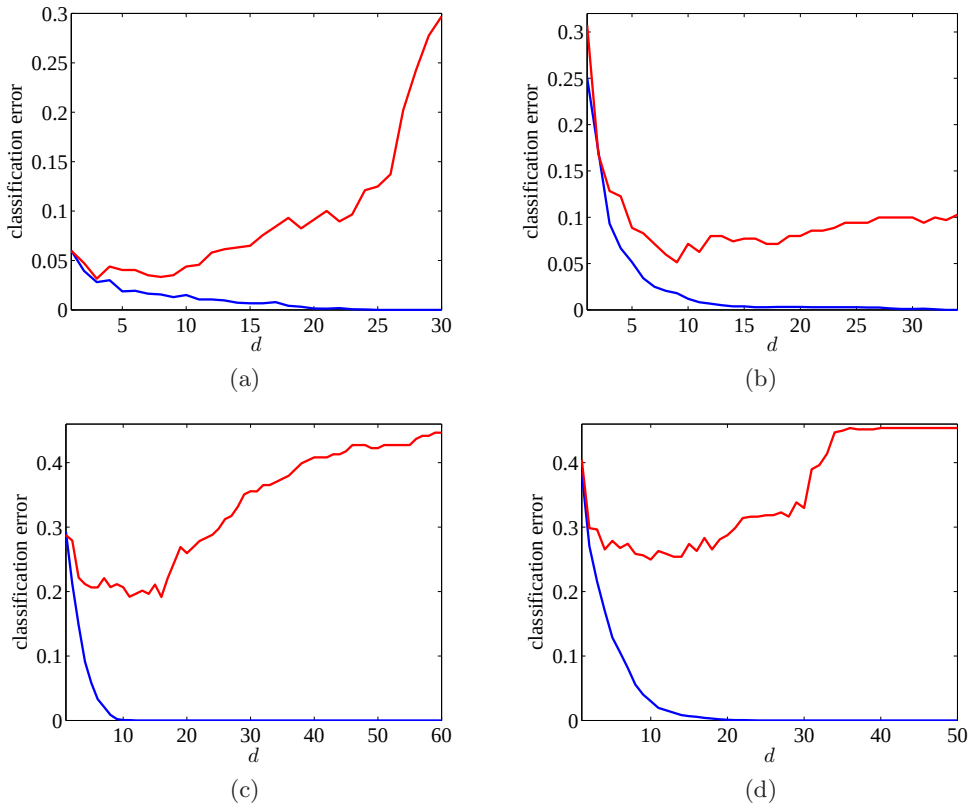


Figure 4.5. Tenfold cross-validation training error (blue line) and test error (red line) on (a) Wisconsin diagnostic breast cancer, (b) ionosphere, (c) sonar, and (d) arrhythmia datasets.

sionality reduction and margin-based classification method matches the structural risk minimization principle. Rademacher complexity analysis supporting these empirical findings is presented in Section 4.3. If generalization error is the only criterion, then any popular model selection method from the statistical learning literature, including those based on cross-validation, bootstrapping, and information criteria, can be used to find a good value for the reduced dimension d. However, other criteria besides generalization error become important in various settings, including sensor networks discussed in Section 4.4.

Test error using the joint linear dimensionality reduction and margin-based classification method may be compared to the test error when dimensionality is first reduced using PCA or FDA followed by classification using the same classifier, the SVM with RBF kernel and default settings. This comparison is given for the ionosphere and sonar datasets in Figure 4.7. The minimum test error is achieved by the joint minimization. This is to be expected, as the dimensionality reduction is matched to the classifier when the joint objective is optimized.

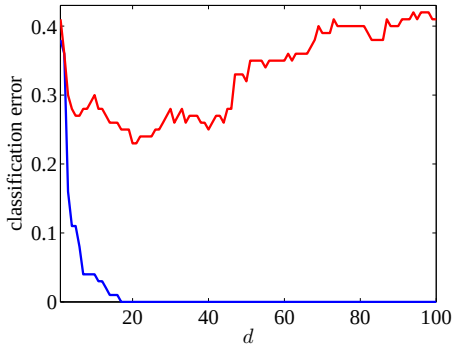


Figure 4.6. Training error (blue line) and test error (red line) on arcene dataset.

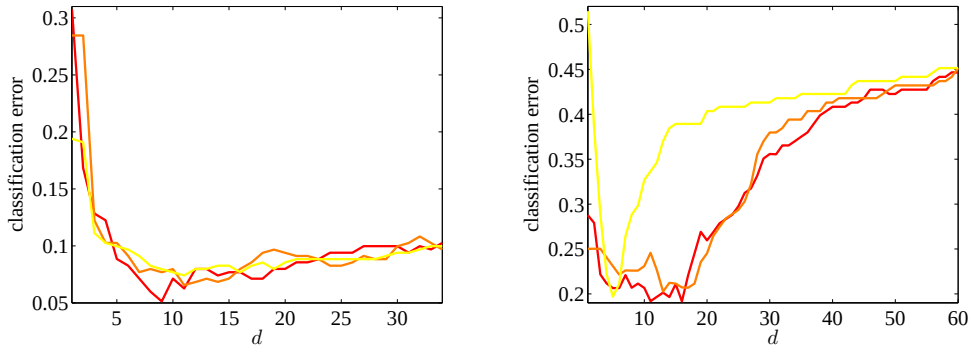


Figure 4.7. Tenfold cross-validation test error on (a) ionosphere and (b) sonar datasets with PCA (yellow line), FDA (orange line), and joint linear dimensionality reduction and margin-based classification (red line).

■ 4.2 Nonlinear Dimensionality Reduction for Margin-Based Classification

This section extends Section 4.1 to *nonlinear* dimensionality reduction. As discussed at the beginning of the chapter, several methods exist in the literature for nonlinear dimensionality reduction in the unsupervised setting, also known as manifold learning, including Isomap [193], locally linear embedding [166], and Laplacian eigenmaps [17]. Methods such as [86] utilize a variational energy minimization formulation and an implicit representation of the low-dimensional manifold, like variational level set methods, but also in the unsupervised setting.

The use of kernel functions allows the extension of linear dimensionality reduction methods to nonlinear dimensionality reduction [176], as has been done for PCA [177] and FDA [15, 128]. Bengio et al. [19] have shown that Isomap, locally linear embedding, and

Laplacian eigenmaps are versions of kernel PCA for different data-dependent kernels. The approach followed in this section is to take the data-dependent kernel for Isomap [19], and use it in the context of the joint dimensionality reduction and margin-based classification problem instead of the PCA problem.

■ 4.2.1 Kernel-Based Nonlinear Formulation

Recall from Section 2.4.3 the data-dependent kernel function corresponding to the manifold learning technique Isomap:

$$\mathrm{K}(\mathbf{w},\mathbf{z}) = -\frac{1}{2}\left(\rho(\mathbf{w},\mathbf{z})^2 - \frac{1}{n}\sum_{j=1}^{n}\rho(\mathbf{w},\mathbf{x}_j)^2 - \frac{1}{n}\sum_{j=1}^{n}\rho(\mathbf{x}_j,\mathbf{z})^2 + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{j'=1}^{n}\rho(\mathbf{x}_j,\mathbf{x}_{j'})^2\right), \tag{4.7}$$

where $\mathbf{x}_1,\ldots,\mathbf{x}_n \in \mathbb{R}^D$ are the given data samples, and $\mathbf{w}$ and $\mathbf{z}$ are general points in $\mathbb{R}^D$. The distance ρ is an approximation to geodesic distance on the manifold calculated by means of a Euclidean k-nearest neighbor graph of the given data samples. The distance from a point not among the set of data samples to another sample is simply calculated by first finding the k-nearest neighbors in the dataset of that point. Then the distance is the minimum, among the neighbors, of the sum of the distance between the point and the neighbor, and the graph-based geodesic distance from the neighbor to the other sample.

Also recall the definition $\tilde{\mathbf{K}}(\mathbf{x}) = \begin{bmatrix}\mathrm{K}(\mathbf{x},\mathbf{x}_1) & \cdots & \mathrm{K}(\mathbf{x},\mathbf{x}_n)\end{bmatrix}^T$. The dimensionality-reduced mapping of the sample $\mathbf{x}$ is $\tilde{\mathbf{x}} = \mathbf{A}^T\tilde{\mathbf{K}}(\mathbf{x})$, where $\mathbf{A}$ is an $n \times d$ matrix with orthogonal columns. In applications such as supervised classification where the scaling of the different dimensions on the reduced-dimensional manifold is not important, $\mathbf{A}$ can be taken to be a member of $\mathcal{V}(n,d)$. Note that by using this nonlinear kernel function representation, the matrix $\mathbf{A}$ has n rows as opposed to D rows in the linear dimensionality reduction case. If n is greater than D, it is possible to use this formulation for dimensionality expansion rather than dimensionality reduction, but that is not pursued here.

For the purposes of margin-based classification, a joint objective similar to the linear dimensionality reduction functional (4.2) is proposed for nonlinear dimensionality reduction as follows.

$$\mathrm{L}(\varphi,\mathbf{A}) = \sum_{j=1}^{n}\ell(y_j\varphi(\mathbf{A}^T\tilde{\mathbf{K}}(\mathbf{x}_j))) + \lambda\,\mathrm{J}(\varphi), \tag{4.8}$$

with the constraint $\mathbf{A} \in \mathcal{V}(n,d)$. As in the linear dimensionality reduction case, the decision function φ is defined in the reduced d-dimensional space. The kernel functions $\tilde{\mathbf{K}}$ remain fixed throughout. A coordinate descent minimization procedure is followed as for linear dimensionality reduction. Since $\tilde{\mathbf{K}}(\mathbf{x})$ is fixed, the optimization is no different than in the linear case.

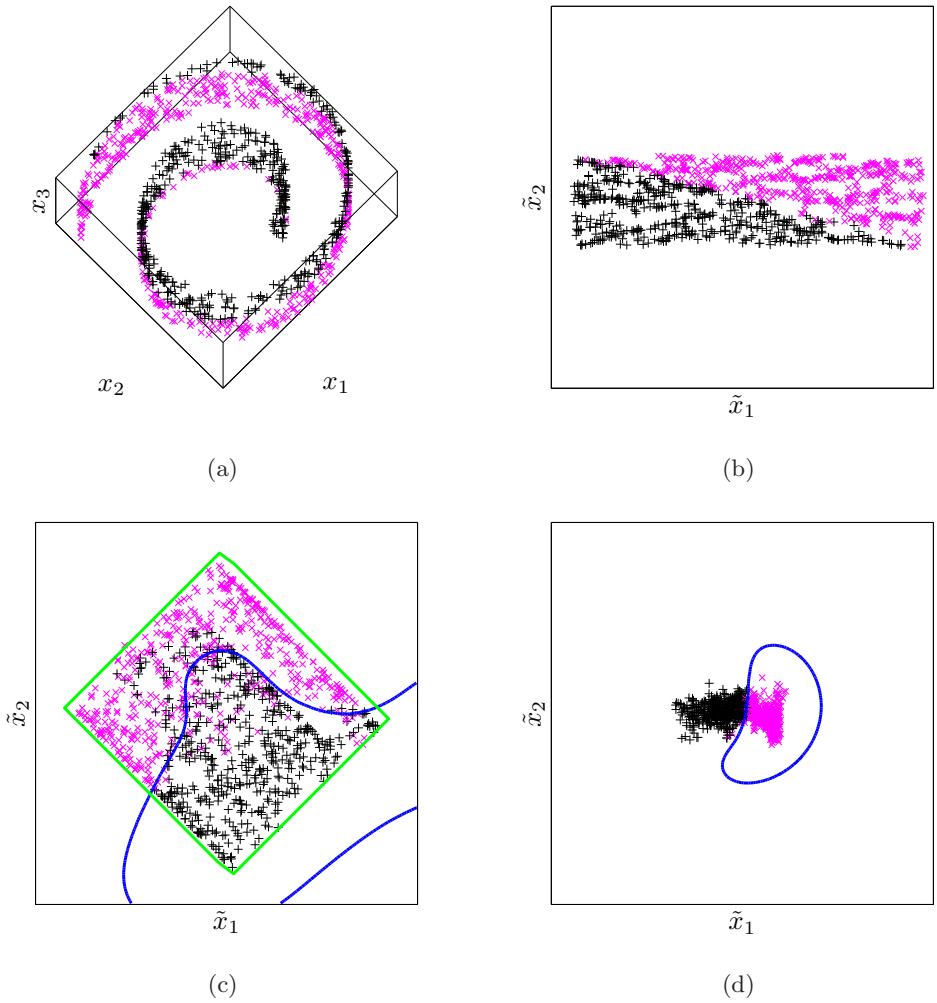


Figure 4.8. The (a) swiss roll dataset, (b) its Isomap embedding, (c) joint linear dimensionality reduction and SVM classification solution, and (d) joint nonlinear dimensionality reduction and SVM classification solution. The magenta $\times$ markers indicate class label -1 and the black $+$ markers indicate class label $+1$. The blue line is the decision boundary. The green line outlines the zonotope $Z(\mathbf{A})$ in (c).

Belkin et al. [18] have proposed a statistical learning methodology involving manifold regularization that is applicable to supervised margin-based classification. The manifold regularization in that framework exploits the concentration of samples in a high-dimensional space on a low-dimensional manifold, but does not provide a mapping or embedding to the low-dimensional manifold. The decision function φ is defined in the D-dimensional space rather than in the reduced d-dimensional space as in the method proposed in this section. In fact, the manifold regularization term proposed in

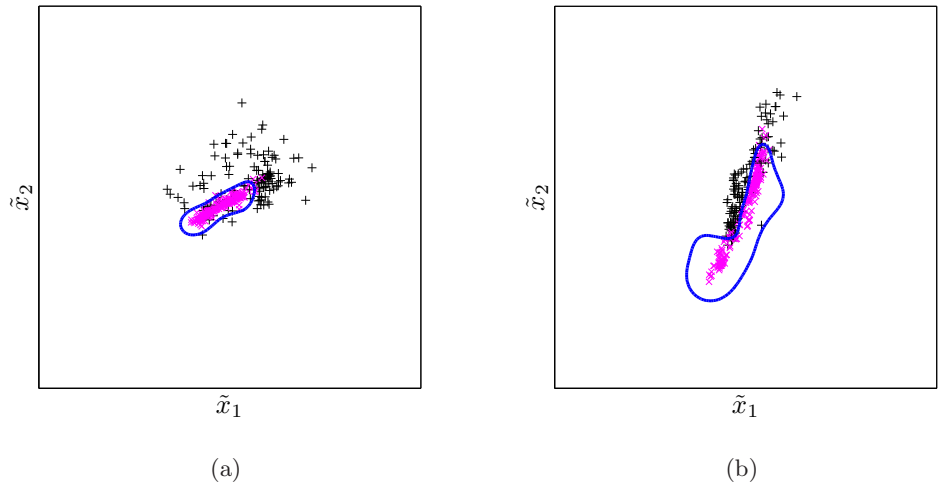


Figure 4.9. Joint (a) linear and (b) nonlinear dimensionality reduction and SVM classification solutions for the ionosphere dataset. The magenta $\times$ markers indicate class label -1 and the black $+$ markers indicate class label $+1$. The blue line is the decision boundary.

[18] parallels the feature subset selection term proposed in Section 3.5.

■ 4.2.2 Examples

Examples of nonlinear dimensionality reduction for margin-based classification are presented on one synthetic dataset and two real-world datasets. The synthetic dataset, a modification of the swiss roll dataset presented in [193], consists of a two-dimensional manifold embedded in three dimensions. The other two examples are the 34-dimensional ionosphere dataset and 60-dimensional sonar dataset from the UCI Repository [10]. For all three, the dimensionality is reduced nonlinearly to $d = 2$.

The first $n = 1000$ samples of the swiss roll dataset from [193] are taken as the $\mathbf{x}_j \in \mathbb{R}^3$, with the dimensions normalized to have unit variance and zero mean. These samples lie on a two-dimensional plane that has been rolled into a spiral in the three-dimensional space. The class labels are assigned so that $y_j = -1$ on one side of a diagonal line on that two-dimensional plane and $y_j = +1$ on the other side. The dataset is shown in Figure 4.8(a). The Isomap embedding of the points without regard to class label is shown in Figure 4.8(b).

Joint dimensionality reduction and margin-based classification is run on the dataset with the SVM with RBF kernel. The solution obtained with linear dimensionality reduction is given in Figure 4.8(c) whereas the solution with nonlinear dimensionality reduction is given in Figure 4.8(d). The example has been constructed so that it is not very suitable to linear dimensionality reduction to $d = 2$ dimensions, but is suited to nonlinear dimensionality reduction. The classifier with the linear mapping misclassifies

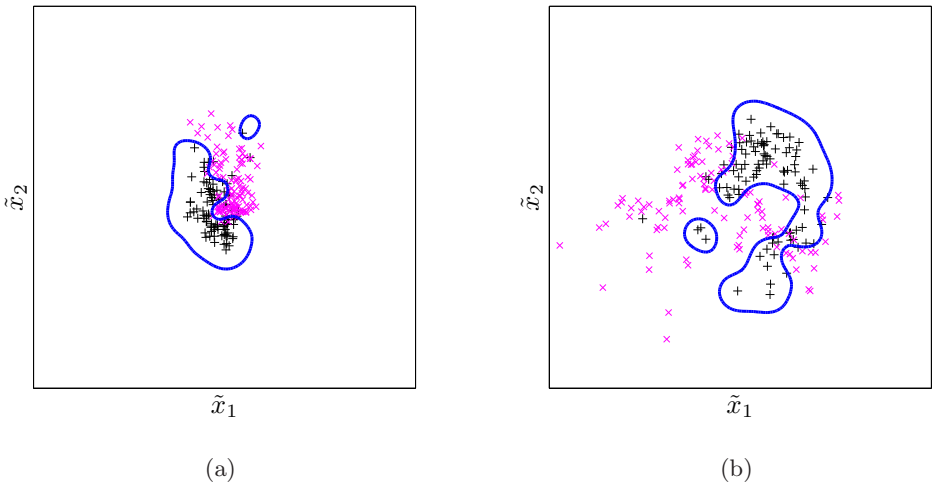


Figure 4.10. Joint (a) linear and (b) nonlinear dimensionality reduction and SVM classification solutions for the sonar dataset. The magenta $\times$ markers indicate class label -1 and the black $+$ markers indicate class label $+1$. The blue line is the decision boundary.

many samples, but the classifier with nonlinear mapping does not. This dataset is an instance where nonlinear dimensionality reduction is vastly superior to linear dimensionality reduction. Importantly, nonlinear dimensionality reduction for margin-based classification with the Isomap kernel does not produce the same embedding as the Isomap algorithm itself.

Figure 4.9 and Figure 4.10 show solutions to joint margin-based classification with both linear and nonlinear dimensionality reduction on the ionosphere and sonar datasets respectively. With these real-world datasets, the first thing to notice is that the linear and nonlinear mappings are different from each other. With the ionosphere dataset for example, the linear mapping has the $+1$ labeled samples surrounding the -1 labeled samples on three sides. The nonlinear mapping, in contrast, has wrapped the space so that the $+1$ labeled samples are on one or two sides of the -1 labeled samples. The linear and nonlinear mappings with the sonar datasets are also different from each other.

However, with the real-world datasets, it is not clear whether the linear or nonlinear dimensionality reduction is superior for use alongside the nonlinear kernel SVM classifier or other nonlinear margin-based classifiers. Since real-world datasets are neither arbitrary nor adversarial to linear dimensionality reduction, but have structure associated with them, linear dimensionality reduction with nonlinear margin-based classification may provide a good balance for generalization. An investigation of classification error as a function of nonlinearly reduced dimension and a comparison to linear dimensionality reduction is given next.

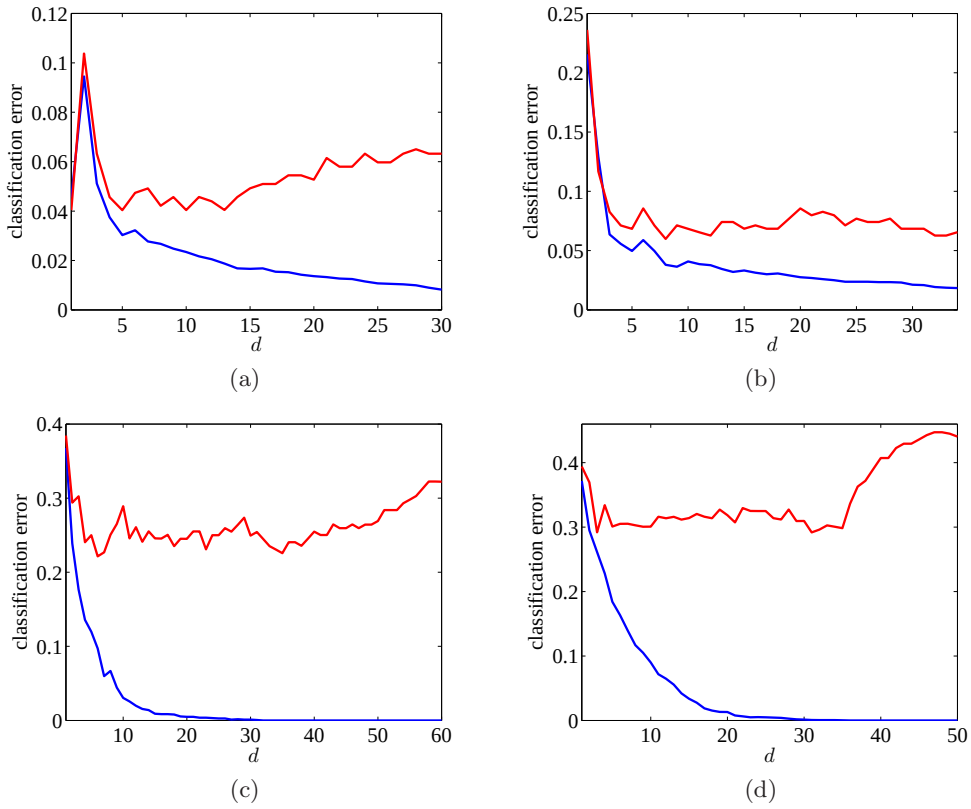


Figure 4.11. Tenfold cross-validation training error (blue line) and test error (red line) on (a) Wisconsin diagnostic breast cancer, (b) ionosphere, (c) sonar, and (d) arrhythmia datasets with nonlinear Isomap kernel.

■ 4.2.3 Classification Error for Different Reduced Dimensions

Training and test error as a function of the reduced dimension d is presented for the same five datasets considered in Section 4.1.4: Wisconsin diagnostic breast cancer, ionosphere, sonar, arrhythmia, and arcene. Also as in that section, the margin-based classifier is the SVM with RBF kernel and default parameters from the Matlab bioinformatics toolbox. The same mutual information-based initialization for $\mathbf{A}$ is used as well.

The approximation to geodesic distance used in the Isomap kernel is calculated using a k-nearest neighbor graph. The value of k has not been found to have much of an effect on training and test error. Tenfold cross-validation errors for the first four datasets with $k = 7$ neighbors in the graph are plotted in Figure 4.11. Figure 4.12 gives the training and test performance for the arcene dataset with $k = 11$ neighbors in the graph. Smaller values of k with the arcene dataset produce a graph with more than one connected component, which is not desirable. Training error tends to decrease with d,

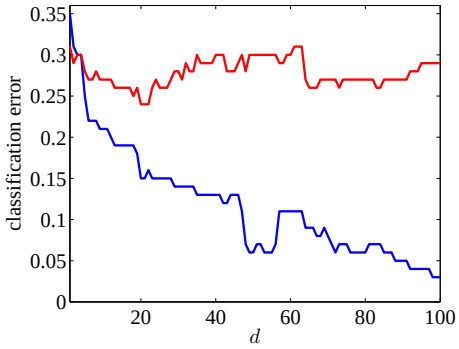


Figure 4.12. Training error (blue line) and test error (red line) on arcene dataset with nonlinear Isomap kernel.

whereas test error tends to first decrease and then increase. These plots indicate that the structural risk minimization principle and overfitting are in play with nonlinear dimensionality reduction just as they are with linear dimensionality reduction.

A comparison of test error when using nonlinear dimensionality reduction and when using linear dimensionality reduction is also provided. Figure 4.13 plots tenfold cross-validation test error as a function of d for the first four datasets with the red line being the test error for nonlinear dimensionality reduction and the magenta line being the test error for linear dimensionality reduction. This figure shows the same values as Figure 4.5 and Figure 4.11, but on the same axis for easy comparison. The test errors are compared for the arcene dataset in Figure 4.14. The values in this figure also appear in Figure 4.6 and Figure 4.12.

Nonlinear dimensionality reduction has smaller test error than linear dimensionality reduction for larger d. However, linear dimensionality reduction has smaller minimum test error. At optimal reduced dimensionality, linear dimensionality reduction is better on these real-world datasets, but overfits more when d is large. Linear dimensionality reduction with a nonlinear margin-based classifier seems to be well-suited for small classification generalization error.

■ 4.3 Complexity and Consistency Analysis

This section provides a theoretical characterization of the Rademacher complexity of dimensionality-reduced GLS classifiers that minimize the functional (4.2). It also shows that dimensionality-reduced GLS classifiers are consistent. The main tool used in both analyses is the ϵ-entropy of the function class that contains the decision function φ [106]. The analysis mirrors that given in Chapter 3, but with the additional ingredient of the dimensionality reduction mapping, which necessitates the use of zonotope content in calculating ϵ-entropy.

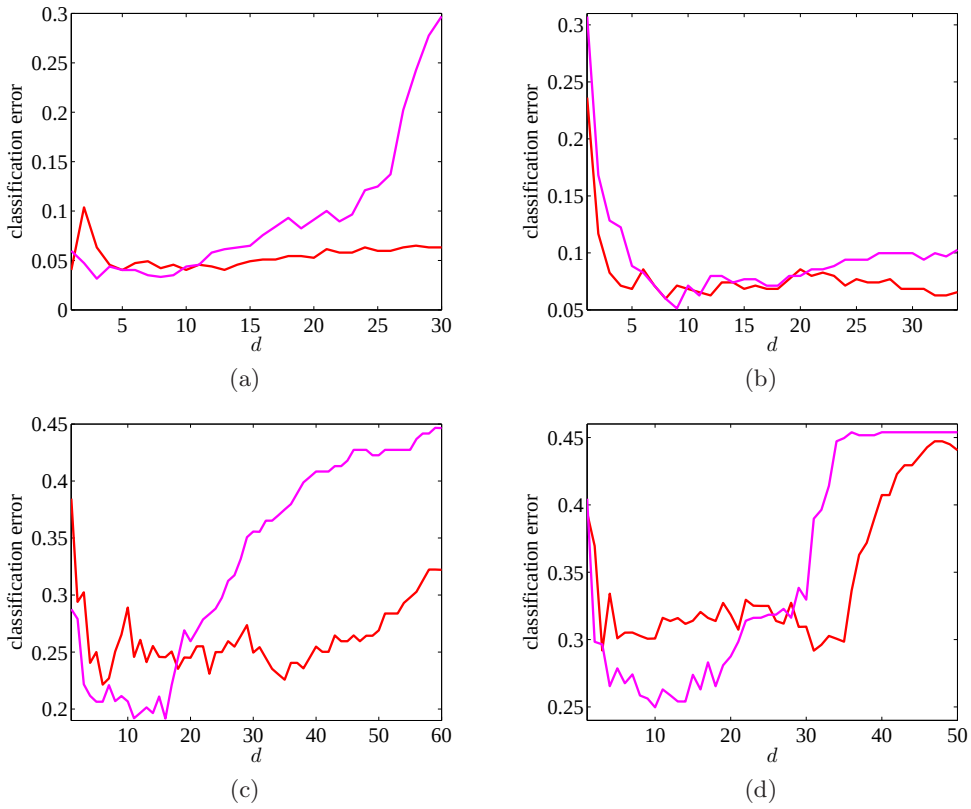


Figure 4.13. Comparison of tenfold cross-validation test error with nonlinear (red line) and linear (magenta line) dimensionality reduction on (a) Wisconsin diagnostic breast cancer, (b) ionosphere, (c) sonar, and (d) arrhythmia datasets.

■ 4.3.1 Epsilon Entropy

As discussed in Chapter 3, scaling and shifting of the data is often performed in classification, and as in that chapter, the domain of the unreduced measurement vectors is taken to be the unit hypercube, that is $\mathbf{x} \in \Omega = [0,1]^D$. The reduced-dimensional domain is then the zonotope $Z = \mathbf{A}^T\Omega \subset \mathbb{R}^d$, where $\mathbf{A}$ is on the Stiefel manifold. The set of signed distance functions φ defined on Ω is denoted $\mathcal{F}_\Omega$ and the set defined on Z is denoted $\mathcal{F}_Z$. The surface area constraint included in the analysis of Chapter 3 is not included in the analysis in this chapter for simplicity, but could be included in a straightforward manner.

For GLS classification without dimensionality reduction, it is shown in Section 3.4 that

$$H_{\rho_\infty,\epsilon}(\mathcal{F}_\Omega) \leq \nu^D, \tag{4.9}$$

where $\nu = \lceil 1/\epsilon \rceil$. This result follows from the fact that ν^D D-dimensional hypercubes

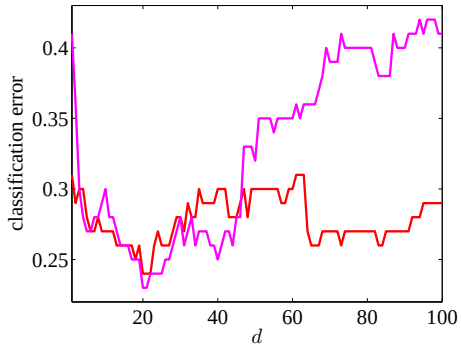


Figure 4.14. Comparison of test error with nonlinear (red line) and linear (magenta line) dimensionality reduction on arcene dataset.

with side of length ϵ fit as a Cartesian grid into $\Omega = [0,1]^D$. To find an expression for the ϵ-entropy of the dimensionality-reduced GLS classifier, analysis of the same type applies. Consequently, the number of d-dimensional hypercubes with side of length ϵ that fit into Z must be determined.

The number of small hypercubes that fit inside Z is related to its content $V(Z)$. Based on the zonotope content inequality (2.82) given in Section 2.4, it is found that

$$H_{\rho_\infty,\epsilon}(\mathcal{F}_Z) \leq V(Z)\nu^d \leq \omega_d \left(\frac{\omega_{d-1}}{\omega_d} \sqrt{\frac{D}{d}} \right)^d \nu^d, \tag{4.10}$$

where ω_d is the content of the d-dimensional unit hypersphere. For fixed reduced dimension d, $H_{\rho_\infty,\epsilon}(\mathcal{F}_Z)$ increases as a function of the measurement dimension D, i.e., the classifier function class is richer for larger measurement dimension with the same reduced-dimension. Importantly, $H_{\rho_\infty,\epsilon}(\mathcal{F}_Z)$ increases as a function of d for fixed D.

The function class $\mathcal{F}_Z$, and consequently $H_{\rho_\infty,\epsilon}(\mathcal{F}_Z)$ is tied to the specific margin-based classification method employed. The GLS classifier has been selected in order to make concrete statements. Similar analysis may be performed for other margin-based classifiers such as the kernel SVM.

■ 4.3.2 Rademacher Complexity

It has been seen empirically that training error decreases as a function of d, but that test error (a surrogate for generalization error) first decreases and then increases. The generalization bound (2.32) discussed in Section 2.2.2 suggests that this increase in test error is due to overfitting caused by high complexity of the classifier function class $\mathcal{F}_Z$ when d is large. As discussed in [25, 131, 230], dimensionality reduction reduces classifier complexity and thus prevents overfitting. Here, the Rademacher complexity term $C_n^{\text{Rad}}(\mathcal{F}_Z)$ is analytically characterized for the joint dimensionality reduction and margin-based classification method proposed in this chapter.

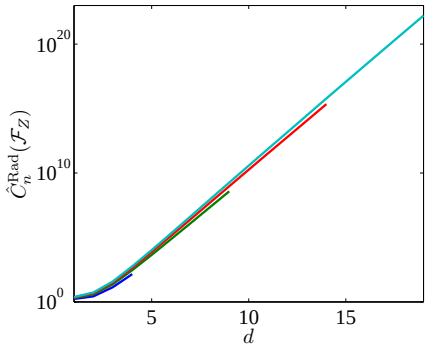


Figure 4.15. Rademacher complexity as a function of the reduced dimension d for $D = 5$ (blue line), $D = 10$ (green line), $D = 15$ (red line), and $D = 20$ (cyan line) for $\epsilon = 0.01$ and $n = 1000$.

As also noted in Section 3.4.2, von Luxburg and Bousquet [121] show that the Rademacher complexity of a function class $\mathcal{F}_Z$ satisfies:

$$\hat{C}_n^{\mathrm{Rad}}(\mathcal{F}_Z) \leq 2\epsilon + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\frac{\epsilon}{4}}^{\infty} \sqrt{H_{\rho_\infty,\epsilon'}(\mathcal{F}_Z)}d\epsilon'. \tag{4.11}$$

Substituting the $H_{\rho_\infty,\epsilon}(\mathcal{F}_Z)$ expression (4.10) into (4.11), it is found that for a fixed measurement dimension D, the more the dimensionality is reduced, that is the smaller the value of d, the smaller the Rademacher complexity. This is shown in Figure 4.15, a plot of the complexity value as a function of d for different values of D. Although larger measurement dimension D does result in larger complexity, the effect is minor in comparison to the effect of d.

Since training error increases as d decreases, and the generalization error is related to the sum of the Rademacher complexity and the training error: the more the dimension is reduced, the more that overfitting is prevented. However, if the dimension is reduced too much, the classifier ends up underfitting the data; the training error component of the generalization error becomes large. There is an optimal reduced dimension that balances the training error and the complexity components of the generalization error. This theoretical conclusion of having an intermediate reduced dimension at which generalization error is minimized agrees with the empirical findings of Section 4.1.4.

■ 4.3.3 Consistency

As discussed in Section 2.2.1, with a training dataset of cardinality n drawn from $f_{\mathbf{x},y}(\mathbf{x}, y)$, a consistent classifier is one whose generalization error converges in the limit as n goes to infinity to the probability of error of the Bayes optimal decision rule. For consistency to be at all meaningful, it is assumed in this analysis that there is a reduced-dimensional statistic $\tilde{\mathbf{x}} = \mathbf{A}^T\mathbf{x}$ so that the optimal Bayes decision rule based on this statistic achieves the same performance as the optimal decision rule based on

the complete data $\mathbf{x}$. I.e., it is assumed that there exists at least one $\mathbf{A}^* \in \mathcal{V}(D,d)$ such that $\mathrm{R}(\hat{y}^*(\mathbf{A}^{*T}\mathbf{x})) = \mathrm{R}(\hat{y}^*(\mathbf{x}))$, where $\hat{y}^*$ takes the appropriate-dimensional argument, and d is known.[3]

The question is whether for a sequence of classifiers learned from training data $\hat{y}^{(n)}(\mathbf{x}) = \mathrm{sign}(\varphi^{(n)}(\mathbf{A}^{(n)T}\mathbf{x}))$, where

$$(\mathbf{A}^{(n)}, \varphi^{(n)}) = \arg \min_{\mathbf{A}\in\mathcal{V}(D,d)} \min_{\varphi\in\mathcal{F}_{Z(\mathbf{A})}} \sum_{j=1}^{n} \ell(y_j \varphi(\mathbf{A}^T\mathbf{x}_j)),$$

does $\mathrm{R}(\hat{y}^{(n)}) - \mathrm{R}(\hat{y}^*)$ converge in probability to zero. The answer follows the same development as in Section 3.4.3.

First, the margin-based loss function ℓ must be a Fisher-consistent loss function [116]. Given that ℓ is Fisher-consistent, Theorem 4.1 of [116] is applied to show consistency. Noting that signed distance functions on Z are bounded in the L_∞ norm, and that there exists a constant $B > 0$ such that $H_{\rho_\infty,\epsilon}(\mathcal{F}_Z) \leq B\epsilon^{-d}$, which follows from (4.10), applying the theorem yields

$$\mathrm{R}(\hat{y}^{(n)}) - \mathrm{R}(\hat{y}^*) = \mathrm{O}_P(n^{-\tau}), \tag{4.12}$$

where

$$\tau = \begin{cases} \frac{1}{3}, & d = 1 \\ \frac{1}{4} - \frac{\log\log n}{2\log n}, & d = 2 \\ \frac{1}{2d}, & d \geq 3 \end{cases}.$$

The dimensionality reduction and classification method is consistent: $\mathrm{R}(\hat{y}^{(n)}) - \mathrm{R}(\hat{y}^*)$ goes to zero as n goes to infinity because $n^{-\tau}$ goes to zero.

■ 4.4 Application to Sensor Networks

The problem of *distributed* detection has been an object of study during the last thirty years [38, 194, 203, 213], but the majority of the work has focused on the situation when either the joint probability distribution of the measurements and labels or the likelihood functions of the measurements given the labels are assumed known. Recently, there has been some work on supervised classification for distributed settings when only training samples, not the distributions, are available [138, 157, 158]. Sensor networks are systems used for distributed detection and data fusion that operate with severe resource limitations; consequently, minimizing complexity in terms of communication and computation is critical [36]. A current interest is in deploying sensor networks with nodes that take measurements using many heterogeneous modalities such as acoustic, infrared, seismic, and video [226, 227]. The sensors measure high-dimensional data, but

[3] Consistency statements similar to the one presented here may be made for cases when such an $\mathbf{A}^* \in \mathcal{V}(D,d)$ does not exist; then convergence of $\mathrm{R}(\hat{y}^{(n)})$ is to the probability of error of the Bayes optimal decision rule in the best linear subspace of dimension d.

it is not known in advance which dimensions or combination of dimensions are most useful for the detection or classification task. Resources can be conserved if sensors do not transmit irrelevant or redundant data. The transmission of irrelevant and redundant data can be avoided through dimensionality reduction [210, 212]. Previous work on the linear dimensionality reduction of sensor measurements in distributed settings, including [77, 167, 175] and references therein, have estimation rather than detection or classification as the objective.

A classification paradigm that intelligently reduces the dimensionality of measurements locally at sensors before transmitting them is critical in sensor network settings. In this section, it is shown how the dimensionality reduction of heterogeneous data specifically for margin-based classification may be distributed in a tree-structured multisensor data fusion network with a fusion center via individual Stiefel manifold matrices at each sensor. The coordinate descent learning algorithm proposed in this chapter is amenable to distributed implementation. In particular, the coordinate descent procedure is extended so that it can be implemented in tree-structured sensor networks through a message-passing approach with the amount of communication related to the reduced dimension rather than the full measurement dimension.

Multisensor networks lead to issues that do not typically arise in statistical learning, where generalization error is the only criterion. In sensor networks, resource usage presents an additional criterion to be considered, and the architecture of the network presents additional design freedom. In wireless sensor networks, the distance between nodes affects energy usage in communication, and must therefore be considered in selecting network architecture. Classification results are given for different network architectures and these issues are touched on empirically.

For ease of exposition, the discussion begins by first considering a setup with a single sensor, and then comes to the general setting with m sensors networked according to a tree graph with a fusion center at the root of the tree. Also for simplicity of exposition, it is assumed that the fusion center does not take measurements, that it is not also a sensor; this assumption is by no means necessary. Additionally, only binary classification and linear dimensionality reduction is described, but multicategory classification and the nonlinear extension developed in Section 4.2 may be used in the sensor network setting as well. It is assumed, as in [138, 157, 158], that the class labels of the training set are available at the fusion center.

■ 4.4.1 Network with Fusion Center and Single Sensor

Consider a network with a single sensor and a fusion center. The sensor measures data vector $\mathbf{x} \in \mathbb{R}^D$ and reduces its dimensionality using $\mathbf{A}$. The sensor transmits $\tilde{\mathbf{x}}_{\mathrm{s}\to\mathrm{fc}} = \mathbf{A}^T\mathbf{x} \in \mathbb{R}^d$ to the fusion center, which applies decision rule $\mathrm{sign}(\varphi(\tilde{\mathbf{x}}_{\mathrm{s}\to\mathrm{fc}}))$ to obtain a classification for $\mathbf{x}$. Clearly in its operational phase, the dimensionality reduction reduces the amount of transmission required from the sensor to the fusion center.

Moreover, the communication required in training depends on the reduced dimension

d rather than the dimension of the measurements D. The coordinate descent procedure described in Section 4.1.2 is naturally implemented in this distributed setting. With $\mathbf{A}$ fixed, the optimization for φ occurs at the fusion center. The information needed by the fusion center to perform the optimization for φ are the $\tilde{\mathbf{x}}_{\mathrm{s}\to\mathrm{fc},j}$, the dimensionality-reduced training examples. With φ fixed, the optimization for $\mathbf{A}$ occurs at the sensor. Looking at the expression for the matrix derivative $\mathbf{L}_{\mathbf{A}}$ (4.5), it is seen that the information required by the sensor from the fusion center to optimize $\mathbf{A}$ includes only the scalar value $y_j\,\ell'(y_j\varphi(\tilde{\mathbf{x}}_{\mathrm{s}\to\mathrm{fc},j}))$ and the column vector $\left[\varphi_{\tilde{x}_1}(\tilde{\mathbf{x}}_{\mathrm{s}\to\mathrm{fc},j}) \quad \cdots \quad \varphi_{\tilde{x}_d}(\tilde{\mathbf{x}}_{\mathrm{s}\to\mathrm{fc},j})\right]^T$, denoted in this section as $\tilde{\boldsymbol{\varphi}}'_{\mathrm{fc}\to\mathrm{s},j} \in \mathbb{R}^d$, for $j = 1, \ldots, n$.

Thus the alternating minimizations of the coordinate descent are accompanied by the alternating communication of messages $\tilde{\mathbf{x}}_{\mathrm{s}\to\mathrm{fc},j}$ and $\tilde{\boldsymbol{\varphi}}'_{\mathrm{fc}\to\mathrm{s},j}$. The more computationally demanding optimization for φ (the application of a margin-based classification algorithm) takes place at the fusion center. A computationally simple Stiefel manifold gradient update occurs at the sensor. This scheme extends to the more interesting case of *multisensor* networks, as described next.

■ 4.4.2 Multisensor Networks

Now consider networks with m sensors connected in a tree topology with the fusion center at the root. Denote the χ_{fc} children of the fusion center as $\mathrm{child}_1(\mathrm{fc}), \ldots, \mathrm{child}_{\chi_{\mathrm{fc}}}(\mathrm{fc})$. Denote the χ_i children of sensor i as $\mathrm{child}_1(i), \ldots, \mathrm{child}_{\chi_i}(i)$. Denote the parent of sensor i as $\mathrm{parent}(i)$. Training data vector $\mathbf{x}_{i,j} \in \mathbb{R}^{D_i}$ is measured by sensor i. The sensor receives dimensionality-reduced measurements from its children, combines them with its own measurements, and transmits a dimensionality-reduced version of this combination to its parent. Mathematically, the transmission from sensor i to its parent is:

$$\tilde{\mathbf{x}}_{i\to\mathrm{parent}(i),j} = \mathbf{A}_i^T \begin{bmatrix} \mathbf{x}_{i,j} \\ \tilde{\mathbf{x}}_{\mathrm{child}_1(i)\to i,j} \\ \vdots \\ \tilde{\mathbf{x}}_{\mathrm{child}_{\chi_i}(i)\to i,j} \end{bmatrix}, \tag{4.13}$$

where $\mathbf{A}_i \in \mathcal{V}\left(D_i + \sum_{k=1}^{\chi_i} d_{\mathrm{child}_k(i)}, d_i\right)$.

As an extension to the margin-based classification and linear dimensionality reduction objective (4.2), the following objective is proposed for sensor networks:

$$\mathrm{L}(\varphi, \mathbf{A}_1, \ldots, \mathbf{A}_m) = \sum_{j=1}^{n} \ell\left(y_j\varphi\left(\begin{bmatrix} \tilde{\mathbf{x}}_{\mathrm{child}_1(\mathrm{fc})\to\mathrm{fc},j} \\ \vdots \\ \tilde{\mathbf{x}}_{\mathrm{child}_{\chi_{\mathrm{fc}}}(\mathrm{fc})\to\mathrm{fc},j} \end{bmatrix}\right)\right) + \lambda\,\mathrm{J}(\varphi). \tag{4.14}$$

Just as in the single sensor network in which the fusion center needed to receive the message $\tilde{\mathbf{x}}_{\mathrm{s}\to\mathrm{fc},j}$ from its child in order to optimize φ, in the multisensor network the fusion center needs to receive the messages $\tilde{\mathbf{x}}_{\mathrm{child}_1(\mathrm{fc})\to\mathrm{fc},j}, \ldots, \tilde{\mathbf{x}}_{\mathrm{child}_{\chi_{\mathrm{fc}}}(\mathrm{fc})\to\mathrm{fc},j}$ from all of its children in order to optimize φ. The messages coming from the children of the

fusion center are themselves simple linear functions of the messages coming from their children, as given in (4.13). The same holds down the tree to the leaf sensors. Thus, to gather the information required by the fusion center to optimize φ, a message-passing sweep occurs from the leaf nodes in the tree up to the root.

For fixed φ and optimization of the $\mathbf{A}_i$, there is also message passing, this time sweeping back from the fusion center toward the leaves that generalizes what occurs in the single sensor network. Before finding the partial derivative of $\mathrm{L}(\varphi, \mathbf{A}_1, \ldots, \mathbf{A}_m)$ with respect to $\mathbf{A}_i$, further notation is first introduced. Slice $\mathbf{A}_i$ into blocks as follows:

$$\mathbf{A}_i = \begin{bmatrix} \mathbf{A}_{i,\mathrm{self}} \\ \mathbf{A}_{i,\mathrm{child}_1} \\ \vdots \\ \mathbf{A}_{i,\mathrm{child}_{\chi_i}} \end{bmatrix},$$

where $\mathbf{A}_{i,\mathrm{self}} \in \mathbb{R}^{D_i \times d_i}$ and $\mathbf{A}_{i,\mathrm{child}_k} \in \mathbb{R}^{d_{\mathrm{child}_k(i)} \times d_i}$. Also,

$$\tilde{\boldsymbol{\varphi}}'_{\mathrm{fc}\to\mathrm{child}_k(\mathrm{fc}),j} = \begin{bmatrix} \varphi_{\tilde{x}_{\sum_{\kappa=1}^{k-1} d_{\mathrm{child}_\kappa(\mathrm{fc})}+1}}\left(\begin{bmatrix} \tilde{\mathbf{x}}_{\mathrm{child}_1(\mathrm{fc})\to\mathrm{fc},j} \\ \vdots \\ \tilde{\mathbf{x}}_{\mathrm{child}_{\chi_{\mathrm{fc}}}(\mathrm{fc})\to\mathrm{fc},j} \end{bmatrix}\right) \\ \vdots \\ \varphi_{\tilde{x}_{\sum_{\kappa=1}^{k} d_{\mathrm{child}_\kappa(\mathrm{fc})}}}\left(\begin{bmatrix} \tilde{\mathbf{x}}_{\mathrm{child}_1(\mathrm{fc})\to\mathrm{fc},j} \\ \vdots \\ \tilde{\mathbf{x}}_{\mathrm{child}_{\chi_{\mathrm{fc}}}(\mathrm{fc})\to\mathrm{fc},j} \end{bmatrix}\right) \end{bmatrix}$$

is the slice of the decision function gradient corresponding to the dimensions transmitted by $\mathrm{child}_k(\mathrm{fc})$ to the fusion center. Additionally, let:

$$\tilde{\boldsymbol{\varphi}}'_{i\to\mathrm{child}_k(i),j} = \mathbf{A}_{i,\mathrm{child}_k} \tilde{\boldsymbol{\varphi}}'_{\mathrm{parent}(i)\to i,j}. \tag{4.15}$$

Then, the matrix partial derivative of the objective function (4.14) with respect to $\mathbf{A}_i$ is:

$$\mathbf{L}_{\mathbf{A}_i} = \sum_{j=1}^{n} y_j \, \ell'\left(y_j \varphi\left(\begin{bmatrix} \tilde{\mathbf{x}}_{\mathrm{child}_1(\mathrm{fc})\to\mathrm{fc},j} \\ \vdots \\ \tilde{\mathbf{x}}_{\mathrm{child}_{\chi_{\mathrm{fc}}}(\mathrm{fc})\to\mathrm{fc},j} \end{bmatrix}\right)\right) \begin{bmatrix} \mathbf{x}_{i,j} \\ \tilde{\mathbf{x}}_{\mathrm{child}_1(i)\to i,j} \\ \vdots \\ \tilde{\mathbf{x}}_{\mathrm{child}_{\chi_i}(i)\to i,j} \end{bmatrix} \tilde{\boldsymbol{\varphi}}'^{T}_{\mathrm{parent}(i)\to i,j}. \tag{4.16}$$

Like in the single sensor network, the information required at sensor i to optimize $\mathbf{A}_i$ that it does not already have consists of a scalar and a vector. The scalar value $y_j \, \ell'(y_j \varphi)$ is common throughout the network. The vector message $\tilde{\boldsymbol{\varphi}}'_{\mathrm{parent}(i)\to i,j}$ has length d_i and is received from $\mathrm{parent}(i)$. As seen in (4.15), the message a sensor passes onto its child is a simple linear function of the message received from its parent. To optimize all of

the $\mathbf{A}_i$, a message-passing sweep starting from the fusion center and going down to the leaves is required. Simple gradient descent along Stiefel manifold geodesics is then performed locally at each sensor. Overall, the coordinate descent training proceeds along with the passing of messages $\tilde{\mathbf{x}}_{i\to\text{parent}(i),j}$ and $\tilde{\boldsymbol{\varphi}}'_{i\to\text{child}_k(i),j}$, which are functions of incoming messages as seen in (4.13) and (4.15).

The data vector that is received by the fusion center is reduced from $\sum_{i=1}^{m} D_i$ dimensions to $\sum_{k=1}^{\chi_{\text{fc}}} d_{\text{child}_k(\text{fc})}$ dimensions. The fact that the composition of linear dimensionality reduction by two matrices on the Stiefel manifold can be represented by a single matrix on the Stiefel manifold leads to the observation that the dimensionality reduction performed by the sensor network has an equivalent matrix $\mathbf{A} \in \mathcal{V}\left(\sum_{i=1}^{m} D_i, \sum_{k=1}^{\chi_{\text{fc}}} d_{\text{child}_k(\text{fc})}\right)$. However, $\mathbf{A}$ has further constraints than just the Stiefel manifold constraint due to the topology of the network. For example, the equivalent $\mathbf{A}$ of the network in which the fusion center has two child sensors must be block-diagonal with two blocks.

Thus in the tree-structured sensor network, there is an equivalent matrix $\mathbf{A} \in \mathcal{T}\left(\sum_{i=1}^{m} D_i, \sum_{k=1}^{\chi_{\text{fc}}} d_{\text{child}_k(\text{fc})}\right) \subset \mathcal{V}\left(\sum_{i=1}^{m} D_i, \sum_{k=1}^{\chi_{\text{fc}}} d_{\text{child}_k(\text{fc})}\right)$, where $\mathcal{T}$ is a subset determined by the tree topology. The consistency analysis of Section 4.3.3 holds under the assumption that there exists an $\mathbf{A}^* \in \mathcal{T}\left(\sum_{i=1}^{m} D_i, \sum_{k=1}^{\chi_{\text{fc}}} d_{\text{child}_k(\text{fc})}\right)$ such that $\text{R}(\hat{y}^*(\mathbf{A}^{*T}\mathbf{x})) = \text{R}(\hat{y}^*(\mathbf{x}))$.

The constrained set of dimensionality reduction matrices $\mathcal{T}$ may have a smaller maximum zonotope content $V(Z)$ than the full Stiefel manifold, which would in turn mean a smaller Rademacher complexity. The fusion center receives the χ_{fc}-ary Cartesian product of dimensionality-reduced data from its children. The content of the Cartesian product is the product of the individual contents, and thus:

$$V(Z) \leq \prod_{k=1}^{\chi_{\text{fc}}} \omega_{d_{\text{child}_k(\text{fc})}} \left(\frac{\omega_{d_{\text{child}_k(\text{fc})}-1}}{\omega_{d_{\text{child}_k(\text{fc})}}} \sqrt{\frac{D_k}{d_{\text{child}_k(\text{fc})}}} \right)^{d_{\text{child}_k(\text{fc})}},$$

which is less than or equal to the bound (2.82) for $\mathcal{Z}\left(\sum_{i=1}^{m} D_i, \sum_{k=1}^{\chi_{\text{fc}}} d_{\text{child}_k(\text{fc})}\right)$. A more refined upper bound may be developed based on the specifics of the tree topology.

The tree-structured network has smaller Rademacher complexity due to further constraints to the classifier function space resulting from the network structure. However, similar to D having a minor effect on complexity seen in Figure 4.15, this smaller complexity for $\mathcal{T}\left(\sum_{i=1}^{m} D_i, \sum_{k=1}^{\chi_{\text{fc}}} d_{\text{child}_k(\text{fc})}\right)$ is not much less than the complexity for the system without network constraints $\mathcal{V}\left(\sum_{i=1}^{m} D_i, \sum_{k=1}^{\chi_{\text{fc}}} d_{\text{child}_k(\text{fc})}\right)$. The network constraints, however, may increase the training error. The generalization error, being composed of both the training error and the complexity, increases with network constraints due to increases in training error that are not offset by decreases in complexity, resulting in worse classification performance. However, for sensor networks, the performance criterion of interest is generally a combination of generalization error *and* power expenditure in communication. This idea is illustrated in the remainder of this section.

■ 4.4.3 Physical Network Model

In Section 4.4.2, dimensionality reduction for margin-based classification in sensor networks is described abstractly, without considering the physical implementation or specific tree topologies. A model of a wireless sensor network is presented here, which is then used in Section 4.4.4 to report classification error as a function of transmission power expended by the network.

Consider m sensors and a fusion center in the plane that communicate wirelessly. The distance between sensor i and its parent is $r_{i\leftrightarrow \text{parent}(i)}$, and the power required for communication from i to its parent is $d_i r^2_{i\leftrightarrow \text{parent}(i)}$, where as before, d_i is the reduced dimension output by the sensor. The model arises by the common assumption of signal attenuation according to the square of the distance.[4] The total transmission power used by the network is then:

$$\text{transmission power} = \sum_{i=1}^{m} d_i r^2_{i\leftrightarrow \text{parent}(i)}. \tag{4.17}$$

Consider three network structures: parallel architecture, serial or tandem architecture, and binary tree architecture. In the parallel architecture, all m sensors are direct children of the fusion center. In the serial architecture, the fusion center has a single child, which in turn has a single child, and so on. In the binary tree architecture, the fusion center has two children, each of whom have two children on down the tree. When the number of sensors is such that a perfect binary tree is not produced, i.e., $m + 2$ is not a power of two, the bottom level of the tree remains partially filled.

The sensor and fusion center locations are modeled as follows. The fusion center is fixed at the center of a circle with unit area and the m sensor locations are uniformly distributed over that circle. Given the sensor node locations and desired network topology, it is assumed that parent-child links and corresponding $r_{i\leftrightarrow \text{parent}(i)}$ are chosen to minimize (4.17). There is only one parallel network, so optimization is not required. Exact minimization of (4.17) for the other architectures may not be tractable in deployed ad hoc wireless sensor networks because it involves solving a version of the traveling salesman problem for the serial architecture and a version of the minimum spanning tree problem for the binary tree architecture. Nevertheless, it is assumed that the minimization has been performed in the following results; this assumption is commented upon later. For the parallel architecture, the distances are [22]:

$$r^{(\text{parallel})}_{i\leftrightarrow \text{fc}} = \frac{\Gamma\left(i+\frac{1}{2}\right)\Gamma\left(m+1\right)}{\sqrt{\pi}\Gamma(i)\Gamma\left(m+\frac{3}{2}\right)}, \tag{4.18}$$

where sensor i is the ith closest sensor to the fusion center. There is no closed form expression for the $r_{i\leftrightarrow \text{parent}(i)}$ in the serial or binary tree architectures, but this quantity can be estimated through Monte Carlo simulation.

[4] The model $r^{\alpha}_{i\leftrightarrow \text{parent}(i)}$ for values of α other than two could also be considered.

To fully specify the network, the reduced dimensions of the sensors d_i must also be set. The choice made is to set d_i proportional to the number of descendants of sensor i plus one for itself. This choice implies that all d_i are equal in the parallel network, and that d_i is proportional to $m - i + 1$ in the serial network so that the number of dimensions passed up the chain to the fusion center increases the closer one gets to the fusion center. As seen in Section 4.4.4, with this choice of d_i, all three topologies have essentially the same classification performance. This is not, however, generally true for different d_i assignments; for example, if all d_i are taken to be equal in the serial network, the classification performance is quite poor.

■ 4.4.4 Classification Error for Different Networks

Given the sensor network model of Section 4.4.3, classification performance is investigated for the three different network architectures with different amounts of transmission power. The phenomenon of overfitting seen in the centralized case has an important counterpart and implication for wireless sensor networks: increasing the total allowed transmission power—manifested either by increases in the number of sensors or increases in the number of transmitted dimensions per sensor—does not necessarily result in improved classification performance. The examples in this section illustrate several tradeoffs and suggest further lines of research.

Different transmission powers are obtained by varying the number of sensors and scaling the d_i values. Data coming from a sensor network is emulated by slicing the dimensions of the ionosphere and sonar datasets and assigning the different dimensions to different sensors. With $D_i = 5$ for all sensors in the network, the dimensions are assigned in the order given in the UCI Repository, so the first sensor 'measures' the first five dimensions listed, the second sensor 'measures' dimensions six through ten, and so on. The dimensions are not ordered according to relevance for classification in any way.

Results for the ionosphere dataset are plotted in Figure 4.16. Figure 4.16(a) shows tenfold cross-validation training and test error obtained from the algorithm described in Section 4.4.2 with the parallel network as a function of transmission power. Each training and test error pair corresponds to a different value of $m = 1, 2, \ldots, 6$ and $d_i = 1, 2, \ldots, 5$. In Section 4.1.4, classification performance is plotted as a function of the reduced dimension, but here the horizontal axis is transmission power, taking the distance between sensor nodes into account. As in Section 4.1.4, the phenomenon of overfitting is quite apparent.

In Figure 4.16(b), classification error is plotted as a function of transmission power for the serial architecture. The points in the plot are for different numbers of sensors $m = 1, 2, \ldots, 6$ and different scalings of the reduced dimension $d_i = (m - i + 1), 2(m - i + 1), \ldots, 5(m - i + 1)$. The classification error values in Figure 4.16(b) are quite similar to the ones for the parallel case.[5] The plot for the parallel architecture appearing to be a horizontally compressed version of the serial architecture plot indicates that to

[5]In fact, they are the same for the five pairs of points when $m = 1$ because the parallel and serial networks are the same when there is a single sensor.

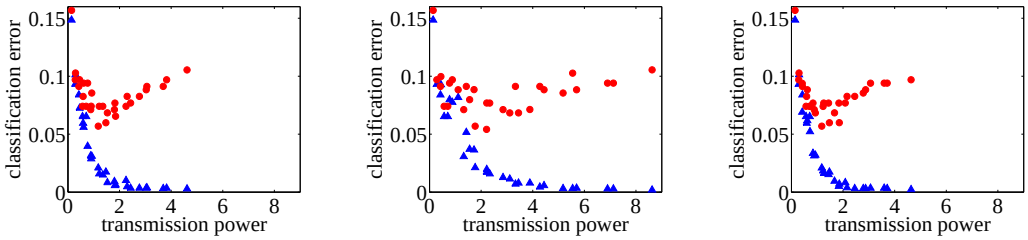


Figure 4.16. Tenfold cross-validation training error (blue triangle markers) and test error (red circle markers) on ionosphere dataset for (a) parallel, (b) serial, and (c) binary tree network architectures.

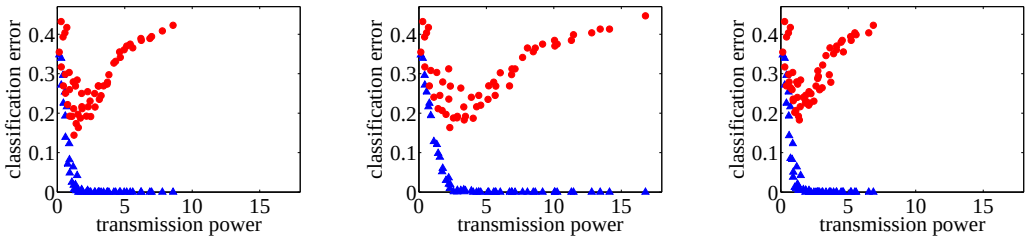


Figure 4.17. Tenfold cross-validation training error (blue triangle markers) and test error (red circle markers) on sonar dataset for (a) parallel, (b) serial, and (c) binary tree network architectures.

achieve those similar classification performances, more transmission power is required by the serial architecture. Although the distances between parents and children tends to be smaller in the serial architecture, the chosen d_i are larger closer to the fusion center leading to higher transmission power.

The binary tree architecture's classification error plot is given in Figure 4.16(c). The training and test error values are similar to the other two architectures.[6] The transmission power needed to achieve the given classification errors is similar to that of the parallel architecture and less than the serial architecture. Among the three architectures with the d_i assigned as described in Section 4.4.3, all have approximately the same classification performance, but the serial network uses more power.

The same experiments are repeated for the sonar dataset with plots given in Figure 4.17. For this dataset, the number of sensors m varies from one to eleven. The same trends can be observed as in the ionosphere dataset. All three network topologies produce similar classification errors, but the serial network uses more power.

Some overall observations for wireless sensor networks are the following. There exist some optimal parameters of the network with a finite number of sensors and some

[6]The binary tree is the same as the parallel network for $m = 1, 2$ and the serial network for $m = 1$.

dimensionality reduction. For a fixed number of samples n, one may be tempted to think that deploying more sensors always helps classification performance since the total number of measured dimensions increases, but this is not always true due to overfitting. That a small number of sensors, which perform dimensionality reduction, yield optimal classification performance is good from the perspective of resource usage. Among different possible choices of network architectures, three particular choices have been compared. Others are certainly possible, including the investigated topologies but with different d_i proportions. For the chosen d_i proportions, all three network topologies have essentially the same classification performance, but this is not true for other choices.

In this empirical investigation of classification performance versus resource usage, the main observation is that the two are not at odds. The decrease of resource usage is coincident with the prevention of overfitting, which leads to improved classification performance. Oftentimes there is a tradeoff between resource usage and performance, but that is not the case in the overfitting regime. Additionally, among the network architectures compared, the parallel and binary tree architectures use less power in communication than the serial architecture for equivalent classification performance. The plotted transmission power values, however, are based on choosing the parent-child links to exactly minimize (4.17); in practice, this minimization will only be approximate for the binary tree architecture and will require a certain amount of communication overhead. Therefore, the parallel architecture, which requires no optimization, is recommended for this application. This new distributed dimensionality reduction formulation and empirical study suggests a direction for future research, namely the problem of finding the number of sensors, the network structure, and the set of d_i that optimize generalization error in classification for a given transmission power budget and given number of training samples n.

■ 4.5 Chapter Summary

In this chapter, a formulation for linear and nonlinear dimensionality reduction driven by the objective of margin-based classification has been presented. An optimization approach has been developed that involves alternation between two minimizations: one to update a classifier decision function and the other to update a matrix on the Stiefel manifold. The phenomenon of overfitting has been examined both analytically and empirically: analytically through the Rademacher complexity, and empirically through experiments on several real-world datasets, illustrating that dimensionality reduction is an important component in improving classification accuracy. The consistency of the dimensionality-reduced classifier has also been analytically characterized. It has been described how the proposed optimization scheme can be distributed in a network containing a single sensor through a message-passing approach, with the classifier decision function updated at the fusion center and the dimensionality reduction matrix updated at the sensor. Additionally, the formulation has been extended to tree-structured fusion

networks.

Papers such as [138, 158] have advocated nonparametric learning, of which margin-based classification is a subset, for inference in distributed settings such as wireless sensor networks. Reducing the amount of communication is an important consideration is these settings, which has been addressed in this chapter through a joint linear dimensionality reduction and margin-based classification method applicable to networks in which sensors measure more than one variable. Reducing communication is often associated with a degradation in performance, but in this application it is not the case in the regime when dimensionality reduction prevents overfitting. Thus, dimensionality reduction is important for two distinct reasons: reducing the amount of resources consumed, and obtaining good generalization.

Chapter 5

Precision of the Prior Probabilities

THE optimal decision rule in Bayesian hypothesis testing is the likelihood ratio test with threshold set according to the prior probabilities of the hypotheses along with the costs of false alarm and missed detection. Many times, hypothesis testing arises when identifiable objects with precisely specified prior probabilities are measured to determine their state. For each object in the population, the Bayes optimal procedure sets the threshold based precisely on the prior probabilities of the object. The frugality pursued in this chapter is to limit the precision of the prior probabilities and therefore of the threshold in the likelihood ratio test. The precision is limited through quantization or clustering optimized for a novel distortion criterion, the mean Bayes risk error [207, 208].

Frugality in prior probability precision is motivated by scenarios in which the decision maker has finite memory or limited information processing resources. The decision maker maps the true prior probability vector of each object in the population to one of a few representation points, which requires less memory than storing the true prior probabilities. Then, when performing the likelihood ratio test, the representation point corresponding to the true prior of the measured object is used in the threshold.

Although not the only such scenario, one example is human decision making. Specifically, consider a referee deciding whether a player has committed a foul using his or her noisy observation as well as prior experience. Players commit fouls at different rates. It is this rate which is the prior probability for the 'foul committed' hypothesis. If the referee tunes the prior probability to the particular player on whose action the decision is to be made, decision-making performance is improved. Human decision makers, however, are limited in their information processing capability and tend to categorize objects [129]. Consequently, the referee is limited and categorizes players into a small number of levels, with associated representative prior probabilities, exactly the scenario described above.

This chapter, unlike Chapter 3 and Chapter 4 which focus on supervised classification, deals with decision making with a known likelihood ratio function. The population of objects is represented by a probability density function of prior probabilities. The design of the mapping from prior probability vectors in the population to representative probability vectors is approached through quantization when this probability density function is given and through k-means clustering when only samples from it are avail-

able. Mean Bayes risk error (MBRE) is defined as a fidelity criterion and conditions are derived for a minimum MBRE quantizer. In the quantization problem, with known population probability distribution, increasing the number of quantization levels always results in improved detection performance. The best detection performance is achieved with an infinite number of quantization levels, which is equivalent to not quantizing. Clustering, however, is not always suboptimal; with access only to samples, there is an intermediate number of clusters that optimizes detection performance. This behavior is akin to overfitting and the structural risk minimization principle seen in Chapter 3 and Chapter 4.

Previous work that combines detection and quantization looks at the quantization of the measurements, not the prior probabilities, and also only approximates the Bayes risk function instead of working with it directly, e.g. [90, 102, 155] and references cited in [90]. In such work, there is a communication constraint between the sensor and the decision maker, but the decision maker has unconstrained processing capability. The work here deals with the opposite case, where there is no communication constraint between the sensor and the decision maker, but the decision maker is constrained. A brief look at imperfect priors appears in [93], but optimal quantization is not considered. It is shown in [83, 104] that small deviations from the true prior yield small deviations in the Bayes risk. There does not seem to be any previous work that has looked at the quantization, clustering, or categorization of prior probabilities or the threshold in likelihood ratio tests.

The chapter is organized in the following manner. Section 5.1 defines the problem of quantizing prior probabilities along with the Bayes risk error distortion. Optimality conditions are derived and some examples of MBRE-optimal quantizers are given along with their performance in the low-rate quantization regime. Section 5.2 discusses the high-rate quantization regime and gives distortion–rate functions. Section 5.3 discusses the clustering problem and presents examples that show that frugality in the precision of prior probabilities may be advantageous in terms of detection performance. Certain human decision-making tasks, as mentioned previously, may be modeled by quantized prior hypothesis testing due to certain features of human information processing. In Section 5.4, human decision making is analyzed in detail for segregated populations, revealing a mathematical model of social discrimination. A brief summary of the chapter appears in Section 5.5.

■ 5.1 Quantization of Prior Probabilities for Hypothesis Testing

Many different distortion criteria can be used in quantization, including absolute error and squared error. When it is prior probabilities in a detection problem that are being quantized, the distortion function should take the Bayes risk into account. In this section, such a distortion function is proposed, some of its properties are derived, and quantizer optimality conditions with this distortion function are also derived. Examples with comparison to absolute error quantization are presented as well.

■ 5.1.1 Bayes Risk Error Distortion

Recall from Section 2.1 the Bayes optimal decision rule, the likelihood ratio test

$$\frac{f_{\mathbf{x}|y}(\mathbf{x}|y=+1)}{f_{\mathbf{x}|y}(\mathbf{x}|y=-1)} \underset{\hat{y}(\mathbf{x})=-1}{\overset{\hat{y}(\mathbf{x})=+1}{\lessgtr}} \frac{p_- c_{+-}}{(1-p_-)c_{-+}}. \tag{5.1}$$

with Bayes risk performance

$$\mathrm{R}(p_-) = c_{+-} p_- p_F(p_-) + c_{-+}[1-p_-]p_M(p_-), \tag{5.2}$$

where

$$p_F = \Pr[\hat{y}(\mathbf{x}) = +1 | y = -1],$$
$$p_M = \Pr[\hat{y}(\mathbf{x}) = -1 | y = +1].$$

In (5.2), the error probabilities p_F and p_M depend on p_- through $\hat{y}(\cdot)$, given in (5.1). $\mathrm{R}(p_-)$ is zero at the points $p_- = 0$ and $p_- = 1$ and is positive-valued, strictly concave, and continuous in the interval $(0,1)$ [49, 218, 221].

Recall from Section 2.5 that a quantizer for a scalar random variable $\mathsf{p}_- \in [0,1]$ with probability density function $f_{\mathsf{p}_-}(p_-)$ is a function $q_k(p_-)$ chosen to minimize the expected distortion

$$\varrho = \mathrm{E}[\rho(\mathsf{p}_-, q_k(\mathsf{p}_-))] = \int_0^1 \rho(p_-, q_k(p_-)) f_{\mathsf{p}_-}(p_-) dp_-, \tag{5.3}$$

where ρ is a distortion function. In the decision-making setup considered, a quantized version of the prior probability is used in the likelihood ratio test threshold rather than the true prior probability. The decision rule considered is

$$\frac{f_{\mathbf{x}|y}(\mathbf{x}|y=+1)}{f_{\mathbf{x}|y}(\mathbf{x}|y=-1)} \underset{\hat{y}(\mathbf{x})=-1}{\overset{\hat{y}(\mathbf{x})=+1}{\lessgtr}} \frac{q_k(p_-)c_{+-}}{(1-q_k(p_-))c_{-+}}. \tag{5.4}$$

There is mismatch when the true prior probability is p_-, but some other value a is substituted for p_- in the threshold in (5.1), i.e.,

$$\frac{f_{\mathbf{x}|y}(\mathbf{x}|y=+1)}{f_{\mathbf{x}|y}(\mathbf{x}|y=-1)} \underset{\hat{y}(\mathbf{x})=-1}{\overset{\hat{y}(\mathbf{x})=+1}{\lessgtr}} \frac{a c_{+-}}{(1-a)c_{-+}}. \tag{5.5}$$

This is the case in (5.4), where $a = q_k(p_-)$. The Bayes risk when there is mismatch is

$$\tilde{\mathrm{R}}(p_-, a) = c_{+-} p_- p_F(a) + c_{-+}[1-p_-]p_M(a). \tag{5.6}$$

$\tilde{\mathrm{R}}(p_-, a)$ is a linear function of p_- with slope $(c_{+-}p_F(a) - c_{-+}p_M(a))$ and intercept $c_{-+}p_M(a)$. Note that $\tilde{\mathrm{R}}(p_-, a)$ is tangent to $\mathrm{R}(p_-)$ at a and that $\tilde{\mathrm{R}}(p_-, p_-) = \mathrm{R}(p_-)$.

Based on the mismatched Bayes risk, a distortion function can be defined that is appropriate for the quantization of prior probabilities in hypothesis testing.

Definition 5.1.1. *Let Bayes risk error $\rho_B(p_-, a)$ be the difference between the mismatched Bayes risk function $\tilde{\mathrm{R}}(p_-, a)$ and the Bayes risk function $\mathrm{R}(p_-)$:*

$$\begin{aligned}\rho_B(p_-, a) &= \tilde{\mathrm{R}}(p_-, a) - \mathrm{R}(p_-) \\ &= c_{+-}p_-p_F(a) + c_{-+}[1-p_-]p_M(a) - c_{+-}p_-p_F(p_-) - c_{-+}[1-p_-]p_M(p_-).\end{aligned} \tag{5.7}$$

A few properties of $\rho_B(p_-, a)$ are derived that are used in stating quantizer optimality conditions in Section 5.1.2.

Theorem 5.1.1. *The Bayes risk error $\rho_B(p_-, a)$ is nonnegative and only equal to zero when $p_- = a$.*

Proof. $\mathrm{R}(p_-)$ is a continuous and strictly concave function, and lines $\tilde{\mathrm{R}}(p_-, a)$ are tangent to $\mathrm{R}(p_-)$. Therefore, $\tilde{\mathrm{R}}(p_-, a) \geq \mathrm{R}(p_-)$ for all p_- and a, with equality only when $p_- = a$. Consequently, $\rho_B(p_-, a)$ is nonnegative and only equal to zero when $p_- = a$. ■

Theorem 5.1.2. *The Bayes risk error $\rho_B(p_-, a)$, as a function of $p_- \in (0, 1)$, is continuous and strictly convex for all a.*

Proof. The Bayes risk error $\rho_B(p_-, a)$ is the difference of a continuous linear function and a continuous strictly concave function. Therefore, it is continuous and strictly convex. ■

Properties of $\rho_B(p_-, a)$ as a function of a involve the complementary receiver operating characteristic (CROC) described in Section 2.1.4. The CROC is traced out as the threshold in the likelihood ratio test is varied. Equivalently, the CROC is traced out as a is varied in the likelihood ratio test (5.5), which is the parameterization considered in the following.

Lemma 5.1.1. *There exists a unique point $a^\dagger$ on the CROC of a deterministic likelihood ratio test at which $\frac{dp_M}{dp_F} = -1$. Also,*

$$\begin{cases} -\infty < \frac{dp_M}{dp_F} < -1, & a > a^\dagger, \\ -1 < \frac{dp_M}{dp_F} < 0, & a < a^\dagger. \end{cases}$$

Proof. As discussed in Section 2.1.4, the CROC at its endpoints takes values $(p_F = 0, p_M = 1)$ when $a = 1$ and $(p_F = 1, p_M = 0)$ when $a = 0$. Therefore, the CROC has average slope -1. The CROC is a strictly convex function for deterministic likelihood ratio tests, as also discussed in Section 2.1.4. The result follows from the mean value theorem and strict convexity. ■

Lemma 5.1.2. *For the CROC of a deterministic likelihood ratio test and positive constants β and γ,*

$$\begin{cases} \beta\frac{dp_F}{da} + \gamma\frac{dp_M}{da} > 0, & \frac{\gamma}{\beta}\frac{dp_M}{dp_F} < -1, \\ \beta\frac{dp_F}{da} + \gamma\frac{dp_M}{da} < 0, & \frac{\gamma}{\beta}\frac{dp_M}{dp_F} > -1. \end{cases}$$

Proof. If $\frac{\gamma}{\beta}\frac{dp_M}{dp_F} < -1$, then $\frac{\gamma dp_M}{da}\frac{da}{\beta dp_F} < -1$ by the chain rule of differentiation. A property shown in Section 2.1.4 is that $\frac{dp_F}{da} < 0$ and $\frac{dp_M}{da} > 0$ for all $a \in (0,1)$. Thus, by rearranging terms in the inequality, it is found that $\gamma\frac{dp_M}{da} > -\beta\frac{dp_F}{da}$. Consequently $\beta\frac{dp_F}{da} + \gamma\frac{dp_M}{da} > 0$. The opposite case is shown in the same manner. ■

Theorem 5.1.3. *The Bayes risk error $\rho_B(p_-, a)$, as a function of $a \in (0,1)$ for all p_-, has exactly one stationary point that is a minimum.*

Proof. Combining the first cases of Lemma 5.1.1 and Lemma 5.1.2, there exists an $a^\star$ such that for all $a > a^\star$, the slope of $\beta p_F(a) + \gamma p_M(a)$ is positive. Combining the second cases of Lemma 5.1.1 and Lemma 5.1.2, for all $a < a^\star$, the slope of $\beta p_F(a) + \gamma p_M(a)$ is negative. Therefore, $\beta p_F(a) + \gamma p_M(a)$ has exactly one stationary point $a^\star$ in $(0,1)$, which is a minimum.

As a function of a, the Bayes risk error (5.7) is of the form $\beta p_F(a) + \gamma p_M(a) + C$, where C is a constant that does not depend on a. Hence, it also has exactly one stationary point in $(0,1)$, which is a minimum. ■

Corollary 5.1.1. *The Bayes risk error $\rho_B(p_-, a)$, as a function of $a \in (0,1)$ for all p_-, is quasiconvex.*

■ 5.1.2 Minimum Mean Bayes Risk Error Quantization

The conditions necessary for the optimality of a quantizer, discussed in Section 2.5.2, are derived when the distortion function is Bayes risk error. Recall that a k-point quantizer for $f_{P_-}(p_-)$ partitions the interval $[0,1]$ into k cells $\mathcal{Q}_1$, $\mathcal{Q}_2$, $\mathcal{Q}_3$, ..., $\mathcal{Q}_k$. For each of these quantization cells $\mathcal{Q}_i$, there is a representation point a_i to which elements are mapped. For regular quantizers, the regions are subintervals $\mathcal{Q}_1 = [0, b_1]$, $\mathcal{Q}_2 = (b_1, b_2]$, $\mathcal{Q}_3 = (b_2, b_3]$, ..., $\mathcal{Q}_k = (b_{k-1}, 1]$ and the representation points a_i are elements of $\mathcal{Q}_i$.[1] The nearest neighbor and centroid conditions are developed for MBRE in the following.

Nearest Neighbor Condition

With the representation points a_i fixed, an expression for the interval boundaries b_i is derived. Given any $p_- \in [a_i, a_{i+1}]$, if $\tilde{\mathrm{R}}(p_-, a_i) < \tilde{\mathrm{R}}(p_-, a_{i+1})$ then Bayes risk error is minimized if p_- is represented by a_i, and if $\tilde{\mathrm{R}}(p_-, a_i) > \tilde{\mathrm{R}}(p_-, a_{i+1})$ then Bayes risk error is minimized if p_- is represented by a_{i+1}. The boundary point $b_i \in [a_i, a_{i+1}]$ is the abscissa of the point at which the lines $\tilde{\mathrm{R}}(p_-, a_i)$ and $\tilde{\mathrm{R}}(p_-, a_{i+1})$ intersect. The idea is illustrated graphically in Figure 5.1.

[1] Due to the strict convexity of $\rho_B(p_-, a)$ in p_- for all a, as shown in Theorem 5.1.2, quantizers that satisfy the necessary conditions for MBRE optimality are regular, see [79, Lemma 6.2.1]. Therefore, only regular quantizers are considered.

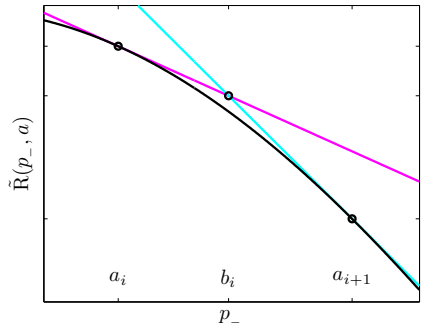


Figure 5.1. The intersection of the lines $\tilde{\mathrm{R}}(p_-, a_i)$ (magenta line) and $\tilde{\mathrm{R}}(p_-, a_{i+1})$ (cyan line), both tangent to $\mathrm{R}(p_-)$ (black line), is the optimal interval boundary b_i.

By manipulating the slopes and intercepts of $\tilde{\mathrm{R}}(p_-, a_i)$ and $\tilde{\mathrm{R}}(p_-, a_{i+1})$, the point of intersection is found to be:

$$b_i = \frac{c_{-+}\left(p_M(a_{i+1}) - p_M(a_i)\right)}{c_{-+}\left(p_M(a_{i+1}) - p_M(a_i)\right) - c_{+-}\left(p_F(a_{i+1}) - p_F(a_i)\right)}. \tag{5.8}$$

Centroid Condition

With the quantization cells fixed, the MBRE is to be minimized over the a_i. Here, the MBRE is expressed as the sum of integrals over quantization cells:

$$\varrho_B = \sum_{i=1}^{k} \int_{b_{i-1}}^{b_i} \left[\tilde{\mathrm{R}}(p_-, a_i) - \mathrm{R}(p_-)\right] f_{\mathsf{p}_-}(p_-) dp_-. \tag{5.9}$$

Because the cells are fixed, the minimization may be performed for each interval separately.

Define conditional means $I_i^F = \int_{b_{i-1}}^{b_i} p_- f_{\mathsf{p}_-}(p_-) dp_-$ and $I_i^M = \int_{b_{i-1}}^{b_i} [1-p_-] f_{\mathsf{p}_-}(p_-) dp_-$. Then:

$$a_i = \arg\min_a \left\{ c_{+-} I_i^F p_F(a) + c_{-+} I_i^M p_M(a) \right\}. \tag{5.10}$$

Since $\beta p_F(a) + \gamma p_M(a)$ has exactly one stationary point, which is a minimum (cf. Theorem 5.1.3), equation (5.10) is uniquely minimized by setting its derivative equal to zero. Thus, a_i is the solution to:

$$c_{+-} I_i^F \left.\frac{dp_F(a)}{da}\right|_{a_i} + c_{-+} I_i^M \left.\frac{dp_M(a)}{da}\right|_{a_i} = 0. \tag{5.11}$$

Commonly, differentiation of the two error probabilities is tractable; they are themselves integrals of the likelihood functions and the differentiation is with respect to some function of the limits of integration.

Lloyd–Max Algorithm

Algorithm 2.5.1, discussed in Section 2.5.3, which alternates between the nearest neighbor and centroid conditions, may be used to find minimum MBRE quantizers. Trushkin [200] shows that the conditions necessary for optimality of the quantizer are also sufficient conditions for local optimality if the following hold. The first condition is that $f_{\mathsf{p}_-}(p_-)$ must be positive and continuous in $(0,1)$. The second condition is that

$$\int_0^1 \rho(p_-, a) f_{\mathsf{p}_-}(p_-) dp_-$$

must be finite for all a. The first and second conditions are met by common distributions such as the beta distribution [66].

The third condition is that the distortion function $\rho(p_-, a)$ satisfy three properties. First, it must be zero only for $p_- = a$, which is shown for $\rho_B(p_-, a)$ in Theorem 5.1.1. Second, it must be continuous in p_- for all a, which is shown for $\rho_B(p_-, a)$ in Theorem 5.1.2. The third of the properties, as listed by Trushkin [200], is that $\rho(p_-, a)$ must be convex in a. Convexity in a, does not hold for Bayes risk error in general, but the convexity is only used by Trushkin [200] to show that a unique minimum exists; in fact, quasiconvexity is enough. As shown in Theorem 5.1.3 and mentioned in Corollary 5.1.1, $\rho_B(p_-, a)$ has a unique stationary point that is a minimum and is quasiconvex in a. Therefore, the analysis of [200] applies to Bayes risk error distortion. Thus, if $f_{\mathsf{p}_-}(p_-)$ satisfies the first and second conditions, then the algorithm is guaranteed to converge to a local optimum. The algorithm may be run many times with different initializations to find the global optimum.

Further conditions on $\rho(p_-, a)$ and $f_{\mathsf{p}_-}(p_-)$ are given in [200] for there to be a unique locally optimal quantizer, i.e. the global optimum. If these further conditions for unique local optimality hold, then Algorithm 2.5.1 is guaranteed to find the globally minimum MBRE quantizer.

■ 5.1.3 Examples

A few examples are presented that show minimum MBRE quantizers of prior probabilities. The examples have scalar measurements x with Gaussian likelihood functions under the two hypotheses with the same variance σ^2 and different means. The likelihood functions are:

$$\begin{aligned} f_{\mathsf{x}|\mathsf{y}}(x|\mathsf{y}=-1) &= \mathcal{N}(x; 0, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}, \\ f_{\mathsf{x}|\mathsf{y}}(x|\mathsf{y}=+1) &= \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}. \end{aligned} \tag{5.12}$$

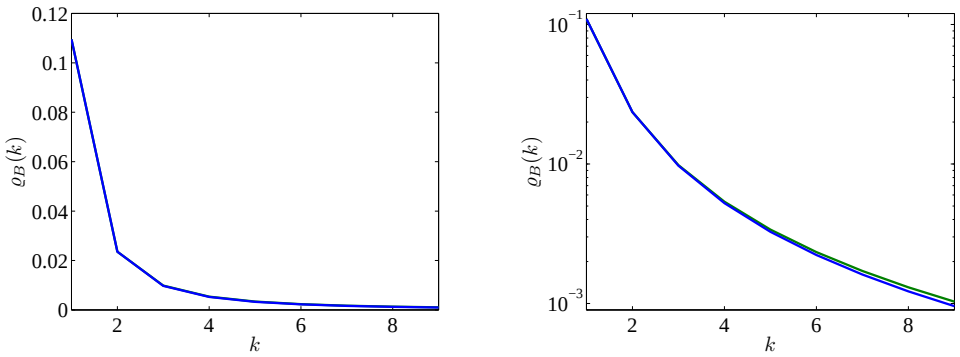


Figure 5.2. Mean Bayes risk error for MBRE-optimal quantizer (blue line) and MAE-optimal quantizer (green line) on (a) linear scale, and (b) logarithmic scale with uniform p_- and $c_{+-} = 1$, $c_{-+} = 1$.

The mean under hypothesis $y = -1$ is zero and the mean under hypothesis $y = +1$ is μ. The two error probabilities are:

$$p_F(a) = Q\left(\frac{\mu}{2\sigma} + \frac{\sigma}{\mu}\ln\left(\frac{c_{+-}a}{c_{-+}(1-a)}\right)\right),$$
$$p_M(a) = Q\left(\frac{\mu}{2\sigma} - \frac{\sigma}{\mu}\ln\left(\frac{c_{+-}a}{c_{-+}(1-a)}\right)\right), \tag{5.13}$$

where

$$Q(z) = \frac{1}{\sqrt{2\pi}}\int_z^\infty e^{-x^2/2}dx.$$

Finding the centroid condition, the derivatives of the error probabilities are:

$$\left.\frac{dp_F}{da}\right|_{a_i} = -\frac{1}{\sqrt{2\pi}}\frac{\sigma}{\mu}\frac{1}{a_i(1-a_i)}e^{-\frac{1}{2}\left(\frac{\mu}{2\sigma}+\frac{\sigma}{\mu}\ln\left(\frac{c_{+-}a_i}{c_{-+}(1-a_i)}\right)\right)^2}, \tag{5.14}$$

$$\left.\frac{dp_M}{da}\right|_{a_i} = +\frac{1}{\sqrt{2\pi}}\frac{\sigma}{\mu}\frac{1}{a_i(1-a_i)}e^{-\frac{1}{2}\left(\frac{\mu}{2\sigma}-\frac{\sigma}{\mu}\ln\left(\frac{c_{+-}a_i}{c_{-+}(1-a_i)}\right)\right)^2}. \tag{5.15}$$

By substituting these derivatives into (5.11) and simplifying, the following expression is obtained for the representation points:

$$a_i = \frac{I_i^F}{I_i^F + I_i^M}. \tag{5.16}$$

Examples with different distributions $f_{p_-}(p_-)$, and different costs c_{+-} and c_{-+} are now presented. All of the examples have $\mu = 1$ and $\sigma^2 = 1$. As a point of reference, a

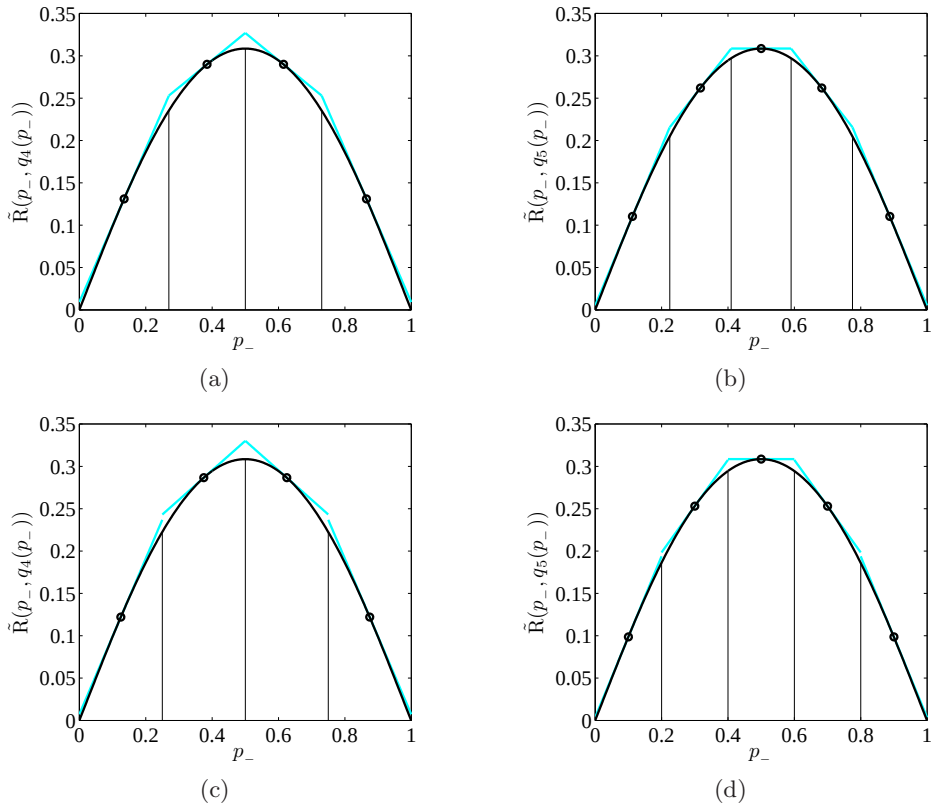


Figure 5.3. (a)–(b) MBRE-optimal, and (c)–(d) MAE-optimal quantizers for $k = 4, 5$ with uniform p_- and $c_{+-} = 1$, $c_{-+} = 1$.

comparison is made to quantizers designed under absolute error [103], i.e. $\rho_1(p_-, a) = |p_- - a|$, an objective that does not account for Bayesian hypothesis testing.

The first example has p_- uniformly distributed over $[0, 1]$ so that all prior probabilities are equally likely in the population, and the Bayes costs c_{+-} and c_{-+} both equal to one. The MBRE of the MBRE-optimal quantizer and of a quantizer designed to minimize mean absolute error (MAE) with respect to $f_{\mathsf{p}_-}(p_-)$ is plotted in Figure 5.2. The plots show MBRE on both linear and logarithmic scales as a function of the number of quantization levels k. The blue line is the MBRE-optimal quantizer and the green line is the MAE-optimal quantizer. The performance of both quantizers is similar, but the MBRE-optimal quantizer always performs better or equally. Each increment of k is associated with a large reduction in Bayes risk. There is a very large performance improvement from $k = 1$ to $k = 2$.

The plots in Figure 5.3 show $\tilde{\mathrm{R}}(p_-, q_k(p_-))$ with cyan lines for the MBRE- and MAE-optimal quantizers. The markers are the representation points and the vertical lines indicate the quantization cell boundaries. The black line is $\mathrm{R}(p_-)$, the Bayes

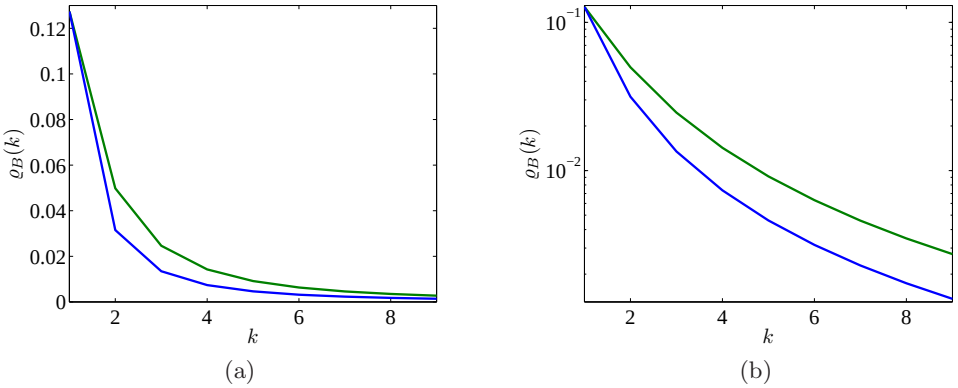


Figure 5.4. Mean Bayes risk error for MBRE-optimal quantizer (blue line) and MAE-optimal quantizer (green line) on (a) linear scale, and (b) logarithmic scale with uniform p_- and $c_{+-} = 1$, $c_{-+} = 4$.

risk with unquantized prior probabilities. The MAE-optimal quantizer for the uniform distribution has quantization cells of equal width and representation points in their centers. The representation points for the MBRE-optimal quantizer are closer to $p_- = \frac{1}{2}$ than the uniform quantizer. This is because the area under the Bayes risk function is concentrated in the center.

Similar plots to those in the first example are given for a second example with unequal Bayes costs in Figure 5.4 and Figure 5.5. This example is also with a uniform population distribution, and has costs $c_{+-} = 1$ and $c_{-+} = 4$. The unequal costs skew the Bayes risk function and consequently the representation point locations. The difference in performance between the MBRE-optimal and MAE-optimal quantizers is greater in this example because the MAE criterion does not incorporate the Bayes costs, which factor into MBRE calculation. It can be clearly seen in Figure 5.5 that the area between the cyan lines and black lines is greater for the MAE-optimal quantizers than for the MBRE-optimal quantizers.

The third example examines a nonuniform distribution for p_-, in particular the beta(5,2) distribution. The probability density function is shown in Figure 5.6. The MBRE of the MBRE-optimal and MAE-optimal quantizers is in Figure 5.7. There are also large improvements in performance with an increase in k here. The representation points a_i are most densely distributed where $f_{p_-}(p_-)$ has mass. In particular, more representation points are in the right half of the domain than in the left, as seen in Figure 5.8.

■ 5.2 High-Rate Quantization Analysis

In this section, minimum MBRE quantization is studied asymptotically from the perspective of high-rate quantization theory [89]. Li et al. [113] propose a high-rate quan-

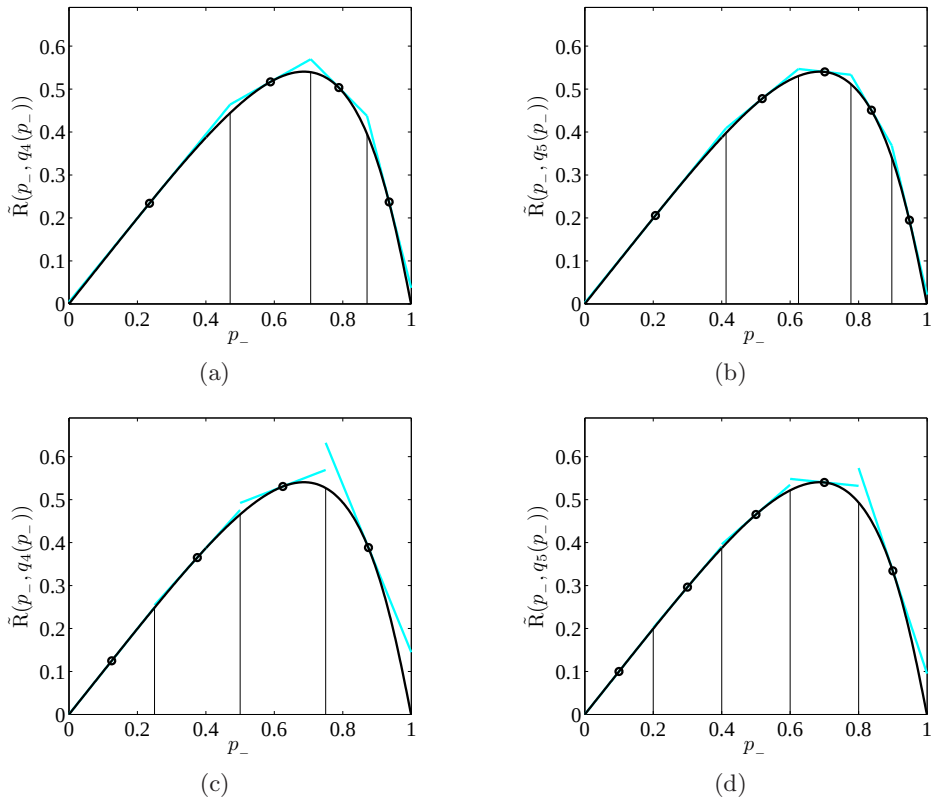


Figure 5.5. (a)–(b) MBRE-optimal, and (c)–(d) MAE-optimal quantizers for $k = 4, 5$ with uniform $p_{\_}$ and $c_{+-} = 1$, $c_{-+} = 4$.

tization analysis based on a locally quadratic approximation. That locally quadratic analysis is applied to Bayes risk error distortion in what follows.

■ 5.2.1 Locally Quadratic Approximation

In high-rate quantization, the objective is to characterize the expected error ϱ as a function of the number of quantization levels k for large k. The expected error is the MBRE for the purposes here. The quantizer point density $\lambda(p_{\_})$ is obtained as part of the analysis. Recall from Section 2.5 that integrating a quantizer point density over an interval yields the fraction of the representation points a_i that are in that interval as k goes to infinity.

The Bayes risk error distortion function has a positive second derivative in $p_{\_}$ due to strict convexity shown in Theorem 5.1.2, and has a continuous third derivative for many families of likelihood functions. The third derivative of $\rho_B(p_{\_}, a)$ with respect to

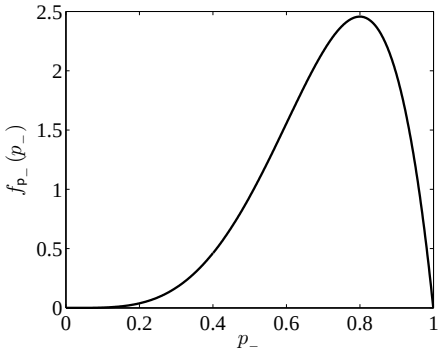


Figure 5.6. The beta(5,2) probability density function.

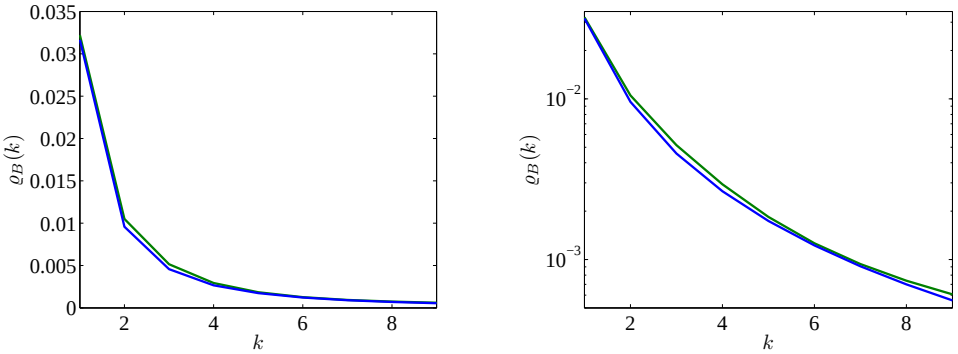


Figure 5.7. Mean Bayes risk error for MBRE-optimal quantizer (blue line) and MAE-optimal quantizer (green line) on (a) linear scale, and (b) logarithmic scale with beta(5,2) p_- and $c_{+-} = 1$, $c_{-+} = 1$.

p_- is

$$\frac{\partial^3 \rho_B}{\partial p_-^3} = -c_{+-} p_- \frac{d^3 p_F}{dp_-^3} - 3c_{+-} \frac{d^2 p_F}{dp_-^2} - c_{-+}(1 - p_-) \frac{d^3 p_M}{dp_-^3} + 3c_{-+} \frac{d^2 p_M}{dp_-^2}, \tag{5.17}$$

when the constituent derivatives exist. When this third derivative (5.17) exists and is continuous, $\rho_B(p_-, a)$ is locally quadratic in the sense of [113], and the high-rate quantization analysis of Li et al. [113] may be applied.

For large k, let

$$B(p_-) = -\frac{c_{+-} p_-}{2} \frac{d^2 p_F}{dp_-^2} - c_{+-} \frac{dp_F}{dp_-} - \frac{c_{-+}(1 - p_-)}{2} \frac{d^2 p_M}{dp_-^2} + c_{-+} \frac{dp_M}{dp_-}. \tag{5.18}$$

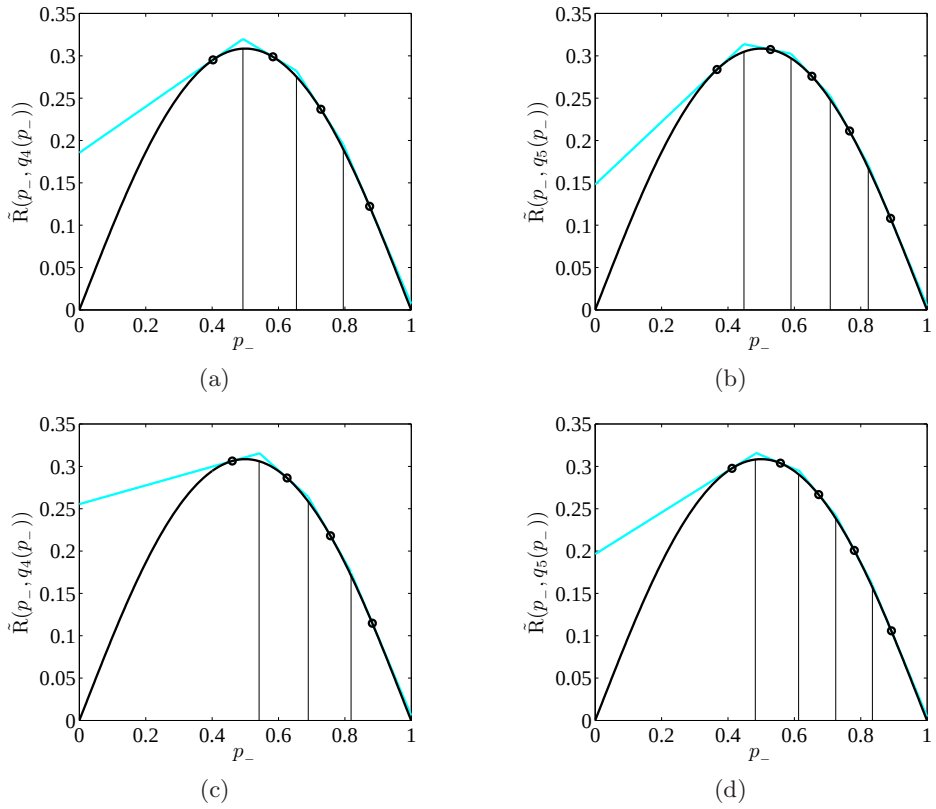


Figure 5.8. (a)–(b) MBRE-optimal, and (c)–(d) MAE-optimal quantizers for $k = 4, 5$ with beta(5,2) p_- and $c_{+-} = 1$, $c_{-+} = 1$.

Then $\rho_B(p_-, a_i)$ is approximated by the following second order Taylor expansion:

$$\rho_B(p_-, a_i) \approx B(a_i) \left(p_- - a_i\right)^2, \quad p_- \in \mathcal{Q}_i. \tag{5.19}$$

Under the assumption made in all studies of asymptotic quantization that $f_{\mathsf{p}_-}(\cdot)$ is sufficiently smooth to be effectively constant over small bounded sets, substituting (5.19) into the quantization objective (5.9) yields the following approximation to the MBRE:

$$\varrho_B \approx \sum_{i=1}^{k} f_{\mathsf{p}_-}(a_i) B(a_i) \int_{b_{i-1}}^{b_i} \left(p_- - a_i\right)^2 dp_-. \tag{5.20}$$

Additionally, the MBRE is asymptotically greater than and approximately equal to the following lower bound, derived in [113] by relationships involving normalized moments of inertia of intervals $\mathcal{Q}_i$ and by Hölder's inequality:

$$\varrho_B \leq \frac{1}{12k^2} \int_0^1 B(p_-) f_{\mathsf{p}_-}(p_-) \lambda(p_-)^{-2} dp_-, \tag{5.21}$$

where the optimal quantizer point density for the Bayes risk error is:

$$\lambda_B(p_-) = \frac{\Big(B(p_-)f_{\mathsf{p}_-}(p_-)\Big)^{1/3}}{\int_0^1 \Big(B(p_-)f_{\mathsf{p}_-}(p_-)\Big)^{1/3} dp_-}. \tag{5.22}$$

Substituting (5.22) into (5.21) yields

$$\varrho_B \leq \frac{1}{12k^2}\|B(p_-)f_{\mathsf{p}_-}(p_-)\|_{1/3}. \tag{5.23}$$

■ 5.2.2 Examples

The examples of Section 5.1.3 are continued here. The likelihood functions are univariate Gaussians with the same variance and the difference of the means being one. The function $B(p_-)$ requires the first and second derivatives of the false alarm and missed detection probabilities. The first derivatives are given in (5.14) and (5.15). The second derivatives are as follows.

$$\frac{d^2p_F}{dp_-^2} =$$

$$-\tfrac{1}{\sqrt{8\pi}}\tfrac{\sigma}{\mu}\tfrac{1}{p_-^2(1-p_-)^2}e^{-\frac{1}{8\mu^2\sigma^2}\left(\mu^2+2\sigma^2\ln\left(\frac{c_{+-}p_-}{c_{-+}(1-p_-)}\right)\right)^2}\left[-3+4p_- - \tfrac{2\sigma^2}{\mu^2}\ln\left(\tfrac{c_{+-}p_-}{c_{-+}(1-p_-)}\right)\right], \tag{5.24}$$

and

$$\frac{d^2p_M}{dp_-^2} =$$

$$+\tfrac{1}{\sqrt{8\pi}}\tfrac{\sigma}{\mu}\tfrac{1}{p_-^2(1-p_-)^2}e^{-\frac{1}{8\mu^2\sigma^2}\left(\mu^2-2\sigma^2\ln\left(\frac{c_{+-}p_-}{c_{-+}(1-p_-)}\right)\right)^2}\left[-1+4p_- - \tfrac{2\sigma^2}{\mu^2}\ln\left(\tfrac{c_{+-}p_-}{c_{-+}(1-p_-)}\right)\right]. \tag{5.25}$$

By inspection, the third derivatives are continuous and thus for the Gaussian likelihood examples, $\rho_B(p_-, a)$ is locally quadratic. An expression for $B(p_0)$, and consequently for $\lambda_B(p_-)$ and the right side of (5.23), is obtained by substituting the first derivatives and second derivatives into (5.18).

A comparison to absolute error quantization is given here as well. The optimal quantizer point density for MAE is [87]:

$$\lambda_1(p_-) = \frac{f_{\mathsf{p}_-}(p_-)^{1/2}}{\int_0^1 f_{\mathsf{p}_-}(p_-)^{1/2}dp_-}. \tag{5.26}$$

Using this point density in (5.21) instead of $\lambda_B(p_-)$ gives the asymptotic MBRE of the MAE-optimal quantizer.

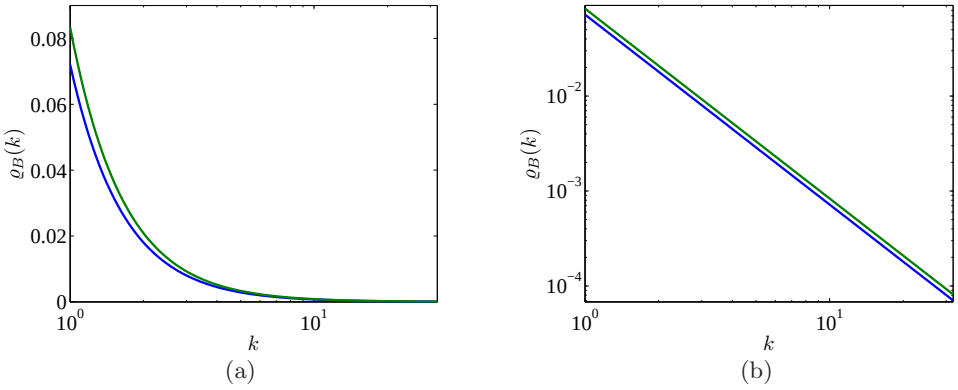


Figure 5.9. High-rate approximation to mean Bayes risk error for MBRE-optimal quantizer (blue line) and MAE-optimal quantizer (green line) on (a) linear scale, and (b) logarithmic scale with uniform p_- and $c_{+-} = 1$, $c_{-+} = 1$.

As in Section 5.1.3, the first example has a uniform population distribution and equal Bayes costs. In Figure 5.9, the right side of (5.21) is plotted as a function of k for both the MBRE-optimal $\lambda_B(p_-)$ and the MAE-optimal $\lambda_1(p_-)$. The horizontal axis is on a logarithmic scale. The quantity $\log_2(k)$ is generally known as the *rate* of the quantizer. With this terminology, the plots in Figure 5.9 are known as distortion–rate functions. The vertical axis is given on both linear and logarithmic scales. On the logarithmic scale, there is a constant gap between the distortion–rate functions of the MBRE-optimal and MAE-optimal quantizer. This difference is:

$$\mathrm{rate}_{\mathrm{MBRE}}(\varrho_B) - \mathrm{rate}_{\mathrm{MAE}}(\varrho_B) = \frac{1}{2}\log_2\left(\frac{\|f_{\mathsf{p}_-}(p_-)B(p_-)\|_{1/3}}{\|f_{\mathsf{p}_-}(p_-)\|_{1/2}\int_0^1 B(p_-)dp_-}\right).$$

The closer the ratio inside the logarithm is to one, the closer the MBRE- and MAE-optimal quantizers. The quantizer point densities are shown in Figure 5.10. As also seen in Figure 5.3, the representation points of the MBRE-optimal quantizer are more concentrated around $p_- = \frac{1}{2}$ than of the MAE-optimal quantizer because of the shape of the Bayes risk function.

The second example has $c_{+-} = 1$ and $c_{-+} = 4$, also with uniformly distributed p_-. Figure 5.11 shows the distortion–rate functions for this case. The gap is much larger in this example because as mentioned previously, the MAE criterion does not take the Bayes costs into account. The difference between the optimal quantizer point densities is quite pronounced in this example, as seen in Figure 5.12. Whereas the MAE-optimal quantizer has representation points uniformly distributed, the MBRE-optimal quantizer has much greater representation point density for $p_- > \frac{1}{2}$ due to the cost of a false alarm being greater than the cost of a missed detection.

With a beta(5,2) distribution for $f_{\mathsf{p}_-}(p_-)$ and equal Bayes costs, the gap is more

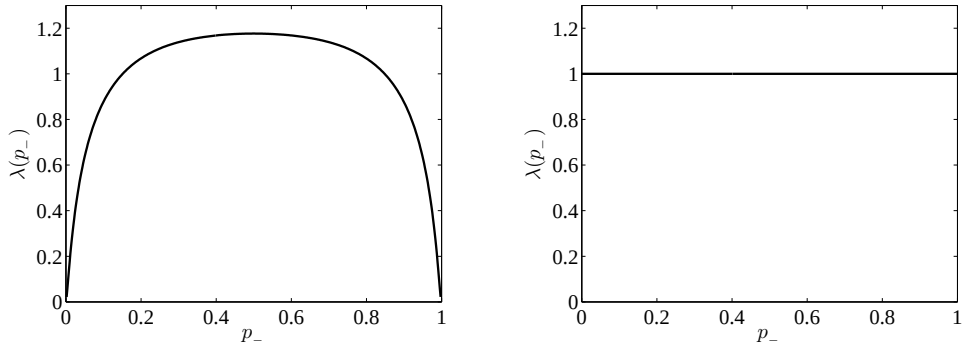


Figure 5.10. (a) MBRE, and (b) MAE quantizer point density with uniform $\boldsymbol{p}_-$ and $c_{+-} = 1$, $c_{-+} = 1$.

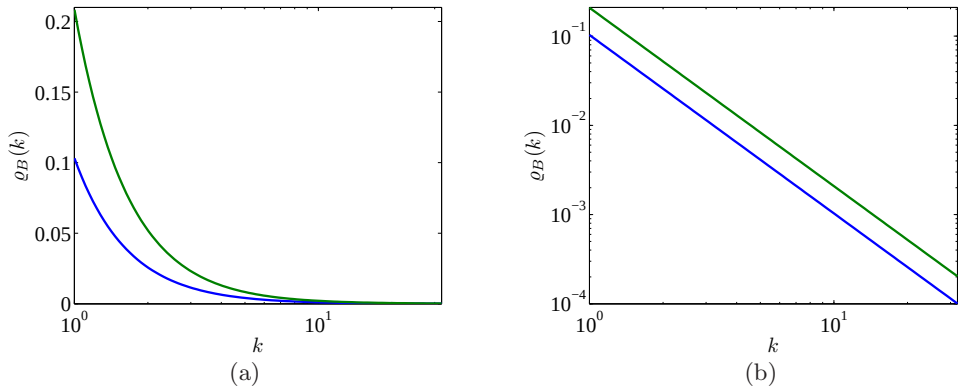


Figure 5.11. High-rate approximation to mean Bayes risk error for MBRE-optimal quantizer (blue line) and MAE-optimal quantizer (green line) on (a) linear scale, and (b) logarithmic scale with uniform $\boldsymbol{p}_-$ and $c_{+-} = 1$, $c_{-+} = 4$.

similar to the first example which also has equal Bayes costs, as seen in Figure 5.13. As expected, the quantizer point densities, shown in Figure 5.14, tend to match the beta(5,2) distribution. When using quantized priors in setting the threshold in a likelihood ratio test, the asymptotic analysis reveals that the detection performance reduction due to quantization approaches zero exponentially in the rate. The MAE-optimal quantizer is not bad for the MBRE distortion criterion when the Bayes costs are equal, but suffers when the Bayes costs are unequal.

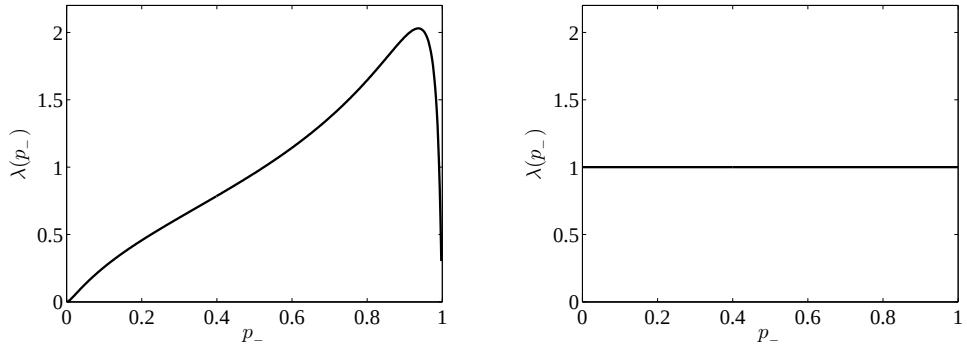


Figure 5.12. (a) MBRE, and (b) MAE quantizer point density with uniform p_- and $c_{+-} = 1$, $c_{-+} = 4$.

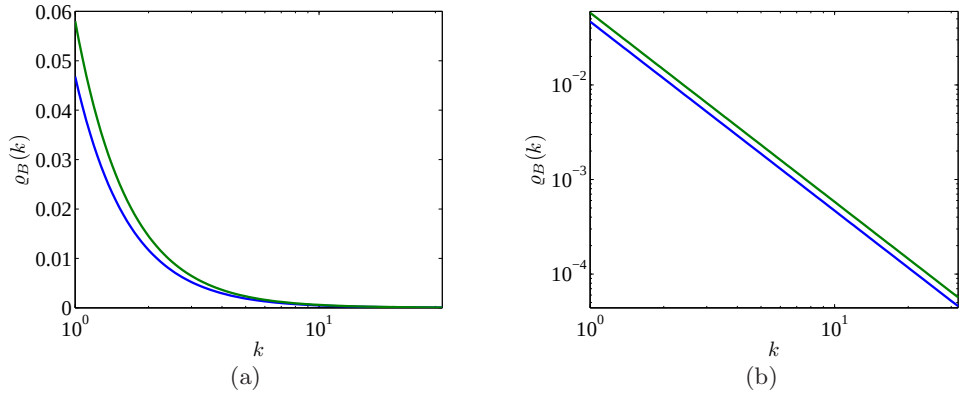


Figure 5.13. High-rate approximation to mean Bayes risk error for MBRE-optimal quantizer (blue line) and MAE-optimal quantizer (green line) on (a) linear scale, and (b) logarithmic scale with beta(5,2) p_- and $c_{+-} = 1$, $c_{-+} = 1$.

■ 5.3 Detection Performance with Empirical Priors

The analysis and conclusions of Section 5.1 and Section 5.2 are in the case that $f_{p_-}(p_-)$ is known. This section deals with the case when this population distribution is not known. The k-means algorithm, Algorithm 2.5.2 presented in Section 2.5.3, can be applied to data to find representation points and limit the precision of prior probabilities for hypothesis testing. In contrast to the previous two sections, limiting the precision may prove beneficial in the same manner that frugality improves performance in situations of finite data encountered in Chapter 3 and Chapter 4.

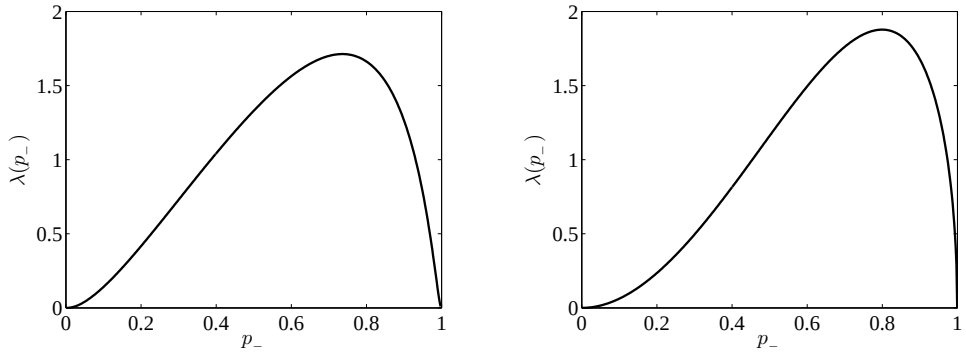


Figure 5.14. (a) MBRE, and (b) MAE quantizer point density with beta(5,2) p_- and $c_{+-} = 1$, $c_{-+} = 1$.

■ 5.3.1 k-Means Clustering of Prior Probabilities

Recall k-means clustering from Section 2.5.3. Given $n > k$ samples $p_{-1}, \ldots, p_{-n}$, k-means clustering partitions the samples into k clusters each with a representation point or cluster center a_i, $i = 1, \ldots, k$. The sequence of k-means clusterings from samples of $f_{\mathsf{p}_-}(p_-)$ converges as n grows to the quantizer designed using $f_{\mathsf{p}_-}(p_-)$ as long as the distortion function meets certain conditions [88, 169]. The Bayes risk error $\rho_B(p_-, a)$ meets the conditions on the distortion function given in [169] except for convexity in a. However, as in the sufficiency of the Lloyd–Max conditions, the quasiconvexity of the Bayes risk error is enough. This sort of consistency is important to note, but the interest in this section is to focus on the finite data regime.

It is easy to imagine systems in which constrained decision makers do not have access to the full population distribution $f_{\mathsf{p}_-}(p_-)$, but only to training data related to that distribution. In the particular setup considered in this section, the decision maker does not have direct access to samples $p_{-1}, \ldots, p_{-n}$, but to estimates of those samples $\hat{p}_{-1}, \ldots, \hat{p}_{-n}$ obtained in the following manner. For each sample p_{-j}, which corresponds to one object in the population, the decision maker observes m_j samples of the random variable $\boldsymbol{\upsilon}_j$. The random variable $\boldsymbol{\upsilon}_j$ has mean 0 with probability p_{-j} and mean 1 with probability $(1 - p_{-j})$; the distribution besides the mean is the same. A simple unbiased, consistent estimate[2] is

$$\hat{p}_{-j} = 1 - \frac{1}{m_j} \sum_{i=1}^{m_j} \upsilon_{j,i}. \tag{5.27}$$

The n samples $\hat{p}_{-1}, \ldots, \hat{p}_{-n}$ are then clustered to minimize MBRE. Likelihood ratio tests involving object j use a threshold set according to the cluster center to which $\hat{p}_{-j}$ is assigned.

[2]The estimate (5.27) is the maximum likelihood estimate in the case that $\boldsymbol{\upsilon}_j$ is Bernoulli.

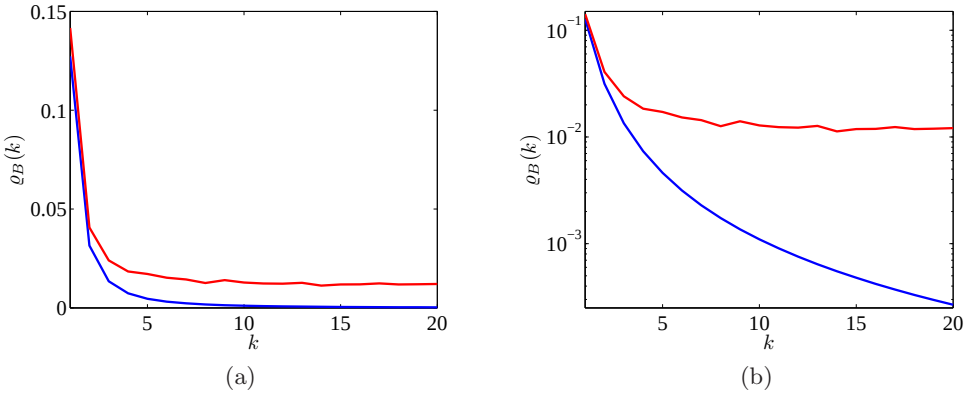


Figure 5.15. Mean Bayes risk error for MBRE-optimal quantizer (blue line) and empirical MBRE-optimal k-means clustering (red line) on (a) linear scale, and (b) logarithmic scale with $m = 100$, uniform $\boldsymbol{p}_-$, and $c_{+-} = 1$, $c_{-+} = 4$.

The proposed model of samples may be understood by considering what data could be available to the decision maker when learning about the population. Populations contain a finite number of objects, which is why the finite number n is included in the model. In training, the decision maker can only observe each object a finite number of times. This number of times could be object-dependent, which motivates the parameter m_j. When the decision maker is learning about the population, perfect measurements of the object state or hypothesis $y_{j,i} \in \{-1, +1\}$ may not be available. These measurements will generally be noisy. The variables $v_{j,i}$ can capture any noise in the measurements. The remainder of the section illustrates the implications of this model through example.

■ 5.3.2 Example

The example presented here is a continuation of the example in Section 5.1.3 and Section 5.2.2 in which the population distribution is uniform over $[0, 1]$ and the Bayes costs are $c_{+-} = 1$ and $c_{-+} = 4$. The likelihood functions $f_{x|y}(x|y = -1)$ and $f_{x|y}(x|y = +1)$ are Gaussian with means zero and one, and the same variance. The noisy measurements in training, the $\boldsymbol{v}_j$, are a mixture of two Gaussians with means zero and one, and the same variance. The mixture weights are p_{-j} and $(1 - p_{-j})$. Thus the noisy measurements used by the decision maker in learning about the population are of the same type as used when doing the hypothesis test. The number of observations per object, m_j, is taken to be the same value m across all n objects in the population.

The MBRE of the empirical formulation is plotted as a function of the number of clusters k in Figure 5.15 for $m = 100$, and in Figure 5.16 for $m = 200$. Zoomed in portions of the plots appear in Figure 5.17. The red lines are the MBRE values of the empirical formulation and the blue lines are the MBRE values of the MBRE-optimal

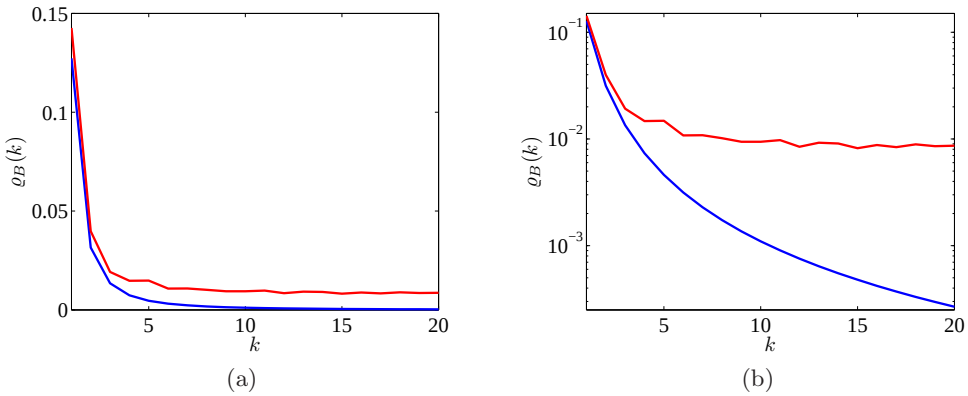


Figure 5.16. Mean Bayes risk error for MBRE-optimal quantizer (blue line) and empirical MBRE-optimal k-means clustering (red line) on (a) linear scale, and (b) logarithmic scale with $m = 200$, uniform p_-, and $c_{+-} = 1$, $c_{-+} = 4$.

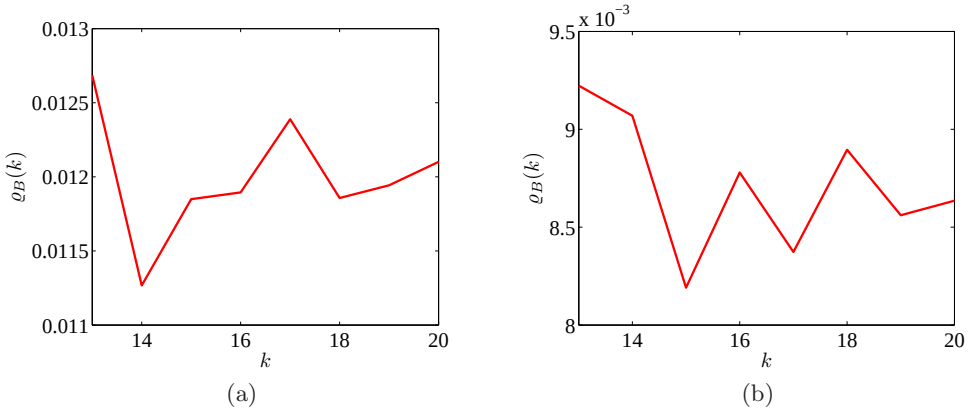


Figure 5.17. Mean Bayes risk error for empirical MBRE-optimal k-means clustering with (a) $m = 100$ and (b) $m = 200$, uniform p_-, and $c_{+-} = 1$, $c_{-+} = 4$.

quantizer designed using known $f_{p_-}(p_-)$. The blue lines here are identical to the blue lines in Figure 5.4.

Several interesting observations may be made about these empirical results. First, the empirical MBRE is greater than the optimal quantizer MBRE. Also, whereas the optimal quantizer MBRE goes to zero as k goes to infinity, the empirical MBRE does not go to zero. In fact, there is an intermediate value of k at which the MBRE is minimized: $k = 14$ for $m = 100$ and $k = 15$ for $m = 200$. This is most obviously seen in the zoomed in plots shown in Figure 5.17. This phenomenon parallels overfitting and the structural risk minimization principle observed in Chapter 3 and Chapter 4. Frugality through limited prior probability precision improves decision-making performance in the setup

of this section.

It is also seen that the MBRE with $m = 200$ is less than the MBRE with $m = 100$: more training helps. More subtly, the optimal k is greater for $m = 200$ than the optimal k for $m = 100$. This observation may be reinforced by examining other values of m as well. For $m = 50$, $k = 9$ is optimal; $k = 14$ for $m = 100$ and $k = 15$ for $m = 200$ are optimal as already noted; $k = 17$ is optimal for $m = 400$. More training also implies that more clusters should be used. Since the estimate $\hat{p}_{\_j}$ gets closer to $p_{\_j}$ as m increases, less regularization in the form of clustering is needed. This behavior also occurs in supervised classification—less regularization is required when more labeled training samples are available, which is noted in the consistency analysis of Section 3.4.3. The empirical formulation of this section exhibits many of the behaviors associated with overfitting in statistical learning.

■ 5.4 Application to Human Decision Making

This section considers the situation in which the decision maker is a human and the objects in the population whose state is to be determined are also humans. For example, as mentioned at the beginning of the chapter, the decision maker and population could be a sports referee and players. The fraction of plays in which player j does not commit a foul is $p_{\_j}$. The population of players is characterized by the distribution $f_{\mathbf{p}\_}(p_{\_})$. Human decision makers categorize into a small number of categories; it has been observed that people can only carry around seven, plus or minus two, categories without getting confused [129]. Thus decisions by humans on large populations of humans that they know may be modeled using the formulation developed in this chapter.

Human decision making is often studied in microeconomics. The decision maker is assumed to be rational so that he or she optimizes decision-making performance. Models of decision making in economics have recently started considering *bounded rationality* [46, 161, 168, 215], including the bounded rationality of memory constraints or information processing constraints [54, 132]. Frugality in prior probability precision described in this chapter can be viewed as a form of bounded rationality. There are two models of bounded rationality, termed *truly bounded rationality* and *costly rationality* [161]. In the truly bounded rationality model, the decision maker is not aware of his or her limitation, whereas in the costly rationality model, the decision maker is aware of the limitation and finds an optimal strategy under the limitation. The quantized prior hypothesis testing model developed in this chapter falls under costly rationality because the quantization or clustering is optimized for decision-making performance through the minimization of MBRE. It may be viewed as an example of categorical and coarse thinking [72, 134] and case-based decision theory [82].

Racial bias in decision making has been observed in society, including in foul calls by National Basketball Association (NBA) referees [159], arrests for minor offences by police [53], searches of stopped vehicles by police [8], and hiring by human resources professionals [190]. Specifically, a decision maker of one race has different decision-

making performance on members of the same race and members of another race. Social cognition theory describes a form of categorization different from the categorization due to limited information processing capability: people tend to automatically categorize others according to race and may use different decision rules for different racial groups [122]. Segregation in social life leads to more interactions within a race than interactions between members of different races, cf. [59] and references therein, which may have an effect on decision making, as human decision makers are not able to completely discount their social interactions in determining how to deploy their limited decision-making resources [72].

Becker [16] discusses two types of racial discrimination: 'a taste for discrimination' and statistical discrimination. Taste-based discrimination is what is usually thought of as discrimination—the explicit preference by a decision maker for or against a group. Statistical discrimination, in contrast, is not based on preferences but arises due to bounded rationality. The model of decision making with frugality in prior probability precision proposed in this chapter is extended in this section to include social factors, leading to a model that generates racial bias without appealing to 'a taste for discrimination' or to different $f_{\mathsf{p}_-}(p_-)$ among different racial groups. The modeled racial bias occurs only because of prior probability precision frugality, automatic categorization by race, and segregation.

Fryer and Jackson [72] discuss how human decision makers use categorization for information processing, how categories are trained, how decisions for members of a minority group are less accurate, and how this may lead to discrimination against minority groups even without malevolent intent. The same conclusion is reached in this section by analyzing the proposed minimum MBRE quantization model. The work in this section, like [72], falls under information-based discrimination [9, 21] as quantization reduces information. Unlike other studies of information-based discrimination, it is assumed here that different populations have the same distribution $f_{\mathsf{p}_-}(p_-)$, the same likelihood functions $f_{\mathsf{x}|y}(\mathbf{x}|y)$, and that there are no dynamic effects. Phelps [152] assumes that the two populations are not identical and that there is a different amount of noise in measuring different populations. In [1, 120], the different populations have the same $f_{\mathsf{p}_-}(p_-)$, but different estimator performances. Discrimination is explained in [9, 44, 187] by a dynamic process in which the minority group does not invest in human capital because it is not valued by the decision maker.

Dow [54] looks at a sequential decision-making scenario, in which the human decision maker is deciding whether to purchase an object from one vendor or another. He or she first observes the price of the first vendor, but does not purchase. Then, he or she observes the price of the second vendor, compares the two prices, and purchases from the lower-priced vendor. However, due to bounded rationality, the decision maker only remembers a quantized version of the first price when making the comparison. The problem setup in that work is different than the scenario discussed here, but the analysis is of a similar flavor.

Mullainathan [133] also considers a sequential scenario and is concerned with learn-

ing beliefs from data. (Decision making is not a part of the framework.) In a sequence of observations, the decision maker perfectly knows the state of an object. If a rational decision maker keeps making observations for a long time, the empirical frequencies of the observations converge to the true probabilities. The model in [133], however, inserts bounded rationality into the learning process. Quantization cells partition $p_- \in [0,1]$; sequential updates to the probabilities are based not on likelihoods from data, but on quantized versions of likelihoods from data. The quantization cells and representation points are maximum a posteriori given the data. Because decision making is not a part of the work, the optimization criterion is not correctly matched; general learning and learning for a particular purpose such as decision making are not always equivalent.

Foster and Vohra [71] discuss calibrated forecasting in which the goal is to learn prior probabilities in a game-theoretic setting. Their method requires discretizing $p_- \in [0,1]$ but does not appeal to quantization theory. Also, observations are not corrupted by noise.

■ 5.4.1 Two Population Model

Consider two populations b and w with identical population distributions $f_{\mathsf{p}_-}(p_-)$ and identical measurement models for Bayesian hypothesis testing $f_{\mathbf{x}|y}(\mathbf{x}|y=-1)$ and $f_{\mathbf{x}|y}(\mathbf{x}|y=+1)$. A rational decision maker who is frugal with prior probability precision ought to ignore the population labels b and w because they are irrelevant for the hypothesis testing task. However, due to automatic categorization resulting from social cognition, the decision maker quantizes the two populations separately. The total quota on representation points, k_t, is split into some number of points for population b and some number for population w, denoted k_b and k_w respectively. The separate quantizers may then be denoted $q_{k_b}(\cdot)$ and $q_{k_w}(\cdot)$.

The definition of MBRE is extended to two populations as follows.

$$\varrho_B^{(2)} = \frac{m_b}{m_b+m_w}\,\mathrm{E}[\tilde{\mathrm{R}}(\mathsf{p}_-, q_{k_b}(\mathsf{p}_-))] + \frac{m_w}{m_b+m_w}\,\mathrm{E}[\tilde{\mathrm{R}}(\mathsf{p}_-, q_{k_w}(\mathsf{p}_-))] - \mathrm{E}[\mathrm{R}(\mathsf{p}_-)], \quad (5.28)$$

where m_b is the number of population b members with whom the decision maker socially interacts, and m_w is the number of population w members with whom the decision maker socially interacts. In order to find the optimal allocation of the total quota of representation points $k_t = k_b + k_w$, $\varrho_B^{(2)}$ is minimized for all $k_t - 1$ possible allocations and the best one is chosen; more sophisticated algorithms developed for bit allocation to subbands in transform coding may also be used [183].

Fryer and Jackson [72] have previously suggested that it is better to allocate more representation points to the majority population than to the minority population. With two separate quantizers and a single size constraint, optimizing $\varrho_B^{(2)}$ over $q_{k_b}(\cdot)$ and $q_{k_w}(\cdot)$ yields the same result. Due to the monotonicity result in Section 2.5.4 that more quantization levels implies smaller MBRE, the MBRE of population b is smaller than the MBRE of population w if $m_b > m_w$ and the MBRE of population w is smaller than the MBRE of population b if $m_w > m_b$. An example of optimal allocation as a function

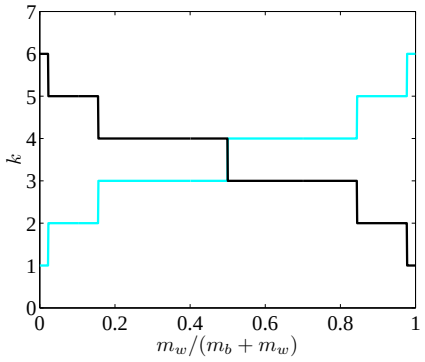


Figure 5.18. Optimal allocation of quantizer sizes to population b (black line) and population w (cyan line) for $k_t = 7$ as a function of $m_w/(m_b + m_w)$ with beta(5,2) p_- and $c_{+-} = 1$, $c_{-+} = 1$.

of $m_w/(m_b + m_w)$ for $k_t = 7$ is shown in Figure 5.18 with the beta(5,2)-distributed p_- example from previous sections. The empirical formulation presented in Section 5.3 also suggests the same thing. It is seen in that section that the greater the number of training samples m per member of the population, the smaller the MBRE and the greater the optimal k.

Due to segregation in social interactions, it is expected that decision makers from population w have a greater $m_w/(m_b+m_w)$ value than decision makers from population b. The model predicts that decision makers from population w have worse expected Bayes risk than decision makers from population b when dealing with population b. Symmetrically, the model predicts that decision makers from population b have worse expected Bayes risk than decision makers from population w when dealing with population w. The model predictions are seen experimentally. A large body of literature in face recognition shows the predicted bias effect; specifically, both p_F and p_M increase when trying to recognize members of the opposite population [126]. Bias in favor of the population of which the decision maker is a member is verified in face recognition by controlled laboratory experimentation. A difficulty in interpreting natural experiments examined through econometric studies, however, is that the ground truth is not known.

■ 5.4.2 Nature of Discrimination Due To Bayes Costs

Since ground truth is not available in econometric studies, it is not clear how to interpret a finding that referees from population w call more fouls on players from population b and that referees from population b call more fouls on players from population w. This phenomenon cannot simply be explained by a larger probability of error. The false alarm probability and the missed detection probability must be disentangled and the Bayes costs must be examined in detail.

In most decision-making scenarios, one of the hypotheses leads to no action and is the default. For example, the hypothesis 'not a crime' or 'not a foul' is the default and

associated with hypothesis $y = -1$. It is shown here that the Bayes costs of the decision maker must be such that the ratio c_{-+}/c_{+-} is large in order to explain the empirical observations in [8, 53, 159, 190].

Using basketball fouls as the running example, the measurable quantity in an econometrics study is the probability that a foul is called. This rate of fouls is:

$$\Pr[\hat{y}(\mathbf{x}) = +1] = p_- p_F + (1 - p_-)(1 - p_M). \tag{5.29}$$

Looking at the average performance of a referee over the b and w populations, the expected foul rates on population b and population w can be compared. Let $\hat{y}_b$ denote the decision rule with priors quantized using k_b quantization levels, and $\hat{y}_w$ denote the decision rule with priors quantized using k_w quantization levels. If

$$\Delta = \mathrm{E}\left[\Pr[\hat{y}_b = +1] - \Pr[\hat{y}_w = +1]\right] \tag{5.30}$$

is greater than zero, then the referee calls more fouls on population b. If Δ is less than zero, then the referee calls more fouls on population w. The Δ expression may be written as:

$$\begin{aligned}\Delta(c_{+-}, c_{-+}) = \mathrm{E}[p_- p_F(q_{k_b}(p_-)) - [1 - p_-]p_M(q_{k_b}(p_-))] \\ - \mathrm{E}[p_- p_F(q_{k_w}(p_-)) - [1 - p_-]p_M(q_{k_w}(p_-))].\end{aligned} \tag{5.31}$$

The dependence of Δ on c_{+-} and c_{-+} is explicit on the left side of (5.31) and is implicit in the false alarm and missed detection probabilities on the right side. The value of Δ also depends on the population distribution $f_{p_-}(p_-)$, the values of k_w and k_b, and the measurement model.

Fixing $f_{p_-}(p_-)$, k_w, k_b, and the measurement model, regions of the plane spanned by c_{+-} and c_{-+} can be determined in which a referee calls more fouls on population b and in which a referee calls more fouls on population w. This is shown in Figure 5.19. For the uniform population distribution, the two regions are divided by the line $c_{-+} = c_{+-}$. For the beta(5,2) distribution, the dividing line is $c_{-+} = \alpha c_{+-}$, where $\alpha > 1$. For any population and measurement model, there is one half-plane in which a referee calls more fouls on population b players. In the other half-plane, the referee calls more fouls on population w players. To reiterate, just because the Bayes risk for foul-calling on population b players is greater than that for population w players, it does not automatically imply that the foul call rate for population b is higher. The high Bayes risk could well be the result of a preponderance of missed foul calls. The choice of Bayes costs with c_{-+} greater than c_{+-} implies that a referee can tolerate more false alarms than missed detections. This assignment of costs has been termed *precautionary* in some contexts. Interpreting Figure 5.19, a referee with $k_w > k_b$ that calls more fouls on population b players is precautionary. A referee with $k_w < k_b$ that calls more fouls on population w players is also precautionary.

Econometric studies often give differences of differences to show racial bias. The first difference is the difference in foul call rate between population b and population w,

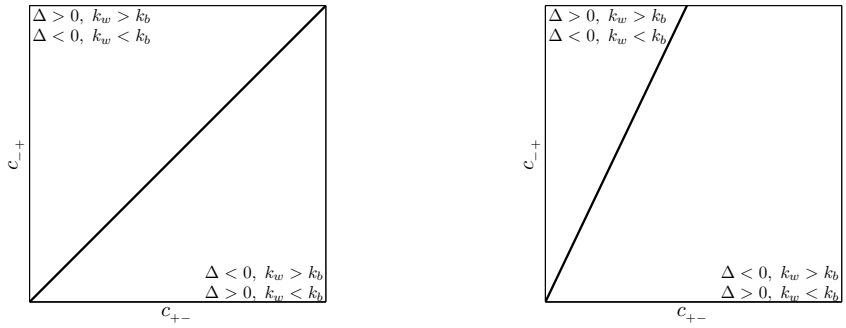


Figure 5.19. Dividing line between Bayes cost region in which referee calls more fouls on population b and region in which referee calls more fouls on population w. A referee with $k_b < k_w$ calls more fouls on population b in the upper left region and more fouls on population w in the lower right region, which correspond to precautionary and anti-precautionary cost settings respectively. The opposite is true for $k_w < k_b$. The population distribution is (a) uniform, and (b) beta(5,2).

which is Δ. The second difference is the difference in Δ between referees belonging to population b and referees belonging to population w. Denote the foul call rate difference of a referee who is a member of population b by Δ_b and the foul call rate difference of a referee who is a member of population w by Δ_w. Then the difference of differences is $\Delta_w - \Delta_b$. Figure 5.20 plots the difference of differences as a function of the ratio c_{-+}/c_{+-} for two different population distributions, the uniform distribution and the beta(5,2) distribution. The right side of each plot is the precautionary regime, in which population w referees call more fouls on population b players than population b referees. For the particular examples, if $c_{-+}/c_{+-} = 10$, then the population w referee has a foul call rate 0.0132 greater than the population b referee on population b players for the beta(5,2) distribution and 0.0142 greater for the uniform distribution. The left side of each plot is the anti-precautionary regime, in which population w referees call fewer fouls on population b players than population b referees. For the particular examples, if $c_{-+}/c_{+-} = 0.1$, then the population w referee has a foul call rate 0.0013 less than the population b referee on population b players for the beta(5,2) distribution and 0.0142 less for the uniform distribution. In these examples, the population w referee has $k_w = 4$, $k_b = 3$, and the population b referee has $k_w = 3$, $k_b = 4$.[3]

Differences of differences calculated in econometric studies are equivalent to the difference between the Δ for population w decision makers and the Δ for population b decision makers. It has been found that the addition of police officers of a given race is

[3]There is no requirement for the population w referee to have $k_w > k_b$ and the population b referee to have $k_w < k_b$. It is only required that the k_w of the population w referee be greater than the k_w of the population b referee (assuming the same k_t). A plot qualitatively similar to Figure 5.20 is obtained if the population w referee has $k_w = 5$, $k_b = 2$, and the population b referee has $k_w = 4$, $k_b = 3$.

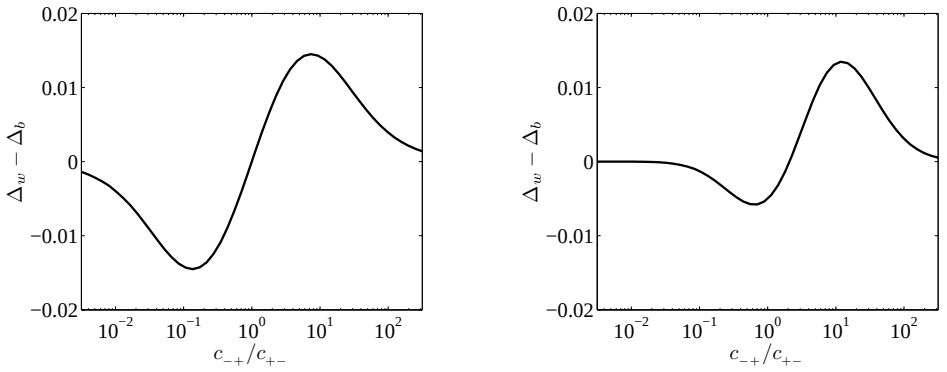


Figure 5.20. Difference of differences in foul calling as a function of the Bayes cost ratio. The population w referee has $k_w = 4$, $k_b = 3$ and the population b referee has $k_w = 3$, $k_b = 4$. The population distribution is (a) uniform, and (b) beta(5,2).

associated with an increase in the number of arrests of suspects of a different race but has little impact on same-race arrests. There are similar own-race bias effects in the decision by police to search a vehicle during a traffic stop [8], in the decision of human resource professionals to not hire [190], and in the decision of NBA referees to call a foul [159]. The rate of searching, the rate of not hiring, and the rate of foul calling are all greater when the decision maker is of a different race than the driver, applicant, and player, respectively. These studies are consistent with model predictions if decision makers are precautionary. Situations in which the model proposed in this section suggests decision makers are anti-precautionary have also been observed occasionally, for example [11] which deals with Canadian juries and is an example in which population members of the same race as the decision maker have less desirable decisions. The proposed model generates the interesting phenomenon that the cost function of the decision maker has a fundamental effect on the nature of racial discrimination. The Bayes costs of human decision makers are revealed in their bias.

■ 5.5 Chapter Summary

Bayesian hypothesis testing is examined when there is a distribution of prior probabilities, but the decision maker may only use a quantized version of the true prior probability in designing a decision rule. Considering the problem of finding the optimal quantizer for this purpose, a new distortion criterion is defined based on the Bayes risk function. For this criterion, the mean Bayes risk error, conditions that an optimal quantizer satisfies are determined. A high-rate approximation to the distortion is also worked through. Previous, though significantly different, work on quantization for hypothesis testing is unable to directly minimize the Bayes risk, as is accomplished

in this work. This formulation of hypothesis testing is also examined when empirical data samples related to the distribution of prior probabilities are given instead of the distribution itself. Overfitting is observed with empirical data, and clustering the prior probabilities aids detection performance.

The mathematical theory of quantized prior hypothesis testing formulated here leads to a generative model of discriminative behavior when combined with theories of social cognition and facts about social segregation. This biased decision making arises despite having identical distributions for different populations and despite no malicious intent on the part of the decision maker. Precautionary settings of Bayes costs lead to a higher probability of declaring the positive hypothesis for the opposite race, whereas the opposite setting of costs leads to a higher probability of declaring the positive hypothesis for the own race. Such a phenomenon of precaution fundamentally altering the nature of discrimination seems not to have been described before.

Chapter 6

Conclusion

SEVERAL aspects of frugal decision making are studied in the preceding chapters, namely margin-based classification with decision boundary surface area regularization, dimensionality reduction for margin-based classification, and the quantization or clustering of priors for Bayesian hypothesis testing. These contributions of the thesis are summarized in this chapter. The chapter also documents directions for further research suggested by the work described in the thesis.

6.1 Summary of Contributions

The thesis examines rules for detection that are limited in complexity. One novel measure of complexity that is investigated is the surface area of the decision boundary. Another novel way to limit complexity that is proposed is through the quantization of prior probabilities appearing in the threshold of the likelihood ratio test. In addition to these two new perspectives on complexity control in decision making, dimensionality reduction for classification is studied as well. Dimensionality reduction specifically within a margin-based classification objective has not received attention before, and neither has distributed dimensionality reduction for classification in sensor networks, as studied in the thesis.

More specifically, a new classifier is developed, the geometric level set (GLS) classifier. A method for joint dimensionality reduction and margin-based classification via optimization on the Stiefel manifold is developed. Also developed is a new distortion criterion for quantization, the mean Bayes risk error (MBRE). Two application areas, sensor networks and human decision making in segregated populations, are looked at in greater depth and lead to extensions of the formulations.

Lehman [112] lists life in the age of the genome, wisdom in the age of digital information, and sustainability in the age of global development as significant research challenges. The work in this thesis contributes to the second of these challenges through the theme of frugality in the age of excess. The essence of wisdom is captured by the generalizability of a decision rule. Generalization is improved by controlling the complexity of a decision rule learned from noisy samples of information. This improvement is illustrated on datasets throughout the thesis, whether the complexity is measured by surface area, dimensionality, or number of clusters. An added benefit of limiting

the complexity of a decision rule is the accompanying reduction of physical costs and resource usage that often occurs.

■ 6.1.1 Geometric Level Set Classifier

Supervised classification is an important learning problem that occurs in many application domains. Many approaches to the problem have been developed, including decision trees, k-nearest neighbor classifiers, and neural networks [56]. Nonparametric methods based on margin-based loss functions including logistic regression and the support vector machine (SVM) are quite popular [99, 176]. No one classifier is always superior [222]; classifier performance is dataset-dependent and thus developing new classifiers is useful.

Viewed from a regularization perspective, existing margin-based classifiers including the SVM often have a squared Hilbert space norm as the regularizer [99], which penalizes properties of the entire decision function. The GLS classifier is proposed as an alternative margin-based classifier in which the regularization term looks at only the zero level set of the decision function. The zero level set is the decision boundary and the only part of the decision function that affects classification performance. Specifically, the regularization term of the GLS classifier is the surface area of the zero level set of the decision function, which promotes smooth decision boundaries when minimized. This geometric regularization term for margin-based classification has not appeared in the statistical learning literature before. This sort of geometric regularization opens up possibilities for other decision boundary preferences to be encoded in the classification objective. One of these possibilities for feature subset selection is described in the thesis.

Unlike SVM training [99], GLS classifier training cannot be expressed in a quadratic programming or other convex optimization form. However, GLS classifier training is ideally suited to variational level set methods [141]. The training is carried out with contour evolution, an Euler–Lagrange descent procedure that is not typically used in statistical learning. In contrast to kernel methods, the GLS classifier with level set formulation finds nonlinear decision boundaries directly in the input space and respects distance relationships in the input space [2]. The level set formulation also enables a new multicategory classification approach in which the number of decision functions is the logarithm of the number of classes. As shown in the thesis, there are real-world datasets for which the GLS classifier outperforms several popular classifiers from the literature.

A new classifier requires new statistical learning theory analysis. The VC dimension of the GLS classifier is measured empirically in the thesis for use in generalization bounds [205]. Also for use in generalization bounds, the ϵ-entropy [106] of the GLS classifier is calculated and used to analytically characterize its Rademacher complexity [13]. Additionally, the GLS classifier is shown to be consistent as the number of training samples goes to infinity as long as the margin-based loss function is Fisher-consistent [116].

■ 6.1.2 Joint Dimensionality Reduction and Margin-Based Classification

Dimensionality reduction is useful to eliminate irrelevant and redundant dimensions of data [149]. However, irrelevance and redundancy cannot be defined without context, as the final use of the data determines what is irrelevant and redundant. When this final use is known, the mapping to a reduced-dimensional space should be optimized for that purpose [188]. Thus when the reduced-dimensional data is to be used for supervised classification, the reduced-dimensional space and the classifier should be learned jointly.

Among existing supervised dimensionality reduction methods for classification, many are founded on strong assumptions about the data generating process such as Gaussianity [15, 69, 119, 128, 195]. Other methods have objectives beyond explicit classification performance, specifically that likelihood functions conditioned on full-dimensional data match those conditioned on reduced-dimensional data [42, 74, 170]. Dimensionality reduction within a nonparametric classification framework has not received as much attention [150, 151, 202]. In the thesis, a formulation is proposed for learning a nonparametric margin-based classifier defined in a reduced-dimensional space along with learning the reduced-dimensional space. When the reduced-dimensional space is a linear subspace, the objective for margin-based classification is extended to include a dimensionality reduction matrix constrained to the Stiefel manifold that multiplies data vectors within the argument of the decision function. As shown on several datasets, linear dimensionality reduction improves classification performance by supressing noise in the data and preventing overfitting.

Optimization in the proposed formulation is approached through coordinate descent with alternating minimizations for the classifier decision function and the dimensionality reduction matrix. Optimization for the dimensionality reduction matrix with a fixed decision function involves gradient descent along Stiefel manifold geodesics [60], whereas optimization for the decision function with a fixed matrix is performed in the typical manner for the margin-based classifier, such as quadratic programming techniques for the SVM and contour evolution for the GLS classifier. Analyses of consistency and Rademacher complexity are provided for joint dimensionality reduction and GLS classification. Like those for the GLS classifier without dimensionality reduction, the analyses are based on ϵ-entropy [106]. Calculation of ϵ-entropy with dimensionality reduction involves the additional ingredient of zonotope content [37, 65].

For certain data distributions, limiting the reduced-dimensional space to be linear is restrictive. An extension is provided in the thesis that allows the reduced-dimensional space to be a nonlinear manifold. For nonlinear dimensionality reduction, the Stiefel manifold-constrained matrix multiplies the vector-valued result of a data-dependent nonlinear kernel function [19], also within the argument of the decision function. However, results with several real-world datasets seem to indicate that linear dimensionality reduction is sufficient for optimal classification performance; the added power of nonlinear dimensionality reduction is not necessary.

Networks of power-limited sensor nodes are deployed for tasks such as environmental monitoring and surveillance. The sensors often measure vector-valued data and a

fusion center node in the network often performs detection or classification [226, 227]. Limits on battery power translate into limits on communication; communication may be reduced if the dimensionality of measurements is reduced locally at sensors prior to transmission. The coordinate descent alternating minimizations for joint dimensionality reduction and margin-based classification are structured in a way that permits processing to be distributed between the fusion center and sensor nodes. Messages communicated between nodes in both training and operation are vectors of the reduced dimension rather than the full measurement dimension. The thesis shows how the distributed supervised dimensionality reduction is applicable to tree-structured, fusion center-rooted sensor networks with data fusion at intermediate layers of the tree. Joint dimensionality reduction and margin-based classification both reduces communication and improves classification performance in sensor networks.

■ 6.1.3 Minimum Mean Bayes Risk Error Distortion

The likelihood ratio test threshold of a Bayesian detector is based on the class prior probabilities of the measured object [221]. A population of many objects with different prior probabilities may confound or overwhelm a detector that is unable to precisely adapt the threshold for each object due to processing limitations. The prior probabilities of the many objects are set based on noisy observations in several scenarios, in which case some form of regularization may be beneficial. Quantization or clustering of the prior probabilities across the population of objects serves to reduce information storage and recall requirements, and also provides a form of regularization. This mode of complexity control in decision making has not been studied before.

Several distortion criteria exist for quantization and clustering [79]. As also noted with dimensionality reduction, if the final use of complexity-reduced data is known, then the complexity reduction mapping should be optimized for that final use. MBRE, a new distortion criterion for quantization and clustering, is proposed in the thesis. This distortion criterion arises directly from the Bayes risk of a likelihood ratio test detection rule, and is thus matched to the decision-making application of interest. It is shown in examples containing prior probabilities estimated from observations with additive Gaussian noise that MBRE-optimal clustering improves decision-making performance.

Several properties of the novel MBRE distortion criterion are derived in the thesis. Specifically, it is shown that Bayes risk error is continuous and strictly convex as a function of the unquantized variable and quasiconvex as a function of the quantized variable. The properties are then used in deriving quantizer optimality conditions and showing that the conditions are both necessary and sufficient [200]. Quantization analysis in the high-resolution limit is also provided in the thesis [113], and is of the same nature as consistency analysis for classification.

The final contribution of the thesis is a model of decision making on the actions of humans by humans. The model is an extension of the quantization/clustering framework that includes more than one population type or race, with decision makers having more training or experience with members within their own population type. The model

predicts discrimination against population members who are of a different type than the decision maker. Significantly more false alarms on members of other population types are noted in empirical studies of society [8, 53, 159, 190]. This finding is reproduced by optimal decision making with quantization when the cost of missed detections is higher than the cost of false alarms in the Bayes risk of the hypothesis test.

6.2 Recommendations for Future Research

The contributions summarized in the first half of this chapter are not only worthwhile intrinsically, but also because they illuminate avenues for further research. Several directions for future work are described in this section. Some are obvious extensions and questions that arise in previous chapters; others are farther afield.

6.2.1 Model Selection

Two model selection questions are not fully resolved in the thesis: how should the complexity level of the decision rule be set, and when should the GLS classifier be used instead of other classifiers from the literature. These are both operationally important questions.

Selecting Complexity Level

As seen throughout the thesis and articulated by the structural risk minimization principle, there is an intermediate decision rule complexity at which generalization error is minimized. The question of how to best set the GLS regularization parameter, the reduced dimension, or the number of clusters based on the available training data is not answered in the thesis (except by cross-validation for the GLS regularization parameter) and is an interesting direction for future work. This complexity level choice may be confounded by resource usage constraints, for example in sensor networks.

Similar to other classifiers, the generalization bounds obtained using Rademacher complexity for the GLS classifier and for the GLS classifier with dimensionality reduction are not tight, and not useful in setting the regularization parameter or number of dimensions. Methods based on cross-validation, bootstrapping, and information criteria may be used [3, 62, 178]. An interesting way forward is through Bayesian nonparametric methods, such as automatic relevance determination in dimensionality reduction [23] and the Dirichlet process mixture model in clustering [64]. In fact, Canini [32] has shown that the Dirichlet process mixture is a rational model for the categorization performed by humans. It has also been used in the economics literature to model human decision making and choice [30].

Selecting the GLS Classifier

As seen in the table of results, Table 3.1 in Section 3.3, the relative performance of different classifiers is dataset-dependent. A classifier may work well on some datasets

and poorly on others. These differences are because different classifiers make different implicit or explicit assumptions on how to generalize from the finite samples of the training set. The question then is which classifier should be used for a given dataset.

Several geometric and topological properties that delineate domains of classifier competence have been catalogued [14, 94]. Different regions of data complexity space correspond to the types of datasets for which different classifiers work well or poorly. This line of work also includes classification with feature selection or dimensionality reduction [156]. It would be interesting to determine the dataset properties for which the GLS classifier, including with dimensionality reduction, is preferred.

■ 6.2.2 Extensions to Other Learning Scenarios

There are several other statistical learning scenarios besides the batch supervised classification problem considered in Chapter 3 and Chapter 4, including active learning [45], reinforcement learning [101], and sequential supervised learning [51]. An interesting future research path is to extend the GLS classifier to three other learning scenarios: online learning, semisupervised learning, and Neyman–Pearson learning.

Online Learning

The GLS classifier, both with and without joint dimensionality reduction, is optimized in a gradient descent manner. As described in the thesis, the gradient descent is a batch operation for all training samples at once. An alternative to batch training is online training [28], which is closely related to adaptive filtering [217]. In online learning, gradient updates are based on single training samples. Modification of GLS classification to online learning should be straightforward, but stochastic approximation analysis of the dynamics and convergence may be more involved.

Online learning makes computation manageable in large-scale problems with very large training sets. It is also useful in tracking data distributions that are not stationary. For example, in classifying internet traffic, the joint distribution of measurements and class labels changes over time and also has periodicity by the time of day, day of week, and seasonally. A geological example is in classifying the type of rock from measurements in wells—rock properties change with depth under the surface of the earth. In fact, variational level set methods are regularly used for tracking deformable objects in time sequences [145, 191].

An online learning paradigm may also be useful in the geological example if training the system prior to deployment under the earth is inexpensive compared to training during deployment. This is more apt to occur when also considering dimensionality reduction. The training data available before deployment may not have the same distribution as the data encountered when deployed, but may provide a good initialization. The same scenario applies elsewhere, such as in deep space exploration and in wireless sensor networks.

Semisupervised Learning

Semisupervised learning is important because obtaining measurement vectors for training is often inexpensive, but obtaining labels for those vectors is costly [39]. As discussed in the thesis, the variational level set framework of the GLS classifier is amenable to the inclusion of additional geometric regularization terms in the objective. The semisupervised learning approach of Belkin et al. [18] with manifold regularization has in its objective a margin-based loss term, a regularization term on the decision function defined in the ambient space, and a regularization term related to the intrinsic geometry of the distribution of the measurements. This geometric third term is based on the training samples, but does not utilize the class labels. It seems quite likely that a close variant of the third term in [18] could be adopted in a variational level set formulation to extend GLS classification for semisupervised learning.

Neyman–Pearson Learning

False alarms and missed detections often have different consequences. In certain decision-making scenarios, it is more natural to specify a maximum false alarm rate rather than the cost ratio c_{-+}/c_{+-}. A fairly new paradigm when the decision rule is to be learned from training data for this case is Neyman–Pearson classification [179]. Statistical learning bounds have been developed for this scenario, including ones based on Rademacher complexity [91]. However, practical classification methods are not as developed. In future work, it would be interesting to modify the GLS classifier for Neyman–Pearson classification.

■ 6.2.3 Quantization/Clustering of Prior Probabilities

The minimum mean Bayes risk error quantization and clustering framework of Chapter 5 suggests some directions for future work.

Statistical Learning Theory Analysis

In Section 5.3, the decision maker learns about the prior probabilities in the population from a finite number of noisy measurements. Further work includes rigorously showing decision-making consistency with this setup. Several ingredients for this are discussed in the thesis, including the consistency of the estimator that yields a prior probability estimate for a single object from m noisy measurements as m grows, the convergence of the clustering to the quantizer as the number of objects n grows [169], and the high-resolution analysis showing that the excess Bayes risk goes to zero as k, the number of quantization levels, grows. Another statistical learning theory analysis of interest for future work is to develop generalization bounds as a function of finite m, n, and k.

Min-Max Representation Point

For the quantizer with $k = 1$, an alternative to the MBRE-optimal representation point

$$a^*_{\text{MBRE}} = \arg\min_a \left\{ \int \tilde{\text{R}}(p_-, a) f_{\mathsf{p}_-}(p_-) dp_- \right\} \tag{6.1}$$

is the min-max hypothesis testing representation point [221]:

$$a^*_{\text{min-max}} = \arg\min_a \left\{ \max_{p_-} \tilde{\text{R}}(p_-, a) \right\}. \tag{6.2}$$

The two representation points a^*_{MBRE} and $a^*_{\text{min-max}}$ are only equivalent in special cases. A distribution on the prior probabilities is needed to specify a^*_{MBRE}, but not to specify $a^*_{\text{min-max}}$. An interesting future direction is to extend the min-max idea to $k > 1$. This would involve finding a covering consisting of k sets of the form $\mathcal{Q}_i = \{p_- | \tilde{\text{R}}(p_-, a_i) \leq \varrho\}$, where all p_- in $\mathcal{Q}_i$ map to a_i and the radius ϱ is the same for all $\mathcal{Q}_i$.

■ 6.2.4 Posterior Probability and Confidence

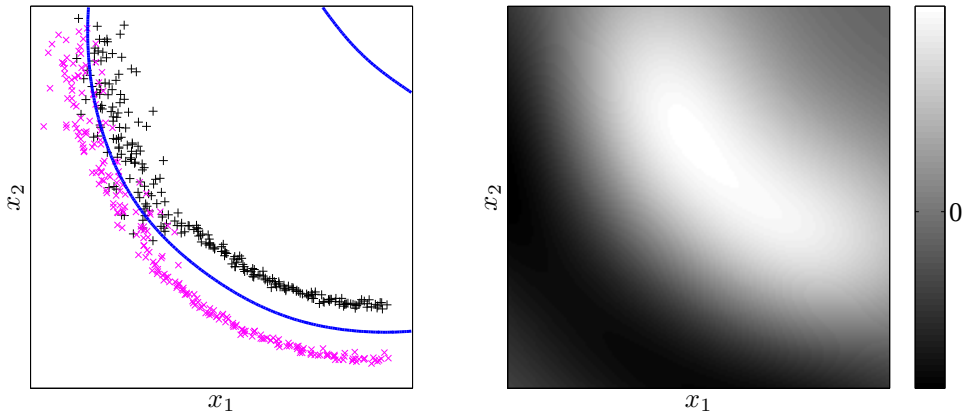
It is useful for a classification algorithm to provide the posterior probability $f_{\mathsf{y}|\mathbf{x}}(y|\mathbf{x} = \mathbf{x})$ in addition to just the decision boundary in many decision-making tasks, including those in which several different classifiers are to be combined and in those where the costs c_{+-} and c_{-+} are unequal. The posterior probability is also useful in showing where in the measurement space the classification is more or less uncertain and to provide confidence intervals.

The margin-based classifier $\hat{y}(\mathbf{x}) = \text{sign}(\varphi(\mathbf{x}))$ does not provide this posterior probability. Platt [153] suggests that a sigmoid function of the decision function φ may be used as a surrogate for the posterior probability:

$$f_{\mathsf{y}|\boldsymbol{\varphi}}(y = +1|\boldsymbol{\varphi} = \varphi) = \frac{1}{1 + \exp(A\varphi + B)}, \tag{6.3}$$

where A and B are constants to be fit using the training data. This form (6.3) is supported by the theoretical foundations of logistic regression as well. However, this surrogate based on the margin can be quite poor. As an example, consider the dataset and SVM decision function shown in Figure 6.1. The decision boundary in the upper left part of the domain is more uncertain than the decision boundary in the lower right, but since the decision function in both parts is essentially the same, the posterior probability using (6.3) is also incorrectly the same.

Future work suggested by this issue is to use a Monte Carlo approach to sample from $f_{\mathsf{y}|\mathbf{x}}(y|\mathbf{x} = \mathbf{x})$. Markov chain Monte Carlo curve sampling, a recently developed approach to image segmentation, may be adapted for this purpose [41, 63]. Given many samples from the posterior distribution, level sets of the histogram may then be used to show confidences. These different level sets may also be used as decision boundaries for classification with $c_{+-} \neq c_{-+}$.



(a) (b)

Figure 6.1. Illustration that the decision function φ of a margin-based classifier may not be a good surrogate for the posterior probability. (a) The magenta $\times$ markers indicate class label -1 and the black $+$ markers indicate class label $+1$. The blue line is the decision boundary of the SVM with radial basis function kernel. (b) The decision function $\varphi(\mathbf{x})$ of the SVM.

6.2.5 Nonlinear Dimensionality Reduction

The classification results with nonlinear dimensionality reduction presented in Section 4.2 utilize the SVM with radial basis function kernel. This nonlinear dimensionality reduction with nonlinear classification does not perform as well as linear dimensionality reduction, which may be due to too much freedom in both nonlinearities. Braun et al. [29] have studied this issue and found that the use of linear decision boundaries with nonlinear dimensionality reduction is often sufficient. It would be interesting to systematically study the proposed joint nonlinear dimensionality reduction framework with simpler margin-based classifiers having linear or quadratic decision boundaries.

Additionally, the data-dependent Isomap kernel requires an $n \times d$ dimensionality reduction matrix, where n is the number of training samples. If n is large, it is possible to use the Landmark Isomap idea, which uses a subsampling of the training set in representing the nonlinear manifold [184]. It would be interesting to systematically investigate the effect of subsampling on classification performance. With labeled samples, it is also possible to subsample different classes differently and affect false alarm and missed detection rates differently.

6.2.6 Sensor Networks

The distributed setting of the sensor network application discussed in the context of supervised dimensionality reduction and information fusion provides problems for further study.

General Message-Passing Approach

The dimensionality reduction extension for sensor networks in Section 4.4 focuses on the margin-based classification objective. The optimization reveals an efficient message-passing structure. Objectives other than margin-based classification can also be considered within the sensor network structure, for example independent component analysis. It would be interesting to develop a general message-passing approach for optimizing dimensionality reduction functions $f(\mathbf{A}^T\mathbf{x})$ in distributed sensor fusion settings. It would also be interesting to investigate the behavior of the message passing in network structures that are not trees, to see if the computations converge, and if so, converge to correct solutions.

Network Architecture Optimization

The plots of classification performance as a function of transmission power for wireless sensor networks given in Section 4.4.4 are for three specific choices of network architecture: serial, parallel, and binary tree, with the reduced dimensionality at a node proportional to the number of descendents plus one. The problem of determining the reduced dimensionality at each sensor as well as the network structure that minimizes generalization error for given training measurements and a given transmission power budget is an interesting research direction. Prior work on sensor network architecture optimization has tended to focus on detection scenarios with known probability distributions [7]; the statistical learning formulation with associated overfitting effects has not been the focus.

■ 6.2.7 PDE Methods

Variational and geometric partial differential equation (PDE) methods abound in the field of image processing. Level set image segmentation is one such example, which is adapted in the development of the GLS classifier in Chapter 3. A future research direction is to examine PDE-based methods for other image processing tasks including inpainting and interpolation, denoising, deblurring and deconvolution, contrast enhancement, anisotropic diffusion, warping, and blending [142, 173] to see if there are statistical learning problems for which they may be adapted. Such an inquiry may be quite fruitful. Image processing methods generally assume that data samples lie on a pixel or voxel grid, but as in developing the GLS classifier, this should not be a major hurdle.

It is mentioned at the beginning of Chapter 3 that spatially continuous and spatially discrete representations lead to different types of segmentation approaches. There have been recent attempts to reconcile the different approaches by extending the definitions of PDE methods to graphs, including papers by Zhou and Schölkopf [224] and Gilboa and Osher [81]. These works have also bridged image processing and statistical learning. In future work, it would be interesting to develop a discrete formulation for GLS classification based on these new formulations of PDEs on graphs.

Pixels are the data for PDE methods in image processing, whereas training samples in Euclidean space are the data for the GLS classifier. An interesting direction is to look at running variational PDE methods with the data being probability distribution sample points on manifolds defined by information geometry [6]. Problems of interpolating a probability distribution between given probability distributions, and denoising a collection of empirical probability distributions do occasionally arise. The blurring and deblurring of probability distributions could arise in multiscale and multiresolution models. It does not seem as though there has been work along these lines.

Bibliography

[1] Dennis J. Aigner and Glen G. Cain. Statistical theories of discrimination in labor markets. *Industrial and Labor Relations Review*, 30(2):175–187, January 1977. 138

[2] Shotaro Akaho. SVM that maximizes the margin in the input space. *Systems and Computers in Japan*, 35(14):78–86, December 2004. 58, 146

[3] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19(6):716–723, December 1974. 149

[4] Erin L. Allwein, Robert E. Schapire, and Yoram Singer. Reducing multiclass to binary: A unifying approach to margin classifiers. *Journal of Machine Learning Research*, 1:113–141, December 2000. 63

[5] Daniel Aloise, Amit Deshpande, Pierre Hansen, and Preyas Popat. NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248, May 2009. 52

[6] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*. American Mathematical Society, Providence, Rhode Island, 2000. 155

[7] Animashree Anandkumar. *Scalable Algorithms for Distributed Statistical Inference*. PhD thesis, Cornell University, Ithaca, New York, August 2009. 154

[8] Kate Antonovics and Brian G. Knight. A new look at racial profiling: Evidence from the Boston Police Department. *The Review of Economics and Statistics*, 91 (1):163–177, February 2009. 24, 137, 141, 143, 149

[9] Kenneth Arrow. The theory of discrimination. In Orley Ashenfelter and Albert Rees, editors, *Discrimination in Labor Markets*, pages 3–33. Princeton University Press, Princeton, New Jersey, 1973. 138

[10] Arthur Asuncion and David J. Newman. *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences, `http://archive.ics.uci.edu/ml`, 2007. 66, 69, 94, 100

[11] R. Michael Bagby, James D. Parker, Neil A. Rector, and Valery Kalemba. Racial prejudice in the Canadian legal system: Juror decisions in a simulated rape trial. *Law and Human Behavior*, 18(3):339–350, June 1994. 143

[12] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473): 138–156, March 2006. 39, 74, 78

[13] Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3: 463–482, November 2002. 19, 21, 22, 34, 35, 72, 146

[14] Mitra Basu and Tin Kam Ho, editors. *Data Complexity in Pattern Recognition*. Springer-Verlag, London, 2006. 150

[15] Gaston Baudat and Fatiha Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12(10):2385–2404, October 2000. 85, 86, 97, 147

[16] Gary S. Becker. *The Economics of Discrimination*. University of Chicago Press, Chicago, 1957. 138

[17] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, June 2003. 47, 86, 97

[18] Mikhail Belkin, Patha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, November 2006. 99, 100, 151

[19] Yoshua Bengio, Olivier Delalleau, Nicolas Le Roux, Jean-François Paiement, Pascal Vincent, and Marie Ouimet. Learning eigenfunctions links spectral embedding and kernel PCA. *Neural Computation*, 16(10):2197–2219, October 2004. 20, 22, 48, 86, 97, 98, 147

[20] José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. John Wiley & Sons, Chichester, United Kingdom, 2000. 17

[21] Marianne Bertrand, Dolly Chugh, and Sendhil Mullainathan. Implicit discrimination. *The American Economic Review*, 95(2):94–98, May 2005. 138

[22] Pratip Bhattacharyya and Bikas K. Chakrabarti. The mean distance to the nth neighbour in a uniform distribution of random points: An application of probability theory. *European Journal of Physics*, 29(3):639–645, May 2008. 112

[23] Christopher M. Bishop. Bayesian PCA. In Michael S. Kearns, Sara A. Solla, and David A. Cohn, editors, *Advances in Neural Information Processing Systems 11*, pages 382–388. MIT Press, Cambridge, Massachusetts, 1999. 149

[24] Gilles Blanchard, Christin Schäfer, Yves Rozenholc, and Klaus-Robert Müller. Optimal dyadic decision trees. *Machine Learning*, 66(2–3):209–241, March 2007. 69

[25] Gilles Blanchard and Laurent Zwald. Finite-dimensional projection for classification and statistical learning. *IEEE Transactions on Information Theory*, 54(9): 4169–4182, September 2008. 85, 105

[26] David M. Blei and Jon D. McAuliffe. Supervised topic models. In John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 121–128, Cambridge, Massachusetts, 2008. MIT Press. 88

[27] Erik M. Boczko, Todd R. Young, Minhui Xie, and Di Wu. Comparison of binary classification based on signed distance functions with support vector machines. In *Proceedings of the Ohio Collaborative Conference on Bioinformatics*, Athens, Ohio, June 2006. 56

[28] Léon Bottou. Stochastic learning. In Olivier Bousquet, Ulrike von Luxburg, and Gunnar Rätsch, editors, *Advanced Lectures on Machine Learning*, pages 146–168, Tübingen, Germany, August 2003. 150

[29] Mikio L. Braun, Joachim M. Buhmann, and Klaus-Robert Müller. On relevant dimensions in kernel feature spaces. *Journal of Machine Learning Research*, 9: 1875–1908, August 2008. 153

[30] Martin Burda, Matthew Harding, and Jerry Hausman. A Bayesian mixed logit-probit model for multinomial choice. *Journal of Econometrics*, 147(2):232–246, December 2008. 149

[31] Xiongcai Cai and Arcot Sowmya. Level learning set: A novel classifier based on active contour models. In Joost N. Kok, Jacek Koronacki, Ramon Lopez de Mantaras, Stan Matwin, Dunja Mladenič, and Andrzej Skowron, editors, *Proceedings of the 18th European Conference on Machine Learning*, pages 79–90, Warsaw, Poland, 2007. 56, 70, 71

[32] Kevin Canini. Modeling categorization as a Dirichlet process mixture. Technical Report UCB/EECS-2007-69, Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, May 2007. 149

[33] Kevin M. Carter, Raviv Raich, and Alfred O. Hero, III. An information geometric approach to supervised dimensionality reduction. In *Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing*, Taipei, Taiwan, April 2009. 86

[34] Vicent Caselles, Ron Kimmel, and Guillermo Sapiro. Geodesic active contours. *International Journal of Computer Vision*, 22(1):61–79, February 1997. 41

[35] Thomas Cecil, Jianliang Qian, and Stanley Osher. Numerical methods for high dimensional Hamilton–Jacobi equations using radial basis functions. *Journal of Computational Physics*, 196(1):327–347, May 2004. 66

[36] Müjdat Çetin, Lei Chen, John W. Fisher, III, Alexander T. Ihler, Randolph L. Moses, Martin J. Wainwright, and Alan S. Willsky. Distributed fusion in sensor networks. *IEEE Signal Processing Magazine*, 23(4):42–55, July 2006. 18, 107

[37] G. Donald Chakerian and Paul Filliman. The measures of the projections of a cube. *Studia Scientiarum Mathematicarum Hungarica*, 21(1–2):103–110, 1986. 20, 22, 46, 47, 147

[38] Jean-François Chamberland and Venugopal V. Veeravalli. Decentralized detection in sensor networks. *IEEE Transactions on Signal Processing*, 51(2):407–416, February 2003. 22, 107

[39] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, Massachusetts, 2006. 151

[40] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998. 80

[41] Siqi Chen and Richard J. Radke. Markov chain Monte Carlo shape sampling using level sets. In *Proceedings of the Second Workshop on Non-Rigid Shape Analysis and Deformable Image Alignment*, Kyoto, Japan, September 2009. 152

[42] Francesca Chiaromonte and R. Dennis Cook. Sufficient dimension reduction and graphics in regression. *Annals of the Institute of Statistical Mathematics*, 54(4): 768–795, December 2002. 86, 147

[43] Yasuko Chikuse. *Statistics on Special Manifolds*. Springer-Verlag, New York, 2002. 47

[44] Stephen Coate and Glenn C. Loury. Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review*, 83(5):1220–1240, December 1993. 138

[45] David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, January 1996. 150

[46] John Conlisk. Why bounded rationality? *Journal of Economic Literature*, 34(2): 669–700, June 1996. 18, 137

[47] Trevor F. Cox and Michael A. A. Cox. *Multidimensional Scaling*. Chapman & Hall/CRC, Boca Raton, Florida, 2001. 48

[48] Daniel Cremers, Mikael Rousson, and Rachid Deriche. A review of statistical approaches to level set segmentation: Integrating color, texture, motion and shape. *International Journal of Computer Vision*, 72(2):195–215, April 2007. 56

[49] Morris H. DeGroot. *Optimal Statistical Decisions*. Wiley-Interscience, Hoboken, New Jersey, 2004. 29, 119

[50] Michel C. Delfour and Jean-Paul Zolésio. *Shapes and Geometries: Analysis, Differential Calculus, and Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 2001. 41, 75

[51] Thomas G. Dietterich. Machine learning for sequential data: A review. In Terry Caelli, Adnan Amin, Robert P. W. Duin, Mohamed Kamel, and Dick de Ridder, editors, *Proceedings of the Joint IAPR International Workshops on Structural, Syntactic, and Statistical Pattern Recognition*, pages 15–30, Windsor, Canada, August 2002. 150

[52] Carlotta Domeniconi, Dimitrios Gunopulos, and Jing Peng. Large margin nearest neighbor classifiers. *IEEE Transactions on Neural Networks*, 16(4):899–909, July 2005. 79, 80

[53] John J. Donohue, III and Steven D. Levitt. The impact of race on policing and arrests. *Journal of Law and Economics*, 44(2):367–394, October 2001. 24, 137, 141, 149

[54] James Dow. Search decisions with limited memory. *The Review of Economic Studies*, 58(1):1–14, January 1991. 137, 138

[55] Petros Drineas, Alan Frieze, Ravi Kannan, Santosh Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition. *Machine Learning*, 56 (1–3):9–33, July 2004. 52

[56] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. John Wiley & Sons, New York, 2001. 146

[57] Richard M. Dudley. Metric entropy of some classes of sets with differentiable boundaries. *Journal of Approximation Theory*, 10(3):227–236, March 1974. 74

[58] Richard M. Dudley. Correction to "metric entropy of some classes of sets with differentiable boundaries". *Journal of Approximation Theory*, 26(2):192–193, June 1979. 74

[59] Federico Echenique and Roland G. Fryer, Jr. A measure of segregation based on social interactions. *Quarterly Journal of Economics*, 122(2):441–485, May 2007. 24, 138

[60] Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, January 1998. 18, 20, 46, 147

[61] Bradley Efron. The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70(352):892–898, December 1975. 55

[62] Bradley Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331, June 1983. 149

[63] Ayres C. Fan, John W. Fisher, III, William M. Wells, III, James J. Levitt, and Alan S. Willsky. MCMC curve sampling for image segmentation. In Nicholas Ayache, Sébastien Ourselin, and Anthony Maeder, editors, *Proceedings of the 10th International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume 2, pages 477–485, Brisbane, Australia, October–November 2007. 152

[64] Thomas S. Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1(2):209–230, March 1973. 149

[65] Paul Filliman. Extremum problems for zonotopes. *Geometriae Dedicata*, 27(3): 251–262, September 1988. 20, 22, 46, 147

[66] Terrence L. Fine. *Probability and Probabilistic Reasoning for Electrical Engineering*. Prentice Hall, Upper Saddle River, New Jersey, 2006. 123

[67] Bruno de Finetti. La prévision: Ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7:1–68, 1937. 17

[68] Ross L. Finney, Maurice D. Weir, and Frank R. Giordano. *Thomas' Calculus*. Addison Wesley, Boston, 2001. 41

[69] Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936. 27, 45, 86, 147

[70] Andrew R. Forsyth. *Calculus of Variations*. Dover Publications, New York, 1960. 41

[71] Dean P. Foster and Rakesh V. Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21(1–2):40–55, October 1997. 139

[72] Roland Fryer and Matthew O. Jackson. A categorical model of cognition and biased decision-making. *The B. E. Journal of Theoretical Economics*, 8(1), January 2008. 50, 137, 138, 139

[73] Kenji Fukumizu, Francis R. Bach, and Michael I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, January 2004. 86

[74] Kenji Fukumizu, Francis R. Bach, and Michael I. Jordan. Kernel dimension reduction in regression. *Annals of Statistics*, 37(4):1871–1905, August 2009. 86, 147

[75] Robert G. Gallager. *Principles of Digital Communication*. Cambridge University Press, Cambridge, United Kingdom, 2008. 50

[76] Michael R. Garey, David S. Johnson, and Hans S. Witsenhausen. The complexity of the generalized Lloyd-Max problem. *IEEE Transactions on Information Theory*, IT-28(2):255–256, March 1982. 52

[77] Michael Gastpar, Pier Luigi Dragotti, and Martin Vetterli. The distributed Karhunen–Loève transform. *IEEE Transactions on Information Theory*, 52(12): 5177–5196, December 2006. 108

[78] Arnaud Gelas, Olivier Bernard, Denis Friboulet, and Rémy Prost. Compactly supported radial basis functions based collocation method for level-set evolution in image segmentation. *IEEE Transactions on Image Processing*, 16(7):1873–1887, July 2007. 21, 66, 67

[79] Allen Gersho and Robert M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, Boston, 1992. 20, 27, 50, 121, 148

[80] Michael Gibbert, Martin Hoegl, and Liisa Välikangas. In praise of resource constraints. *MIT Sloan Management Review*, 48(3):15–17, Spring 2007. 17

[81] Guy Gilboa and Stanley Osher. Nonlocal operators with applications to image processing. *Multiscale Modeling and Simulation*, 7(3):1005–1028, 2008. 154

[82] Itzhak Gilboa and David Schmeidler. Case-based decision theory. *The Quarterly Journal of Economics*, 110(3):605–639, August 1995. 137

[83] Dennis C. Gilliland and Merrilee K. Helmers. On continuity of the Bayes response. *IEEE Transactions on Information Theory*, IT-24(4):506–508, July 1978. 118

[84] Malcolm Gladwell. *Blink: The Power of Thinking Without Thinking*. Little, Brown and Company, New York, 2005. 17

[85] Yair Goldberg and Ya'acov Ritov. Local Procrustes for manifold embedding: A measure of embedding quality and embedding algorithms. *Machine Learning*, 77 (1):1–25, October 2009. 47

[86] José Gomes and Aleksandra Mojsilović. A variational approach to recovering a manifold from sample points. In Anders Heyden, Gunnar Sparr, Mads Nielsen, and Peter Johansen, editors, *Proceedings of the 7th European Conference on Computer Vision*, volume 2, pages 3–17, Copenhagen, Denmark, May 2002. 97

[87] Robert M. Gray and Augustine H. Gray, Jr. Asymptotically optimal quantizers. *IEEE Transactions on Information Theory*, IT-23(1):143–144, January 1977. 130

[88] Robert M. Gray, John C. Kieffer, and Yoseph Linde. Locally optimal block quantizer design. *Information and Control*, 45(2):178–198, May 1980. 53, 134

[89] Robert M. Gray and David L. Neuhoff. Quantization. *IEEE Transactions on Information Theory*, 44(6):2325–2383, October 1998. 23, 50, 52, 126

[90] Riten Gupta and Alfred O. Hero, III. High-rate vector quantization for detection. *IEEE Transactions on Information Theory*, 49(8):1951–1969, August 2003. 118

[91] Min Han, Di Rong Chen, and Zhao Xu Sun. Rademacher complexity in Neyman–Pearson classification. *Acta Mathematica Sinica, English Series*, 25(5):855–868, May 2009. 151

[92] Matthias Hein. *Geometrical Aspects of Statistical Learning Theory*. Dr. rer. nat. dissertation, Technische Universität Darmstadt, Germany, November 2005. 20

[93] Clifford Hildreth. Bayesian statisticians and remote clients. *Econometrica*, 31(3): 422–438, July 1963. 118

[94] Tin Kam Ho and Mitra Basu. Complexity measures of supervised classification problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24 (3):289–300, March 2002. 56, 150

[95] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, September 1933. 27, 45, 85

[96] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(7):498–520, October 1933. 27, 45, 85

[97] Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, March 2002. 62

[98] Ke Huang and Selin Aviyente. Sparse representation for signal classification. In Bernhard Schölkopf, John C. Platt, and Thomas Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 609–616. MIT Press, Cambridge, Massachusetts, 2007. 88

[99] Tommi S. Jaakkola. *6.867 Machine Learning Course Notes*. Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, Fall 2007. 19, 35, 37, 146

[100] C. Richard Johnson, Jr. and Ella Hendriks. Background. In *Proceedings of the 1st International Workshop on Image Processing for Artist Identification*, pages 15–34, Amsterdam, Netherlands, May 2007. 17

[101] Leslie Pack Kaelbling, Michael L. Littman, and Andrew W. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, May 1996. 150

[102] Saleem A. Kassam. Optimum quantization for signal detection. *IEEE Transactions on Communications*, COM-25(5):479–484, May 1977. 118

[103] Saleem A. Kassam. Quantization based on the mean-absolute-error criterion. *IEEE Transactions on Communications*, COM-26(2):267–270, February 1978. 50, 125

[104] Richard E. Kihlstrom. The use of approximate prior distributions in a Bayesian decision model. *Econometrica*, 39(6):899–910, November 1971. 118

[105] Junmo Kim. *Nonparametric Statistical Methods for Image Segmentation and Shape Analysis*. PhD thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, February 2005. 41

[106] Andrey N. Kolmogorov and Vladimir M. Tihomirov. ϵ-entropy and ϵ-capacity of sets in functional spaces. *American Mathematical Society Translations Series 2*, 17:277–364, 1961. 21, 22, 72, 74, 75, 103, 146, 147

[107] Vladimir Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, July 2001. 34, 72

[108] Sanjeev R. Kulkarni. On metric entropy, Vapnik–Chervonenkis dimension, and learnability for a class of distributions. Technical Report P-1910, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts, September 1989. 74

[109] Simon Lacoste-Julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In Daphne Koller, Dale Schuurmans, Yoshua Bengio, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 21*. MIT Press, Cambridge, Massachusetts, 2009. 88

[110] Svetlana Lazebnik and Maxim Raginsky. Supervised learning of quantizer codebooks by information loss minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(7):1294–1309, July 2009. 88

[111] Guy Lebanon. *Riemannian Geometry and Statistical Machine Learning*. PhD thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania, January 2005. 20

[112] Jeffrey Sean Lehman. *An Optimistic Heart: What Great Universities Give Their Students ... and the World*. Cornell University, Ithaca, New York, 2008. 145

[113] Jia Li, Navin Chaddha, and Robert M. Gray. Asymptotic performance of vector quantizers with a perceptual distortion measure. *IEEE Transactions on Information Theory*, 45(4):1082–1091, May 1999. 23, 126, 128, 129, 148

[114] Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, June 1991. 86

[115] Ker-Chau Li. On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association*, 87(420):1025–1039, December 1992. 86

[116] Yi Lin. A note on margin-based loss functions in classification. *Statistics & Probability Letters*, 68(1):73–82, June 2004. 21, 22, 39, 74, 78, 79, 107, 146

[117] Xiuwen Liu, Anuj Srivastava, and Kyle Gallivan. Optimal linear representations of images for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):662–666, May 2004. 22, 87

[118] Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, IT-28(2):129–137, March 1982. 20

[119] Rohit Lotlikar and Ravi Kothari. Adaptive linear dimensionality reduction for classification. *Pattern Recognition*, 33(2):185–194, February 2000. 86, 147

[120] Shelly J. Lundberg and Richard Startz. Private discrimination and social intervention in competitive labor market. *The American Economic Review*, 73(3): 340–347, June 1983. 138

[121] Ulrike von Luxburg and Olivier Bousquet. Distance-based classification with Lipschitz functions. *Journal of Machine Learning Research*, 5:669–695, June 2004. 21, 57, 74, 106

[122] C. Neil Macrae and Galen V. Bodenhausen. Social cognition: Thinking categorically about others. *Annual Review of Psychology*, 51:93–120, February 2000. 24, 138

[123] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Discriminative learned dictionaries for local image analysis. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Anchorage, Alaska, June 2008. 88

[124] Jonathan H. Manton. Optimization algorithms exploiting unitary constraints. *IEEE Transactions on Signal Processing*, 50(3):635–650, March 2002. 46

[125] Joel Max. Quantizing for minimum distortion. *IRE Transactions on Information Theory*, 6(1):7–12, March 1960. 20

[126] Christian A. Meissner and John C. Brigham. Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, 7(1):3–35, January 2001. 140

[127] Diana L. Miglioretti, Rebecca Smith-Bindman, Linn Abraham, R. James Brenner, Patricia A. Carney, Erin J. Aiello Bowles, Diana S. M. Buist, and Joann G. Elmore. Radiologist characteristics associated with interpretive performance of diagnostic mammography. *Journal of the National Cancer Institute*, 99(24):1854–1863, December 2007. 17

[128] Sebastian Mika, Gunnar Rätsch, Jason Weston, Bernhard Schölkopf, and Klaus-Robert Müller. Fisher discriminant analysis with kernels. In *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing*, pages 41–48, Madison, Wisconsin, August 1999. 85, 86, 97, 147

[129] George A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63:81–97, 1956. 23, 24, 117, 137

[130] Tom M. Mitchell. The need for biases in learning generalizations. Technical Report CBM-TR-117, Department of Computer Science, Rutgers University, New Brunswick, New Jersey, May 1980. 18

[131] Sofia Mosci, Lorenzo Rosasco, and Alessandro Verri. Dimensionality reduction and generalization. In Zoubin Ghahramani, editor, *Proceedings of the 24th International Conference on Machine Learning*, pages 657–664, Corvallis, Oregon, June 2007. 85, 105

[132] Sendhil Mullainathan. A memory-based model of bounded rationality. *The Quarterly Journal of Economics*, 117(3):735–774, August 2002. 137

[133] Sendhil Mullainathan. Thinking through categories. April 2002. 138, 139

[134] Sendhil Mullainathan, Joshua Schwartzstein, and Andrei Shleifer. Coarse thinking and persuasion. *The Quarterly Journal of Economics*, 123(2):577–619, May 2008. 137

[135] Zoran Nenadic. Information discriminant analysis: Feature extraction with an information-theoretic objective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(8):1394–1407, August 2007. 86

[136] John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, New Jersey, 1944. 17

[137] Andrew Y. Ng. Feature selection, L_1 vs. L_2 regularization, and rotational invariance. In Russell Greiner and Dale Schuurmans, editors, *Proceedings of the 21st International Conference on Machine Learning*, pages 79–90, Banff, Canada, July 2004. 79, 80, 81

[138] XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Nonparametric decentralized detection using kernel methods. *IEEE Transactions on Signal Processing*, 53(11):4053–4066, November 2005. 107, 108, 116

[139] XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. On surrogate loss functions and f-divergences. *Annals of Statistics*, 37(2):876–904, April 2009. 87

[140] Yasunori Nishimori and Shotaro Akaho. Learning algorithms utilizing quasi-geodesic flows on the Stiefel manifold. *Neurocomputing*, 67:106–135, August 2005. 46

[141] Stanley Osher and Ronald Fedkiw. *Level Set Methods and Dynamic Implicit Surfaces*. Springer-Verlag, New York, 2003. 18, 19, 20, 27, 41, 42, 146

[142] Stanley Osher and Nikos Paragios, editors. *Geometric Level Set Methods in Imaging, Vision, and Graphics*. Springer-Verlag, New York, 2003. 20, 27, 41, 154

[143] Stanley Osher and James A. Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton–Jacobi formulations. *Journal of Computational Physics*, 79(1):12–49, November 1988. 19, 27, 55, 66

[144] Nikos Paragios and Rachid Deriche. Geodesic active regions and level set methods for supervised texture segmentation. *International Journal of Computer Vision*, 46(3):223–247, February 2002. 63

[145] Nikos Paragios and Rachid Deriche. Geodesic active regions and level set methods for motion estimation and tracking. *Computer Vision and Image Understanding*, 97(3):259–282, March 2005. 150

[146] Nikos K. Paragios. *Geodesic Active Regions and Level Set Methods: Contributions and Applications in Artificial Vision*. PhD thesis, Université de Nice Sophia Antipolis, France, January 2000. 41

[147] Edward A. Patrick and Frederic P. Fischer, II. Nonparametric feature selection. *IEEE Transactions on Information Theory*, IT-15(5):577–584, September 1969. 86

[148] Karl Pearson. On lines and planes of closest fit to systems of points in space. *London, Edinburgh and Dublin Philosophical Magazine and Journal of Science, Sixth Series*, 2:559–572, 1901. 18, 27, 45, 85

[149] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, August 2005. 147

[150] Francisco Pereira and Geoffrey Gordon. The support vector decomposition machine. In William W. Cohen and Andrew Moore, editors, *Proceedings of the 23rd International Conference on Machine Learning*, pages 689–696, Pittsburgh, Pennsylvania, June 2006. 22, 87, 147

[151] Duc-Son Pham and Svetha Venkatesh. Robust learning of discriminative projection for multicategory classification on the Stiefel manifold. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Anchorage, Alaska, June 2008. 22, 87, 147

[152] Edmund S. Phelps. The statistical theory of racism and sexism. *The American Economic Review*, 62(4):659–661, September 1972. 138

[153] John C. Platt. Probabilities for SV machines. In Alexander J. Smola, Peter L. Bartlett, Bernhard Schölkopf, and Dale Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, Cambridge, Massachusetts, 2000. 152

[154] Georg Pölzlbauer, Thomas Lidy, and Andreas Rauber. Decision manifolds—a supervised learning algorithm based on self-organization. *IEEE Transactions on Neural Networks*, 19(9):1518–1530, September 2008. 57

[155] H. Vincent Poor and John B. Thomas. Applications of Ali–Silvey distance measures in the design of generalized quantizers for binary decision systems. *IEEE Transactions on Communications*, COM-25(9):893–900, September 1977. 118

[156] Erinija Pranckeviciene, Tin Kam Ho, and Ray Somorjai. Class separability in spaces reduced by feature selection. In *Proceedings of the 18th International Conference on Pattern Recognition*, pages 254–257, Hong Kong, 2006. 150

[157] Joel B. Predd, Sanjeev R. Kulkarni, and H. Vincent Poor. Consistency in models for distributed learning under communication constraints. *IEEE Transactions on Information Theory*, 52(1):52–63, January 2006. 107, 108

[158] Joel B. Predd, Sanjeev R. Kulkarni, and H. Vincent Poor. Distributed learning in wireless sensor networks. *IEEE Signal Processing Magazine*, 23(4):56–69, July 2006. 107, 108, 116

[159] Joseph Price and Justin Wolfers. Racial discrimination among NBA referees. Working Paper 13206, National Bureau of Economic Research, June 2007. 17, 24, 137, 141, 143, 149

[160] Jose C. Principe, Dongxin Xu, and John W. Fisher, III. Information-theoretic learning. In Simon Haykin, editor, *Unsupervised Adaptive Filtering*, volume 1, pages 265–320. John Wiley & Sons, New York, 2000. 86

[161] Roy Radner. Costly and bounded rationality in individual and team decision-making. In Giovanni Dosi, David J. Teece, and Josef Chytry, editors, *Understanding Industrial and Corporate Change*, pages 3–35. Oxford University Press, New York, 2005. 137

[162] Frank P. Ramsey. Truth and probability. In Richard B. Braithwaite, editor, *The Foundation of Mathematics and Other Logical Essays*, pages 156–198. Routledge & Kegan Paul, London, 1931. 17

[163] Ryan Rifkin and Aldebaro Klautau. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5:101–141, January 2004. 20, 63

[164] Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958. 55

[165] Mikael Rousson and Nikos Paragios. Prior knowledge, level set representations & visual grouping. *International Journal of Computer Vision*, 76(3):231–243, 2008. 89

[166] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, December 2000. 47, 86, 97

[167] Olivier Roy and Martin Vetterli. Dimensionality reduction for distributed estimation in the infinite dimensional regime. *IEEE Transactions on Information Theory*, 54(4):1655–1669, April 2008. 108

[168] Ariel Rubinstein. *Modeling Bounded Rationality*. MIT Press, Cambridge, Massachusetts, 1998. 18, 137

[169] Michael J. Sabin and Robert M. Gray. Global convergence and empirical consistency of the generalized Lloyd algorithm. *IEEE Transactions on Information Theory*, IT-32(2):148–155, March 1986. 53, 134, 151

[170] Sajama and Alon Orlitsky. Supervised dimensionality reduction using mixture models. In Luc De Raedt and Stefan Wrobel, editors, *Proceedings of the 22nd International Conference on Machine Learning*, pages 768–775, Bonn, Germany, August 2005. 86, 147

[171] Christophe Samson, Laure Blanc-Féraud, Gilles Aubert, and Josiane Zerubia. A level set model for image classification. *International Journal of Computer Vision*, 40(3):187–197, December 2000. 63

[172] Sujay R. Sanghavi, Vincent Y. F. Tan, and Alan S. Willsky. Learning graphical models for hypothesis testing. In *Proceedings of the 14th IEEE/SP Workshop on Statistical Signal Processing*, pages 69–73, Madison, Wisconsin, August 2007. 88

[173] Guillermo Sapiro. *Geometric Partial Differential Equations and Image Analysis*. Cambridge University Press, Cambridge, United Kingdom, 2001. 20, 154

[174] Leonard J. Savage. *The Foundations of Statistics*. Wiley, New York, 1954. 17

[175] Ioannis D. Schizas, Georgios B. Giannakis, and Zhi-Quan Luo. Distributed estimation using reduced-dimensionality sensor observations. *IEEE Transactions on Signal Processing*, 55(8):4284–4299, August 2007. 108

[176] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, Massachusetts, 2002. 19, 35, 37, 55, 57, 85, 97, 146

[177] Bernhard Schölkopf, Alexander J. Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5): 1299–1319, July 1998. 18, 85, 97

[178] Gideon Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2): 461–464, March 1978. 149

[179] Clayton Scott and Robert D. Nowak. A Neyman–Pearson approach to statistical learning. *IEEE Transactions on Information Theory*, 51(11):3806–3819, November 2005. 151

[180] Clayton Scott and Robert D. Nowak. Minimax-optimal classification with dyadic decision trees. *IEEE Transactions on Information Theory*, 52(4):1335–1353, April 2006. 57, 69

[181] Xiaotong Shen and Wing Hung Wong. Convergence rate of sieve estimates. *Annals of Statistics*, 22(2):580–615, June 1994. 79

[182] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, August 2000. 56

[183] Yair Shoham and Allen Gersho. Efficient bit allocation for an arbitrary set of quantizers. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36 (9):1445–1453, September 1988. 139

[184] Vin de Silva and Joshua B. Tenenbaum. Global versus local methods in nonlinear dimensionality reduction. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 705–712. MIT Press, Cambridge, Massachusetts, 2003. 153

[185] Vikas Sindhwani, Mikhail Belkin, and Partha Niyogi. The geometric basis of semi-supervised learning. In Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors, *Semi-Supervised Learning*, pages 209–226. MIT Press, Cambridge, Massachusetts, 2006. 20

[186] Gregory G. Slabaugh, H. Quynh Dinh, and Gözde B. Unal. A variational approach to the evolution of radial basis functions for image segmentation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Minneapolis, Minnesota, June 2007. 66, 67

[187] Michael Spence. Job market signaling. *The Quarterly Journal of Economics*, 87 (3):355–374, August 1973. 138

[188] Anuj Srivastava and Xiuwen Liu. Tools for application-driven linear dimension reduction. *Neurocomputing*, 67:136–160, August 2005. 18, 20, 22, 45, 85, 86, 87, 147

[189] Ingo Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, 51(1):128–142, January 2005. 74, 78

[190] Michael A. Stoll, Steven Raphael, and Harry J. Holzer. Black job applicants and the hiring officer's race. *Industrial and Labor Relations Review*, 57(2):267–287, January 2004. 24, 137, 141, 143, 149

[191] Walter Sun, Müjdat Çetin, Raymond Chan, and Alan S. Willsky. Learning the dynamics and time-recursive boundary estimation of deformable objects. *IEEE Transactions on Image Processing*, 17(11):2186–2200, November 2008. 150

[192] Mark Sussman, Peter Smereka, and Stanley Osher. A level set approach for computing solutions to incompressible two-phase flow. *Journal of Computational Physics*, 114(1):146–159, September 1994. 44

[193] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, December 2000. 20, 22, 47, 48, 85, 97, 100

[194] Robert R. Tenney and Nils R. Sandell, Jr. Detection with distributed sensors. *IEEE Transactions on Aerospace and Electronic Systems*, AES-17(4):501–510, July 1981. 22, 107

[195] Madan Thangavelu and Raviv Raich. Multiclass linear dimension reduction via a generalized Chernoff bound. In *Proceedings of the IEEE Workshop on Machine Learning for Signal Processing*, pages 350–355, Cancún, Mexico, October 2008. 86, 147

[196] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. 80

[197] Arkadiusz Tomczyk and Piotr S. Szczepaniak. On the relationship between active contours and contextual classification. In Marek Kurzyński, Edward Puchała, Michał Woźniak, and Andrzej Żołnierek, editors, *Proceedings of the 4th International Conference on Computer Recognition Systems*, pages 303–310, Rydzyna, Poland, 2005. 56

[198] Arkadiusz Tomczyk, Piotr S. Szczepaniak, and Michal Pryczek. Active contours as knowledge discovery methods. In Vincent Corruble, Masayuki Takeda, and Einoshin Suzuki, editors, *Proceedings of the 10th International Conference on Discovery Science*, pages 209–218, Sendai, Japan, 2007. 57

[199] Kari Torkkola. Feature extraction by non-parametric mutual information maximization. *Journal of Machine Learning Research*, 3:1415–1438, March 2003. 86

[200] Alexander V. Trushkin. Sufficient conditions for uniqueness of a locally optimal quantizer for a class of convex error weighting functions. *IEEE Transactions on Information Theory*, IT-28(2):187–198, March 1982. 23, 51, 52, 123, 148

[201] Andy Tsai. *Curve Evolution and Estimation-Theoretic Techniques for Image Processing*. PhD thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, September 2001. 41

[202] Ivor Wai-Hung Tsang, András Kocsor, and James Tin-Yau Kwok. Large-scale maximum margin discriminant analysis using core vector machines. *IEEE Transactions on Neural Networks*, 19(4):610–624, April 2008. 22, 87, 147

[203] John N. Tsitsiklis. Decentralized detection. Technical Report P-1913, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts, September 1989. 22, 107

[204] Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, United Kingdom, 1998. 79

[205] Vladimir Vapnik, Esther Levin, and Yann Le Cun. Measuring the VC-dimension of a learning machine. *Neural Computation*, 6(5):851–876, September 1994. 21, 72, 73, 146

[206] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 2000. 17, 19, 27, 34, 35, 55, 72, 85

[207] Kush R. Varshney and Lav R. Varshney. Minimum mean Bayes risk error quantization of prior probabilities. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3445–3448, Las Vegas, Nevada, April 2008. 25, 117

[208] Kush R. Varshney and Lav R. Varshney. Quantization of prior probabilities for hypothesis testing. *IEEE Transactions on Signal Processing*, 56(10):4553–4562, October 2008. 25, 117

[209] Kush R. Varshney and Alan S. Willsky. Supervised learning of classifiers via level set segmentation. In *Proceedings of the IEEE Workshop on Machine Learning for Signal Processing*, pages 115–120, Cancún, Mexico, October 2008. 25, 55

[210] Kush R. Varshney and Alan S. Willsky. Learning dimensionality-reduced classifiers for information fusion. In *Proceedings of the 12th International Conference on Information Fusion*, pages 1881–1888, Seattle, Washington, July 2009. 25, 85, 108

[211] Kush R. Varshney and Alan S. Willsky. Classification using geometric level sets. *Journal of Machine Learning Research*, 11:491–516, February 2010. 25, 55

[212] Kush R. Varshney and Alan S. Willsky. Linear dimensionality reduction for margin-based classification: High-dimensional data and sensor networks. *IEEE Transactions on Signal Processing*, submitted. 25, 85, 108

[213] Pramod K. Varshney. *Distributed Detection and Data Fusion*. Springer-Verlag, New York, 1996. 22, 107

[214] Luminita A. Vese and Tony F. Chan. A multiphase level set framework for image segmentation using the Mumford and Shah model. *International Journal of Computer Vision*, 50(3):271–293, December 2002. 20, 62, 63

[215] Paul Weirich. *Realistic Decision Theory*. Oxford University Press, New York, 2004. 17, 137

[216] Holger Wendland. *Scattered Data Approximation*. Cambridge University Press, Cambridge, United Kingdom, 2005. 21, 66

[217] Bernard Widrow and Marcian E. Hoff. Adaptive switching circuits. In *IRE Western Electronic Show and Convention Record*, volume 4, pages 96–104, Los Angeles, August 1960. 150

[218] Robert A. Wijsman. Continuity of the Bayes risk. *Annals of Mathematical Statistics*, 41(3):1083–1085, June 1970. 29, 119

[219] Rebecca Willett and Robert D. Nowak. Minimax optimal level-set estimation. *IEEE Transactions on Image Processing*, 16(12):2965–2979, December 2007. 57

[220] Robert C. Williamson, Alexander J. Smola, and Bernhard Schölkopf. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. *IEEE Transactions on Information Theory*, 47(6):2516–2532, September 2001. 74

[221] Alan S. Willsky, Gregory W. Wornell, and Jeffrey H. Shapiro. *Stochastic Processes, Detection and Estimation 6.432 Course Notes*. Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, Fall 2003. 18, 23, 27, 29, 30, 32, 119, 148, 152

[222] David H. Wolpert. The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7):1341–1390, October 1996. 146

[223] Andy M. Yip, Chris Ding, and Tony F. Chan. Dynamic cluster formation using level set methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(6):877–889, June 2006. 56

[224] Dengyong Zhou and Bernhard Schölkopf. Regularization on discrete spaces. In Walter G. Kropatsch, Robert Sablatnig, and Allan Hanbury, editors, *Proceedings of the 27th DAGM Symposium on Pattern Recognition*, pages 361–368, Vienna, Austria, August–September 2005. 154

[225] Song Chun Zhu and Alan Yuille. Region competition: Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(9):884–900, September 1996. 41

[226] Zhigang Zhu and Thomas S. Huang, editors. *Multimodal Surveillance: Sensors, Algorithms, and Systems*. Artech House, Boston, 2007. 107, 148

[227] Lei Zong, Jeff Houser, and T. Raju Damarla. Multi-modal unattended ground sensor (MMUGS). In *Proceedings of SPIE*, volume 6231, page 623118, April 2006. 107, 148

[228] Hui Zou, Ji Zhu, and Trevor Hastie. The margin vector, admissible loss and multiclass margin-based classifiers. Technical report, School of Statistics, University of Minnesota, Minneapolis, Minnesota, 2006. 63, 64

[229] Hui Zou, Ji Zhu, and Trevor Hastie. New multicategory boosting algorithms based on multicategory Fisher-consistent losses. *Annals of Applied Statistics*, 2 (4):1290–1306, December 2008. 63

[230] Laurent Zwald, Régis Vert, Gilles Blanchard, and Pascal Massart. Kernel projection machine: A new tool for pattern recognition. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1649–1656. MIT Press, Cambridge, Massachusetts, 2005. 85, 105