Learning in decentralized systems: A nonparametric approach

by

XuanLong Nguyen

B.S. (Pohang University of Science and Technology) 1999
M.S. (Arizona State University) 2001
M.A. (University of California, Berkeley) 2007

A dissertation submitted in partial satisfaction of the
requirements for the degree of

Doctor of Philosophy

in

Computer Science
and the Designated Emphasis
in
Communication, Computation, and Statistics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Michael I. Jordan, Chair
Professor Martin J. Wainwright
Professor Peter L. Bartlett
Professor Peter J. Bickel

Fall 2007

The dissertation of XuanLong Nguyen is approved:

Professor Michael I. Jordan, Chair Date

Professor Martin J. Wainwright Date

Professor Peter L. Bartlett Date

Professor Peter J. Bickel Date

University of California, Berkeley

Fall 2007

Learning in decentralized systems: A nonparametric approach

by

XuanLong Nguyen

Abstract

Learning in decentralized systems: A nonparametric approach

by

XuanLong Nguyen

Doctor of Philosophy in Computer Science

and the Designated Emphasis

in

Communication, Computation, and Statistics

University of California, Berkeley

Professor Michael I. Jordan, Chair

Rapid advances in information technology result in increased deployment of decentralized decision-making systems embedded within large-scale infrastructures consisting of data collection and processing devices. In such a system, each statistical decision is performed on the basis of limited amount of data due to constraints given by the decentralized system. For instance, the constraints may be imposed by limits in energy source, communication bandwidth, computation or time budget. A fundamental problem arised in decentralized systems involves the development of methods that take into account not only the statistical accuracy of decision-making procedures, but also the constraints imposed by the system limits. It is this general problem that drives the focus of this thesis. In particular, we focus on the development and analysis of statistical learning methods for decentralized decision-making by employing a nonparametric approach. The nonparametric approach imposes very little a priori assumption on the data; such flexibility allows it to be applicable to a wide range of applications. Coupled with tools from convex analysis and empirical process theory we develop computationally efficient algorithms and analyze their statistical behavior both theoretically and empirically.

Our specific contributions include the following. We develop a novel kernel-based algorithm for centralized detection and estimation in the ad hoc sensor networks through the challenging task of sensor mote localization. Next, we develop and analyze a nonparametric decentralized detection algorithm using the methodology of convex surrogate loss functions and marginalized kernels. The analysis of this algorithm leads to an in-depth study of the correspondence between the class of surrogate loss functions widely used in statistical machine learning and the class of divergence functionals widely used in information theory. This correspondence allows us to provide an interesting decision-theoretic justification to a given choice of divergence functionals, which often arise from asymptotic analysis. In addition, this correspondence also motivates the development and the analysis of a novel M-estimation procedure for estimating divergence functionals and the likelihood ratio. Finally, we also investigate a sequential setting of the decentralized detection algorithm, and settle an open question regarding the characterization of optimal decision rules in such a setting.

Professor Michael I. Jordan, Chair Date

Acknowledgements

A journey is closing to an end, and a new one is just about to begin. During the course of my graduate study I have benefited from interactions with many people. It is a pleasure to acknowledge all of them on this page.

At Berkeley I have the fortune to have enjoyed the interaction and support of two research advisors, Michael Jordan and Martin Wainwright. I have learned a great deal from Mike, be it machine learning or statistics, but am most impressed with and influenced by his relentless enthusiasm for expanding and deepening one's boundary of knowledge, and his inspiring vision for statistical research. I am grateful for the freedom he has given me in pursuing my work. Still, as I've moved from one topic to another, always constant are Mike's warm encouragement, advices and very helpful pointers. I am astonished to find from time to time that the tools I've learned from his courses and group meetings always find themselves handy at some point in my research paths.

Martin Wainwright has instilled in me an early interest in signal processing and information theory. I will not forget our many hours in offices and cafés: Martin's sharpness and humor make working with him very enjoyable, and keep me even more motivated by the end of the sessions. I am also greatly influenced by Martin's approach to research, as well as many other aspects such as presentation and teaching.

I am grateful to Professor Subbarao Kambhampati, who took me on-board and taught me the first research skills during my two years at ASU. Rao's infectious enthusiasm for research and teaching will remain with me for a long time to come. I thank Professors Peter Bartlett and Peter Bickel, both of whom are members of my thesis committee, and Professor Bin Yu for their fantastic courses in statistics, and for their insightful advices and useful pointers for my work.

I have the fortune to have interacted with a number of people in the EECS and Statistics departments, as well as colleagues in the wider academic community. Ling Huang, Ram Rajagopal and Bruno Sinopoli have actively worked with me on a number of projects on detection and sequential analysis in systems and sensor networks. Some of these work serve either as a starting point or an outgrowth of the results presented in this thesis. I have benefited from interactions with Barbara Engelhardt, Eyal Amir, Tijl De Bie, Hung H. Bui, Dave Blei, Francis Bach, Ben Blum, Simon Lacoste-Julien, Percy Liang, Gert Lanckriet, Greg Lawrence, Brian Milch, Bhaskara Marthi, Kurt Miller, Andrew Ng, Guillaume Obozinski, Songhwai Oh, T.K. Satish Kumar, Tye Rattenbury, Mathias Seeger, Fei Sha, Erik Sudderth, Yee-Whye Teh, Ambuj Tewari, Romain Thibaux, Eric Xing and Alice Zheng. Thank you to La Shana Porlaris, who is always cheerful and helpful in getting me through Berkeley's paperworks seamlessly.

My time at Berkeley was enriched by a circle of good friends in and out of the campus. I am thankful for their camaraderie and friendships.

Finally, this thesis is dedicated to the memory of my late mother, and to my family in Vietnam, for the boundless love, encouragement and belief in me. To my child Thaian, who will be happy to learn that I managed to finish the writing of this thesis a day before she was born. To my wife Lan Nguyen; I can't be happier for having her on this journey and on the road ahead.

Dedicated to my family

Contents

Chapter 1

Introduction

Research in the area of decentralized systems focuses on problems in which measurements are collected by a collection of devices distributed over space (e.g., arrays of cameras, acoustic sensors, wireless nodes). Due to power and bandwidth limitations, these devices cannot simply relay their measurements to the common site where a centralized decision is to be made; rather, the measurements must be compressed prior to transmission, and the statistical decision-making at the central site is performed on the transformed data [Tsitsiklis, 1993b; Blum *et al.*, 1997]. A fundamental problem in decentralized systems research is that of designing local decision rules at individual data collection/transmission devices and global decision rules at some fusion center(s) so as to optimize an objective function of interest.

The problems of decentralized decision making have been the focus of considerable research in the past three decades (see, e.g., [Tenney and Sandell, 1981; Tsitsiklis, 1993b; Blum *et al.*, 1997; Viswanathan and Varshney, 1997; Chong and Kumar, 2003; Chamberland and Veeravalli, 2003; Chen *et al.*, 2006]). Indeed, decentralized systems arise in a variety of important applications, ranging from sensor networks, in which each sensor operates under severe power or bandwidth constraints, to the modeling of human decision-making, in which high-level executive decisions are frequently based on lower-level summaries. In applied databases and computer systems, there has been growing interest in building large-scale distributed monitoring systems of sensor, enterprise and ISP networks. In such settings, data are typically collected by distributed monitoring devices, transmitted through the network to a central network operation center for aggregation and analysis (see, e.g., [Cormode and Garofalakis, 2005; Olston *et al.*, 2003; Padmanabhan *et al.*, 2005; Xie *et al.*, 2004; Yegneswaran *et al.*, 2004; Huang *et al.*, 2007]).

From a broad statistical perspective, the variety of learning and decision-making problems in decentralized systems can be viewed as interesting and highly nontrivial extensions of basic statistical analysis tasks that involve aspects of experiment design, where each particular decentralized system impose a different type of constraints on the experiment design space. There is a vast statistical literature on experiment designs going back to David Blackwell and others [Blackwell, 1951; Blackwell, 1953; Bradt and Karlin, 1956; Lindley, 1956; Goel and DeGroot, 1979; Chernoff, 1972; Steinberg and Hunter, 1985; Ford *et al.*, 1989; Pukelsheim, 1993; Chaloner and Verdinelli, 1995]. When applied to decentralized systems, an experiment design is translated to a variety of types of decision rules: For example, for a sensor network, it is a collection of compression rules at each individual sensors [Tsitsiklis, 1993b]. For a distributed enterprise network, it is the so-called filtering scheme at monitoring devices [Olston *et al.*, 2003]. It is worth noting that removing the trapping of "decentralized systems" terminologies, the problems of learning and decision making in such settings share much in common with many fundamental problems in modern data analysis, such as dimensionality reduction, feature selection, independent component analysis, because the latter can also be viewed as instances of the problem of experiment design.

Despite having strong roots in the classical statistics literature, problems of decentralized decision making exhibit unique challenges that typically render a large portion of existing methods inapplicable. From a computational viewpoint, the high dimensionality of data (e.g., large number of sensors in a large-scale monitoring infrastructure) and a variety of decentralization constraints imposed on the way such devices can communicate result in an exponentially large space for possible designs. Indeed, with the exception of special cases, it is known that the problem of computing decentralized decision rules is NP-hard [Tsitsiklis and Athans, 1985], even under assumption that the underlying distribution generating the data is completely known. From a statistical viewpoint, however, the assumption that the underlying distribution generating the data is known is rather unrealistic in real applications such as sensor networks or large-scale distributed systems. This necessitates the need to develop methods that impose minimal assumptions on the practitioner's statistical knowledge of the data. Instead of computing optimal decentralized decision rules given the knowledge of relevant probability distributions, one could aim to estimate optimal decentralized decision rules directly on the basis of the empirical data samples which are more readily available. Therefore, this thesis is motivated by the need of developing a *nonparametric* approach to decentralized decision-making, where the optimal decision rules have to be estimated from empirical data. The nonparametric approach studies learning procedures that aim to capture large, open-ended classes of functions of interest for our decision-making purposes.

This chapter is devoted to an overview of a nonparametric approach to a number of decision-making problems arised from decentralized systems. We shall start by a more detailed review of existing approaches to decentralized decision-making in Section 1.1. In Section 1.2 we summarize briefly key ingredients of our nonparametric approach. Section 1.3 discusses the main problems and contributions of the thesis in some detail.

1.1 Existing parametric frameworks and methods

1.1.1 Decentralized decision making

To be more concrete, let us state a basic problem of decentralized detection in the language of discriminant analysis augmented with a component of experiment design. In particular, throughout this thesis our focus will be that of *binary* discriminants. Let X be a covariate taking values in space $\mathcal{X}$, and let $Y \in \mathcal{Y} := \{-1, +1\}$ be a binary random variable. The joint vector (X, Y) is drawn from some probability distribution $\mathbb{P}$. A *discriminant function* is a measurable function f mapping from $\mathcal{X}$ to the real line, whose sign is used to make a detection/classification decision. The standard goal of discriminant analysis is to choose the discriminant function f so as to minimize the probability of making the incorrect detection, also known as the *Bayes risk*, $\mathbb{P}(Y \neq \text{sign}(f(X)))$.

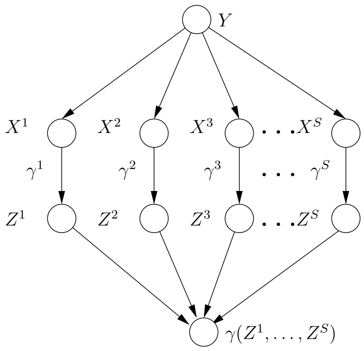


Figure 1.1. Decentralized detection system with S sensors, in which Y is the unknown hypothesis, $X = (X^1, \ldots, X^S)$ is the vector of sensor observations; and $Z = (Z^1, \ldots, Z^S)$ are the quantized messages transmitted from sensors to the fusion center.

An elaboration of this basic problem in which the decision-maker, rather than having direct access to X, observes a random variable variable $Z \in \mathcal{Z}$ that is obtained via a (possibly stochastic) mapping $Q : \mathcal{X} \to \mathcal{Z}$. In a statistical context, the choice of the mapping Q can be viewed as choosing a particular *experiment*; in the signal processing literature, where $\mathcal{Z}$ is generally taken to be discrete, the mapping Q is often referred to as a *quantizer*. A decentralized system naturally imposes constraints on the class of quantizer Q that need to be taken into account in the decision making process.

When the underlying experiment Q is fixed, then we simply have a centralized binary classification problem on the space $\mathcal{Z}$: that is, our goal is to find a real-valued discriminant function γ on $\mathcal{Z}$ so as to minimize the Bayes risk $\mathbb{P}(Y \neq \text{sign}\gamma(Z))$. On the other hand, the basic issue in decentralized detection is the problem of determining *jointly* the classifier $\gamma \in \Gamma$, as well as the experiment choice $Q \in \mathcal{Q}$ in the following decision-making scheme:

$$X \xrightarrow{Q} Z \xrightarrow{\gamma} Y$$

The problem of designing such compression rules is of substantial current interest in the field of sensor networks [Chong and Kumar, 2003; Chamberland and Veeravalli, 2003]. There has also significant amount of work devoted to criteria other than the Bayes error, such as criteria based on Neyman-Pearson or minimax formulations [Tsitsiklis, 1993b]. A closely related set of "signal selection" problems, arising for instance in radar array processing, also blend discriminant analysis with aspects of experimental design [Kailath, 1967].

It is well-known that the optimal decision rule (Q, γ) has to be a threshold rule on some likelihood ratios [Tsitsiklis, 1993b]. On the algorithmic front, the large majority of the literature is based on the assumption that the probability distributions $\mathbb{P}(X|Y)$ lie within some known parametric family (e.g., Gaussian and conditional independent), and seeks to characterize the optimal decision rules under such assumptions. Despite such rather strong assumptions, the standard formulation rarely leads to computationally tractable algorithms. One main source of difficulty is the intractability of minimizing the probability of error, whether as a functional of the discriminant function or of the compression rule. Consequently, it is natural to consider loss functions that act as surrogates for the probability of error, and lead to practical algorithms. For example, the Hellinger distance has been championed for decentralized detection problems [Longo *et al.*, 1990; Kailath, 1967], due to the fact that it yields a tractable algorithm both for the experimental design aspect of the problem (i.e., the choice of compression rule) and the discriminant analysis aspect of the problem. Chernoff's distance was used as a surrogate loss in conjunction with Gaussian and conditional independence assumptions on $\mathbb{P}(X, Y)$ [Chamberland and Veeravalli, 2003]. More broadly, a class of functions known as *Ali-Silvey distances* or *f-divergences* [Ali and Silvey, 1966; Csiszaŕ, 1967]— which includes not only the Hellinger distance, but also the variational distance, Kullback-Leibler (KL) divergence and Chernoff distance—have been explored as surrogate loss functions for the probability of error in a wide variety of applied discrimination problems. An f-divergence $D_\phi(\mathbb{P}, \mathbb{Q})$ captures a kind of "distance" between two distributions $\mathbb{P}$ and $\mathbb{Q}$, and has the following form:

$$D_\phi(\mathbb{P}, \mathbb{Q}) = \int p_0 \phi(q_0/p_0)\, d\mu,$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is a convex function.

Theoretical support for the use of f-divergences in discrimination problems comes from two main sources. First, a classical result of [Blackwell, 1951] asserts that if procedure A has a smaller f-divergence than procedure B (for some particular f-divergence), then there exists some set of prior probabilities such that procedure A has a smaller probability of error than procedure B. This fact, though a relatively weak justification, has nonetheless proven useful in designing signal selection and quantization rules [Kailath, 1967; Poor and Thomas, 1977; Longo *et al.*, 1990]. Second, f-divergences often arise as exponents in asymptotic (large-deviation) characterizations of the optimal rate of convergence in hypothesis-testing problems; examples include Kullback-Leibler divergence for the Neyman-Pearson formulation, and the Chernoff distance for the Bayesian formulation [Cover and Thomas, 1991].

1.1.2 Decentralized detection in sequential setting

An interesting variant of the decentralized detection problem is its extension to an-online setting: more specifically, the *sequential decentralized detection* problem [Tsitsiklis, 1986; Veeravalli, 1999; Mei, 2003] involves a data sequence, $\{X_1, X_2, \ldots\}$, and a corresponding sequence of summary statistics, $\{U_1, U_2, \ldots\}$, determined by a sequence of local decision rules $\{Q_1, Q_2, \ldots\}$. The goal is to design both the local decision functions and to specify a global decision rule so as to predict H in a manner that optimally trades off accuracy and delay. In short, the sequential decentralized detection problem is the communication-constrained extension of classical formulation of sequential centralized decision-making problems (see, e.g., [Wald, 1947; Chernoff, 1972; Shiryayev, 1978; Siegmund, 1985; Lai, 2001]) to the decentralized setting.

The bulk of the literature so far is confined to setting up general framework for studying sequential decentralized detection and studying the structure of the optimal solutions. In setting up a general framework for studying sequential decentralized problems, Veeravalli et al. [Veeravalli *et al.*, 1993] defined five problems, denoted "Case A" through "Case E", distinguished from one another by the amount of information available to the local sensors. For example, in Case E, the local sensors are provided with memory and with feedback from the global decision-maker (also known as the *fusion center*), so that each sensor has available to it the current data, X_n, as well as all of the summary statistics from all of the other local sensors. In other words, each local sensor has the same snapshot of past state as the fusion center; this is an instance of a so-called "quasi-classical information structure" [Ho, 1980] for which dynamic programming (DP) characterizations of the optimal decision functions are available. Veeravalli et al. [Veeravalli *et al.*, 1993] exploit this fact to show that the decentralized case has much in common with the centralized case, in particular that likelihood ratio tests are optimal local decision functions at the sensors and that a variant of a sequential probability ratio test is optimal for the decision-maker.

Unfortunately, however, part of the spirit of the decentralized detection is arguably lost in Case E, which requires full feedback. In applications such as power-constrained sensor networks, we generally do not wish to assume that there are high-bandwidth feedback channels from the decision-maker to the sensors, nor do we wish to assume that the sensors have unbounded memory. Most suited to this perspective—and the focus of this thesis—is Case A, in which the local decisions are of the simplified form $U_n = Q_n(X_n)$; i.e., neither local memory nor feedback are assumed to be available.

Noting that Case A is not amenable to dynamic programming and is presumably intractable, Veeravalli et al. [Veeravalli *et al.*, 1993] suggested restricting the analysis to the class of *stationary* local decision functions; i.e., local decision functions Q_n that are independent of n. They conjectured that stationary decision functions may actually be optimal in the setting of Case A (given the intuitive symmetry and high degree of independence of the problem in this case), even though it is not possible to verify this optimality via DP arguments. This conjecture has remained open since it was first posed by Veeravalli et al. [Veeravalli *et al.*, 1993; Veeravalli, 1999].

In comparison to (non-sequential) decentralized detection, since little is known about the nature of optimal decision rules Q_n in the aforementioned setting of sequential decentralized detection, much less is known about an algorithmic solutions for such problems, even in a parametric setting.

1.2 Nonparametric framework

Despite enormous advances in the area of (parametric) decentralized decision making, strong parametric assumptions of data make existing methods inappropriate in a wide range of application domains. For example, in realistic monitoring infrastructure such as sensor networks, it is well-known that idealized parametric models can be highly inaccurate due to variability caused by multipath effects and ambient noise interference as well as device-specific factors such as the frequencies of node radios, physical antenna orientation, and fluctuations in the power source (e.g., see [Bulusu *et al.*, 2000; Priyantha *et al.*, 2000]). What is clearly needed is a flexible framework that requires only minimal assumtions on the data, and let the computation tasks of decision rules in decentralized systems be done through estimation/learning from empirical data, where the learning is performed under the constraints imposed by the decentralized systems. Nonparametric statistics [Wasserman, 2005] provide a suitable framework for this goal.

In the context of *centralized* signal detection problems, there is an extensive line of research on nonparametric techniques, in which no specific parametric form for the joint distribution $P(X,Y)$ is assumed (see, e.g., Kassam [Kassam, 1993] for a survey). In the decentralized setting, however, it is only relatively recently that nonparametric methods for detection have been explored. Several authors have taken classical nonparametric methods from the centralized setting, and shown how they can also be applied in a decentralized system. Such methods include schemes based on Wilcoxon signed-rank test statistic [Viswanathan and Ansari, 1989; Nasipuri and Tantaratana, 1997], as well as the sign detector and its extensions [Han *et al.*, 1990; Al-Ibrahim and Varshney, 1989; Hussaini *et al.*, 1995]. These methods have been shown to be quite effective for certain types of joint distributions.

The overarching theme in this thesis is the development of a general nonparametric framework for decision making in a decentralized systems. Restricting ourselves for a moment to the basic setup stated in Section 1.1.1 our framework can be succintly described as follows. Let X be a covariate taking values in space $\mathcal{X}$, and let $Y \in \mathcal{Y} := \{-1, +1\}$ be a binary random variable. The joint vector (X, Y) is drawn from some *unknown* probability distribution $\mathbb{P}$. Given classes $\mathcal{Q}$ and $\mathcal{F}$ of decision rules Q and γ, respectively, and that the knowledge of the distribution $\mathbb{P}(X, Y)$ is given through the basis of independent and identically distributed (i.i.d.) sample $((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n))$, our goal is to learn the discriminant function $\gamma \in \Gamma$ and decision rule $Q \in \mathcal{Q}$ so as to minimize the probability of making the incorrect detection, i.e., Bayes risk: $\mathbb{P}(Y \neq \text{sign}(\gamma(Q(X))))$.

There is a suite of important issues underlying our framework:

- What is the appropriate learning procedure for estimating the discriminant functions and the quantization rules?
- What are the representations of the discriminant functions and the quantization rules?
- How are the constraints imposed by decentralized systems taken into account?
- What optimization techniques can be employed to improve the computational efficiency of the algorithm?
- What are the statistical and computational properties of the algorithm?

Addressing these issues forms the core part of this thesis. Moreover, as we shall elaborate in Section 1.3, the answers to some of these issues are also of independent interest in contexts beyond the realm of decentralized systems.

1.2.1 Classification methods

At a very high level, our development is partly motivated by the recent advances in the statistical classification literature. By classification we refer to a class of problem of learning discriminant functions from empirical (training) data. The classification literature has enjoyed intense research in the past half century with contributions from a variety of disciplines, including statistics (see, e.g., [Bickel and Doksum, 2006; Hastie *et al.*, 2001]), engineering (e.g., [Duda *et al.*, 2000; Fukunaga, 1990]), artificial intelligence and machine learning (e.g., [Bishop, 1995; Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004]).

Early research on classification focused on learning linear discriminants underlying certain parametric models, resulting in classical methods such as linear discriminant analysis and linear logistic regression, which have become standard toolboxes in a wide variety of applied fields. At the same time, more flexible alternatives have been proposed to explicitly model the linear discrinant function in a direct manner. In particular, for the binary discrinant functions, the well-known perceptron algorithm, due to Rosenblatt, was proposed to find separating hyperplane in the empirical data [Rosenblatt, 1958]. Another key idea, due to Vapnik [Vapnik, 1998], was to finds an optimally separating hyperplane using some measure of loss. These new methods were particularly interesting because they paved ways for later classification methods that are flexible enough not to rely on strong assumptions on the underlying distribution generating the empirical data.

The second strand of progress focused on moving from linear classification to nonlinear classification. A significant development was the resurgence of neural networks, which allow for representing arbitrary nonlinear discriminant functions by compositions of simpler linear functions and threshold functions via the network's multiple layers, coupled with the well-known backpropagation learning algorithm [Rumelhart *et al.*, 1986; Werbos, 1974]. Another important development was the adoption of the kernel method in representing discriminant functions via the support vector machine algorithm [Cortes and Vapnik, 1995]. The AdaBoost algorithm [Freund and Schapire, 1997] introduced a novel way of constructing more complex discriminant functions out of simpler classification algorithms. These are examples of nonparametric classification algorithms that have enjoyed a significant level of popularity in the past decade.[1]

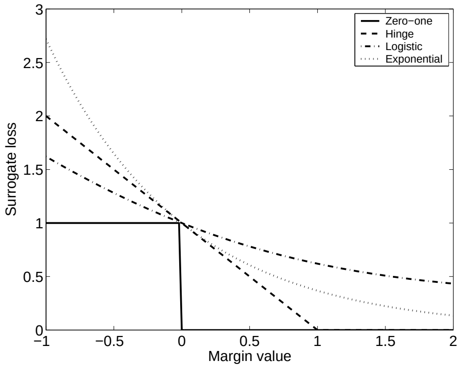


Figure 1.2. Illustrations of the 0-1 loss function, and three surrogate loss functions: hinge loss, logistic loss, and exponential loss.

The third strand of progress in the field of classification, through the work of many researchers, includes an improved understanding of the statistical and computational behavior of the proposed learning algorithms, the recognition of the important role of efficient computation via convex optimization (e.g., [Boyd and Vandenberghe, 2004; Bertsekas, 1995b]),

[1] Strictly speaking, the discriminant functions considered in the methods of [Cortes and Vapnik, 1995] and [Freund and Schapire, 1997] are linear in some function spaces.

and the (re)integration of the field with the existing statistics literature of nonparametric estimation (e.g., [Silverman, 1986; Wahba, 1990]). It is now well-understood that the vast arsenal of classification algorithms can be characterized in terms of two key components: (1) The use of computationally-motivated surrogate loss functions and (2) the choice of a function class representing the class of discriminant functions.

Indeed, in the decision-theoretic formulation of the classification problem, the Bayes error is interpreted as risk under 0-1 loss. The algorithmic goal is to design discriminant functions by minimizing the empirical expectation of 0-1 loss. In this setting, the non-convexity of the 0-1 loss renders intractable a direct minimization of the probability of error, so that a variety of algorithm can be viewed as replacing the 0-1 loss with "surrogate loss functions." These alternative loss functions are convex, and represent approximations to the 0-1 loss (see Figure 1.2 for an illustration). A wide variety of practically successful machine learning algorithms are based on such a strategy, including support vector machines [Cortes and Vapnik, 1995; Schölkopf and Smola, 2002], the AdaBoost algorithm [Freund and Schapire, 1997], the X4 method [Breiman, 1998], and logistic regression [Friedman *et al.*, 2000].

Although the use of kernel methods in classification problems is relatively recent, they has been studied extensively in the past three decades in the nonparametric statistics literature, mostly in the context of regression (i.e., function estimation) and density estimation. On the algorithmic side, kernel methods are almost synonymous to density and function estimation algorithms – see [Silverman, 1986] for an introduction. The use of reproducing kernel Hilbert space in general and smoothing splines in particular in both estimation tasks were pioneered by the work of Wahba and others [Wahba, 1990]. This is related to but different from the use of kernels in classical kernel density estimation methods (e.g., see [Scott, 1992]).

There has been a significant amount of research effort devoting to the theoretical analysis of classification algorithms [Vapnik, 1998; Barron, 1993; Bartlett, 1998; Breiman, 1998; Jiang, 2004; Lugosi and Vayatis, 2004; Mannor *et al.*, 2003; Zhang, 2004; Bartlett *et al.*, 2006; Steinwart, 2005]. These work provide theoretical support for modern classification algorithms, in particular by characterizing statistical consistency and convergence rates of the resulting estimation procedures in terms of the properties of surrogate loss functions. The methods of analysis fall largely within the framework of M-estimation analysis [van de Geer, 1999; van der Vaart, 1998] using empirical process theory [van der Vaart and Wellner, 1996; Pollard, 1984]. [Zhang and Yu, 2005] analyzes the interplay between statistical convergences and computational properties of boosting algorithms.

1.2.2 Key ideas in our framework

In this section we shall outline at a high level several key ideas in our nonparametric approach to learning in decentralized systems. An elaboration of these ideas are described in the next section. Furthermore, to keep this summary relatively focused, the discussion in this section is confined only to the class of non-sequential decentralized detection problems. The sequential counterpart shall be discussed in detail in the next section as well.

As in standard classification settings, we deal with the Bayes error as the objective function. Thus, a natural idea is to replace the 0-1 loss by a convex surrogate loss function ϕ. In contrast to the standard classification settings, learning is required for both the discriminant function and quantization rules at individual monitoring devices. Our algorithm is a realization of the following M-estimation procedure, i.e., the decision rules are obtained by minimizing an empirical version of the risk functional:

$$\begin{aligned}(f, Q) &= \operatorname{argmin}_{(\gamma,Q)\in\{\Gamma,\mathcal{Q}\}} \hat{\mathbb{E}}\phi(Y, \gamma(Z)) \\ &= \operatorname{argmin}_{(\gamma,Q)\in\{\Gamma,\mathcal{Q}\}} \frac{1}{n}\sum_{i=1}^{n}\sum_{z\in\mathcal{Z}} Q(z|X_i)\phi(Y_i, \gamma(z)).\end{aligned}$$

In terms of representation, we apply the kernel methods by letting Γ be a reproducing kernel Hilbert space. By a small abuse of notation, we use $Q(Z|X)$ to denote the conditional probability representing the (stochastic) decision rule $x \mapsto z$ at individual monitoring devices: x is mapped to z with probability $Q(z|x)$.

In contrast to standard classification algorithms, replacing 0-1 loss by some convex surrogate ϕ helps but does not completely resolve the computational challenges inherent in our problem. Although the empirical risk functional can be made convex with respect to either γ or Q, it is not a convex function with respect to the joint vector (γ, Q). Nonetheless, this suggests that an efficient optimization procedure by coordinate-wise optimization is possible. A more challeging issue is that the risk functional itself is difficult to evaluate, because it involves summing over an exponential number of possible values of $z \in \mathcal{Z}$. The exponentiality is with respect to the number of dimensions of z, i.e., the number of monitoring devices in the decentralized system. To resolve this computational difficulty, we propose a method for approximating the risk functional. Our approximation method exploits the decentralization constraints implicitly imposed on the decision rule Q and the use of a *marginalized* kernel [Tsuda *et al.*, 2002], where the marginalization is defined naturally based on the conditional distribution $Q(Z|X)$. The theory of duality in convex analysis is utilized to great effect to ensure that the overall optimization algorithm can be performed efficiently to overcome the curse of dimensionality presented by X and Z.

From a statistical viewpoint, does the use of surrogate loss function ϕ still yield consistent answers in the sense of the 0-1 loss? It is worth emphazing again that the existing theory of classification is not adequate to provide an answer to this question, because our problem involves learning both the discriminant function γ and the decentralized decision rule Q. It has been proved that the broad class of so-called *classification-calibrated* loss functions [Bartlett *et al.*, 2006], including the hinge loss, exponential loss and logistic loss, all yield consistency in the classification context. We show that in our problem, i.e., classification plus experiment design, among these three loss functions, only the hinge loss yields consistent learning procedure. Furthermore, it is possible to construct a class of convex losses that have the same consistency property.

The proof of these consistency results hinge on a deeper fact about the correspondence between the class of surrogate loss functionals in binary classification, which is a decision-theoretic concept, and the class of f-divergence functionals, an information-theoretic concept arising mostly in the asymptotics. This correspondence allows us to catergorize the class of surrogate of losses into "equivalent" subclasses by examining at equivalent subclasses of f-divergences. It turns out that only those loss functions which are equivalent to the 0-1 loss can produce a consistent learning procedure. This correspondence extends the early work on the relationship between 0-1 loss and f-divergence in experiment design [Blackwell, 1951; Blackwell, 1953]. It also provides concrete decision-theoretic justifications for certain choices of divergence functionals used in existing (parametric) decentralized detection literature [Kailath, 1967; Poor and Thomas, 1977; Longo *et al.*, 1990; Chamberland and Veeravalli, 2003], as well as other experiment design contexts such as dimensionality reduction and feature selection in machine learning [Tishby *et al.*, 1999]. For instance, the choice of mutual information in the information bottleneck method [Tishby *et al.*, 1999] implies an underlying logistic loss function. The choice of Hellinger distance in [Longo *et al.*, 1990] implies an underlying exponential loss.

The correspondence between surrogate losses and f-divergences also provides a nonparametric estimation method for f-divergence functionals, by turning the estimation problem into a convex risk minimization problem. It is worth noting that the problem of estimating divergences is significant from both theoretical and practical standpoints. As will be shown in this thesis, our method for estimating f-divergence functionals link together several interesting estimation problems: estimation of integral functionals of unknown densities, estimation of function (the likelihood ratio of two unknown distributions), and classification problem (estimating the classifier).f-divergences play important roles in many practical contexts: They are the rate of various coding and compression scheme. They are also the objective functionals in the estimation procedures for many statistical tasks, including dimensionality reduction and feature selection, independent component analysis, and so on. As we shall elaborate in the sequel, they play key roles in not only (non-sequential) detection problems, but also sequential detection problems as well.

1.3 Main problems and contributions

In this section we shall elaborate on the main problems considered in this thesis and our key contributions to such problems.

- a nonparametric approach to centralized detection and estimation tasks and its application to the problem of localization in ad hoc sensor network

- a nonparametric approach to decentralized detection problem.
- a characterization of optimal decision rules of sequential decentralized detection problem
- a characterization of the correspondence between surrogate loss and divergence functionals.
- a nonparametric estimation method for divergence functionals and the likelihood ratio

At a a high level, underlying much of our thesis is an insight about the relationship between loss functions and divergence functionals. This relationship is exploited to characterize optimal decision rules in various decision-making settings of decentralized systems, to devise efficient algorithms for learning such decision rules, and to provide statistical analyses of such learning algorithms.

1.3.1 Nonparametric centralized detection and estimation

Before focusing our main attention to decentralized systems, in Chapter 2 we consider an application of a nonparametric approach to detection and estimation tasks within a centralized setting of sensor networks. This also provides a concrete platform from which we investigate and demonstrate in the sequel our nonparametric approach to decentralized systems.

A sensor network can be viewed as a pattern recognition device. Rather than transforming sensor locations and sensor readings into Euclidean, world-centric coordinates, we work directly with the (non-Euclidean) coordinate system given by the physical sensor readings themselves. Using the methodology of "kernel functions," the topology implicit in sets of sensor readings can be exploited in the construction of signal-based function spaces that are useful for the prediction of various extrinsic quantities of interest, using any of a variety of statistical algorithms for regression and classification. In Chapter 2 we illustrate this approach in a novel setting of a localization problem [Hightower and Borriello, 2000; Bulusu *et al.*, 2000; Savarese *et al.*, 2002].

The localization problem that we study is that of determining the location of a (large) number of sensors of unknown location, based on the known location of a (small) number of base sensors. Let $X_1, \ldots, X_m$ denote a set of m sensors, and let x_i denote the position in $\mathbb{R}^2$ of sensor X_i. Suppose that the locations of the first n sensors are known, i.e., $X_1 = x_1, \ldots, X_n = x_n$, where $n \ll m$. We want to estimate the positions of $X_{n+1}, \ldots, X_m$ solely on the basis of the receive/transmit signals $s(x_i, x_j)$ between pairs of sensors.

An important characteristic of radio or light signal strength is the relationship of the signal attenuation as a function of distance [Seidel and Rappaport, 1992]. For instance, for radio signals in an idealized environment, given that the sending and receiving antennas are focused on the same radio frequency, we have:

$$s \propto Pd^{-\eta}, \tag{1.1}$$

where $\eta > 2$ is a constant, and P is the sending signal voltage. Such relationships provide the basis for a variety of localization algorithms in the literature, which consist of two main steps: (1) a ranging procedure which involves estimating the distance from a sensor to another sensor based on the signal strength of the signals transmitted/received between the two, and (2) a procedure that recovers the locations of the sensors based on their pairwise distance estimates either by triangulation or by least-squares methods [Priyantha *et al.*, 2000; Girod and Estrin, 2001; Savvides *et al.*, 2001; Whitehouse, 2002]. However, the idealized model in Eq. (1.1) can be highly inaccurate due to variability caused by multipath effects and ambient noise interference as well as device-specific factors such as the frequencies of node radios, physical antenna orientation, and fluctuations in the power source [Bulusu *et al.*, 2000; Priyantha *et al.*, 2000]. Methods based on ranging inherit these inaccuracies and improvements are possible only if difficult problems in signal modeling are addressed.

We propose a method that bypasses the ranging step altogether. We show that it is possible to pose a coarse-grained localization problem as a detection (classification) problem. Fine-grained localization is then achieved by a second application of the coarse-grained localization technique. Our localization algorithm thus involves two phases. First, there is a training phase that chooses discriminant functions for classifying positions using arbitrarily constructed target regions. This phase is performed either on-line at the base stations, or taken off-line, and takes $O(n^3)$ computational time, where n is the number of base sensors. Second, once the training phase is completed, other location-unknown low-power sensors can determine their own position locally, and the computation takes only $O(n)$ time for each of these sensors.

Our approach makes use of kernel methods for classification and regression, an example of which is the “support vector machine (SVM)” [Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004]. (See Section 1.2.1 on a brief account of classification algorithms and the kernel methods developed in statistics and machine learning literature). Central to this approach is the notion of a *kernel function*, which provides a generalized measure of similarity for any pair of entities (e.g., sensor locations). The functions that are output by the SVM and other kernel methods are sums of kernel functions, with the number of terms in the sum equal to the number of data points.

Kernel functions typically used in practice include Gaussian kernels and polynomial kernels. A technical requirement of these functions is that they are positive semidefinite, which is equivalent to the requirement that the $n \times n$ *Gram matrix* formed by evaluating the kernel on all pairs of n data points is a positive semidefinite matrix. Intuitively, this requirement allows a kernel function to be interpreted as a generalized measure of similarity. The kernel function imposes a topology on the data points which is assumed to be useful for the prediction of extrinsic quantities such as classification labels.

Given that the raw signal readings in a sensor network implicitly capture topological relations among the sensors, kernel methods would seem to be particularly natural in the sensor network setting. In the simplest case, the signal strength would itself be a kernel function and the *signal matrix* $(s(x_i, x_j))_{ij}$ would be a positive semidefinite matrix. Alternatively, the matrix may be well approximated by a positive semidefinite matrix (e.g., a simple transformation that symmetrizes the signal matrix and adds a scaled identity matrix may be sufficient). More generally, and more realistically, derived kernels can be defined based on the signal matrix. In particular, inner products between vectors of received signal strengths necessarily define a positive semidefinite matrix and can be used in kernel methods. Alternatively, generalized inner products of these vectors can be computed—this simply involves the use of higher-level kernels whose arguments are transformations induced by lower-level kernels. In general, hierarchies of kernels can be defined to convert the initial topology provided by the raw sensor readings into a topology more appropriate for the classification or regression task at hand. This can be done with little or no knowledge of the physical sensor model.

1.3.2 Nonparametric decentralized detection

Consider a decentralized sensor network system, which typically involves a set of sensors that receive observations from the environment, but are permitted to transmit only a summary message (as opposed to the full observation) back to a fusion center. On the basis of its received messages, this fusion center then chooses a final decision from some number of alternative hypotheses about the environment. The problem of decentralized detection is to design the local decision rules at each sensor, which determine the messages that are relayed to the fusion center, as well a decision rule for the fusion center itself [Tsitsiklis, 1993b]. A key aspect of the problem is the presence of *communication constraints*, meaning that the sizes of the messages sent by the sensors back to the fusion center must be suitably "small" relative to the raw observations, whether measured in terms of either bits or power. The decentralized nature of the system is to be contrasted with a centralized system, in which the fusion center has access to the full collection of raw observations. See Section 1.1.1 for a review of existing approaches to the problem of decentralized detection, and Section 1.2.2 for an overview of our key ideas.

Recalling our setup, let $Y \in \{-1, +1\}$ be a random variable, representing the two possible hypotheses in a binary hypothesis-testing problem. Moreover, suppose that the system consists of S sensors, each of which observes a single component of the S-dimensional vector $X = (X^1, \ldots, X^S)$. One starting point is to assume that the joint distribution $P(X, Y)$ falls within some parametric family. Of course, such an assumption raises the modeling issue of how to determine an appropriate parametric family, and how to estimate parameters. Both of these problems are very challenging in contexts such as sensor networks, given highly inhomogeneous distributions and a large number S of sensors. Our focus in this thesis is on relaxing this assumption, and developing a nonparametric method in which no assumption about the joint distribution $P(X, Y)$ is required. Instead, we posit that a number of empirical samples $(X_i, Y_i)_{i=1}^n$ are given.

Our approach, to be described in Chapter 3, is based on a combination of ideas from reproducing-kernel Hilbert spaces [Aronszajn, 1950; Saitoh, 1988], and the framework of empirical risk minimization from nonparametric statistics. Methods based on reproducing-kernel Hilbert spaces (RKHSs) have figured prominently in the literature on centralized signal detection and estimation for several decades (e.g., [Weinert, 1982; Kailath, 1971]). More recent work in statistical machine learning (e.g., [Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004]) has demonstrated the power and versatility of kernel methods for solving classification or regression problems on the basis of empirical data samples. Roughly speaking, kernel-based algorithms in statistical machine learning involve choosing a function, which though linear in the RKHS, induces a nonlinear function in the original space of observations. A key idea is to base the choice of this function on the minimization of a *regularized empirical risk* functional. This functional consists of the empirical expectation of a convex loss function ϕ, which represents an upper bound on the 0-1 loss (the 0-1 loss corresponds to the probability of error criterion), combined with a regularization term that restricts the optimization to a convex subset of the RKHS. It has been shown that suitable choices of margin-based convex loss functions lead to algorithms that are robust both computationally [Schölkopf and Smola, 2002], as well as statistically [Zhang, 2004; Bartlett *et al.*, 2006]. The use of kernels in such empirical loss functions greatly increases their flexibility, so that they can adapt to a wide range of underlying joint distributions.

We show how kernel-based methods and empirical risk minimization are naturally suited to the decentralized detection problem. More specifically, a key component of the methodology that we propose involves the notion of a marginalized kernel, where the marginalization is induced by the transformation from the observations X to the local decisions Z. The decision rules at each sensor, which can be either probabilistic or deterministic, are defined by conditional probability distributions of the form $Q(Z|X)$, while the decision at the fusion center is defined in terms of $Q(Z|X)$ and a linear function over the corresponding RKHS. We develop and analyze an algorithm for optimizing the design of these decision rules. It is interesting to note that this algorithm is similar in spirit to a suite of *locally optimum* detectors in the literature (e.g, [Blum *et al.*, 1997]), in the sense that one step consists of optimizing the decision rule at a given sensor while fixing the decision rules of the rest, whereas another step involves optimizing the decision rule of the fusion center while holding fixed the local decision rules at each sensor. Our development relies heavily on the convexity of the loss function ϕ, which allows us to leverage results from convex analysis [Rockafellar, 1970] so as to derive an efficient optimization procedure. In addition, we analyze the statistical properties of our algorithm, and provide probabilistic bounds on its performance.

1.3.3 Surrogate losses and f-divergence

In Chapter 4 we study the roles of and relationships between surrogate losses and f-divergences in the context of centralized and decentralized detection problems. As mathematical objects, the f-divergences studied in information theory and the surrogate loss functions studied in statistical machine learning are fundamentally different: the former are functions on pairs of measures, whereas the latter are functions on values of discriminant functions and class labels. However, their underlying role in obtaining computationally-tractable algorithms for discriminant analysis suggests that they should be related. Indeed, Blackwell's result hints at such a relationship, but its focus on 0-1 loss does not lend itself to developing a general relationship between f-divergences and surrogate loss functions. The primary contribution of Chapter 4 is to provide a detailed analysis of the relationship between f-divergences and surrogate loss functions, developing a full characterization of the connection, and explicating its consequences. We show that for any surrogate loss, regardless of its convexity, there exists a corresponding convex f such that minimizing the expected loss is equivalent to maximizing the f-divergence. We also provide necessary and sufficient conditions for an f-divergence to be realized from some (decreasing) convex loss function. More precisely, given a convex f, we provide a constructive procedure to generate *all* decreasing convex loss functions for which the correspondence holds.

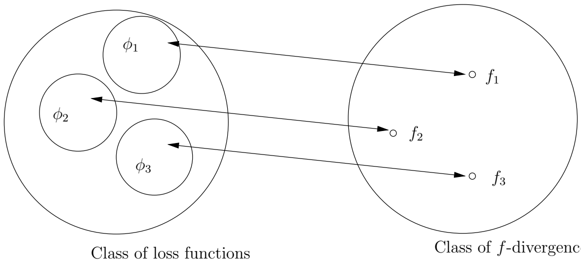


Figure 1.3. Illustration of the correspondence between f-divergences and loss functions. For each loss function ϕ, there exists exactly one corresponding f-divergence (induced by some underlying convex function f) such that the ϕ-risk is equal to the negative f-divergence. Conversely, for each f-divergence, there exists a whole set of surrogate loss functions ϕ for which the correspondence holds. Within the class of convex loss functions and the class of f-divergences, one can construct equivalent loss functions and equivalent f-divergences, respectively. For the class of classification-calibrated decreasing convex loss functions, we can characterize the correspondence precisely.

The relationship is illustrated in Figure 4.1; whereas each surrogate loss ϕ induces only one f-divergence, note that in general there are many surrogate loss functions that correspond to the same f-divergence. As particular examples of the general correspondence established in this paper, we show that the hinge loss corresponds to the variational distance, the exponential loss corresponds to the Hellinger distance, and the logistic loss corresponds to the capacitory discrimination distance.

This correspondence—in addition to its intrinsic interest as an extension of Blackwell's work—has several specific consequences. First, there are numerous useful inequalities relating the various f-divergences [Topsoe, 2000]; our theorem allows these inequalities to be exploited in the analysis of loss functions. Second, the minimizer of the Bayes error and the maximizer of f-divergences are both known to possess certain extremal properties [Tsitsiklis, 1993a]; our result provides a natural connection between these properties. Third, our theorem allows a notion of equivalence to be defined among loss functions: in particular, we say that loss functions are equivalent if they induce the same f-divergence. We then exploit the constructive nature of our theorem to exhibit all possible convex loss functions that are equivalent (in the sense just defined) to the 0-1 loss. Finally, we illustrate the application of this correspondence to the problem of decentralized detection. Whereas the more classical approach to this problem is based on f-divergences [Kailath, 1967; Poor and Thomas, 1977], our method instead builds on the framework of statistical machine learning. The correspondence allows us to establish consistency results for the algorithmic framework for decentralized detection described in Chapter 3: in particular, we prove that for any surrogate loss function equivalent to 0-1 loss, our estimation procedure is consistent in the strong sense that it will asymptotically choose Bayes-optimal quantization rules.

1.3.4 Sequential decentralized detection

In Chapter 5 we take a detour from the non-sequential setting, and consider instead the sequential setting of the decentralized detection problem. The reader is refered to Section 1.1.2 for a brief background of this problem.

We are interested in particular the following problem of sequential decentralized detection, which is "Case A" in the framework of [Veeravalli, 1999]: Let $X_1, X_2, \ldots$ be a data sequence drawn i.i.d. by either probability distribution $\mathbb{P}_0$ or $\mathbb{P}_1$, which correspond to the two hypotheses $H = 0$ or 1, with prior π^1 and π^0, respectively. Note that the random X_i can be multivariate; each variate is collected by a sensor in a sensor network. Due to communication constraints, however, given a data X_i, each sensor transmits a message $U_i = Q_i(X_i)$ to a fusion center. Thus, the fusion center receives a sequence of (possibly multivariate) messages $U_1, U_2, \ldots$, one at a time, and has to decide when to stop receiving data based on a stopping time N [2], and to determine the hypothesis H via an estimate

[2] In technical terms, a stopping time N is a random variable defined with respect to the sigma-field

$\hat{H} = \gamma(U_1, \ldots, U_N)$. In a Bayesian setting of the problem, the performance measure is *sequential cost* made up by a weighted sum of the detection error, and the expected time delay:

$$\mathbb{P}(H \neq \hat{H}) + c\mathbb{E}N,$$

where c denotes the cost of each extra sample U. The overall goal of a sequential detection problem is to determine the decision triple (Q, N, γ) so as to minimize the sequential cost. In the sensor network setting, the decision rule Q is also called the quantization rule at sensors.

Note that when Q is fixed, we are reduced to a classical sequential detection problem, which was well-understood [Wald, 1947; Shiryayev, 1978; Siegmund, 1985; Lai, 2001]. Thus the key issues lie in the characterization of the optimal quantization rules $(Q_1, Q_2, \ldots$. Veeravalli et al [Veeravalli *et al.*, 1993; Veeravalli, 1999] conjectured that the optimal decision rule Q is stationary, e.g., the quantization rule Q_n at each time step n is independent of n, at least in the asymptotic setting as $c \rightarrow 0$. This is due to an observation that as $c \rightarrow 0$, the stopping time tends to infinity. Thus, each sample at a time step may have the same role in the asymptotic setting. The stationary conjecture has remained open since it was first posed.

Characterizing the optimal rules Q has important implication if we are to take the sequential detection problem beyond the original parametric setting in the existing literature. Indeed, if we drop the assumptions that the distributions $\mathbb{P}_0$ and $\mathbb{P}_1$ are known, and now view $U_i = Q_i(X_i)$ as a summarizing statistic, then a key issue would be how choose the best class of statistical functions Q_i's in a sequential estimation procedure.

One primary contribution in this chapter is to show that stationary decision functions are, in fact, *not* optimal. Our argument is based on an asymptotic characterization of the optimal Bayesian risk as the cost per sample goes to zero. In this asymptotic regime, the optimal cost can be expressed as a simple function of priors and Kullback-Leibler (KL) divergences. This characterization allows us to construct counterexamples to the stationarity conjecture, both in an exact and an asymptotic setting. In the latter setting, we present a broad class of problems in which there always exists a range of prior probabilities for which stationary strategies, either deterministic or randomized, are suboptimal. We note in passing that an intuition for the source of this suboptimality is easily provided—it is due to the asymmetry of the KL divergence.

It is well known that optimal quantizers when unrestricted are necessarily likelihood-based threshold rules [Tsitsiklis, 1986]. Our counterexamples and analysis imply that optimal thresholds are not generally stationary (i.e., the threshold may differ from sample to sample). We also provide a partial converse to this result: specifically, if we restrict ourselves to stationary (or blockwise stationary) quantizer designs, then there exists an optimal

$\sigma(U_1, \ldots, U_N)$ generated by the random sequence $U_1, U_2, \ldots$ [Durrett, 1995].

design that is a threshold rule based on the likelihood ratio. We prove this result by establishing a quasiconcavity result for the asymptotically optimal cost function. In this thesis, this result is proven for the space of deterministic quantizers with arbitrary output alphabets, as well as for the space of randomized quantizers with binary ouputs. We conjecture that the same result holds more generally for randomized quantizers with arbitrary output alphabets.

1.3.5 Nonparametric estimation of f-divergence functionals and the likelihood ratio

One consequence of the relationship between surrogate losses and f-divergences studied in Chapter 4 is a non-asymptotic decision-theoretic variational characterization of f-divergence functionals. This allows us to devise and analyze a nonparametric estimation method for f-divergence functionals and the likelihood ratio. Recall that an f-divergence $D_\phi(\mathbb{P}, \mathbb{Q})$ captures a "distance" between two distributions $\mathbb{P}$ and $\mathbb{Q}$:

$$D_\phi(\mathbb{P}, \mathbb{Q}) = \int p_0 \phi(q_0/p_0)\ d\mu,$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is a convex function.

This problem estimating D_ϕ has important applications. As noted earlier, divergences play important roles not only in learning in (non-sequential and sequential) decentralized systems. They also have a fundamental role as an objective to optimize in various other data analysis and learning tasks, including dimentionality reduction and feature selection. An important quantity in information theory, the Shannon mutual information, can be viewed as a KL divergence. The KL divergence is used as the bit rate in several compression schemes. Mutual information is often used as a measure of independence to be minimized such as in the problem of independent component analysis [Hyvarinen *et al.*, 2001]. If the divergences are to be used as objective functional in such tasks, one has to be able to estimate them efficiently from empirical data.

We propose a novel M-estimator for the likelihood ratio and the family of f-divergences based on the variational characterization of f-divergence as explained above. Our estimation procedure is inherently nonparametric: $\mathbb{P}$ and $\mathbb{Q}$ are not known. Nor do we make strong assumptions on the forms of the densities for $\mathbb{P}$ and $\mathbb{Q}$. The estimation procedure is based on i.i.d. empirical samples $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ drawn from $\mathbb{P}$ and $\mathbb{Q}$, respectively.

We provide a consistency and convergence analysis for our estimators. For the analysis, we make assumptions on the boundedness of the *density ratio*, which can be relaxed in some cases. The maximization procedure is cast over a whole function class $\mathcal{G}$ of density ratio, thus our tool is based on results from the theory of empirical processes. Our method of proof is based on the analysis of M-estimation for nonparametric density estimation [van de Geer, 1999; van der Vaart and Wellner, 1996]. The key issue essentially hinges on the modulus of continuity of the suprema of two empirical processes (defined on $\mathbb{P}$ and $\mathbb{Q}$ measures) with respect to a metric defined on the class $\mathcal{G}$. This metric turns out to be a surrogate lower bound of a Bregman divergence defined on a pair of density ratios. Our choice of metrics include the Hellinger distance and L_2 norm.

We provide an efficient implementation of our estimation procedures using RKHS as the relevant function classes. Our estimation method compares favorably againts existing approaches in the literature.

1.4 Thesis organization

The remainder of this thesis is organized as follows.

Chapter 2: Nonparametric centralized detection and estimation

This chapter introduces the use of kernel methods in centralized detection and estimation by considering a challenging problem of localization in ad hoc sensor network. It also provides a concrete application of our nonparametric approach as we go decentralized in the sequel.

Chapter 3: Nonparametric decentralized detection

This chapter considers the problem of decentralized detection, proposes a nonparametric learning algorithm and describes its computational and statistical properties.

Chapter 4: Surrogate losses and f-divergence functionals

This chapter investigates the correspondence of surrogate loss functions and divergence funtionals and the implications of this correspondence. As an application we prove the consistency of the learning algorithm proposed in Chapter 3.

Chapter 5: Optimal quantization rules in sequential decentralized detection

This chapter studies the structure of optimal decision rules in a sequential decentralized detection problem.

Chapter 6: Nonparametric estimation of divergences and the likelihood ratio

This chapter introduces and analyzes a nonparametric estimation procedure for divergence functionals and the likelihood ratio.

Chapter 7: Contributions and suggestions

This chapter summarizes the contributions of the thesis, and discusses several directions for future research.

All background knowledge are included in each individual chapter, making each chapter sufficiently self-contained. Nonetheless, Chapter 2 is a good warm-up for the materials developed in Chapter 3, especially for the readers who are new to kernel methods and their application to detection and estimation problems. For readers who are interested in the motivation of the theoretical study of losses and divergence funtionals it is useful to start with Chapter 3 before going into Chapter 4. Chapter 5 focuses on sequential detection problems and can be read independently from the rest. Techniques introduced in Chapter 6 have useful applications that go beyond the context of decentralized systems and can also be read independently without the background in the previous chapters.

Chapter 2

Nonparametric centralized detection and estimation

This chapter demonstrates the use of kernel methods in a challenging problem of localization in sensor networks. We show that the coarse-grained and fine-grained localization problems for ad hoc sensor networks can be posed and solved as a pattern recognition problem using kernel methods from statistical learning theory. This stems from an observation that the kernel function, which is a similarity measure critical to the effectiveness of a kernel-based learning algorithm, can be naturally defined in terms of the matrix of signal strengths received by the sensors. Thus we work in the natural coordinate system provided by the physical devices. This not only allows us to sidestep the difficult ranging procedure required by many existing localization algorithms in the literature, but also enables us to derive a simple and effective localization algorithm. The algorithm is particularly suitable for networks with densely distributed sensors, most of whose locations are unknown. The computations are initially performed at the base sensors and the computation cost depends only on the number of base sensors. The localization step for each sensor of unknown location is then performed locally in linear time. We present an analysis of the localization error bounds, and provide an evaluation of our algorithm on both simulated and real sensor networks. [1]

2.1 Introduction

A sensor network can be viewed as a distributed pattern recognition device. In the pattern recognition approach, rather than transforming sensor locations and sensor readings into Euclidean, world-centric coordinates, we work directly with the (non-Euclidean) coordinate system given by the physical sensor readings themselves. Using the methodology

[1] This work has been published in [Nguyen *et al.*, 2005a].

of "kernel functions," the topology implicit in sets of sensor readings can be exploited in the construction of signal-based function spaces that are useful for the prediction of various extrinsic quantities of interest, using any of a variety of statistical algorithms for regression and classification. In the current chapter we illustrate this approach in the setting of a localization problem [Hightower and Borriello, 2000; Bulusu *et al.*, 2000; Savarese *et al.*, 2002].

The localization problem that we study is that of determining the location of a (large) number of sensors of unknown location, based on the known location of a (small) number of base sensors. Let $X_1, \ldots, X_m$ denote a set of m sensors, and let x_i denote the position in $\mathbb{R}^2$ of sensor X_i. Suppose that the locations of the first n sensors are known, i.e., $X_1 = x_1, \ldots, X_n = x_n$, where $n \ll m$. We want to recover the positions of $X_{n+1}, \ldots, X_m$ solely on the basis of the receive/transmit signals $s(x_i, x_j)$ between pairs of sensors.

An important characteristic of radio or light signal strength is the relationship of the signal attenuation as a function of distance [Seidel and Rappaport, 1992]. For instance, for radio signals in an idealized environment, given that the sending and receiving antennas are focused on the same radio frequency, we have:

$$s \propto P d^{-\eta}, \tag{2.1}$$

where $\eta > 2$ is a constant, and P is the sending signal voltage. Such relationships provide the basis for a variety of localization algorithms in the literature, which consist of two main steps: (1) a ranging procedure which involves estimating the distance from a sensor to another sensor based on the signal strength of the signals transmitted/received between the two, and (2) a procedure that recovers the locations of the sensors based on their pairwise distance estimates either by triangulation or by least-squares methods [Priyantha *et al.*, 2000; Girod and Estrin, 2001; Savvides *et al.*, 2001; Whitehouse, 2002]. Unfortunately, however, the idealized model in Eq. (1.1) can be highly inaccurate due to variability caused by multipath effects and ambient noise interference as well as device-specific factors such as the frequencies of node radios, physical antenna orientation, and fluctuations in the power source [Bulusu *et al.*, 2000; Priyantha *et al.*, 2000]. Methods based on ranging inherit these inaccuracies and improvements are possible only if difficult problems in signal modeling are addressed.

In this chapter we propose a method that bypasses the ranging step altogether. We show that it is possible to pose a coarse-grained localization problem as a discriminative classification problem that can be solved using tools from the statistical machine learning literature. Fine-grained localization is then achieved by a second application of the coarse-grained localization technique. Our localization algorithm thus involves two phases. First, there is a training phase that chooses discriminant functions for classifying positions using arbitrarily constructed target regions. This phase is performed either on-line at the base stations, or taken off-line, and takes $O(n^3)$ computational time, where n is the number of base sensors. Hence, our assumption is that the base sensors have sufficient power and processing capability (indeed, these are also the nodes that might have GPS-capability to determine their own exact locations). Second, once the training phase is completed, other location-unknown low-power sensors can determine their own position locally, and the computation takes only $O(n)$ time for each of these sensors.

Our approach makes use of kernel methods for statistical classification and regression [Schölkopf and Smola, 2002], an example of which is the "support vector machine (SVM)." Central to this approach is the notion of a *kernel function*, which provides a generalized measure of similarity for any pair of entities (e.g., sensor locations). The functions that are output by the SVM and other kernel methods are sums of kernel functions, with the number of terms in the sum equal to the number of data points. Kernel methods are examples of *nonparametric* statistical procedures—procedures that aim to capture large, open-ended classes of functions.

Kernel functions typically used in practice include Gaussian kernels and polynomial kernels. A technical requirement of these functions is that they are positive semidefinite, which is equivalent to the requirement that the $n \times n$ *Gram matrix* formed by evaluating the kernel on all pairs of n data points is a positive semidefinite matrix. Intuitively, this requirement allows a kernel function to be interpreted as a generalized measure of similarity. The kernel function imposes a topology on the data points which is assumed to be useful for the prediction of extrinsic quantities such as classification labels.

Given that the raw signal readings in a sensor network implicitly capture topological relations among the sensors, kernel methods would seem to be particularly natural in the sensor network setting. In the simplest case, the signal strength would itself be a kernel function and the *signal matrix* $(s(x_i, x_j))_{ij}$ would be a positive semidefinite matrix. Alternatively, the matrix may be well approximated by a positive semidefinite matrix (e.g., a simple transformation that symmetrizes the signal matrix and adds a scaled identity matrix may be sufficient). More generally, and more realistically, derived kernels can be defined based on the signal matrix. In particular, inner products between vectors of received signal strengths necessarily define a positive semidefinite matrix and can be used in kernel methods. Alternatively, generalized inner products of these vectors can be computed—this simply involves the use of higher-level kernels whose arguments are transformations induced by lower-level kernels. In general, hierarchies of kernels can be defined to convert the initial topology provided by the raw sensor readings into a topology more appropriate for the classification or regression task at hand. This can be done with little or no knowledge of the physical sensor model.

Our focus is on the discriminative classification problem of locating sensors in an ad hoc sensor network. It is worth noting that similar methods have been explored recently in the context of tracking one or more objects (e.g., mobile robots) that move through a wireless sensor field.[2] Systems of this type include Active Badge [Want *et al.*, 1992], [Ward *et al.*, 1997], RADAR [Bahl and Padmanabhan, 2000], Cricket [Priyantha *et al.*, 2000], and UW-CSP (cf. [Li *et al.*, 2002]). In [Bahl and Padmanabhan, 2000], the authors describe a simple nearest neighbor classification algorithm to obtain coarse localization of objects. Most closely related to our approach is the work of [Li *et al.*, 2002] in which a number of classification algorithms are used for tracking moving vehicles, including k-nearest neighbor and support vector machines. We elaborate on the connections between this work and ours in the description of our algorithm.

The chapter is organized as follows. We begin with a brief background of classification using kernel methods, and motivate our application of kernel methods to the localization problem based on sensor signal strength. Next, the localization algorithm and its error analysis are described. We then present details of the implementation of the algorithm and its computational cost, followed by an evaluation of our algorithm with simulated and real sensor networks. Finally, we present our discussions in the final section.

2.2 Classification using kernel methods

In a classification algorithm, we are given as training data n samples $(x_i, y_i)_{i=1}^n$ in $\mathcal{X} \times \{\pm 1\}$, where $\mathcal{X}$ denotes the input space. Each y_i specifies whether the data point $x_n \in \mathcal{X}$ lies in a class $C \subseteq \mathcal{X}$ ($y_i = 1$) or not ($y_i = -1$). A classification algorithm involves finding a discriminant function $y = \text{sign}(f(x))$ that minimizes the classification error $P(Y \neq \text{sign}(f(X)))$.

Central to a kernel-based classification algorithm (e.g., the SVM) is the notion of a kernel function $K(x, x')$ that provides a measure of similarity between two data points x and x' in $\mathcal{X}$. Technically, K is required to be a symmetric positive semidefinite function.[3] For such a function, Mercer's theorem implies that there must exist a feature space $\mathcal{H}$ in which K acts as an inner product, i.e., $K(x, x') = \langle \Phi(x), \Phi(x') \rangle$ for some mapping $\Phi(x)$. The SVM and related kernel-based algorithms choose a linear function $f(x) = \langle \mathrm{w}, \Phi(x) \rangle$ in this feature space. That is, they find a vector w which minimizes the loss

$$\sum_{i=1}^{n} \phi(y_i f(x_i))$$

subject to $||\mathrm{w}|| \leq B$ for some constant B. Here ϕ denotes a convex function that is an

[2] The alternative to discriminative classification is classification using *generative* probabilistic models. This is a well-explored area that dates back to contributors such as Wiener and Kalman. Recent work in this vein focuses on the distributed and power-constrained setting of wireless sensor networks (e.g. [Sheng and Hu, 2003; D'Costa and Sayeed, 2003]).

[3] For a translation-invariant kernel, i.e., $K(x, x') = h(x - x')$ for some function h, K is a positive semidefinite kernel if the Fourier transform of h is non-negative.

upper bound on the 0-1 loss $\mathbb{I}(y \neq \text{sign}(f(x)))$.[4] In particular, the SVM algorithm is based on the hinge loss $\phi(yf(x)) = (1 - yf(x))_+$.[5] By the Representer Theorem (cf. [Schölkopf and Smola, 2002]), it turns out that the minimizing f can be expressed directly in terms of the kernel function K:

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) \tag{2.2}$$

for an optimizing choice of coefficients α_i.

There are a large number of kernel functions that satisfy the positive semidefinite property required by the SVM algorithm. Examples include the Gaussian kernel:

$$K(x, x') = \exp -(||x - x'||^2/\sigma)$$

as well as the polynomial kernel:

$$K(x, x') = (\gamma + ||x - x'||)^{-\sigma},$$

for parameters σ and γ. Both of these kernel functions decay with respect to the distance $||x - x'||$, a property that is shared by most idealized signal strength models. In particular, the radio signal model (1.1) has a form similar to that of a polynomial kernel. In [Sheng and Hu, 2003], the authors justify the use of an acoustic energy model for localization that has the form of the Gaussian kernel above. These relationships suggest a basic connection between kernel methods and sensor networks. In particular, a naive usage of kernel methods could be envisaged in which signal strength is used directly to define a kernel function. In general, however, signal strength in real sensor networks need not define a positive semidefinite function. Nonetheless, it is the premise of this chapter that signal strength matrices provide a useful starting point for defining kernel-based discriminant functions. We show how to define derived kernels which are stacked on top of signal strength measurements in the following section.

Finally, it is worth noting that multi-modal signals are naturally accommodated within the kernel framework. Indeed, suppose that we have D types of sensory signals, each of which can be used to define a kernel function $K_d(x, x')$ for $d = 1, \ldots, D$. Then any conic combination of K_d yields a new positive semidefinite function:

$$K(x, x') = \sum_{d=1}^{D} \beta_d K_d(x, x').$$

There are methods for choosing the parameters $\beta_d > 0$ based on empirical data [Lanckriet

[4] The indicator function is defined as $\mathbb{I}(A) = 1$ if A is true, and 0 otherwise.

[5] The subscript + notation means that $x_+ = \max(x, 0)$.

et al., 2004].

2.3 Localization in ad hoc sensor network

2.3.1 Problem statement

We assume that a large number of sensors are deployed in a geographical area. The input to our algorithm is a set of m sensors, denoted by $X_1, \ldots, X_m$. For each i we denote by x_i the position in $\mathbb{R}^2$ of sensor X_i. Suppose that the first n sensor locations are known, i.e., $X_1 = x_1, \ldots, X_n = x_n$, where $n \ll m$. For every pair of sensors X_i and X_j, we are given the signal $s(x_i, x_j)$ that sensor X_j receives from X_i. We want to recover the positions of $X_{n+1}, \ldots, X_m$.

2.3.2 Algorithm description

We first aim to obtain a coarse location estimate for $X_{n+1}, \ldots, X_m$. Given an arbitrarily constructed region $C \subseteq \mathbb{R}^2$, we ask whether $X_i \in C$ or not, for $i = n+1, \ldots, m$. This can be readily formulated as a classification problem. Indeed, since the location of the base sensors $X_1, \ldots, X_n$ are known, we know whether or not each of these base sensors are in C. Hence we have as our training data n pairs $(x_i, y_i = \mathrm{sign}(x_i \in C))_{i=1}^n$. For any sensor X_j, $j = n+1, \ldots, m$, we can predict whether $X_j \in C$ or not based on the sign of the discriminant function $f(x_j)$:

$$f(x_j) = \sum_{i=1}^{n} \alpha_i K(x_i, x_j). \tag{2.3}$$

We emphasize that the value of $f(x_j)$ is known because the values of the kernels, $K(x_i, x_j)$, are known, despite the fact that we do not know the position x_j per se.

Next, we turn to the definition of the kernel matrix $K = (K(x_i, x_j))_{1 \le i,j \le m}$. In general we envision a hierarchy of kernels based on the signal matrix. An example of such a hierarchy is as follows:

1. We might simply define $K(x_i, x_j) = s(x_i, x_j)$. We call this naive choice a *first-tier* kernel. If the signal matrix $S = (s(x_i, x_j))_{1 \le i,j \le m}$ is a symmetric positive semidefinite Gram matrix then this approach is mathematically correct although it may not yield optimal performance. If S is not symmetric positive semidefinite, then a possible approximation is $(S + S^T)/2 + \delta I$. This matrix is symmetric, and is positive semidefinite for sufficiently large $\delta > 0$ (in particular, for δ larger in absolute value than the most negative eigenvalue of $(S + S^T)/2$).

2. Alternatively, define $K = S^T S$, to be refered to as a *second-tier* linear kernel. K is always symmetric positive semidefinite. This kernel can be interpreted as an inner product for a feature space $\mathcal{H}$ which is spanned by vectors of the form:

   $$\Phi(x) = (s(x, x_1), s(x, x_2), \ldots, s(x, x_m)).$$

   Specifically, we define:

   $$K(x_i, x_j) = \sum_{t=1}^{m} s(x_i, x_t) s(x_j, x_t).$$

   Intuitively, the idea is that sensors that are associated with similar vectors of sensor readings are likely to be nearby in space.

3. Finally, it is also possible to evaluate any kernel function (e.g., Gaussian) on the feature space $\mathcal{H}$ induced by the second-tier kernel. This yields a symmetric positive semidefinite matrix, to be refered to as a *third-tier* kernel. Specifically, a third-tier Gaussian kernel has the following form, for a parameter σ:

   $$\begin{aligned} K(x_i, x_j) &= \exp\left\{ - \frac{\|\Phi(x_i) - \Phi(x_j)\|^2}{\sigma} \right\} \\ &= \exp\left\{ - \frac{\sum_{t=1}^{m} (s(x_i, x_t) - s(x_j, x_t))^2}{\sigma} \right\}. \end{aligned}$$

Given training data $(x_i, y_i)_{i=1}^n$ and a kernel function K, we apply the SVM algorithm to learn a discriminant function $f(x)$ as in Eq. (2.2). The algorithmic details and computational costs are discussed in Section 2.4.

Our classification formulation has several noteworthy characteristics. First, the training points correspond to the base sensors, and thus may be limited in number, making the learning problem nominally a difficult one. However, because we are free to choose the target region C, the problem can in fact be made easy. This ability to design the geometry of the boundary to fit the geometry of the classifier distinguishes this problem from a traditional pattern recognition problem.

The second characteristic is that we require that the network be relatively dense. As seen in Eq. (2.3), the prediction of position is based on a sum over sensors, and an accurate prediction can be achieved in general only if there are enough non-zero terms in the sum for it to be statistically stable.

A related point is that it is not necessary that the network be completely connected. If the sensor reading $s(x_i, x_j)$ is generally small or zero for a pair of sensors, then that term does not perturb the kernel calculation or the discriminant calculation. If readings fluctuate between small values and large non-zero values, then the prediction will generally be degraded. Given that the approach is a statistical approach, however, with predictions based on an aggregation over neighboring sensors, it should be expected to exhibit a certain degree of robustness to fluctuations. This robustness should be enhanced by the protocol for fine-grained estimation, as we now discuss.

We now turn to the fine-grained estimate of sensor positions. We use the coarse-grained solution presented above as a subroutine for a localization algorithm for sensors $X_j (j = n+1, \ldots, m)$. The idea is as follows: We fix a number of overlapping regions $C_\beta (\beta = 1, \ldots, U)$ in the geographical region containing the sensor network. For each β, we formulate a corresponding classification problem with respect to class C_β and predict whether or not $X_j \in C_\beta$. Hence, X_j has to be in the intersection of regions that contain it. We might, for example, assign its location x_j to be the centroid of such an intersection. Given an appropriate choice of granularity and shapes for the regions C_β, if most of the classification labels are correct we expect to be able to obtain a good estimate of x_j.

As we have seen in our experiments on both simulated data (using a Gaussian or polynomial kernel) and real sensor data (using kernels that are constructed directly from the signal matrix), given a sufficient number of base sensors (i.e., training data points), the SVM algorithm can fit regions of arbitrary shape and size with reasonable accuracy. When the number of base sensors is limited, it is found that the SVM algorithm can still fit elliptic shapes very well. This can be turned to our advantage for fine-grained localization: By picking appropriate regions C_β such as ellipses that are easy to classify, we do not need many base sensors to achieve reasonable localization performance for the entire network. In the sequel, we will show that this intuition can be quantified to give an upper bound on the expected (fine-grained) localization error with respect to the number of base sensors.

2.3.3 Localization error analysis

Suppose that the sensor network of size $L \times L$ is covered uniformly by k^2 discs with radius R. Then any given point in the sensor network is covered by approximately $\pi(Rk/L)^2$ discs. Each of these discs are used to define the region for a region classification problem. To obtain a fine-grained location estimate for all remaining sensors, X_j for $j = n+1, \ldots, m$, we need to solve k^2 region classification problems. Let e_β be the training error for each of these problems, for $\beta = 1, \ldots, k^2$. That is,

$$e_\beta = \sum_{i=1}^{n} \phi(\text{sign}(x_i \in C_\beta) f(x_i)).$$

Since the size and shape of the regions are ours to decide, it is reasonable to assume that the training error for these classification problems are small. For instance, the circle/elliptic shape is particularly suited for Gaussian or polynomial kernels. Define $\epsilon(R)$ to be the upper bound for all training errors:

$$\epsilon(R) = \max_{\beta} e_{\beta}.$$

Our analysis needs the following assumption:

Assumption 2.1. *If a sensor is correctly classified with respect to all covering discs, then it is also correctly classified with respect to all remaining discs.*

This assumption is reasonable and follows from an observation that the covering discs imply boundaries that are closer to a given sensor location. Thus the classification problems with respect to the covering discs tend to be more difficult than with respect to other dics located farther away from the sensor location.

Using a generalization error bound for margin-based classification [Koltchinskii and Panchenko, 2002], for each $\beta = 1, \ldots, k^2$, the probability of misclassification for each new sensor X_j and region C_β is $e_\beta + O(1/\sqrt{n})$, where n is the number of training points (i.e., number of base sensors). Since each location is covered by $\pi R^2 k^2 / L^2$ discs, the probability of misclassification for at least one these covering discs is, by the union bound, less than $\frac{\pi R^2 k^2}{L^2}(\epsilon(R) + O(1/\sqrt{n}))$. If a given sensor is correctly classified with respect to all of its covering discs, then we assign the sensor's location to be the center of the intersection of all these discs, in which case the localization error is bounded by $O(L/k)$.

Hence, the expectation of the localization error is bounded by

$$O\left(\frac{L}{k}\right) + \frac{\pi R^2 k^2}{L}\left(\epsilon(R) + O(1/\sqrt{n})\right).$$

This asymptotic bound is minimized by letting $k \propto L^{2/3} R^{-2/3} (\epsilon(R) + O(1/\sqrt{n}))^{-1/3}$. The bound then becomes $O(L^{1/3} R^{2/3} (\epsilon(R) + O(1/\sqrt{n}))^{1/3})$.

In summary, we have proved the following:

Proposition 2.2. *Assume that all sensor locations are independently and identically distributed according to an (unknown) distribution. For any sensor location x, let $\hat{x}$ be the location estimate given by our algorithm. Then, under Assumption 2.1 there holds:*

$$\mathbb{E}||x - \hat{x}|| \leq O(L^{1/3} R^{2/3} (\epsilon(R) + O(1/\sqrt{n}))^{1/3}).$$

This result has the following consequences for the expected variation of the fine-grained localization error as a function of the parameters n (the number of base sensors), R (the size of the discs), and k^2 (the number of discs):

1. The fine-grained localization error decreases as sensor network becomes more densely distributed (i.e. n increases). In addition, the localization error increases with the size of the network, but this increase is at most linear.

2. The fine-grained localization error increases as R increases; on the other hand, as R increases, the optimal value of k decreases, resulting in a smaller computational cost, because there are k^2 discs to classify. Hence, variation in R induces a tradeoff between localization accuracy and computational complexity.

3. We would expect the localization error to increase at a rate $O(R^{2/3})$ if $\epsilon(R)$ were to remain constant. However, as R increases, the length of the boundary of the regions C_β also increases, and the training error $\epsilon(R)$ is expected to increase as well. As a result, we expect the localization error to actually increase faster than $O(R^{2/3})$.

4. The training error $\epsilon(R)$ depends on the distribution of sensor location, but which is unknown.

Note that our analysis makes some simplifying assumptions—it assumes a uniform distribution for the locations of regions C_β and it assumes circular shapes. While the analysis can be readily generalized to other specific choices, it would be of substantial interest to develop a general optimization-theoretic approach to the problem of choosing the regions.

2.4 Algorithm details and computational cost

During the training phase associated with each coarse localization subroutine, i.e., classification with respect to a fixed region C_β, we construct the training data set based on the locations of the base sensors as described in the previous section. This is achieved by having all base stations send the signal matrix entries $s(x_i, x_j)$ and their known locations to a central station, a procedure which involves receiving and storing $n^2 + n$ numbers at the central station. The central station then solves the following optimization problem:

$$\min_{\mathrm{w}} ||\mathrm{w}||^2 + \frac{c}{n}\sum_{i=1}^{n} \phi(y_i f(x_i)),$$

where $f(x) = \langle \mathrm{w}, \Phi(x)\rangle$, $\phi(yf(x)) = (1 - yf(x))_+$, and c is a fixed parameter.[6] This is a convex optimization problem, which has the following dual form [Schölkopf and Smola, 2002]:

$$\max_{0\le\alpha\le c} 2\sum_{i=1}^{n} \alpha_i - \sum_{1\le i,j\le n} \alpha_i\alpha_j y_i y_j K(x_i, x_j). \tag{2.4}$$

The algorithm finds optimizing values of $\{\alpha_i\}$, which are then used to form the discriminant function in Eq. (2.2).

[6] The parameter c is a regularization parameter associated with the SVM algorithm. In all our experiments, we fix $c = 10$.

```
Coarse localization algorithm
Input: $X_i = x_i \in \mathbb{R}^2$ for $i = 1, \dots, n$; signal matrix $[s(x_i, x_j)]_{1 \le i,j \le m}$ where $n \ll m$; a region $C \subseteq \mathbb{R}^2$.
Output: $y_j \in \{\pm 1\}$ for $j = n+1, \dots, m$.

1. For $i = 1, \dots, n$, let $y_i = \mathrm{sign}(x_i \in C)$.
2. Define a positive semidefinite kernel matrix $[K(x_i, x_j)]_{1 \le i,j \le m}$ based upon $[s(x_i, x_j)]_{ij}$.
3. Solve the optimization problem (2.4) for optimum $\{\alpha_i\}_{i=1}^n$.
4. For $j = n+1, \dots, m$, $y_j = \mathrm{sign}\left(\sum_{i=1}^n \alpha_i K(x_i, x_j)\right)$.
```

Figure 2.1: Summary of the coarse localization algorithm.

It is known that the solution to this optimization problem can be found in the worst case in $O(n^3)$ computational time. Thus if there are k^2 regions to classify, this result suggests a total training time of $O(n^3 k^2)$. However, this generic worst-case estimate is overly conservative in our setting. Indeed, an estimate based on the number of support vectors n_s returned by each classification algorithm (those X_i such that $\alpha_i \neq 0$) reveals that the computational complexity is $O(n_s^3 + n_s^2 n)$ instead of $O(n^3)$. Usually $n_s \ll n$. Our simulation experience (to be presented in the next section) shows that when discs with radius R are used, the support vectors reside mostly along the boundaries of the discs, hence $n_s \approx O(\min(n\pi R^2/L^2, 2\pi R))$, in which case the overall training phase takes only $O(R^2 n k^2)$ time. Note also that this training phase is the most expensive part of our algorithm and is performed at a central station.

Once the training phase is complete, each base sensor is required to store the n parameters $(\alpha_1, \dots, \alpha_n)$ for the purpose of classification of the remaining (location-unknown) sensors. If the first-tier kernel is used, a new sensor X_j for $j = n+1, \dots, m$ records the signal $s(x_i, x_j)$ from the n_s base sensors $i \in \{1, \dots, n\}$, and combines these with the non-zero values α_i, resulting in a cost of $O(n_s)$ in time and storage. If a second-tier linear kernel or a third-tier Gaussian kernel is used, a new sensor X_j records n-element signal vectors $(s(x_j, x_1), \dots, s(x_j, x_n))$ from the n_s base stations, resulting in a $O(n_s n)$ cost in time and storage. The kernel values $K(x_i, x_j)$ are then readily computable from the received signals $s(x_i, x_j)$ in $O(1), O(n), O(n^2)$ time for the first-tier, second-tier and third-tier kernel, respectively. Then a simple computation (Equation 2.3) determines for sensor X_j whether it resides in the region C or not. The attractive feature of the localizing step is that it is done locally (in a distributed fashion), taking only linear (for the first-tier and second-tier kernels) or quadratic (for the third-tier Gaussian kernel) time and storage space (in terms of n). Since the localization is done on an as-needed basis, its time and storage cost do not depend on the total number of sensors m in the network. A summary of our algorithm is provided in Figure 2.1.

Now we turn to fine-grained localization. At both algorithmic and system levels, this involves invoking the coarse localization subroutine k^2 times with respect to regions $C_1, \ldots, C_{k^2}$. Therefore, for each region $\beta = 1, \ldots, k^2$ we have a set of parameters $(\alpha_i)_{i=1}^n$. Each sensor X_j can then determine its location by setting x_j to be the centroid of the intersection of all regions C_β that it finds itself residing in. In the case in which C_β are discs with centers c_β, this yields:

$$x_j := \frac{\sum_{\beta=1}^{k^2} c_\beta \mathbb{I}(X_j \in C_\beta)}{\sum_{\beta=1}^{k^2} \mathbb{I}(X_j \in C_\beta)}.$$

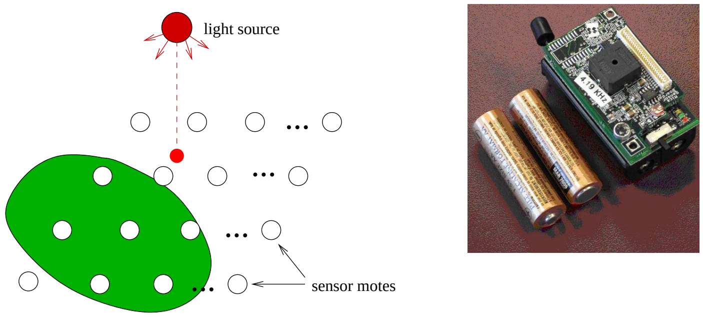
Clearly, the computational cost of a fine-grained localization algorithm is k^2 times as much as the computational cost of each coarse localization step. In summary, our fine-grained localization algorithm is shown in Figure 2.2.

```
Fine-grained localization algorithm
Input: $X_i = x_i \in \mathbb{R}^2$ for $i = 1, \ldots, n$; signal matrix $[s(x_i, x_j)]_{1 \leq i,j \leq m}$ where $n \ll m$; $k$; $R$.
Output: $x_j \in \mathbb{R}^2$ for $j = n+1, \ldots, m$.

1. Let $A_1 = \min\{(x_i)_1\}_{i=1}^n$; $B_1 = \max\{(x_i)_1\}_{i=1}^n$; $A_2 = \min\{(x_i)_2\}_{i=1}^n$;
   $B_2 = \max\{(x_i)_2\}_{i=1}^n$.
2. Let $C_\beta$ for $\beta = 1, \ldots, k^2$ be $k^2$ discs with radius $R$ distributed uniformly
   in a grid of coordinates $[A_1, B_1] \times [A_2, B_2]$.
3. For $\beta = 1, \ldots, k^2$, run the coarse localization algorithm with respect to
   region $C_\beta$ to get values $\{y_{\beta,j}\}_{j=n+1}^m$.
4. Letting $c_\beta$ be the center of $C_\beta$ for $\beta = 1 \ldots, k^2$, then:
   $x_j = \frac{\sum_{\beta=1}^{k^2} c_\beta \mathbb{I}(y_{\beta,j}=1)}{\sum_{\beta=1}^{k^2} \mathbb{I}(y_{\beta,j}=1)}$.
```

Figure 2.2: Summary of the fine-grained localization algorithm.

2.5 Experimental Results



(a) (b)

Figure 2.3: (a) Illustration of a sensor field. (b) a Mica sensor mote.

We evaluate our algorithm on simulated sensor networks in the first two subsections, and then on a real network using Berkeley sensor motes.

2.5.1 Coarse localization

Simulation set-up:

We consider a network of size 10×10 square units. The base sensors are distributed uniformly in a grid-like structure. There are a total of n such sensors. We are concerned with recognizing whether a sensor position x, characterized by the signal reading $s(x_i, x)$ for $i = 1, \ldots, n$, lies in a region C or not.

We first define a signal model: Each sensor location x is assumed to receive from a sensor located at x' a signal value following a fading channel model: $s(x, x') = \exp -\frac{||x-x'||^2}{\sigma} + N(0, \tau)$, where $N(0, \tau)$ denotes an independently generated normal random variable with standard deviation τ. This signal model is a randomized version of a Gaussian kernel. We have also experimented with a signal strength model that is a randomized version of the polynomial kernel: $s(x, x') = (||x - x'||)^{-\sigma} + N(0, \tau)$. The results for the polynomial kernels are similar to the Gaussian kernels, and are not presented here. It is emphasized that although the use of these models have been motivated elsewhere as signal models [Seidel and Rappaport, 1992; Sheng and Hu, 2003], in our case they are used merely to generate the signal matrix S. Our algorithm is not provided with any knowledge of the procedure that generates S.

Next, we define a region C to be recognized. In particular, C consists of all locations x that satisfy the following equations: $(x - v)^T H_1 (x - v) \leq R$ and $(x - v)^T H_2 (x - v) \leq R$,

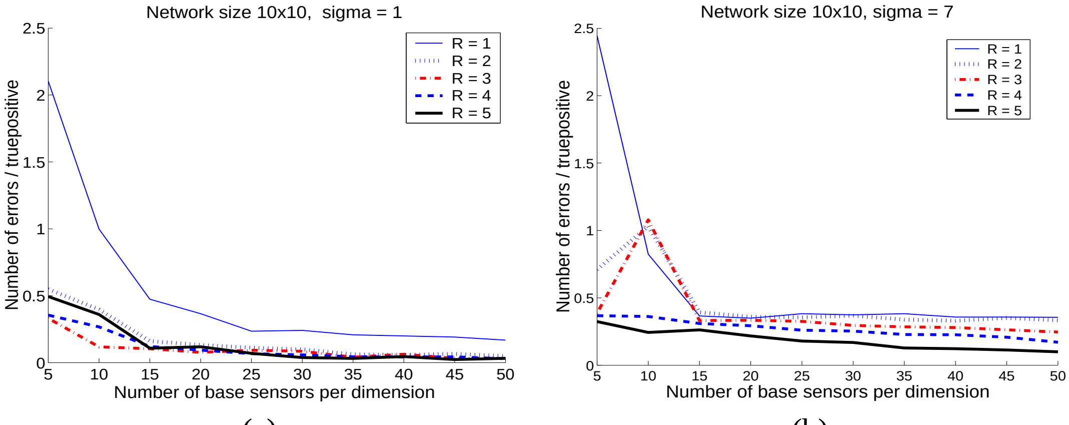


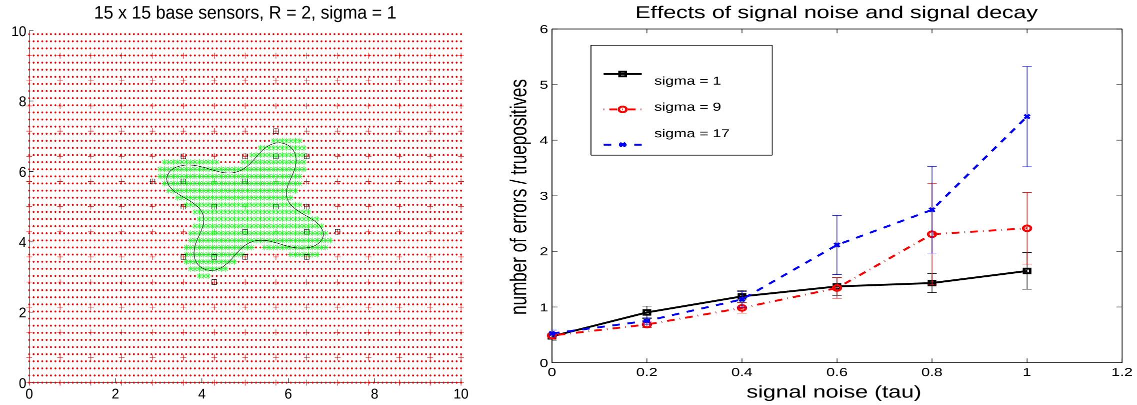
Figure 2.4. Simulation results with (randomized) Gaussian models. The x-axis shows the number of sensors employed along each dimension of the network. The y-axis shows the ratio between the number of incorrectly classified points and the number of points inside the area to be recognized. (Note that this ratio is larger than the overall failure rate; in the latter the denominator includes the points outside the area to be recognized).

where $v = [5\ 5]^T$, $H_1 = [2\ -1; -1\ 1]$ and $H_2 = [2\ 1; 1\ 1]$. The radius R is used to describe the size of C. For each simulation set-up (n, R, σ, τ), we learn a discriminant function f for the region C using the training data given by the base sensor positions. Once f is learned, we test the classification at 100×100 sensor locations distributed uniformly in the region containing the network.

Figure 2.5(a) illustrates C as a green shaded region, for $R = 2$, while the black boundary represents the region learned by our localization algorithm. Qualitatively, the algorithm has captured the shape of the target region C. We now present a quantitative analysis on the effects of n, R, σ, and τ on the localization (i.e., classification) performance:

Effects of n: The plots in Figure 2.4 show the localization (test) error with respect to the number of base sensors deployed in the network. The test error is defined to be the ratio between the number of misclassified points and the number of points located within the area $C(R)$ (out of 100×100 locations distributed uniformly in the grid). In this set of simulations, we fix the noise parameter $\tau = 0$, and let $\sigma = 1$ and $\sigma = 7$, while varying n. The plots confirm that the localization error tends to decrease as the sensor network becomes more densely distributed. Note that if we need to recognize a particular area, we only need to plant base sensors in the area near the boundary, because these are the likely locations of support vectors. Of course, in our context coarse-grained localization is only a subroutine for fine-grained localization, and it is our interest to have base sensors spread throughout the whole geographical area.

Effects of σ and τ: The parameter σ is used to describe the sensitivity of the signal strength with respect to the sensor distance. In particular, for a Gaussian signal function, a



(a) (b)

Figure 2.5. (a) Illustration of a simulated sensor network with 15×15 base sensors, and the recognized boundary in black (with $R = 2$) using a Gaussian kernel with $\sigma = 1$. The black squares are the support vector base sensors. The test error in this figure is 0.27. (b) Plots show the effect of the sensor fading signal parameter σ and signal noise parameter τ on coarse localization performance.

small value of σ implies that the signal strength fades very quickly for distant sensors. The plots in Figure 2.5(b) display the effects of both σ and τ on the localization performance. In this set of simulations, we fix the number of base sensors along each dimension to be 10, and set the radius of C to be $R = 2$, while varying σ and τ. It is seen that the localization performance degrades as we increase the noise parameter τ, and the degradation is more severe for the least sensitive signal, i.e., when σ is large.

2.5.2 Fine-grained localization

Simulation set-up: The network set-up is the same as the previous section, except that the n base sensors are now distributed approximately uniformly at random in the whole area. By this we mean that each base sensor is initially planted at a grid point in the $L \times L$ square, where $L = 10$, and then perturbed by Gaussian noise $N(0, L/(2\sqrt{n}))$. There are 400 other sensors whose locations are to be determined using our algorithm. These 400 sensors are deployed uniformly in the grid. Again, we assume the signal strength follows a Gaussian signal model, with noise parameter $\tau = 0.2$.

We applied the algorithm described in Section 2.3 for fine-grained localization. The algorithm involves repeated coarse localizations with respect to a set of regions that cover the whole network area. We choose these regions to be discs of radius R, and distributed uniformly over the network. Let k be the number of discs along each dimension, such that there are a total k^2 discs to recognize. In this simulation we study the effects of R, k and the number of base sensors n on the localization performance. Specifically, we examine the

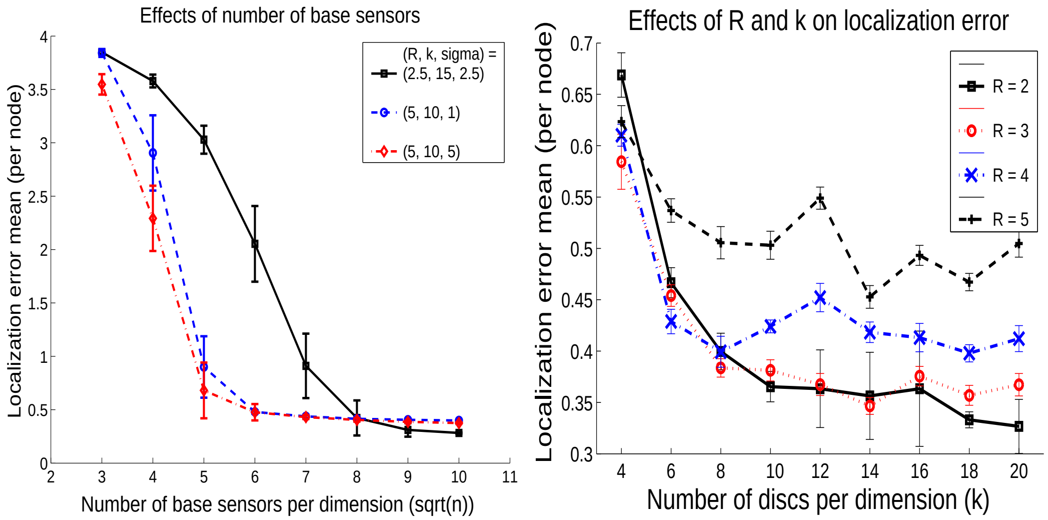


Figure 2.6. The left panel shows the effect of the number of base sensors n on fine-grained localization error mean and standard deviation (for all nodes). The right panel shows the effects of the size of discs (by radius R) and the number of discs (k^2) distributed uniformly on the field. The means and variances are collected after performing the simulation on 20 randomly generated sensor networks.

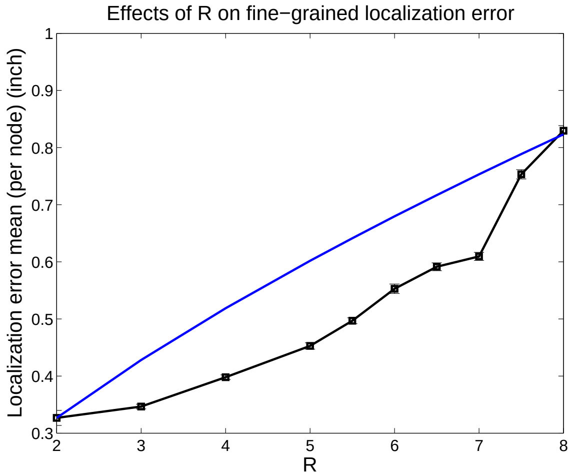
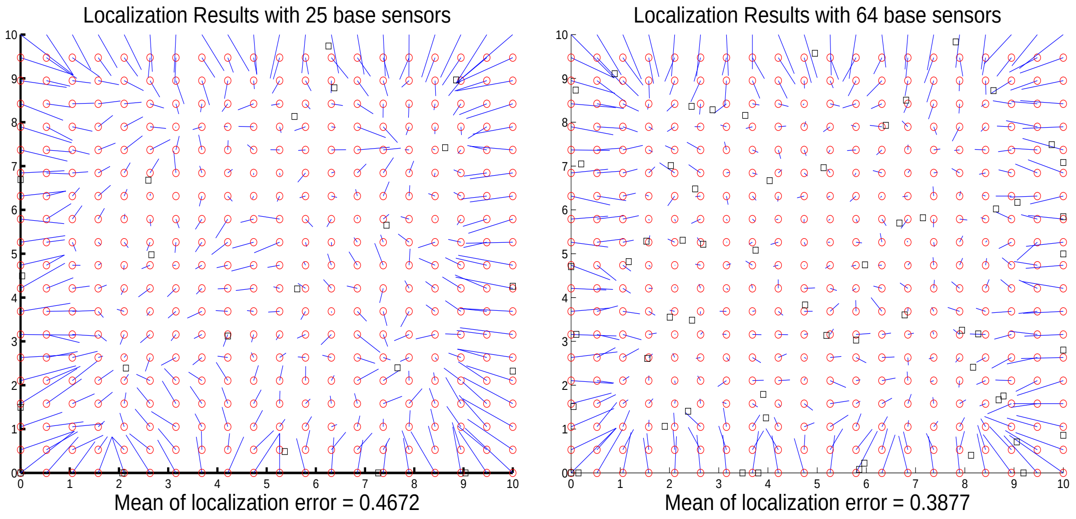


Figure 2.7. This figure shows the effects of the size of discs (R) on the fine-grained localization error. The number of disks (k^2) is chosen so that the mean localization error (per node) is smallest. The error rate is compared with the curve $O(R^{2/3})$ plotted in blue.



(a)

(b)

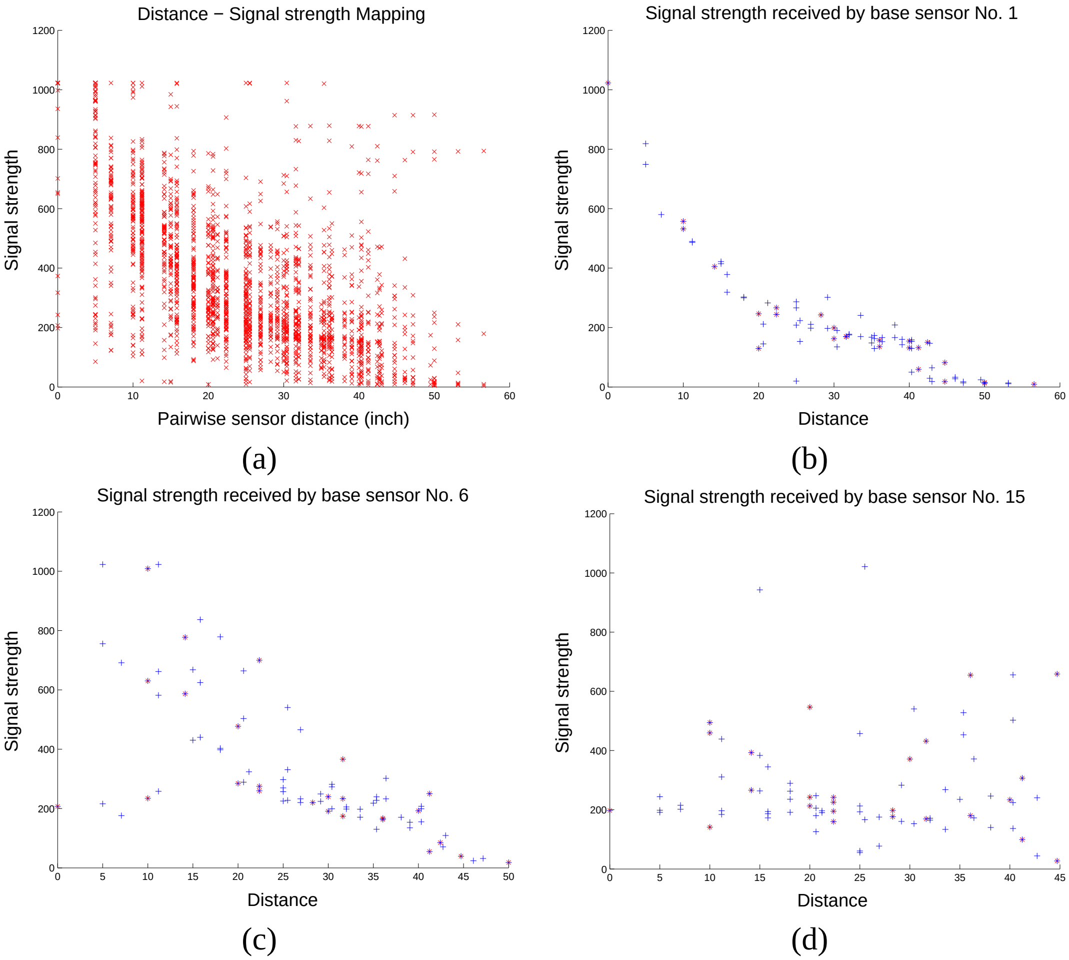
Figure 2.8. Localization results for a simulated sensor network of size 10×10 square units with 25 base sensors (left figure) and 64 base sensors (right figure). The base sensors are the black squares. Each blue line connects a true sensor position (in circle) and its estimate. The signal model is Gaussian. The mean localization error is 0.4672 in the left figure and 0.3877 in the right figure.

tradeoff between computational cost and localization accuracy of our algorithm by varying these parameters, as suggested by the theoretical analysis in Section 2.3.3.

Effects of n: Figure 2.6(a) shows that the mean localization error (averaged over all sensor networks and over all sensors) decreases monotonically as more base sensors are added to the network. This agrees with the theoretical result presented in Section 2.3.3. Figure 2.8 illustrates the localization results for each node in the networks with 25 and 64 base sensors. The mean localization error (averaging over all sensors) for these two networks are 0.47 and 0.39, respectively.

Effects of R and k: Figure 2.6(b) shows the effects of R and k on the localization performance. In this set of simulations, we fix $\tau = 0.2$, $\sigma = 2$, and $n = 100$, while varying R and k. The analysis in Section 2.3.3 suggests that for each value of R, there exists an optimal value for k that increases as R decreases. Since there are k^2 classification problems to solve, the computational cost generally increases as R decreases. However, the mean localization error improves as R decreases. Hence, there is a tradeoff between computational cost and localization accuracy as manifested by the behavior of R and k.

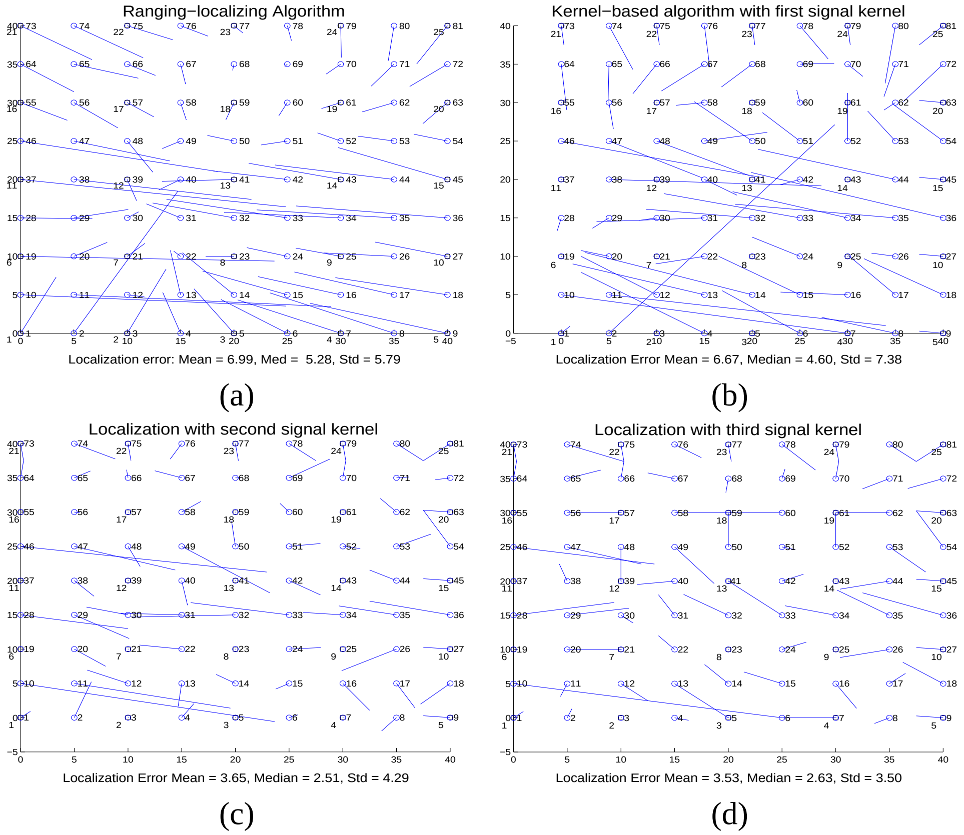
To gain more insight of the effects of the size of discs (R) on the fine-grained localization error, in Figure 2.7, we plot the mean localization error for the optimal value of k. This figure shows that the optimal mean localization error increases as R increases. We also compare the rate of increase with that of $R^{2/3}$. As shown in Figure 2.7 the rate is



(a) (b)

(c) (d)

Figure 2.9. Panel (a) shows the noisy relationship between signal strength received by sensors and the distances. Few sensors exhibit a clear signal strength-distance functional pattern as in panel (b), while most are like those in panels (c) and (d). Note that only data points marked with x in red are available for regression training.


Ranging−localizing Algorithm

Localization error: Mean = 6.99, Med = 5.28, Std = 5.79

(a)

Kernel−based algorithm with first signal kernel

Localization Error Mean = 6.67, Median = 4.60, Std = 7.38

(b)

Localization with second signal kernel

Localization Error Mean = 3.65, Median = 2.51, Std = 4.29

(c)

Localization with third signal kernel

Localization Error Mean = 3.53, Median = 2.63, Std = 3.50

(d)

Figure 2.10. Localization result for a real sensor networks covering a 40×40 square-inch area. There are 25 base sensors (Berkeley motes) spaced in a 5 × 5 grid. Each line connects a true position (in circle) and its estimate. Panel (a) shows the results given by a traditional 2-step localization algorithm, while panels (b,c,d) show the localization results obtained by our algorithm using three different kernels (the first-tier, second-tier and third-tier Gaussian kernel, respectively).

approximately that of $R^{2/3}$ in a middle range and eventually surpasses $R^{2/3}$. Recall from the analysis in Section 2.3.3 that we expect this increase in rate due to the increase in $\epsilon(R)$. On the other hand, the analysis does not predict the smaller rate of increase observed for small values of R.

2.5.3 Localization with Berkeley sensor motes

Experiment set-up: We evaluated our algorithm on a real sensor network using Berkeley tiny sensor motes (Mica motes) as the base stations. The goal of the experiment is to estimate the positions of light sources given the light signal strength received by a number of base sensors deployed in the network. Our hardware platform consists of 25 base sensors

Method	Mean	Median	Std
Two-step ranging-based	6.99	5.28	5.79
First-tier signal kernel	6.67	4.60	7.38
Second-tier linear kernel	3.65	2.51	4.29
Third-tier Gaussian kernel	3.53	2.63	3.50

Table 2.1. Comparison between a two-step ranging-based algorithm and our kernel-based localization algorithm in a sensor network with 25 base sensors covering a 40×40 square-inch area. The localization error mean, median and standard deviation are taken over all position estimates, and measured in inches.

placed 10 inches apart on a 5×5 grid in a flat indoor environment. Each sensor mote is composed of one Atmel ATmega 103 8-bit processor running at 4MHz, with 128Kb of flash and 4Kb of RAM, RFM TR1000 radio, EEprom and a sensor board which includes light, temperature, microphone sensors and a sounder. Our experiment makes use of light sensor data received by the motes. The measured signals are a scalar field produced by a light source shining on the sensor network from above; the height and intensity of the light source were constant. Only the position of light sources placed at the base sensors are given as training data. To be estimated are 81 light source positions distributed uniformly in a 9×9 grid spread over the whole network.

A range-based algorithm: We compared our algorithm to a state-of-the-art algorithm that epitomizes a majority of localization algorithms in the literature. This algorithm was described in [Whitehouse, 2002], and consists of two main steps: (1) a ranging procedure aimed at establishing a mapping between the signal strength received by a base sensor and the distance to the light source, and (2) a localization procedure giving the distance estimates using least-squares methods.

Figure 2.9 illustrates the difficulty of the ranging problem—the functional relationship between distances and signal strengths is very noisy. Much of this noise is device-specific; as shown in Figure 2.9, a few sensors exhibit a clear distance-to-signal-strength pattern, while most others exhibit a very noisy pattern. As shown in [Whitehouse, 2002], improvement in the ranging step can be achieved by accounting for properties of specific base sensors. This is done by introducing regression coefficients for each of these base sensors. Once the ranging step is completed, we have estimates of the distance between the base sensors and the positions of the light source. The initial position estimates are obtained using the Bounding-Box algorithm, and are then iteratively updated using a least-squares method (see [Whitehouse, 2002; Savvides *et al.*, 2001]). Figure 2.5.1(a) shows the localization results for this algorithm.

Results for the kernel-based algorithm: Three different kernels are used in our algorithm. The first is a first-tier symmetric positive semidefinite approximation of the signal matrix. In particular, as discussed in Section 2.3, given a signal matrix S, we define

$S' := (S + S^T)/2 + \delta I$. The remaining kernels are a second-tier linear and third-tier Gaussian kernel, with the parameter σ fixed to 0.5 in the latter case. For fine-grained localization, coarse localization is repeatedly applied for discs of radius $R = L/2 = 20$ inches that cover part of the network area. The centers of these discs are five inches apart in both dimensions, and there are 10 discs along each dimension (i.e., $k = 10$).

Table 2.1 shows that the localization error achieved by the kernel-based approach is smaller than that of the two-step algorithm. Among the three choices of signal kernels, the second-tier kernels are much better than the simple first-tier kernel. The localization results are depicted spatially in Figure 2.5.1. Note that the minimum distance between two neighboring base sensors is about 10 inches, and the localization error of our algorithm (using second-tier kernels) is slightly over one third of that distance.

2.6 Discussions

We have presented a nonparametric learning algorithm for coarse-grained and fine-grained localization for ad hoc wireless sensor networks. Our approach treats the signal strength as measured by sensor motes as a natural coordinate system in which to deploy statistical classification and regression methods. For the localization problem, this approach avoids the ranging computation, a computation which requires accurate signal models that are difficult to calibrate. Instead, we use signal strength either directly to define basis functions for kernel-based classification algorithms, or indirectly via derived kernels that operate on top of the signal strength measurements. We show how a kernel-based classification algorithm can be invoked multiple times to achieve accurate localization results, and we present an error analysis for the accuracy that can be achieved as a function of base sensor density. Our algorithm is particularly suitable for densely distributed sensor networks, and is appealing for its computational scaling in such networks: The preprocessing computations are performed at the base sensors, which are assumed to have sufficient processing and power capability, while the localizing step at location-unknown sensors can be achieved in linear time.

We have argued for a simple approach to localization that dispenses with ranging computations and sensor modeling. We do not necessarily believe, however, that our statistical approach is always to be preferred. In particular, the level of accuracy that we appear to be able to obtain with our approach is on the order of one third the distance between the motes. While this accuracy is sufficient for many potential applications of sensor networks, in some applications higher accuracy may be required. In this case, ranging-based approaches offer an alternative, but only in the setting in which highly accurate models of the relationship between sensor signals and distances are available.

Chapter 3

Nonparametric decentralized detection using kernel methods

We consider the problem of decentralized detection under constraints on the number of bits that can be transmitted by each sensor. In contrast to most previous work, in which the joint distribution of sensor observations is assumed to be known, we address the problem when only a set of empirical samples is available. We propose a nonparametric approach using the framework of empirical risk minimization and marginalized kernels, and analyze its computational and statistical properties both theoretically and empirically. We provide a computationally efficient algorithm, and demonstrate its performance on both simulated and real data sets.[1]

3.1 Introduction

A decentralized detection system typically involves a set of sensors that receive observations from the environment, but are permitted to transmit only a summary message (as opposed to the full observation) back to a fusion center. On the basis of its received messages, this fusion center then chooses a final decision from some number of alternative hypotheses about the environment. The problem of decentralized detection is to design the local decision rules at each sensor, which determine the messages that are relayed to the fusion center, as well a decision rule for the fusion center itself [Tsitsiklis, 1993b]. A key aspect of the problem is the presence of *communication constraints*, meaning that the sizes of the messages sent by the sensors back to the fusion center must be suitably "small" relative to the raw observations, whether measured in terms of either bits or power. The *decentralized* nature of the system is to be contrasted with a centralized system, in which the fusion center has access to the full collection of raw observations.

[1] This chapter has been published in [Nguyen *et al.*, 2005b].

Such problems of decentralized decision-making have been the focus of considerable research in the past two decades [Tenney and Sandell, 1981; Tsitsiklis, 1993b; Blum *et al.*, 1997; Chamberland and Veeravalli, 2003]. Indeed, decentralized systems arise in a variety of important applications, ranging from sensor networks, in which each sensor operates under severe power or bandwidth constraints, to the modeling of human decision-making, in which high-level executive decisions are frequently based on lower-level summaries. The large majority of the literature is based on the assumption that the probability distributions of the sensor observations lie within some known parametric family (e.g., Gaussian and conditionally independent), and seek to characterize the structure of optimal decision rules. The probability of error is the most common performance criterion, but there has also been a significant amount of work devoted to other criteria, such as criteria based on Neyman-Pearson or minimax formulations. See Tsitsiklis [Tsitsiklis, 1993b] and Blum et al. [Blum *et al.*, 1997] for comprehensive surveys of the literature.

More concretely, let $Y \in \{-1, +1\}$ be a random variable, representing the two possible hypotheses in a binary hypothesis-testing problem. Moreover, suppose that the system consists of S sensors, each of which observes a single component of the S-dimensional vector $X = \{X^1, \ldots, X^S\}$. One starting point is to assume that the joint distribution $P(X, Y)$ falls within some parametric family. Of course, such an assumption raises the modeling issue of how to determine an appropriate parametric family, and how to estimate parameters. Both of these problems are very challenging in contexts such as sensor networks, given highly inhomogeneous distributions and a large number S of sensors. Our focus in this chapter is on relaxing this assumption, and developing a method in which no assumption about the joint distribution $P(X, Y)$ is required. Instead, we posit that a number of empirical samples $(x_i, y_i)_{i=1}^n$ are given.

In the context of *centralized* signal detection problems, there is an extensive line of research on nonparametric techniques, in which no specific parametric form for the joint distribution $P(X, Y)$ is assumed (see, e.g., Kassam [Kassam, 1993] for a survey). In the decentralized setting, however, it is only relatively recently that nonparametric methods for detection have been explored. Several authors have taken classical nonparametric methods from the centralized setting, and shown how they can also be applied in a decentralized system. Such methods include schemes based on Wilcoxon signed-rank test statistic [Viswanathan and Ansari, 1989; Nasipuri and Tantaratana, 1997], as well as the sign detector and its extensions [Han *et al.*, 1990; Al-Ibrahim and Varshney, 1989; Hussaini *et al.*, 1995]. These methods have been shown to be quite effective for certain types of joint distributions.

Our approach to decentralized detection in this chapter is based on a combination of ideas from *reproducing-kernel Hilbert spaces* [Aronszajn, 1950; Saitoh, 1988], and the framework of *empirical risk minimization* from nonparametric statistics. Methods based on reproducing-kernel Hilbert spaces (RKHSs) have figured prominently in the literature on centralized signal detection and estimation for several decades [Weinert, 1982;

Kailath, 1971, e.g.,]. More recent work in statistical machine learning [Schölkopf and Smola, 2002, e.g.,] has demonstrated the power and versatility of kernel methods for solving classification or regression problems on the basis of empirical data samples. Roughly speaking, kernel-based algorithms in statistical machine learning involve choosing a function, which though linear in the RKHS, induces a nonlinear function in the original space of observations. A key idea is to base the choice of this function on the minimization of a *regularized empirical risk* functional. This functional consists of the empirical expectation of a convex loss function ϕ, which represents an upper bound on the 0-1 loss (the 0-1 loss corresponds to the probability of error criterion), combined with a regularization term that restricts the optimization to a convex subset of the RKHS. It has been shown that suitable choices of margin-based convex loss functions lead to algorithms that are robust both computationally [Schölkopf and Smola, 2002], as well as statistically [Zhang, 2004; Bartlett *et al.*, 2006]. The use of kernels in such empirical loss functions greatly increases their flexibility, so that they can adapt to a wide range of underlying joint distributions.

In this chapter, we show how kernel-based methods and empirical risk minimization are naturally suited to the decentralized detection problem. More specifically, a key component of the methodology that we propose involves the notion of a *marginalized kernel*, where the marginalization is induced by the transformation from the observations X to the local decisions Z. The decision rules at each sensor, which can be either probabilistic or deterministic, are defined by conditional probability distributions of the form $Q(Z|X)$, while the decision at the fusion center is defined in terms of $Q(Z|X)$ and a linear function over the corresponding RKHS. We develop and analyze an algorithm for optimizing the design of these decision rules. It is interesting to note that this algorithm is similar in spirit to a suite of *locally optimum* detectors in the literature [Blum *et al.*, 1997, e.g.,], in the sense that one step consists of optimizing the decision rule at a given sensor while fixing the decision rules of the rest, whereas another step involves optimizing the decision rule of the fusion center while holding fixed the local decision rules at each sensor. Our development relies heavily on the convexity of the loss function ϕ, which allows us to leverage results from convex analysis [Rockafellar, 1970] so as to derive an efficient optimization procedure. In addition, we analyze the statistical properties of our algorithm, and provide probabilistic bounds on its performance.

While the thrust of this chapter is to explore the utility of recently-developed ideas from statistical machine learning for distributed decision-making, our results also have implications for machine learning. In particular, it is worth noting that most of the machine learning literature on classification is abstracted away from considerations of an underlying communication-theoretic infrastructure. Such limitations may prevent an algorithm from aggregating all relevant data at a central site. Therefore, the general approach described in this chapter suggests interesting research directions for machine learning—specifically, in designing and analyzing algorithms for communication-constrained environments.[2]

The remainder of the chapter is organized as follows. In Section 3.2, we provide a formal statement of the decentralized decision-making problem, and show how it can be cast as a learning problem. In Section 3.3, we present a kernel-based algorithm for solving the problem, and we also derive bounds on the performance of this algorithm. Section 3.4 is devoted to the results of experiments using our algorithm, in application to both simulated and real data. Finally, we conclude the chapter with a discussion of future directions in Section 3.5.

3.2 Problem formulation and a simple strategy

In this section, we begin by providing a precise formulation of the decentralized detection problem to be investigated in this chapter, and show how it can be cast in a statistical learning framework. We then describe a simple strategy for designing local decision rules, based on an optimization problem involving the empirical risk. This strategy, though naive, provides intuition for our subsequent development based on kernel methods.

3.2.1 Formulation of the decentralized detection problem

Suppose Y is a discrete-valued random variable, representing a hypothesis about the environment. Although the methods that we describe are more generally applicable, the focus of this chapter is the binary case, in which the hypothesis variable Y takes values in $\mathcal{Y} := \{-1, +1\}$. Our goal is to form an estimate $\widehat{Y}$ of the true hypothesis, based on observations collected from a set of S sensors. More specifically, for each $t = 1, \ldots, S$, let $X^t \in \mathcal{X}$ represent the observation at sensor t, where $\mathcal{X}$ denotes the observation space. The full set of observations corresponds to the S-dimensional random vector $X = (X^1, \ldots, X^S) \in \mathcal{X}^S$, drawn from the conditional distribution $P(X|Y)$.

We assume that the global estimate $\widehat{Y}$ is to be formed by a *fusion center*. In the *centralized setting*, this fusion center is permitted access to the full vector $X = (X^1, \ldots, X^S)$ of observations. In this case, it is well-known [van Trees, 1990] that optimal decision rules, whether under Bayes error or Neyman-Pearson criteria, can be formulated in terms of the likelihood ratio $P(X|Y = 1)/P(X|Y = -1)$. In contrast, the defining feature of the *decentralized setting* is that the fusion center has access only to some form of summary of each observation X^t, for $t = 1, \ldots S$. More specifically, we suppose that each sensor $t = 1 \ldots, S$ is permitted to transmit a *message* Z^t, taking values in some space $\mathcal{Z}$. The fusion center, in turn, applies some decision rule γ to compute an estimate $\widehat{Y} = \gamma(Z^1, \ldots, Z^S)$ of Y based on its received messages.

[2] For a related problem of distributed learning under communication constraints and its analysis, see a recent paper by Predd et al. [Predd *et al.*, 2004].

In this chapter, we focus on the case of a discrete observation space—say $\mathcal{X} = \{1, 2, \ldots, M\}$. The key constraint, giving rise to the decentralized nature of the problem, is that the corresponding message space $\mathcal{Z} = \{1, \ldots, L\}$ is considerably smaller than the observation space (i.e., $L \ll M$). The problem is to find, for each sensor $t = 1, \ldots, S$, a decision rule $\gamma^t : \mathcal{X}^t \to \mathcal{Z}^t$, as well as an overall decision rule $\gamma : \mathcal{Z}^S \to \{-1, +1\}$ at the fusion center so as to minimize the *Bayes risk* $P(Y \neq \gamma(Z))$. We assume that the joint distribution $P(X, Y)$ is unknown, but that we are given n independent and identically distributed (i.i.d.) data points $(x_i, y_i)_{i=1}^n$ sampled from $P(X, Y)$.

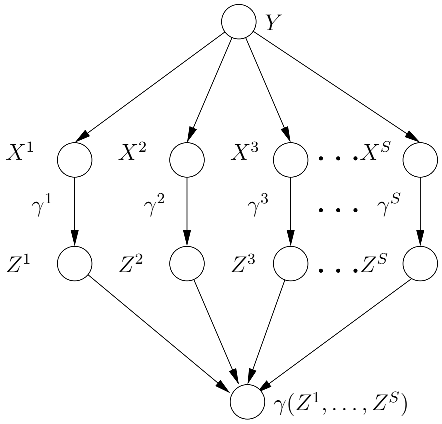


Figure 3.1. Decentralized detection system with S sensors, in which Y is the unknown hypothesis, $X = (X^1, \ldots, X^S)$ is the vector of sensor observations; and $Z = (Z^1, \ldots, Z^S)$ are the quantized messages transmitted from sensors to the fusion center.

Figure 3.1 provides a graphical representation of this decentralized detection problem. The single node at the top of the figure represents the hypothesis variable Y, and the outgoing arrows point to the collection of observations $X = (X^1, \ldots, X^S)$. The local decision rules γ^t lie on the edges between sensor observations X^t and messages Z^t. Finally, the node at the bottom is the fusion center, which collects all the messages.

Although the Bayes-optimal risk can always be achieved by a deterministic decision rule [Tsitsiklis, 1993b], considering the larger space of stochastic decision rules confers some important advantages. First, such a space can be compactly represented and parameterized, and prior knowledge can be incorporated. Second, the optimal deterministic rules are often very hard to compute, and a probabilistic rule may provide a reasonable approximation in practice. Accordingly, we represent the rule for the sensors $t = 1, \ldots, S$ by a conditional probability distribution $Q(Z|X)$. The fusion center makes its decision by applying a deterministic function $\gamma(z)$ of z. The overall decision rule (Q, γ) consists of the individual sensor rules and the fusion center rule.

The decentralization requirement for our detection/classification system—i.e., that the decision or quantization rule for sensor t must be a function only of the observation x^t—can be translated into the probabilistic statement that $Z^1, \ldots, Z^S$ be conditionally independent given X:

$$Q(Z|X) = \prod_{t=1}^{S} Q^t(Z^t|X^t). \tag{3.1}$$

In fact, this constraint turns out to be advantageous from a computational perspective, as will be clarified in the sequel. We use $\mathcal{Q}$ to denote the space of all factorized conditional distributions $Q(Z|X)$, and $\mathcal{Q}_0$ to denote the subset of factorized conditional distributions that are also deterministic.

3.2.2 A simple strategy based on minimizing empirical risk

Suppose that we have as our training data n pairs (x_i, y_i) for $i = 1, \ldots, n$. Note that each x_i, as a particular realization of the random vector X, is an S dimensional signal vector $x_i = (x_i^1, \ldots, x_i^S) \in \mathcal{X}^S$. Let P be the unknown underlying probability distribution for (X, Y). The probabilistic set-up makes it simple to estimate the Bayes risk, which is to be minimized.

Consider a collection of local quantization rules made at the sensors, which we denote by $Q(Z|X)$. For each such set of rules, the associated Bayes risk is defined by:

$$R_{opt} \quad := \quad \frac{1}{2} - \frac{1}{2}\mathbb{E}\Big|P(Y=1|Z) - P(Y=-1|Z)\Big|. \tag{3.2}$$

Here the expectation $\mathbb{E}$ is with respect to the probability distribution $P(X, Y, Z) := P(X, Y)Q(Z|X)$. It is clear that no decision rule at the fusion center (i.e., having access only to z) has Bayes risk smaller than R_{opt}. In addition, the Bayes risk R_{opt} can be achieved by using the decision function

$$\gamma_{opt}(z) = \text{sign}(P(Y=1|z) - P(Y=-1|z)).$$

It is key to observe that this optimal decision rule *cannot* be computed, because $P(X, Y)$ is not known, and $Q(Z|X)$ is to be determined. Thus, our goal is to determine the rule $Q(Z|X)$ that minimizes an empirical estimate of the Bayes risk based on the training data $(x_i, y_i)_{i=1}^n$. In Lemma 3.1 we show that the following is one such unbiased estimate of the Bayes risk:

$$R_{emp} \quad := \quad \frac{1}{2} - \frac{1}{2n}\sum_{z} \Big|\sum_{i=1}^{n} Q(z|x_i)y_i\Big|. \tag{3.3}$$

In addition, $\gamma_{opt}(z)$ can be estimated by the decision function

$$\gamma_{emp}(z) = \operatorname{sign}\big(\sum_{i=1}^{n} Q(z|x_i)y_i\big).$$

Since Z is a discrete random vector, the following lemma, proved in the Appendix, shows that the optimal Bayes risk can be estimated easily, regardless of whether the input signal X is discrete or continuous:

Lemma 3.1. *(a) If $P(z) > 0$ for all z and $\kappa(z) = \frac{\sum_{i=1}^{n} Q(z|x_i)\mathbb{I}(y_i=1)}{\sum_{i=1}^{n} Q(z|x_i)}$, then $\lim_{n\to\infty} \kappa(z) = P(Y=1|z)$ almost surely.*

(b) As $n \to \infty$, R_{emp} and $\gamma_{emp}(z)$ tend to R_{opt} and $\gamma_{opt}(z)$, respectively, almost surely.

The significance of Lemma 3.1 is in motivating the goal of finding decision rules $Q(Z|X)$ to minimize the empirical error R_{emp}. It is equivalent, using equation (3.3), to maximize

$$C(Q) = \sum_{z} \Big|\sum_{i=1}^{n} Q(z|x_i)y_i\Big|, \tag{3.4}$$

subject to the constraints that define a probability distribution:

$$\begin{cases} Q(z|x) = \prod_{t=1}^{S} Q^t(z^t|x^t) & \text{for all values of } z \text{ and } x, \\ \sum_{z^t} Q^t(z^t|x^t) = 1, \quad \text{and } Q^t(z^t|x^t) \in [0,1] & \text{for } t = 1,\ldots,S. \end{cases} \tag{3.5}$$

The major computational difficulty in the optimization problem defined by Eqs (3.4) and (3.5) lies in the summation over all L^S possible values of $z \in \mathcal{Z}^S$. One way to avoid this obstacle is by maximizing instead the following function:

$$C_2(Q) \quad := \quad \sum_{z} \Big(\sum_{i=1}^{n} Q(z|x_i)y_i\Big)^2.$$

Expanding the square and using the conditional independence condition (3.1) leads to the following equivalent form for C_2:

$$C_2(Q) \quad = \quad \sum_{i,j} y_i y_j \prod_{t=1}^{S} \sum_{z^t=1}^{L} Q^t(z^t|x_i^t) Q^t(z^t|x_j^t). \tag{3.6}$$

Note that the conditional independence condition (3.1) on Q allow us to compute $C_2(Q)$ in $O(SL)$ time, as opposed to $O(L^S)$.

While this simple strategy is based directly on the empirical risk, it does not exploit any prior knowledge about the class of discriminant functions for $\gamma(z)$. As we discuss in the following section, such knowledge can be incorporated into the classifier using kernel methods. Moreover, the kernel-based decentralized detection algorithm that we develop turns out to have an interesting connection to the simple approach based on $C_2(Q)$.

3.3 A kernel-based algorithm

In this section, we turn to methods for decentralized detection based on empirical risk minimization and kernel methods [Aronszajn, 1950; Saitoh, 1988; Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004]. We begin by introducing some background and definitions necessary for subsequent development. We then motivate and describe a central component of our decentralized detection system—namely, the notion of a *marginalized kernel*. Our method for designing decision rules is based on an optimization problem, which we show how to solve efficiently. Finally, we derive theoretical bounds on the performance of our decentralized detection system.

3.3.1 Empirical risk minimization and kernel methods

In this section, we provide some background on empirical risk minimization and kernel methods. We refer the reader to the books [Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004; Saitoh, 1988; Weinert, 1982] for more details. Our starting point is to consider estimating Y with a rule of the form $\widehat{y}(x) = \text{sign} f(x)$, where $f : \mathcal{X} \rightarrow \mathbb{R}$ is a *discriminant function* that lies within some function space to be specified. The ultimate goal is to choose a discriminant function f to minimize the Bayes error $P(Y \neq \widehat{Y})$, or equivalently to minimize the expected value of the following *0-1 loss*:

$$\phi_0(yf(x)) \quad := \quad \mathbb{I}[y \neq \text{sign}(f(x))]. \tag{3.7}$$

This minimization is intractable, both because the function ϕ_0 is not well-behaved (i.e., non-convex and non-differentiable), and because the joint distribution P is unknown. However, since we are given a set of i.i.d. samples $\{(x_i, y_i)\}_{i=1}^n$, it is natural to consider minimizing a loss function based on an *empirical expectation*, as motivated by our development in Section 3.2.2. Moreover, it turns out to be fruitful, for both computational and statistical reasons, to design loss functions based on *convex surrogates* to the 0-1 loss.

Indeed, a variety of classification algorithms in statistical machine learning have been shown to involve loss functions that can be viewed as convex upper bounds on the 0-1 loss. For example, the support vector machine (SVM) algorithm [Schölkopf and Smola, 2002] uses a *hinge loss* function:

$$\phi_1(yf(x)) \quad := \quad (1 - yf(x))_+ \;\equiv\; \max\{1 - yf(x), 0\}. \tag{3.8}$$

On the other hand, the logistic regression algorithm [Friedman *et al.*, 2000] is based on the *logistic loss* function:

$$\phi_2(yf(x)) \quad := \quad \log\left(1 + \exp^{-yf(x)}\right). \tag{3.9}$$

Finally, the standard form of the boosting classification algorithm [Freund and Schapire, 1997] uses a *exponential loss* function:

$$\phi_3(yf(x)) \quad := \quad \exp(-yf(x)). \tag{3.10}$$

Intuition suggests that a function f with small ϕ-risk $\mathbb{E}\phi(Yf(X))$ should also have a small Bayes risk $P(Y \neq \text{sign}(f(X)))$. In fact, it has been established rigorously that convex surrogates for the (non-convex) 0-1 loss function, such as the hinge (3.8) and logistic loss (3.9) functions, have favorable properties both computationally (i.e., algorithmic efficiency), and in a statistical sense (i.e., bounds on both approximation error and estimation error) [Zhang, 2004; Bartlett *et al.*, 2006].

We now turn to consideration of the function class from which the discriminant function f is to be chosen. Kernel-based methods for discrimination entail choosing f from within a function class defined by a positive semidefinite kernel, defined as follows (see [Saitoh, 1988]):

Definition 3.2. *A real-valued kernel function is a symmetric bilinear mapping $K_x : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$. It is positive semidefinite, which means that for any subset $\{x_1, \ldots, x_n\}$ drawn from $\mathcal{X}$, the Gram matrix $K_{ij} = K_x(x_i, x_j)$ is positive semidefinite.*

Given any such kernel, we first define a vector space of functions mapping $\mathcal{X}$ to the real line $\mathbb{R}$ through all sums of the form

$$f(\cdot) = \sum_{j=1}^{m} \alpha_j K_x(\cdot, x_j), \tag{3.11}$$

where $\{x_j\}_{j=1}^{m}$ are arbitrary points from $\mathcal{X}$, $m \in \mathbb{N}$, and $\alpha_j \in \mathbb{R}$. We can equip this space with a *kernel-based inner product* by defining $\langle K_x(\cdot, x_i),\ K_x(\cdot, x_j)\rangle := K_x(x_i, x_j)$, and then extending this definition to the full space by bilinearity. Note that this inner product induces, for any function of the form (3.11), the kernel-based norm $\|f\|_{\mathcal{H}}^2 = \sum_{i,j=1}^{m} \alpha_i \alpha_j K_x(x_i, x_j)$.

Definition 3.3. *The* reproducing kernel Hilbert space $\mathcal{H}$ *associated with a given kernel K_x consists of the kernel-based inner product, and the closure (in the kernel-based norm) of all functions of the form* (3.11)*.*

As an aside, the term “reproducing” stems from the fact for any $f \in \mathcal{H}$, we have

$$\langle f,\, K_x(\cdot, x_i)\rangle = f(x_i),$$

showing that the kernel acts as the representer of evaluation [Saitoh, 1988].

In the framework of empirical risk minimization, the discriminant function $f \in \mathcal{H}$ is chosen by minimizing a cost function given by the sum of the *empirical ϕ-risk* $\widehat{E}\phi(Yf(X))$ and a suitable regularization term

$$\min_{f\in\mathcal{H}} \sum_{i=1}^{n} \phi(y_i f(x_i)) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2, \tag{3.12}$$

where $\lambda > 0$ is a regularization parameter that serves to limit the richness of the class of discriminant functions. The Representer Theorem (Thm. 4.2; [Schölkopf and Smola, 2002]) guarantees that the optimal solution to problem (3.12) can be written in the form

$$\widehat{f}(x) = \sum_{i=1}^{n} \alpha_i y_i K_x(x, x_i),$$

for a particular vector $\alpha \in \mathbb{R}^n$. The key here is that sum ranges *only* over the observed data points $\{(x_i, y_i)\}_{i=1}^n$.

For the sake of development in the sequel, it will be convenient to express functions $f \in \mathcal{H}$ as linear discriminants involving the the *feature map* $\Phi(x) := K_x(\cdot, x)$. (Note that for each $x \in \mathcal{X}$, the quantity $\Phi(x) \equiv \Phi(x)(\cdot)$ is a function from $\mathcal{X}$ to the real line $\mathbb{R}$.) Any function f in the Hilbert space can be written as a linear discriminant of the form $\langle w,\, \Phi(x)\rangle$ for some function $w \in \mathcal{H}$. (In fact, by the reproducing property, we have $f(\cdot) = w(\cdot)$). As a particular case, the Representer Theorem allows us to write the optimal discriminant as

$$\widehat{f}(x) = \langle \widehat{w},\, \Phi(x)\rangle,$$

where $\widehat{w} = \sum_{i=1}^{n} \alpha_i y_i \Phi(x_i)$.

3.3.2 Fusion center and marginalized kernels

With this background, we first consider how to design the decision rule γ at the fusion center for a *fixed* setting $Q(Z|X)$ of the sensor quantization rules. Since the fusion center rule can only depend on $z = (z^1, \ldots, z^S)$, our starting point is a feature space $\{\Phi'(z)\}$ with associated kernel K_z. Following the development in the previous section, we consider fusion center rules defined by taking the sign of a linear discriminant of the form

$$\gamma(z) := \langle w, \Phi'(z)\rangle.$$

We then link the performance of γ to another kernel-based discriminant function f that acts *directly* on $x = (x^1, \ldots, x^S)$, where the new kernel K_Q associated with f is defined as a *marginalized kernel* in terms of $Q(Z|X)$ and K_z.

The relevant optimization problem is to minimize (as a function of w) the following regularized form of the empirical ϕ-risk associated with the discriminant γ

$$\min_w \ \Big\{ \sum_z \sum_{i=1}^n \phi(y_i \gamma(z)) Q(z|x_i) + \frac{\lambda}{2} ||w||^2 \Big\}, \tag{3.13}$$

where $\lambda > 0$ is a regularization parameter. In its current form, the objective function (3.13) is intractable to compute (because it involves summing over all L^S possible values of z of a loss function that is generally non-decomposable). However, exploiting the convexity of ϕ allows us to perform the computation exactly for deterministic rules in $\mathcal{Q}_0$, and also leads to a natural relaxation for an arbitrary decision rule $Q \in \mathcal{Q}$. This idea is formalized in the following:

Proposition 3.4. *Define the quantities*

$$\Phi_Q(x) \ := \sum_z Q(z|x) \Phi'(z), \quad \text{and} \quad f(x;Q) \ := \ \langle w, \, \Phi_Q(x) \rangle. \tag{3.14}$$

For any convex ϕ, the optimal value of the following optimization problem is a lower bound on the optimal value in problem (3.13)*:*

$$\min_w \ \sum_i \phi(y_i f(x_i; Q)) + \frac{\lambda}{2} ||w||^2. \tag{3.15}$$

Moreover, the relaxation is tight for any deterministic rule $Q(Z|X)$.

Proof. The lower bound follows by applying Jensen's inequality to the function ϕ yields $\phi(y_i f(x_i;Q)) \leq \sum_z \phi(y_i \gamma(z)) Q(z|x_i)$ for each $i = 1, \ldots n$. □ □

A key point is that the modified optimization problem (3.15) involves an ordinary regularized empirical ϕ-loss, but in terms of a linear discriminant function

$$f(x;Q) = \langle w, \, \Phi_Q(x) \rangle$$

in the *transformed* feature space $\{\Phi_Q(x)\}$ defined in equation (3.14). Moreover, the corresponding *marginalized kernel* function takes the form:

$$K_Q(x, x') := \sum_{z,z'} Q(z|x) Q(z'|x') \ K_z(z, z'), \tag{3.16}$$

where $K_z(z, z') := \langle \Phi'(z),\, \Phi'(z') \rangle$ is the kernel in $\{\Phi'(z)\}$-space. It is straightforward to see that the positive semidefiniteness of K_z implies that K_Q is also a positive semidefinite function.

From a computational point of view, we have converted the marginalization over loss function values to a marginalization over kernel functions. While the former is intractable, the latter marginalization can be carried out in many cases by exploiting the structure of the conditional distributions $Q(Z|X)$. (In Section 3.3.3, we provide several examples to illustrate.) From the modeling perspective, it is interesting to note that marginalized kernels, like that of equation (3.16), underlie recent work that aims at combining the advantages of graphical models and Mercer kernels [Jaakkola and Haussler, 1999; Tsuda *et al.*, 2002].

As a standard kernel-based formulation, the optimization problem (3.15) can be solved by the usual Lagrangian dual formulation [Schölkopf and Smola, 2002], thereby yielding an optimal weight vector w. This weight vector defines the decision rule for the fusion center by taking the sign of discriminant function $\gamma(z) := \langle w,\, \Phi'(z) \rangle$. By the Representer Theorem [Schölkopf and Smola, 2002], the optimal solution w to problem (3.15) has an expansion of the form

$$w = \sum_{i=1}^{n} \alpha_i y_i \Phi_Q(x_i) \;\; = \;\; \sum_{i=1}^{n} \sum_{z'} \alpha_i y_i Q(z'|x_i) \Phi'(z'),$$

where α is an optimal dual solution, and the second equality follows from the definition of $\Phi_Q(x)$ given in equation (3.14). Substituting this decomposition of w into the definition of γ yields

$$\gamma(z) := \sum_{z'} \sum_{i=1}^{n} \alpha_i y_i Q(z'|x_i) K_z(z, z'). \tag{3.17}$$

Note that there is an intuitive connection between the discriminant functions f and γ. In particular, using the definitions of f and K_Q, it can be seen that

$$f(x) = \mathbb{E}[\gamma(Z)|x],$$

where the expectation is taken with respect to $Q(Z|X = x)$. The interpretation is quite natural: when conditioned on some x, the average behavior of the discriminant function $\gamma(Z)$, which does *not* observe x, is equivalent to the optimal discriminant $f(x)$, which does have access to x.

3.3.3 Design and computation of marginalized kernels

As seen in the previous section, the representation of discriminant functions f and γ depends on the kernel functions $K_z(z, z')$ and $K_Q(x, x')$, and *not* on the explicit representation of the underlying feature spaces $\{\Phi'(z)\}$ and $\{\Phi_Q(x)\}$. It is also shown in the next section that our algorithm for solving f and γ requires only the knowledge of the kernel functions K_z and K_Q. Indeed, the effectiveness of a kernel-based algorithm typically hinges heavily on the design and computation of its kernel function(s).

Accordingly, let us now consider the computational issues associated with marginalized kernel K_Q, assuming that K_z has already been chosen. In general, the computation of $K_Q(x, x')$ entails marginalizing over the variable Z, which (at first glance) has computational complexity on the order of $O(L^S)$. However, this calculation fails to take advantage of any structure in the kernel function K_z. More specifically, it is often the case that the kernel function $K_z(z, z')$ can be decomposed into local functions, in which case the computational cost is considerably lower. Here we provide a few examples of computationally tractable kernels.

Computationally tractable kernels:

Perhaps the simplest example is the *linear kernel* $K_z(z, z') = \sum_{t=1}^{S} z^t z'^t$, for which it is straightforward to derive $K_Q(x, x') = \sum_{t=l}^{S} \mathbb{E}[z^t|x^t]\, \mathbb{E}[z'^t|x'^t]$.

A second example, natural for applications in which X^t and Z^t are discrete random variables, is the *count kernel*. Let us represent each discrete value $u \in \{1, \ldots, M\}$ as a M-dimensional vector $(0, \ldots, 1, \ldots, 0)$, whose u-th coordinate takes value 1. If we define the first-order count kernel $K_z(z, z') := \sum_{t=1}^{S} \mathbb{I}[z^t = z'^t]$, then the resulting marginalized kernel takes the form:

$$K_Q(x, x') \;=\; \sum_{z,z'} Q(z|x) Q(z'|x') \sum_{t=1}^{S} \mathbb{I}[z^t = z'^t] \;=\; \sum_{t=1}^{S} Q(z^t = z'^t | x^t, x'^t). \quad (3.18)$$

A natural generalization is the *second-order count kernel* $K_z(z, z') = \sum_{t,r=1}^{s} \mathbb{I}[z^t = z'^t] \mathbb{I}[z^r = z'^r]$ that accounts for the pairwise interaction between coordinates z^t and z^r. For this example, the associated marginalized kernel $K_Q(x, x')$ takes the form:

$$2 \sum_{1 \leq t < r \leq S} Q(z^t = z'^t | x^t, x'^t) Q(z^r = z'^r | x^r, x'^r). \quad (3.19)$$

Remarks: First, note that even for a linear base kernel K_z, the kernel function K_Q inherits additional (nonlinear) structure from the marginalization over $Q(Z|X)$. As a consequence, the associated discriminant functions (i.e., γ and f) are certainly not linear. Second, our formulation allows any available prior knowledge to be incorporated into K_Q in at least two possible ways: (i) The base kernel representing a similarity measure in the quantized space of z can reflect the structure of the sensor network, or (ii) More structured decision rules $Q(Z|X)$ can be considered, such as chain or tree-structured decision rules.

3.3.4 Joint optimization

Our next task is to perform joint optimization of both the fusion center rule, defined by w (or equivalently α, as in equation (3.17)), and the sensor rules Q. Observe that the cost function (3.15) can be re-expressed as a function of both w and Q as follows:

$$G(w;Q) \ := \ \frac{1}{\lambda}\sum_i \phi\Big(y_i\langle w, \sum_z Q(z|x_i)\Phi'(z)\rangle\Big) + \frac{1}{2}||w||^2. \tag{3.20}$$

Of interest is the joint minimization of the function G in both w and Q. It can be seen easily that

(a) G is convex in w with Q fixed; and

(b) G is convex in Q^t, when both w and all other $\{Q^r, r \neq t\}$ are fixed.

These observations motivate the use of blockwise coordinate gradient descent to perform the joint minimization.

Optimization of w: As described in Section 3.3.2, when Q is fixed, then $\min_w G(w;Q)$ can be computed efficiently by a dual reformulation. Specifically, as we establish in the following result using ideas from convex duality [Rockafellar, 1970], a dual reformulation of $\min_w G(w;Q)$ is given by

$$\max_{\alpha\in\mathbb{R}^n}\Big\{-\frac{1}{\lambda}\sum_{i=1}^n \phi^*(-\lambda\alpha_i) - \frac{1}{2}\alpha^T\left[(yy^T)\circ K_Q\right]\alpha\Big\}, \tag{3.21}$$

where $\phi^*(u) := \sup_{v\in\mathbb{R}}\{u\cdot v - \phi(v)\}$ is the conjugate dual of ϕ, $[K_Q]_{ij} := K_Q(x_i,x_j)$ is the empirical kernel matrix, and $\circ$ denotes Hadamard product.

Proposition 3.5. *For each fixed $Q \in \mathcal{Q}$, the value of the primal problem $\inf_w G(w;Q)$ is attained and equal to its dual form (3.21). Furthermore, any optimal solution α to problem (3.21) defines the optimal primal solution $w(Q)$ to $\min_w G(w;Q)$ via*

$$w(Q) = \textstyle\sum_{i=1}^n \alpha_i y_i \Phi_Q(x_i).$$

Proof. It suffices for our current purposes to restrict to the case where the functions w and $\Phi_Q(x)$ can be viewed as vectors in some finite-dimensional space—say $\mathbb{R}^m$. However, it is possible to extend this approach to the infinite-dimensional setting by using conjugacy in general normed spaces [Luenberger, 1969].

A remark on notation before proceeding: since Q is fixed, we drop Q from G for notational convenience (i.e., we write $G(w) \equiv G(w;Q)$). First, we observe that $G(w)$ is convex with respect to w and that $G \to \infty$ as $||w|| \to \infty$. Consequently, the infimum defining the primal problem $\inf_{w\in\mathbb{R}^m} G(w)$ is attained. We now re-write this primal problem as $\inf_{w\in\mathbb{R}^m} G(w) = \inf_{w\in\mathbb{R}^m}\{G(w) - \langle w,\, 0\rangle\} = -G^*(0)$, where $G^* : \mathbb{R}^m \to \mathbb{R}$ denotes the conjugate dual of G.

Using the notation $g_i(w) := \frac{1}{\lambda}\phi(\langle w,\, y_i\Phi_Q(x_i)\rangle)$ and $\Omega(w) := \frac{1}{2}||w||^2$, we can decompose G as the sum $G(w) = \sum_{i=1}^n g_i(w) + \Omega(w)$. This decomposition allows us to compute the conjugate dual G^* via the inf-convolution theorem (Thm. 16.4; Rockafellar [Rockafellar, 1970]) as follows:

$$G^*(0) \quad = \quad \inf_{u_i, i=1,\ldots,n}\left\{\sum_{i=1}^n g_i^*(u_i) + \Omega^*(-\sum_{i=1}^n u_i)\right\}. \tag{3.22}$$

The function g_i is the composition of a convex function ϕ with the linear function $w \mapsto \langle w, y_i\Phi_Q(x_i)\rangle$, so that Theorem 16.3 of Rockafellar [Rockafellar, 1970] yields the conjugate dual as follows:

$$g_i^*(u_i) \quad = \quad \begin{cases} \frac{1}{\lambda}\phi^*(-\lambda\alpha_i) & \text{if } u_i = -\alpha_i(y_i\Phi_Q(x_i)) \text{ for some } \alpha_i \in \mathbb{R} \\ +\infty & \text{otherwise.} \end{cases} \tag{3.23}$$

A straightforward calculation yields $\Omega^*(v) = \sup_w\{\langle v,\, w\rangle - \frac{1}{2}||w||^2\} = \frac{1}{2}||v||^2$. Substituting these expressions into equation (3.22) leads to:

$$G^*(0) \quad = \quad \inf_{\alpha\in\mathbb{R}^n} \sum_{i=1}^n \frac{1}{\lambda}\phi^*(-\lambda_i\alpha_i) + \frac{1}{2}\left\|\sum_i^n \alpha_i y_i \Phi_Q(x_i)\right\|^2,$$

from which it follows that

$$\inf_w G(w) \quad = \quad -G^*(0) \quad = \quad \sup_{\alpha\in\mathbb{R}^n}\left\{-\frac{1}{\lambda}\sum_{i=1}^n \phi^*(-\lambda\alpha_i) - \frac{1}{2}\sum_{1\leq i,j\leq n} \alpha_i\alpha_j y_i y_j K_x(x_i, x_j)\right\}.$$

Thus, we have derived the dual form (3.21). See the Appendix for the remainder of the proof, in which we derive the link between $w(Q)$ and the dual variables α. □

This proposition is significant in that the dual problem involves only the kernel matrix $(K_Q(x_i, x_j))_{1\leq i,j\leq n}$. Hence, one can solve for the optimal discriminant functions $y = f(x)$

or $y = \gamma(z)$ without requiring explicit knowledge of the underlying feature spaces $\{\Phi'(z)\}$ and $\{\Phi_Q(x)\}$. As a particular example, consider the case of hinge loss function (3.8), as used in the SVM algorithm [Schölkopf and Smola, 2002]. A straightforward calculation yields

$$\phi^*(u) = \begin{cases} u & \text{if } u \in [-1, 0] \\ +\infty & \text{otherwise.} \end{cases}$$

Substituting this formula into (3.21) yields, as a special case, the familiar dual formulation for the SVM:

$$\max_{0 \leq \alpha \leq 1/\lambda} \left\{ \sum_i^n \alpha_i - \frac{1}{2} \alpha^T \left[(yy^T) \circ K_Q \right] \alpha \right\}.$$

Optimization of Q: The second step is to minimize G over Q^t, with w and all other $\{Q^r, r \neq t\}$ held fixed. Our approach is to compute the derivative (or more generally, the subdifferential) with respect to Q^t, and then apply a gradient-based method. A challenge to be confronted is that G is defined in terms of feature vectors $\Phi'(z)$, which are typically high-dimensional quantities. Indeed, although it is intractable to evaluate the gradient at an arbitrary w, the following result, proved in the Appendix, establishes that it can always be evaluated at the point $(w(Q), Q)$ for any $Q \in \mathcal{Q}$.

Lemma 3.6. *Let $w(Q)$ be the optimizing argument of $\min_w G(w; Q)$, and let α be an optimal solution to the dual problem* (3.21)*. Then the following element*

$$-\lambda \sum_{(i,j)(z,z')} \alpha_i \alpha_j Q(z'|x_j) \frac{Q(z|x_i)}{Q^t(z^t|x_i^t)} K_z(z, z') \mathbb{I}[x_i^t = \bar{x}^t]\, \mathbb{I}[z^t = \bar{z}^t]$$

is an element of the subdifferential $\partial_{Q^t(\bar{z}^t|\bar{x}^t)} G$ evaluated at $(w(Q), Q)$. [3]

Note that this representation of the (sub)gradient involves marginalization over Q of the kernel function K_z, and therefore can be computed efficiently in many cases, as described in Section 3.3.3. Overall, the blockwise coordinate descent algorithm for optimizing the local quantization rules has the form:

[3] The *subgradient* is a generalized counterpart of gradient for non-differentiable convex functions [Rockafellar, 1970; Hiriart-Urruty and Lemaréchal, 2001]; in particular, a vector $s \in \mathbb{R}^m$ is a *subgradient* of a convex function $f : \mathbb{R}^m \to \mathbb{R}$ means $f(y) \geq f(x) + \langle s, y - x \rangle$ for all $y \in \mathbb{R}^m$. The *subdifferential* at a point x is the set of all subgradients. In our cases, G is non-differentiable when ϕ is the hinge loss (3.8), and differentiable when ϕ is the logistic loss (3.9) or exponential loss (3.10).

Kernel quantization (KQ) algorithm:

(a) With Q fixed, compute the optimizing $w(Q)$ by solving the dual problem (3.21).

(b) For some index t, fix $w(Q)$ and $\{Q^r, r \neq t\}$ and take a gradient step in Q^t using Lemma 3.6.

Upon convergence, we define a deterministic decision rule for each sensor t via:

$$\gamma^t(x^t) \quad := \quad \operatorname{argmax}_{z^t \in \mathcal{Z}} Q(z^t|x^t). \tag{3.24}$$

First, note that the updates in this algorithm consist of alternatively updating the decision rule for a sensor while fixing the decision rules for the remaining sensors and the fusion center, and updating the decision rule for the fusion center while fixing the decision rules for all other sensors. In this sense, our approach is similar in spirit to a suite of practical algorithms [Tsitsiklis, 1993b, e.g.,] for decentralized detection under particular assumptions on the joint distribution $P(X, Y)$. Second, using standard results [Bertsekas, 1995b], it is possible to guarantee convergence of such coordinate-wise updates when the loss function ϕ is strictly convex and differentiable (e.g., logistic loss (3.9) or exponential loss (3.10)). In contrast, the case of non-differentiable ϕ (e.g., hinge loss (3.8)) requires more care. We have, however, obtained good results in practice even in the case of hinge loss. Third, it is interesting to note the connection between the KQ algorithm and the naive approach considered in Section 3.2.2. More precisely, suppose that we fix w such that all α_i are equal to one, and let the base kernel K_z be constant (and thus entirely uninformative). Under these conditions, the optimization of G with respect to Q reduces to exactly the naive approach.

3.3.5 Estimation error bounds

This section is devoted to analysis of the statistical properties of the KQ algorithm. In particular, our goal is to derive bounds on the performance of our classifier (Q, γ) when applied to new data, as opposed to the i.i.d. samples on which it was trained. It is key to distinguish between two forms of ϕ-risk:

(a) the *empirical ϕ-risk* $\widehat{E}\phi(Y\gamma(Z))$ is defined by an expectation over $\widehat{P}(X, Y)Q(Z|X)$, where $\widehat{P}$ is the empirical distribution given by the i.i.d. samples $\{(x_i, y_i)\}_{i=1}^n$.

(b) the *true ϕ-risk* $\mathbb{E}\phi(Y\gamma(Z))$ is defined by taking an expectation over the joint distribution $P(X, Y)Q(Z|X)$.

In designing our classifier, we made use of the empirical ϕ-risk as a proxy for the actual risk. On the other hand, the appropriate metric for assessing performance of the designed classifier is the true ϕ-risk $\mathbb{E}\phi(Y\gamma(Z))$. At a high level, our procedure for obtaining performance bounds can be decomposed into the following steps:
(1) First, we relate the true ϕ-risk $\mathbb{E}\phi(Y\gamma(Z))$ to the true ϕ-risk $\mathbb{E}\phi(Yf(X))$ for the functions $f \in \mathcal{F}$ (and $f \in \mathcal{F}_0$) that are computed at intermediate stages of our algorithm. The latter quantities are well-studied objects in statistical learning theory.
(2) The second step to relate the empirical ϕ-risk $\widehat{E}(Yf(X))$ to the true ϕ-risk $\mathbb{E}(Yf(X))$. In general, the true ϕ-risk for a function f in some class $\mathcal{F}$ is bounded by the empirical ϕ-risk plus a complexity term that captures the "richness" of the function class $\mathcal{F}$ [Zhang, 2004; Bartlett *et al.*, 2006]. In particular, we make use of the *Rademacher complexity* as a measure of this richness.
(3) Third, we combine the first two steps so as to derive bounds on the true ϕ-risk $\mathbb{E}\phi(Y\gamma(Z))$ in terms of the empirical ϕ-risk of f and the Rademacher complexity.
(4) Finally, we derive bounds on the Rademacher complexity in terms of the number of training samples n, as well as the number of quantization levels L and M.

Step 1: For each $Q \in \mathcal{Q}$, the class of functions $\mathcal{F}_Q$ over which we optimize is given by:

$$\big\{f : x \mapsto \langle w, \Phi_Q(x)\rangle = \sum_i \alpha_i y_i K_Q(x, x_i) \;\big|\; \text{s. t. } ||w|| \leq B\big\}, \tag{3.25}$$

where $B > 0$ is a constant. Note that $\mathcal{F}_Q$ is simply the class of functions associated with the marginalized kernel K_Q. The function class over which our algorithm performs the optimization is defined by the union $\mathcal{F} := \cup_{Q\in\mathcal{Q}}\mathcal{F}_Q$, where $\mathcal{Q}$ is the space of all factorized conditional distributions $Q(Z|X)$. Lastly, we define the function class $\mathcal{F}_0 := \cup_{Q\in\mathcal{Q}_0}\mathcal{F}_Q$, corresponding to the union of the function spaces defined by marginalized kernels with deterministic distributions Q.

Any discriminant function $f \in \mathcal{F}$ (or $\mathcal{F}_0$), defined by a vector α, induces an associated discriminant function γ_f via equation (3.17). Relevant to the performance of the classifier γ_f is the expected ϕ-loss $\mathbb{E}\phi(Y\gamma_f(Z))$, whereas the algorithm actually minimizes (the empirical version of) $\mathbb{E}\phi(Yf(X))$. The relationship between these two quantities is expressed in the following proposition.

Proposition 3.7.
(a) We have $\mathbb{E}\phi(Y\gamma_f(Z)) \geq \mathbb{E}\phi(Yf(X))$, with equality when $Q(Z|X)$ is deterministic.
(b) Moreover, there holds

$$\inf_{f\in\mathcal{F}} \mathbb{E}\phi(Yf(X)) \overset{(i)}{\leq} \inf_{f\in\mathcal{F}} \mathbb{E}\phi(Y\gamma_f(Z)) \overset{(ii)}{\leq} \inf_{f\in\mathcal{F}_0} \mathbb{E}\phi(Yf(X)) \tag{3.26}$$

The same statements also hold for empirical expectations.

Proof. Applying Jensen's inequality to the convex function ϕ yields

$$\mathbb{E}\phi(Y\gamma_f(Z)) \;=\; \mathbb{E}_{XY}\mathbb{E}[\phi(Y\gamma_f(Z))|X,Y] \;\geq\; \mathbb{E}_{XY}\phi(\mathbb{E}[Y\gamma_f(Z)|X,Y]) = \mathbb{E}\phi(Yf(X)),$$

where we have used the conditional independence of Z and Y given X. This establishes inequality (ii), and the lower bound (i) follows directly. Moreover, part (a) also implies that $\inf_{f\in\mathcal{F}_0}\mathbb{E}\phi(Y\gamma_f(Z)) = \inf_{f\in\mathcal{F}_0}\mathbb{E}\phi(Yf(X))$, and the upper bound (3.26) follows since $\mathcal{F}_0 \subset \mathcal{F}$. □ □

Step 2: The next step is to relate the empirical ϕ-risk for f (i.e., $\widehat{\mathbb{E}}(Yf(X))$) to the true ϕ-risk (i.e., $\mathbb{E}(Yf(X))$). Recall that the *Rademacher complexity* of the function class $\mathcal{F}$ is defined [van der Vaart and Wellner, 1996] as

$$R_n(\mathcal{F}) = \mathbb{E}\sup_{f\in\mathcal{F}}\frac{2}{n}\sum_{i=1}^{n}\sigma_i f(X_i),$$

where the *Rademacher variables* $\sigma_1,\ldots,\sigma_n$ are independent and uniform on $\{-1,+1\}$, and $X_1,\ldots,X_n$ are i.i.d. samples selected according to distribution P. In the case that ϕ is Lipschitz with constant ℓ, the empirical and true risk can be related via the Rademacher complexity as follows [Koltchinskii and Panchenko, 2002]. With probability at least $1-\delta$ with respect to training samples $(X_i,Y_i)_{i=1}^n$, drawn according to the empirical distribution P^n, there holds

$$\sup_{f\in\mathcal{F}}|\widehat{E}\phi(Yf(X)) - \mathbb{E}\phi(Yf(X))| \leq 2\ell R_n(\mathcal{F}) + \sqrt{\frac{\ln(2/\delta)}{2n}}. \tag{3.27}$$

Moreover, the same bound applies to $\mathcal{F}_0$.

Step 3: Combining the bound (3.27) with Proposition 3.7 leads to the following theorem, which provides generalization error bounds for the optimal ϕ-risk of the decision function learned by our algorithm in terms of the Rademacher complexities $R_n(\mathcal{F}_0)$ and $R_n(\mathcal{F})$:

Theorem 3.8. *Given n i.i.d. labeled data points $(x_i,y_i)_{i=1}^n$, with probability at least $1-2\delta$,*

$$\inf_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\phi(y_if(x_i)) - 2\ell R_n(\mathcal{F}) - \sqrt{\frac{\ln(2/\delta)}{2n}} \leq \inf_{f\in\mathcal{F}}\mathbb{E}\phi(Y\gamma_f(Z))$$

$$\inf_{f\in\mathcal{F}}\mathbb{E}\phi(Y\gamma_f(Z)) \leq \inf_{f\in\mathcal{F}_0}\frac{1}{n}\sum_{i=1}^{n}\phi(y_if(x_i)) + 2\ell R_n(\mathcal{F}_0) + \sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Proof. Using bound (3.27), with probability at least $1-\delta$, for any $f \in \mathcal{F}$,

$$\mathbb{E}\phi(Yf(X)) \geq \frac{1}{n}\sum_{i=1}^{n}\phi(y_i f(x_i)) - 2\ell R_n(\mathcal{F}) - \sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Combining with bound (i) in equation (3.26), we have, with probability $1-\delta$,

$$\inf_{f\in\mathcal{F}} \mathbb{E}\phi(Y\gamma_f(Z)) \;\geq\; \inf_{f\in\mathcal{F}} \mathbb{E}\phi(Yf(X)) \;\geq\; \inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\phi(y_i f(x_i)) - 2\ell R_n(\mathcal{F}) - \sqrt{\frac{\ln(2/\delta)}{2n}},$$

which proves the lower bound of the theorem with probability at least $1-\delta$. The upper bound is similarly true with probability at least $1-\delta$. Hence, both are true with probability at least $1-2\delta$, by the union bound. □ □

Step 4: So that Theorem 3.8 has useful meaning, we need to derive upper bounds on the Rademacher complexity of the function classes $\mathcal{F}$ and $\mathcal{F}_0$. Of particular interest is the decrease in the complexity of $\mathcal{F}$ and $\mathcal{F}_0$ with respect to the number of training samples n, as well as their growth rate with respect to the number of discrete signal levels M, number of quantization levels L, and the number of sensors S. The following proposition, proved in the Appendix, derives such bounds by exploiting the fact that the number of 0-1 conditional probability distributions $Q(Z|X)$ is finite (namely, (L^{MS})).

Proposition 3.9.

$$R_n(\mathcal{F}_0) \leq \frac{2B}{n}\left[\mathbb{E}\sup_{Q\in\mathcal{Q}_0}\sum_{i=1}^{n}K_Q(X_i,X_i) + 2(n-1)\sqrt{n/2}\sup_{z,z'}K_z(z,z')\sqrt{2MS\log L}\right]^{1/2}. \tag{3.28}$$

Note that the upper bound involves a linear dependence on constant B, assuming that $\|w\| \leq B$—this provides a statistical justification of minimizing $\|w\|^2$ in the formulation (3.13). Although the rate given in equation (3.28) is not tight in terms of the number of data samples n, the bound is nontrivial and is relatively simple. (In particular, it depends directly on the kernel function K, the number of samples n, quantization levels L, number of sensors S, and size of observation space M.)

We can also provide a more general and possibly tighter upper bound on the Rademacher complexity based on the concept of *entropy number* [van der Vaart and Wellner, 1996]. Indeed, an important property of the Rademacher complexity is that it can be estimated reliably from a single sample $(x_1, \ldots, x_n)$. Specifically, if we define

$$\widehat{R}_n(\mathcal{F}) := \mathbb{E}[\frac{2}{n} \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i f(x_i)]$$

(where the expectation is w.r.t. the Rademacher variables $\{\sigma_i\}$ only), then it can be shown using McDiarmid's inequality that $\widehat{R}_n(\mathcal{F})$ is tightly concentrated around $R_n(\mathcal{F})$ with high probability [Bartlett and Mendelson, 2002]. Concretely, assuming that $\|f\|_\infty$ is bounded from above for all $\in \mathcal{F}$, then for any $\eta > 0$, there holds:

$$P\left\{|R_n(\mathcal{F}) - \widehat{R}_n(\mathcal{F})| \quad \geq \quad \eta\right\} \leq 2e^{-\eta^2 n/8}. \tag{3.29}$$

Hence, the Rademacher complexity is closely related to its empirical version $\widehat{R}_n(\mathcal{F})$, which can be related to the concept of entropy number. In general, define the covering number $N(\epsilon, S, \rho)$ for a set S to be the minimum number of balls of diameter ϵ that completely cover S (according to a metric ρ). The ϵ-entropy number of S is then defined as $\log N(\epsilon, S, \rho)$. In particular, if we define the $L_2(P_n)$ metric on an empirical sample $(x_1, \ldots, x_n)$ as

$$\|f_1 - f_2\|_{L_2(P_n)} := \left[\frac{1}{n} \sum_{i=1}^{n} (f_1(x_i) - f_2(x_i))^2\right]^{1/2},$$

then it is well known [van der Vaart and Wellner, 1996] that for some absolute constant C, there holds:

$$\widehat{R}_n(\mathcal{F}) \leq C \int_0^\infty \sqrt{\frac{\log N(\epsilon, \mathcal{F}, L_2(P_n))}{n}} \mathrm{d}\epsilon. \tag{3.30}$$

The following result, proved in the Appendix, relates the entropy number for $\mathcal{F}$ to the supremum of the entropy number taken over a restricted function class $\mathcal{F}_Q$.

Proposition 3.10. *The entropy number* $\log N(\epsilon, \mathcal{F}, L_2(P_n))$ *of* $\mathcal{F}$ *is bounded above by*

$$\sup_{Q \in \mathcal{Q}} \log N(\epsilon/2, \mathcal{F}_Q, L_2(P_n)) + (L-1)MS \log \frac{2L^S \sup ||\alpha||_1 \sup_{z,z'} K_z(z, z')}{\epsilon}. \tag{3.31}$$

Moreover, the same bound holds for $\mathcal{F}_0$.

This proposition guarantees that the increase in the entropy number in moving from some $\mathcal{F}_Q$ to the larger class $\mathcal{F}$ is only $O((L-1)MS \log(L^S/\epsilon))$. Consequently, we incur at most an $O([MS^2(L-1)\log L/n]^{\frac{1}{2}})$ increase in the upper bound (3.30) for $R_n(\mathcal{F})$ (as well as $R_n(\mathcal{F}_0)$). Moreover, the Rademacher complexity increases with the square root of the number $L \log L$ of quantization levels L.

3.4 Experimental Results

We evaluated our algorithm using both data from simulated and real sensor networks and real-world data sets. First, we consider three types of simulated sensor network configurations:

Conditionally independent observations: In this example, the observations $X^1, \ldots, X^S$ are independent conditional on Y, as illustrated in Figure 3.1. We consider networks with 10 sensors ($S = 10$), each of which receive signals with 8 levels ($M = 8$). We applied the algorithm to compute decision rules for $L = 2$. In all cases, we generate $n = 200$ training samples, and the same number for testing. We performed 20 trials on each of 20 randomly generated models $P(X, Y)$.

Chain-structured dependency: A conditional independence assumption for the observations, though widely employed in most work on decentralized detection, may be unrealistic in many settings. For instance, consider the problem of detecting a random signal in noise [van Trees, 1990], in which $Y = 1$ represents the hypothesis that a certain random signal is present in the environment, whereas $Y = -1$ represents the hypothesis that only i.i.d. noise is present. Under these assumptions $X^1, \ldots, X^S$ will be conditionally independent given $Y = -1$, since all sensors receive i.i.d. noise. However, conditioned on $Y = +1$ (i.e., in the presence of the random signal), the observations at spatially adjacent sensors will be dependent, with the dependence decaying with distance.

In a 1-D setting, these conditions can be modeled with a chain-structured dependency, and the use of a count kernel to account for the interaction among sensors. More precisely, we consider a set-up in which five sensors are located in a line such that only adjacent sensors interact with each other. More specifically, the sensors X_{t-1} and X_{t+1} are independent given X_t *and* Y, as illustrated in Figure 3.2. We implemented the kernel-based quantization algorithm using either first- or second-order count kernels, and the hinge loss function (3.8), as in the SVM algorithm. The second-order kernel is specified in equation (3.19) but with the sum taken over only t, r such that $|t - r| = 1$.

Spatially-dependent sensors: As a third example, we consider a 2-D layout in which, conditional on the random target being present ($Y = +1$), all sensors interact but with the strength of interaction decaying with distance. Thus $P(X|Y = 1)$ is of the form:

$$P(X|Y = 1) \propto \exp\Big\{ \sum_t h_{t;u} \mathbb{I}_u(X^t) + \sum_{t \neq r; uv} \theta_{tr;uv} \mathbb{I}_u(X^t) \mathbb{I}_v(X^r) \Big\}.$$

Here the parameter h represents observations at individual sensors, whereas θ controls the dependence among sensors. The distribution $P(X|Y = -1)$ can be modeled in the same way with observations h', and setting $\theta' = 0$ so that the sensors are conditionally independent. In simulations, we generate $\theta_{tr;uv} \sim N(1/d_{tr}, 0.1)$, where d_{tr} is the distance between sensor t and r, and the observations h and h' are randomly chosen in $[0, 1]^S$. We

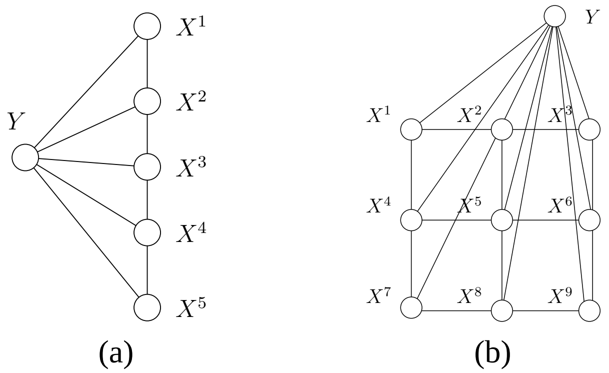


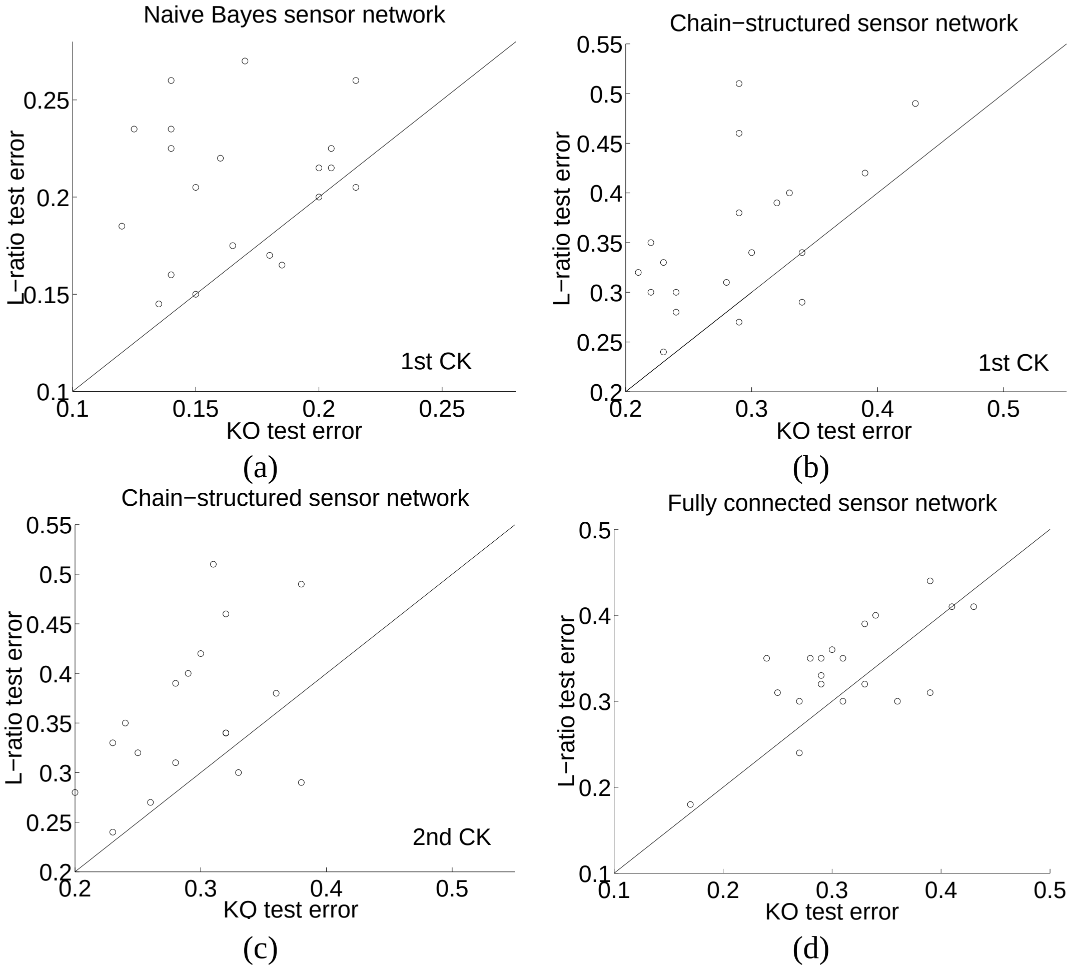
Figure 3.2. Examples of graphical models $P(X, Y)$ of our simulated sensor networks. (a) Chain-structured dependency. (b) Fully connected (not all connections shown).

consider a sensor network with 9 nodes (i.e., $S = 9$), arrayed in the 3×3 lattice illustrated in Figure 3.2(b). Since computation of this density is intractable for moderate-sized networks, we generated an empirical data set (x_i, y_i) by Gibbs sampling.

We compare the results of our algorithm to an alternative decentralized classifier based on performing a likelihood-ratio (LR) test at each sensor. Specifically, for each sensor t, the estimates $\frac{\hat{P}(X^t=u|Y=1)}{\hat{P}(X^t=u|Y=-1)}$ for $u = 1, \ldots, M$ of the likelihood ratio are sorted and grouped evenly into L bins, resulting in a simple and intuitive likelihood-ratio based quantization scheme. Note that the estimates $\hat{P}$ are obtained from the training data. Given the quantized input signal and label Y, we then construct a naive Bayes classifier at the fusion center. This choice of decision rule provides a reasonable comparison, since thresholded likelihood ratio tests are optimal in many cases [Tsitsiklis, 1993b].

The KQ algorithm generally yields more accurate classification performance than the likelihood-ratio based algorithm (LR). Figure 3.3 provides scatter plots of the test error of the KQ versus LQ methods for four different set-ups, using $L = 2$ levels of quantization. Panel (a) shows the naive Bayes setting and the KQ method using the first-order count kernel. Note that the KQ test error is below the LR test error on the large majority of examples. Panels (b) and (c) show the case of chain-structured dependency, as illustrated in Figure 3.2(a), using a first- and second-order count kernel respectively. Again, the performance of KQ in both cases is superior to that of LR in most cases. Finally, panel (d) shows the fully-connected case of Figure 3.2(b) with a first-order kernel. The performance of KQ is somewhat better than LR, although by a lesser amount than the other cases.

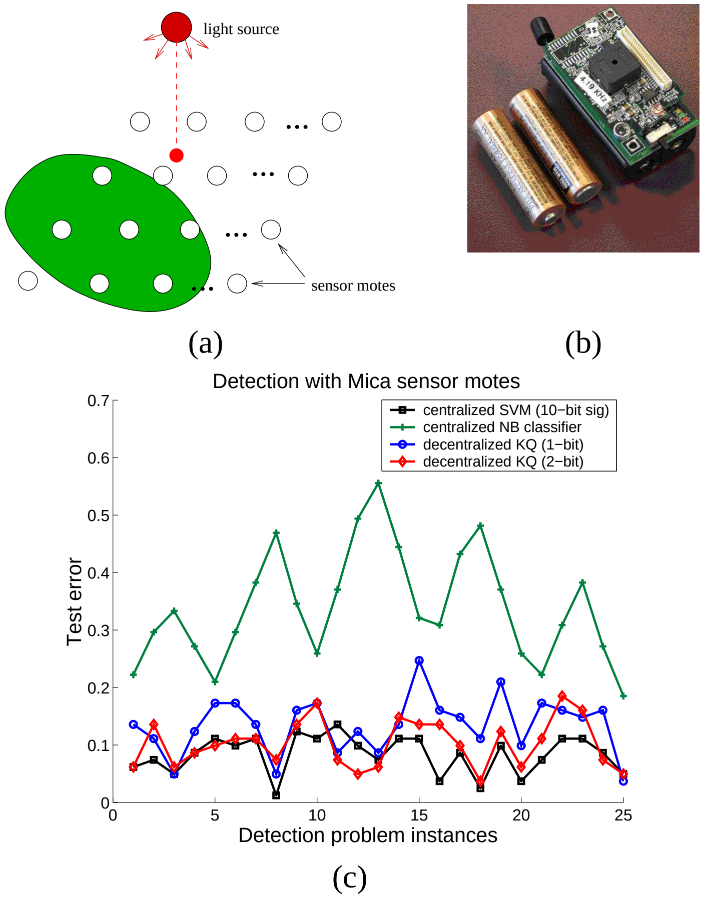
Real sensor network data set: We evaluated our algorithm on a real sensor network using Berkeley tiny sensor motes (Mica motes) as the base stations. The goal of the experiment is to determine the locations of light sources given the light signal strength received



(a) (b) (c) (d)

Figure 3.3. Scatter plots of the test error of the LR versus KQ methods. (a) Conditionally independent network. (b) Chain model with first-order kernel. (c) Chain model with second-order kernel. (d) Fully connected model.

by a number of sensors deployed in the network. Specifically, we fix a particular region in the plane (i.e., sensor field) and ask whether the light source's projection onto the plane is within this region or not (see Figure 3.4(a)). The light signal strength received by each sensor mote requires 10 bits to store, and we wish to reduce the size of each sensor message being sent to the fusion center to only 1 or 2 bits. Our hardware platform consists of 25 sensors placed 10 inches apart on a 5×5 grid in an indoor environment. We performed 25 detection problems corresponding to 25 circular regions of radius 30 inches distributed uniformly over the sensor field. For each problem instance, there are 25 training positions (i.e., empirical samples), and 81 test positions.



(a) (b)

(c)

Figure 3.4. (a) Illustration of a sensor field. (b) a Mica sensor mote. (c) Comparison of test errors of the decentralized KQ algorithm and centralized SVM and NBC algorithms on different problem instances.

The performance of the KQ algorithm is compared to *centralized* detection algorithms based on a Naive Bayes classifier (NBC), and the SVM algorithm using a Gaussian kernel.[4] The test errors of these algorithms are shown in Figure 3.4(b). Note that the test algorithm of the KQ algorithm improves considerably by relaxing the communication constraints from 1 to 2 bits. Furthermore, with the 2-bit bandwidth constraint, the KQ's test errors are comparable to that of the centralized SVM algorithm on most problem instances. On the other hand, the centralized NBC algorithm does not perform well on this data set.

UCI repository data sets: We also applied our algorithm to several data sets from the machine learning data repository at the University of California Irvine [Blake and Merz, 1998]. In contrast to the sensor network detection problem, in which communication constraints must be respected, the problem here can be viewed as that of finding a good quantization scheme that retains information about the class label. Thus, the problem is similar in spirit to work on discretization schemes for classification [Dougherty *et al.*, 1995]. The difference is that we assume that the data have already been crudely quantized (we use $m = 8$ levels in our experiments), and that we retain no topological information concerning the relative magnitudes of these values that could be used to drive classical discretization algorithms. Overall, the problem can be viewed as hierarchical decision-making, in which a second-level classification decision follows a first-level set of decisions concerning the features. We used 75% of the data set for training and the remainder for testing. The results for our algorithm with $L = 2, 4$, and 6 quantization levels are shown in Table 3.1. Note that in several cases the quantized algorithm actually outperforms a naive Bayes algorithm (NB) with access to the real-valued features. This result may be due in part to the fact that our quantizer is based on a discriminative classifier, but it is worth noting that similar improvements over naive Bayes have been reported in earlier empirical work using classical discretization algorithms [Dougherty *et al.*, 1995].

Data	$L = 2$	4	6	NB	CK
Pima	0.212	0.217	0.212	0.223	0.212
Iono	0.091	0.034	0.079	0.056	0.125
Bupa	0.368	0.322	0.345	0.322	0.345
Ecoli	0.082	0.176	0.176	0.235	0.188
Yeast	0.312	0.312	0.312	0.303	0.317
Wdbc	0.083	0.097	0.111	0.083	0.083

Table 3.1: Experimental results for the UCI data sets.

[4] The sensor observations are initially quantized into $m = 10$ bins, which then serves as input to the NBC and KQ algorithm.

3.5 Discussions

We have presented a new approach to the problem of decentralized decision-making under constraints on the number of bits that can be transmitted by each of a distributed set of sensors. In contrast to most previous work in an extensive line of research on this problem, we propose a nonparametric solution: in particular, we assume that the joint distribution of sensor observations is unknown, and that a set of data samples is available. We have proposed a novel algorithm based on kernel methods, and shown that it is quite effective on both simulated and real-world data sets.

This line of work described here can be extended in a number of directions. First, although we have focused on discrete observations X, it is natural to consider continuous signal observations. Doing so would require considering parameterized distributions $Q(Z|X)$. Second, our kernel design so far makes use of only rudimentary information from the sensor observation model, and could be improved by exploiting such knowledge more thoroughly. Third, we have considered only the so-called *parallel* configuration of the sensors, which amounts to the conditional independence of $Q(Z|X)$. One direction to explore is the use of kernel-based methods for richer configurations, such as tree-structured and *tandem* configurations [Tsitsiklis, 1993b]. Finally, the work described here falls within the area of *fixed sample size* detectors. An alternative type of decentralized detection procedure is a *sequential* detector, in which there is usually a large (possibly infinite) number of observations that can be taken in sequence (e.g. [Veeravalli *et al.*, 1993]). It is also interesting to consider extensions our method to this sequential setting.

On the theoretical front, although we have provided an estimation error analysis with respect to the surrogate ϕ-risk, no guarantee is given with respect to the Bayes error per se. Specifically, does the quantizer-classifier pair (Q, γ) obtained our learning procedure is (asymptotically) optimal in the sense of 0-1 loss? A complete answer to this question is given in Chapter 4.

Appendix 3.A Proof of Lemma 3.1

(a) Since $x_1, \ldots, x_n$ are independent realizations of the random vector X, the quantities $Q(z|x_1), \ldots, Q(z|x_n)$ are independent realizations of the random variable $Q(z|X)$. (This statement holds for each fixed $z \in \mathcal{Z}^S$.) The strong law of large numbers yields

$$\frac{1}{n}\sum_{i=1}^{n} Q(z|x_i) \xrightarrow{a.s.} \mathbb{E}Q(z|x_i) = P(z)$$

as $n \to +\infty$. Similarly, we have

$$\frac{1}{n}\sum_{i=1}^{n} Q(z|x_i)\mathbb{I}(y_i = 1) \xrightarrow{a.s.} \mathbb{E}Q(z|X)\mathbb{I}(Y = 1).$$

Therefore, as $n \to \infty$,

$$\kappa(z) \quad \xrightarrow{a.s.} \quad \frac{\mathbb{E}Q(z|X)\mathbb{I}(Y=1)}{P(z)} \quad = \quad \sum_{x} \frac{Q(z|X=x)P(X=x, Y=1)}{P(z)} = P(Y=1|z),$$

here we have exploited the fact that Z is independent of Y given X.

(b) For each $z \in \mathcal{Z}^S$, we have

$$\begin{aligned} & \operatorname{sign}\left(\frac{\sum_{i=1}^{n} Q(z|x_i)\mathbb{I}(y_i=1)}{\sum_{i=1}^{n} Q(z|x_i)} - \frac{\sum_{i=1}^{n} Q(z|x_i)\mathbb{I}(y_i=-1)}{\sum_{i=1}^{n} Q(z|x_i)}\right) \\ = \quad & \operatorname{sign}\left(\frac{\sum_{i=1}^{n} Q(z|x_i)y_i}{\sum_{i=1}^{n} Q(z|x_i)}\right) \quad = \gamma_{emp}(z). \end{aligned}$$

Thus, part (a) implies $\gamma_{emp}(z) \to \gamma_{opt}(z)$ for each z. Similarly, $R_{emp} \to R_{opt}$.

Appendix 3.B Proof of Proposition 3.5

Here we complete the proof of Proposition 3.5. It remains to show that the optimum $w(Q)$ of the primal problem is related to the optimal α of the dual problem via $w(Q) = \sum_{i=1}^{n} \alpha_i y_i \Phi_Q(x_i)$. Indeed, since $G(w)$ is a convex function with respect to w, $w(Q)$ is an optimum solution for $\min_w G(w; Q)$ if and only if $0 \in \partial_w G(w(Q))$. By definition of the conjugate dual, this condition is equivalent to $w(Q) \in \partial G^*(0)$.

Recall that G^* is an inf-convolution of n functions $g_1^*, \ldots, g_n^*$ and Ω^*. Let $\widehat{\alpha} := (\widehat{\alpha_1}, \ldots, \widehat{\alpha_n})$ be an optimum solution to the dual problem, and $\widehat{u} := (\widehat{u_1}, \ldots, \widehat{u_n})$ be the corresponding value in which the infimum operation in the definition of G^* is attained. Applying the subdifferential operation rule on a inf-convolution function (Cor. 4.5.5, [Hiriart-

Urruty and Lemaréchal, 2001]), we have

$$\partial G^*(0) = \partial g_1^*(\widehat{u}_1) \cap \ldots \cap \partial g_n^*(\widehat{u_n}) \cap \partial\Omega^*(-\sum_{i=1}^n \widehat{u}_i).$$

But $\Omega^*(v) = \frac{1}{2}\|v\|^2$, and so $\partial\Omega^*(-\sum_{i=1}^n \widehat{u}_i)$ reduces to a singleton

$$-\sum_{i=1}^n \widehat{u}_i = \sum_{i=1}^n \widehat{\alpha}_i y_i \Phi_Q(x_i).$$

This implies that $w(Q) = \sum_{i=1}^n \widehat{\alpha}_i y_i \Phi_Q(x_i)$ is the optimum solution to the primal problem.

To conclude, it will be useful for the proof of Lemma 3.6 to calculate $\partial g_i^*(\widehat{u}_i)$, and derive several additional properties relating $w(Q)$ and $\widehat{\alpha}$. The expression for g_i^* in equation (3.23) shows that it is the image of the function $\frac{1}{\lambda}\phi^*$ under the linear mapping $\alpha_i \mapsto \frac{1}{\lambda}\alpha_i(y_i\Phi_Q(x_i))$. Consequently, by Theorem 4.5.1 of Urruty and Lemarechal [Hiriart-Urruty and Lemaréchal, 2001]), we have $\partial g_i^*(\widehat{u}_i) = \{w : \langle w,\, y_i\Phi_Q(x_i)\rangle \in \partial\phi^*(-\lambda\widehat{\alpha}_i)\}$, which implies that $b_i := \langle w(Q),\, y_i\Phi_Q(x_i)\rangle \in \partial\phi^*(-\lambda\widehat{\alpha}_i)$ for each $i = 1,\ldots,n$. By convex duality, this also implies that $-\lambda\widehat{\alpha}_i \in \partial\phi(b_i)$ for $i = 1,\ldots,n$.

Appendix 3.C Proof of Lemma 3.6

We shall show that the subdifferential $\partial_{Q^t(\bar{z}^t|\bar{x}^t)}G$ can be computed directly in terms of the optimal solution α of the dual optimization problem (3.21) and the kernel function K_z. Our approach is to first derive a formula for $\partial_{Q(\bar{z}|\bar{x})}G$, and then to compute $\partial_{Q^t(\bar{z}^t|\bar{x}^t)}G$ by applying the chain rule.

Define $b_i := \langle w(Q),\, y_i\Phi_Q(x_i)\rangle$. Using Theorem 23.8 of Rockafellar [Rockafellar, 1970], the subdifferential $\partial_{Q(\bar{z}|\bar{x})}G$ evaluated at $(w(Q); Q)$ can be expressed as

$$\partial_{Q(\bar{z}|\bar{x})}G \;=\; \sum_{i=1}^n \partial_{Q(\bar{z}|\bar{x})}g_i \;=\; \sum_{i=1}^n \partial\phi(b_i)y_i\langle w,\, \Phi'(\bar{z})\rangle\mathbb{I}[x_i = \bar{x}].$$

Earlier in the proof of Proposition 3.5 we proved that $-\lambda\alpha_i \in \partial\phi(b_i)$ for each $i = 1,\ldots,n$, where α is the optimal solution of (3.21). Therefore, $\partial_{Q(\bar{z}|\bar{x})}G$ evaluated at $(w(Q); Q)$ contains the element:

$$\sum_{i=1}^n -\lambda\alpha_i y_i\langle w(Q),\, \Phi'(\bar{z})\rangle\mathbb{I}[x_i = \bar{x}] \;=\; \sum_{i,j} -\lambda\alpha_i\alpha_j y_i y_j\mathbb{I}[x_i = \bar{x}]\sum_z K(z,\bar{z})Q(z|x_j).$$

For each $t = 1, \ldots, S$, $\partial_{Q^t(\bar{z}^t|\bar{x}^t)} G$ is related to $\partial_{Q(\bar{z}|\bar{x})} G$ by the chain rule. Note that for $Q(\bar{z}|\bar{x}) = \prod_{t=1}^{S} Q^t(\bar{z}^t|\bar{x}^t)$, we have

$$\partial_{Q^t(\bar{z}^t|\bar{x}^t)} G \;=\; \sum_{z,x} \partial_{Q^t(\bar{z}^t|\bar{x}^t)} Q(z|x) \partial_{Q(z|x)} G \;=\; \sum_{z,x} \frac{Q(z|x)}{Q^t(\bar{z}^t|\bar{x}^t)} \mathbb{I}[x^t = \bar{x}^t] \mathbb{I}[z^t = \bar{z}^t] \partial_{Q(z|x)} G,$$

which contains the following element as one of its subgradients:

$$\begin{aligned} &\sum_{z,x} \frac{Q(z|x)}{Q^t(\bar{z}^t|\bar{x}^t)} \mathbb{I}[x^t = \bar{x}^t] \mathbb{I}[z^t = \bar{z}^t] \left\{ \sum_{i,j} -\lambda \alpha_i \alpha_j y_i y_j \mathbb{I}[x_i = x] \sum_{z'} K_z(z', z) Q(z'|x_j) \right\} \\ &= \sum_{i,j,z,z'} -\lambda \alpha_i \alpha_j y_i y_j \mathbb{I}[x_i^t = \bar{x}^t] \mathbb{I}[z^t = \bar{z}^t] \frac{Q(z|x_i)}{Q^t(\bar{z}^t|\bar{x}_i^t)} Q(z'|x_j) K_z(z', z). \end{aligned}$$

This completes the proof of the lemma.

Appendix 3.D Proof of Proposition 3.9

By definition [van der Vaart and Wellner, 1996], the Rademacher complexity $R_n(\mathcal{F}_0)$ is given by

$$\begin{aligned} \mathbb{E} \sup_{f \in \mathcal{F}_0} \frac{2}{n} \sum_{i=1}^{n} \sigma_i f(X_i) &= \mathbb{E} \sup_{\|w\| \le B; Q \in \mathcal{Q}_0} \frac{2}{n} \sum_{i=1}^{n} \sigma_i \langle w, \, \Phi_Q(X_i) \rangle \\ &= \frac{2B}{n} \mathbb{E} \sup_{Q \in \mathcal{Q}_0} \| \sum_{i=1}^{n} \sigma_i \Phi_Q(X_i) \|. \end{aligned}$$

Applying the Cauchy-Schwarz inequality yields that $R_n(\mathcal{F}_0)$ is upper bounded as

$$\begin{aligned} &\frac{2B}{n} \sqrt{\mathbb{E} \sup_{Q \in \mathcal{Q}_0} || \sum_{i=1}^{n} \sigma_i \Phi_Q(X_i) ||^2} \\ &\qquad = \frac{2B}{n} \left(\mathbb{E} \sup_{Q \in \mathcal{Q}_0} \sum_{i=1}^{n} K_Q(X_i, X_i) + 2\mathbb{E} \sup_{Q \in \mathcal{Q}_0} \sum_{1 \le i < j \le n} \sigma_i \sigma_j K_Q(X_i, X_j) \right)^{1/2}. \end{aligned}$$

It remains to upper bound the second term inside the square root in the RHS. The trick is to partition the $n(n-1)/2$ pairs of (i, j) into $n - 1$ subsets each of which has $n/2$ pairs of different i and j (assuming n is even for simplicity). The existence of such a partition can be shown by induction on n. Now, for each $i = 1, \ldots, n-1$, denote the subset indexed by

i by $n/2$ pairs $(\pi_i(j), \pi'_i(j))_{j=1}^{n/2}$, where all

$$\{\pi_i(1), \ldots, \pi_i(n/2)\} \cap \{\pi'_i(1), \ldots, \pi'_i(n/2)\} = \emptyset.$$

Therefore,

$$\begin{aligned}
\mathbb{E} \sup_{Q \in \mathcal{Q}_0} \sum_{1 \le i < j \le n} \sigma_i \sigma_j K_Q(X_i, X_j) &= \mathbb{E} \sup_{Q \in \mathcal{Q}_0} \sum_{i=1}^{n-1} \sum_{j=1}^{n/2} \sigma_{\pi_i(j)} \sigma_{\pi'_i(j)} K_Q(X_{\pi_i(j)}, X_{\pi'_i(j)}) \\
&\le \sum_{i=1}^{n-1} \mathbb{E} \sup_{Q \in \mathcal{Q}_0} \sum_{j=1}^{n/2} \sigma_{\pi_i(j)} \sigma_{\pi'_i(j)} K_Q(X_{\pi_i(j)}, X_{\pi'_i(j)}).
\end{aligned}$$

Our final step is to bound the terms inside the summation over i by invoking Massart's lemma [Massart, 2000] for bounding Rademacher averages over a finite set $A \subset \mathbb{R}^d$ to conclude that $\mathbb{E} \sup_{a \in A} \sum_{i=1}^d \sigma_i a_i \le \max ||a||_2 \sqrt{2 \log |A|}$. Now, for each i and a realization of $X_1, \ldots, X_n$, treat $\sigma_{\pi_i(j)} \sigma_{\pi'_i(j)}$ for $j = 1, \ldots, n/2$ as $n/2$ Rademacher variables, and the $n/2$ dimensional vector $(K_Q(X_{\pi_i(j)}, X_{\pi'_i(j)}))_{j=1}^{n/2}$ takes on only L^{MS} possible values (since there are L^{MS} possible choices for $Q \in \mathcal{Q}_0$). Then we have,

$$\mathbb{E} \sup_{Q \in \mathcal{Q}_0} \sum_{j=1}^{n/2} \sigma_{\pi_i(j)} \sigma_{\pi'_i(j)} K_Q(X_{\pi_i(j)}, X_{\pi'_i(j)}) \le \sqrt{n/2} \sup_{z,z'} K_z(z, z') \sqrt{2 \log(L^{MS})},$$

from which the lemma follows.

Appendix 3.E Proof of Proposition 3.10

We treat each $Q(Z|X) \in \mathcal{Q}$ as a function over all possible values (z, x). Recall that X is an S-dimensional vector $X = (X^1, \ldots, X^S)$. For each fixed realization x^t of X^t, for $t = 1, \ldots, S$, the set of all discrete conditional probability distributions $Q(Z^t|x^t)$ is a $(L-1)$ simplex Δ_L. Since each X^t takes on M possible values, and X has S dimensions, we have: $N(\epsilon, \mathcal{Q}, L_\infty) \le N(\epsilon, \Delta_L, l_\infty)^{MS} \le (1/\epsilon)^{(L-1)MS}$. Recall that each $f \in \mathcal{F}$ can be written as:

$$f(x) = \sum_{i=1}^n \alpha_i \sum_{z,z_i} Q(z|x) Q(z_i|x_i) K_z(z, z_i). \tag{3.32}$$

We now define $\epsilon_0 := \epsilon\, [2L^S \sup ||\alpha||_1 \sup_{z,z'} K_z(z, z')]^{-1}$. Given each fixed conditional distribution Q in the ϵ_0-covering $G(\epsilon_0, \mathcal{Q}, L_\infty)$ for $\mathcal{Q}$, we can construct an $\epsilon/2$-covering in $L_2(P_n)$ for $\mathcal{F}_Q$. It is straightforward to verify that the union of all coverings for $\mathcal{F}_Q$ indexed by $Q \in G(\epsilon_0, \mathcal{Q}, L_\infty)$ forms an ϵ-covering for $\mathcal{F}$. Indeed, given any function

$f \in \mathcal{F}$ that is expressed in the form (3.32) with a corresponding $Q \in \mathcal{Q}$, there exists some $Q^* \in G(\epsilon_0, \mathcal{Q}, L_\infty)$ such that $\|Q - Q^*\|_\infty \leq \epsilon_0$. Let f_1 be a function in $\mathcal{F}_{Q^*}$ using the same coefficients α as those of f. Given Q^* there exists some $f_2 \in \mathcal{F}_{Q^*}$ such that $\|f_1 - f_2\|_{L_2(P_n)} \leq \epsilon/2$. The triangle inequality yields that $\|f - f_2\|_{L_2(P_n)}$ is upper bounded by

$$\begin{aligned}\|f - f_1\|_{L_2(P_n)} + \|f_1 - f_2\|_{L_2(P_n)} &\leq \|f - f_1\|_\infty + \epsilon/2 \\ &\leq L^S \sup ||\alpha||_1 \sup_{z,z'} K_z(z,z') \|Q - Q^*\|_\infty + \epsilon/2,\end{aligned}$$

which is less than ϵ. In summary, we have constructed an ϵ-covering in $L_2(P_n)$ for $\mathcal{F}$ whose number of coverings is no more than $N(\epsilon_0, \mathcal{Q}, L_\infty) \sup_Q N(\epsilon/2, \mathcal{F}_Q, L_2(P_n))$. This implies that

$$\begin{aligned}&\log N(\epsilon, \mathcal{F}, L_2(P_n)) \leq \log\left\{ N(\epsilon_0, \mathcal{Q}, L_\infty) \sup_Q N(\epsilon/2, \mathcal{F}_Q, L_2(P_n)) \right\} \\ &\leq \log\left\{ \left(\frac{2L^S \sup ||\alpha||_1 \sup_{z,z'} K_z(z,z')}{\epsilon} \right)^{(L-1)MS} \sup_Q N(\epsilon/2, \mathcal{F}_Q, L_2(P_n)) \right\} \\ &= \sup_{Q \in \mathcal{Q}} \log N(\epsilon/2, \mathcal{F}_Q, L_2(P_n)) + (L-1)MS \log \frac{2L^S \sup ||\alpha||_1 \sup_{z,z'} K_z(z,z')}{\epsilon},\end{aligned}$$

which completes the proof.

Chapter 4

Surrogate convex losses and f-divergences

In this chapter we develop a general correspondence between a family of loss functions that act as surrogates to 0-1 loss, and the class of Ali-Silvey or f-divergence functionals. This correspondence provides the basis for choosing and evaluating various surrogate losses frequently used in statistical learning (e.g., hinge loss, exponential loss, logistic loss); conversely, it provides a decision-theoretic framework for the choice of divergences in signal processing and quantization theory. We exploit this correspondence to characterize the statistical behavior of the nonparametric decentralized detection algorithm described in Chapter 3 that operate by minimizing convex surrogate loss functions. In particular, we specify the family of loss functions that are equivalent to 0-1 loss in the sense of producing the same quantization rules and discriminant functions.

4.1 Introduction

Over the past several decades, the classical topic of discriminant analysis has undergone significant and sustained development in various scientific and engineering fields. Much of this development has been driven by the physical, informational and computational constraints imposed by specific problem domains. Incorporating such constraints leads to interesting extensions of the basic discriminant analysis paradigm that involve aspects of experimental design. As one example, research in the area of "decentralized detection" focuses on problems in which measurements are collected by a collection of devices distributed over space (e.g., arrays of cameras, acoustic sensors, wireless nodes). Due to power and bandwidth limitations, these devices cannot simply relay their measurements to the common site where a hypothesis test is to be performed; rather, the measurements must be compressed prior to transmission, and the statistical test at the central site is performed on the transformed data [Tsitsiklis, 1993b; Blum *et al.*, 1997]. The problem of designing such compression rules is of substantial current interest in the field of sensor networks [Chong and Kumar, 2003; Chamberland and Veeravalli, 2003]. A closely related set of "signal selection" problems, arising for instance in radar array processing, also blend discriminant analysis with aspects of experimental design [Kailath, 1967].

The standard formulation of these problems—namely, as hypothesis-testing within either a Neyman-Pearson or Bayesian framework—rarely leads to computationally tractable algorithms. The main source of difficulty is the intractability of minimizing the probability of error, whether as a functional of the discriminant function or of the compression rule. Consequently, it is natural to consider loss functions that act as surrogates for the probability of error, and lead to practical algorithms. For example, the Hellinger distance has been championed for decentralized detection problems [Longo *et al.*, 1990], due to the fact that it yields a tractable algorithm both for the experimental design aspect of the problem (i.e., the choice of compression rule) and the discriminant analysis aspect of the problem. More broadly, a class of functions known as *Ali-Silvey distances* or *f-divergences* [Ali and Silvey, 1966; Csiszaŕ, 1967]—which includes not only the Hellinger distance, but also the variational distance, Kullback-Leibler (KL) divergence and Chernoff distance—have been explored as surrogate loss functions for the probability of error in a wide variety of applied discrimination problems.

Theoretical support for the use of f-divergences in discrimination problems comes from two main sources. First, a classical result of [Blackwell, 1951] asserts that if procedure A has a smaller f-divergence than procedure B (for some particular f-divergence), then there exists some set of prior probabilities such that procedure A has a smaller probability of error than procedure B. This fact, though a relatively weak justification, has nonetheless proven useful in designing signal selection and quantization rules [Kailath, 1967; Poor and Thomas, 1977; Longo *et al.*, 1990]. Second, f-divergences often arise as exponents in asymptotic (large-deviation) characterizations of the optimal rate of convergence in hypothesis-testing problems; examples include Kullback-Leibler divergence for the Neyman-Pearson formulation, and the Chernoff distance for the Bayesian formulation [Cover and Thomas, 1991].

A parallel and more recent line of research in the field of statistical machine learning has also focused on computationally-motivated surrogate functions in discriminant analysis. In statistical machine learning, the formulation of the discrimination problem (also known as *classification*) is decision-theoretic, with the Bayes error interpreted as risk under a 0-1 loss. The algorithmic goal is to design discriminant functions by minimizing the empirical expectation of 0-1 loss, wherein empirical process theory provides the underlying analytic framework. In this setting, the non-convexity of the 0-1 loss renders intractable a direct minimization of probability of error, so that various researchers have studied algorithms based on replacing the 0-1 loss with "surrogate loss functions." These alternative loss functions are convex, and represent upper bounds or approximations to the 0-1 loss (see Figure 4.2 for an illustration). A wide variety of practically successful machine learning algorithms are based on such a strategy, including support vector machines [Cortes and Vapnik, 1995; Schölkopf and Smola, 2002], the AdaBoost algorithm [Freund and Schapire, 1997], the X4 method [Breiman, 1998], and logistic regression [Friedman *et al.*, 2000]. More recent work by [Jiang, 2004], [Lugosi and Vayatis, 2004], [Mannor *et al.*, 2003], [Zhang, 2004], [Bartlett *et al.*, 2006], [Steinwart, 2005] and others provides theoretical support for these algorithms, in particular by characterizing statistical consistency and convergence rates of the resulting estimation procedures in terms of the properties of surrogate loss functions.

4.1.1 Our contributions

As mathematical objects, the f-divergences studied in information theory and the surrogate loss functions studied in statistical machine learning are fundamentally different: the former are functions on pairs of measures, whereas the latter are functions on values of discriminant functions and class labels. However, their underlying role in obtaining computationally-tractable algorithms for discriminant analysis suggests that they should be related. Indeed, Blackwell's result hints at such a relationship, but its focus on 0-1 loss does not lend itself to developing a general relationship between f-divergences and surrogate loss functions. The primary contribution of this chapter is to provide a detailed analysis of the relationship between f-divergences and surrogate loss functions, developing a full characterization of the connection, and explicating its consequences. We show that for any surrogate loss, regardless of its convexity, there exists a corresponding convex f such that minimizing the expected loss is equivalent to maximizing the f-divergence. We also provide necessary and sufficient conditions for an f-divergence to be realized from some (decreasing) convex loss function. More precisely, given a convex f, we provide a constructive procedure to generate *all* decreasing convex loss functions for which the correspondence holds.

The relationship is illustrated in Figure 4.1; whereas each surrogate loss ϕ induces only one f-divergence, note that in general there are many surrogate loss functions that correspond to the same f-divergence. As particular examples of the general correspondence established in this chapter, we show that the hinge loss corresponds to the variational distance, the exponential loss corresponds to the Hellinger distance, and the logistic loss corresponds to the capacitory discrimination distance.

This correspondence—in addition to its intrinsic interest as an extension of Blackwell's work—has several specific consequences. First, there are numerous useful inequalities relating the various f-divergences [Topsoe, 2000]; our theorem allows these inequalities to be exploited in the analysis of loss functions. Second, the minimizer of the Bayes error

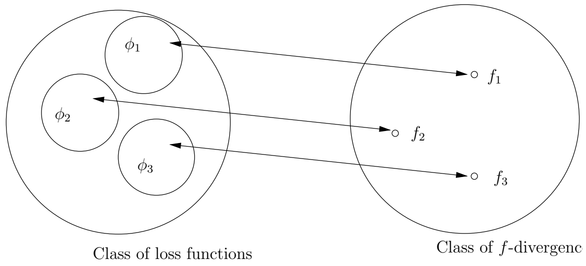


Figure 4.1. Illustration of the correspondence between f-divergences and loss functions. For each loss function ϕ, there exists exactly one corresponding f-divergence (induced by some underlying convex function f) such that the ϕ-risk is equal to the negative f-divergence. Conversely, for each f-divergence, there exists a whole set of surrogate loss functions ϕ for which the correspondence holds. Within the class of convex loss functions and the class of f-divergences, one can construct equivalent loss functions and equivalent f-divergences, respectively. For the class of classification-calibrated decreasing convex loss functions, we can characterize the correspondence precisely.

and the maximizer of f-divergences are both known to possess certain extremal properties [Tsitsiklis, 1993a]; our result provides a natural connection between these properties. Third, our theorem allows a notion of equivalence to be defined among loss functions: in particular, we say that loss functions are equivalent if they induce the same f-divergence. We then exploit the constructive nature of our theorem to exhibit all possible convex loss functions that are equivalent (in the sense just defined) to the 0-1 loss. Finally, we illustrate the application of this correspondence to the problem of decentralized detection. Whereas the more classical approach to this problem is based on f-divergences [Kailath, 1967; Poor and Thomas, 1977], our method instead builds on the framework of statistical machine learning. The correspondence allows us to establish consistency results for a novel algorithmic framework for decentralized detection: in particular, we prove that for any surrogate loss function equivalent to 0-1 loss, our estimation procedure is consistent in the strong sense that it will asymptotically choose Bayes-optimal quantization rules.

The remainder of the chapter is organized as follows. In Section 4.2, we define a version of discriminant analysis that is suitably general so as to include problems that involve a component of experiment design (such as in decentralized detection, and signal selection). We also provide a formal definition of surrogate loss functions, and present examples of optimized risks based on these loss functions. In Section 4.3, we state and prove the correspondence theorem between surrogate loss functions and f-divergences. Section 4.4 illustrates the correspondence using well-known examples of loss functions and their f-divergence counterparts. In Section 4.5, we discuss connections between the choice of quantization designs and Blackwell's classic results on comparisons of experiments. We introduce notions of equivalence among surrogate loss functions, and explore their properties. In Section 4.6, we establish the consistency of schemes for choosing Bayes-optimal classifiers based on surrogate loss functions that are equivalent to 0-1 loss. We conclude with a discussion in Section 4.7.

4.2 Background and problem set-up

4.2.1 Binary classification and its extension

We begin by defining a classical discriminant analysis problem, in particular the *binary classification problem*. Let X be a covariate taking values in a compact topological space $\mathcal{X}$, and let $Y \in \mathcal{Y} := \{-1, +1\}$ be a binary random variable. The product space $(X \times Y)$ is assumed to be endowed with a Borel regular probability measure $\mathbb{P}$. A *discriminant function* is a measurable function f mapping from $\mathcal{X}$ to the real line, whose sign is used to make a classification decision. The goal is to choose the discriminant function f so as to minimize the probability of making the incorrect classification, also known as the *Bayes risk*. This risk is defined as follows

$$\mathbb{P}(Y \neq \text{sign}(f(X))) = \mathbb{E}\big[\mathbb{I}[Y \neq \text{sign}(f(X))\big], \tag{4.1}$$

where $\mathbb{I}$ is a 0-1-valued indicator function.

The focus of this chapter is an elaboration of this basic problem in which the decision-maker, rather than having direct access to X, observes a random variable Z with range $\mathcal{Z}$ that is obtained via a (possibly stochastic) mapping $Q : \mathcal{X} \to \mathcal{Z}$. In a statistical context, the choice of the mapping Q can viewed as choosing a particular *experiment*; in the signal processing literature, where $\mathcal{Z}$ is generally taken to be discrete, the mapping Q is often referred to as a *quantizer*. In any case, the mapping Q can be represented by conditional probabilities $Q(z|x)$.

Let $\mathcal{Q}$ denote the space of all stochastic Q, and let $\mathcal{Q}_0$ denote the subset of deterministic mappings. When the underlying experiment Q is fixed, then we simply have a binary classification problem on the space $\mathcal{Z}$: that is, our goal is to find a real-valued measurable function γ on $\mathcal{Z}$ so as to minimize the Bayes risk $\mathbb{P}(Y \neq \text{sign}(\gamma(Z)))$. We use Γ to represent the space of all such possible discriminant functions on $\mathcal{Z}$. This chapter is motivated by the problem of specifying the classifier $\gamma \in \Gamma$, as well as the experiment choice $Q \in \mathcal{Q}$, so as to minimize the Bayes risk.

Throughout the chapter, we assume that $\mathcal{Z}$ is a discrete space for simplicity. We note in passing that this requirement is not essential to our analysis. It is only needed in Section 4.6, where we require that $\mathcal{Q}$ and $\mathcal{Q}_0$ be compact. This condition is satisfied when Z is discrete.

4.2.2 Surrogate loss functions

As shown in equation (4.1), the Bayes risk corresponds to the expectation of the 0-1 loss

$$\phi(y, \gamma(z)) = \mathbb{I}[y \neq \operatorname{sign}(\gamma(z))]. \tag{4.2}$$

Given the nonconvexity of this loss function, it is natural to consider a surrogate loss function ϕ that we optimize in place of the 0-1 loss. In particular, we focus on loss functions of the form $\phi(y, \gamma(z)) = \phi(y\gamma(z))$, where $\phi : \mathbb{R} \to \mathbb{R}$ is typically a convex upper bound on the 0-1 loss. In the statistical learning literature, the quantity $y\gamma(z)$ is known as the *margin* and $\phi(y\gamma(z))$ is often referred to as a "margin-based loss function." Given a particular loss function ϕ, we denote the associated ϕ*-risk* by $R_\phi(\gamma, Q) := \mathbb{E}\phi(Y\gamma(Z))$.

A number of such loss functions are used commonly in the statistical learning literature. See Figure 4.2 for an illustration of some different surrogate functions, as well as the original 0-1 loss. First, the *hinge loss* function

$$\phi_{hinge}(y\gamma(z)) \quad := \quad \max\{1 - y\gamma(z), 0\} \tag{4.3}$$

underlies the so-called support vector machine (SVM) algorithm [Schölkopf and Smola, 2002]. Second, the *logistic loss* function

$$\phi_{log}(y\gamma(z)) \quad := \quad \log\left(1 + \exp^{-y\gamma(z)}\right) \tag{4.4}$$

forms the basis of logistic regression [Friedman *et al.*, 2000]. As a third example, the Adaboost algorithm [Freund and Schapire, 1997] uses a *exponential loss* function:

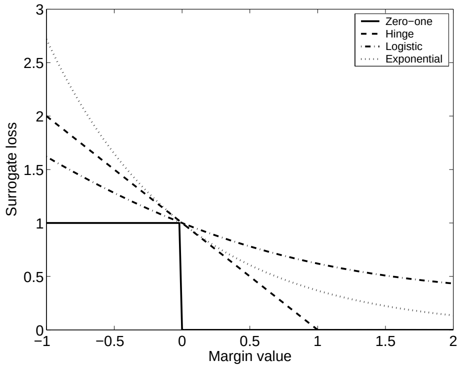


Figure 4.2. Illustrations of the 0-1 loss function, and three surrogate loss functions: hinge loss, logistic loss, and exponential loss.

$$\phi_{exp}(y\gamma(z)) \quad := \quad \exp(-y\gamma(z)). \tag{4.5}$$

Finally, another possibility (though less natural for a classification problem) is the *least squares* function:

$$\phi_{sqr}(y\gamma(z)) := (1 - y\gamma(z))^2. \tag{4.6}$$

[Bartlett *et al.*, 2006] have provided a general definition of surrogate loss functions. Their definition is crafted so as to permit the derivation of a general bound that links the ϕ-risk and the Bayes risk, thereby permitting an elegant general treatment of the consistency of estimation procedures based on surrogate losses. The definition is essentially a pointwise form of a Fisher consistency condition that is appropriate for the classification setting; in particular, it takes the following form:

Definition 4.1. *A loss function ϕ is* classification-calibrated *if for any $a, b \geq 0$ and $a \neq b$:*

$$\inf_{\{\alpha \in \mathbb{R} \,|\, \alpha\,(a-b)<0\}} \big[\phi(\alpha)a + \phi(-\alpha)b\big] \quad > \quad \inf_{\alpha \in \mathbb{R}} \big[\phi(\alpha)a + \phi(-\alpha)b\big]. \tag{4.7}$$

As will be clarified subsequently, this definition ensures that the decision rule γ behaves equivalently to the Bayes decision rule in the (binary) classification setting.

For our purposes we will find it useful to consider a somewhat more restricted definition of surrogate loss functions. In particular, we impose the following three conditions on any surrogate loss function $\phi : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$:

A1: ϕ is classification-calibrated.

A2: ϕ is continuous and convex.

A3: Let $\alpha^* = \inf\big\{\alpha \in \mathbb{R} \cup \{+\infty\} \;\big|\; \phi(\alpha) = \inf \phi\big\}$. If $\alpha^* < +\infty$, then for any $\epsilon > 0$,

$$\phi(\alpha^* - \epsilon) \geq \phi(\alpha^* + \epsilon). \tag{4.8}$$

The interpretation of Assumption A3 is that one should penalize deviations away from α^* in the negative direction at least as strongly as deviations in the positive direction; this requirement is intuitively reasonable given the margin-based interpretation of α. Moreover, this assumption is satisfied by all of the loss functions commonly considered in the literature; in particular, any decreasing function ϕ (e.g., hinge loss, logistic loss, exponential loss) satisfies this condition, as does the least squares loss (which is not decreasing).

[Bartlett *et al.*, 2006] also derived a simple lemma that characterizes classification-calibration for convex functions:

Lemma 4.2. *Let ϕ be a convex function. Then ϕ is classification-calibrated if and only if it is differentiable at 0 and $\phi'(0) < 0$.*

Consequently, Assumption A1 is equivalent to requiring that ϕ be differentiable at 0 and $\phi'(0) < 0$. These facts also imply that the quantity α^* defined in Assumption A3 is strictly positive. Finally, although ϕ is not defined for $-\infty$, we shall use the convention that $\phi(-\infty) = +\infty$.

4.2.3 Examples of optimum ϕ-risks

For each fixed experiment Q, we define the *optimal ϕ-risk* (a function of Q) as follows:

$$R_\phi(Q) := \inf_{\gamma \in \Gamma} R_\phi(\gamma, Q). \tag{4.9}$$

Let $p = \mathbb{P}(Y = 1)$ and $q = \mathbb{P}(Y = -1)$, where $p, q > 0$ and $p + q = 1$, define a prior on the hypothesis space. Any fixed experiment Q induces positive measures μ and π over $\mathcal{Z}$ as follows:

$$\mu(z) \;:=\; \mathbb{P}(Y = 1, Z = z) = p \int_x Q(z|x) d\mathbb{P}(x|Y = 1) \tag{4.10a}$$

$$\pi(z) \;:=\; \mathbb{P}(Y = -1, Z = z) = q \int_x Q(z|x) d\mathbb{P}(x|Y = -1). \tag{4.10b}$$

The integrals are defined with respect to a dominating measure, e.g., $\mathbb{P}(x|Y = 1) + \mathbb{P}(x|Y = -1)$. It can be shown using Lyapunov's theorem that the space of $\{(\mu, \pi)\}$ by varying $Q \in \mathcal{Q}$ (or $\mathcal{Q}_0$) is both convex and compact under an appropriately defined topology(see, [Tsitsiklis, 1993a]).

For simplicity, in this chapter, we assume that the spaces $\mathcal{Q}$ and $\mathcal{Q}_0$ are restricted such that both μ and π are strictly positive measures. Note that the measures μ and π are constrained by the following simple relations:

$$\sum_{z \in \mathcal{Z}} \mu(z) \;=\; \mathbb{P}(Y = 1), \sum_{z \in \mathcal{Z}} \pi(z) \;=\; \mathbb{P}(Y = -1), \text{and} \mu(z) + \pi(z) = \mathbb{P}(z) \text{ for each } z \in \mathcal{Z}\,.$$

Note that Y and Z are independent conditioned on X. Therefore, letting $\eta(x) = \mathbb{P}(Y = 1|x)$, we can write

$$R_\phi(\gamma, Q) = \mathbb{E}_X \big[\sum_z \phi(\gamma(z)) \eta(X) Q(z|X) + \phi(-\gamma(z))(1 - \eta(X)) Q(z|X) \big]. \tag{4.11}$$

On the basis of this equation, the ϕ-risk can be written in the following way:

$$\begin{aligned} R_\phi(\gamma, Q) &= \mathbb{E}\phi(Y\gamma(Z)) && (4.12) \\ &= \sum_z \phi(\gamma(z))\mathbb{E}_X\left[\eta(X)Q(z|X)\right] + \phi(-\gamma(z))\mathbb{E}_X\left[(1-\eta(X))Q(z|X)\right] \\ &= \sum_z \phi(\gamma(z))\mu(z) + \phi(-\gamma(z))\pi(z). && (4.13) \end{aligned}$$

This representation allows us to compute the optimal value for $\gamma(z)$ for all $z \in \mathcal{Z}$, as well as the optimal ϕ-risk for a fixed Q. We illustrate this procedure with some examples:

0-1 loss. In this case, it is straightforward to see from equation (4.12) that $\gamma(z) = \text{sign}(\mu(z) - \pi(z))$. As a result, the optimal Bayes risk given a fixed Q takes the form:

$$\begin{aligned} R_{bayes}(Q) &= \sum_{z\in\mathcal{Z}} \min\{\mu(z), \pi(z)\} = \frac{1}{2} - \frac{1}{2}\sum_{z\in\mathcal{Z}} |\mu(z) - \pi(z)| \\ &= \frac{1}{2}(1 - V(\mu, \pi)), \end{aligned}$$

where $V(\mu, \pi)$ denotes the variational distance $V(\mu, \pi) := \sum_{z\in\mathcal{Z}} |\mu(z) - \pi(z)|$ between the two measures μ and π.

Hinge loss. If ϕ is hinge loss, then equation (4.12) again yields that $\gamma(z) = \text{sign}(\mu(z) - \pi(z))$. As a result, the optimal risk for hinge loss takes the form:

$$\begin{aligned} R_{hinge}(Q) &= \sum_{z\in\mathcal{Z}} 2\min\{\mu(z), \pi(z)\} = 1 - \sum_{z\in\mathcal{Z}} |\mu(z) - \pi(z)| \\ &= 1 - V(\mu, \pi) = 2R_{bayes}(Q). \end{aligned}$$

Least squares loss. If ϕ is least squares loss, then $\gamma(z) = \frac{\mu(z)-\pi(z)}{\mu(z)+\pi(z)}$, so that the optimal risk for least squares loss takes the form:

$$\begin{aligned} R_{sqr}(Q) &= \sum_{z\in\mathcal{Z}} \frac{4\mu(z)\pi(z)}{\mu(z)+\pi(z)} = 1 - \sum_{z\in\mathcal{Z}} \frac{(\mu(z)-\pi(z))^2}{\mu(z)+\pi(z)} \\ &= 1 - \Delta(\mu, \pi), \end{aligned}$$

where $\Delta(\mu, \pi)$ denotes the *triangular discrimination* distance defined by $\Delta(\mu, \pi) := \sum_{z\in\mathcal{Z}} \frac{(\mu(z)-\pi(z))^2}{\mu(z)+\pi(z)}$.

Logistic loss. If ϕ is logistic loss, then $\gamma(z) = \log\frac{\mu(z)}{\pi(z)}$. As a result, the optimal risk for logistic loss takes the form:

$$\begin{aligned}
R_{log}(Q) &= \sum_{z \in \mathcal{Z}} \mu(z) \log \frac{\mu(z) + \pi(z)}{\mu(z)} + \pi(z) \log \frac{\mu(z) + \pi(z)}{\pi(z)} \\
&= \log 2 - KL(\mu || \frac{\mu + \pi}{2}) - KL(\pi || \frac{\mu + \pi}{2}) \\
&= \log 2 - C(\mu, \pi),
\end{aligned}$$

where $KL(U, V)$ denotes the Kullback-Leibler divergence between two measures U and V, and $C(U, V)$ denotes the *capacitory discrimination* distance defined by

$$C(U, V) \; := \; KL(U || \frac{U + V}{2}) + KL(V || \frac{U + V}{2}).$$

Exponential loss. If ϕ is exponential loss, then $\gamma(z) = \frac{1}{2} \log \frac{\mu(z)}{\pi(z)}$. The optimal risk for exponential loss takes the form:

$$\begin{aligned}
R_{exp}(Q) &= \sum_{z \in \mathcal{Z}} 2\sqrt{\mu(z)\pi(z)} = 1 - \sum_{z \in \mathcal{Z}} (\sqrt{\mu(z)} - \sqrt{\pi(z)})^2 \\
&= 1 - 2h^2(\mu, \pi),
\end{aligned}$$

where $h(\mu, \pi) := \frac{1}{2} \sum_{z \in \mathcal{Z}} (\sqrt{\mu(z)} - \sqrt{\pi(z)})^2$ denotes the Hellinger distance between measures μ and π.

It is worth noting that in all of these cases, the optimum ϕ-risk takes the form of a well-known "distance" or "divergence" function. This observation motivates a more general investigation of the relationship between surrogate loss functions and the form of the optimum risk.

4.3 Correspondence between surrogate loss functions and divergences

The correspondence exemplified in the previous section turns out to be quite general. So as to make this connection precise, we begin by defining the class of f*-divergence functions*, which includes all of the examples discussed above as well as numerous others [Csiszaŕ, 1967; Ali and Silvey, 1966]:

Definition 4.3. *Given any continuous convex function* $f : [0, +\infty) \to \mathbb{R} \cup \{+\infty\}$*, the*

f-divergence between measures μ and π is given by

$$I_f(\mu, \pi) := \sum_z \pi(z) f\left(\frac{\mu(z)}{\pi(z)}\right). \tag{4.14}$$

As particular cases, the variational distance is given by $f(u) = |u-1|$, Kullback-Leibler divergence by $f(u) = u \ln u$, triangular discrimination by $f(u) = (u-1)^2/(u+1)$, and Hellinger distance by $f(u) = \frac{1}{2}(\sqrt{u}-1)^2$. Other well-known f-divergences include the (negative) Bhattacharyya distance ($f(u) = -2\sqrt{u}$), and the (negative) harmonic distance ($f(u) = -\frac{4u}{u+1}$).

As discussed in the introduction, these functions are widely used in the engineering literature to solve problems in decentralized detection and signal selection. Specifically, for a pre-specified joint distribution $\mathbb{P}(X, Y)$ and a given quantizer Q, one defines an f-divergence on the class-conditional distributions $\mathbb{P}(Z|Y = 1)$ and $\mathbb{P}(Z|Y = -1)$. This f-divergence is then viewed as a function of the underlying Q, and the optimum quantizer is chosen by maximizing the f-divergence. Typically, the discriminant function γ—which acts on the quantized space $\mathcal{Z}$— has an explicit form in terms of the distributions $P(Z|Y = 1)$ and $P(Z|Y = -1)$. As we have discussed, the choice of the class of f-divergences as functions to optimize is motivated both by Blackwell's classical theorem [Blackwell, 1951] on the design of experiments, as well as by the computational intractability of minimizing the probability of error, a problem rendered particularly severe in practice when X is high dimensional [Kailath, 1967; Poor and Thomas, 1977; Longo *et al.*, 1990].

4.3.1 From ϕ-risk to f-divergence

In the following two sections, we develop a general relationship between optimal ϕ-risks and f-divergences. The easier direction, on which we focus in the current section, is moving from ϕ-risk to f-divergence. In particular, we begin with a simple result that shows that any ϕ-risk induces a corresponding f-divergence.

Proposition 4.4. *For each fixed Q, let γ_Q be the optimal decision rule for the fusion center. Then the ϕ-risk for (Q, γ_Q) is a* f-divergence *between μ and π, as defined in equation* (4.10)*, for some convex function f:*

$$R_\phi(Q) = -I_f(\mu, \pi). \tag{4.15}$$

Moreover, this relation holds whether or not ϕ is convex.

Proof. The optimal ϕ-risk has the form

$$R_\phi(Q) = \sum_{z \in \mathcal{Z}} \min_\alpha (\phi(\alpha)\mu(z) + \phi(-\alpha)\pi(z)) = \sum_z \pi(z) \min_\alpha \left(\phi(-\alpha) + \phi(\alpha)\frac{\mu(z)}{\pi(z)}\right).$$

For each z, define $u := \frac{\mu(z)}{\pi(z)}$. With this notation, the function $\min_\alpha(\phi(-\alpha) + \phi(\alpha)u)$ is concave as a function of u (since the minimum of a collection of linear functions is concave). Thus, if we define

$$f(u) := -\min_{\alpha}(\phi(-\alpha) + \phi(\alpha)u). \tag{4.16}$$

then the claim follows. Note that the argument does not require convexity of ϕ. □

Remark: We can also write $I_f(\mu, \pi)$ in terms of an f-divergence between the two conditional distributions $\mathbb{P}(Z|Y = 1) \sim \mathbb{P}_1(Z)$ and $\mathbb{P}(Z|Y = -1) \sim \mathbb{P}_{-1}(Z)$. Recalling the notation $q = \mathbb{P}(Y = -1)$, we have:

$$I_f(\mu, \pi) = q \sum_z \mathbb{P}_{-1}(z) f\left(\frac{(1-q)\mathbb{P}_1(z)}{q\mathbb{P}_{-1}(z)}\right) = I_{f_q}(\mathbb{P}_1, \mathbb{P}_{-1}), \tag{4.17}$$

where $f_q(u) := qf((1-q)u/q)$. Although it is equivalent to study either form of divergences, we focus primarily on the representation (4.15) because the prior probabilities are absorbed in the formula. It will be convenient, however, to use the alternative (4.17) when the connection to the general theory of comparison of experiments is discussed.

4.3.2 From f-divergence to ϕ-risk

In this section, we develop the converse of Proposition 4.4. Given a divergence $I_f(\mu, \pi)$ for some convex function f, does there exists a loss function ϕ for which $R_\phi(Q) = -I_f(\mu, \pi)$? We establish that such a correspondence indeed holds for a general class of margin-based convex loss functions; in such cases, it is possible to construct ϕ to induce a given f-divergence.

4.3.2.1 Some intermediate functions

Our approach to establishing the desired correspondence proceeds via some intermediate functions, which we define in this section. First, let us define, for each β, the inverse mapping

$$\phi^{-1}(\beta) := \inf\{\alpha : \phi(\alpha) \leq \beta\}, \tag{4.18}$$

where $\inf \emptyset := +\infty$. The following result summarizes some useful properties of ϕ^{-1}:

Lemma 4.5. *(a) For all $\beta \in \mathbb{R}$ such that $\phi^{-1}(\beta) < +\infty$, the inequality $\phi(\phi^{-1}(\beta)) \leq \beta$ holds. Furthermore, equality occurs when ϕ is continuous at $\phi^{-1}(\beta)$.*

(b) The function $\phi^{-1} : \mathbb{R} \to \overline{\mathbb{R}}$ is strictly decreasing and convex.

Proof. See Appendix 4.A. □

Using the function ϕ^{-1}, we then define a new function $\Psi : \mathbb{R} \to \overline{\mathbb{R}}$ by

$$\Psi(\beta) \;:=\; \begin{cases} \phi(-\phi^{-1}(\beta)) & \text{if } \phi^{-1}(\beta) \in \mathbb{R}, \\ +\infty & \text{otherwise.} \end{cases} \tag{4.19}$$

Note that the domain of Ψ is $\text{Dom}(\Psi) = \{\beta \in \mathbb{R} : \phi^{-1}(\beta) \in \mathbb{R}\}$. Now define

$$\beta_1 \;:=\; \inf\{\beta : \Psi(\beta) < +\infty\} \qquad \text{and} \qquad \beta_2 \;:=\; \inf\{\beta : \Psi(\beta) = \inf \Psi\}. \tag{4.20}$$

It is simple to check that $\inf \phi = \inf \Psi = \phi(\alpha^*)$, and $\beta_1 = \phi(\alpha^*)$, $\beta_2 = \phi(-\alpha^*)$. Furthermore, by construction, we have $\Psi(\beta_2) = \phi(\alpha^*) = \beta_1$, as well as $\Psi(\beta_1) = \phi(-\alpha^*) = \beta_2$. The following properties of Ψ are particularly useful for our main results.

Lemma 4.6. *(a) Ψ is strictly decreasing in (β_1, β_2). If ϕ is decreasing, then Ψ is also decreasing in $(-\infty, +\infty)$. In addition, $\Psi(\beta) = +\infty$ for $\beta < \beta_1$.*

(b) Ψ is convex in $(-\infty, \beta_2]$. If ϕ is a decreasing function, then Ψ is convex in $(-\infty, +\infty)$.

(c) Ψ is lower semi-continuous, and continuous in its domain.

(d) For any $\alpha \geq 0$, $\phi(\alpha) = \Psi(\phi(-\alpha))$. In particular, there exists $u^* \in (\beta_1, \beta_2)$ such that $\Psi(u^*) = u^*$.

(e) The function Ψ satisfies $\Psi(\Psi(\beta)) \leq \beta$ for all $\beta \in$ Dom(Ψ). Moreover, if ϕ is a continuous function on its domain $\{\alpha \in \mathbb{R} \,|\, \phi(\alpha) < +\infty\}$, then $\Psi(\Psi(\beta)) = \beta$ for all $\beta \in (\beta_1, \beta_2)$.

Proof. See Appendix 4.B. □

Remark: With reference to statement (b), if ϕ is not a decreasing function, then the function Ψ need not be convex on the entire real line. See Appendix 4.B for an example.

The following result provides the necessary connection between the function Ψ and the f-divergence associated with ϕ, as defined in equation (4.16):

Proposition 4.7. *(a) Given a loss function ϕ, the associated f-divergence (4.16) satisfies the relation*

$$f(u) = \Psi^*(-u), \tag{4.21}$$

where Ψ^* denotes the conjugate dual of Ψ. If the surrogate loss ϕ is decreasing, then $\Psi(\beta) = f^*(-\beta)$.

(b) For a given Ψ, there exists a point $u^* \in (\beta_1, \beta_2)$ such that $\Psi(u^*) = u^*$. All loss functions ϕ that induce Ψ via (4.19) take the form:

$$\phi(\alpha) \quad = \quad \begin{cases} u^* & \text{if } \alpha = 0 \\ \Psi(g(\alpha + u^*)) & \text{if } \alpha > 0 \\ g(-\alpha + u^*) & \text{if } \alpha < 0, \end{cases} \tag{4.22}$$

where $g : [u^*, +\infty) \to \overline{\mathbb{R}}$ is some increasing continuous convex function such that $g(u^*) = u^*$, and g is right-differentiable at u^* with $g'(u^*) > 0$.

Proof. (a) From equation (4.16), we have

$$f(u) \quad = \quad -\inf_{\alpha \in \mathbb{R}} \Big(\phi(-\alpha) + \phi(\alpha) u \Big) = - \inf_{\left\{\alpha, \beta \,\middle|\, \phi^{-1}(\beta) \in \mathbb{R}, \ \phi(\alpha) = \beta\right\}} \Big(\phi(-\alpha) + \beta u \Big).$$

For β such that $\phi^{-1}(\beta) \in \mathbb{R}$, there might be more than one α such that $\phi(\alpha) = \beta$. However, our assumption (4.8) ensures that $\alpha = \phi^{-1}(\beta)$ results in minimum $\phi(-\alpha)$. Hence,

$$\begin{aligned} f(u) \quad &= \quad -\inf_{\beta : \phi^{-1}(\beta) \in \mathbb{R}} \Big(\phi(-\phi^{-1}(\beta)) + \beta u \Big) = -\inf_{\beta \in \mathbb{R}} (\beta u + \Psi(\beta)) \\ &= \quad \sup_{\beta \in \mathbb{R}} (-\beta u - \Psi(\beta)) = \Psi^*(-u). \end{aligned}$$

If ϕ is decreasing, then Ψ is convex. By convex duality and the lower semicontinuity of Ψ (from Lemma 4.6), we can also write:

$$\Psi(\beta) = \Psi^{**}(\beta) = f^*(-\beta). \tag{4.23}$$

(b) From Lemma 4.6, we have $\Psi(\phi(0)) = \phi(0) \in (\beta_1, \beta_2)$. As a consequence, $u^* := \phi(0)$ satisfies the relation $\Psi(u^*) = u^*$. Since ϕ is decreasing and convex on the interval $(-\infty, 0]$, for any $\alpha \geq 0$, $\phi(-\alpha)$ can be written as the form:

$$\phi(-\alpha) = g(\alpha + u^*),$$

where g is some increasing convex function. From Lemma 4.6, we have $\phi(\alpha) = \Psi(\phi(-\alpha)) = \Psi(g(\alpha + u^*))$ for $\alpha \geq 0$. To ensure the continuity at 0, there holds $u^* = \phi(0) = g(u^*)$. To ensure that ϕ is classification-calibrated, we require that ϕ is differentiable at 0 and $\phi'(0) < 0$. These conditions in turn imply that g must be right-differentiable at u^* with $g'(u^*) > 0$. □

4.3.2.2 A converse theorem

One important aspect of Proposition 4.7(a) is that it suggests a route—namely via convex duality [Rockafellar, 1970]—to recover the function Ψ from f, assuming that Ψ is lower semi-continuous. We exploit this intuition in the following:

Theorem 4.8. *Given a lower semicontinuous convex function $f : \mathbb{R} \to \overline{\mathbb{R}}$, consider the function:*

$$\Psi(\beta) = f^*(-\beta). \tag{4.24}$$

Define $\beta_1 := \inf\{\beta : \Psi(\beta) < +\infty\}$ and $\beta_2 := \inf\{\beta : \Psi(\beta) \le \inf \Psi\}$, and suppose that Ψ is decreasing, and satisfies $\Psi(\Psi(\beta)) = \beta$ for all $\beta \in (\beta_1, \beta_2)$.

(a) Then any *continuous loss function ϕ of the form (4.22) must induce f-divergence with respect to f in the sense of (4.15) and (4.16).*

(b) Moreover, if Ψ is differentiable at the point $u^* \in (\beta_1, \beta_2)$ such that $\Psi(u^*) = u^*$, then any such ϕ is classification-calibrated.

Proof. (a) Since f is lower semicontinuous by assumption, convex duality allows us to write

$$f(u) = f^{**}(u) = \Psi^*(-u) = \sup_{\beta \in \mathbb{R}}(-\beta u - \Psi(\beta)) = -\inf_{\beta \in \mathbb{R}}(\beta u + \Psi(\beta)).$$

Proposition 4.7(b) guarantees that all convex loss function ϕ for which equations (4.15) and (4.16) hold must have the form (4.22). Note that Ψ is lower semicontinuous and convex by definition. It remains to show that any convex loss function ϕ of form (4.22) must be linked to Ψ via the relation

$$\Psi(\beta) = \begin{cases} \phi(-\phi^{-1}(\beta)) & \text{if } \phi^{-1}(\beta) \in \mathbb{R}, \\ +\infty & \text{otherwise.} \end{cases} \tag{4.25}$$

Since Ψ is assumed to be a decreasing function, the function ϕ defined in equation (4.22) is also a decreasing function. By assumption, we also have $\Psi(\Psi(\beta)) = \beta$ for any $\beta \in (\beta_1, \beta_2)$. Therefore, it is straightforward to verify that there exists $u^* \in (\beta_1, \beta_2)$ such that $\Psi(u^*) = u^*$. Using the value u^*, we divide our analysis into three cases:

- For $\beta \ge u^*$, there exists $\alpha \ge 0$ such that $g(\alpha + u^*) = \beta$. Choose the largest α that is so. From our definition of ϕ, $\phi(-\alpha) = \beta$. Thus $\phi^{-1}(\beta) = -\alpha$. It follows that $\phi(-\phi^{-1}(\beta)) = \phi(\alpha) = \Psi(g(\alpha + u^*)) = \Psi(\beta)$.
- For $\beta < \beta_1 = \inf_{u \in \mathbb{R}} \Psi(u)$, we have $\Psi(\beta) = +\infty$.

- Lastly, for $\beta_1 \leq \beta < u^* < \beta_2$, then there exists $\alpha > 0$ such that $g(\alpha + u^*) \in (u^*, \beta_2)$ and $\beta = \Psi(g(\alpha + u^*))$, which implies that $\beta = \phi(\alpha)$ from our definition. Choose that smallest α that satisfies these conditions. Then $\phi^{-1}(\beta) = \alpha$, and it follows that $\phi(-\phi^{-1}(\beta)) = \phi(-\alpha) = g(\alpha + u^*) = \Psi(\Psi(g(\alpha + u^*))) = \Psi(\beta)$, where we have used the fact that $g(\alpha + u^*) \in (\beta_1, \beta_2)$).

The proof is complete.

(b) From Lemma 4.6(e), we have $\Psi(\Psi(\beta)) = \beta$ for $\beta \in (\beta_1, \beta_2)$. This fact, in conjunction with the assumption that Ψ is differentiable at u^*, implies that $\Psi'(u^*) = -1$. Therefore, by choosing g to be differentiable at u^* with $g'(u^*) > 0$, as dictated by Proposition 4.7(b), ensures that ϕ is also differentiable at 0 and $\phi'(0) < 0$. Thus, by Lemma 4.2, the function ϕ is classification-calibrated. □

Remark: One interesting consequence of Theorem 4.8 that any f-divergence can be obtained from a fairly large set of surrogate loss functions. More precisely, from the procedure (4.22), we see that any valid ϕ is specified by a function g that need satisfy only a mild set of conditions. It is important to note that not all ϕ losses of the form (4.22) are convex, but they still satisfy (4.16). We illustrate this flexibility with several examples in Section 4.4.

4.3.2.3 Some additional properties

Theorem 4.8 provides one set of conditions for an f-divergence to be realized by some surrogate loss ϕ, as well as a constructive procedure for finding all such loss functions. The following result provides a related set of conditions that can be easier to verify. We say that an f-divergence is *symmetric* if $I_f(\mu, \pi) = I_f(\pi, \mu)$ for any measures μ and π. With this definition, we have the following:

Corollary 4.9. *The following are equivalent:*

(a) f is realizable by some surrogate loss function ϕ (via Proposition 4.4).

(b) f-divergence I_f is symmetric.

(c) For any $u > 0$, $f(u) = uf(1/u)$.

Proof. (a) ⇒ (b): From Proposition 4.4, we have the representation $R_\phi(Q) = -I_f(\mu, \pi)$. Alternatively, we can write:

$$R_\phi(Q) = \sum_z \mu(z) \min_\alpha \left(\phi(\alpha) + \phi(-\alpha) \frac{\pi(z)}{\mu(z)} \right) \; = \; -\sum_z \mu(z) f\left(\frac{\pi(z)}{\mu(z)} \right),$$

which is equal to $-I_f(\pi, \mu)$, thereby showing that the f-divergence is symmetric.

(b) ⇒ (c): By assumption, the following relation holds for any measures μ and π:

$$\sum_z \pi(z) f(\mu(z)/\pi(z)) = \sum_z \mu(z) f(\pi(z)/\mu(z)). \tag{4.26}$$

Take any instance of $z = l \in \mathcal{Z}$, and consider measures μ' and π', which are defined on the space $\mathcal{Z} - \{l\}$ such that $\mu'(z) = \mu(z)$ and $\pi'(z) = \pi(z)$ for all $z \in \mathcal{Z} - \{l\}$. Since Equation (4.26) also holds for μ' and π', it follows that

$$\pi(z) f(\mu(z)/\pi(z)) = \mu(z) f(\pi(z)/\mu(z))$$

for all $z \in \mathcal{Z}$ and any μ and π. Hence, $f(u) = uf(1/u)$ for any $u > 0$.

(c) ⇒ (a): It suffices to show that all sufficient conditions specified by Theorem 4.8 are satisfied.

Since any f-divergence is defined by applying f to a likelihood ratio (see definition (4.14)), we can assume $f(u) = +\infty$ for $u < 0$ without loss of generality. Since $f(u) = uf(1/u)$ for any $u > 0$, it can be verified using subdifferential calculus [Rockafellar, 1970] that for any $u > 0$, there holds:

$$\partial f(u) = f(1/u) + \partial f(1/u) \frac{-1}{u}. \tag{4.27}$$

Given some $u > 0$, consider any $v_1 \in \partial f(u)$. Combined with (4.27), we have

$$f(u) - v_1 u \in \partial f(1/u). \tag{4.28}$$

By definition of conjugate duality,

$$f^*(v_1) \quad = \quad v_1 u - f(u).$$

Define $\Psi(\beta) = f^*(-\beta)$. Then,

$$\begin{aligned} \Psi(\Psi(-v_1)) \quad &= \quad \Psi(f^*(v_1)) = \Psi(v_1 u - f(u)) \\ &= \quad f^*(f(u) - v_1 u) = \sup_{\beta \in \mathbb{R}} (\beta f(u) - \beta v_1 u - f(\beta)). \end{aligned}$$

Note that the supremum is achieved at $\beta = 1/u$ because of (4.28). Therefore, $\Psi(\Psi(-v_1)) = -v_1$ for any $v_1 \in \partial f(u)$ for $u > 0$. In other words, $\Psi(\Psi(\beta)) = \beta$ for any $\beta \in \{-\partial f(u), u > 0\}$. By convex duality, $\beta \in -\partial f(u)$ for some $u > 0$ if and only if $-u \in \partial\Psi(\beta)$ for some $u > 0$ [Rockafellar, 1970]. This condition on β is equivalent to $\partial\Psi(\beta)$ containing some negative value. This is satisfied by any $\beta \in (\beta_1, \beta_2)$. Hence, $\Psi(\Psi(\beta)) = \beta$ for $\beta \in (\beta_1, \beta_2)$. In addition, $f(u) = +\infty$ for $u < 0$, Ψ is a decreasing function. Now, as an application of Theorem 4.8, I_f is realizable by some (decreasing) surrogate loss function.

□

Remarks. It is worth noting that not all f-divergences are symmetric; well-known cases of asymmetric divergences include the Kullback-Leibler divergences $KL(\mu, \pi)$ and $KL(\pi, \mu)$, which correspond to the functions $f(u) = -\log u$ and $f(u) = u \log u$, respectively. Corollary 4.9 establishes that such asymmetric f-divergences cannot be generated by *any* (margin-based) surrogate loss function ϕ. Therefore, margin-based surrogate losses can be considered as symmetric loss functions. It is important to note that our analysis can be extended to show that asymmetric f-divergences can be realized by general (asymmetric) loss functions. Finally, from the proof of Corollary 4.9, it can be deduced that if an f-divergence is realized by some surrogate loss function, it is also realized by some decreasing surrogate loss function.

Most surrogate loss functions ϕ considered in statistical learning are bounded from below (e.g., $\phi(\alpha) \geq 0$ for all $\alpha \in \mathbb{R}$). The following result establishes a link between (un)boundedness and the properties of the associated f:

Corollary 4.10. *Assume that ϕ is a decreasing (continuous convex) loss function corresponding to an f-divergence, where f is a continuous convex function that is bounded from below by an affine function. Then ϕ is* unbounded *from below if and only if f is 1-coercive, i.e., $f(x)/||x|| \to +\infty$ as $||x|| \to \infty$.*

Proof. ϕ is unbounded from below if and only if $\Psi(\beta) = \phi(-\phi^{-1}(\beta)) \in \mathbb{R}$ for all $\beta \in \mathbb{R}$, which is equivalent to the dual function $f(\beta) = \Psi^*(-\beta)$ being 1-coercive(cf. [Hiriart-Urruty and Lemaréchal, 2001]). □

Therefore, for any decreasing and lower-bounded ϕ loss (which includes the hinge, logistic and exponential losses), the associated f-divergence is *not* 1-coercive. Other interesting f-divergences such as the *symmetric* KL divergence considered in [Bradt and Karlin, 1956] are 1-coercive, meaning that any associated surrogate loss ϕ cannot be bounded below. We illustrate such properties of f-divergences and their corresponding loss functions in the following section.

4.4 Examples of loss functions and f-divergences

In this section, we consider a number of specific examples in order to illustrate the correspondence developed in the previous section. As a preliminary remark, it is simple to check that if f_1 and f_2 are related by $f_1(u) = cf_2(u) + au + b$ for some constants $c > 0$ and a, b, then $I_{f_1}(\mu, \pi) = I_{f_2}(\mu, \pi) + a\mathbb{P}(Y = 1) + b\mathbb{P}(Y = -1)$. This relationship implies that the f-divergences I_{f_1} and I_{f_2}, when viewed as functions of Q, are equivalent (up to an additive constant). For this reason, in the following development, we consider divergences so related to be equivalent. We return to a more in-depth exploration of this notion of equivalence in Section 4.5.

Example 1 (Hellinger distance). The Hellinger distance is equivalent to the negative of the Bhattacharyya distance, which is an f-divergence with $f(u) = -2\sqrt{u}$ for $u \geq 0$. Let us augment the definition of f with $f(u) = +\infty$ for $u < 0$; doing so does not alter the Hellinger (or Bhattacharyya) distances. Following the constructive procedure of Theorem 4.8, we begin by recovering Ψ from f:

$$\Psi(\beta) = f^*(-\beta) = \sup_{u \in \mathbb{R}}(-\beta u - f(u)) = \begin{cases} 1/\beta & \text{when } \beta > 0 \\ +\infty & \text{otherwise.} \end{cases}$$

Thus, we see that $u^* = 1$. If we let $g(u) = u$, then a possible surrogate loss function that realizes the Hellinger distance takes the form:

$$\phi(\alpha) \quad = \quad \begin{cases} 1 & \text{if } \alpha = 0 \\ \frac{1}{\alpha+1} & \text{if } \alpha > 0 \\ -\alpha + 1 & \text{if } \alpha < 0. \end{cases}$$

On the other hand, if we set $g(u) = \exp(u - 1)$, then we obtain the exponential loss $\phi(\alpha) = \exp(-\alpha)$, agreeing with what was shown in Section 4.2.3. See Figure 4.4 for illustrations of these loss functions using difference choices of g.

Example 2 (Variational distance). In Section 4.2.3, we established that the hinge loss as well as the 0-1 loss both generate the variational distance. This f-divergence is based on the function $f(u) = -2\min(u, 1)$ for $u \geq 0$. As before, we can augment the definition by setting $f(u) = +\infty$ for $u < 0$, and then proceed to recover Ψ from f:

$$\Psi(\beta) = f^*(-\beta) = \sup_{u \in \mathbb{R}}(-\beta u - f(u)) = \begin{cases} 0 & \text{if } \beta > 2 \\ 2 - \beta & \text{if } 0 \leq \beta \leq 2 \\ +\infty & \text{if } \beta < 0. \end{cases}$$

By inspection, we see that $u^* = 1$. If we set $g(u) = u$, then we recover the hinge loss $\phi(\alpha) = (1 - \alpha)_+$. On the other hand, choosing $g(u) = e^{u-1}$ leads to the following loss:

$$\phi(\alpha) \quad = \quad \begin{cases} (2 - e^{\alpha})_+ & \text{for } \alpha \leq 0 \\ e^{-\alpha} & \text{for } \alpha > 0. \end{cases} \tag{4.29}$$

Note that this choice of g does not lead to a convex loss ϕ, although this non-convex loss still induces f in the sense of Proposition 4.4. To ensure that ϕ is convex, g is any increasing convex function in $[1, +\infty)$ such that $g(u) = u$ for $u \in [1, 2]$. See Figure 4.4 for illustrations.

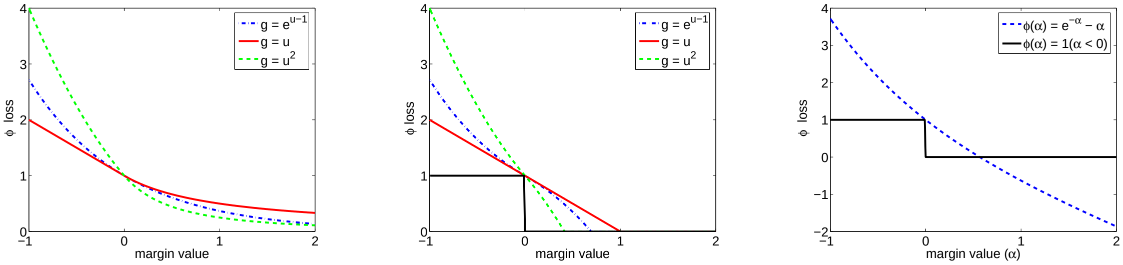


Figure 4.3. Panels (a) and (b) show examples of ϕ losses that induce the Hellinger distance and variational distance, respectively, based on different choices of the function g. Panel (c) shows a loss function that induces the symmetric KL divergence; for the purposes of comparison, the 0-1 loss is also plotted.

Example 3 (Capacitory discrimination distance). The capacitory discrimination distance is equivalent to an f-divergence with $f(u) = -u \log \frac{u+1}{u} - \log(u+1)$, for $u \geq 0$. Augmenting this function with $f(u) = +\infty$ for $u < 0$, we have

$$\Psi(\beta) = \sup_{u \in \mathbb{R}} -\beta u - f(u) = \begin{cases} \beta - \log(e^{\beta} - 1) & \text{for } \beta \geq 0 \\ +\infty & \text{otherwise.} \end{cases}$$

This representation shows that $u^* = \log 2$. If we choose $g(u) = \log(1 + \frac{e^u}{2})$, then we obtain the logistic loss $\phi(\alpha) = \log(1 + e^{-\alpha})$.

Example 4 (Triangular discrimination distance). Triangular discriminatory distance is equivalent to the negative of the harmonic distance; it is an f-divergence with $f(u) = -\frac{4u}{u+1}$ for $u \geq 0$. Let us augment f with $f(u) = +\infty$ for $u < 0$. Then we can write

$$\Psi(\beta) = \sup_{u \in \mathbb{R}} (-\beta u - f(u)) = \begin{cases} (2 - \sqrt{\beta})^2 & \text{for } \beta \geq 0 \\ +\infty & \text{otherwise.} \end{cases}$$

Clearly $u^* = 1$. In this case, setting $g(u) = u^2$ gives the least square loss $\phi(\alpha) = (1 - \alpha)^2$.

Example 5 (Another Kullback-Leibler based divergence). Recall that both the KL divergences (i.e., $KL(\mu||\pi)$ and $KL(\pi||\pi)$) are asymmetric; therefore, Corollary 4.9(b) implies that they are *not* realizable by any margin-based surrogate loss. However, a closely related functional is the *symmetric Kullback-Leibler* divergence [Bradt and Karlin, 1956]:

$$KL_s(\mu, \pi) := KL(\mu||\pi) + KL(\pi||\mu). \tag{4.30}$$

It can be verified that this symmetrized KL divergence is an f-divergence, generated by the function $f(u) = -\log u + u \log u$ for $u \geq 0$, and $+\infty$ otherwise. Therefore, Corollary 4.9(a) implies that it can be generated by some surrogate loss function, but the form of this loss function is not at all obvious. Therefore, in order to recover an explicit form for some ϕ, we follow the constructive procedure of Theorem 4.8, first defining

$$\Psi(\beta) \;=\; \sup_{u \geq 0} \big\{ -\beta u + \log u - u \log u \big\}.$$

In order to compute the value of this supremum, we take the derivative with respect to u and set it to zero; doing so yields the zero-gradient condition $-\beta + 1/u - \log u - 1 = 0$. To capture this condition, we define a function $r : [0, +\infty) \to [-\infty, +\infty]$ via $r(u) = 1/u - \log u$. It is easy to see that $r(u)$ is a strictly decreasing function whose range covers the whole real line; moreover, the zero-gradient condition is equivalent to $r(u) = \beta + 1$. We can thus write $\Psi(\beta) = u + \log u - 1$ where $u = r^{-1}(\beta + 1)$, or equivalently

$$\Psi(\beta) \;=\; r(1/u) - 1 \;=\; r\left(\frac{1}{r^{-1}(\beta+1)}\right) - 1.$$

It is straightforward to verify that the function Ψ thus specified is strictly decreasing and convex with $\Psi(0) = 0$, and that $\Psi(\Psi(\beta)) = \beta$ for any $\beta \in \mathbb{R}$. Therefore, Proposition 4.7 and Theorem 4.8 allow us to specify the form of any convex surrogate loss function that generate the symmetric KL divergence; in particular, any such functions must be of the form (4.22):

$$\phi(\alpha) = \begin{cases} g(-\alpha) & \text{for } \alpha \leq 0 \\ \Psi(g(\alpha)) & \text{otherwise,} \end{cases}$$

where $g : [0, +\infty) \to [0, +\infty)$ is some increasing convex function satisfying $g(0) = 0$. As a particular example (and one that leads to a closed form expression for ϕ), let us choose $g(u) = e^u + u - 1$. Doing so leads to the surrogate loss function

$$\phi(\alpha) = e^{-\alpha} - \alpha - 1.$$

It can be verified by some calculations that the optimized ϕ-risk is indeed the symmetrized KL divergence. See Figure 4.4(c) for an illustration of this loss function.

4.5 On comparison of surrogate loss functions and quantizer designs

The previous section was devoted to study of the correspondence between f-divergences and the optimal ϕ-risk $R_\phi(Q)$ for a fixed experiment Q. Our ultimate goal, however, is that of choosing an optimal Q, which can be viewed as a problem of experimental design [Blackwell, 1953]. Accordingly, the remainder of this chapter is devoted to the joint optimization of ϕ-risk (or more precisely, its empirical version) over both the discriminant function γ as well as the choice of experiment Q (hereafter referred to as a quantizer). In particular, we address the fundamental question associated with such an estimation procedure: for what loss functions ϕ does such joint optimization lead to minimum Bayes risk? Note that this question is not covered by standard consistency results [Jiang, 2004; Lugosi and Vayatis, 2004; Zhang, 2004; Steinwart, 2005; Bartlett *et al.*, 2006; Mannor *et al.*, 2003] on classifiers obtained from surrogate loss functions, because the optimization procedure involves both the discriminant function γ and the choice of quantizer Q.

4.5.1 Inequalities relating surrogate losses and f-divergences

The correspondence between surrogate loss functions and f-divergence allows one to compare surrogate ϕ-risks by comparing the corresponding f-divergences, and vice versa. For instance, since the optimal ϕ-risk for hinge loss is equivalent to the optimal ϕ-risk for 0-1 loss, we can say affirmatively that minimizing risk for hinge loss is equivalent to minimizing the Bayes risk. Moreover, it is well-known that the f-divergences are connected via various inequalities, some of which are summarized in the following lemma, proved in Appendix 4.C:

Lemma 4.11. *The following inequalities among f-divergences hold:*

(a) $V^2 \leq \Delta \leq V$.

(b) $2h^2 \leq \Delta \leq 4h^2$. *As a result,* $\frac{1}{2}V^2 \leq 2h^2 \leq V$.

(c) $\frac{1}{2}\Delta \leq C \leq \log 2 \cdot \Delta$. *As a result,* $\frac{1}{2}V^2 \leq C \leq (\log 2)\, V$.

Using this lemma and our correspondence theorem, it is straightforward to derive the following connection between different risks.

Lemma 4.12. *The following inequalities among optimized ϕ-risks hold:*

(a) $R_{hinge}(Q) = 2R_{bayes}(Q)$.

(b) $2R_{bayes}(Q) \leq R_{sqr}(Q) \leq 1 - (1 - 2R_{bayes}(Q))^2$.

(c) $2 \cdot \log 2R_{bayes}(Q) \leq R_{log}(Q) \leq \log 2 - \frac{1}{2}(1 - 2R_{bayes}(Q))^2$.

(d) $2R_{bayes}(Q) \leq R_{exp}(Q) \leq 1 - \frac{1}{2}(1 - 2R_{bayes}(Q))^2$.

Note that Lemma 4.12 shows that all the ϕ-risks considered (i.e., hinge, square, logistic, and exponential) are bounded below by the variational distance (up to some constant multiplicative term). However, with the exception of hinge loss, these results do *not* tell us whether minimizing ϕ-risk leads to a classifier-quantizer pair (γ, Q) with minimal Bayes risk. We explore this issue in more detail in the sequel: more precisely, we specify all surrogate losses ϕ such that minimizing the associated ϕ-risk leads to the same optimal decision rule (Q, γ) as minimizing the Bayes risk.

4.5.2 Connection between 0-1 loss and f-divergences

The connection between f-divergences and 0-1 loss can be traced back to seminal work on comparison of experiments, pioneered by Blackwell and others [Blackwell, 1951; Blackwell, 1953; Bradt and Karlin, 1956].

Definition 4.13. *The quantizer Q_1 dominates Q_2 if $R_{Bayes}(Q_1) \leq R_{Bayes}(Q_2)$ for any choice of prior probabilities $q = \mathbb{P}(Y = -1) \in (0, 1)$.*

Recall that a choice of quantizer design Q induces two conditional distributions $P(Z|Y = 1) \sim P_1$ and $P(Z|Y = -1) \sim P_{-1}$. Hence, we shall use P_{-1}^Q and P_1^Q to denote the fact that both P_{-1} and P_1 are determined by the specific choice of Q. By "parameterizing" the decision-theoretic criterion in terms of loss function ϕ and establishing a precise correspondence between ϕ and the f-divergence, we can derive the following theorem that relates 0-1 loss and f-divergences:

Theorem 4.14. [Blackwell, 1951; Blackwell, 1953] *For any two quantizer designs Q_1 and Q_2, the following statement are equivalent:*

(a) Q_1 dominates Q_2 (i.e., $R_{bayes}(Q_1) \leq R_{bayes}(Q_2)$ for any prior probabilities $q \in (0, 1)$).

(b) $I_f(P_1^{Q_1}, P_{-1}^{Q_1}) \geq I_f(P_1^{Q_2}, P_{-1}^{Q_2})$, for all functions f of the form $f(u) = -\min(u, c)$ for some $c > 0$.

(c) $I_f(P_1^{Q_1}, P_{-1}^{Q_1}) \geq I_f(P_1^{Q_2}, P_{-1}^{Q_2})$, for all convex functions f.

We include a short proof of this result in Appendix 4.D, using the tools developed in this chapter. In conjunction with our correspondence between f-divergences and ϕ-risks, this theorem implies the following

Corollary 4.15. *The quantizer Q_1 dominates Q_2 if and only if $R_\phi(Q_1) \leq R_\phi(Q_2)$ for any loss function ϕ.*

Proof. By Proposition 4.4, we have $R_\phi(Q) = -I_f(\mu, \pi) = -I_{f_q}(P_1, P_{-1})$, from which the corollary follows using Theorem 4.14. □

Corollary 4.15 implies that if

$$R_\phi(Q_1) \leq R_\phi(Q_2)$$

for some loss function ϕ, then

$$R_{bayes}(Q_1) \leq R_{bayes}(Q_2)$$

for some set of prior probabilities on the hypothesis space. This implication justifies the use of a given surrogate loss function ϕ in place of the 0-1 loss for *some* prior probability; however, for a given prior probability, it gives no guidance on how to choose ϕ. Moreover, in many applications (e.g., decentralized detections), it is usually the case that the prior probabilities on the hypotheses are fixed, and the goal is to determine optimum quantizer design Q for this fixed set of priors. In such a setting, the Blackwell's notion of Q_1 dominating Q_2 has limited usefulness. With this motivation in mind, the following section is devoted to development of a more stringent method for assessing equivalence between loss functions.

4.5.3 Universal equivalence

In the following definition, the loss functions ϕ_1 and ϕ_2 realize the f-divergences associated with the convex function f_1 and f_2, respectively.

Definition 4.16. *The surrogate loss functions ϕ_1 and ϕ_2 are* universally equivalent, *denoted by $\phi_1 \overset{u}{\approx} \phi_2$, if for any $\mathbb{P}(X, Y)$ and quantization rules Q_1, Q_2, there holds:*

$$R_{\phi_1}(Q_1) \leq R_{\phi_1}(Q_2) \Leftrightarrow R_{\phi_2}(Q_1) \leq R_{\phi_2}(Q_2).$$

In terms of the corresponding f-divergences, this relation is denoted by $f_1 \overset{u}{\approx} f_2$.

Observe that this definition is very stringent, in that it requires that the ordering between optimized ϕ_1 and ϕ_2 risks holds for all probability distributions $\mathbb{P}$ on $\mathcal{X} \times \mathcal{Y}$. However, this notion of equivalence is needed for nonparametric approaches to classification, in which the underlying distribution $\mathbb{P}$ is not available in parametric form.

The following result provides necessary and sufficient conditions for two f-divergences to be universally equivalent:

Theorem 4.17. *Let f_1 and f_2 be convex functions on $[0, +\infty) \to \mathbb{R}$ and differentiable almost everywhere. Then $f_1 \overset{u}{\approx} f_2$ if and only if $f_1(u) = cf_2(u) + au + b$ for some constants $c > 0$ and a, b.*

Proof. One direction of the theorem ("if") is easy. We focus on the other direction. The proof relies on the following technical result (see Appendix 4.E for a proof):

Lemma 4.18. *Given a continuous convex function $f : \mathbb{R}^+ \to \mathbb{R}$, define, for any $u, v \in \mathbb{R}^+$, define:*

$$T_f(u, v) := \left\{ \frac{u\alpha - v\beta - f(u) + f(v)}{\alpha - \beta} = \frac{f^*(\alpha) - f^*(\beta)}{\alpha - \beta} \middle| \alpha \in \partial f(u), \beta \in \partial f(v), \alpha \neq \beta \right\}.$$

If $f_1 \overset{u}{\approx} f_2$, then for any $u, v > 0$, one of the following must be true:

1. *$T_f(u, v)$ are non-empty for both f_1 and f_2, and $T_{f_1}(u, v) = T_{f_2}(u, v)$.*
2. *Both f_1 and f_2 are linear in (u, v).*

Note that if function f is differentiable at u and v and $f'(u) \neq f'(v)$, then $T_f(u, v)$ is reduced to a number:

$$\frac{uf'(u) - vf'(v) - f(u) + f(v)}{f'(u) - f'(v)} = \frac{f^*(\alpha) - f^*(\beta)}{\alpha - \beta},$$

where $\alpha = f'(u)$, $\beta = f'(v)$, and f^* denotes the conjugate dual of f.
Let v is a point where both f_1 and f_2 are differentiable. Let $d_1 = f_1'(v)$, $d_2 = f_2'(v)$. Without loss of generality, assume $f_1(v) = f_2(v) = 0$ (if not, we can consider functions $f_1(u) - f_1(v)$ and $f_2(u) - f_2(v)$).

Now, for any u where both f_1 and f_2 are differentiable, applying Lemma 4.18 for v and u, then either f_1 and f_2 are both linear in $[v, u]$ (or $[u, v]$ if $u < v$), in which case $f_1(u) = cf_2(u)$ for some constant c, or the following is true:

$$\frac{uf_1'(u) - f_1(u) - vd_1}{f_1'(u) - d_1} = \frac{uf_2'(u) - f_2(u) - vd_2}{f_2'(u) - d_2}.$$

In either case, we have

$$(uf_1'(u) - f_1(u) - vd_1)(f_2'(u) - d_2) = (uf_2'(u) - f_2(u) - vd_2)(f_1'(u) - d_1).$$

Let $f_1(u) = g_1(u) + d_1 u$, $f_2(u) = g_2(u) + d_2 u$. Then, $(ug_1'(u) - g_1(u) - vd_1)g_2'(u) = (ug_2'(u) - g_2(u) - vd_2)g_1'(u)$, implying that $(g_1(u) + vd_1)g_2'(u) = (g_2(u) + vd_2)g_1'(u)$ for any u where f_1 and f_2 are both differentiable. It follows that $g_1(u) + vd_1 = c(g_2(u) + vd_2)$ for some constant c and this constant c has to be the same for any u due to the continuity of f_1 and f_2. Hence, we have $f_1(u) = g_1(u) + d_1 u = cg_2(u) + d_1 u + cvd_2 - vd_1 = cf_2(u) + (d_1 - cd_2)u + cvd_2 - vd_1$. It is now simple to check that $c > 0$ is necessary and sufficient for I_{f_1} and I_{f_2} to have the same monotonicity. □

An important special case is when one of the f-divergences is the variational distance. In this case, we have the following

Proposition 4.19. *(a) All f-divergences based on continuous convex $f : [0, +\infty) \to \infty$ that are universally equivalent to the variational distance have the form*

$$f(u) = -c \min(u, 1) + au + b \qquad \textit{for some } c > 0. \tag{4.31}$$

(b) The 0-1 loss is universally equivalent only to those loss functions whose corresponding f-divergence is based on a function of the form (4.31).

Proof. Note that statement (b) follows immediately from statement (a). The proof in Theorem 4.17 does not exactly apply here, because it requires both f_1 and f_2 to be differentiable almost everywhere. We provide a modified argument in Appendix 4.F. □

Theorem 4.17 shows that each class of equivalent f-divergences are restricted by a strong linear relationship. It is important to note, however, that this restrictiveness does *not* translate over to the classes of universally equivalent loss functions (by Theorem 4.8).

4.5.4 Convex loss functions equivalent to 0-1 loss

This section is devoted to a more in-depth investigation of the class of surrogate loss functions ϕ that are universally equivalent to the 0-1 loss.

4.5.4.1 Explicit construction

We begin by presenting several examples of surrogate loss functions equivalent to 0-1 loss. From Proposition 4.19, any such loss must realize an f-divergence based on a function of the form (4.31). For simplicity, we let $a = b = 0$; these constants do not have any significant effect on the corresponding loss functions ϕ (only simple shifting and translation operations). Hence, we will be concerned only with loss functions whose corresponding f has the form $f(u) = -c\min(u, 1)$ for $u \geq 0$. Suppose that we augment the definition by setting $f(u) = +\infty$ for $u < 0$; with this modification, f remains a lower semicontinuous convex function. In Section 4.4, we considered this particular extension, and constructed all loss functions that were equivalent to the 0-1 loss (in particular, see equation (4.29)). As a special case, this class of loss functions includes the hinge loss function.

Choosing an alternative extension of f for $u < 0$ leads to a different set of loss functions, also equivalent to 0-1 loss. For example, if we set $f(u) = -k \min(u, 1)$ for $u < 0$ where $k \geq c$, then the resulting Ψ takes the form

$$\Psi(\beta) = \begin{cases} (c-\beta)_+ & \text{for } 0 \leq \beta \leq k \\ +\infty & \text{otherwise.} \end{cases}$$

In this case, the associated loss functions ϕ has the form:

$$\phi(\alpha) = \begin{cases} g(c/2 - \alpha) & \text{for } \alpha \leq 0 \\ (c - g(c/2 + \alpha))_+ & \text{when } g(c/2+\alpha) \leq k \\ +\infty & \text{otherwise,} \end{cases} \tag{4.32}$$

where g is a increasing convex function such that $g(c/2) = c/2$. However, to ensure that ϕ is a convex function, it is simple to see that g has to be linear in the interval $[c/2, u]$ for some u such that $g(u) = k$.

4.5.4.2 A negative result

Thus, varying the extension of f for $u < 0$ (and subsequently the choice of g) leads to a large class of possible loss functions equivalent to the 0-1 loss. What are desirable properties of a surrogate loss function? Properties can be desirable either for computational reasons (e.g., convexity, differentiablity), or for statistical reasons (e.g., consistency). Unfortunately, in this regard, the main result of this section is a negative one: in particular, we prove that there is no differentiable surrogate loss that is universally equivalent to the 0-1 loss.

Proposition 4.20. *There does not exist a continuous and differentiable convex loss function ϕ that is universally equivalent to the 0-1 loss.*

Proof. From Proposition 4.19, any ϕ that is universally equivalent to 0-1 loss must generate an f-divergence of the form (4.31). Let $a = b = 0$ without loss of generality; the proof proceeds in the same way for the general case. First, we claim that regardless of how f is augmented for $u < 0$, the function Ψ always has the following form:

$$\Psi(\beta) = f^*(-\beta) \;=\; \sup_{u \in \mathbb{R}} \big\{ -\beta u - f(u) \big\} \;=\; \begin{cases} +\infty & \text{for } \beta < 0 \\ c - \beta & \text{for } 0 \leq \beta \leq c \\ \geq 0 & \text{otherwise.} \end{cases} \tag{4.33}$$

Indeed, for $\beta < 0$, we have

$$\Psi(\beta) \ \geq \sup_{u \geq 0} \big\{ -\beta u + c\min(u,1) \big\} \ = \ +\infty.$$

Turning to the case $\beta \in [0, c]$, we begin by observing that we must have $f(u) \geq -cu$ for $u \leq 0$ (since f is a convex function). Therefore,

$$\sup_{u<0} \big\{ -\beta u - f(u) \big\} \ \leq \ \sup_{u<0} \big\{ -\beta u + cu \big\} \ = \ 0.$$

On the other hand, we have $\sup_{u\geq 0} \big\{ -\beta u + c\min(u,1) \big\} = c - \beta \geq 0$, so that we conclude that $\Psi(\beta) = c - \beta$ for $\beta \in [0, c]$. Finally, for $\beta \geq c$, we have $\Psi(\beta) \geq \sup_{u\geq 0} \big\{ -\beta u + c\min(u,1) \big\} = 0$.

Given the form (4.33), Theorem 4.7 implies that the loss function ϕ must have the following form:

$$\phi(\alpha) = \begin{cases} g(c/2 - \alpha) & \text{when } \alpha \leq 0 \\ (c - g(c/2 + \alpha))_+ & \text{when } \alpha > 0 \text{ and } g(c/2+\alpha) \leq c, \\ \geq 0 & \text{otherwise,} \end{cases} \tag{4.34}$$

where g is an increasing continuous convex function from $[c/2, +\infty)$ to $\mathbb{R}$ satisfying $g(c/2) = c/2$.

For ϕ to be differentiable, the function g has to be differentiable everywhere in its domain. Let $a > 0$ be the value such that $c = g(c/2 + a)$. Since ϕ achieves its minimum at a, $\phi'(a) = 0$. This implies that g has to satisfy $g'(c/2 + a) = 0$. That would imply that g attains its minimum at $c/2 + a$, but $g(c/2 + a) = c > g(c/2)$, which leads to a contradiction. □

4.6 Empirical risk minimization with surrogate convex loss functions

As discussed in Section 4.1, surrogate loss functions are widely used in statistical learning theory, where the goal is to learn a discriminant function given only indirect access to the distribution $\mathbb{P}(X, Y)$ via empirical samples. In this section, we demonstrate the utility of our correspondence between f-divergences and surrogate loss functions in the setting of the elaborated version of the classical discriminant problem, in which the goal is to choose both a discriminant function γ as well as a quantizer Q. As described in the previous chapter, our strategy is the natural one given empirical data: in particular, we choose (Q, γ) by minimizing the empirical version of the ϕ-risk. It is worthwhile noting that without direct access to the distribution $\mathbb{P}(X, Y)$, it is impossible to compute or manipulate the associated f-divergences. In particular, without closed form knowledge of $\mu(z)$ and $\pi(z)$, it is impossible to obtain closed-form solution for the optimal discriminant γ, as required to compute the f-divergence (see Proposition 4.4). Nonetheless, the correspondence to f-divergences turns out to be useful, in that it allows us to establish Bayes consistency of the procedure based on ϕ-risks for choosing the quantizer and discriminant function.

4.6.1 Decentralized detection problem

We begin by recalling the set-up and notations for the decentralized detection problem; see the previous chapter for further details. Let S be an integer, representing some number of sensors that collect observations from the environment. More precisely, for each $t = 1, \ldots, S$, let $X^t \in \mathcal{X}^t$ represent the observation at sensor t, where $\mathcal{X}^t$ denotes the observation space. The covariate vector $X = (X^t,\ t = 1, \ldots, S)$ is obtained by concatenating all of these observations together. We assume that the global estimate $\widehat{Y}$ is to be formed by a *fusion center*. In the *centralized setting*, this fusion center is permitted access to the full vector X of observations. In this case, it is well-known [van Trees, 1990] that optimal decision rules, whether under Bayes error or Neyman-Pearson criteria, can be formulated in terms of the likelihood ratio $\mathbb{P}(X|Y = 1)/\mathbb{P}(X|Y = -1)$. In contrast, the defining feature of the *decentralized setting* is that the fusion center has access only to some form of summary of each observation X^t. More specifically, we suppose that each sensor $t = 1 \ldots, S$ is permitted to transmit a *message* Z^t, taking values in some space $\mathcal{Z}^t$. The fusion center, in turn, applies some decision rule γ to compute an estimate $\widehat{Y} = \gamma(Z^1, \ldots, Z^S)$ of Y based on its received messages.

For simplicity, let us assume that the input space $\mathcal{X}^t$ is identical for each $t = 1, \ldots, S$, and similarly, that the quantized space $\mathcal{Z}^t$ is the same for all t. The original observation space $\mathcal{X}^t$ can be either finite (e.g, having M possible values), or continuous (e.g., Gaussian measurements). The key constraint, giving rise to the decentralized nature of the problem, is that the corresponding message space $\mathcal{Z} = \{1, \ldots, L\}^S$ is discrete with finite number of values, and hence "smaller" than the observation space (i.e., $L \ll M$ in the case of discrete $\mathcal{X}$). The problem is to find, for each sensor $t = 1, \ldots, S$, a decision rule represented as a measurable function $Q^t : \mathcal{X}^t \to \mathcal{Z}^t$, as well as an overall decision rule represented by a measurable function $\gamma : \mathcal{Z} \to \{-1, +1\}$ at the fusion center so as to minimize the *Bayes risk* $\mathbb{P}(Y \neq \gamma(Z))$.

Figure 4.4 provides a graphical representation of this decentralized detection problem. The single node at the top of the figure represents the hypothesis variable Y, and the outgoing arrows point to the collection of observations $X = (X^1, \ldots, X^S)$. The local decision rules Q^t lie on the edges between sensor observations X^t and messages Z^t. Finally, the node at the bottom is the fusion center, which collects all the messages.

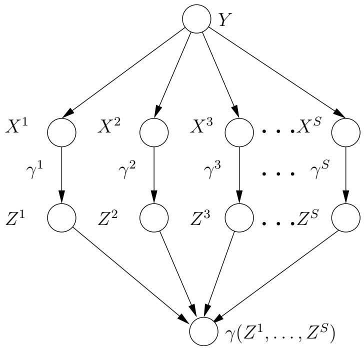


Figure 4.4. Decentralized detection system with S sensors, in which Y is the unknown hypothesis, $X = (X^1, \dots, X^S)$ is the vector of sensor observations; and $Z = (Z^1, \dots, Z^S)$ are the quantized messages transmitted from sensors to the fusion center.

Recall that the quantizer Q can be conveniently viewed as conditional probability distribution $Q(z|x)$, which implies that an aggregate observation x is mapped to an aggregate quantized message z with probability $Q(z|x)$. In particular, the decentralization constraints require that the conditional probability distributions $Q(z|x)$ factorize; i.e., for any realization z of Z, $Q(z|X) = \prod_{t=1}^{S} Q^t(z^t|X^t)$ with probability one. For the remainder of this section, however, we shall use $Q_z(x)$ to denote $Q(z|x)$, to highlight the formal view that the quantizer rule Q is a collection of measurable functions $Q_z : \mathcal{X} \to \mathbb{R}$ for $z \in \mathcal{Z}$.

In summary, our decentralized detection problem is a particular case of the elaborated discriminant problem—namely, a hypothesis testing problem with an additional component of experiment design, corresponding to the choice of the quantizer Q.

A learning algorithm for decentralized detection. In Chapter 3 we introduced an algorithm for designing a decentralized detection system (i.e., both the quantizer and the classifier at the fusion center) based on surrogate loss functions. The algorithm operates on an i.i.d. set of data samples, and makes no assumptions about the underlying probability distribution $\mathbb{P}(X, Y)$. Such an approach is fundamentally different from the bulk of previous work on decentralized decentralization, which typically are based on restrictive parametric assumptions. This type of nonparametric approach is particularly useful in practical applications of decentralized detection (e.g., wireless sensor networks), where specifying an accurate parametric model for the probability distribution $\mathbb{P}(X, Y)$ may be difficult or infeasible.

Let $(x_i, y_i)_{i=1}^n$ be a set of i.i.d. samples from the (unknown) underlying distribution $\mathbb{P}(X, Y)$ over the covariate X and hypothesis $Y \in \{-1, +1\}$. Let $\mathcal{C}_n \subseteq \Gamma$ and $\mathcal{D}_n \subseteq \mathcal{Q}$ represent subsets of classifiers and quantizers, respectively. The algorithm chooses an optimum decision rule $(\gamma, Q) \in (\mathcal{C}_n, \mathcal{D}_n)$ by minimizing an empirical version of ϕ-risk:

$$\hat{R}_\phi(\gamma, Q) := \frac{1}{n} \sum_{i=1}^{n} \sum_{z} \phi(y_i \gamma(z)) Q_z(x_i). \tag{4.35}$$

It is worth noting that the perspective of surrogate ϕ-loss (as opposed to f-divergence) is the most natural in this nonparametric setting. Given that the minimization takes place over the subset $(\mathcal{C}_n, \mathcal{D}_n)$, there is no closed-form solution for the minimizer $\gamma \in \mathcal{C}_n$ of problem (4.35) (even when the optimum Q is known). Hence, it is not even possible to formulate an equivalent closed-form problem in terms of f-divergences. Despite this fact, we demonstrate that the connection to f-divergences is nonetheless useful, in that it allows to address the consistency of the estimation procedure (4.35). In particular, we prove that for all ϕ that are universally equivalent to the 0-1 loss, this estimation procedure is indeed consistent (for suitable choices of the sequences of function classes $\mathcal{C}_n$ and $\mathcal{D}_n$). The analysis is inspired by frameworks recently developed by a number of authors (see, e.g., [Zhang, 2004; Steinwart, 2005; Bartlett *et al.*, 2006]) for the standard case of classification (i.e., without any component of experiment design) in statistical machine learning.

4.6.2 A consistency theorem

For each $z \in \mathcal{Z}$, let us endow the space of functions $Q_z : \mathcal{X} \to \mathbb{R}$ with an appropriate topology, specifically that defined in the proof of Proposition 2.1 in [Tsitsiklis, 1993a], and endow the space of $\mathcal{Q}$ with the product topology, under which it is shown to be compact [Tsitsiklis, 1993a]. In addition, the space of measurable functions $\gamma : \mathcal{Z} \to \{-1, 1\}$ is endowed with the uniform-norm topology.

Consider sequences of increasing compact function classes $\mathcal{C}_1 \subseteq \mathcal{C}_2 \subseteq \ldots \subseteq \Gamma$ and $\mathcal{D}_1 \subseteq \mathcal{D}_2 \subseteq \ldots \subseteq \mathcal{Q}$. This analysis supposes that there exists oracle that outputs an optimal solution to the minimization problem

$$\min_{(\gamma, Q) \in (\mathcal{C}_n, \mathcal{D}_n)} \hat{R}_\phi(\gamma, Q), \tag{4.36}$$

and let (γ_n^*, Q_n^*) denote one such solution. Let R_{bayes}^* denote the minimum Bayes risk achieved over the space of decision rules $(\gamma, Q) \in (\Gamma, \mathcal{Q})$. We refer to the non-negative quantity $R_{bayes}(\gamma_n^*, Q_n^*) - R_{bayes}^*$ the *excess Bayes risk* of our estimation procedure. We say that such an estimation procedure is *universally consistent* if the excess Bayes risk converges to zero (in probability) as $n \to \infty$. More precisely, we require that for any (unknown) Borel probability measure $\mathbb{P}(X, Y)$

$$\lim_{n \to \infty} R_{bayes}(\gamma_n^*, Q_n^*) = R_{bayes}^*. \tag{4.37}$$

In order to analyze statistical behavior of this algorithm and to establish universal consistency for appropriate sequences $(\mathcal{C}_n, \mathcal{D}_n)$ of function classes, we follow a standard strategy of decomposing the Bayes error in terms of two types of errors:

- the *approximation error* introduced by the bias of the function classes $\mathcal{C}_n \subseteq \Gamma$, and $\mathcal{D}_n \subseteq \mathcal{Q}$, and
- the *estimation error* introduced by the variance of using finite sample size n.

These quantities are defined as follows:

Definition 4.21. *The approximation error of the procedure is given by*

$$\mathcal{E}_0(\mathcal{C}_n, \mathcal{D}_n) = \inf_{(\gamma,Q)\in(\mathcal{C}_n,\mathcal{D}_n)} \{R_\phi(\gamma, Q)\} - R_\phi^*, \tag{4.38}$$

where $R_\phi^* := \inf_{(\gamma,Q)\in(\Gamma,\mathcal{Q})} R_\phi(\gamma, Q)$.

Definition 4.22. *The estimation error is given by*

$$\mathcal{E}_1(\mathcal{C}_n, \mathcal{D}_n) = \mathbb{E} \sup_{(\gamma,Q)\in(\mathcal{C}_n,\mathcal{D}_n)} \left| \hat{R}_\phi(\gamma, Q) - R_\phi(\gamma, Q) \right|, \tag{4.39}$$

where the expectation is taken with respect to the (unknown) measure $\mathbb{P}(X, Y)$.

Conditions on loss function ϕ. Our consistency result applies to the class of surrogate losses that are universally equivalent to the 0-1 loss. From Proposition 4.19, all such loss functions ϕ correspond to an f-divergence of the form

$$f(u) = -c \min(u, 1) + au + b, \tag{4.40}$$

for some constants $c > 0, a, b$. For any such ϕ, a straightforward calculation (see the proof of Proposition 4.4) shows that the optimum risk (for fixed quantizer Q) takes the form

$$R_\phi(Q) = -I_f(\mu, \pi) = c \sum_{z\in\mathcal{Z}} \min\{\mu(z), \pi(z)\} - ap - bq, \tag{4.41}$$

where $p = \mathbb{P}(Y = 1)$ and $q = \mathbb{P}(Y = -1) = 1 - p$.

Recall that any surrogate loss ϕ is assumed to be continuous, convex, and classification-calibrated (see Definition 4.1). For our proof, we require the additional technical conditions, expressed in terms of ϕ as well as its induced f-divergence (4.40):

$$(a - b)(p - q) \geq 0 \qquad \text{and} \qquad \phi(0) \geq 0. \tag{4.42}$$

Intuitively, these technical conditions are needed so that the approximation error due to varying Q dominates the approximation error due to varying γ (because the optimum γ is determined only after Q is). Simply letting, say, $a = b$ would suffice.

Any surrogate loss that satisfies all of these conditions (continuous, convex, classification-calibrated, universally equivalent to 0-1 loss, and condition (4.42)) is said to satisfy *property $\mathcal{P}$*. Throughout this section, we shall assume that the loss function ϕ has property $\mathcal{P}$. In addition, for each $n = 1, 2, \ldots$, we assume that

$$M_n \;:=\; \max_{y \in \{-1,+1\}} \; \sup_{(\gamma,Q)\in(\mathcal{C}_n,\mathcal{D}_n)} \; \sup_{z\in\mathcal{Z}} |\phi(y\gamma(z))| \;<\; +\infty. \tag{4.43}$$

The following theorem ties together the Bayes error with the approximation error and estimation error, and provides sufficient conditions for universal consistency:

Theorem 4.23. *Let $\mathcal{C}_1 \subseteq \mathcal{C}_2 \subseteq \ldots \subseteq \Gamma$ and $\mathcal{D}_1 \subseteq \mathcal{D}_2 \subseteq \ldots \subseteq \mathcal{Q}$ be nested sequences of compact function classes, and consider the estimation procedure* (4.36) *using a surrogate loss ϕ that satisfies property $\mathcal{P}$.*

(a) For any Borel probability measure $\mathbb{P}(X,Y)$, with probability at least $1-\delta$, there holds:

$$R_{bayes}(\gamma_n^*, Q_n^*) - R_{bayes}^* \;\;\leq\;\; \frac{2}{c}\left\{2\mathcal{E}_1(\mathcal{C}_n,\mathcal{D}_n) + \mathcal{E}_0(\mathcal{C}_n,\mathcal{D}_n) + 2M_n\sqrt{2\frac{\ln(2/\delta)}{n}}\right\}.$$

(b) Universal Consistency: *Suppose that the function classes satisfy the following properties:*

Approximation condition: $\lim_{n\to\infty} \mathcal{E}_0(\mathcal{C}_n,\mathcal{D}_n) = 0.$

Estimation condition: $\lim_{n\to\infty} \mathcal{E}_1(\mathcal{C}_n,\mathcal{D}_n) = 0$ *and* $\lim_{n\to\infty} M_n\sqrt{\ln n/n} = 0.$

Then the estimation procedure (4.36) *is universally consistent:*

$$\lim_{n\to\infty} R_{bayes}(\gamma_n^*, Q_n^*) = R_{bayes}^* \qquad \textit{in probability}. \tag{4.44}$$

The proof of this theorem relies on an auxiliary result that is of independent interest. In particular, we prove that for any function classes $\mathcal{C}$ and $\mathcal{D}$, and surrogate loss satisfying property $\mathcal{P}$, the excess ϕ-risk is related to the excess Bayes risk as follows:

Proposition 4.24. *Let ϕ be a loss function that has property $\mathcal{P}$. Then any classifier-quantizer pair $(\gamma, Q) \in (\mathcal{C}, \mathcal{D})$, we have*

$$\frac{c}{2}\left[R_{bayes}(\gamma,Q) - R_{bayes}^*\right] \;\;\leq\;\; R_\phi(\gamma,Q) - R_\phi^*. \tag{4.45}$$

See Appendix 4.G for a proof of this result. A consequence of equation (4.45) is that in order to achieve Bayes consistency (i.e., driving the excess Bayes risk to zero), it suffices to drive the excess ϕ-risk to zero.

With Proposition 4.24, we are now equipped to prove Theorem 4.23:

Proof. (a) First observe that the value of $\sup_{\gamma \in \mathcal{C}_n, Q \in \mathcal{D}_n} |\hat{R}_\phi(\gamma, Q) - R_\phi(\gamma, Q)|$ varies by at most $2M_n/n$ if one changes the values of (x_i, y_i) for some index $i \in \{1, \ldots, n\}$. Hence, applying McDiarmid's inequality yields concentration around the expected value, or (alternatively stated) that with probability at least $1 - \delta$,

$$\left| \sup_{\gamma \in \mathcal{C}_n, Q \in \mathcal{D}_n} |\hat{R}_\phi(\gamma, Q) - R_\phi(\gamma, Q)| - \mathcal{E}_1(\mathcal{C}_n, \mathcal{D}_n) \right| \leq M_n \sqrt{2 \ln(1/\delta)/n}. \tag{4.46}$$

Suppose that $R_\phi(\gamma, Q)$ attains its minimum over the compact subset $(\mathcal{C}_n, \mathcal{D}_n)$ at $(\gamma_n^\dagger, Q_n^\dagger)$. Then, using Proposition 4.24, we have

$$\begin{aligned} \frac{c}{2}(R_{bayes}(\gamma_n^*, Q_n^*) - R_{bayes}^*) &\leq R_\phi(\gamma_n^*, Q_n^*) - R_\phi^* \\ &= R_\phi(\gamma_n^*, Q_n^*) - R_\phi(\gamma_n^\dagger, Q_n^\dagger) + R_\phi(\gamma_n^\dagger, Q_n^\dagger) - R_\phi^*) \\ &= R_\phi(\gamma_n^*, Q_n^*) - R_\phi(\gamma_n^\dagger, Q_n^\dagger) + \mathcal{E}_0(\mathcal{C}_n, \mathcal{D}_n) \end{aligned}$$

Hence, using equation (4.46), we have with probability at least $1 - \delta$:

$$\begin{aligned} \frac{c}{2}(R_{bayes}(\gamma_n^*, Q_n^*) - R_{bayes}^*) \leq \hat{R}_\phi(\gamma_n^*, Q_n^*) - \hat{R}_\phi(\gamma_n^\dagger, Q_n^\dagger) + 2\mathcal{E}_1(\mathcal{C}_n, \mathcal{D}_n) \\ + 2M_n \sqrt{2 \ln(2/\delta)/n} + \mathcal{E}_0(\mathcal{C}_n, \mathcal{D}_n) \\ \leq 2\mathcal{E}_1(\mathcal{C}_n, \mathcal{D}_n) + \mathcal{E}_0(\mathcal{C}_n, \mathcal{D}_n) + 2M_n \sqrt{2 \ln(2/\delta)/n}, \end{aligned}$$

from which Theorem 4.23(a) follows.

(b) This statement follows by applying (a) with $\delta = 1/n$, and noting that $R_{bayes}(\gamma_n^*, Q_n^*) - R_{bayes}^*$ is bounded. □

A natural question is under what conditions the approximation and estimation conditions of Theorem 4.23 hold. We conclude this section by stating some precise conditions on the function classes that ensure that the approximation condition holds. Let U be a Borel subset of $\mathcal{X}$ such that $\mathbb{P}_X(U) = 1$, and let $C(U)$ denote the Banach space of continuous functions $Q_z(x)$ mapping U to $\mathbb{R}$. If $\cup_{n=1}^\infty \mathcal{D}_n$ is dense in $\mathcal{Q} \cap C(U)$ and if $\cup_{n=1}^\infty \mathcal{C}_n$ is dense in Γ, then the approximation condition in Theorem 4.23 holds. In order to establish this fact, note that $R_\phi(\gamma, Q)$ is a continuous function with respect to (γ, Q) over the compact space

$(\Gamma, \mathcal{Q})$. (Here compactness is defined with respect to the topology defined in the proof of Proposition 2.1 in [Tsitsiklis, 1993a].) The approximation condition then follows by applying Lusin's approximation theorem for regular measures, using an argument similar to the proof of Theorem 4.1 in [Zhang, 2004].

4.6.3 Estimation error for kernel classes

For the estimation condition in Theorem 4.23(b) to hold the sequence of function classes $(\mathcal{C}_n, \mathcal{D}_n)_{n=1}^{\infty}$ has to increase sufficiently slowly in "size" with respect to n. In this section, we analyze the behavior of this estimation error for a certain kernel-based function class. Throughout this section, in addition to the conditions imposed on ϕ in the preceding section, we assume that the loss function ϕ is Lipschitz with constant L_ϕ. We also assume without loss of generality that $\phi(0) = 0$ (otherwise, one could consider the modified loss function $\phi(\alpha) - \phi(0)$).

First of all, we require a technical definition of a particular measure of function class complexity:

Definition 4.25. *Let $\mathcal{F}$ be a class of measurable functions mapping from its domain to $\mathbb{R}$. The* Rademacher complexity *of $\mathcal{F}$ is given by*

$$R_n(\mathcal{F}) = \frac{2}{n}\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i f(X_i)\right|, \tag{4.47}$$

where $\sigma_i,\ i = 1,\ldots n$ are i.i.d. Bernoulli variables (taking values $\{-1,+1\}$ equiprobably), and the expectation is taken over both $\sigma_1,\ldots,\sigma_n$ and $X_1,\ldots,X_n$.

For analyzing the estimation error, the relevant class of functions takes the form

$$\mathcal{G} \ := \ \left\{g : \mathcal{X} \to \mathbb{R} \middle| g(x) = \gamma(\mathrm{argmax}_z Q_z(x)) \text{ for some } \ (\gamma, Q) \in (\mathcal{C}, \mathcal{D} \cap \mathcal{Q}_0)\right\}. \tag{4.48}$$

We now show that the Rademacher complexity of this class can be used to upper bound the estimation error:

Lemma 4.26. *For a Lipschitz ϕ (with constant L_ϕ), the estimation error is upper bounded by the Rademacher complexity of $\mathcal{G}$ as follows:*

$$\mathcal{E}_1(\mathcal{C}, \mathcal{D}) \leq 2L_\phi R_n(\mathcal{G}). \tag{4.49}$$

Proof. Using the standard symmetrization method [van der Vaart and Wellner, 1996], we have:

$$\begin{aligned}\mathcal{E}_1(\mathcal{C},\mathcal{D}) &\leq R_n(\mathcal{H}) \\ &= \frac{2}{n}\mathbb{E}\sup_{(\gamma,Q)\in(\mathcal{C},\mathcal{D})}\left|\sum_{i=1}^{n}\sigma_i\sum_{z\in\mathcal{Z}}\phi(y_i\gamma(z))Q_z(x_i)\right|\end{aligned}$$

where $\mathcal{H}$ is the function class given by

$$\mathcal{H} \;:=\; \Big\{h:\mathcal{X}\times\{\pm1\}\to\mathbb{R}\;\Big|\;h(x,y)=\sum_{z\in\mathcal{Z}}\phi(y\gamma(z))Q_z(x)\text{ for some}(\gamma,Q)\in(\mathcal{C},\mathcal{D})\Big\}$$

Let $\mathcal{H}_0$ be the subset of $\mathcal{H}$ defined by restricting to $Q\in\mathcal{Q}_0$. Since $\mathcal{Q}=\text{co}\mathcal{Q}_0$ (where co denotes the convex hull), it follows that $\mathcal{H}=\text{co}\mathcal{H}_0$, from which it follows from a result in [Bartlett and Mendelson, 2002] that $R_n(\mathcal{H})=R_n(\mathcal{H}_0)$. For $h\in\mathcal{H}_0$, we have $h(x,y)=\phi(y\gamma(\text{argmax}_z Q_z(x)))$. Using results from [Bartlett and Mendelson, 2002] again, we conclude that $R_n(\mathcal{H}_0)\leq 2L_\phi R_n(\mathcal{G})$, □

Using Lemma 4.26, in order for the estimation condition to hold, it is sufficient to choose the function classes so that the Rademacher complexity converges to zero as n tends to infinity. The function classes used in practice often correspond to classes defined by reproducing kernel Hilbert spaces (RKHS). Accordingly, herein we focus our analysis on such a kernel class.

Briefly, a kernel class of functions is defined as follows. Let $K:\mathcal{Z}\times\mathcal{Z}\to\mathbb{R}$ be a positive semidefinite kernel function with $\sup_{z,z'}K(z,z')<+\infty$. Given a kernel function K, we can associate a feature map $\Phi:\mathcal{Z}\to\mathcal{H}$, where $\mathcal{H}$ is a Hilbert space with inner product $\langle .,.\rangle$ and for all $z,z'\in\mathcal{Z}$, $K(z,z')=\langle\Phi(z),\,\Phi(z')\rangle$. As a reproducing kernel Hilbert space, any function $\gamma\in\mathcal{H}$ can be expressed as an inner product $\gamma(z)=\langle w,\,\Phi(z)\rangle$, where w can be expressed as $w=\sum_{i=1}^m\alpha_i\Phi(z_i)$ for some $\alpha_1,\ldots,\alpha_m$ and $z_1,\ldots,z_m\in\mathcal{Z}$ for some m. See [Aronszajn, 1950] and [Saitoh, 1988] for general mathematical background on reproducing kernel Hilbert spaces, and [Schölkopf and Smola, 2002] for more details on learning approaches using kernel methods.

If we use this type of kernel class, then the classification rule γ can be written as $\gamma(z)=\sum_{i=1}^m\alpha_iK(z,z_i)$. Suppose that $\mathcal{C}$ is the subset of $\mathcal{H}$ given by

$$\mathcal{C} \;:=\; \Big\{\gamma\;\Big|\;\gamma(z)=\langle w,\,\Phi(z)\rangle,\;\;||w||\leq B\Big\}, \tag{4.50}$$

where $B>0$ is a constant that controls the “size” of the space. Assume further that the space $\mathcal{X}$ is discrete with M^S possible values, and that $\mathcal{Z}$ has L^S possible values. (Recall that S is the total number of covariates $(X_1,\ldots,X_S)$). In Chapter 3, Prop. 3.9 we proved that for the function class $\mathcal{G}$ defined in (4.48), the Rademacher complexity $R_n(\mathcal{G})$ is upper bounded by

$$\frac{2B}{n}\Bigg[\mathbb{E}\sup_{Q\in\mathcal{D}_0}\sum_{i=1}^{n}K(\text{argmax}_z Q_z(X_i),\text{argmax}_z Q_z(X_i))+ \\ 2(n-1)\sqrt{n/2}\sup_{z,z'}K(z,z')\sqrt{2MS\log L}\Bigg]^{1/2}, \quad (4.51)$$

which decays with order $O(1/n^{1/4})$. (We note in passing that this $O(1/n^{1/4})$ rate is not tight, but the bound is nonetheless useful for its particularly simple form).

It follows from Lemma 4.26 and equation (4.51) that $\mathcal{E}_1(\mathcal{C},\mathcal{D}) = O(B/n^{1/4})$, where B is the constant used to control the "size" of the function class $\mathcal{C}$ defined in equation (4.50). Let B_n denote the constant for the corresponding function class $\mathcal{C}_n$, and let $(B_n)_{n=1}^{\infty}$ be an increasing sequence such that $B_n \to +\infty$. Then, we see from the bound (4.51) that if B_n increases sufficiently slowly (i.e., slower than $n^{1/4}$), then the estimation error $\mathcal{E}_1(\mathcal{C}_n,\mathcal{D}_n) \to 0$. Note also that

$$|\gamma(z)| \;\leq\; ||w||\cdot||\Phi(z)|| \;=\; O(B_n),$$

so that $M_n = O(B_n)$ (where M_n is defined in equation (4.43)). As a consequence, we have $M_n\sqrt{\ln n/n} \to 0$, so that the estimation condition of condition of Theorem 4.23(b) holds.

4.7 Discussions

The main contribution of this chapter is a precise explication of the correspondence between loss functions that act as surrogates to the 0-1 loss (which are widely used in statistical machine learning), and the class of f-divergences (which are widely used in information theory and signal processing, and arise as error exponents in the large deviations setting). The correspondence helps explicate the use of various divergences in signal processing and quantization theory, as well as explain the behavior of surrogate loss functions often used in machine learning and statistics. Building on this foundation, we defined the notion of universal equivalence among divergences (and their associated loss functions). As an application of these ideas, we investigated the statistical behavior of a practical nonparametric kernel-based algorithm for designed decentralized hypothesis testing rules proposed in Chapter 3, and in particular proved that it is strongly consistent under appropriate conditions.

Appendix 4.A Proof of Lemma 4.5

(a) Since $\phi^{-1}(\beta) < +\infty$, we have $\phi(\phi^{-1}(\beta)) = \phi(\inf\{\alpha : \phi(\alpha) \leq \beta\}) \leq \beta$, where the final inequality follows from the lower semi-continuity of ϕ. If ϕ is continuous at $\phi^{-1}(\beta)$, then we have $\phi^{-1}(\beta) = \min\{\alpha : \phi(\alpha) = \beta\}$, in which case we have $\phi(\phi^{-1}(\beta)) = \beta$.
(b) Due to convexity and the inequality $\phi'(0) < 0$, it follows that ϕ is a strictly decreasing function in $(-\infty, \alpha^*]$. Furthermore, for all $\beta \in \mathbb{R}$ such that $\phi^{-1}(\beta) < +\infty$, we must have $\phi^{-1}(\beta) \leq \alpha^*$. Therefore, definition (4.18) and the (decreasing) monotonicity of ϕ imply that for any $a, b \in \mathbb{R}$, if $b \geq a \geq \inf \phi$, then $\phi^{-1}(a) \geq \phi^{-1}(b)$, which establishes that ϕ^{-1} is a decreasing function. In addition, we have $a \geq \phi^{-1}(b)$ if and only if $\phi(a) \leq b$.

Now, due to the convexity of ϕ, applying Jensen's inequality for any $0 < \lambda < 1$, we have $\phi(\lambda\phi^{-1}(\beta_1) + (1-\lambda)\phi^{-1}(\beta_2)) \leq \lambda\phi(\phi^{-1}(\beta_1) + (1-\lambda)\phi(\phi^{-1}(\beta_2)) \leq \lambda\beta_1 + (1-\lambda)\beta_2$. Therefore,

$$\lambda\phi^{-1}(\beta_1) + (1-\lambda)\phi^{-1}(\beta_2) \geq \phi^{-1}(\lambda\beta_1 + (1-\lambda)\beta_2),$$

implying the convexity of ϕ^{-1}.

Appendix 4.B Proof of Lemma 4.6

(a) We first prove the statement for the case of a decreasing function ϕ. First, if $a \geq b$ and $\phi^{-1}(a) \notin \mathbb{R}$, then $\phi^{-1}(b) \notin \mathbb{R}$, hence $\Psi(a) = \Psi(b) = +\infty$. If only $\phi^{-1}(b) \notin \mathbb{R}$, then clearly $\Psi(b) \geq \Psi(a)$ (since $\Psi(b) = +\infty$). If $a \geq b$, and both $\phi^{-1}(\alpha), \phi^{-1}(\beta) \in \mathbb{R}$, then from the previous lemma, $\phi^{-1}(a) \leq \phi^{-1}(b)$, so that $\phi(-\phi^{-1}(a)) \leq \phi(-\phi^{-1}(b))$, implying that Ψ is a decreasing function.

We next consider the case of a general function ϕ. For $\beta \in (\beta_1, \beta_2)$, we have $\phi^{-1}(\beta) \in (-\alpha^*, \alpha^*)$, and hence $-\phi^{-1}(\beta) \in (-\alpha^*, \alpha^*)$. Since ϕ is strictly decreasing in $(-\infty, \alpha^*]$, then $\phi(-\phi^{-1}(\beta))$ is strictly decreasing in (β_1, β_2). Finally, when $\beta < \inf \Psi = \phi(\alpha^*)$, $\phi^{-1}(\beta) \notin \mathbb{R}$, so $\Psi(\beta) = +\infty$ by definition.

(b) First of all, assume that ϕ is decreasing. By applying Jensen's inequality, for any $0 < \lambda < 1$, and γ_1, γ_2, we have:

$$\begin{aligned}
\lambda\Psi(\gamma_1) + (1-\lambda)\Psi(\gamma_2)) &= \lambda\phi(-\phi^{-1}(\gamma_1)) + (1-\lambda)\phi(-\phi^{-1}(\gamma_2) \\
&\geq \phi(-\lambda\phi^{-1}(\gamma_1) - (1-\lambda)\phi^{-1}(\gamma_2)) \qquad \text{due to convexity of } \phi \\
&\geq \phi(-\phi^{-1}(\lambda\gamma_1 + (1-\lambda)\gamma_2)) \\
&= \Psi(\lambda\gamma_1 + (1-\lambda)\gamma_2),
\end{aligned}$$

where the last inequality is due to the convexity of ϕ^{-1} and decreasing ϕ. Hence, Ψ is a convex function.

In general, the above arguments go through for any $\gamma_1, \gamma_2 \in [\beta_1, \beta_2]$. Since $\Psi(\beta) = +\infty$ for $\beta < \beta_1$, this implies that Ψ is convex in $(-\infty, \beta_2]$.

(c) For any $a \in \mathbb{R}$, from the definition of ϕ^{-1}, and due to the continuity of ϕ,

$$\begin{aligned} \{\beta \,\big|\, \Psi(\beta) = \phi(-\phi^{-1}(\beta)) \leq a\} &= \{\beta \,\big|\, -\phi^{-1}(\beta) \geq \phi^{-1}(a)\} \\ &= \{\beta \,\big|\, \phi^{-1}(\beta) \leq -\phi^{-1}(a)\} \\ &= \{\beta \,\big|\, \beta \geq \phi(-\phi^{-1}(a))\} \end{aligned}$$

is a closed set. Similarly, $\{\beta \in \mathbb{R} \,\big|\, \Psi(\beta) \geq a\}$ is a closed set. Hence Ψ is continuous in its domain.

(d) Since ϕ is assumed to be classification-calibrated, Lemma 4.2 implies that ϕ is differentiable at 0 and $\phi'(0) < 0$. Since ϕ is convex, this implies that ϕ is strictly decreasing for $\alpha \leq 0$. As a result, for any $\alpha \geq 0$, let $\beta = \phi(-\alpha)$, then we obtain $\alpha = -\phi^{-1}(\beta)$. Since $\Psi(\beta) = \phi(-\phi^{-1}(\beta))$, we have $\Psi(\beta) = \phi(\alpha)$. Hence, $\Psi(\phi(-\alpha)) = \phi(\alpha)$. Letting $u^* = \phi(0)$, then we have $\Psi(u^*) = u^*$, and $u^* \in (\beta_1, \beta_2)$.

(e) Let $\alpha = \Psi(\beta) = \phi(-\phi^{-1}(\beta)$. Then from (4.18), $\phi^{-1}(\alpha) \leq -\phi^{-1}(\beta)$. Therefore,

$$\Psi(\Psi(\beta)) = \Psi(\alpha) = \phi(-\phi^{-1}(\alpha)) \leq \phi(\phi^{-1}(\beta)) \leq \beta.$$

We have proved that Ψ is strictly decreasing for $\beta \in (\beta_1, \beta_2)$. As such, $\phi^{-1}(\alpha) = -\phi^{-1}(\beta)$. We also have $\phi(\phi^{-1}(\beta)) = \beta$. It follows that $\Psi(\Psi(\beta)) = \beta$ for all $\beta \in (\beta_1, \beta_2)$.
Remark: With reference to statement (b), if ϕ is not a decreasing function, then the function Ψ need not be convex on the entire real line. For instance, the following loss function generates a function Ψ that is not convex:

$$\phi(\alpha) = \begin{cases} (1-\alpha)^2 & \text{when } \alpha \leq 1 \\ 0 & \text{when } 1 \leq \alpha \leq 2 \\ \alpha - 2 & \text{otherwise.} \end{cases}$$

We have $\Psi(9) = \phi(2) = 0$, $\Psi(16) = \phi(3) = 1$, $\Psi(25/2) = \phi(-1+5/\sqrt{2}) = -3+5/\sqrt{2} > (\Psi(9) + \Psi(16))/2$.

Appendix 4.C Proof of Lemma 4.11

(a) The inequality $\Delta \leq V$ is trivial. On the other hand, the inequality $V^2 \leq \Delta$ follows by applying the Cauchy-Schwarz inequality:

$$\Delta = \sum_z \left(\frac{|\mu(z) - \pi(z)|}{\sqrt{\mu(z) + \pi(z)}}\right)^2 \sum_z \left(\sqrt{\mu(z) + \pi(z)}\right)^2 \geq \left(\sum_z |\mu(z) - \pi(z)|\right)^2 = V^2(\mu, \pi).$$

(b) Note that for any $z \in \mathcal{Z}$, we have $1 \leq \frac{(\sqrt{\mu(z)}+\sqrt{\pi(z)})^2}{\mu(z)+\pi(z)} \leq 2$. Applying these inequalities in the following expression

$$\Delta(\mu, \pi) = \sum_{z \in \mathcal{Z}} (\sqrt{\mu(z)} - \sqrt{\pi(z)})^2 \frac{(\sqrt{\mu(z)} + \sqrt{\pi(z)})^2}{\mu(z) + \pi(z)}$$

yields $2h^2 \leq \Delta \leq 4h^2$.
(c) See [Topsoe, 2000] for a proof.

Appendix 4.D Proof of Theorem 4.14

We first establish the equivalence (a) ⇔ (b). By the correspondence between 0-1 loss and an f-divergence with $f(u) = -\min(u, 1)$, and the remark following Proposition 4.4, we have $R_{bayes}(Q) = -I_f(\mu, \pi) = -I_{f_q}(P_1, P_{-1})$, where $f_q(u) := qf(\frac{1-q}{q}u) = -(1 - q)\min(u, \frac{q}{1-q})$. Hence, (a) ⇔ (b).

Next, we prove the equivalence (b) ⇔ (c). The implication (c) ⇒ (b) is immediate. Considering the reverse implication (b) ⇒ (c), we note that any convex function $f(u)$ can be uniformly approximated over a bounded interval as a sum of a linear function and $-\sum_k \alpha_k \min(u, c_k)$, where $\alpha_k > 0, c_k > 0$ for all k. For a linear function f, $I_f(P_{-1}, P_1)$ does not depend on P_{-1}, P_1. Using these facts, Statement (c) follows from Statement (b).

Appendix 4.E Proof of Lemma 4.18

Consider a joint distribution $\mathbb{P}(X, Y)$ defined by $\mathbb{P}(Y = -1) = q = 1 - \mathbb{P}(Y = 1)$ and

$$\mathbb{P}(X|Y = -1) \sim \text{Uniform}[0, b], \quad \text{and} \quad \mathbb{P}(X|Y = 1) \sim \text{Uniform}[a, c],$$

where $0 < a < b < c$. Let $Z \in \{1, 2\}$ be a quantized version of X. We assume Z is produced by a deterministic quantizer design Q specified by a threshold $t \in (a, b)$; in particular, we set $Q(z = 1|x) = 1$ when $x \geq t$, and $Q(z = 2|x) = 1$ when $x < t$. Under this quantizer design, we have

$$\begin{aligned}\mu(1) &= (1-q)\frac{t-a}{c-a}; \;\; \mu(2) = (1-q)\frac{c-t}{c-a}\\ \pi(1) &= q\frac{t}{b}; \;\; \pi(2) = q\frac{b-t}{b}.\end{aligned}$$

Therefore, the f-divergence between μ and π takes the form:

$$I_f(\mu,\pi) = \frac{qt}{b} f\left(\frac{(t-a)b(1-q)}{(c-a)tq}\right) + \frac{q(b-t)}{b} f\left(\frac{(c-t)b(1-q)}{(c-a)(b-t)q}\right).$$

If $f_1 \overset{u}{\approx} f_2$, then $I_{f_1}(\mu,\pi)$ and $I_{f_1}(\mu,\pi)$ have the same monotonicity property for any $q \in (0,1)$ as well for for any choice of the parameters q and $a < b < c$. Let $\gamma = \frac{b(1-q)}{(c-a)q}$, which can be chosen arbitrarily positive, and then define the function

$$F(f,t) = tf\left(\frac{(t-a)\gamma}{t}\right) + (b-t)f\left(\frac{(c-t)\gamma}{b-t}\right).$$

Note that the functions $F(f_1,t)$ and $F(f_2,t)$ have the same monotonicity property, for any positive parameters γ and $a < b < c$.

We now claim that $F(f,t)$ is a convex function of t. Indeed, using convex duality [Rockafellar, 1970], $F(f,t)$ can be expressed as follows:

$$\begin{aligned}F(f,t) &= t\sup_{r\in\mathbb{R}}\left\{\frac{(t-a)\gamma}{t}r - f^*(r)\right\} + (b-t)\sup_{s\in\mathbb{R}}\left\{\frac{(c-t)\gamma}{b-t}s - f^*(s))\right\}\\ &= \sup_{r,s}\left\{\frac{(t-a)r}{\gamma} - tf^*(r) + \frac{(c-t)s}{\gamma} - tf^*(s)\right\},\end{aligned}$$

which is a supremum over a linear function of t, thereby showing that $F(f,t)$ is convex of t.

It follows that both $F(f_1,t)$ and $F(f_2,t)$ are subdifferentiable everywhere in their domains; since they have the same monotonicity property, we must have

$$0 \in \partial F(f_1,t) \Leftrightarrow 0 \in \partial F(f_2,t). \tag{4.52}$$

It can be verified using subdifferential calculus (e.g,, [Hiriart-Urruty and Lemaréchal, 2001]) that:

$$\partial F(f,t) = \frac{a\gamma}{t}\partial f\left(\frac{(t-a)\gamma}{t}\right) + f\left(\frac{(t-a)\gamma}{t}\right) - f\left(\frac{(c-t)\gamma}{b-t}\right) + \frac{(c-b)\gamma}{b-t}\partial f\left(\frac{(c-t)\gamma}{b-t}\right).$$

Letting $u = \frac{(t-a)\gamma}{t}$, $v = \frac{(c-t)\gamma}{b-t}$, we have

$$\begin{aligned} 0 \in \partial F(f,t) \quad &\Leftrightarrow \quad 0 \in (\gamma - u)\partial f(u) + f(u) - f(v) + (v-\gamma)\partial f(v) && \text{(4.53a)} \\ &\Leftrightarrow \quad \exists \alpha \in \partial f(u), \beta \in \partial f(v) \ \text{ s.t. } 0 = (\gamma - u)\alpha + f(u) - f(v) + (v-\gamma)\beta && \text{(4.53b)} \\ &\Leftrightarrow \quad \exists \alpha \in \partial f(u), \beta \in \partial f(v) \ \text{ s.t. } \gamma(\alpha - \beta) = u\alpha - f(u) + f(v) - v\beta && \text{(4.53c)} \\ &\Leftrightarrow \quad \exists \alpha \in \partial f(u), \beta \in \partial f(v) \ \text{ s.t. } \gamma(\alpha - \beta) = f^*(\alpha) - f^*(\beta). && \text{(4.53d)} \end{aligned}$$

By varying our choice of $q \in (0,1)$, the number γ can take any positive value. Similarly, by choosing different positive values of a, b, c (such that $a < b < c$), we can ensure that u and v can take on any positive real values such that $u < \gamma < v$. Since equation (4.52) holds for any t, it follows that for any triples $u < \gamma < v$, equation (4.53d) holds for f_1 if and only if it also holds for f_2.

Considering a fixed pair $u < v$, first suppose that the function f_1 is linear on the interval $[u,v]$ with a slope s. In this case, equation (4.53d) holds for f_1 and any γ by choosing $\alpha = \beta = s$, which implies that equation (4.53d) also holds for f_2 for any γ. Thus, we deduce that f_2 is also a linear function on the interval $[u,v]$.

Suppose, on the other hand, that f_1 and f_2 are both non-linear in $[u,v]$. Due to the monotonicity of subdifferentials, we have $\partial f_1(u) \cap \partial f_1(v) = \emptyset$ and $\partial f_2(u) \cap \partial f_2(v) = \emptyset$. Consequently, it follows that both $T_{f_1}(u,v)$ and $T_{f_2}(u,v)$ are non-empty. If $\gamma \in T_{f_1}(u,v)$, then (4.53d) holds for f_1 for some γ. Thus, it must also hold for f_2 using the same γ, which implies that $\gamma \in T_{f_2}(u,v)$. The same argument can also be applied with the roles of f_1 and f_2 reversed, so that we conclude that $T_{f_1}(u,v) = T_{f_2}(u,v)$.

Appendix 4.F Proof of Proposition 4.19

Using Lemma 4.18, the proof of Proposition 4.19 follows relatively easily. Note that the variational distance corresponds to $f_1(u) = |u-1| = u + 1 - 2\min\{u,1\}$, which is linear above and below 1. Therefore, the same must be true for any continuous convex function f_2. All such functions can indeed be written as $-c\min(u,1) + au + b$, for some constant c, a, b. In order for f_2 to have the same monotonicity as f_1, it is necessary and sufficient that $c > 0$.

Appendix 4.G Proof of Proposition 4.24

Following a similar construction as in the proof of Proposition 4.20, all ϕ satisfying property $\mathcal{P}$ have $\phi(0) = (c-a-b)/2$. Now, note that

$$R_{bayes}(\gamma, Q) - R^*_{bayes} = R_{bayes}(\gamma, Q) - R_{bayes}(Q) + R_{bayes}(Q) - R^*_{bayes}$$
$$= \sum_{z \in \mathcal{Z}} \pi(z)\mathbb{I}(\gamma(z) > 0) + \mu(z)\mathbb{I}(\gamma(z) < 0) - \min\{\mu(z), \pi(z)\} + R_{bayes}(Q) - R^*_{bayes}$$
$$= \sum_{z:(\mu(z)-\pi(z))\gamma(z)<0} |\mu(z) - \pi(z)| + R_{bayes}(Q) - R^*_{bayes}.$$

In addition,

$$R_\phi(\gamma, Q) - R^*_\phi = R_\phi(\gamma, Q) - R_\phi(Q) + R_\phi(Q) - R^*_\phi.$$

By Proposition 4.4,

$$\begin{aligned} R_\phi(Q) - R^*_\phi &= -I_f(\mu, \pi) - \inf_{Q \in \mathcal{Q}} (-I_f(\mu, \pi)) \\ &= c \sum_{z \in \mathcal{Z}} \min\{\mu(z), \pi(z)\} - \inf_{Q \in \mathcal{Q}} c \sum_{z \in \mathcal{Z}} \min\{\mu(z), \pi(z)\} \\ &= c(R_{bayes}(Q) - R^*_{bayes}). \end{aligned}$$

Therefore, the lemma would be immediate once we could show that

$$\frac{c}{2} \sum_{z:(\mu(z)-\pi(z))\gamma(z)<0} |\mu(z) - \pi(z)| \leq R_\phi(\gamma, Q) - R_\phi(Q)$$
$$= \sum_{z \in \mathcal{Z}} \pi(z)\phi(-\gamma(z)) + \mu(z)\phi(\gamma(z)) - c\min\{\mu(z), \pi(z)\} + ap + bq. \quad (4.54)$$

It is simple to check that for any $z \in \mathcal{Z}$ such that $(\mu(z) - \pi(z))\gamma(z) < 0$, there holds:

$$\pi(z)\phi(-\gamma(z)) + \mu(z)\phi(\gamma(z)) \geq \pi(z)\phi(0) + \mu(z)\phi(0). \quad (4.55)$$

Indeed, w.o.l.g., suppose $\mu(z) > \pi(z)$. Since ϕ is classification-calibrated, the convex function (with respect to α) $\pi(z)\phi(-\alpha) + \mu(z)\phi(\alpha)$ achieves its minimum at some $\alpha \geq 0$. Hence, for any $\alpha \leq 0$, $\pi(z)\phi(-\alpha) + \mu(z)\phi(\alpha) \geq \pi(z)\phi(0) + \mu(z)\phi(0)$. Hence, (4.55) is proven. The RHS of Eqn. (4.54) is lower bounded by:

$$\begin{aligned}
&\sum_{z:(\mu(z)-\pi(z))\gamma(z)<0} (\pi(z)+\mu(z))\phi(0) - c\min\{\mu(z),\pi(z)\} + ap + bq \\
&= \sum_{z:(\mu(z)-\pi(z))\gamma(z)<0} (\pi(z)+\mu(z))\frac{c-a-b}{2} - c\min\{\mu(z),\pi(z)\} + ap + bq \\
&= \frac{c}{2}\sum_{z:(\mu(z)-\pi(z))\gamma(z)<0} |\mu(z)-\pi(z)| - (a+b)(p+q)/2 + ap + bq \\
&= \frac{c}{2}\sum_{z:(\mu(z)-\pi(z))\gamma(z)<0} |\mu(z)-\pi(z)| + \frac{1}{2}(a-b)(p-q) \\
&\geq \frac{c}{2}\sum_{z:(\mu(z)-\pi(z))\gamma(z)<0} |\mu(z)-\pi(z)|.
\end{aligned}$$

This completes the proof.

Chapter 5

Decentralized sequential detection

In this chapter we consider the problem of sequential decentralized detection, a problem that entails several interdependent choices: the choice of a stopping rule (specifying the sample size), a global decision function (a choice between two competing hypotheses), and a set of quantization rules (the local decisions on the basis of which the global decision is made). We resolve an open problem concerning whether optimal local decision functions for the Bayesian formulation of sequential decentralized detection can be found within the class of stationary rules. We develop an asymptotic approximation to the optimal cost of stationary quantization rules and show how this approximation yields a negative answer to the stationarity question. We also consider the class of blockwise stationary quantizers and show that asymptotically optimal quantizers are likelihood-based threshold rules.

5.1 Introduction

In Chapter 3 and Chapter 4 we have studied the problem of non-sequential decentralized detection. Detection is a classical discrimination or hypothesis-testing problem, in which observations $\{X_1, X_2, \ldots\}$ are assumed to be drawn i.i.d. from the (multivariate) conditional distribution $\mathbb{P}(\cdot \mid H)$ and the goal is to infer the value of the random variable H, which takes values in $\{0, 1\}$. In a typical engineering application, the case $\{H = 1\}$ represents the presence of some target to be detected, whereas $\{H = 0\}$ represents its absence. Placing this problem in a communication-theoretic context, a decentralized detection problem is a hypothesis-testing problem in which the decision-maker is not given access to the raw data points X_n, but instead must infer H based only on the output of a set of quantization rules or local decision functions, say $\{U_n = \phi_n(X_n)\}$, which map the raw data to quantized values. Of interest in this chapter is the extension to an-online setting: more specifically, the *sequential decentralized detection* problem [Tsitsiklis, 1986; Veeravalli, 1999; Mei, 2003] involves a data sequence, $\{X_1, X_2, \ldots\}$, and a corresponding sequence of summary statistics, $\{U_1, U_2, \ldots\}$, determined by a sequence of local decision rules $\{\phi_1, \phi_2, \ldots\}$. The goal is to design both the local decision functions and to specify a global decision rule so as to predict H in a manner that optimally trades off accuracy and delay. In short, the sequential decentralized detection problem is the communication-constrained extension of classical formulation of sequential centralized decision-making problems; see, e.g., [Chernoff, 1972; Shiryayev, 1978; Lai, 2001] to the decentralized setting.

In setting up a general framework for studying sequential decentralized problems, Veeravalli et al. [Veeravalli *et al.*, 1993] defined five problems, denoted "Case A" through "Case E", distinguished from one another by the amount of information available to the local sensors. In applications such as power-constrained sensor networks, we generally do not wish to assume that there are high-bandwidth feedback channels from the decision-maker to the sensors, nor do we wish to assume that the sensors have unbounded memory. Most suited to this perspective—and the focus of this thesis—is Case A, in which the local decisions are of the simplified form $\phi_n(X_n)$; i.e., neither local memory nor feedback are assumed to be available. Noting that Case A is not amenable to dynamic programming and is presumably intractable, Veeravalli et al. [Veeravalli *et al.*, 1993] suggested restricting the analysis to the class of *stationary* local decision functions; i.e., local decision functions ϕ_n that are independent of n. They conjectured that stationary decision functions may actually be optimal in the setting of Case A (given the intuitive symmetry and high degree of independence of the problem in this case), even though it is not possible to verify this optimality via DP arguments. This conjecture has remained open since it was first posed by Veeravalli et al. [Veeravalli *et al.*, 1993; Veeravalli, 1999].

The main contribution of this chapter is to resolve this question by showing that stationary decision functions are, in fact, *not* optimal for decentralized problems of type A. Our argument is based on an asymptotic characterization of the optimal Bayesian risk as the cost per sample goes to zero. In this asymptotic regime, the optimal cost can be expressed as a simple function of priors and Kullback-Leibler (KL) divergences. This characterization allows us to construct counterexamples to the stationarity conjecture, both in an exact and an asymptotic setting. In the latter setting, we present a class of problems in which there always exists a range of prior probabilities for which stationary strategies, either deterministic or randomized, are suboptimal. We note in passing that an intuition for the source of this suboptimality is easily provided—it is due to the asymmetry of the KL divergence.

It is well known that optimal quantizers when unrestricted are necessarily likelihood-based threshold rules [Tsitsiklis, 1986]. Our counterexamples and analysis imply that optimal thresholds are not generally stationary (i.e., the threshold may differ from sample to sample). We also provide a partial converse to this result: specifically, if we restrict ourselves to stationary (or blockwise stationary) quantizer designs, then there exists an optimal design that is a threshold rule based on the likelihood ratio. We prove this result by establishing a quasiconcavity result for the asymptotically optimal cost function. In this chapter, this result is proven for the space of deterministic quantizers with arbitrary output alphabets, as well as for the space of randomized quantizers with binary ouputs. We conjecture that the same result holds more generally for randomized quantizers with arbitrary output alphabets.

The remainder of this chapter is organized as follows. We begin in Section 5.2 with background on the Bayesian formulation of sequential detection problems, and Wald's approximation. Section 5.3 provides a simple asymptotic approximation of the optimal cost that underlies our main analysis in Section 5.4. In Section 5.5, we establish the existence of optimal decision rules that are likelihood-based threshold rules, under the restriction to blockwise stationarity. We conclude with a discussion in Section 5.6 [1]

5.2 Background

This chapter provides background on the Bayesian formulation of sequential (centralized) detection problems. Of particular use in our subsequent analysis is Wald's approximation of the cost of optimal sequential test.

Let $\mathbb{P}_0$ and $\mathbb{P}_1$ represent the distribution of X, when conditioned on $\{H = 0\}$ and $\{H = 1\}$ respectively. Assume that $\mathbb{P}_0$ and $\mathbb{P}_1$ are absolutely continuous with respect to one another. We use $f^0(x)$ and $f^1(x)$ to denote the respective density functions with respect to some dominating measure (e.g., Lebesgue for continuous variables, or counting measure for discrete-valued variables).

Our focus is the Bayesian formulation of the sequential detection problem [Shiryayev, 1978; Veeravalli, 1999]; accordingly, we let $\pi^1 = \mathbb{P}(H = 1)$ and $\pi^0 = \mathbb{P}(H = 0)$ denote the prior probabilities of the two hypotheses. Let $X_1, X_2, \ldots$ be a sequence of conditionally i.i.d. realizations of X. A sequential decision rule consists of a *stopping time* N defined with respect to the sigma field $\sigma(X_1, \ldots, X_N)$, and a decision function γ measurable with respect to $\sigma(X_1, \ldots, X_N)$. The cost function is the expectation of a weighted sum of the sample size N and the probability of incorrect decision—namely

$$J(N, \gamma) := \mathbb{E}\big\{cN + \mathbb{I}[\gamma(X_1, \ldots, X_N) \neq H]\big\}, \tag{5.1}$$

where $c > 0$ is the incremental cost of each sample. The overall goal is to choose the pair (N, γ) so as to minimize the expected loss (5.1).

It is well known that the optimal solution of the sequential decision problem can be characterized recursively using dynamic programming (DP) arguments [Arrow *et al.*, 1949; Wald and Wolfowitz, 1948; Shiryayev, 1978; Bertsekas, 1995a]. Although useful in classical (centralized) sequential detection, the DP approach is not always straightforward to

[1] This work has been published in [Nguyen *et al.*, 2006].

apply to *decentralized* versions of sequential detection [Veeravalli, 1999]. In the remainder of this section, we describe an asymptotic approximation of the optimal sequential cost, originally due to Wald (cf. [Siegmund, 1985]), valid as $c \to 0$. To sketch out Wald's approximation, we begin by noting the optimal stopping rule for the cost function (5.1) takes the form

$$N = \inf\left\{n \geq 1 \;\middle|\; L_n(X_1, \ldots, X_n) := \sum_{i=1}^{n} \log \frac{f^1(X_i)}{f^0(X_i)} \notin (a, b)\right\}, \tag{5.2}$$

for some real numbers $a < b$. Given this stopping rule, the optimal decision function has the form

$$\gamma(L_N) = \begin{cases} 1 & \text{if } L_N \geq b, \\ 0 & \text{if } L_N \leq a. \end{cases} \tag{5.3}$$

Consider the two types of error:

$$\begin{aligned} \alpha &= \mathbb{P}_0(\gamma(L_N) \neq H) = \mathbb{P}_0(L_N \geq b) \\ \beta &= \mathbb{P}_1(\gamma(L_N) \neq H) = \mathbb{P}_1(L_N \leq a). \end{aligned}$$

As $c \to 0$, it can be shown that the optimal choice of a and b satisfies $a \to -\infty, b \to \infty$, and the corresponding α, β satisfy $\alpha + \beta \to 0$. Ignoring the overshoot of L_N upon the optimal stopping time N (i.e., instead assuming L_N attains precisely the value a or b) we can express a, b, $\mathbb{E}N$ and the cost function J in terms of α and β as follows [Wald, 1947]:

$$a \approx a(\alpha, \beta) := \log \frac{\beta}{1-\alpha} \quad \text{and} \quad b \approx b(\alpha, \beta) := \log \frac{1-\beta}{\alpha} \tag{5.4}$$

$$\mathbb{E}_0[L_N] \approx (1-\alpha)a + \alpha b \quad \text{and} \quad \mathbb{E}_1[L_N] \approx (1-\beta)b + \beta a \tag{5.5}$$

Now define the Kullback-Leibler divergences

$$\mu^1 = \mathbb{E}_1[\log \frac{f^1(X_1)}{f^0(X_1)}] = D(f^1||f^0), \qquad \text{and} \qquad \mu^0 = -\mathbb{E}_0[\log \frac{f^1(X_1)}{f^0(X_1)}] = D(f^0||f^1). \tag{5.6}$$

With a slight abuse of notation, we shall also use $D(\alpha, \beta)$ to denote a function in $[0, 1]^2 \to \mathbb{R}$ such that:

$$D(\alpha, \beta) := \alpha \log \frac{\alpha}{\beta} + (1-\alpha) \log \frac{1-\alpha}{1-\beta}.$$

With the above approximations, the cost function J of the decision rule based on envelopes

a and b can be written as

$$\begin{aligned} J &= \pi^1\mathbb{E}_1(cN + \mathbb{I}[L_N \leq a]) + \pi^0\mathbb{E}_0(cN + \mathbb{I}[L_N \geq b]) \\ &= c\pi^1\frac{\mathbb{E}_1 L_N}{\mu^1} + c\pi^0\frac{\mathbb{E}_0 L_N}{-\mu^0} + \pi^0\alpha + \pi^1\beta, \qquad (5.7) \\ &\approx c\pi^0\frac{D(\alpha, 1-\beta)}{\mu^0} + c\pi^1\frac{D(1-\beta,\alpha)}{\mu^1} + \pi^0\alpha + \pi^1\beta, \qquad (5.8) \end{aligned}$$

where the third line follows from Wald's equation [Wald, 1947].

Let J^* denote the cost of an optimal sequential test. Since $\alpha + \beta \to 0$, $D(1-\beta,\alpha) = \log(1/\alpha) + o(1)$, and $D(1-\alpha,\beta) = \log(1/\beta) + o(1)$. We approximate J^* by minimizing J over α and β. The minimum is achieved at $\alpha^* = \frac{c\pi^1}{\mu^1_\phi \pi^0}$ and $\beta^* = \frac{c\pi^0}{\mu^0_\phi\pi^1}$, yielding:

$$\begin{aligned} J^* &\approx \inf_{\alpha,\beta}\left\{\pi^0\alpha + \pi^1\beta + c\pi^0\frac{\log(1/\beta)}{\mu^0} + c\pi^1\frac{\log(1/\alpha)}{\mu^1}\right\} + o(c) \\ &\approx (\frac{\pi^0}{\mu^0} + \frac{\pi^1}{\mu^1})c\log c^{-1} + O(c). \qquad (5.9) \end{aligned}$$

The approximations described here can be made rigorous using the results of Chernoff [Chernoff, 1959].

5.3 Characterization of optimal stationary quantizers

Turning now to the decentralized setting, the primary challenge lies in the design of the quantization rules ϕ_n applied to data X_n. When X_n is univariate, a deterministic quantization rule ϕ_n is a function that maps $\mathcal{X}$ to the discrete space $\mathcal{U} = \{0,\ldots,K-1\}$ for some natural number K. For multivariate X_n with d dimension arising from the multiple sensor setting, a deterministic quantizer ϕ_n is defined as a mapping from the d-dim product space $\mathcal{X}$ to $\mathcal{U} = \{0,\ldots,K-1\}^d$. In the decentralized problem defined as Case A by Veeravalli et al. [Veeravalli *et al.*, 1993], the function ϕ_n is composed of d separate quantizer functions, one each for each dimension. A randomized quantizer ϕ_n is obtained by placing a distribution over the space of deterministic quantizers.

Any fixed set of quantization rules ϕ_n yields a sequence of compressed data $U_n = \phi_n(X_n)$, to which the classical theory can be applied. We are thus interested in choosing quantization rules $\phi_1, \phi_2, \ldots$ so that the error resulting from applying the optimal sequential test to the sequence of statistics $U_1, U_2, \ldots$, is minimized over some space Φ of quantization rules. For a given quantizer ϕ_n we use

$$f^i_{\phi_n}(u) \quad := \quad \mathbb{P}_i(\phi_n(X_n) = u), \qquad \text{for} \quad i = 0, 1,$$

to denote the distributions of the compressed data, conditioned on the hypothesis. In general, when randomized quantizers are allowed, the vector $(f^0_{\phi_n}(.), f^1_{\phi_n}(.))$ ranges over a convex set, denoted $\text{Conv}\Phi$, whose extreme points correspond to deterministic quantizers based on likelihood ratio threshold rules [Tsitsiklis, 1993a].

We say that a quantizer design is *stationary* if the rule ϕ_n is independent of n; in this case, we simplify the notation to f^1_ϕ and f^0_ϕ. In addition, we define the KL divergences $\mu^1_\phi := D(f^1_\phi || f^0_\phi)$ and $\mu^0_\phi := D(f^0_\phi || f^1_\phi)$. Moreover, let J_ϕ and J^*_ϕ denote the analogues of the functions J in Eq. (5.7) and J^* in (5.9), respectively, defined using μ^i_ϕ, for $i = 0, 1$. In this scenario, the sequence of compressed data $U_1, \ldots, U_n, \ldots$ are drawn i.i.d. from either f^0_ϕ or f^1_ϕ. Thus we can use the approximation (5.9) to characterize the asymptotically optimal stationary quantizer design. This is stated formally in the lemma to follow.

We begin by stating the assumptions underlying the lemma. For a given class of quantizers Φ, we assume that the Kullback-Leibler divergences are uniformly bounded away from zero

$$D(f^1_\phi || f^0_\phi) > 0, D(f^0_\phi || f^1_\phi) > 0 \text{ for all } \phi \in \Phi \tag{5.10}$$

and moreover that the variance of the log likelihood ratios are bounded

$$\sup_{\phi \in \Phi} \text{Var}_{f^1_\phi} \log(f^1_\phi / f^0_\phi)) < \infty, \text{ and } \sup_{\phi \in \Phi} \text{Var}_{f^0_\phi} \log(f^1_\phi / f^0_\phi)) < \infty. \tag{5.11}$$

Examples of distributions that satisfy these assumptions include pairs of discrete distributions, pairs of Gaussian distributions, and so on.

Lemma 5.1. *Under assumptions* (5.10) *and* (5.11), *the optimal stationary cost takes the form*

$$J^*_\phi = \left(\frac{\pi^0}{\mu^0_\phi} + \frac{\pi^1}{\mu^1_\phi} \right) c \log c^{-1} \; (1 + o(1)) \tag{5.12}$$

as $c \to 0$.

Proof: We prove the lemma using results originally due to Chernoff [Chernoff, 1959], restricted to a simple binary hypothesis test between f^0_ϕ and f^1_ϕ. By Theorem 1 from Chernoff [Chernoff, 1959], under conditions (5.10) and (5.11), there is a sequential test (N, γ) for which:

$$\begin{aligned} J^* \leq J(N, \gamma) \quad &= \quad \pi^0(\alpha + c\mathbb{E}_0 N) + \pi^1(\beta + c\mathbb{E}_1 N) \\ &\leq \quad \pi^0(1 + o(1)) c \log c^{-1} / \mu^0_\phi + \pi^1(1 + o(1)) c \log c^{-1} / \mu^1_\phi. \end{aligned}$$

But then the optimal test with the cost J^* (i.e., the likelihood ratio based test) must satisfies that $\alpha + c\mathbb{E}_0 N = O(c\log c^{-1})$ and $\beta + c\mathbb{E}_1 N = O(c\log c^{-1})$. Theorem 2 of Chernoff [Chernoff, 1959] implies that

$$J^* \geq \left(\frac{\pi^0}{\mu^0_\phi} + \frac{\pi^1}{\mu^1_\phi}\right)(1+o(1))c\log c^{-1},$$

concluding the proof.
Remarks:

1. The preceding approximation of the optimal cost essentially ignores the overshoot of the likelihood ratio L_N. While it is possible to analyze this overshoot to obtain a finer approximation (cf. [Lorden, 1970; Siegmund, 1985; Lai, 2001; Poor, 1994]), we see that this is not needed for our purpose. Lemma 5.1 shows that given a fixed prior (π^0, π^1), among all stationary quantizer designs in Φ, ϕ is optimal for sufficiently small c if *and* only if ϕ minimizes what we shall call the *sequential cost coefficient*:

$$G_\phi := \frac{\pi^0}{\mu^0_\phi} + \frac{\pi^1}{\mu^1_\phi}.$$

2. As a consequence of Lemma 5.7 to be proved in the sequel, if we consider the class Φ of all binary randomized quantizers, then sequential cost coefficient G_ϕ is a quasiconcave function with respect to $(f^0_\phi(.), f^1_\phi(.))$. (A function F is quasiconcave if and only if for any η, the level set $\{F(x) \geq \eta\}$ is a convex set; see Boyd and Vandenberghe [Boyd and Vandenberghe, 2004] for further background). The minimum of a quasiconcave function lies in the set of extreme points in its domain. For the set $\mathrm{Conv}\Phi$, these extreme points correspond to deterministic quantizers based on likelihood ratios [Tsitsiklis, 1993b]. Consequently, we conclude that for quantizers with binary outputs, the optimal cost is not decreased by considering randomized quantizers. We conjecture that this statement also holds beyond the binary case.

Section 5.5 is devoted to a more detailed study of asymptotically optimal stationary quantizers. In the meantime, we turn to the question whether stationary quantizers are optimal in either finite-sample or asymptotic settings.

5.4 Suboptimality of stationary designs

It was shown by Tsitsiklis [Tsitsiklis, 1986] that optimal quantizers ϕ_n take the form of threshold rules based on the likelihood ratio $f^1(X_n)/f^0(X_n)$. Veeravalli et al. [Veeravalli *et al.*, 1993; Veeravalli, 1999] asked whether these rules can always be taken to be stationary, a conjecture that has remained open. In this section, we resolve this question with a negative answer for both the finite-sample and asymptotic settings.

5.4.1 Suboptimality in exact setting

We begin by providing a numerical counterexample for which stationary designs are suboptimal. Consider a problem in which $X \in \mathcal{X} = \{1, 2, 3\}$ and the conditional distributions take the form

$$f^0(x) = \begin{bmatrix} \frac{8}{10} & \frac{1999}{10000} & \frac{1}{10000} \end{bmatrix} \text{ and } f^1(x) = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}.$$

Suppose that the prior probabilities are $\pi^1 = \frac{8}{100}$ and $\pi^0 = \frac{92}{100}$, and that the cost for each sample is $c = \frac{1}{100}$.

If we restrict to binary quantizers (i.e., $\mathcal{U} = \{0, 1\}$), by the symmetric roles of the output alphabets there are only three possible deterministic quantizers:

1. Design A: $\phi_A(X_n) = 0 \iff X_n = 1$. As a result, the corresponding distribution for U_n is specified by $f^0_{\phi_A}(u_n) = [\frac{4}{5}\ \frac{1}{5}]$ and $f^1_{\phi_A}(u) = [\frac{1}{3}\ \frac{2}{3}]$.
2. Design B: $\phi_B(X_n) = 0 \iff X_n \in \{1, 2\}$. The corresponding distribution for U_n is given by $f^0_{\phi_B}(u) = [\frac{9999}{10000}\ \frac{1}{10000}]$ and $f^1_{\phi_B}(u) = [\frac{2}{3}\ \frac{1}{3}]$.
3. Design C: $\phi_C(X_n) = 0 \iff X_n \in \{1, 3\}$. The corresponding distribution for U_n is specified by $f^0_{\phi_C} \sim [\frac{8001}{10000}\ \frac{1999}{10000}]$ and $f^1_{\phi_C}(u) = [\frac{2}{3}\ \frac{1}{3}]$.

Now consider the three stationary strategies, each of which uses only one fixed design, A, B or C. For any given stationary quantization rule ϕ, we have a classical centralized sequential problem, for which the optimal cost (achieved by a sequential probability ratio test) can be computed using a dynamic-programming procedure [Wald and Wolfowitz, 1948; Arrow *et al.*, 1949]. Accordingly, for each stationary strategy, we compute the optimal cost function J for 10^6 points on the p-axis by performing 300 updates of Bellman's equation (cf. [Bertsekas, 1995a]). In all cases, the difference in cost between the 299th and 300th updates is less than 10^{-6}. Let J_A, J_B and J_C denote the optimal cost function for sequential tests using all A's, all B's, and all C's, respectively. When evaluated at $\pi^1 = 0.08$, these computations yield $J_A = 0.0567$, $J_B = 0.0532$ and $J_C = 0.08$.

Finally, we consider a non-stationary rule obtained by applying design A for only the first sample, and applying design B for the remaining samples. Again using Bellman's equation, we find that the cost for this design is

$$\begin{aligned} J_* = \min\{\min\{\pi^1, 1 - \pi^1\}, c + J_B(P(H = 1|u_1 = 0))P(u_1 = 0) + \\ J_B(P(H = 1|u_1 = 1))P(u_1 = 1)\} = 0.052767, \end{aligned}$$

which is better than any of the stationary strategies.

In this particular example, the cost J^* of the non-stationary quantizer yields a slim improvement (0.0004) over the best stationary rule J_B. This slim margin is due in part to the choice of a small per-sample cost $c = 0.01$; however, larger values of c do not yield counterexample when using the particular distributions specified above. A more significant factor is that our non-stationary rule differs from the optimal stationary rule B only in its treatment of the first sample. This fact suggests that one might achieve better cost by alternating between using design A and design B on the odd and even samples, respectively. Our analysis of the asymptotic setting in the next section confirms this intuition.

5.4.2 Asymptotic suboptimality for both deterministic and randomized quantizers

We now prove that in a broad class of examples, there is a range of prior probabilities for which stationary quantizer designs are suboptimal. Our result stems from the following observation: Lemma 5.1 implies that in order to achieve a small cost we need to choose a quantizer ϕ for which the KL divergences $\mu_\phi^0 := D(f_\phi^0 || f_\phi^1)$ and $\mu_\phi^1 := D(f_\phi^1 || f_\phi^0)$ are both as large as possible. Due to the asymmetry of the KL divergence, however, these maxima are not necessarily achieved by a single quantizer ϕ. This suggests that one could improve upon stationary designs applying different quantizers to different samples, as the following lemma shows.

Lemma 5.2. *Let ϕ_1 and ϕ_2 be any two quantizers. If the following inequalities hold*

$$\mu_{\phi_1}^0 < \mu_{\phi_2}^0 \textit{ and } \mu_{\phi_1}^1 > \mu_{\phi_2}^1 \tag{5.13}$$

then there exists a non-empty interval $(U, V) \subseteq (0, +\infty)$ such that as $c \to 0$,

$$\begin{aligned} J_{\phi_1}^* \leq J_{\phi_1,\phi_2}^* \leq J_{\phi_2}^* \quad &\textit{if} \quad \frac{\pi^0}{\pi^1} \leq U \\ J_{\phi_1,\phi_2}^* < \min\{J_{\phi_1}^*, J_{\phi_2}^*\} - \Theta(c \log c^{-1}) \quad &\textit{if} \quad \frac{\pi^0}{\pi^1} \in (U, V) \\ J_{\phi_1}^* \geq J_{\phi_1,\phi_2}^* \geq J_{\phi_2}^* \quad &\textit{if} \quad \frac{\pi^0}{\pi^1} \geq V, \end{aligned}$$

where J_{ϕ_1,ϕ_2}^* denotes the optimal cost of a sequential test that alternates between using ϕ_1 and ϕ_2 on odd and even samples respectively.

Proof: According to Lemma 5.1, we have

$$J_{\phi_i}^* \quad = \quad \left(\frac{\pi^0}{\mu_{\phi_i}^0} + \frac{\pi^1}{\mu_{\phi_i}^1}\right) c \log c^{-1}(1 + o(1)), \quad i = 0, 1. \tag{5.14}$$

Now consider the sequential test that applies quantizers ϕ_1 and ϕ_2 alternately to odd and even samples. Furthermore, let this test consider two samples at a time. Let $f^0_{\phi_1\phi_2}$ and $f^1_{\phi_1\phi_2}$ denote the induced conditional probability distributions, jointly on the odd-even pairs of quantized variables. From the additivity of the KL divergence and assumption (5.13), there holds:

$$D(f^0_{\phi_1\phi_2}||f^1_{\phi_1\phi_2}) \;=\; \mu^0_{\phi_1} + \mu^0_{\phi_2} > 2\mu^0_{\phi_1} \tag{5.15a}$$

$$D(f^1_{\phi_1\phi_2}||f^0_{\phi_1\phi_2}) \;=\; \mu^1_{\phi_1} + \mu^1_{\phi_2} < 2\mu^1_{\phi_1}. \tag{5.15b}$$

Clearly, the cost of the proposed sequential test is an upper bound for $J^*_{\phi_1,\phi_2}$. Furthermore, the gap between this upper bound and the true optimal cost is no more than $O(c)$. Hence, as in the proof of Lemma 5.1, as $c \to 0$, the optimal cost $J^*_{\phi_1,\phi_2}$ can be written as

$$\left(\frac{2\pi^0}{\mu^0_{\phi_1} + \mu^0_{\phi_2}} + \frac{2\pi^1}{\mu^1_{\phi_1} + \mu^1_{\phi_2}}\right) c \log c^{-1}(1 + o(1)). \tag{5.16}$$

From equations (5.14) and (5.16), simple calculations yield the claim with

$$U = \frac{\mu^0_{\phi_1}(\mu^1_{\phi_1} - \mu^1_{\phi_2})(\mu^0_{\phi_1} + \mu^0_{\phi_2})}{\mu^1_{\phi_1}(\mu^1_{\phi_1} + \mu^1_{\phi_2})(\mu^0_{\phi_2} - \mu^0_{\phi_1})} < V = \frac{\mu^0_{\phi_2}(\mu^1_{\phi_1} - \mu^1_{\phi_2})(\mu^0_{\phi_1} + \mu^0_{\phi_2})}{\mu^1_{\phi_2}(\mu^1_{\phi_1} + \mu^1_{\phi_2})(\mu^0_{\phi_2} - \mu^0_{\phi_1})}. \tag{5.17}$$

□

Example: Let us return to the example provided in the previous section. Note that the two quantizers ϕ_A and ϕ_B satisfy assumption (5.13), since $D(f^0_{\phi_B}||f^1_{\phi_B}) = 0.4045 < D(f^0_{\phi_A}||f^1_{\phi_A}) = 0.45$ and $D(f^1_{\phi_B}||f^0_{\phi_B}) = 2.4337 > D(f^1_{\phi_A}||f^0_{\phi_A}) = 0.5108$. Furthermore, both quantizers dominates ϕ_C in terms of KL divergences: $D(f^0_{\phi_C}||f^1_{\phi_C}) = 0.0438$, $D(f^0_{\phi_C}||f^1_{\phi_C}) = 0.0488$. As a result, there exist a range of priors for which a sequential test using stationary quantizer design (either ϕ_A, ϕ_B or ϕ_C for all samples) is not optimal.

Theorem 5.3. *Suppose that Φ is a finite collection of quantizers, and that there is no single quantizer ϕ that dominates all other quantizers in Φ in the sense that*

$$\mu^0_\phi \;\geq\; \mu^0_{\phi'} \quad \textit{and} \quad \mu^1_\phi \;\geq\; \mu^1_{\phi'} \qquad \textit{for all} \quad \phi' \in \Phi. \tag{5.18}$$

Then there exists a non-empty range of prior probabilities for which no stationary design based on a quantizer in Φ is optimal.

Proof. Since there are a finite number of quantizers in Φ and no quantizer dominates all others, the interval $(0, \infty)$ is divided into at least two adjacent non-empty intervals, each of which corresponds to a range of prior probability ratios π^0/π^1 for which a quantizer is strictly optimal (asymptotically) among all stationary designs. Let them be (δ_1, δ) and (δ, δ_2), for two quantizers, namely, ϕ_1 and ϕ_2. In particular, δ is the value for π^0/π^1 for which the sequential cost coefficients are equal—viz. $G_{\phi_1} = G_{\phi_2}$—which happens only if assumption (5.13) holds. Some calculations verify that

$$\delta = \frac{\mu^0_{\phi_1}\mu^0_{\phi_2}(\mu^1_{\phi_2} - \mu^1_{\phi_1})}{\mu^1_{\phi_1}\mu^1_{\phi_2}(\mu^0_{\phi_1} - \mu^0_{\phi_2})}. \tag{5.19}$$

By Lemma 5.2, a non-stationary design by alternating between ϕ_1 and ϕ_2 has smaller sequential cost than both ϕ_1 and ϕ_2 for $\pi^0/\pi^1 \in (U, V)$, where U and V are given in equation (5.17). Since it can be verified that δ as defined (5.19) belongs to the interval (U, V), we conclude that for $\pi^0/\pi^1 \in (U, V) \cap (\delta_1, \delta_2)$, this non-stationary design has smaller cost than any stationary design using $\phi \in \Phi$. □

Remarks:

1. Suppose that Φ is restricted a finite class of deterministic quantizers with binary outputs. By the second remark following Lemma 5.1, it follows that stationary randomized quantizers are not optimal under the assumptions of Theorem 5.3.

2. It is interesting to contrast the Bayesian formulation of the problem of quantizer design with the Neyman-Pearson formulation. Our results on the suboptimality of stationary quantizer design in the Bayesian formulation repose on the asymmetry of the Kullback-Leibler divergence, as well as the sensitivity of the optimal quantizers on the prior probability. We note that Mei [Mei, 2003] (see p. 58) considered the Neyman-Pearson formulation of this problem. In this formulation, it can be shown that for all sequential tests for which the type 1 and type 2 errors are bounded by α and β, respectively, then as $\alpha + \beta \to 0$, the expected stopping time $\mathbb{E}_0 N$ under hypothesis $H = 0$ is asymptotically minimized by applying a stationary quantizer ϕ^* that maximizes $D(f^0_\phi || f^1_\phi)$. Similarly, the expected stopping time $\mathbb{E}_1 N$ under hypothesis $H = 1$ is asymptotically minimized by the stationary quantizer ϕ^{**} that maximizes $D(f^1_\phi || f^0_\phi)$ [Mei, 2003]. In this context, the example in section 5.4.1 provides a case in which the asymptotically minimal KL divergences ϕ^* and ϕ^{**} are not the same, due to the asymmetry, which suggests that there may not exist a stationary quantizer that simultaneously minimizes both $\mathbb{E}_1 N$ and $\mathbb{E}_0 N$.

5.4.3 Asymptotic suboptimality in multiple sensor setting

Our analysis thus far has established that with a single sensor per time step ($d = 1$), applying multiple quantizers to different samples can reduce the sequential cost. It is natural to ask whether the same phenomenon persists in the case of multiple sensors ($d > 1$). In this section, we show that the phenomenon does indeed carry over, more specifically by providing an example in which stationary strategies are still sub-optimal in comparison to non-stationary ones. The key insight is that we have only a fixed number of dimensions, whereas as $c \to 0$ we are allowed to take more samples, and each sample can act as an extra dimension, providing more flexibility for non-stationary strategies.

Suppose that the observation vector X_n at time n is d-dimensional, with each component corresponding to a sensor in a typical decentralized setting. Suppose that the observations from each sensor are assumed to be independent and identically distributed according to the conditional distributions defined in our earlier example (see section 5.4.1). Of interest are the optimal deterministic binary quantizer designs for all d sensors. Although there are three possible choices ϕ_A, ϕ_B and ϕ_C for each sensor, the quantizer ϕ_C is dominated by the other two, so each sensor should choose either ϕ_A and ϕ_B. Suppose that among these sensors, a subset of size t choose ϕ_A and whereas the remaining $d-t$ sensors choose ϕ_B for $0 \leq k \leq d$. We thus have $d+1$ possible stationary designs to consider. For each t, the sequential cost coefficient corresponding to the associated stationary design takes the form

$$G_k := \frac{\pi^0}{t\mu^0_{\phi_A} + (d-t)\mu^0_{\phi_B}} + \frac{\pi^1}{t\mu^1_{\phi_A} + (d-t)\mu^1_{\phi_B}}. \tag{5.20}$$

Now consider the following non-stationary design: the first sensor alternates between decision rules ϕ_A and ϕ_B, while the remaining $d-1$ sensors simply apply the stationary design based on ϕ_B. For this design, the associated sequential cost coefficient is given by

$$G := \frac{2\pi^0}{\mu^0_{\phi_A} + (2d-1)\mu^0_{\phi_B}} + \frac{2\pi^1}{\mu^1_{\phi_A} + (2d-1)\mu^1_{\phi_B}}. \tag{5.21}$$

Consider the interval (U, V), where the interval has endpoints

$$U = \frac{(\mu^1_{\phi_B} - \mu^1_{\phi_A})(\mu^0_{\phi_A} + (2d-1)\mu^0_{\phi_B})\mu^0_{\phi_B}}{(\mu^0_{\phi_A} - \mu^0_{\phi_B})(\mu^1_{\phi_A} + (2d-1)\mu^1_{\phi_B})\mu^1_{\phi_B}} < \\ V = \frac{(\mu^1_{\phi_B} - \mu^1_{\phi_A})(\mu^0_{\phi_A} + (2d-1)\mu^0_{\phi_B})(\mu^0_{\phi_A} + (d-1)\mu^0_{\phi_B})}{(\mu^0_{\phi_A} - \mu^0_{\phi_B})(\mu^1_{\phi_A} + (2d-1)\mu^1_{\phi_B})(\mu^1_{\phi_A} + (d-1)\mu^1_{\phi_B})}. \tag{5.22}$$

Since $\mu^0_{\phi_A} > \mu^0_{\phi_B}$ and $\mu^1_{\phi_B} > \mu^1_{\phi_A}$, straightforward calculations yield that for any prior likelihood $\pi^0/\pi^1 \in (U, V)$, the minimal cost over stationary designs $\min_{t=0,\ldots,d} G_k$ is strictly larger than the sequential cost G of the non-stationary design, previously defined in equation (5.21).

5.5 On asymptotically optimal blockwise stationary designs

Despite the possible loss in optimality, it is useful to consider some form of stationarity in order to reduce computational complexity of the optimization and decision process. In this section, we consider the class of *blockwise stationary* designs, meaning that there exists some natural number T such that $\phi_{T+1} = \phi_1, \phi_{T+2} = \phi_2$, and so on. For each T, let C_T denote the class of all blockwise stationary designs with period T. We assume throughout the analysis that each decision rule ϕ_n $(n = 1, \ldots, T)$ satisfies conditions (5.10) and (5.11). Thus, as T increases, we have a hierarchy of increasingly rich quantizer classes that will be seen to yield progressively better approximations to the optimal solution.

For a fixed prior (π^0, π^1) and $T > 0$, let $(\phi_1, \ldots, \phi_T)$ denote a quantizer design in C_T. As before, the cost J^*_ϕ of an asymptotically optimal sequential test using this quantizer design is of order $c \log c^{-1}$ with the sequential cost coefficient

$$G_\phi = \frac{T\pi^0}{\mu^0_{\phi_1} + \ldots + \mu^0_{\phi_T}} + \frac{T\pi^1}{\mu^1_{\phi_1} + \ldots + \mu^1_{\phi_T}}. \tag{5.23}$$

G_ϕ is a function of the vector of probabilities introduced by the quantizer: $(f^0_\phi(.), f^1_\phi(.))$. We are interested in the properties of a quantization rule ϕ that minimizes J^*_ϕ.

It is well known [Tsitsiklis, 1986] that optimal quantizers—*when unrestricted*—can be expressed as threshold rules based on the log likelihood ratio (LLR). Our counterexamples in the previous sectionimply that the thresholds need not be stationary (i.e., the threshold may differ from sample to sample). In the remainder of this section, we addresses a partial converse to this issue: specifically, if we restrict ourselves to stationary (or blockwise stationary) quantizer designs, then there exists an optimal design consisting of LLR-based threshold rules.

In the analysis to follow, it is sufficient to assume $T = 1$ so as to simplify the exposition. Our main result, stated below as Theorem 5.8, provides a characterization of the optimal quantizer ϕ^*_1, denoted more simply by ϕ^*. For $T > 1$, due to the symmetry in the roles of individual quantizer functions, ϕ_n, for $n = 1, \ldots, T$, result can be obtained by proving for each n while the quantizer rules for other time steps in the period remain fixed. Indeed, fixing the rules for other time steps except for $n = 1$, the sequential cost coefficient has the form:

$$G_\phi = \frac{T\pi^0}{\mu^0_{\phi_1} + d_0} + \frac{T\pi^1}{\mu^1_{\phi_1} + d_1},$$

for some non-negative constants d_0 and d_1. It can be verified that our proof for the case $T = 1$, which corresponds to $d_0 = d_1 = 0$, can be extended to the general case of $d_0, d_1 \geq 0$ in a straightforward manner.

Definition 5.4. *The quantizer design function* $\phi : \mathcal{X} \to \mathcal{U}$ *is said to be a* likelihood ratio threshold rule *if there are thresholds* $d_0 = -\infty < d_1 < \ldots < d_K = +\infty$*, and a permutation* $(u_1, \ldots, u_K)$ *of* $(0, 1, \ldots, K-1)$ *such that for* $l = 1, \ldots, K$*, with* $\mathbb{P}_0$*-probability 1, we have:*

$$\phi(X) = u_l \text{ if } d_{l-1} \leq f^1(X)/f^0(X) \leq d_l,$$

When $f^1(X)/f^0(X) = d_{l-1}$*, set* $\phi(X) = u_{l-1}$ *or* $\phi(X) = u_l$ *with* $\mathbb{P}_0$*-probability 1.*[2]

Previous work on the extremal properties of likelihood ratio based quantizers guarantees that the Kullback-Leibler divergence is maximized by a LLR-based quantizer [Tsitsiklis, 1993a]. In our case, however, the sequential cost coefficient G_ϕ involves a pair of KL divergences, μ^0_ϕ and μ^1_ϕ, which are related to one another in a nontrivial manner. Hence, establishing asymptotic optimality of LLR-based rules for this cost function does not follow from existing results, but rather requires further understanding of the interplay between these two KL divergences.

The following lemma concerns certain "unnormalized" variants of the Kullback-Leibler (KL) divergence. Given vectors $a = (a_0, a_1)$ and $b = (b_0, b_1)$, we define functions $\tilde{D}^0$ and $\tilde{D}^1$ mapping from $\mathbb{R}^4_+$ to the real line as follows:

$$\tilde{D}^0(a, b) \quad := \quad a_0 \log \frac{a_0}{a_1} + b_0 \log \frac{b_0}{b_1} \tag{5.24a}$$

$$\tilde{D}^1(a, b) \quad := \quad a_1 \log \frac{a_1}{a_0} + b_1 \log \frac{b_1}{b_0}. \tag{5.24b}$$

These functions are related to the standard (normalized) KL divergence via the relations $\tilde{D}^0(a, 1-a) \equiv D(a_0, a_1)$, and $\tilde{D}^1(a, 1-a) \equiv D(a_1, a_0)$.

Lemma 5.5. *For any positive scalars* $a_1, b_1, c_1, a_0, b_0, c_0$ *such that* $\frac{a_1}{a_0} < \frac{b_1}{b_0} < \frac{c_1}{c_0}$, at least one *of the two following conditions must hold:*

$$\tilde{D}^0(a, b+c) > \tilde{D}^0(b, c+a) \quad \textit{and} \quad \tilde{D}^1(a, b+c) > \tilde{D}^0(b, c+a), \quad \textit{or} \tag{5.25a}$$

$$\tilde{D}^0(c, a+b) > \tilde{D}^0(b, c+a) \quad \textit{and} \quad \tilde{D}^1(c, a+b) > \tilde{D}^0(b, c+a). \tag{5.25b}$$

This lemma implies that under certain conditions on the ordering of the probability ratios, one can increase *both* KL divergences by re-quantizing. This insight is used in the following lemma to establish that the optimal quantizer ϕ behaves almost like a likelihood ratio rule. To state the result, recall that the *essential supremum* is the infimum of the set of all η such that $f(x) \leq \eta$ for $\mathbb{P}_0$-almost all x in the domain, for a measurable function f.

[2] This last requirement of the definition is termed the *canonical* likelihood ratio quantizer by Tsitsiklis [Tsitsiklis, 1993a]. Although one could consider performing additional randomization when there are ties, our later results (in particular, Lemma 5.7) establish that in this case, randomization will not further decrease the optimal cost J^*_ϕ.

Lemma 5.6. *If ϕ is an asymptotically optimal quantizer, then for all pairs $(u_1, u_2) \in \mathcal{U}$, $u_1 \neq u_2$, there holds:*

$$\frac{f^1(u_1)}{f^0(u_1)} \notin \left(\text{ess} \inf_{x:\phi(x)=u_2} \frac{f^1(x)}{f^0(x)}, \; \text{ess} \sup_{x:\phi(x)=u_2} \frac{f^1(x)}{f^0(x)} \right).$$

Note that a likelihood ratio rule guarantees something stronger: For $\mathbb{P}_0$-almost all x such that $\phi(x) = u_1$, $f^1(x)/f^0(x)$ takes a value either to the left or to the right, but not to both sides, of the interval specified above. As we shall show, the proof that there exists an optimal LLR-based rule turns out to reduce to the problem of showing that the sequential cost coefficient G_ϕ is a *quasiconcave* function with respect to $(f^0_\phi(.), f^1_\phi(.))$. Since the minima of a quasiconcave function are extreme points of the function's domain [Boyd and Vandenberghe, 2004], and the extreme points in the quantizer space are LLR-based rules [Tsitsiklis, 1993a], we deduce that there exists an optimal quantizer that is LLR-based.

Lemma 5.7 stated below guarantees quasiconcavity for the case of binary quantizers. To state the result, et $F : [0, 1]^2 \to R$ be given by

$$F(a_0, a_1) = \frac{c_0}{D(a_0, a_1) + d_0} + \frac{c_1}{D(a_1, a_0) + d_1}. \tag{5.26}$$

Lemma 5.7. *For any non-negative constants c_0, c_1, d_0, d_1, the function F defined in* (5.26) *is quasiconcave.*

We provide a proof of this result in the Appendix. An immediate consequence of Lemma 5.7 that LLR-based quantizers exists for the class of randomized quantizers with binary outputs. It turns out that the same statement can also be proved for deterministic quantizers with arbitrary output alphabets:

Theorem 5.8. *Restricting to the class of (blockwise) stationary and deterministic decision rules, then there exists an asymptotically optimal quantizer ϕ that is a likelihood ratio threshold rule.*

We present the full proof of this theorem in the Appendix 5.E. The proof exploits both Lemma 5.6 and Lemma 5.7.

5.6 Discussions

In this chapter, we have considered the problem of sequential decentralized detection. More specifically, focusing on the case of quantization rules with neither memory nor feedback, we have analyzed the (sub)-optimality of stationary quantizer designs. For quantizers with neither local memory nor feedback (Case A in the taxonomy of Veeravalli et al. [Veeravalli *et al.*, 1993]), we have established that stationary designs are not optimal in general. Moreover, we have shown that in the asymptotic setting (i.e., when the cost per sample goes to zero), there is a class of problems for which there exists a range of prior probabilities over which stationary strategies are suboptimal.

There are a number of open questions raised by the analysis in this chapter. First, our analysis has shown only that the best stationary rule from finite sets of deterministic quantizers need not be optimal. Is there a corresponding example with an infinite number of deterministic stationary quantizer designs for which none is optimal? Second, Theorem 5.8 establishes the optimality of likelihood ratio rules for randomized decision rules based on binary outputs. Is the sequential cost coefficient G_ϕ also a quasiconcave function for quantizers other than binary ones? Such quasiconcavity would establish the validity of Theorem 5.8 for the general class of randomized quantizers.

Appendix 5.A Dynamic-programming characterization

In this appendix, we describe how the optimal solution of the sequential decision problem can be characterized recursively using dynamic programming (DP) arguments [Arrow *et al.*, 1949; Wald and Wolfowitz, 1948]. We assume that $X_1, X_2, \ldots$ are independent but not identically distributed conditioned on H. We use subscript n in $f_n^0(x)$ and $f_n^1(x)$ to denote the probability mass (or density) function conditioned on $H = 0$ and $H = 1$, respectively. It has been shown that the sufficient statistic for the DP analysis is the posterior probability $p_n = P(H = 1|X_1, \ldots, X_n)$, which can be updated as by:

$$p_0 = \pi^1; p_{n+1} = \frac{p_n f_{n+1}^1(X_{n+1})}{p_n f_{n+1}^1(X_{n+1}) + (1-p_n) f_{n+1}^0(X_{n+1})}.$$

Finite horizon: First, let us restrict the stopping time N to a finite interval $[0, T]$ for some T. At each time step n, define $J_n^T(p_n)$ to be the minimum expected cost-to-go. At $n = T$, it is easily seen that

$$J_T^T(p_T) = g(p_T),$$

where $g(p) := \min\{p, 1-p\}$. In addition, the optimal decision function γ at time step T, which is a function of p_T, has the following form: $\gamma_T(p_T) = 1$ if $p \geq 1/2$ and 0 otherwise.

For $0 \leq n \leq T-1$, a standard DP argument gives the following backward recursion:

$$J_n^T(p_n) = \min\{g(p_n), c + A_n^T(p_n)\},$$

where

$$A_n^T(p_n) = \mathbb{E}\{J_{n+1}^T(p_{n+1})|X_1, \ldots, X_n\} = \sum_{x_{n+1}} J_{n+1}^T(p_{n+1})(p_n f_{n+1}^1(x_{n+1}) + (1-p_n) f_{n+1}^0(x_{n+1})).$$

The decision whether to stop depends on p_n: If $g(p_n) \leq c + A_n^T(p_n)$, there is no additional benefit of making one more observation, thus we stop. The final decision $\gamma(p_n)$ takes value 1 if $p_n \geq 1/2$ and 0 otherwise. The overall optimal cost function for the sequential test just described is J_0^T.

It is known that the functions J_n^T and A_n^T are concave and continuous in p that take value 0 when $p = 0$ and $p = 1$ [Arrow *et al.*, 1949]. Furthermore, the optimal region for which we decide $\hat{H} = 1$ is a convex set that contains $p_n = 1$, and the optimal region for which we decide $\hat{H} = 0$ is a convex set that contains $p_n = 0$. Hence, we stop as soon as either $p_n \leq p_n^+$ or $p_n \geq p_n^-$ for some $0 < p_n^+ < p_n^-$. This corresponds to a likelihood ratio test: For some threshold $a_n < 0 < b_n$, let:

$$N = \inf\{n \geq 1 \,|L_n := \sum_{i=1}^{n} \log \frac{f_i^1(X_i)}{f_i^0(X_i)} \leq a_n \text{ or } L_n \geq b_n\}. \tag{5.27}$$

Set $\gamma(L_N) = 1$ if $L_n \geq b_n$ and 0 otherwise.

Infinite horizon: The original problem is solved by relaxing the restriction that the stopping time is bounded by a constant T. Letting $T \to \infty$, for each n, the optimal expected cost-to-go $J_n^T(p_n)$ decreases and tends to a limit denoted by $J(p_n) := \lim_{T\to\infty} J_n(p_n)$.

Note that since $X_1, X_2, \ldots$ are i.i.d. conditionally on a hypothesis H, the two functions $J_n^T(p)$ and $J_{n+1}^{T+1}(p)$ are equivalent. As a result, by lettting $T \to \infty$, $J_n(p)$ independent of n and can be denoted as $J(p)$. A similar time-shift argument also yields that the cost function $\lim_{T\to\infty} A_n^T(p)$ is independent of n. We denote this limit by $A(p)$. It is then easily seen that the optimal stopping time N is a likelihood ratio test where the thresholds a_n and b_n are independent of n. We use a to denote the former and b the latter. The functions $J(p)$ and $A(p)$ are related by the following Bellman equation [Bertsekas, 1995a]:

$$J(p) = \min\{g(p), c + A(p)\} \text{ for all } p \in [0,1]. \tag{5.28}$$

The cost of the optimal sequential test of the problem is $J(\pi_1)$.

Appendix 5.B Proof of Lemma 5.5

By renormalizing, we can assume w.l.o.g. that $a_1 + b_1 + c_1 = a_0 + b_0 + c_0 = 1$. Also w.l.o.g, assume that $b_1 \geq b_0$. Thus, $c_1 > c_0$ and $a_1 < a_0$. Replacing $c_1 = 1 - a_1 - b_1$ and $c_0 = 1 - a_0 - b_0$, the inequality $c_1/c_0 > b_1/b_0$ is equivalent to $a_1 < a_0 b_1/b_0 - (b_1 - b_0)/b_0$.

We fix values of b, and consider varying $a \in A$, where A denotes the domain for (a_0, a_1) governed by the following equality and inequality constraints: $0 < a_1 < 1 - b_1;\ 0 < a_0 < 1 - b_0;\ a_1 < a_0$ and

$$a_1 < a_0 b_1/b_0 - (b_1 - b_0)/b_0. \tag{5.29}$$

Note that the third constraint is redundant due to the other three constraints. In particular, constraint (5.29) corresponds to a line passing through $((b_1 - b_0)/b_1, 0)$ and $(1 - b_0, 1 - b_1)$ in the (a_0, a_1) coordinates. As a result, A is the interior of the triangle defined by this line and two other lines given by $a_1 = 0$ and $a_0 = 1 - b_0$ (see Figure 5.B).

It is straightforward to check that both $\tilde{D}^0(a, 1-a)$ and $\tilde{D}^1(a, 1-a)$ are convex functions with respect to (a_0, a_1). In addition, the derivatives with respect to a_1 are $\frac{a_1 - a_0}{a_1(1-a_1)} < 0$ and $\log \frac{a_1(1-a_0)}{a_0(1-a_1)} < 0$, respectively. Hence, both functions can be (strictly) bounded from below by increasing a_1 while keeping a_0 unchanged, i.e., by replacing a_1 by a_1' so that

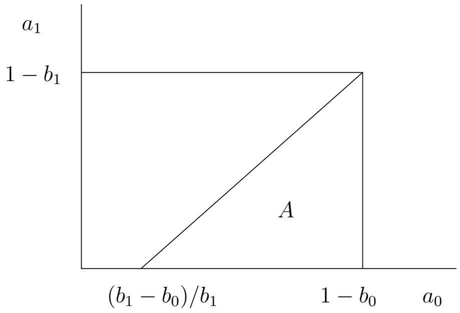


Figure 5.1: Illustration of the domain A.

(a_0, a_1') lies on the line given by (5.29), which is equivalent to the constraint $c_1/c_0 = b_1/b_0$. Let $c_1' = 1 - b_1 - a_1'$, then $c_1'/c_0 = b_1/b_0$.

We have

$$\begin{aligned}
\tilde{D}^0(a, b + c) &\overset{(a)}{>} a_1' \log \frac{a_1'}{a_0} + (b_1 + c_1') \log \frac{b_1 + c_1'}{b_0 + c_0} && (5.30a)\\
&\overset{(b)}{=} a_1' \log \frac{a_1'}{a_0} + c_1' \log \frac{c_1'}{c_0} + b_1 \log \frac{b_1}{b_0} && (5.30b)\\
&\overset{(c)}{\geq} (a_1' + c_1') \log \frac{a_1' + c_1'}{a_0 + c_0} + b_1 \log \frac{b_1}{b_0} && (5.30c)\\
&= \tilde{D}^0(a + c, b), && (5.30d)
\end{aligned}$$

where inequality (c) follows from an application of the log-sum inequality [Cover and Thomas, 1991]. A similar conclusion holds for $\tilde{D}^1(a, b + c)$ as well.

Appendix 5.C Proof of Lemma 5.6

Suppose the opposite is true, that there exist two sets S_1, S_2 with positive $\mathbb{P}_0$-measure such that $\phi(X) = u_2$ for any $X \in S_1 \cup S_2$, and

$$\frac{f^1(S_1)}{f^0(S_1)} < \frac{f^1(u_1)}{f^0(u_1)} < \frac{f^1(S_2)}{f^0(S_2)}. \tag{5.31}$$

By reassigning S_1 or S_2 to the quantile u_1, we are guaranteed to have a new quantizer ϕ' such that $\mu^0_{\phi'} > \mu^0_{\phi^*}$ and $\mu^1_{\phi'} > \mu^1_{\phi^*}$, thanks to Lemma 5.5. As a result, ϕ' has a smaller sequential cost $J^*_{\phi'}$, which is a contradiction.

Appendix 5.D Proof of Lemma 5.7

The proof of this lemma is conceptually straightforward, but the algebra is involved. To simplify the notation, we replace a_0 by x, a_1 by y, the function $D(a_0, a_1)$ by $f(x, y)$, and the function $D(a_1, a_0)$ by $g(x, y)$. Finally, we assume that $d_0 = d_1 = 0$; the proof will reveal that this case is sufficient to establish the more general result with arbitrary non-negative scalars d_0 and d_1.

We have $f(x, y) = x \log(x/y) + (1 - x) \log(1 - x/1 - y)$ and $g(x, y) = y \log(y/x) + (1 - y) \log(1 - y/1 - x)$. Note that both f and g are convex functions and are non-negative in their domains, and moreover that we have $F(x, y) = c_0/f(x, y) + c_1/g(x, y)$. In order to establish the quasiconcavity of F, it suffices to show that for any (x, y) in the domain of F, whenever vector $h = [h_0\ h_1] \in \mathbb{R}^2$ such that $h^T \nabla F(x, y) = 0$, there holds

$$h^T \nabla^2 F(x, y)\ h \leq 0. \tag{5.32}$$

Here we adopt the standard notation of ∇F for the gradient vector of F, and $\nabla^2 F$ for its Hessian matrix. We also use F_x to denote the partial derivative with respect to variable x, F_{xy} to denote the partial derivative with respect to x and y, and so on.

We have $\nabla F = -\frac{c_0 \nabla f}{f^2} - \frac{c_1 \nabla g}{g^2}$. Thus, it suffices to prove relation (5.32) for vectors of the form

$$h = \left[\left(-\tfrac{c_0 f_y}{f^2} - \tfrac{c_1 g_y}{g^2} \right) \quad \left(\tfrac{c_0 f_x}{f^2} + \tfrac{c_1 g_x}{g^2} \right) \right]^T.$$

It is convenient to write $h = c_0 v_0 + c_1 v_1$, where $v_0 = [-f_y/f^2 \quad f_x/f^2]^T$ and $v_1 = [-g_y/g^2 \quad g_x/g^2]^T$.

The Hessian matrix $\nabla^2 F$ can be written as $\nabla^2 F = c_0 H_0 + c_0 H_1$, where

$$H_0 = -\frac{1}{f^3} \begin{bmatrix} f_{xx} f - 2f_x^2 & f_{xy} f - 2 f_x f_y \\ f_{xy} f - 2 f_x f_y & f_{yy} f - 2 f_y^2 \end{bmatrix},$$

and

$$H_1 = -\frac{1}{g^3} \begin{bmatrix} g_{xx} g - 2g_x^2 & g_{xy} g - 2 g_x g_y \\ g_{xy} g - 2 g_x g_y & g_{yy} g - 2 g_y^2 \end{bmatrix}.$$

Now observe that

$$h^T \nabla^2 F h \quad = \quad (c_0 v_0 + c_1 v_1)^T (c_0 H_0 + c_1 H_1)(c_0 v_0 + c_1 v_1),$$

which can be simplified to

$$h^T \nabla^2 F h = c_0^3 v_0^T H_0 v_0 + c_1^3 v_1^T H_1 v_1 + c_0^2 c_1 (2 v_0^T H_0 v_1 + v_0^T H_1 v_0) + c_0 c_1^2 (2 v_0^T H_1 v_1 + v_1^T H_0 v_1).$$

This function is a polynomial in c_0 and c_1, which are restricted to be non-negative scalars.

Therefore, it suffices to prove that all the coefficients of this polynomial (with respect to c_0 and c_1) are non-positive. In particular, we shall show that

(i) $v_0^T H_0 v_0 \leq 0$, and

(ii) $2v_0^T H_0 v_1 + v_0^T H_1 v_0 \leq 0$.

The non-positivity of the other two coefficients follows from entirely analogous arguments.

First, some straightforward algebra shows that inequality (i) is equivalent to the relation

$$f_{xx} f_y^2 + f_{yy} f_x^2 \geq 2 f_x f_y f_{xy}.$$

But note that f is a convex function, so $f_{xx} f_{yy} \geq f_{xy}^2$. Hence, we have

$$f_{xx} f_y^2 + f_{yy} f_x^2 \overset{(a)}{\geq} 2\sqrt{f_{xx} f_{yy}} f_x f_y \overset{(b)}{\geq} 2 f_x f_y f_{xy},$$

thereby proving (i). (In this argument, inequality (a) follows from the fact that $a^2 + b^2 \geq 2ab$, whereas inequality (b) follows from the convexity of f.)

Regarding (ii), some further algebra reduces it to the inequality

$$G_1 + G_2 - G_3 \geq 0, \tag{5.33}$$

where

$$\begin{aligned} G_1 &= 2(f_y g_y f_{xx} + f_x g_x f_{yy} - (f_y g_x + f_x g_y) f_{xy}), \\ G_2 &= f_y^2 g_{xx} + f_x^2 g_{yy} - 2 f_x f_y g_{xy}, \\ G_3 &= \frac{2}{g}(f_y g_x - f_x g_y)^2. \end{aligned}$$

At this point in the proof, we need to exploit specific information about the functions f and g, which are defined in terms of KL divergences. To simplify notation, we let $u = x/y$ and $v = (1-x)/(1-y)$. Computing derivatives, we have

$$\begin{aligned} f_x(x,y) &= \log(x/y) - \log((1-x)/(1-y)) = \log(u/v), \\ f_y(x,y) &= (1-x)/(1-y) - x/y = v - u, \\ g_x(x,y) &= (1-y)/(1-x) - y/x = 1/v - 1/u, \\ g_y(x,y) &= \log(y/x) - \log((1-y)/(1-x)) = \log(v/u), \end{aligned}$$

$$\nabla^2 f(x,y) = \begin{bmatrix} \frac{1}{x(1-x)} & -\frac{1}{y(1-y)} \\ -\frac{1}{x(1-x)} & \frac{1-x}{(1-y)^2} + \frac{x}{y^2} \end{bmatrix}, \text{ and } \nabla^2 g(x,y) = \begin{bmatrix} \frac{1-y}{(1-x)^2} + \frac{y}{x^2} & -\frac{1}{x(1-x)} \\ -\frac{1}{x(1-x)} & \frac{1}{y(1-y)} \end{bmatrix}.$$

Noting that $f_x = -g_y;\ g_{xy} = -f_{xx};\ f_{xy} = -g_{yy}$, we see that equation (5.33) is equivalent to

$$2(f_x g_x f_{yy} + f_y g_x g_{yy}) - f_x^2 g_{yy} + f_y^2 g_{xx} \ \geq\ \frac{2}{g}(f_y g_x - f_x g_y)^2. \tag{5.34}$$

To simplify the algebra further, we shall make use of the inequality $(\log t^2)^2 \leq (t - 1/t)^2$, which is valid for any t. This implies that

$$f_y g_x = (v-u)(1/v - 1/u) \leq f_x g_y = -(\log(u/v))^2 = -f_x^2 = -g_y^2 \leq 0.$$

Thus, $-f_x^2 g_{yy} \geq f_y g_x g_{yy}$, and $\frac{2}{g}(f_y g_x - f_x g_y)^2 \leq \frac{2}{g} f_y g_x (f_y g_x - f_x g_y)$. As a result, (5.34) would follow if we can show that

$$2(f_x g_x f_{yy} + f_y g_x g_{yy}) + f_y g_x g_{yy} + f_y^2 g_{xx} \geq \frac{2}{g} f_y g_x (f_y g_x - f_x g_y).$$

For all $x \neq y$, we may divide both sides by $-f_y(x,y) g_x(x,y) > 0$. Consequently, it suffices to show that:

$$-2 f_x f_{yy}/f_y - f_y g_{xx}/g_x - 3 g_{yy} \geq \frac{2}{g}(f_x g_y - g_x f_y),$$

or, equivalently,

$$2\log(u/v)\left(\frac{v}{u-1} + \frac{u}{1-v}\right) + \left(\frac{u}{1-x} + \frac{v}{x}\right) - \frac{3}{y(1-y)} \geq \frac{2}{g}\left(\frac{(u-v)^2}{uv} - (\log\frac{u}{v})^2\right),$$

or, equivalently,

$$2\log(u/v)\frac{(u-v)(u+v-1)}{(u-1)(1-v)} + \frac{(u-v)^2(u+v-4uv)}{uv(u-1)(1-v)} \geq \frac{2}{g}\left(\frac{(u-v)^2}{uv} - (\log\frac{u}{v})^2\right). \tag{5.35}$$

Due to the symmetry, it suffices to prove (5.35) for $x < y$. In particular, we shall use the following inequality for logarithm mean [Mitrinović *et al.*, 1993], which holds for $u \neq v$:

$$\frac{3}{2\sqrt{uv} + (u+v)/2} < \frac{\log u - \log v}{u - v} < \frac{1}{(uv(u+v)/2)^{1/3}}.$$

We shall replace $\frac{\log(u/v)}{u-v}$ in (5.35) by appropriate upper and lower bounds. In addition, we shall also bound $g(x,y)$ from below, using the following argument. When $x < y$, we have

$u < 1 < v$, and

$$\begin{aligned} g(x,y) &= y \log \frac{y}{x} + (1-y) \log \frac{1-y}{1-x} \\ &> \frac{3y(y-x)}{2\sqrt{xy} + (x+y)/2} + \frac{(1-y)(x-y)}{[(1-x)(1-y)(1-(x+y)/2)]^{1/3}} \\ &= \frac{3(1-v)(1-u)}{(u-v)(2\sqrt{u} + \frac{u+1}{2})} + \frac{(u-1)(1-v)}{(u-v)(v(v+1)/2)^{1/3}} > 0. \end{aligned}$$

Let us denote this lower bound by $q(u,v)$.

Having got rid of the logarithm terms, (5.35) will hold if we can prove the following:

$$\frac{6(u-v)^2(u+v-1)}{(2\sqrt{uv} + (u+v)/2)(u-1)(1-v)} + \frac{(u-v)^2(u+v-4uv)}{uv(u-1)(1-v)} \geq \frac{2}{q(u,v)} \left(\frac{(u-v)^2}{uv} - \frac{9(u-v)^2}{(2\sqrt{uv} + (u+v)/2)^2} \right), \tag{5.36}$$

or equivalently,

$$\left(\frac{6(u+v-1)}{(2\sqrt{uv} + (u+v)/2)} + \frac{(u+v-4uv)}{uv} \right) \left(\frac{3}{(v-u)(2\sqrt{u} + \frac{u+1}{2})} - \frac{1}{(v-u)(v(v+1)/2)^{1/3}} \right) \geq 2 \left(\frac{1}{uv} - \frac{9}{(2\sqrt{uv} + (u+v)/2)^2} \right), \tag{5.37}$$

which is equivalent to

$$\frac{(u+v-2\sqrt{uv})((u+v)/2 + 3\sqrt{uv} + 4uv)}{(2\sqrt{uv} + (u+v)/2)uv} \frac{3(v(v+1)/2)^{1/3} - (2\sqrt{u} + (u+1)/2)}{(v-u)(2\sqrt{u} + (u+1)/2)(v(v+1)/2)^{1/3}} \geq \frac{(u+v-2\sqrt{uv})((u+v)/2 + 5\sqrt{uv})}{uv(2\sqrt{uv} + (u+v)/2)^2} \tag{5.38}$$

and also equivalent to

$$\begin{aligned} &((u+v)/2 + 2\sqrt{uv})((u+v)/2 + 3\sqrt{uv} + 4uv)[3(v(v+1)/2)^{1/3} - (2\sqrt{u} + (u+1)/2)] \\ &\geq (2\sqrt{u} + (u+1)/2)(v(v+1)/2)^{1/3}((u+v)/2 + 5\sqrt{uv})(v-u) \end{aligned} \tag{5.39}$$

It can be checked by tedious but straightforward calculus that inequality (5.39) holds for any $u \leq 1 \leq v$, and equality holds when $u = 1 = v$, i.e., $x = y$.

Appendix 5.E Proof of Theorem 5.8

Suppose that ϕ is not a likelihood ratio rule. Then there exist positive $\mathbb{P}_0$-probability disjoint sets S_1, S_2, S_3 such that for any $X_1 \in S_1, X_2 \in S_2, X_3 \in S_3$,

$$\phi(X_1) = \phi(X_3) = u_1 \tag{5.40a}$$

$$\phi(X_2) = u_2 \neq u_1 \tag{5.40b}$$

$$\frac{f^1(X_1)}{f^0(X_1)} < \frac{f^1(X_2)}{f^0(X_2)} < \frac{f^1(X_3)}{f^0(X_3)}. \tag{5.40c}$$

Define the probability of the quantiles as:

$$\begin{aligned} f^0(u_1) &:= \mathbb{P}_0(\phi(X) = u_1), \quad \text{and} \quad f^0(u_2) := \mathbb{P}_0(\phi(X) = u_2), \\ f^1(u_1) &:= \mathbb{P}_1(\phi(X) = u_1), \quad \text{and} \quad f^1(u_2) := \mathbb{P}_1(\phi(X) = u_2). \end{aligned}$$

Similarly, for the sets S_1, S_2 and S_3, we define

$$\begin{aligned} a_0 = f^0(S_1), \quad b_0 = f^0(S_2) \quad \text{and} \quad c_0 = f^0(S_3), \\ a_1 = f^1(S_1), \quad b_1 = f^1(S_2), \quad \text{and} \quad c_1 = f^1(S_3). \end{aligned}$$

Finally, let p_0, p_1, q_0 and q_1 denote the probability measures of the "residuals":

$$\begin{aligned} p_0 &= f^0(u_2) - b_0, \qquad p_1 = f^1(u_2) - b_1, \\ q_0 &= f^0(u_1) - a_0 - c_0, \qquad q_1 = f^1(u_1) - a_1 - c_1. \end{aligned}$$

Note that we have $\frac{a_1}{a_0} < \frac{b_1}{b_0} < \frac{c_1}{c_0}$. In addition, the sets S_1 and S_3 were chosen so that $\frac{a_1}{a_0} \leq \frac{q_1}{q_0} \leq \frac{c_1}{c_0}$. From Lemma 5.6, there holds $\frac{p_1+b_1}{p_0+b_0} = \frac{f^1(u_2)}{f^0(u_2)} \notin \left(\frac{a_1}{a_0}, \frac{c_1}{c_0}\right)$. We may assume without loss of generality that $\frac{p_1+b_1}{p_0+b_0} \leq \frac{a_1}{a_0}$. Then, $\frac{p_1+b_1}{p_0+b_0} < \frac{b_1}{b_0}$, so $\frac{p_1}{p_0} < \frac{p_1+b_1}{p_0+b_0}$. Overall, we are guaranteed to have the ordering

$$\frac{p_1}{p_0} < \frac{p_1+b_1}{p_0+b_0} \leq \frac{a_1}{a_0} < \frac{b_1}{b_0} < \frac{c_1}{c_0}. \tag{5.41}$$

Our strategy will be to modify the quantizer ϕ only for those X for which $\phi(X)$ takes the values u_1 or u_2, such that the resulting quantizer is defined by a LLR-based threshold, and has a smaller (or equal) value of the corresponding cost J^*_ϕ. For simplicity in notation, we use $\mathcal{A}$ to denote the set with measures under $\mathbb{P}_0$ and $\mathbb{P}_1$ equal to a_0 and a_1; the sets $\mathcal{B}$, $\mathcal{C}$, $\mathcal{P}$ and $\mathcal{Q}$ are defined in an analogous manner. We begin by observing that we have either $\frac{a_1}{a_0} \leq \frac{q_1+a_1}{q_0+a_0} < \frac{b_1}{b_0}$ or $\frac{b_1}{b_0} < \frac{q_1+c_1}{q_0+c_0} \leq \frac{c_1}{c_0}$. Thus, in our subsequent manipulation of sets, we always bundle $\mathcal{Q}$ with either $\mathcal{A}$ or $\mathcal{C}$ accordingly without changing the ordering of the probability ratios. Without loss of generality, then, we may disregard the corresponding residual set corresponding to $\mathcal{Q}$ in the analysis to follow.

In the remainder of the proof, we shall show that either one of the following two modifications of the quantizer ϕ will improve (decrease) the sequential cost J^*_ϕ:

(i) Assign $\mathcal{A}, \mathcal{B}$ and $\mathcal{C}$ to the same quantization level u_1, and leave $\mathcal{P}$ to the level u_2, or

(ii) Assign $\mathcal{P}$, $\mathcal{A}$ and $\mathcal{B}$ to the same level u_2, and leave c to the level u_1.

It is clear that this modified quantizer design respects the likelihood ratio rule for the quantization indices u_1 and u_2. By repeated application of this modification for every such pair, we are guaranteed to arrive at a likelihood ratio quantizer that is optimal, thereby completing the proof.

Let a'_0, b'_0, c'_0, p'_0 be normalized versions of a_0, b_0, c_0, p_0, respectively (i.e., $a'_0 = a_0/(p_0 + a_0 + b_0 + c_0)$, and so on). Similarly, let a'_1, b'_1, c'_1, p'_1 be normalized versions of a_1, b_1, c_1, p_1, respectively. With this notation, we have the relations

$$\begin{aligned}
\mu^0_\phi &= \sum_{u \neq u_1, u_2} f^0(u) \log \frac{f^0(u)}{f^1(u)} + (p_0 + b_0) \log \frac{p_0 + b_0}{p_1 + b_1} + (a_0 + c_0) \log \frac{a_0 + c_0}{a_1 + c_1} \\
&= A_0 + (f^0(u_1) + f^0(u_2)) \left((p'_0 + b'_0) \log \frac{p'_0 + b'_0}{p'_1 + b'_1} + (a'_0 + c'_0) \log \frac{a'_0 + c'_0}{a'_1 + c'_1} \right) \\
&= A_0 + (f^0(u_1) + f^0(u_2)) \tilde{D}^0(p' + b', a' + c'), \\
\mu^1_\phi &= \sum_{u \neq u_1, u_2} f^1(u) \log \frac{f^1(u)}{f^0(u)} + (p_1 + b_1) \log \frac{p_1 + b_1}{p_0 + b_0} + (a_1 + c_1) \log \frac{a_1 + c_1}{a_0 + c_0} \\
&= A_1 + (f^1(u_1) + f^1(u_2)) \tilde{D}^1(p' + b', a' + c'),
\end{aligned}$$

where we define

$$\begin{aligned}
A_0 &:= \sum_{u \neq u_1, u_2} f^0(u) \log \frac{f^0(u)}{f^1(u)} + (f^0(u_1) + f^0(u_2)) \log \frac{f^0(u_1) + f^0(u_2)}{f^1(u_1) + f^1(u_2)} \geq 0, \\
A_1 &:= \sum_{u \neq u_1, u_2} f^1(u) \log \frac{f^1(u)}{f^0(u)} + (f^1(u_1) + f^1(u_2)) \log \frac{f^1(u_1) + f^1(u_2)}{f^0(u_1) + f^0(u_2)} \geq 0
\end{aligned}$$

due to the non-negativity of the KL divergences.

Note that from (5.41) we have

$$\frac{p'_1}{p'_0} < \frac{p'_1 + b'_1}{p'_0 + b'_0} \leq \frac{a'_1}{a'_0} < \frac{b'_1}{b'_0} < \frac{c'_1}{c'_0},$$

in addition to the normalization constraints that $p'_0 + a'_0 + b'_0 + c'_0 = p'_1 + a'_1 + b'_1 + c'_1 = 1$. It follows that $\frac{p'_1+b'_1}{p'_0+b'_0} < \frac{p'_1+a'_1+b'_1+c'_1}{p'_0+a'_0+b'_0+c'_0} = 1$.

Let us consider varying the values of a'_1, b'_1, while fixing all other variables and ensuring that all the above constraints hold. Then, $a'_1 + b'_1$ is constant, and both $\tilde{D}^0(p' + b', a' + c')$ and $\tilde{D}^1(p' + b', a' + c')$ increase as b_1 decreases and a_1 increases. In other words, if we define $a''_0 = a'_0$, $b''_0 = b'_0$ and a''_1 and b''_1 such that

$$\frac{a''_1}{a'_0} = \frac{b''_1}{b'_0} = \frac{1 - p'_1 - c'_1}{1 - p'_0 - c'_0},$$

then we have

$$\tilde{D}^0(p'+b', a'+c') \leq \tilde{D}^0(p'+b'', a''+c') \text{ and } \tilde{D}^1(p'+b', a'+c') \leq \tilde{D}^1(p'+b'', a''+c'). \quad (5.42)$$

Now note that vector (b''_0, b''_1) in $\mathbb{R}^2$ is a convex combination of $(0, 0)$ and $(a''_0 + b''_0, a''_1 + b''_1)$. It follows that $(p'_0 + b''_0, p'_1 + b''_1)$ is a convex combination of (p'_0, p'_1) and $(p'_0 + a''_0 + b''_0, p'_1 + a''_1 + b''_1) = (p'_0 + a'_0 + b'_0, p'_1 + a'_1 + b'_1)$.

By (5.42) we have:

$$\begin{aligned}
G_\phi &= \frac{\pi^0}{\mu^0_\phi} + \frac{\pi^1}{\mu^1_\phi} \\
&= \frac{\pi^0}{A_0 + (f^0(u_1) + f^0(u_2))\tilde{D}^0(p' + b', a' + c')} + \frac{\pi^1}{A_1 + (f^1(u_1) + f^1(u_2))\tilde{D}^1(p' + b', a' + c')} \\
&\geq \frac{\pi^0}{A_0 + (f^0(u_1) + f^0(u_2))\tilde{D}^0(p' + b'', a'' + c')} + \frac{\pi^1}{A_1 + (f^1(u_1) + f^1(u_2))\tilde{D}^1(p' + b'', a'' + c')} \\
&= \frac{\pi^0}{A_0 + (f^0(u_1) + f^0(u_2))D(p'_0 + b''_0, p'_1 + b''_1)} + \frac{\pi^1}{A_1 + (f^1(u_1) + f^1(u_2))D(p'_1 + b''_1, p'_0 + b''_0)}.
\end{aligned}$$

Now, by the quasiconcavity result in Lemma 5.7,

$$\begin{aligned}
G_\phi \geq \min\Bigg\{ & \frac{\pi^0}{A_0 + (f^0(u_1) + f^0(u_2))D(p'_0, p'_1)} + \frac{\pi^1}{A_1 + (f^1(u_1) + f^1(u_2))D(p'_1, p'_0)}, \\
& \frac{\pi^0}{A_0 + (f^0(u_1) + f^0(u_2))D(p'_0 + a'_0 + b'_0, p'_1 + a'_1 + b'_1)} + \\
& \frac{\pi^1}{A_1 + (f^1(u_1) + f^1(u_2))D(p'_1 + a'_1 + b'_1, p'_0 + a'_0 + b'_0)} \Bigg\}.
\end{aligned}$$

But the two arguments of the minimum in the final equation are the sequential cost coefficient corresponding to the two possible modifications of ϕ. Hence, the proof is complete.

Chapter 6

Estimation of divergence functionals and the likelihood ratio

We present a novel M-estimation method for the divergence functionals and the density ratios of two probability distributions. Our method is based on a non-asymptotic variational characterization of f-divergences, which turns the problem of estimating divergences to a convex risk optimization. We present an analysis of consistency and convergence for our estimator. Given conditions only on the ratios of densities, we show that our estimators can achieve optimal minimax rates for the likelihood ratio in some regime. Finally, we present an efficient optimization algorithm for our estimator and demonstrate its convergence behavior and practical viability by simulations.[1]

6.1 Introduction

Given empirical samples from two (multivariate) probability distributions $\mathbb{P}$ and $\mathbb{Q}$, we are interested in estimating a divergence functional between $\mathbb{P}$ and $\mathbb{Q}$. We consider in particular Kullback-Leibler divergence, and then all divergences in the class of Ali-Silvey distance, also known as f-divergences [Ali and Silvey, 1966; Csiszaŕ, 1967]. This family of divergence, which shall be defined formally in the sequel, is of the form $D_\phi(\mathbb{P},\mathbb{Q}) = \int \phi(d\mathbb{Q}/d\mathbb{P})d\mathbb{P}$, where ϕ is a convex function of the likelihood ratio $d\mathbb{Q}/d\mathbb{P}$.

The divergences have a fundamental role as an objective to optimize in various data analysis and learning tasks. Divergences are used as a measure to distinguish between two hypotheses. In experiment design for binary hypothesis testing and classification applications, the experiments are designed so that the divergence between two underlying hypothesis distributions are maximized. Problems of this type can be seen in signal selection [Kailath, 1967], decentralized detection [Nguyen *et al.*, 2005c] (see Chapter 4). An

[1] Part of this chapter has been published in [Nguyen *et al.*, 2007].

important quantity in information theory, the Shanon mutual information, can be viewed as a KL divergence. Mutual information is often used as a measure of independence to be minimized such as in the problem of independent component analysis [Hyvarinen *et al.*, 2001]. If the divergences are to be used as objective functional in such tasks, one has to be able to estimate them efficiently from empirical data.

There are two ways in which divergences can be characterized. Taking the KL divergence in particular, in the Neyman-Pearson setting of a binary hypothesis testing problem, the KL divergence emerges as the correct asympotic rate of the probability error, a result known as Stein's lemma. On the other hand, a non-asymptotic view of KL divergence emerges through Fano's lemma, which provide a lower bound for the error probability for decoding/hypothesis test in terms of KL divergence (cf. [Cover and Thomas, 1991]). Note that there are a multitude of results in the same vein for other divergences, due to statisticians such as Cramér, Chernoff, Le Cam, and others [van der Vaart, 1998].

In this chapter, we shall present an estimation method that is motivated by a non-asymptotic characterization of f-divergence that was explicated in Theorem 4.8. Roughly speaking, this theorem states that that there is a correspondence between the family of f-divergences and a family of losses such that the minimum risk is equal to the negative of the divergence. In other words, any negative f-divergence can serve as a lower bound of a risk minimization problem. While this result deals only with binary hypotheses (as opposed to Fano's lemma) it goes significantly further than Fano's lemma in that it covers a whole class of losses and divergences. This correspondence provides what we shall call a *variational characterization* of divergence: One can write a divergence $D_\phi(\mathbb{P}, \mathbb{Q})$ as the maximum of an Bayes decision problem involving two hypotheses $\mathbb{P}$ and $\mathbb{Q}$. This characterization is stated in Lemma 6.1. As a result, one can estimate $D_\phi(\mathbb{P}, \mathbb{Q})$ by solving the Bayes decision (maximization) problem. Not surprisingly, we show how the problem of estimating f-divergence is intrinsically linked to that of estimating the likelihood ratio $g_0 = d\mathbb{P}/d\mathbb{Q}$. As a result we obtain an M estimator for the likelihood ratio, from which one can obtain an estimation of the divergences by a plug-in procedure.

Our contributions are three-fold:

- We propose a novel M-estimator for the likelihood ratio and the family of f divergences based on a variational characterization of f-divergence as explained above. Our estimation procedure is inherently nonparametric. We make no strong assumption on the form of the densities for $\mathbb{P}$ and $\mathbb{Q}$.

- We provide a consistency and convergence analysis for our estimators. For the analysis, we make assumptions on the boundedness of the *density ratio*, which can be relaxed in some cases. The maximization procedure is cast over a whole function class $\mathcal{G}$ of density ratio, thus our tool is based on results from the theory of empirical processes. Our method of proof is based on the analysis of M-estimation for nonparametric density estimation [van de Geer, 1999; van der Vaart and Wellner, 1996].

The key issue essentially hinges on the modulus of continuity of the suprema of two empirical processes (defined on $\mathbb{P}$ and $\mathbb{Q}$ measures) with respect to a metric defined on the class $\mathcal{G}$. This metric turns out to be a surrogate lower bound of a Bregman divergence defined on a pair of density ratios. Our choice of metrics include the Hellinger distance and L_2 norm.

- We provide an efficient algorithm for our estimation procedure. In particular, we approximate $\mathcal{G}$ by a reproducing kernel Hilbert space given a positive definite kernel function $K(u, v)$ [Saitoh, 1988]. We control the size of the function class $\mathcal{G}$ by introducing a penalty term for the RKHS norm of $g \in \mathcal{G}$. The estimation problem is converted into a convex optimization problem, which is then turned into a dual form involving only the Gram matrix $K(u_i, v_j)$, where u_i and v_j are drawn from either $\mathbb{P}$ or $\mathbb{Q}$. This kernel-based method has been widely used in statistical learning tasks [Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004]. Finally, we demonstrate our estimator in a large number of simulation runs on a number of pairs of probability distributions.

Several interesting properties of this estimator is worth highlighting.

- First, in terms of convergence rates. When the likehood ratio g_0 lies in a function class $\mathcal{G}$ of smoothness α with $\alpha > d/2$, where d is the number of dimensions of the data, our estimation of the likelihood ratio achieves the optimal minimax rate $n^{-\alpha/(2\alpha+d)}$ according to the Hellinger metric, and divergence estimator achieves the same rate. It remains an open question what is the optimal minimax rate for the divergence estimation.

- An obvious alternative approach to our problem would be to separately estimate the densities for $\mathbb{P}$ and $\mathbb{Q}$ and then use an appropriate plug-in estimator for the divergences. As we shall see in our analysis, estimating directly the density ratio has several distinct advantages. Firstly, from computational viewpoint, it is more efficient to perform one estimation procedure instead of two. Comparing to an M-estimator for density estimation (e.g, [Silverman, 1982]), there is no need to enforce the constraint that the estimated function is a valid density. Secondly, from a statistical viewpoint, we achieve the same estimation efficiency without making individual assumptions on each density. Assumptions are made only on the density ratios.

- Finally, if we use small function classes $\mathcal{G}$ that might not include the true likelihood ratio, our estimator of the divergence has the property of being a lower bound of the true divergence. This might provide additional useful information for a task in hand.

Related work. The variational representation of divergences has been derived independently and exploited by several authors [Broniatowski and Keziou, 2004; Keziou, 2003;

Nguyen *et al.*, 2005c]. Broniatowski and Keziou [Broniatowski and Keziou, 2004] studied testing and estimation problems based on dual representations of f-divergences, but working in a parametric setting as opposed to the nonparametric framework considered here. Nguyen et al. [Nguyen *et al.*, 2005c] established a one-to-one correspondence between the family of f-divergences and the family of surrogate loss functions [Bartlett *et al.*, 2006], through which the (optimum) "surrogate risk" is equal to the negative of an associated f-divergence. Another link is to the problem of estimating integral functionals of a single density, with the Shannon entropy being a well-known example, which has been studied extensively dating back to early work [Ibragimov and Khasminskii, 1978; Levit, 1978] as well as the more recent work [Bickel and Ritov, 1988; Birgé and Massart, 1995; Laurent, 1996]. See also [Gyorfi and van der Meulen, 1987; Joe, 1989; Hall and Morton, 1993] for the problem of (Shannon) entropy functional estimation. In another branch of related work, Wang et al. [Wang *et al.*, 2005] proposed an algorithm for estimating the KL divergence for continuous distributions, which exploits histogram-based estimation of the likelihood ratio by building data-dependent partitions of equivalent (empirical) $\mathbb{Q}$-measure. The estimator was empirically shown to outperform direct plug-in methods, but no theoretical results on its convergence rate were provided.

The chapter is organized as follows. In Sec. 6.2 we describe the variational characterization of f-divergence in general and KL divergence in particular, followed by an M-estimator for the KL divergence and the likelihood ratio. Sec. 6.3 and Sec. 6.4 are devoted to the analysis of consistency and convergence rates of our estimators. In Sec. 6.5 we describe our estimation method and the analysis in a more general light, encompassing virtually all f-divergences. We also consider a general estimation framework based on the delta method, assuming the ϕ is a differentiable function. Sec. 6.7 describe the optimization in detail. In Sec. 6.9 we present our simulation results.

6.2 M-estimators for KL divergence and the density ratio

6.2.1 Variational characterization of f-divergence

Let $X_1, \ldots, X_n$ be n i.i.d. random variables according to a distribution $\mathbb{P}$, and $Y_1, \ldots, Y_n$ be n random variables according to a distribution $\mathbb{Q}$. We assume that $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$, and both are absolutely continuous with respect to Lebesgue measure μ with densities p_0 and q_0, respectively, on some compact domain $\mathcal{X} \subset \mathbb{R}^d$. The Kullback-Leibler divergence between $\mathbb{P}$ and $\mathbb{Q}$ is defined as:

$$D_K(\mathbb{P}, \mathbb{Q}) = \int p_0 \log \frac{p_0}{q_0} \, d\mu.$$

The KL divergence is a special case of a broader class of divergences known as Ali-Silvey distance, or f-divergence [Csiszaŕ, 1967; Ali and Silvey, 1966]:

$$D_\phi(\mathbb{P},\mathbb{Q}) = \int p_0 \phi(q_0/p_0)\, d\mu,$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is a convex function. Different choices of ϕ result in many divergences that play important roles in information theory and statistics, including the variational distance, Hellinger distance, KL divergence and so on (see, e.g., [Topsoe, 2000]).

Since ϕ is a convex function, by Legendre-Fenchel convex duality [Rockafellar, 1970] we can write:

$$\phi(u) = \sup_{v\in\mathbb{R}} uv - \phi^*(v),$$

where ϕ^* is the convex conjugate of ϕ. As a result,

$$\begin{aligned} D_\phi(\mathbb{P},\mathbb{Q}) &= \int p_0 \sup_{f\in\mathbb{R}} (fq_0/p_0 - \phi^*(f))\, d\mu \\ &= \sup_f \int f q_0 - \phi^*(f) p_0 \, d\mu \\ &= \sup_f \int f\, d\mathbb{Q} - \phi^*(f)\, d\mathbb{P}, \end{aligned}$$

where the suppremum is taken over all measurable function $f : \mathcal{X} \to \mathbb{R}$, and $\int f \ d\mathbb{P}$ denotes the expectation of f under distribution $\mathbb{P}$. It is simple to see that equality the supremum is attained at function f such that $q_0/p_0 \in \partial\phi^*(f)$ where q_0, p_0 and f are evaluated at any $x \in \mathcal{X}$. By convex duality, this is true if $f \in \partial\phi(q_0/p_0)$ for any $x \in \mathcal{X}$. Thus, we have proved the following lemma:

Lemma 6.1. *Let $\mathcal{F}$ be any function class $\mathcal{X} \to \mathbb{R}$, there holds:*

$$D_\phi(\mathbb{P},\mathbb{Q}) \geq \sup_{f\in\mathcal{F}} \int f\, d\mathbb{Q} - \phi^*(f)\, d\mathbb{P}. \tag{6.1}$$

Furthermore, equality holds whenever $\mathcal{F} \cap \partial\phi(q_0/p_0) \neq \emptyset$.

Remark. There is an interesting connection between Lemma 6.1 and Fano's lower bound in coding theory. Indeed, consider a Bayesian hypothesis testing problem between two distributions $\mathbb{P}$ and $\mathbb{Q}$, which have equal priors (1/2). Let $-f$ be the loss for incorrectly rejecting $\mathbb{Q}$, and $\phi^*(f)$ the loss for incorrectly rejecting $\mathbb{P}$. Then $-D_\phi(\mathbb{P},\mathbb{Q})$ is nothing but the lower bound of the risk:

$$\inf_f \int (-f)\, d\mathbb{Q} + \phi^*(f)\, d\mathbb{P} = -D_\phi(\mathbb{P},\mathbb{Q}).$$

In other words, for each f divergence $D_\phi(\mathbb{P}, \mathbb{Q})$ there exists a binary classification problem with appropriate loss functions whose optimal risk is characterized by the divergence. For a thorough analysis of this correspondence, see Chapter 4. It can be seen that for some (appropriately parametrized) choices of f and ϕ so that both loss functions $(-f)$ and $\phi^*(f)$ correspond to the 0-1 loss function, D_ϕ becomes the variational distance (plus a constant). As a result, one obtain a special case of Fano's lemma for binary classification. This connection extends to multiple hypothesis testing, but we shall not pursue further here.

6.2.2 An M-estimator of density ratio and KL divergence

Returning to the KL divergence, ϕ has the form $\phi(u) = -\log(u)$ for $u > 0$ and $+\infty$ for $u \leq 0$. The convex dual of ϕ is $\phi^*(v) = \sup_u uv - \phi(u) = -1 - \log(-v)$ if $u < 0$ and $+\infty$ otherwise. By Lemma 6.1,

$$D_K(\mathbb{P}, \mathbb{Q}) = \sup_{f<0} \int f \, d\mathbb{Q} - \int -1 - \log(-f) \, d\mathbb{P} = \sup_{g>0} \int \log g \, d\mathbb{P} - \int g d\mathbb{Q} + 1. \quad (6.2)$$

In addition, the supremum is attained at $g = p_0/q_0$. This motivates our estimator of the KL divergence as follows: Let $\mathcal{G}$ be a function class of $\mathcal{X} \to \mathbb{R}_+$, and $\int \, d\mathbb{P}_n$ and $\int \, d\mathbb{Q}_n$ denote the expectation under empirical measures $\mathbb{P}_n$ and $\mathbb{Q}_n$, respectively, then our estimator has the following form:

$$\hat{D}_K = \sup_{g \in \mathcal{G}} \int \log g \, d\mathbb{P}_n - \int g d\mathbb{Q}_n + 1. \quad (6.3)$$

For the implementation, we shall assume that $\mathcal{G}$ is a convex function class. The above estimator can be posed as a convex optimization problem that can be solved efficiently (see Section X). Suppose that the supremum is attained at $\hat{g}_n$. Then $\hat{g}_n$ is an M-estimator of the density ratio $g_0 = p_0/q_0$.

For the KL divergence estimation, there are two sources of error, namely, approximation error $\mathcal{E}_0(\mathcal{G})$ and estimation error $\mathcal{E}_1(\mathcal{G})$:

$$\begin{aligned} \mathcal{E}_0(\mathcal{G}) &= D_K(\mathbb{P}, \mathbb{Q}) - \sup_{g \in \mathcal{G}} \int (\log g \, d\mathbb{P} - g \, d\mathbb{Q} + 1) \geq 0 && (6.4) \\ \mathcal{E}_1(\mathcal{G}) &= \sup_{g \in \mathcal{G}} \left| \int \log g \, d(\mathbb{P}_n - \mathbb{P}) \, - g d(\mathbb{Q}_n - \mathbb{Q}) \right|. && (6.5) \end{aligned}$$

From (6.2)(6.3)(6.4) and (6.5), it is simple to see that:

$$-\mathcal{E}_1(\mathcal{G}) - \mathcal{E}_0(\mathcal{G}) \leq \hat{D}_K - D_K(\mathbb{P}, \mathbb{Q}) \leq \mathcal{E}_1(\mathcal{G}).$$

For the density ratio estimation, $\hat{D}_K - D_K(\mathbb{P},\mathbb{Q})$ can also be considered as a performance measure. Note that p_0/q_0 can be viewed as a density function with respect to $\mathbb{Q}$ measure. A natural performance measure is the Hellinger distance:

$$h^2_{\mathbb{Q}}(g,g_0) := \frac{1}{2}\int (g^{1/2} - g_0^{1/2})^2 \, d\mathbb{Q}. \tag{6.6}$$

As we shall see, this distance measure is less strong than $\hat{D}_K - D_K(\mathbb{P},\mathbb{Q})$, but it allows us to obtain convergence rate guarantees with less assumption.

6.3 Consistency analysis

In this section we shall prove consistency results and obtain convergence rates of our estimators. Throughout the chapter, the following assumptions are made with respect to $\mathbb{P},\mathbb{Q}$ and the function class $\mathcal{G}$.

Assumptions. (i) $D_K(\mathbb{P},\mathbb{Q}) < \infty$.
(ii) $\mathcal{G}$ is sufficiently rich, i.e., $g_0 \in \mathcal{G}$.

Due to (ii), $\mathcal{E}_0(\mathcal{F}) = 0$. Hence, we shall focus on estimation error $\mathcal{E}_1(\mathcal{G})$ only. Note that if (ii) does not hold, we should obtain instead a lower bound of the KL divergence.

6.3.1 Preliminary lemmas

Define the following processes:

$$v_n(\mathcal{G}) = \sup_{g\in\mathcal{G}}\left|\int \log\frac{g}{g_0} d(\mathbb{P}_n - \mathbb{P}) - \int (g-g_0) d(\mathbb{Q}_n - \mathbb{Q})\right|.$$

$$w_n(g_0) = \left|\int \log g_0 \; d(\mathbb{P}_n - \mathbb{P}) \; - g_0 d(\mathbb{Q}_n - \mathbb{Q})\right|.$$

We have:

$$\mathcal{E}_1(\mathcal{G}) \le v_n(\mathcal{G}) + w_n(g_0). \tag{6.7}$$

Lemma 6.2. $w_n(g_0) \xrightarrow{a.s.} 0$.

Note that in this lemma and other theorems, all almost sure convergence statement can be understood with respect to either $\mathbb{P}$ or $\mathbb{Q}$ because they share the same support.

Proof. This follows immediately from the law of large numbers. We only need to check the condition for which this law applies. Applying the following inequality due to Csiszár (cf. [Gyorfi and van der Meulen, 1987]):

$$\int p_0|log(p_0/q_0)| \leq D_K(\mathbb{P},\mathbb{Q}) + 4\sqrt{D_K(\mathbb{P},\mathbb{Q})}$$

so that $\log g_0$ is $\mathbb{P}$ integrable. In addition, g_0 is $\mathbb{Q}$ integrable, since $\int g_0 d\mathbb{Q} = \int (p_0/q_0) d\mathbb{Q} = 1$. □

Next, we shall relate $v_n(\mathcal{G})$ to the Hellinger distance. This is done through an intermediate term which is also a (pseudo) distance between g_0 and g:

$$d(g_0,g) \;=\; \int (g-g_0)d\mathbb{Q} - \log\frac{g}{g_0}d\mathbb{P}. \qquad (6.8)$$

Lemma 6.3. *(i)* $d(g_0,g) \geq 2h^2_{\mathbb{Q}}(g,g_0)$.
(ii) If $\hat{g}_n$ *is an estimate of* g*, then* $d(g_0,\hat{g}_n) \leq v_n(\mathcal{G})$.

Proof. (i) Note that for $x>0$, $\frac{1}{2}\log x \leq \sqrt{x}-1$. Thus, $\int \log\frac{g}{g_0}\, d\mathbb{P} \leq 2\int (g^{1/2}g_0^{-1/2}-1)\, d\mathbb{P}$. As a result,

$$\begin{aligned} d(g_0,g) &\geq \int (g-g_0)\, d\mathbb{Q} - 2\int (g^{1/2}g_0^{-1/2}-1)\, d\mathbb{P} \\ &= \int (g-g_0)\, d\mathbb{Q} - 2\int (g^{1/2}g_0^{1/2}-g_0)\, d\mathbb{Q} \\ &= \int (g^{1/2}-g_0^{1/2})^2 d\mathbb{Q}. \end{aligned}$$

(ii) By our estimation procedure, we have $\int \hat{g}_n d\mathbb{Q}_n - \int \log\hat{g}_n d\mathbb{P}_n \leq \int g_0 d\mathbb{Q}_n - \int \log g_0 d\mathbb{P}_n$. It follows that

$$\begin{aligned} d(g_0,\hat{g}_n) &= \int (\hat{g}_n-g_0)d\mathbb{Q} - \int (\log\hat{g}_n - \log g_0)d\mathbb{P} \\ &\leq \int (\hat{g}_n-g_0)d(\mathbb{Q}-\mathbb{Q}_n) - \int (\log\hat{g}_n-\log g_0)d(\mathbb{P}-\mathbb{P}_n) \\ &\leq \sup_{g\in\mathcal{G}} \int \log\frac{g}{g_0}d(\mathbb{P}_n-\mathbb{P}) - \int (g-g_0)d(\mathbb{Q}_n-\mathbb{Q}). \end{aligned}$$

□

We can prove the Hellinger consistency using less assumption. For that we shall need the following lemma using a similar idea of using $(g_0+g)/2$ due to Birgé and Massart (cf. [van de Geer, 1999]):

Lemma 6.4. *If $\hat{g}_n$ is an estimate of g, then:*

$$\frac{1}{8}h^2_{\mathbb{Q}}(g_0,\hat{g}_n) \leq 2h^2_{\mathbb{Q}}(g_0,\frac{g_0+\hat{g}_n}{2}) \leq -\int \frac{\hat{g}_n-g_0}{2}d(\mathbb{Q}_n-\mathbb{Q}) + \int \log\frac{\hat{g}_n+g_0}{2g_0}d(\mathbb{P}_n-\mathbb{P}).$$

Proof. The first inequality is straigthforward. We shall focus on the second. By the definition of our estimator, we have:

$$\int \hat{g}_n d\mathbb{Q}_n - \int \log \hat{g}_n d\mathbb{P}_n \leq \int g_0 d\mathbb{Q}_n - \int \log g_0 d\mathbb{P}_n.$$

Both sides are convex functionals of g. Use the following fact: If F is a convex function and $F(u) \leq F(v)$, then $F((u+v)/2) \leq F(v)$. We obtain:

$$\int \frac{\hat{g}_n+g_0}{2}d\mathbb{Q}_n - \int \log\frac{\hat{g}_n+g_0}{2}d\mathbb{P}_n \leq \int g_0 d\mathbb{Q}_n - \int \log g_0 d\mathbb{P}_n.$$

Rearranging,

$$\int \frac{\hat{g}_n-g_0}{2}d(\mathbb{Q}_n-\mathbb{Q}) - \int \log\frac{\hat{g}_n+g_0}{2g_0}d(\mathbb{P}_n-\mathbb{P}) \leq \int \log\frac{\hat{g}_n+g_0}{2g_0}d\mathbb{P} - \int \frac{\hat{g}_n-g_0}{2}d\mathbb{Q}$$
$$= -d(g_0,\frac{g_0+\hat{g}_n}{2}) \leq -2h^2_{\mathbb{Q}}(g_0,\frac{g_0+\hat{g}_n}{2}),$$

where the last inequality is an application of Lemma 6.3.

□

6.3.2 Consistency results

Our analysis shall rely on results from empirical processes theory. We first introduce several standard notions of *entropy* of a function class (see, e.g., [van der Vaart and Wellner, 1996] for more detail). For each $\delta > 0$, a covering for function class $\mathcal{G}$ using metric $L_r(\mathbb{Q})$ is a collection of functions which cover entire $\mathcal{G}$ using $L_r(\mathbb{Q})$ balls of radius δ and centering at these functions. Let $N_\delta(\mathcal{G}, L_r(\mathbb{Q}))$ be the smallest cardinality of such a covering, then $\mathcal{H}_\delta(\mathcal{G}, L_r(\mathbb{Q})) := \log N_\delta(\mathcal{G}, L_r(\mathbb{Q}))$ is called the entropy for $\mathcal{G}$ using $L_r(\mathbb{Q})$ metric. A related notion is *entropy with bracketing*. Let $N^B_\delta(\mathcal{G}, L_r(\mathbb{Q}))$ be the smallest value of N for which there exist pairs of functions $\{g^L_j, g^U_j\}$ such that $\|g^U_j - g^L_j\|_{L_r(\mathbb{Q})} \leq \delta$, and such that for each $g \in \mathcal{G}$ thre is a j such that $g^L_j \leq g \leq g^L_j$. Then $\mathcal{H}^B_\delta(\mathcal{G}, L_r(\mathbb{Q})) := \log N^B_\delta(\mathcal{G}, L_r(\mathbb{Q}))$ is called the entropy with bracketing of $\mathcal{G}$. Define the envelope functions:

$$G_0(x) = \sup_{g\in\mathcal{G}} |g(x)|.$$

$$G_1(x) = \sup_{g\in\mathcal{G}} |\log \frac{g(x)}{g_0(x)}|,$$

Proposition 6.5. *Assume the envelope conditions*

$$\int G_0 d\mathbb{Q} \quad < \quad \infty \tag{6.9a}$$

$$\int G_1 d\mathbb{P} \quad < \quad \infty \tag{6.9b}$$

and suppose that for all $\delta > 0$ *there holds:*

$$\frac{1}{n}\mathcal{H}_\delta(\mathcal{G} - g_0, L_1(\mathbb{Q}_n)) \xrightarrow{\mathcal{P}} \mathbb{Q}0, \tag{6.10a}$$

$$\frac{1}{n}\mathcal{H}_\delta(\log \mathcal{G}/g_0, L_1(\mathbb{P}_n)) \xrightarrow{\mathcal{P}} \mathbb{P}0. \tag{6.10b}$$

Then, $v_n(\mathcal{G}) \xrightarrow{a.s.} 0$*. As a result,* $\mathcal{E}_1(\mathcal{G}) \xrightarrow{a.s.} 0$*, and* $h_{\mathbb{Q}}(g_0, \hat{g}_n) \xrightarrow{a.s.} 0$.

Proof. That $v_n(\mathcal{G}) \xrightarrow{a.s.} 0$ is a direct consequence of Thm 6.15 (see the Appendix). By (6.7) and Lemma 6.2, $\mathcal{E}_1(\mathcal{G}) \xrightarrow{a.s.} 0$. By Lemma 6.3, this would also imply that $h_{\mathbb{Q}}(\hat{g}_n, g_0) \xrightarrow{a.s.} 0$, i.e., our estimation of the ratio p_0/g_0 is consistent in Hellinger sense. □

The envelope condition (6.9a) is satisfied if $\mathcal{G}$ is uniformly bounded from above. The envelope condition (6.9b) is much more severe. Due to logarithm, this can be satisfied if all functions in $\mathcal{G}$ is bounded from *both* above and below. To ensure the Hellinger consistency of the estimation for g_0, however, we can essentially drop the envelope condition (6.9b) as well as the entropy condition (6.10b), which is replaced by a milder entropy condition.

Proposition 6.6. *Assume that* (6.9a) *and* (6.10a) *holds, and*

$$\frac{1}{n}\mathcal{H}_\delta(\log \frac{\mathcal{G} + g_0}{2g_0}, L_1(\mathbb{P}_n)) \xrightarrow{\mathcal{P}} \mathbb{P}0. \tag{6.11}$$

then $h_{\mathbb{Q}}(g_0, \hat{g}_n) \xrightarrow{a.s.} 0$.

Proof. Define $G_2(x) = \sup_{g\in\mathcal{G}} |\log \frac{g(x)+g_0(x)}{2g_0(x)}|$. Due to Lemma 6.4(i) and Thm 6.15 (see the Appendix), it is sufficient to prove that

$$\int G_2 d\mathbb{P} < \infty. \tag{6.12}$$

Indeed,

$$\int G_2 d\mathbb{P} \leq \int \sup_{g\in\mathcal{G}} \max\{\frac{g(x)+g_0(x)}{2g_0(x)} - 1, \log 2\} d\mathbb{P} \leq \log 2 + \int \sup_{g\in\mathcal{G}} |g(x) - g_0(x)| d\mathbb{Q} < \infty,$$

where the last inequality is due to envelope condition (6.9a). □

Remark. Let us now turn to a discussion of the entropy conditions. Note that both entropy conditions (6.10a) and (6.11) can be deduced from the following single condition: For all $\delta > 0$,

$$\mathcal{H}_\delta^B(\mathcal{G}, L_1(\mathbb{Q})) < \infty. \tag{6.13}$$

Indeed, that (6.13) implies (6.10a) is a direct consequence of the law of large numbers (given (6.9a)). To show (6.11), note that (by Taylor's expansion):

$$\left| \log \frac{g_1 + g_0}{2g_0} - \log \frac{g_2 + g_0}{2g_0} \right| \leq \frac{|g_1 - g_2|}{g_0},$$

so $\frac{1}{n}\mathcal{H}_\delta(\log \frac{\mathcal{G}+g_0}{2g_0}, L_1(\mathbb{P}_n)) \leq \frac{1}{n}\mathcal{H}_\delta(\mathcal{G}/g_0, L_1(\mathbb{P}_n))$. Since $G_0 \in L_1(\mathbb{Q})$, we have $G_0/g_0 \in L_1(\mathbb{P})$. In addition, $\mathcal{H}_\delta^B(\mathcal{G}/g_0, L_1(\mathbb{P})) \leq \mathcal{H}_\delta^B(\mathcal{G}, L_1(\mathbb{Q})) < \infty$. By the law of large numbers, $\mathcal{H}_\delta(\mathcal{G}/g_0, L_1(\mathbb{P}_n))$ is bounded in probability, thus (6.11) holds.

In the remaining of this section, we shall consider an example of smooth function classes for which the conditions of Prop. 6.5 and 6.6 hold.

Sobolev spaces. For $x \in \mathbb{R}^d$, an d-dimensional multi-index $\kappa = (\kappa_1, \ldots, \kappa_d)$ (all κ_i are natural numbers), write $x^\kappa = \prod_{i=1}^d x_i^{\kappa_i}$, and $|\kappa| = \sum_{i=1}^d \kappa_i$. Let D^κ denote the differential operator:

$$D^\kappa g(x) = \frac{\partial^{|\kappa|}}{\partial x_1^{\kappa_1} \ldots \partial x_d^{\kappa_d}} g(x_1, \ldots, x_d).$$

We use $W_r^\alpha(\mathcal{X})$ to denote the Sobolev space of functions $f : \mathcal{X} \to \mathbb{R}$. The norm in $W_r^\alpha(\mathcal{X})$ is defined by

$$||f||_{W_r^\alpha(\mathcal{X})} = ||f||_{L_r(\mathcal{X})} + ||f||_{L_r^\alpha(\mathcal{X})},$$

where

$$||f||^r_{L_r^\alpha(\mathcal{X})} = \sum_{|\kappa|=\alpha} \int |D^\kappa f(x)|^r \, dx.$$

Suppose, for simplicity, that the domain $\mathcal{X}$ is a compact set such as a cube $[0, h]^d$. Assume that p_0 and q_0 are bounded from above *and* below by some constants (these assumptions shall be relaxed in the next section). As a result, g_0 is bounded from above and below. Suppose that

$$\eta_1 \leq g_0(x) = \frac{p_0(x)}{q_0(x)} \leq \eta_2 \text{ for all } x \in \mathcal{X}. \tag{6.14}$$

We now restrict our function class to a Sobolev's space of functions that are bounded from above and below:

$$\mathcal{G} = \left\{ g \in W_r^{\alpha}(\mathcal{X}) \text{ such that } ||g||_{W_r^{\alpha}(\mathcal{X})} \leq M \right\} \cap \left\{ g : \; K_1 \leq g(x) \leq K_2 \text{ for all } x \in \mathcal{X} \right\}, \tag{6.15}$$

where K_1 and K_2 are some constants satisfying $K_1 \leq \eta_1 < \eta_2 \leq K_2$. In the algorithmic development and subsequent analysis of our estimator, we typically restrict ourselves to $r = 2$ unless indicated otherwise.

Under the boundedness assumption, the envelope conditions (6.9) hold trivially. For a function class that is sufficiently smooth, i.e., when $r\alpha > d$, then it was shown [Birman and Solomjak, 1967] that

$$\mathcal{H}_{\delta}(\mathcal{G}, L_{\infty}) < c\delta^{-d/\alpha} < \infty,$$

where c is some constant independent of δ. As a result, it is simple to see that the condition (6.13) holds. The entropy condtion (6.10b) also holds due to the boundedness condition.

Finally, while the boundedness conditions are rather severe, we can study the rate of convergence under such conditions. Once having the convergence rates for bounded cases, it would be easy to obtain consistency in more general unbounded cases if we have additional knowledge of the tail condition for the densities.

6.4 Rates of convergence

6.4.1 Convergence rate of the likelihood ratio in Hellinger metric

In this section, we shall obtain the same convergence rate of the likelihood ratio g using Hellinger metric as a performance measure. Our result is based on Lemma 6.4, in which the Hellinger distance is bounded from above by the suprema of two empirical processes.

We shall need the assumption that

$$\sup_{g \in \mathcal{G}} \|g\|_{\infty} < K_2. \tag{6.16}$$

One empirical process in the RHS in Lemma 6.4 involves function class $\mathcal{F} := \log \frac{\mathcal{G}+g_0}{2g_0}$. For each $g \in \mathcal{G}$, let $f_g := \log \frac{g+g_0}{2g_0}$. We endow $\mathcal{F}$ with a new norm, namely, *Bernstein distance*: for a constant $K > 0$,

$$\rho_K(f)^2 := 2K^2 \int (e^{|f|/K} - 1 - |f|/K) d\mathbb{P}.$$

The Bernstein distance is related to the Hellinger distance in several crucial ways (see, e.g., [van de Geer, 1999], page 97):

- $\rho_1(f_g) \leq 4h_{\mathbb{Q}}(g_0, \frac{g+g_0}{2})$.
- The bracket entropy based on Bernstein distance is also related to the bracket entropy based Hellinger distance (i.e., which is the L_2 norm for the square root function):

$$\mathcal{H}^B_{\sqrt{2}\delta}(\mathcal{F}, \rho_1) \leq \mathcal{H}^B_\delta(\bar{\mathcal{G}}, L_2(\mathbb{Q})), \tag{6.17}$$

  where $\bar{\mathcal{G}} := \{((g+g_0)/2)^{1/2}, g \in \mathcal{G}\}$, and $\bar{g} := (g+g_0)/2$.

We shall need an assumption on function class $\bar{\mathcal{G}}$: For some constant $0 < \gamma_{\bar{\mathcal{G}}} < 2$, there holds for any $\delta > 0$,

$$\mathcal{H}^B_\delta(\bar{\mathcal{G}}, L_2(\mathbb{Q})) = O(\delta^{-\gamma_{\bar{\mathcal{G}}}}). \tag{6.18}$$

Combining this condition with (6.16), we deduce that for $\mathcal{G}$,

$$\mathcal{H}^B_\delta(\mathcal{G}, L_2(\mathbb{Q})) \leq O(\delta^{-\gamma_{\bar{\mathcal{G}}}}).$$

In the following theorem, $O_{\mathbb{P}}$ means "bounded in probability" with respect to $\mathbb{P}$ measure.

Theorem 6.7. *Assume* (6.16) *and* (6.18), *then* $h_{\mathbb{Q}}(g_0, \hat{g}_n) = O_{\mathbb{P}}(n^{-1/(\gamma_{\bar{\mathcal{G}}}+2)})$.

Proof. By Lemma 6.4, for any $\delta > 0$, with respect to $\mathbb{P}$ measure:

$$\begin{aligned}
&P(h_{\mathbb{Q}}(g_0, \hat{g}_n) > \delta) \leq P(h_{\mathbb{Q}}(g_0, (\hat{g}_n + g_0)/2) > \delta/4) \\
&\leq P\Big(\sup_{g\in\mathcal{G},\, h_{\mathbb{Q}}(g_0,\bar{g})>\delta/4} - \int (\bar{g} - g_0) d(\mathbb{Q}_n - \mathbb{Q}) + \int f_g \, d(\mathbb{P}_n - \mathbb{P}) - 2h^2_{\mathbb{Q}}(g_0, \bar{g}) \geq 0\Big) \\
&\leq P\Big(\sup_{g\in\mathcal{G},\, h_{\mathbb{Q}}(g_0,\bar{g})>\delta/4} - \int (\bar{g} - g_0) d(\mathbb{Q}_n - \mathbb{Q}) - h^2_{\mathbb{Q}}(g_0, \bar{g}) \geq 0\Big) + \\
&P\Big(\sup_{g\in\mathcal{G},\, h_{\mathbb{Q}}(g_0,\bar{g})>\delta/4} \int f_g \, d(\mathbb{P}_n - \mathbb{P}) - h^2_{\mathbb{Q}}(g_0, \bar{g}) \geq 0\Big) := A + B.
\end{aligned}$$

We need to upper bound the RHS's two quantities A and B, both of which can be handled in a similar manner. Since $\mathcal{H}^B_\delta(\bar{\mathcal{G}}, L_2(\mathbb{Q})) < \infty$ the diameter of $\bar{\mathcal{G}}$ is finite. Let S be the minimum s such that $2^{s+1}\delta/4$ exceeds that diameter. We apply the so-called peeling device: Decompose $\bar{\mathcal{G}}$ into layers of Hellinger balls around g_0 and then applying union bound on the probability of excess. For each layer, one can now apply the modulus of continuity of suprema of an empirical process.

$$B \;\leq\; \sum_{s=0}^{S} P\biggl(\sup_{g\in\mathcal{G},\, h_{\mathbb{Q}}(g_0,\bar{g})\leq 2^{s+1}\delta/4} \int f_g \; d(\mathbb{P}_n - \mathbb{P}) \geq 2^{2s}(\delta/4)^2\biggr).$$

Note that if $h_{\mathbb{Q}}(g_0,\bar{g}) \leq 2^{s+1}\delta/4$ then $\rho_1(f_g) \leq 2^{s+1}\delta$. Note that for any $s = 1,\ldots,S$, the bracket entropy integral can be bounded as:

$$\begin{aligned}
&\int_0^{2^{s+1}\delta} \mathcal{H}^B_{\epsilon}(\mathcal{F}\cap\{h_{\mathbb{Q}}(g_0,\bar{g}) \leq 2^{s+1}\delta/4\},\rho_1)^{1/2}\; d\epsilon \\
\leq\;& \int_0^{2^{s+1}\delta} \mathcal{H}^B_{\epsilon/\sqrt{2}}(\bar{\mathcal{G}}\cap\{h_{\mathbb{Q}}(g_0,\bar{g}) \leq 2^{s+1}\delta/4\},L_2(\mathbb{Q}))^{1/2}\; d\epsilon \\
\leq\;& \int_0^{2^{s+1}\delta} C_9(\epsilon/\sqrt{2})^{-\gamma_{\bar{\mathcal{G}}}/2}\; d\epsilon \\
\leq\;& C_8(2^{s+1}\delta)^{1-\gamma_{\bar{\mathcal{G}}}/2},
\end{aligned}$$

where C_8, C_9 are constants independent of s. Now apply Thm 6.16 (see the Appendix), where $K = 1$, $R = 2^{s+1}\delta$, $a = C_1\sqrt{n}R^2/K = C_1\sqrt{n}2^{2(s+1)}\delta^2$. We need

$$a \geq C_0C_8(2^{s+1}\delta)^{1-\gamma_{\bar{\mathcal{G}}}/2} > C_0R.$$

This is satisfied if $\delta = n^{-1/(\gamma_{\bar{\mathcal{G}}}+2)}$, and $C_1 = C_0C_8$, where C_8 is sufficiently large (independently of s). Finally, $C_0^2 \geq C^2(C_1+1) = C^2(C_0C_8+1)$ if $C_0 := 2C^2C_8 \vee 2C$, where C is some universal constant in Thm 6.16. Applying this theorem, we obtain:

$$B \;\leq\; \sum_{s=0}^{S} C\exp\Bigl[-\frac{C_1^2 n 2^{2(s+1)}\delta^2}{C^2(C_1+1)}\Bigr] \leq c\exp\Bigl[-\frac{n\delta^2}{c^2}\Bigr]$$

for some universal constant c. A similar bound for A, with respect to $\mathbb{Q}$ measure and with $\delta = n^{-1/(\gamma_{\bar{\mathcal{G}}}+2)}$ can be obtained in the same manner. Since p_0/q_0 is bounded from above, this also implies a probability statement with respect to $\mathbb{P}$. Thus, $h_{\mathbb{Q}}(g_0,\hat{g}_n)$ is bounded in $\mathbb{P}$ probability by $n^{-1/(\gamma_{\bar{\mathcal{G}}}+2)}$. □

In the following we note that the rate of convergence with respect to Hellinger metric is also the optimal minimax rate, which is defined as:

$$r_n := \inf_{\hat{g}_n\in\mathcal{G}} \sup_{\mathbb{P},\mathbb{Q}} \mathbb{E}_{\mathbb{P}} h_{\mathbb{Q}}(g_0,\hat{g}_n).$$

First, note that $r_n \geq \inf_{\hat{g}_n\in\mathcal{G}}\sup_{\mathbb{P}} \mathbb{E}h_{\mu}(g_0,\hat{g}_n)$, where we have fixed $\mathbb{Q} = \mu$ the Lebesgue measure on $\mathcal{X}$. Our strategy is to reduce this bound to that minimax lower bound for a nonparametric density estimation problem [Yu, 1996]. Note a technicality here, in which the space $\mathcal{G}$ ranges over smooth functions that need not to be valid probability density. Therefore, an easy-to-use mimimax lower bound such as that of [Yang and Barron, 1999] is not immediately applicable. Nonetheless, we can still apply the hypercube argument and the Assouad lemma to obtain the right minimax rate. See [van der Vaart, 1998] (Sec. 24.3) for a proof for the case of one dimension. The proof goes through for general $d \geq 1$.

Proposition 6.8. *For $\mathcal{G}$ defined in (6.15), $\mathbb{P}, \mathbb{Q}$ satisfy (6.14), $r_n = \Omega(n^{-1/(\gamma+2)})$, where $\gamma = d/\alpha$.*

6.4.2 Convergence rate for divergence estimation

In this section we shall obtain the convergence rate of our estimation procedure for the KL divergence, i.e., $\|\hat{D}_K - D_K(\mathbb{P}, \mathbb{Q}).\|$ We shall need the assumption that all functions in $\mathcal{G}$ are bounded from above *and* below:

$$0 < K_1 \leq g \leq K_2 \text{ for all } g \in \mathcal{G}. \tag{6.19}$$

Theorem 6.9. *Assume (6.19) and (6.18), then $|\hat{D}_K - D_K(\mathbb{P}, \mathbb{Q})| = O_{\mathbb{P}}(n^{-1/(\gamma_{\bar{\mathcal{G}}}+2)})$.*

Proof. Note that

$$\begin{aligned}
|\hat{D}_K - D_K(\mathbb{P}, \mathbb{Q})| &= \left| \int \log \hat{g}_n d\mathbb{P}_n - \int \hat{g}_n d\mathbb{Q}_n - (\int \log g_0 d\mathbb{P} - \int g_0 d\mathbb{Q}) \right| \\
&\leq \left| \int \log \hat{g}_n / g_0 d(\mathbb{P}_n - \mathbb{P}) - \int (\hat{g}_n - g_0) d(\mathbb{Q}_n - \mathbb{Q}) \right| \\
&+ \left| \int \log \hat{g}_n / g_0 d\mathbb{P} - \int (\hat{g}_n - g_0) d\mathbb{Q} \right| \\
&+ \left| \int \log g_0 d(\mathbb{P}_n - \mathbb{P}) - \int g_0 d(\mathbb{Q}_n - \mathbb{Q}) \right| := A + B + C.
\end{aligned}$$

We have $C = O_P(n^{-1/2})$ by the central limit theorem. Using assumption (6.19),

$$\begin{aligned}
B &\leq \int |\hat{g}_n - g_0| \frac{K_2}{K_1} d\mathbb{Q}| + \int |\hat{g}_n - g_0| d\mathbb{Q} \\
&\leq (K_2/K_1 + 1) \|\hat{g}_n - g_0\|_{L_2(\mathbb{Q})} \\
&\leq (K_2/K_1 + 1) \left(\int 4K_2 (\hat{g}_n^{1/2} - g_0^{1/2})^2 d\mathbb{Q} \right)^{1/2} \\
&\leq (K_2/K_1 + 1) K_2^{1/2} 4 h_{\mathbb{Q}}(g_0, \hat{g}_n) = O_{\mathbb{P}}(n^{-1/(2+\gamma_{\bar{\mathcal{G}}})}),
\end{aligned}$$

where the last equality is due to Thm 6.7.

Finally, to bound A, we shall apply a modulus of continuity result on the suprema of empirical processes with respect to function $(g - g_0)$ and $(\log g - \log g_0)$. In particular, due to (6.19), the bracket entropy for both function classes $\mathcal{G}$ and $\log \mathcal{G}$ has the same order as that of $\bar{\mathcal{G}}$, as given in (6.18). Apply Lemma 6.17 (see the Appendix), we obtain that for $\delta_n = n^{-1/(2+\gamma_{\bar{\mathcal{G}}})}$, there holds:

$$A = O_{\mathbb{P}}(n^{-1/2}\|\hat{g}_n - g_0\|_{L_2(\mathbb{Q})}^{1-\gamma_{\bar{\mathcal{G}}}/2} \vee \delta_n^2) = O_{\mathbb{P}}(n^{-2/(2+\gamma_{\bar{\mathcal{G}}})}).$$

The overall estimation error is bounded by the upper bound of B. □

6.5 General methods for estimating f-divergence

In this section, we shall present several general methods for estimating f-divergence, and discuss their properties and limitations.

6.5.1 M-estimator of D_ϕ and p_0/q_0

It is not difficult to see that our method for estimating the KL divergence can be easily applied to any divergence $D_\phi(p_0, q_0)$. In fact, the method for consistency analysis, while tailored to each specific choice of ϕ, is also very similar in spirit. Assume in this section that ϕ is a differentiable (convex) function. Motivated by Lemma 6.1, our estimator has the following form:

$$\hat{D}_\phi := \sup_{f\in\mathcal{F}} \int f \; d\mathbb{Q}_n - \int \phi^*(f) \; d\mathbb{P}_n. \tag{6.20}$$

Let $\hat{f}$ be the supremum of the above optimization. $\hat{f}$ is considered an estimator of $f_0 = \phi'(q_0/p_0)$. As before, we define the estimation and approximation error. The latter is assumed to be 0, i.e., $\phi'(q_0/p_0) \in \mathcal{F}$.

$$\mathcal{E}_0^\phi(\mathcal{F}) \;=\; D_\phi(\mathbb{P}, \mathbb{Q}) - \sup_{f\in\mathcal{F}} \int (f \; d\mathbb{Q} - \phi^*(f) \; d\mathbb{P}) \geq 0 \tag{6.21}$$

$$\mathcal{E}_1^\phi(\mathcal{F}) \;=\; \sup_{f\in\mathcal{F}} \left| \int f d(\mathbb{Q}_n - \mathbb{Q}) - \phi^*(f) \; d(\mathbb{P}_n - \mathbb{P}) \right|. \tag{6.22}$$

From (6.20),(6.4) and (6.5), it is simple to see that: $-\mathcal{E}_1^\phi(\mathcal{F}) - \mathcal{E}_0^\phi(\mathcal{F}) \leq \hat{D}_\phi - D_\phi(\mathbb{P}, \mathbb{Q}) \leq \mathcal{E}_1(\mathcal{F})$. Since $\mathcal{E}_0^\phi(\mathcal{F}) = 0$, our main focus is in analysis of $\mathcal{E}_1^\phi(\mathcal{F})$. As before, define:

$$v_n^\phi(\mathcal{F}) = \sup_{f\in\mathcal{F}} \left| \int (\phi^*(f) - \phi^*(f_0)) d(\mathbb{P}_n - \mathbb{P}) - \int (f - f_0) d(\mathbb{Q}_n - \mathbb{Q}) \right|.$$

$$w_n^\phi(f_0) = \left| \int \phi^*(f_0)\, d(\mathbb{P}_n - \mathbb{P}) \; - f_0 d(\mathbb{Q}_n - \mathbb{Q}) \right|.$$

We have $\mathcal{E}_1^\phi(\mathcal{F}) \leq v_n^\phi(\mathcal{F}) + w_n^\phi(f_0)$. Since $w_n(\mathcal{F})$ converges to 0 almost surely under mild assumptions on g_0, to prove consistency of our estimator, it remains to analyze the convergence of $v_n^\phi(\mathcal{F})$. This can be done in the same manner as in Section 6.3.

To analyze the convergence rate of our estimator, the key idea of our analysis is to exploit the modulus of continuity of the supremum of the empirical processes involved in the definition of $v_n^\phi(\mathcal{F})$ with respect the a notion of distance between g and f_0:

$$\begin{aligned} d_\phi(f_0, f) \;&:=\; D_\phi(\mathbb{P}, \mathbb{Q}) - \int f d\mathbb{Q} - \phi^*(f) d\mathbb{P} &(6.23)\\ &=\; \int (\phi^*(f) - \phi^*(f_0)) d\mathbb{P} - (f - f_0) d\mathbb{Q} &(6.24)\\ &=\; \int (\phi^*(f) - \phi^*(f_0) - \left.\frac{\partial \phi^*}{\partial f}\right|_{f_0} (f - f_0)\; d\mathbb{P} \geq 0. &(6.25)\end{aligned}$$

The last line in the above equation shows that d_ϕ is a *Bregman divergence* using convex function ϕ^*. The following lemma is an analogue of Lemma 6.3(ii) whose proof is straightforward:

Lemma 6.10. *If $\hat{f}$ is an estimation of f_0 by solving* (6.20)*, then $d_\phi(f_0, \hat{f}) \leq v_n^\phi(\mathcal{F})$.*

Since $d_\phi(f_0, f)$ is usually not a proper metric, to apply standard results from empirical process theory one usually needs to replace d_ϕ by a lower bound which is a proper metric (such as L_2 or Hellinger metric). In the case of KL divergence, we have seen that this lower bound is the Hellinger distance.

In the following we shall demonstrate our general method to the estimation yet another f-divergence: the χ-square distance. This divergence is very amenable to the general framework just described. As we shall see, it also plays a special role in another general method for divergence any f-divergence.

The χ-square divergence is defined as $D_\chi(\mathbb{P}, \mathbb{Q}) = \int p_0^2/q_0 d\mu$. It is a f-divergence with $\phi(u) = 1/u$. We have $\phi^*(v) = -2\sqrt{-v}$ if $v < 0$ and $+\infty$ otherwise. As a result, we only need to restrict $\mathcal{F}$ to the subset for which $f < 0$ for any $f \in \mathcal{F}$. Let $g := \sqrt{-f}$ and $\mathcal{G} = \sqrt{-\mathcal{F}}$. $\mathcal{G}$ is a function class of positive functions. We have $g_0 := \sqrt{-f_0} = \sqrt{-\phi'(q_0/p_0)} = p_0/q_0$. We shall also replace notation $d_\phi(f_0, f)$ by $d_\phi(g_0, g)$. For our choice of ϕ, we have:

$$\begin{aligned} d_\chi(g_0, g) &= d_\chi(f_0, f) = \int (-2\sqrt{-f} + 2\sqrt{-f_0}) d\mathbb{P} - (f - f_0) d\mathbb{Q} \\ &= \int (g_0 - g)(2p_0/q_0 - g_0 - g) d\mathbb{Q} \\ &= \int (g - g_0)^2 d\mathbb{Q} \\ v_n^\chi(\mathcal{G}) &= v_n^\chi(\mathcal{F}) = \sup_{g \in \mathcal{G}} \left| \int 2(g^2 - g_0^2) d(\mathbb{Q}_n - \mathbb{Q}) - \int (g - g_0) d(\mathbb{P}_n - \mathbb{P}) \right|. \end{aligned}$$

Assume moreover that for some constant $0 < \gamma < 2$,

$$\mathcal{H}_\delta^B(\mathcal{G}, L_2(\mathbb{Q})) \quad \leq \quad A_\mathcal{G} \delta^{-\gamma} \tag{6.26}$$

The following theorem is a parallel of Thm 6.7; the proof is essentially the same (if not simpler) and therefore not included herein:

Theorem 6.11. *Assume* (6.16) *and* (6.26), *and* $\hat{g}_n$ *be our estimator of* g_0 *then:* $d_\chi(g_0, \hat{g}_n) = O_\mathbb{P}(n^{-2/(\gamma+2)})$.

Remark. Comparing with Thm 6.7 the assumption on the $L_2(\mathbb{Q})$-based entropy and the assertion on the $L_2(\mathbb{Q})$ metric are both weaker, because the $L_2(\mathbb{Q})$ metric is less strong than the Hellinger distance.

Finally, it is straightward to show that our method is applicable to a broader class of functional of the following form:

$$T(\mathbb{P}, \mathbb{Q}) = \int p_0 \phi(q_0/p_0) \psi \, d\mu,$$

where $\psi : \mathcal{X} \to \mathbb{R}_+$ is a known positive function that is also bounded from both above and below (away from 0). All analysis goes through, with the insertion of ψ in all integrals involved. We also obtain the same convergence rate as when $\psi = 1$.

6.5.2 Plug-in estimator based on Taylor expansion

In this section we shall present an estimator based on functional delta method. This idea was also used by [Joe, 1989; Birgé and Massart, 1995] to estimate integral functional of a density function. While $D_\phi(\mathbb{P}, \mathbb{Q})$ is a functional of two densities, we can exploit its special structure and our method of estimating the density ratio to achieve an estimator of with similar effects. Indeed, we can write

$$D_\phi(\mathbb{P},\mathbb{Q}) = \int (p_0/q_0)\phi(q_0/p_0)\, d\mathbb{Q} = \int g_0\phi(1/g_0)\, d\mathbb{Q}.$$

Thus, D_ϕ can be viewed as an integral functional of $g_0 = p_0/q_0$. Of course, the difference here is that the integration is with respect to unknown $\mathbb{Q}$.

Suppose that $\phi : \mathbb{R}_+ \to \mathbb{R}$ is a differentiable convex function up to the third order, $\mathcal{G}$ is a smooth function class bounded from both above and below as in (6.15) (with smooth parameter α). Suppose that $\hat{g}_n$ is an estimator of g_0 such as the one described in the previous section, i.e., $\|\hat{g}_n - g_0\|_{L_2(\mathbb{Q})} = d_\chi(g_0, \hat{g}_n) = O_P(n^{-\alpha/(2\alpha+d)})$. Using a Taylor expansion around $\hat{g}_n$, we obtain:

$$\begin{aligned} g\phi(1/g) &= \hat{g}_n\phi(1/\hat{g}_n) + (g-\hat{g}_n)(\phi(1/\hat{g}_n) - \phi'(1/\hat{g}_n)/\hat{g}_n) + (g-\hat{g}_n)^2\phi''(1/\hat{g}_n)/\hat{g}_n^3 + \\ &\quad O((g-\hat{g}_n)^3) \\ &= \phi'(1/\hat{g}_n) + \phi''(1/\hat{g}_n)/\hat{g}_n + g(\phi(1/\hat{g}_n) - \phi'(1/\hat{g}_n)/\hat{g}_n - 2\phi''(1/\hat{g}_n)/\hat{g}_n^2) + \\ &\quad g^2\phi''(1/\hat{g}_n)/\hat{g}_n^3 + O((g-\hat{g}_n)^3). \end{aligned}$$

We arrive at

$$\begin{aligned} D_\phi(\mathbb{P},\mathbb{Q}) &= \int g\phi(1/g)d\mathbb{Q} \\ &= \int \phi'(1/\hat{g}_n) + \phi''(1/\hat{g}_n)/\hat{g}_n\, d\mathbb{Q} + \\ &\quad \int (\phi(1/\hat{g}_n) - \phi'(1/\hat{g}_n)/\hat{g}_n - 2\phi''(1/\hat{g}_n)/\hat{g}_n^2)\, d\mathbb{P} + \\ &\quad \int p_0^2/q_0\phi''(1/\hat{g}_n)/\hat{g}_n^3\, d\mu + O(\|g_0 - \hat{g}_n\|_3^3). \end{aligned}$$

In the above expression, the first two integrals can be estimated from (other) sets of empirical data drawn from $\mathbb{P}$ and $\mathbb{Q}$. Because of the boundedness assumption, these estimations have at most $O_P(n^{-1/2})$ error. The error of our Taylor approximation is $O(\|g_0 - \hat{g}_n\|_3^3) = O_P(n^{-3\alpha/(2\alpha+d)})$. This rate is less than $O(n^{-1/2})$ for $\alpha \geq d/4$. Thus when $\alpha \geq d/4$, the optimal rate of convergence for estimating D_ϕ hinges on the rate of estimating the integral of the form $\int p_0^2/q_0\psi\, d\mu$.

Before ending this section, it is informative to return to the case of KL divergence, i.e., $\phi(u) = -\log u$. If we use Taylor approximation up to first order (thus guaranteeing an error rate of $O_P(n^{-2\alpha/(2\alpha+d)})$, the estimator has the following form:

$$\begin{aligned}\hat{D}_\phi &= \int (\phi(1/\hat{g}_n) - \phi'(1/\hat{g}_n)/\hat{g}_n)\ d\mathbb{P}_n + \int \phi'(1/\hat{g}_n)\ d\mathbb{Q}_n \\ &= \int \log \hat{g}_n + 1 d\mathbb{P}_n - \hat{g}_n d\mathbb{Q}_n,\end{aligned}$$

which has exactly the same form as our original estimator (6.3), except that here $\hat{g}_n$ can be any estimator of the density ratio. The estimator (6.3) achieves simultaneously both goals (i) estimating the density ratio and (ii) estimating the divergence. While our method for (i) achieves the optimal minimax bound, our method for (ii) can be viewed as only a first-order Taylor expansion based plug-in estimator. As discussed in the previous paragraph, it seems that one might obtain a better rate by using Taylor expansion up to second order. This is, of course, possible only if we can obtain a better rate for estimating the integral of the form $\int p_0^2/q_0\psi\ d\mu$.

6.6 M-estimation with penalties

In practice, the "true" size of $\mathcal{G}$ is not known. Accordingly, our approach in this chapter is an alternative approach based on controlling the size of $\mathcal{G}$ by using penalties. More precisely, let $I(g)$ be a measure of complexity for g. Assume that I is a non-negative functional and $I(g_0) < \infty$. We decompose the function class $\mathcal{G}$ as follows:

$$\mathcal{G} = \cup_{1 \leq M \leq \infty} \mathcal{G}_M, \tag{6.27}$$

where $\mathcal{G}_M := \{g \mid I(g) \leq M\}$ is a ball determined by $I(\cdot)$.

The estimation procedure involves solving the following program:

$$\hat{g}_n = \operatorname{argmin}_{g \in \mathcal{G}} \int g d\mathbb{Q}_n - \int \log g\ d\mathbb{P}_n + \frac{\lambda_n}{2} I^2(g), \tag{6.28}$$

where $\lambda_n > 0$ is a regularization parameter. The minimizing argument $\hat{g}_n$ is plugged into (6.3) to obtain an estimate of the KL divergence D_K.

For the KL divergence, the difference $|\hat{D}_K - D_K(\mathbb{P}, \mathbb{Q})|$ is a natural performance measure. For estimating the density ratio, various metrics are possible. Viewing $g_0 = p_0/q_0$ as a density function with respect to $\mathbb{Q}$ measure, one useful metric is the (generalized) Hellinger distance:

$$h_{\mathbb{Q}}^2(g_0, g) := \frac{1}{2} \int (g_0^{1/2} - g^{1/2})^2\ d\mathbb{Q}. \tag{6.29}$$

For the analysis, several assumptions are in order. First, assume that g_0 (*not* all of $\mathcal{G}$) is bounded from above and below:

$$0 < \eta_0 \leq g_0 \leq \eta_1 \text{ for some constants } \eta_0, \eta_1. \tag{6.30}$$

Next, the uniform norm of $\mathcal{G}_M$ is Lipchitz with respect to the penalty measure $I(g)$, i.e.:

$$\sup_{g \in \mathcal{G}_M} |g|_\infty \leq cM \quad \text{for any } M \geq 1. \tag{6.31}$$

Finally, on the bracket entropy of $\mathcal{G}$ [van der Vaart and Wellner, 1996]: For some $0 < \gamma < 2$,

$$\mathcal{H}^B_\delta(\mathcal{G}_M, L_2(\mathbb{Q})) = O(M/\delta)^\gamma \quad \text{for any } \delta > 0. \tag{6.32}$$

The following is our main theoretical result, whose proof is given in Section 6.8:

Theorem 6.12. *(a) Under assumptions* (6.30) (6.31) (6.32)*, and set $\lambda_n \to 0$ so that:*

$$\lambda_n^{-1} = O_{\mathbb{P}}(n^{2/(2+\gamma)})(1 + I(g_0)),$$

then under $\mathbb{P}$:

$$h_{\mathbb{Q}}(g_0, \hat{g}_n) = O_{\mathbb{P}}(\lambda_n^{1/2})(1 + I(g_0)), \quad I(\hat{g}_n) = O_{\mathbb{P}}(1 + I(g_0)).$$

(b) If, in addition to (6.30) (6.31) (6.32)*, there holds $\inf_{g \in \mathcal{G}} g(x) \geq \eta_0$ for any $x \in \mathcal{X}$, then*

$$|\hat{D}_K - D_K(\mathbb{P}, \mathbb{Q})| = O_{\mathbb{P}}(\lambda_n^{1/2})(1 + I(g_0)). \tag{6.33}$$

6.7 Algorithm: Optimization and dual formulation

$\mathcal{G}$ is an RKHS. Our algorithm involves solving program (6.28), for some choice of function class $\mathcal{G}$. In our implementation, relevant function classes are taken to be a reproducing kernel Hilbert space induced by a Gaussian kernel. The RKHS's are chosen because they are sufficiently rich [Saitoh, 1988], and as in many learning tasks they are quite amenable to efficient optimization procedures [Schölkopf and Smola, 2002].

Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Mercer kernel function [Saitoh, 1988]. Thus, K is associated with a feature map $\Phi : \mathcal{X} \to \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space with inner product $\langle ., . \rangle$ and for all $x, x' \in \mathcal{X}$, $K(x, x') = \langle \Phi(x), \Phi(x') \rangle$. As a reproducing kernel Hilbert space, any function $g \in \mathcal{H}$ can be expressed as an inner product $g(x) = \langle w, \Phi(x) \rangle$, where $\|g\|_{\mathcal{H}} = \|w\|_{\mathcal{H}}$. A kernel used in our simulation is the Gaussian kernel:

$$K(x, y) := e^{-\|x-y\|^2/\sigma},$$

where $\|.\|$ is the Euclidean metric in $\mathbb{R}^d$, and $\sigma > 0$ is a parameter for the function class.

Let $\mathcal{G} := \mathcal{H}$, and let the complexity measure be $I(g) = \|g\|_{\mathcal{H}}$. Thus, Eq. (6.28) becomes:

$$\min_w J := \min_w \frac{1}{n}\sum_{i=1}^n \langle w,\, \Phi(x_i)\rangle - \frac{1}{n}\sum_{j=1}^n \log\langle w,\, \Phi(y_j)\rangle + \frac{\lambda_n}{2}\|w\|_{\mathcal{H}}^2, \tag{6.34}$$

where $\{x_i\}$ and $\{y_j\}$ are realizations of empirical data drawn from $\mathbb{Q}$ and $\mathbb{P}$, respectively. The $\log$ function is extended take value $-\infty$ for negative arguments.

Lemma 6.13. $\min_w J$ *has the following dual form:*

$$-\min_{\alpha>0} \sum_{j=1}^n -\frac{1}{n} - \frac{1}{n}\log n\alpha_j + \frac{1}{2\lambda_n}\sum_{i,j}\alpha_i\alpha_j K(y_i, y_j) + \frac{1}{2\lambda_n n^2}\sum_{i,j} K(x_i, x_j) \\ - \frac{1}{\lambda_n n}\sum_{i,j}\alpha_j K(x_i, y_j).$$

Proof. Let $\psi_i(w) := \frac{1}{n}\langle w,\, \Phi(x_i)\rangle$, $\varphi_j(w) := -\frac{1}{n}\log\langle w,\, \Phi(y_j)\rangle$, and $\Omega(w) = \frac{\lambda_n}{2}\|w\|_{\mathcal{H}}^2$. We have

$$\begin{aligned}\min_w J &= -\max_w(\langle 0,\, w\rangle - J(w)) = -J^*(0)\\ &= -\min_{u_i, v_j} \sum_{i=1}^n \psi_i^*(u_i) + \sum_{j=1}^n \varphi_j^*(v_j) + \Omega^*(-\sum_{i=1}^n u_i - \sum_{j=1}^n v_j),\end{aligned}$$

where the last line is due to the inf-convolution theorem [Rockafellar, 1970]. Simple calculations yield:

$$\begin{aligned}\varphi_j^*(v) &= -\frac{1}{n} - \frac{1}{n}\log n\alpha_j \text{ if } v = -\alpha_j\Phi(y_j) \text{ and } +\infty \text{ otherwise}\\ \psi_i^*(u) &= 0 \text{ if } u = \frac{1}{n}\Phi(x_i) \text{ and } +\infty \text{ otherwise}\\ \Omega^*(v) &= \frac{1}{2\lambda_n}\|v\|_{\mathcal{H}}^2.\end{aligned}$$

So, $\min_w J = -\min_{\alpha_i}\sum_{j=1}^n(-\frac{1}{n} - \frac{1}{n}\log n\alpha_j) + \frac{1}{2\lambda_n}\|\sum_{j=1}^n \alpha_j\Phi(y_j) - \frac{1}{n}\sum_{i=1}^n \Phi(x_i)\|_{\mathcal{H}}^2$, which implies the lemma immediately. □

If $\hat\alpha$ is solution of the dual formulation, it is not difficult to show that the optimal $\hat w$ is attained at $\hat w = \frac{1}{\lambda_n}(\sum_{j=1}^n \hat\alpha_j\Phi(y_j) - \frac{1}{n}\sum_{i=1}^n \Phi(x_i))$.

For an RKHS based on a Gaussian kernel, the entropy condition (6.32) holds for any $\gamma > 0$ [Zhou, 2002]. Furthermore, (6.31) trivially holds via the Cauchy-Schwarz inequality: $|g(x)| = |\langle w, \Phi(x)\rangle| \leq \|w\|_{\mathcal{H}}\|\Phi(x)\|_{\mathcal{H}} \leq I(g)\sqrt{K(x,x)} \leq I(g)$. Thus, by Theorem 6.12(a), $\|\hat{w}\|_{\mathcal{H}} = \|\hat{g}_n\|_{\mathcal{H}} = O_{\mathbb{P}}(\|g_0\|_{\mathcal{H}})$, so the penalty term $\lambda_n\|\hat{w}\|^2$ vanishes at the same rate as λ_n. We have arrived at the following estimator for the KL divergence:

$$\hat{D}_K = 1 + \sum_{j=1}^{n}(-\frac{1}{n} - \frac{1}{n}\log n\hat{\alpha}_j) = \sum_{j=1}^{n} -\frac{1}{n}\log n\hat{\alpha}_j.$$

$\log\mathcal{G}$ **is an RKHS.** Alternatively, we could set $\log\mathcal{G}$ to be the RKHS, letting $g(x) = \exp\langle w, \Phi(x)\rangle$, and letting $I(g) = \|\log g\|_{\mathcal{H}} = \|w\|_{\mathcal{H}}$. Theorem 6.12 is not applicable in this case, because condition (6.31) no longer holds, but this choice nonetheless seems reasonable and worth investigating, because in effect we have a far richer function class which might improve the bias of our estimator when the density ratio is not very smooth.

A derivation similar to the previous case yields the following convex program:

$$\begin{aligned}\min_w J \quad &:= \quad \min_w \frac{1}{n}\sum_{i=1}^{n} e^{\langle w,\, \Phi(x_i)\rangle} - \frac{1}{n}\sum_{j=1}^{n}\langle w,\, \Phi(y_j)\rangle + \frac{\lambda_n}{2}\|w\|_{\mathcal{H}}^2 \\ &= \quad -\min_{\alpha>0} \sum_{i=1}^{n}\alpha_i\log(n\alpha_i) - \alpha_i + \frac{1}{2\lambda_n}\|\sum_{i=1}^{n}\alpha_i\Phi(x_i) - \frac{1}{n}\sum_{j=1}^{n}\Phi(y_j)\|_{\mathcal{H}}^2.\end{aligned}$$

Letting $\hat{\alpha}$ be the solution of the above convex program, the KL divergence can be estimated by:

$$\hat{D}_K = 1 + \sum_{i=1}^{n}\hat{\alpha}_i\log\hat{\alpha}_i + \hat{\alpha}_i\log\frac{n}{e}.$$

6.8 Proof of Theorem 6.12

We now sketch out the proof of the main theorem. The key to our analysis is the following lemma:

Lemma 6.14. *If $\hat{g}_n$ is an estimate of g using* (6.28)*, then:*

$$\frac{1}{4}h_{\mathbb{Q}}^2(g_0,\hat{g}_n) + \frac{\lambda_n}{2}I^2(\hat{g}_n) \leq -\int(\hat{g}_n - g_0)d(\mathbb{Q}_n - \mathbb{Q}) + \int 2\log\frac{\hat{g}_n + g_0}{2g_0}d(\mathbb{P}_n - \mathbb{P}) + \frac{\lambda_n}{2}I^2(g_0).$$

Proof. Define $d_l(g_0, g) = \int(g - g_0)d\mathbb{Q} - \log\frac{g}{g_0}d\mathbb{P}$. Note that for $x > 0$, $\frac{1}{2}\log x \leq \sqrt{x} - 1$. Thus,

$$\int \log\frac{g}{g_0}\, d\mathbb{P} \leq 2\int(g^{1/2}g_0^{-1/2} - 1)\, d\mathbb{P}.$$

As a result, for any g, d_l is related to $h_{\mathbb{Q}}$ as follows:

$$\begin{aligned} d_l(g_0, g) &\geq \int (g - g_0)\, d\mathbb{Q} - 2\int (g^{1/2} g_0^{-1/2} - 1)\, d\mathbb{P} \\ &= \int (g - g_0)\, d\mathbb{Q} - 2\int (g^{1/2} g_0^{1/2} - g_0)\, d\mathbb{Q} = \int (g^{1/2} - g_0^{1/2})^2 d\mathbb{Q} \\ &= 2h^2_{\mathbb{Q}}(g_0, g). \end{aligned}$$

By the definition (6.28) of our estimator, we have:

$$\int \hat{g}_n d\mathbb{Q}_n - \int \log \hat{g}_n d\mathbb{P}_n + \frac{\lambda_n}{2} I^2(\hat{g}_n) \leq \int g_0 d\mathbb{Q}_n - \int \log g_0 d\mathbb{P}_n + \frac{\lambda_n}{2} I^2(g_0).$$

Both sides are convex functionals of g. By Jensen's inequality, if F is a convex function, then $F((u+v)/2) - F(v) \leq (F(u) - F(v))/2$. We obtain:

$$\int \frac{\hat{g}_n + g_0}{2} d\mathbb{Q}_n - \int \log \frac{\hat{g}_n + g_0}{2} d\mathbb{P}_n + \frac{\lambda_n}{4} I^2(\hat{g}_n) \leq \int g_0 d\mathbb{Q}_n - \int \log g_0 d\mathbb{P}_n + \frac{\lambda_n}{4} I^2(g_0).$$

Rearranging, $\int \frac{\hat{g}_n - g_0}{2} d(\mathbb{Q}_n - \mathbb{Q}) - \int \log \frac{\hat{g}_n + g_0}{2g_0} d(\mathbb{P}_n - \mathbb{P}) + \frac{\lambda_n}{4} I^2(\hat{g}_n) \leq$

$$\begin{aligned} \int \log \frac{\hat{g}_n + g_0}{2g_0} d\mathbb{P} - \int \frac{\hat{g}_n - g_0}{2} d\mathbb{Q} + \frac{\lambda_n}{4} I^2(g_0) &= -d_l(g_0, \frac{g_0 + \hat{g}_n}{2}) + \frac{\lambda_n}{4} I^2(g_0) \\ &\leq -2h^2_{\mathbb{Q}}(g_0, \frac{g_0 + \hat{g}_n}{2}) + \frac{\lambda_n}{4} I^2(g_0) \leq -\frac{1}{8} h^2_{\mathbb{Q}}(g_0, \hat{g}_n) + \frac{\lambda_n}{4} I^2(g_0), \end{aligned}$$

where the last inequality is a standard result for the (generalized) Hellinger distance (cf. [van de Geer, 1999]). □

Let us now proceed to part (a) of the theorem. Define $f_g := \log \frac{g + g_0}{2g_0}$, and let $\mathcal{F}_M := \{f_g | g \in \mathcal{G}_M\}$. Since f_g is a Lipschitz function of g, conditions (6.30) and (6.32) imply that

$$\mathcal{H}^B_\delta(\mathcal{F}_M, L_2(\mathbb{P})) = O(M/\delta)^\gamma. \tag{6.35}$$

Apply Lemma 6.18 (see the Appendix) using distance metric $d_2(g_0, g) = \|g - g_0\|_{L_2(\mathbb{Q})}$, the following is true under $\mathbb{Q}$ (and so true under $\mathbb{P}$ as well, since $d\mathbb{P}/d\mathbb{Q}$ is bounded from above),

$$\sup_{g \in \mathcal{G}} \frac{|\int (g - g_0) d(\mathbb{Q}_n - \mathbb{Q})|}{n^{-1/2} d_2(g_0, g)^{1-\gamma/2} (1 + I(g) + I(g_0))^{\gamma/2} \vee n^{-\frac{2}{2+\gamma}} (1 + I(g) + I(g_0))} = O_{\mathbb{P}}(1) \tag{6.36}$$

In the same vein, we obtain that under $\mathbb{P}$ measure:

$$\sup_{g \in \mathcal{G}} \frac{|\int f_g d(\mathbb{P}_n - \mathbb{P})|}{n^{-1/2} d_2(g_0, g)^{1-\gamma/2}(1 + I(g) + I(g_0))^{\gamma/2} \vee n^{-\frac{2}{2+\gamma}}(1 + I(g) + I(g_0))} = O_{\mathbb{P}}(1) \tag{6.37}$$

By condition (6.31), it is easy to see that:

$$d_2(g_0, g) = \|g - g_0\|_{L_2(\mathbb{Q})} \leq 2c^{1/2}(1 + I(g) + I(g_0))^{1/2} h_{\mathbb{Q}}(g_0, g).$$

Combining Lemma 6.14 and Eqs. (6.37), (6.36), we obtain the following:

$$\begin{aligned} &\frac{1}{4} h^2_{\mathbb{Q}}(g_0, \hat{g}_n) + \frac{\lambda_n}{2} I^2(\hat{g}_n) \leq \lambda_n I(g_0)^2/2 + \\ &\quad O_{\mathbb{P}}\left(n^{-1/2} h_{\mathbb{Q}}(g_0, g)^{1-\gamma/2}(1 + I(g) + I(g_0))^{1/2+\gamma/4} \ \vee \ n^{-\frac{2}{2+\gamma}}(1 + I(g) + I(g_0)) \right). \end{aligned} \tag{6.38}$$

From this point, the proof involves simple algebraic manipulation of (6.38). To simplify notation, let $\hat{h} = h_{\mathbb{Q}}(g_0, \hat{g}_n)$, $\hat{I} = I(\hat{g}_n)$, and $I_0 = I(g_0)$. There are four possibilities:

Case a. $\hat{h} \geq n^{-1/(2+\gamma)}(1 + \hat{I} + I_0)^{1/2}$ and $\hat{I} \geq 1 + I_0$. From (6.38), either

$$\hat{h}^2/4 + \lambda_n \hat{I}^2/2 \leq O_{\mathbb{P}}(n^{-1/2}) \hat{h}^{1-\gamma/2} \hat{I}^{1/2+\gamma/4} \ \text{ or } \ \hat{h}^2/4 + \lambda_n \hat{I}^2/2 \leq \lambda_n I_0^2/2,$$

which implies, respectively, either

$$\hat{h} \leq \lambda_n^{-1/2} O_{\mathbb{P}}(n^{-2/(2+\gamma)}), \quad \hat{I} \leq \lambda_n^{-1} O_{\mathbb{P}}(n^{-2/(2+\gamma)}).$$

or

$$\hat{h} \leq O_{\mathbb{P}}(\lambda_n^{1/2} I_0), \quad \hat{I} \leq O_{\mathbb{P}}(I_0).$$

Both scenarios conclude the proof if we set $\lambda_n^{-1} = O_{\mathbb{P}}(n^{2/(\gamma+2)}(1 + I_0))$.

Case b. $\hat{h} \geq n^{-1/(2+\gamma)}(1 + \hat{I} + I_0)^{1/2}$ and $\hat{I} < 1 + I_0$. From (6.38), either

$$\hat{h}^2/4 + \lambda_n \hat{I}^2/2 \leq O_{\mathbb{P}}(n^{-1/2}) \hat{h}^{1-\gamma/2}(1 + I_0)^{1/2+\gamma/4} \ \text{ or } \ \hat{h}^2/4 + \lambda_n \hat{I}^2/2 \leq \lambda_n I_0^2/2,$$

which implies, respectively, either

$$\hat{h} \leq (1 + I_0)^{1/2} O_{\mathbb{P}}(n^{-1/(\gamma+2)}), \quad \hat{I} \leq 1 + I_0$$

or

$$\hat{h} \leq O_{\mathbb{P}}(\lambda_n^{1/2} I_0), \quad \hat{I} \leq O_{\mathbb{P}}(I_0).$$

Both scenarios conclude the proof if we set $\lambda_n^{-1} = O_{\mathbb{P}}(n^{2/(\gamma+2)}(1 + I_0))$.

Case c. $\hat{h} \leq n^{-1/(2+\gamma)}(1+\hat{I}+I_0)^{1/2}$ and $\hat{I} \geq 1+I_0$. From (6.38)

$$\hat{h}^2/4 + \lambda_n \hat{I}^2/2 \leq O_{\mathbb{P}}(n^{-2/(2+\gamma)})\hat{I},$$

which implies that $\hat{h} \leq O_{\mathbb{P}}(n^{-1/(2+\gamma)})\hat{I}^{1/2}$ and $\hat{I} \leq \lambda_n^{-1} O_{\mathbb{P}}(n^{-2/(2+\gamma)})$. This means that

$$\hat{h} \leq O_{\mathbb{P}}(\lambda_n^{1/2})(1+I_0), \quad \hat{I} \leq O_{\mathbb{P}}(1+I_0)$$

if we set $\lambda_n^{-1} = O_{\mathbb{P}}(n^{2/(2+\gamma)})(1+I_0)$.

Case d. $\hat{h} \leq n^{-1/(2+\gamma)}(1+\hat{I}+I_0)^{1/2}$ and $\hat{I} \leq 1+I_0$. Part (a) of the theorem is immediate.

Finally, part (b) is a simple consequence of part (a) using the same argument as in Thm 6.9.

6.9 Simulation results

In this section, we describe the results of various simulations that demonstrate the practical viability of our estimators, as well as their convergence behavior. We experimented with our estimators using various choices of $\mathbb{P}$ and $\mathbb{Q}$, including Gaussian, beta, mixture of Gaussians, and multivariate Gaussian distributions. Here we report results in terms of KL estimation error. For each of the eight estimation problems described here, we experiment with increasing sample sizes (the sample size, n, ranges from 100 to 10^4 or more). Error bars are obtained by replicating each set-up 250 times.

For all simulations, we report our estimator's performance using the the simple fixed rate $\lambda_n \sim 1/n$, noting that this may be a suboptimal rate. We set the kernel width to be relatively small ($\sigma = .1$) for one-dimension data, and larger σ for higher dimensions. We use M1 to denote the method in which $\mathcal{G}$ is the RKHS, and M2 for the method in which $\log \mathcal{G}$ is the RKHS. Our methods are compared to algorithm A in Wang et al [Wang *et al.*, 2005], which was shown empirically to be one of the best methods in the literature. Their method, to be denoted by WKV, is based on data-dependent partitioning of the covariate space. Naturally, the performance of WKV is critically dependent on the amount s of data allocated to each partition; here we report results with $s \sim n^\gamma$, where $\gamma = 1/3, 1/2, 2/3$.

The first four plots present results with univariate distributions. In the first two, our estimators $M1$ and $M2$ appear to have faster convergence rate than WKK. The WKV estimator performs very well in the third example, but rather badly in the fourth example. The next four plots present results with two and three dimensional data. Again, M1 has the best convergence rates in all examples. The M2 estimator does not converge in the last example, suggesting that the underlying function class exhibits very strong bias. WKV have weak convergence rates despite different choices of the partition sizes. It is worth noting that as one increases the number of dimensions, histogram based methods such as WKV become

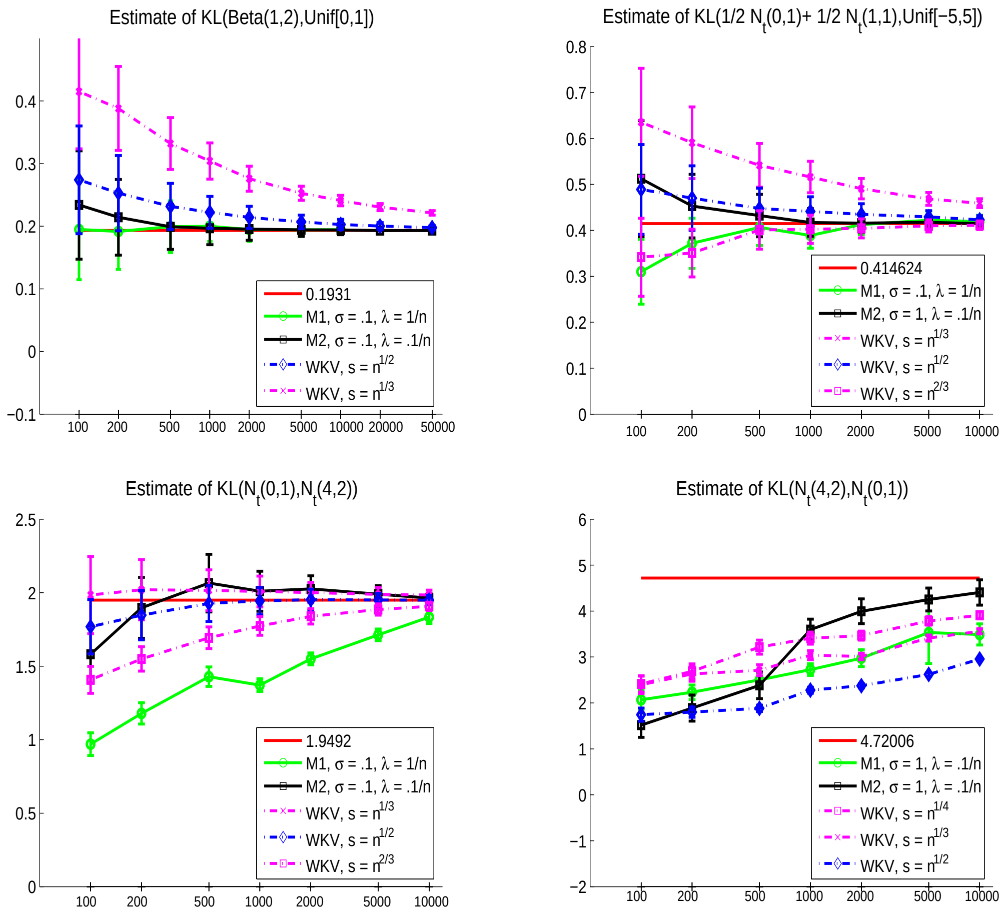


Figure 6.1. Results of estimating KL divergences for various choices of probability distributions. In all plots, the X-axis is the number of data points plotted on a log scale, and the Y-axis is the estimated value. The error bar is obtained by replicating the experiment 250 times. $N_t(a, I_k)$ denotes a truncated normal distribution of k dimensions with mean $(a, \ldots, a)$ and identity covariance matrix.

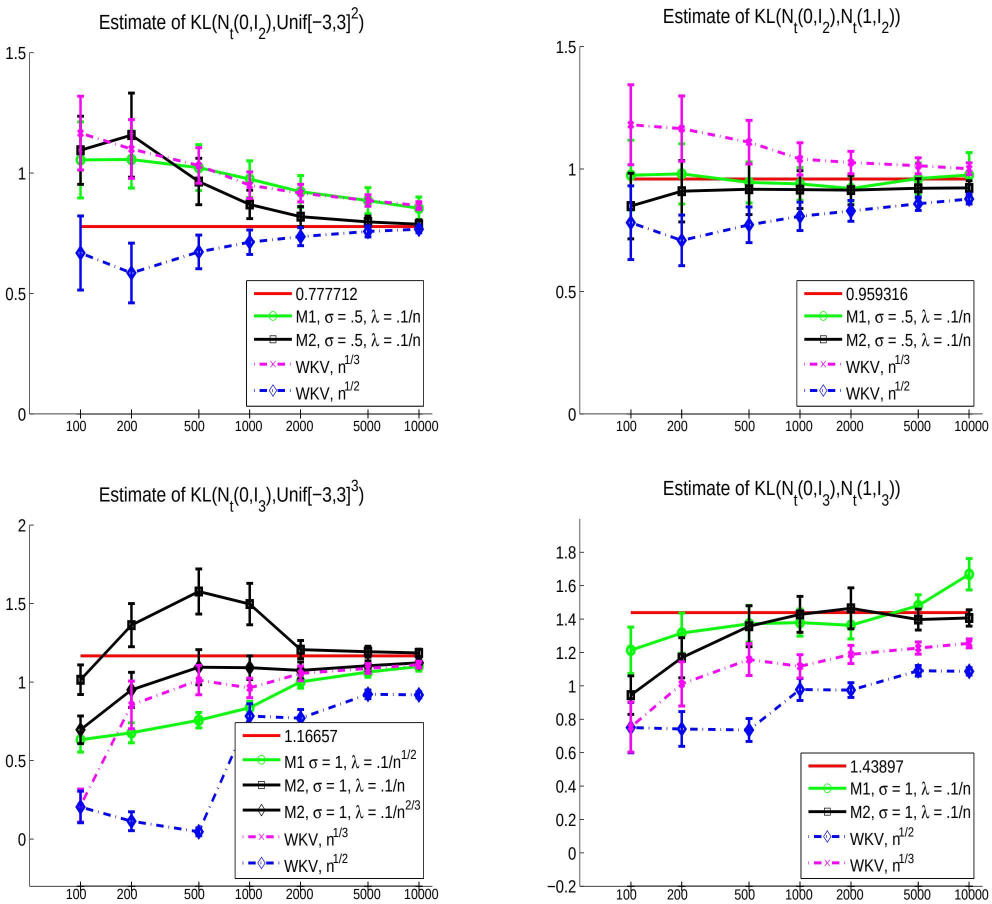


Figure 6.2. Results of estimating KL divergences for various choices of probability distributions. In all plots, the X-axis is the number of data points plotted on a log scale, and the Y-axis is the estimated value. The error bar is obtained by replicating the experiment 250 times. $N_t(a, I_k)$ denotes a truncated normal distribution of k dimensions with mean $(a, \dots, a)$ and identity covariance matrix.

increasingly difficult to implement, whereas increasing dimension has only a mild effect on our method.

Appendix 6.A Increments of empirical processes

In this section we summarize several key theorems from empirical process theory that have been used in the proof of theorems in this chapter.

The following theorem (Thm 3.7 in [van de Geer, 1999]) specifies conditions under which the supremum of an empirical process goes to 0 almost surely.

Theorem 6.15. *Let G be the envelope function for $\mathcal{G}$. Assume that $\int G d\mathbb{P} < \infty$, and suppose that for any $\delta > 0$, $\frac{1}{n}\mathcal{H}_\delta(\mathcal{G}, L_1(\mathbb{P}_n)) \xrightarrow{\mathcal{P}} \mathbb{P}0$, then $\sup_{g\in\mathcal{G}} \int g d(\mathbb{P}_n - \mathbb{P}) \xrightarrow{a.s.} 0$.*

Next, this is a result on the convergence rate of the supremum of an empirical process (Thm 5.11 in [van de Geer, 1999]):

Theorem 6.16. *Let K, R be some constants, $\mathcal{G}$ satisfy $\sup_{g\in\mathcal{G}} \rho_K(g) \leq R$. If there hold, for some sufficiently large universal constant C:*

$$\begin{aligned} a &\leq C_1\sqrt{n}R^2/K \\ a &\geq C_0\left(\int_0^R \mathcal{H}_u^B(\mathcal{G}, \rho_K)^{1/2} du \vee R\right) \\ C_0^2 &\geq C^2(C_1+1), \end{aligned}$$

then

$$P\left(\sup_{g\in\mathcal{G}} |\sqrt{n}\int g d(\mathbb{P}_n - \mathbb{P})| \geq a\right) \leq C\exp\left[-\frac{a^2}{C^2(C_1+1)R^2}\right].$$

Finally, we shall present several results on the modulus of continuity of the supremum of empirical processes. Consider a uniformly bounded class of functions, say:

$$\sup_{g\in\mathcal{G}} |g - g_0|_\infty \leq 1. \tag{6.39}$$

Assume more over that

$$\mathcal{H}_\delta^B(\mathcal{G}, L_2(\mathbb{P})) \leq A\delta^{-\alpha}, \text{ for all } \delta > 0, \tag{6.40}$$

for some constant $0 < \alpha < 2$, and some constant A. A direct consequence of Lemma 5.13 of [van de Geer, 1999] is the following:

Lemma 6.17. *Assume (6.39) and (6.40), then as $n \to \infty$, for $\delta_n = n^{-1/(2+\alpha)}$ there holds:*

$$\sup_{g\in\mathcal{G}} \frac{\sqrt{n}|\int(g-g_0)d(\mathbb{P}_n - \mathbb{P})|}{\|g-g_0\|_{L_2(\mathbb{P})}^{1-\alpha/2} \vee \sqrt{n}\delta_n^2} = O_{\mathbb{P}}(1).$$

This lemma can be extended to function classes of infinite size (in terms of entropy and other metrics), and proves useful in the analysis of M-estimators with penalties. In the remainder of this appendix we state this result.

Suppose that $\mathcal{G}$ has infinite entropy, but that

$$\mathcal{G} = \cup_{1 \leq M \leq \infty} \mathcal{G}_M, \tag{6.41}$$

where $\mathcal{G}_M := \{g \mid I(g) \leq M\}$ is a ball determined by $I(\cdot)$. Here, we think of $I(g)$ as the complexity of irregularity of the function g (e.g., some Sobolev or Besov norm).

We shall present a result of modulus of continuity of the (infinite) function class $\mathcal{G}$ in terms of a general distance function $d(\cdot, \cdot)$ such that:

$$\|g - g_0\|_{L_2(\mathbb{P})} \leq d(g, g_0). \tag{6.42}$$

Our result is also applicable if the above condition holds up to a multiplicative constant.

The following conditions will be used: there exist constants $0 < \alpha < 2,\ 0 \leq \beta \leq 1,\ c_0 > 0$ and $A > 0$ such that for all $M \geq 1$,

$$\sup_{g \in \mathcal{G}_M} d(g, g_0) \leq c_0 M. \tag{6.43}$$

$$\mathcal{H}^B_\delta(\mathcal{G}_M, L_2(\mathbb{P})) \leq A \left(\frac{M}{\delta}\right)^\alpha. \tag{6.44}$$

$$\sup_{g \in \mathcal{G}_M, d(g, g_0) \leq \delta} |g - g_0|_\infty \leq (c_0 \delta)^\beta M^{1-\beta}, \text{ for all } \delta > 0. \tag{6.45}$$

Now we are ready to state Lemma 5.14 of [van de Geer, 1999]:

Lemma 6.18. *Assume (6.42), (6.43), (6.44), (6.45). Then, for some constants c and n_0 depdending on α, β, c_0 and A, we have for all $T \geq c$ and $n \geq n_0$,*

$$\mathbb{P}\left(\sup_{g \in \mathcal{G},\ d(g, g_0) \leq n^{-\frac{1}{2+\alpha-2\beta}} I(g)} \frac{|\int (g - g_0) d(\mathbb{P}_n - \mathbb{P})|}{I(g)} \geq T n^{-\frac{2-\beta}{2+\alpha-2\beta}}\right) \leq c \exp\left[-\frac{T n^{\frac{\alpha}{2+\alpha-2\beta}}}{c^2}\right].$$

Moreover, for $T \geq c, n \geq n_0$,

$$\mathbb{P}\left(\sup_{g \in \mathcal{G},\ d(g, g_0) > n^{-\frac{1}{2+\alpha-2\beta}} I(g)} \frac{\sqrt{n} |\int (g - g_0) d(\mathbb{P}_n - \mathbb{P})|}{d(g, g_0)^{1-\alpha/2} I(g)^{\alpha/2}} \geq T\right) \leq c \exp\left[-\frac{T}{c^2}\right].$$

Chapter 7

Conclusions and suggestions

In this thesis we have investigated several settings of decision-making in decentralized systems. Our main contributions can be summarized as follows:

- a nonparametric approach to centralized detection estimation tasks and its application to the problem of localization in ad hoc sensor network
- a nonparametric aproach to decentralized detection problem
- a characterization of optimal decision rules of sequential decentralized detection problem
- a characterization of the correspondence between surrogate loss and divergence functionals.
- a nonparametric estimation method for divergence functionals and the likelihood ratio

There are a number of issues and open questions arising from this thesis. In the following we shall outline several of such issues, and in some cases suggest possible avenues of attack.

7.1 Tradeoff between quantization rates and statistical error

In Chapter 3 we considered the problem of learning local quantization rules and global decision rule so as to minimize the detection error. The quantization rules are constrained by the number of bits (which is decided *a priori*) to be transmitted by each sensor. There are two key quantities, the communication constraint and the statistical efficiency, whose interplay is of interest. From a practical viewpoint, it is useful for a designer to specify a priori a desirable level of detection error, based from which communication constraints are set and the quantization rules are learned. Thus, one key issue here is to study the tradeoff between the number of bits allowed and the optimal detection error.

In the setting of binary classification, in Chapter 4 we have shown that the optimal detection error is equal to the corresponding f-divergence between the two distributions underlying the binary hypotheses. As a result, the relationship between detection error and bit constraints hinges on the approximation error rate of the f-divergence, since the communication constraints (especially for continuous data) essentially amount to an approximation method using step functions. It appears that results from approximation theory [DeVore and Lorentz, 1993] can be applied. Furthermore, key properties such as the subadditivity of certain f-divergences (e.g., the KL divergence and log-sum inequality, cf. [Cover and Thomas, 1991]) could be exploited (for an example, see [Birman and Solomjak, 1967]).

We could consider alternative routes that are more amenable to the analysis. For instance, the class of quantization rules can be specified up front, and the tradeoff between the quantization rates and the statistical error rate can be studied within this class of quantization rules. Although the optimal quantization rules do not necessarily lie within the specified class, the loss might be negligible in practice, and searching for the optimal rules within the specified class might be a more tractable task. This is the approach taken by [Huang *et al.*, 2007] in the context of an anomaly detection method using principle component analysis. In this work the quantization rules are simple rules based on thresholding the data magnitude. By studying the effects of approximating the covariance matrix the authors are able to characterize the tradeoff between the number of bits and the anomaly detection error. It is of interest to extend this approach to other settings such as classification and regression.

7.2 Nonparametric estimation in sequential detection setting

In Chapter 5 we have studied the sequential setting of the decentralized detection problem. We have obtained results on the characterization of (asymptotically) optimal quantization rules. One key issue is: How can the quantization rules be learned for a decentralized system? In a parametric setting, where the (binary) hypotheses are assumed to be known, the asymptotic formulae given in Lemma 5.1 provides a method for computing the quantization rules by minimizing over a sum of the inverse of two KL divergences. Thus, it is of substantial interest to come up with efficient algorithms for optimizing divergence functionals over a class of quantization rules.

In the nonparametric setting, the underlying distributions are typically not known, but assumed to be within certain function classes. Results in Chapter 5 can be used to justify our focus to classes of stationary quantizers. The objective functional involves the KL divergences, which can be estimated using the nonparametric method developed in Chapter 6. It would be interesting to explore efficient algorithms for learning quantizer rules by optimizing such objective functional.

While the main focus of the thesis is in binary sequential hypothesis testing, it is a promising direction to consider various other statistical tasks (e.g, point estimation, regression, dependence testing) in a decentralized system. For such different tasks, it is promising to consider other statistical functionals (other than f-divergence functionals) and accompanying nonparametric sequential procedures [Sen, 1981]. Many such procedures are not of the M-estimation type, and are potentially more tractable from a computational viewpoint.

7.3 Multiple dependent decentralized subsystems

Throughout the thesis, we considered a decentralized system that consists of multiple measurements (collected by local monitoring devices) and a global fusion center aggregating the local measurements. From the viewpoint of each monitoring device, the processing is distributed, but the whole system is coordinated centrally by the fusion center to solve a *single* statistical task (e.g., detection, estimation, etc). In practice, we may be given a distributed architecture in which there are *multiple* statistical tasks that partially share the set of measurements. In other words, each statistical task corresponds to a subsystem within a decentralized system. It is an important problem to devise distributed protocol that facilitate the performance for the dependent statistical tasks in a computationally efficient manner.

For concreteness, consider the following application. There are a number of sensors placed in a geographical area such as highway or building. We are interested in finding sequential procedures for detecting the failure of these sensors. Typically, only one sensor fails at a time, and so a reasonable statistic to be exploited is the correlations among neighboring sensors. If there is a change in the distribution of the correlation between two neighboring sensors, then at least one of them must have probably failed. Thus, we can have a set-up involving multiple sequential detection problems, each of which is concerned with the status of one sensor. Furthermore, these detection problems are dependent because they have shared measurements.

For simplicity, suppose that we have two sensors, which can be measured by two separate covariates X and Y. Furthermore, these two sensors share a measurement Z. More formally, given three sequences of i.i.d. data $(X_1, X_2, \ldots)$, $(Y_1, Y_2, \ldots)$, $((Z_1, Z_2, \ldots))$. At each time point i we receive the triple (X_i, Y_i, Z_i). If sensor 1 fails at time point k_1, then there is a change in distribution for X_i, $i \geq k_1$ from f_0 to f_1. Similarly, if sensor 2 fails at k_2, then there is a change in distribution for Y_i, $i \geq k_2$ from g_0 to g_1. The change in distribution for sequence Z_i happens at $\min(k_1, k_2)$ from h_0 to h_1. We are interested in a sequential and decentralized detection procedure for sensor 1, i.e., a stopping time $\nu_1(X, Z)$, based on the sequences X and Z, and a sequential and decentralized procedure for sensor 2, i.e., a stopping time $\nu_2(Y, Z)$, based on the sequences Y and Z so as to minimize the delay of the detection of the respective change-points, while maintaining an upper bound α on the false alarm rates.

One could treat these two sequential change-point detection problem as separate, ignoring the shared sequence Z. In a collaboration with Ram Rajagopal, we proposed a decentralized sequential detection method that involve sharing information between the two sensors. Specifically, the two sequential procedures devised for each of two sensors also exploit the information passed by the other sensor. We show that the resulting procedures exhibit shorter detection delay times than the method that treat the two detection problems separately [Rajagopal and Nguyen, 2007]. It is of significant interest to extend this idea to the setting of multiple sensors.

7.4 Minimax rate for divergence estimation

In comparison to the problem of estimating divergence functionals (which are integrals of two densities), the problem of estimating integrals of a single density has been studied more extensively by [Bickel and Ritov, 1988; Donoho and Liu, 1991; Birgé and Massart, 1995; Laurent, 1996] and others.

Let ϕ be a smooth function of one variable, and f belongs to some class of d-dim densities of smoothness α. Then, the optimal minimax rate for estimating $T(f) = \int \phi(f)$ is $n^{-1/2}$ when $\alpha > d/4$, and $n^{-4\alpha/(4\alpha+d)}$ when $\alpha \geq d/4$.

In the case of divergences of the form $\int \phi(f, g)$, it is of interest to find the optimal minimax rate. By fixing a density, say g, to be known, the minimax rate of an integral of two densities cannot be better than that of one density. However, does there still exist a critical threshold of smoothness for both f and g above which the integral can be estimated at the semiparametric rate $n^{-1/2}$? It seems that such a threshold does exist for integrals of two densities.[1] Note that our estimation method developed in Chapter 6 yields only the rate of $n^{-2\alpha/(2\alpha+d)}$, which is always worse than $n^{-1/2}$. This is perhaps due to the fact that our estimator is essentially a linear estimator.

[1] Personal communication with Peter Bickel, who suggested a suite of estimation methods studied in [Bickel *et al.*, 1998].

7.5 Connection to dimensionality reduction and feature selection

Both decentralized detection problem and dimensionality reduction or feature selection problem can be viewed as instances of an experiment design problem: In a decentralized detection problem, the design is the quantization rules applied across dimensions of data; in a dimensionality reduction problem, the design is the transformation of the original data to lower dimensional data; in a feature selection problem, the design is a combinatorial choice of a subset of dimensions.

There is a huge literature on the development of efficient algorithms and their analysis in the context of (parametric) linear regression and basis pursuit (see, e.g., [Tibshirani, 1996; Tropp, 2004; Tropp, 2006; Donoho, 2004; Candes and Tao, 2005; Fu and Knight, 2000; Fan and Li, 2001; Fan and Peng, 2004; Wainwright, 2006]), graph structure learning [Meinshausen and Buhlmann, 2006], classification [van de Geer, to appear], and nonparametric regression and density estimation [Lafferty and Wasserman, 2005; Liu *et al.*, 2007]. With the exception of the last two references, the majority of the cited work considered variations of l_1 relaxation method to obtain computationally efficient algorithms with good statistical properties. It would be interesting to exploit the connection of these problems to decentralized detection problems to devise more efficient procedures that overcome the computational intractability of the learning of quantization rules. At the same time, it is interesting to explore potential applications of the correspondence between surrogate losses and divergences to devise and study alternative surrogate losses for the existing feature selection and dimensionality reduction algorithms in the literature.

Bibliography

[Al-Ibrahim and Varshney, 1989] M. M. Al-Ibrahim and P. K. Varshney. Nonparametric sequential detection based on multisensor data. In *Proc. 23rd Annu. Conf. on Inform. Sci. and Syst.*, pages 157–162, 1989.

[Ali and Silvey, 1966] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *J. Royal Stat. Soc. Series B*, 28:131–142, 1966.

[Aronszajn, 1950] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[Arrow *et al.*, 1949] K. J. Arrow, D. Blackwell, and M. A. Girshick. Bayes and minimax solutions of sequential decision problems. *Econometrica*, 17(3/4):213–244, 1949.

[Bahl and Padmanabhan, 2000] P. Bahl and V. N. Padmanabhan. RADAR: An in-building RF-based user location and tracking system. In *INFOCOM*, pages 775–784, 2000.

[Barron, 1993] A. Barron. Universal approximation bounds for superpositions of a sigmoid function. *IEEE Transactions on Information Theory*, 39:930–945, 1993.

[Bartlett and Mendelson, 2002] P. Bartlett and S. Mendelson. Gaussian and Rademacher complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[Bartlett *et al.*, 2006] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006.

[Bartlett, 1998] Peter L. Bartlett. The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998.

[Bertsekas, 1995a] D. P. Bertsekas. *Dynamic Programming and Stochastic Control*, volume 1. Athena Scientific, Belmont, MA, 1995.

[Bertsekas, 1995b] D.P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 1995.

[Bickel and Doksum, 2006] P. Bickel and K. Doksum. *Mathematical statistics*. Prentice Hall, second edition, 2006.

[Bickel and Ritov, 1988] P. Bickel and Y. Ritov. Estimating integrated squared density derivatives: Sharp best order of convergence estimates. *Sankhy$\bar{a}$ Ser. A*, 50:381–393, 1988.

[Bickel *et al.*, 1998] P. Bickel, C. Klaassen, R. Ya'acov, and J. Wellner. *Efficient and adaptive estimation for semiparametric models*. Springer, first edition, 1998.

[Birgé and Massart, 1995] L. Birgé and P. Massart. Estimation of integral functionals of a density. *Ann. Statist.*, 23(1):11–29, 1995.

[Birman and Solomjak, 1967] M. S. Birman and M. Z. Solomjak. Piecewise-polynomial approximations of functions of the classes W_p^α. *Math. USSR-Sbornik*, 2(3):295–317, 1967.

[Bishop, 1995] C. Bishop. *Neural networks for pattern recognition*. Clarendon Press, Oxford, 1995.

[Blackwell, 1951] D. Blackwell. Comparison of experiments. *Proceeding of 2nd Berkeley Symposium on Probability and Statistics*, 1:93–102, 1951.

[Blackwell, 1953] D. Blackwell. Equivalent comparisons of experiments. *Annals of Statistics*, 24(2):265–272, 1953.

[Blake and Merz, 1998] C.L. Blake and C.J. Merz. UCI repository of machine learning databases, 1998.

[Blum *et al.*, 1997] R. S. Blum, S. A. Kassam, and H. V. Poor. Distributed detection with multiple sensors: Part ii—advanced topics. *Proceedings of the IEEE*, 85:64–79, January 1997.

[Boyd and Vandenberghe, 2004] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[Bradt and Karlin, 1956] R. Bradt and S. Karlin. On the design and comparison of certain dichotomous experiments. *Annals of Statistics*, 27(2):390–409, 1956.

[Breiman, 1998] L. Breiman. Arcing classifiers. *Annals of Statistics*, 26:801–824, 1998.

[Broniatowski and Keziou, 2004] M. Broniatowski and A. Keziou. Parametric estimation and tests through divergences. Technical report, LSTA, Université Pierre et Marie Curie, 2004.

[Bulusu *et al.*, 2000] N. Bulusu, J. Heidemann, and D. Estrin. GPS-less low cost outdoor localization for very small devices. Technical Report 00-729, Computer Science Department, University of Southern California, 2000.

[Candes and Tao, 2005] E. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51:4203–4215, 2005.

[Chaloner and Verdinelli, 1995] K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10:273–304, 1995.

[Chamberland and Veeravalli, 2003] J. F. Chamberland and V. V. Veeravalli. Decentralized detection in sensor networks. *IEEE Transactions on Signal Processing*, 51(2):407–416, 2003.

[Chen *et al.*, 2006] B. Chen, L. Tong, and P. K. Varshney. Channel aware distributed detection in wireless sensor networks. *IEEE Signal Processing Magazine*, 23(4):16–26, July 2006.

[Chernoff, 1959] H. Chernoff. Sequential design of experiments. *Annals of Statistics*, 30(3):755–770, 1959.

[Chernoff, 1972] H. Chernoff. *Sequential Analysis and Optimal Design*. CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM Press, 1972.

[Chong and Kumar, 2003] C. Chong and S. P. Kumar. Sensor networks: Evolution, opportunities, and challenges. *Proceedings of the IEEE*, 91:1247–1256, 2003.

[Cormode and Garofalakis, 2005] Graham Cormode and Minos Garofalakis. Sketching streams through the net: Distributed approximate query tracking. In *Proceedings of VLDB*, pages 13–24, 2005.

[Cortes and Vapnik, 1995] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[Cover and Thomas, 1991] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, 1991.

[Csiszaŕ, 1967] I. Csiszaŕ. Information-type measures of difference of probability distributions and indirect observation. *Studia Sci. Math. Hungar*, 2:299–318, 1967.

[D'Costa and Sayeed, 2003] A. D'Costa and A. Sayeed. Collaborative signal processing for distributed classification in sensor networks. In *2nd International Workshop on Information Processing in Sensor Networks (IPSN)*, pages 193–208, 2003.

[DeVore and Lorentz, 1993] R. DeVore and G. G. Lorentz. *Constructive Approximation*. Springer-Verlag, 1993.

[Donoho and Liu, 1991] D. Donoho and R. Liu. Geometrizing rates of convergence II. *Annals of Statistics*, 19:633–667, 1991.

[Donoho, 2004] D. Donoho. For most large underdetermined systems of equations, the minimal l_1-norm near-solution approximates the sparsest near-solution. Technical report, Statistics Department, Stanford University, 2004.

[Dougherty *et al.*, 1995] J. Dougherty, R. Kohavi, and M. Sahami. Supervised and unsupervised discretization of continuous features. In *Proceedings of the ICML*, 1995.

[Duda *et al.*, 2000] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley, New York, second edition, 2000.

[Durrett, 1995] R. Durrett. *Probability: Theory and Examples*. Duxbury Press, second edition, 1995.

[Fan and Li, 2001] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.

[Fan and Peng, 2004] J. Fan and H. Peng. Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics*, 32:928–961, 2004.

[Ford *et al.*, 1989] I. Ford, C.P. Kitsos, and D.M. Titterington. Recent advances in nonlinear experiment designs. *Technometrics*, 31:49–60, 1989.

[Freund and Schapire, 1997] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[Friedman *et al.*, 2000] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, 28:337–374, 2000.

[Fu and Knight, 2000] W. Fu and K. Knight. Asymptotics for lasso type estimators. *Annals of Statistics*, 28:1356–1378, 2000.

[Fukunaga, 1990] K. Fukunaga. *Statistical Pattern Recognition*. Academic Press, 1990.

[Girod and Estrin, 2001] L. Girod and D. Estrin. Robust range estimation using acoustic and multimodal sensing. In *IEEE/RSI International Conference on Intelligent Robots and Systems (IROS)*, 2001.

[Goel and DeGroot, 1979] P. Goel and M. DeGroot. Comparisons of experiments and information measures. *Annals of Statistics*, 7(2):1066–1077, 1979.

[Gyorfi and van der Meulen, 1987] L. Gyorfi and E.C. van der Meulen. Density-free convergence properties of various estimators of entropy. *Computational Statistics and Data Analysis*, 5:425–436, 1987.

[Hall and Morton, 1993] P. Hall and S. Morton. On estimation of entropy. *Ann. Inst. Statist. Math.*, 45(1):69–88, 1993.

[Han *et al.*, 1990] J. Han, P. K. Varshney, and V. C. Vannicola. Some results on distributed nonparametric detection. In *Proc. 29th Conf. on Decision and Control*, pages 2698–2703, 1990.

[Hastie *et al.*, 2001] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer, 2001.

[Hightower and Borriello, 2000] J. Hightower and G. Borriello. Real-time error in location modeling for ubiquitous computing. In *Location Modeling for Ubiquitous Computing—Ubicomp 2001 Workshop Proceedings*, pages 21–27, 2000.

[Hiriart-Urruty and Lemaréchal, 2001] J. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer, 2001.

[Ho, 1980] Y. C. Ho. Team decision problems and information structures. *Proceedings of the IEEE*, 68:644–654, 1980.

[Huang *et al.*, 2007] L. Huang, X. Nguyen, M. Garofalakis, J. Hellerstein, A. Josepth, M. I. Jordan, and N. Taft. Communication-efficient online detection of network-wide anomalies. In *Proc. of 26th IEEE INFOCOM*, May 2007.

[Hussaini *et al.*, 1995] E. K. Hussaini, A. A. M. Al-Bassiouni, and Y. A. El-Far. Decentralized CFAR signal detection. *Signal Processing*, 44:299–307, 1995.

[Hyvarinen *et al.*, 2001] A. Hyvarinen, J. Karhunen, and E. Oja. *Independent component analysis*. John Wiley & Sons, Inc, 2001.

[Ibragimov and Khasminskii, 1978] I. A. Ibragimov and R. Z. Khasminskii. On the nonparametric estimation of functionals. In *Symposium in Asymptotic Statistics*, pages 41–52, 1978.

[Jaakkola and Haussler, 1999] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems 11*, Cambridge, MA, 1999. MIT Press.

[Jiang, 2004] W. Jiang. Process consistency for adaboost. *Annals of Statistics*, 32:13–29, 2004.

[Joe, 1989] H. Joe. Estimation of entropy and other functionals of a multivariate density. *Ann. Inst. Statist. Math.*, 41:683–697, 1989.

[Kailath, 1967] T. Kailath. The divergence and Bhattacharyya distance measures in signal selection. *IEEE Trans. on Communication Technology*, 15(1):52–60, 1967.

[Kailath, 1971] T. Kailath. RKHS approach to detection and estimation problems—Part I: Deterministic signals in Gaussian noise. *IEEE Trans. Info. Theory.*, 17:530–549, 1971.

[Kassam, 1993] S. A. Kassam. Nonparametric signal detection. In *Advances in Statistical Signal Processing*. JAI Press, 1993.

[Keziou, 2003] A. Keziou. Dual representation of ϕ-divergences and applications. *C. R. Acad. Sci. Paris, Ser. I 336*, pages 857–862, 2003.

[Koltchinskii and Panchenko, 2002] V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30:1–50, 2002.

[Lafferty and Wasserman, 2005] John Lafferty and Larry Wasserman. Rodeo: Sparse nonparametric regression in high dimensions, 2005.

[Lai, 2001] T. L. Lai. Sequential analysis: Some classical problems and new challenges (with discussion). *Statist. Sinica*, 11:303–408, 2001.

[Lanckriet *et al.*, 2004] G. Lanckriet, N. Cristianini, L. El Ghaoui, P. Bartlett, and M.I. Jordan. Learning the kernel matrix with semi-definite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

[Laurent, 1996] B. Laurent. Efficient estimation of integral functionals of a density. *Ann. Statist.*, 24(2):659–681, 1996.

[Levit, 1978] B. Ya. Levit. Asymptotically efficient estimation of nonlinear functionals. *Problems Inform. Transmission*, 14:204–209, 1978.

[Li *et al.*, 2002] D. Li, K. Wong, Y. Hu, and A. Sayeed. Detection, classification, and tracking of targets. *IEEE Signal Processing Magazine*, 19:17–29, 2002.

[Lindley, 1956] D. V. Lindley. On the measure of the information provided by an experiment. *Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

[Liu *et al.*, 2007] Han Liu, John Lafferty, and Larry Wasserman. Sparse nonparametric density estimation in high dimensions using the rodeo. In *Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2007.

[Longo *et al.*, 1990] M. Longo, T. Lookabaugh, and R. Gray. Quantization for decentralized hypothesis testing under communication contraints. *IEEE Trans. on Information Theory*, 36(2):241–255, 1990.

[Lorden, 1970] G. Lorden. On excess over the boundary. *Annals of Statistics*, 41(2):520–527, 1970.

[Luenberger, 1969] D. G. Luenberger. *Optimization by Vector Space Methods*. Wiley, New York, 1969.

[Lugosi and Vayatis, 2004] G. Lugosi and N. Vayatis. On the bayes-risk consistency of regularized boosting methods. *Annals of Statistics*, 32:30–55, 2004.

[Mannor *et al.*, 2003] S. Mannor, R. Meir, and T. Zhang. Greedy algorithms for classification - consistency, convergence rates and adaptivity. *Journal of Machine Learning Research*, 4:713–741, 2003.

[Massart, 2000] P. Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciences de Toulouse*, IX:245–303, 2000.

[Mei, 2003] Y. Mei. *Asymptotically optimal methods for sequential change-point detection*. PhD thesis, California Institute of Technology, 2003.

[Meinshausen and Buhlmann, 2006] N. Meinshausen and P. Buhlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436–1462, 2006.

[Mitrinović *et al.*, 1993] D. S. Mitrinović, J. E. Pecarić, and A. M. Fink. *Classical and New Inqualities in Analysis*. Kluwer Academic Publishers, 1993.

[Nasipuri and Tantaratana, 1997] A. Nasipuri and S. Tantaratana. Nonparametric distributed detection using Wilcoxon statistics. *Signal Processing*, 57(2):139–146, 1997.

[Nguyen *et al.*, 2005a] X. Nguyen, M. I. Jordan, and B. Sinopoli. A kernel-based learning approach to ad hoc sensor network localization. *ACM Transactions on Sensor Networks*, 1:134–152, 2005.

[Nguyen *et al.*, 2005b] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Nonparametric decentralized detection using kernel methods. *IEEE Transactions on Signal Processing*, 53(11):4053–4066, 2005.

[Nguyen *et al.*, 2005c] X. Nguyen, M. J. Wainwright, and M. I. Jordan. On divergences, surrogate losses and decentralized detection. Technical Report 695, Dept of Statistics, UC Berkeley, October 2005.

[Nguyen *et al.*, 2006] X. Nguyen, M. J. Wainwright, and M. I. Jordan. On optimal quantization rules in sequential decision problems. In *International Symposium on Information Theory (ISIT)*, 2006.

[Nguyen *et al.*, 2007] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Nonparametric estimation of the likelihood ratio and divergence functionals. In *International Symposium on Information Theory (ISIT)*, 2007.

[Olston *et al.*, 2003] C. Olston, J. Jiang, and J. Widom. Adaptive filters for continuous queries over distributed data streams. In *Proc. of SIGMOD*, pages 563–574, 2003.

[Padmanabhan *et al.*, 2005] V. N. Padmanabhan, S. Ramabhadran, and J. Padhye. Netprofiler: Profiling wide-area networks using peer cooperation. In *Proceedings of the Fourth International Workshop on Peer-to-Peer Systems (IPTPS)*, Ithaca, NY, USA, 2005.

[Pollard, 1984] D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.

[Poor and Thomas, 1977] H. V. Poor and J. B. Thomas. Applications of Ali-Silvey distance measures in the design of generalized quantizers for binary decision systems. *IEEE Trans. on Communications*, 25:893–900, 1977.

[Poor, 1994] H. V. Poor. *An Introduction to Signal Detection and Estimation*. Springer-Verlag, New York, NY, 1994.

[Predd *et al.*, 2004] J. Predd, S. Kulkarni, and H. V. Poor. Consistency in models for communication constrained distributed learning. In *Proceedings of the COLT*, pages 442–456, 2004.

[Priyantha *et al.*, 2000] N. Priyantha, A. Chakraborty, and H. Balakrishnan. The Cricket location-support system. In *ACM International Conference on Mobile Computing and Networking*, New York, 2000. ACM Press.

[Pukelsheim, 1993] F. Pukelsheim. *Optimal design of experiments*. Wiley, 1993.

[Rajagopal and Nguyen, 2007] R. Rajagopal and X. Nguyen. Multiple failure decentralized detection. Manuscript, July 2007.

[Rockafellar, 1970] G. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

[Rosenblatt, 1958] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–408, 1958.

[Rumelhart *et al.*, 1986] D. Rumelhart, G. Hinton, and R.Williams. Learning internal representations by error propagation in parallel distributed processing. In *Explorations in the microstructure of cognition*, 1986.

[Saitoh, 1988] S. Saitoh. *Theory of Reproducing Kernels and its Applications*. Longman Scientific & Technical, Harlow, UK, 1988.

[Savarese *et al.*, 2002] C. Savarese, J. Rabaey, and K. Langendoen. Robust positioning algorithms for distributed ad-hoc wireless sensor networks. In *USENIX Annual Technical Conference, Monterey, CA*, pages 317–327, 2002.

[Savvides *et al.*, 2001] A. Savvides, C. Han, and M. Srivastava. Dynamic fine grained localization in ad-hoc sensor networks. In *Proceedings of the Fifth International Conference on Mobile Computing and Networking*, pages 166–179, 2001.

[Schölkopf and Smola, 2002] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[Scott, 1992] D. Scott. *Multivariate density estimation: Theory, practice, and visualization*. Wiley, New York, 1992.

[Seidel and Rappaport, 1992] A. Seidel and T. Rappaport. 914MHz path loss prediction models for indoor wireless communications in multi-floored buildings. *IEEE Transactions on Antennas and Propagation*, 40:207–217, 1992.

[Sen, 1981] P. K. Sen. *Sequential nonparametrics*. Wiley, 1981.

[Shawe-Taylor and Cristianini, 2004] J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge Univ Press, 2004.

[Sheng and Hu, 2003] X. Sheng and Y. Hu. Energy based acoustic source localization. In *2nd International Workshop on Information Processing in Sensor Networks (IPSN)*, pages 285–300, 2003.

[Shiryayev, 1978] A. N. Shiryayev. *Optimal Stopping Rules*. Springer-Verlag, 1978.

[Siegmund, 1985] D. Siegmund. *Sequential Analysis*. Springer-Verlag, 1985.

[Silverman, 1982] B. W. Silverman. On the estimation of a probability density function by the maximum penalized likelihood method. *Annals of Statistics*, 10:795–810, 1982.

[Silverman, 1986] B. Silverman. *Density Estimation for Statistics and data analysis*. Chapman and Hall, London, 1986.

[Steinberg and Hunter, 1985] D. M. Steinberg and W. G. Hunter. Experimental design: review and comment. *Technometrics*, 26:71–97, 1985.

[Steinwart, 2005] I. Steinwart. Consistency of support vector machines and other regularized kernel machines. *IEEE Trans. Info. Theory*, 51:128–142, 2005.

[Tenney and Sandell, 1981] R. R. Tenney and N. R. Jr. Sandell. Detection with distributed sensors. *IEEE Trans. Aero. Electron. Sys.*, 17:501–510, 1981.

[Tibshirani, 1996] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B, Methololological*, 58:267–288, 1996.

[Tishby *et al.*, 1999] N. Tishby, F. Pereira, and W. Bialek. The information bottlekneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, 1999.

[Topsoe, 2000] F. Topsoe. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on Information Theory*, 46:1602–1609, 2000.

[Tropp, 2004] J. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50:2231–2241, 2004.

[Tropp, 2006] J. Tropp. Just relax: Convex programming methods for identifying sparse signals. *IEEE Transactions on Information Theory*, 51:1030–1051, 2006.

[Tsitsiklis and Athans, 1985] J. Tsitsiklis and M. Athans. On the complexity of decentralized decision making and detection problems. *IEEE Trans. on Automatic Control*, 30(5):440–446, 1985.

[Tsitsiklis, 1986] J. N. Tsitsiklis. On threshold rules in decentralized detection. In *Proc. 25th IEEE Conf. Decision Control*, pages 232–236, 1986.

[Tsitsiklis, 1993a] J. Tsitsiklis. Extremal properties of likelihood-ratio quantizers. *IEEE Trans. on Communication*, 41(4):550–558, 1993.

[Tsitsiklis, 1993b] J. N. Tsitsiklis. Decentralized detection. In *Advances in Statistical Signal Processing*, pages 297–344. JAI Press, 1993.

[Tsuda *et al.*, 2002] K. Tsuda, T. Kin, and K. Asai. Marginalized kernels for biological sequences. *Bioinformatics*, 18:268–275, 2002.

[van de Geer, 1999] S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 1999.

[van de Geer, to appear] S. A. van de Geer. High-dimensional generalized linear models and the Lasso. *Annals of Statistics*, to appear.

[van der Vaart and Wellner, 1996] A. W. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, NY, 1996.

[van der Vaart, 1998] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[van Trees, 1990] H. L. van Trees. *Detection, Estimation and Modulation Theory*. Krieger Publishing Co., Melbourne, FL, 1990.

[Vapnik, 1998] V. Vapnik. *Statistical Learning Theory*. John Wiley, New York, 1998.

[Veeravalli *et al.*, 1993] V. V. Veeravalli, T. Basar, and H. V. Poor. Decentralized sequential detection with a fusion center performing the sequential test. *IEEE Trans. Info. Theory*, 39(2):433–442, 1993.

[Veeravalli, 1999] V. V. Veeravalli. Sequential decision fusion: theory and applications. *Journal of the Franklin Institute*, 336:301–322, 1999.

[Viswanathan and Ansari, 1989] R. Viswanathan and A. Ansari. Distributed detection of a signal in generalized Gaussian noise. *IEEE Trans. Acoust., Speech, and Signal Process.*, 37:775–778, 1989.

[Viswanathan and Varshney, 1997] R. Viswanathan and P. K. Varshney. Distributed detection with multiple sensors: Part i—fundamentals. *Proceedings of the IEEE*, 85:54–63, January 1997.

[Wahba, 1990] G. Wahba. *Spline models for observational data*. SIAM, Philadelphia, 1990.

[Wainwright, 2006] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy recovery of sparsity. Technical Report 709, Dept of Statistics, UC Berkeley, May 2006.

[Wald and Wolfowitz, 1948] A. Wald and J. Wolfowitz. Optimum character of the sequential probability ratio test. *Annals of Statistics*, 19:326–339, 1948.

[Wald, 1947] A. Wald. *Sequential Analysis*. John Wiley and Sons, Inc., New York, 1947.

[Wang *et al.*, 2005] Q. Wang, S. R. Kulkarni, and S. Verdú. Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Transactions on Information Theory*, 51(9):3064–3074, 2005.

[Want *et al.*, 1992] R. Want, A. Hopper, V. Falcao, and J. Gibbons. The active badge location system. *ACM Transactions on Information Systems*, 10:91–102, 1992.

[Ward *et al.*, 1997] A. Ward, A. Jones, and A. Hopper. A new location technique for the active office. *IEEE Personnel Communications*, 4:42–47, 1997.

[Wasserman, 2005] L. Wasserman. *All of nonparametric statistics*. Springer, 2005.

[Weinert, 1982] H. L. Weinert, editor. *Reproducing Kernel Hilbert Spaces : Applications in Statistical Signal Processing*. Hutchinson Ross Publishing Co., Stroudsburg, PA, 1982.

[Werbos, 1974] P. Werbos. *Beyond regression*. PhD thesis, Harvard Univesity, 1974.

[Whitehouse, 2002] C. Whitehouse. The design of Calamari: An ad-hoc localization system for sensor networks. Master's thesis, Department of Electrical Engineering and Computer Science, University of California, Berkeley, 2002.

[Xie *et al.*, 2004] Y. Xie, H. Kim, D. R. O'Hallaron, M. K. Reiter, and H. Zhang. Seurat: A pointillist approach to anomaly detection. In *Proceedings of the 7th International Symposium on Recent Advances in Intrusion Detection (RAID)*, Sophia Antipolis, France, 2004.

[Yang and Barron, 1999] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.

[Yegneswaran *et al.*, 2004] V. Yegneswaran, P. Barford, and S. Jha. Global intrusion detection in the domino overlay system. In *Proceedings of Network and Distributed Security Symposium (NDSS)*, 2004.

[Yu, 1996] B. Yu. Assouad, Fano and Le Cam. *Research Papers in Probability and Statistics: Festschrift in Honor of Lucien Le Cam*, pages 423–435, 1996.

[Zhang and Yu, 2005] T. Zhang and B. Yu. Boosting with early stopping: convergence and consistency. *Annals of Statistics*, 33:1538–1579, 2005.

[Zhang, 2004] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annal of Statistics*, 53:56–134, 2004.

[Zhou, 2002] D. X. Zhou. The covering number in learning theory. *Journal of Complexity*, 18:739–767, 2002.