Decentralized Detection and Classification using Kernel Methods

XuanLong Nguyen
Computer Science Division
University of California, Berkeley
xuanlong@cs.berkeley.edu

Martin J. Wainwright
Electrical Engineering and Computer Science
University of California, Berkeley
wainwrig@eecs.berkeley.edu

Michael I. Jordan
Computer Science Division and Department of Statistics
University of California, Berkeley
jordan@cs.berkeley.edu

April 30, 2004

Technical Report 658
Department of Statistics
University of California, Berkeley

Abstract

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 novel algorithm using the framework of empirical risk minimization and marginalized kernels, and analyze its computational and statistical properties both theoretically and empirically. We provide an efficient implementation of the algorithm, and demonstrate its performance on both simulated and real data sets.

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 [28]. 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.

Such problems of decentralized decision-making have been the focus of considerable research in the past two decades [e.g., 27, 28, 7, 8]. 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 the Neyman-Pearson or minimax formulations. See Tsitsiklis [28] and Blum et al. [7] 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 paper 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 [19] 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 [33, 23], as well as the sign detector and its extensions [13, 1, 15]. These methods have been shown to be quite effective for certain types of joint distributions.

Our approach to decentralized detection in this paper is based on a combination of ideas from *reproducing-kernel Hilbert spaces* [2, 25], 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., 34, 17, 18]. More recent work in statistical machine learning [e.g., 26] 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 [26], as well as statistically [35, 3]. 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 paper, 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., 7], 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 [24] 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 paper 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 paper suggests interesting research directions for machine learning—specifically, in designing and analyzing algorithms for communication-constrained environments.

The remainder of the paper is organized as follows. In Section 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, we present a kernel-based algorithm for solving the problem, and we also derive bounds on the performance of this algorithm. Section 4 is devoted to the results of experiments using our algorithm, in application to both simulated and real data. Finally, we conclude the paper with a discussion of future directions in Section 5.

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 paper, and show how it can be formulated in terms of statistical learning. 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.

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 paper 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, 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 [31] that optimal decision rules, whether under the Bayes error or the 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 $X^t, t = 1, \ldots S$. More specifically, we suppose that each 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.

In this paper, 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 $P(X, Y)$.

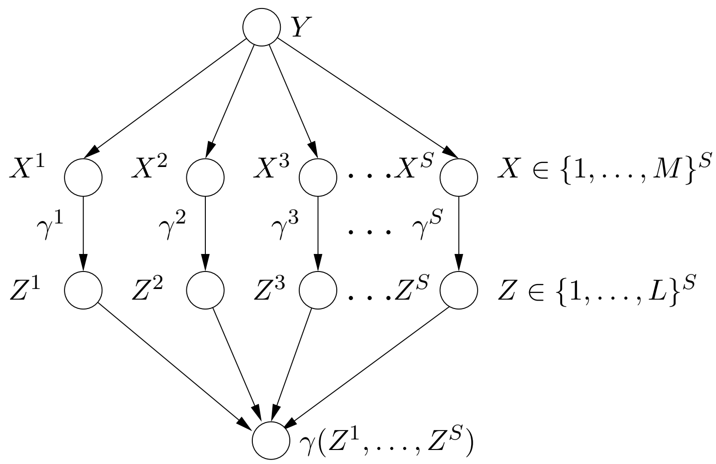

Figure 1. Decentralized detection system with S sensors, in which Y is the unknown hypothesis, $X = (X^1, \ldots, X^S)$ is the vector of sensor observations; and $Z = (Z^1, \ldots, Z^S)$ are the quantized messages transmitted from sensors to the fusion center.

Figure 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 [28], 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 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:

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

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.

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 decision 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:

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

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

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

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 1 we show that the following is one such unbiased estimate of the Bayes risk:

$$R_{emp} \;:=\; \frac{1}{2} - \frac{1}{2n}\sum_z \big|\sum_{i=1}^n Q(z|x_i)y_i\big|. \tag{3}$$

In addition, $\gamma_{opt}(z)$ can be estimated by the decision function $\gamma_{emp}(z) = \mathrm{sign}\big(\sum_{i=1}^n Q(z|x_i)y_i\big)$. Since Z is a discrete random vector, the optimal Bayes risk can be estimated easily, regardless of whether the input signal X is discrete or continuous.

Lemma 1. *(a) Assume that $P(z) > 0$ for all z. Define*

$$\kappa(z) = \frac{\sum_{i=1}^n Q(z|x_i)\mathbb{I}(y_i = 1)}{\sum_{i=1}^n Q(z|x_i)}.$$

Then $\lim_{n\to\infty}\kappa(z) = P(Y=1|z)$.

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

Proof. See Appendix 1. □

The significance of Lemma 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), to maximize

$$C(Q) = \sum_z \Big|\sum_{i=1}^n Q(z|x_i)y_i\Big|, \tag{4}$$

subject to the constraints that define a probability distribution:

$$\begin{cases} Q(z|x) = \prod_{t=1}^{S} Q^t(z^t|x^t) & \text{for all values of } z \text{ and } x. \\ \sum_{z^t} Q^t(z^t|x^t) = 1 & \text{for } t = 1, \ldots, S, \\ Q^t(z^t|x^t) \in [0,1] & \text{for } t = 1, \ldots, S. \end{cases} \quad (5)$$

The major computational difficulty in the optimization problem defined by equations (4) and (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:

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

Expanding the square and using the conditional independence condition (1) leads to the following equivalent form for C_2:

$$C_2(Q) \quad = \quad \sum_{i,j} y_i y_j \prod_{t=1}^{S} \sum_{z^t=1}^{L} Q^t(z^t|x_i^t) Q^t(z^t|x_j^t). \quad (6)$$

Note that the conditional independence condition (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 A kernel-based algorithm

In this section, we turn to methods for decentralized detection based on empirical risk minimization and kernel methods [2, 25, 26]. 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.

3.1 Empirical risk minimization and kernel methods

In this section, we provide some background on empirical risk minimization and kernel methods. The exposition given here is necessarily very brief; we refer the reader to the books [26, 25, 34] 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 $f : \mathcal{X} \to \mathbb{R}$ 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*:

$$\phi_0(yf(x)) \quad := \quad \mathbb{I}[y \neq \text{sign}(f(x))]. \quad (7)$$

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 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 [9, 26] uses a *hinge loss* function:

$$\phi_1(yf(x)) \;\; := \;\; (1 - yf(x))_+ \;\equiv\; \max\{1 - yf(x), 0\}. \tag{8}$$

On the other hand, the logistic regression algorithm [12] is based on the *logistic loss* function:

$$\phi_2(yf(x)) \;\; := \;\; \log\left[1 + \exp^{-yf(x)}\right]^{-1}. \tag{9}$$

Finally, the standard form of the boosting classification algorithm [11] uses a *exponential loss* function:

$$\phi_3(yf(x)) \;\; := \;\; \exp(-yf(x)). \tag{10}$$

Intuition suggests that a function f with small ϕ-risk $\mathbb{E}\phi(Yf(X))$ should also have a small Bayes risk $P(Y \neq \mathrm{sign}(f(X)))$. In fact, it has been established rigorously that convex surrogates for the (non-convex) 0-1 loss function, such as the hinge (8) and logistic loss (9) functions, have favorable properties both computationally (i.e., algorithmic efficiency), and in a statistical sense (i.e., bounds on estimation error) [35, 3].

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 [25]):

Definition 2. *A real-valued kernel function is a symmetric bilinear mapping $K_x : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. 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

$$f(\cdot) = \sum_{j=1}^{m} \alpha_j K_x(\cdot, x_j), \tag{11}$$

where $\{x_j\}_{j=1}^m$ are arbitrary points from $\mathcal{X}$, 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 (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. *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* (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 [25].

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

$$\min_{f \in \mathcal{H}} \sum_{i=1}^{n} \phi(y_i f(x_i)) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2, \tag{12}$$

where $\lambda > 0$ is a regularization parameter. The Representer Theorem (Thm. 4.2; [26]) guarantees that the optimal solution to problem (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 $\widehat{w} = \sum_{i=1}^{n} \alpha_i y_i \Phi(x_i)$.

3.2 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 decision 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

$$\min_{w} \; \Big\{ \sum_{z} \sum_{i=1}^{n} \phi(y_i \gamma(z)) Q(z|x_i) + \frac{\lambda}{2}||w||^2 \Big\}, \tag{13}$$

where $\lambda > 0$ is a regularization parameter. In its current form, the objective function (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 4. *Define the quantities*

$$\Phi_Q(x) := \sum_{z} Q(z|x)\Phi'(z), \quad \text{and} \quad f(x;Q) := \langle w, \Phi_Q(x)\rangle. \tag{14}$$

For any convex ϕ, the optimal value of the following optimization problem is a lower bound on the optimal value in problem (13)*:*

$$\min_{w} \; \sum_{i} \phi(y_i f(x_i;Q)) + \frac{\lambda}{2}||w||^2 \tag{15}$$

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

Proof. 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 $i = 1, \ldots n$, from which the lower bound follows. Equality for deterministic $Q \in \mathcal{Q}_0$ is immediate. $\square$

A key point is that the modified optimization problem (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 (14). Moreover, the corresponding *marginalized kernel* function takes the form:

$$K_Q(x, x') := \sum_{z,z'} Q(z|x) Q(z'|x') \; K_z(z, z'), \tag{16}$$

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, we provide several examples to illustrate.) From the modeling perspective, it is interesting to note that marginalized kernels, like that of equation (16), underlie recent work that aims at combining the advantages of graphical models and Mercer kernels [16, 29].

As a standard kernel-based formulation, the optimization problem (15) can be solved by the usual Lagrangian dual formulation [26], thereby yielding an optimal weight vector w. This weight vector defines the decision rule for the fusion center by $\gamma(z) := \langle w, \, \Phi'(z) \rangle$. By the Representer Theorem [26], the optimal solution w to problem (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 (14). Substituting this decomposition of w into the definition of γ yields

$$\gamma(z) := \sum_{z'} \sum_{i=1}^{n} \alpha_i y_i Q(z'|x_i) K_z(z, z'). \tag{17}$$

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 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:

(a) 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 $K_Q(x, x') = \sum_{t=l}^{S} \mathbb{E}[z^t|x^t]\, \mathbb{E}[z'^t|x'^t]$.

(b) 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:

$$K_Q(x, x') \;=\; \sum_{z,z'} Q(z|x)Q(z'|x') \sum_{t=1}^{S} \mathbb{I}[z^t = z'^t] \;=\; \sum_{t=1}^{S} Q(z^t = z'^t|x^t, x'^t). \tag{18}$$

(c) 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:

$$2 \sum_{1 \le t < r \le S} Q(z^t = z'^t|x^t, x'^t)Q(z^r = z'^r|x^r, x'^r). \tag{19}$$

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.

3.4 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 (17)), and the sensor rules Q. Observe that the cost function (15) can be re-expressed as a function of both w and Q as follows:

$$G(w; Q) \;:=\; \frac{1}{\lambda} \sum_i \phi\Big(y_i \langle w, \sum_z Q(z|x_i)\Phi'(z)\rangle\Big) + \frac{1}{2}||w||^2. \tag{20}$$

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) moreover, 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 3.2, 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 [24], a dual reformulation of $\min_w G(w; Q)$ is given by

$$\max_{\alpha \in \mathbb{R}^n} \Big\{ -\frac{1}{\lambda} \sum_{i=1}^{n} \phi^*(-\lambda \alpha_i) - \frac{1}{2} \alpha^T \big[(yy^T) \circ K_Q \big] \alpha \Big\}, \tag{21}$$

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 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* (21)*. Furthermore, any optimal solution α to problem* (21) *defines the optimal primal solution $w(Q)$ to $\min_w G(w; Q)$ via $w(Q) = \sum_{i=1}^n \alpha_i y_i \Phi_Q(x_i)$.*

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 [21].

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 follows:

$$\inf_{w \in \mathbb{R}^m} G(w) \;=\; \inf_{w \in \mathbb{R}^m} \{G(w) - \langle w,\, 0\rangle\} \;=\; -G^*(0),$$

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 [24]) as follows:

$$G^*(0) \;=\; \inf_{u_i, i=1,\ldots,n} \Big\{ \sum_{i=1}^{n} g_i^*(u_i) + \Omega^*(-\sum_{i=1}^{n} u_i) \Big\}. \tag{22}$$

Applying calculus rules for conjugacy operations (Thm. 16.3; [24]), we obtain:

$$g_i^*(u_i) \;=\; \begin{cases} \frac{1}{\lambda}\phi^*(-\lambda\alpha_i) & \text{if } u_i = -\alpha_i(y_i \Phi_Q(x_i)) \text{ for some } \alpha_i \in \mathbb{R} \\ +\infty & \text{otherwise.} \end{cases} \tag{23}$$

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 (22) leads to:

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

from which it follows that

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

Thus, we have derived the dual form (21). See Appendix 5 for the remainder of the proof, in which we derive the link between $w(Q)$ and the dual variables α. $\square$

This proposition is significant in that the dual problem involves only the kernel matrix $(K_Q(x_i,x_j))_{1\le i,j\le n}$. 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 (8), as used in the SVM algorithm [26]. A straightforward calculation yields

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

Substituting this formula into (21) yields, as a special case, the familiar dual formulation for the SVM:

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

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 establishes that it can always be evaluated at the point $(w(Q), Q)$ for any $Q \in \mathcal{Q}$.

Lemma 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 (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. [1]

Proof. See Appendix 5. $\square$

Observe 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. Overall, the blockwise coordinate descent algorithm for optimizing the choice of local decision rules takes the following form:

[1] *Subgradient* is a generalized counterpart of gradient for non-differentiable convex functions. Briefly, a *subgradient* of a convex function $f : \mathbb{R}^m \to \mathbb{R}$ at x is a vector $s \in \mathbb{R}^m$ satisfying $f(y) \ge f(x) + \langle s,\, y - x\rangle$ for all $y \in \mathbb{R}^m$. The *subdifferential* at a point x is the set of all subgradients; hence, if f is differentiable at x, the subdifferential consists of the single vector $\{\nabla f(x)\}$. In our cases, G is non-differentiable when ϕ is the hinge loss (8), and differentiable when ϕ is the logistic loss (9) or exponential loss (10). $\partial_{Q^t(\bar{z}^t|\bar{x}^t)}G$ evaluated at $(w(Q), Q)$. More details on convex analysis can be found in the books [24, 14].

Kernel quantization (KQ) algorithm:

(a) With Q fixed, compute the optimizing $w(Q)$ by solving the dual problem (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 6.

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

$$\gamma^t(x^t) \quad := \quad \operatorname{argmax}_{z^t \in \mathcal{Z}} Q(z^t | x^t). \tag{24}$$

Remarks: A number of comments about this algorithm are in order. At a high level, the updates 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 [e.g., 28] for decentralized detection under particular assumptions on the joint distribution $P(X,Y)$.

Using standard results [5], it is straightforward to guarantee convergence of such coordinate-wise updates when the loss function ϕ is strictly convex and differentiable (e.g., logistic loss (9) or exponential loss (10)). In contrast, the case of non-differentiable ϕ (e.g., hinge loss (8)) requires more care. We have, however, obtained good results in practice even in the case of hinge loss.

Finally, it is interesting to note the connection between the KQ algorithm and the naive approach considered in Section 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.

3.5 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 $\mathcal{F}$ [35, 3]. 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: We begin by isolating the class of functions over which we optimize. Define, for a fixed $Q \in \mathcal{Q}$, the function space $\mathcal{F}_Q$ as

$$\big\{f : x \mapsto \langle w, \Phi_Q(x)\rangle = \sum_i \alpha_i y_i K_Q(x, x_i) \;\big|\; \text{s.t. } ||w|| \leq B\big\}, \tag{25}$$

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 (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 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

$$\begin{aligned}\inf_{f\in\mathcal{F}} \mathbb{E}\phi(Y\gamma_f(Z)) &\leq \inf_{f\in\mathcal{F}_0} \mathbb{E}\phi(Yf(X) \qquad (26a)\\ \inf_{f\in\mathcal{F}} \mathbb{E}\phi(Y\gamma_f(Z)) &\geq \inf_{f\in\mathcal{F}} \mathbb{E}\phi(Yf(X)). \qquad (26b)\end{aligned}$$

The same statement also holds 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 part (a), and the lower bound (26b) 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 (26a) 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 [30] 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 [20]. 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

$$\sup_{f\in\mathcal{F}} |\widehat{E}\phi(Yf(X)) - \mathbb{E}\phi(Yf(X))| \leq 2\ell R_n(\mathcal{F}) + \sqrt{\frac{\ln(2/\delta)}{2n}}. \tag{27}$$

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

Step 3: Combining the bound (27) with Proposition 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 8. *Given n i.i.d. labeled data points $(x_i, y_i)_{i=1}^n$, with probability at least $1-2\delta$,*

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

Proof. Using bound (27), with probablity 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 (26b), we have, with probability $1-\delta$,

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

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 8 has practical 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 growth in the complexity of $\mathcal{F}$ and $\mathcal{F}_0$ with respect to the number of training samples n, as well as the number of discrete signals L and M. The following proposition derives such bounds, exploiting the fact that the number of 0-1 conditional probability distributions $Q(Z|X)$ is a finite number, (L^{MS}).

Proposition 9.

$$R_n(\mathcal{F}_0) \;\leq\; \frac{2B}{n}\left[\mathbb{E}\sup_{Q\in\mathcal{Q}_0}\sum_{i=1}^n K_Q(X_i,X_i) \;+\; 2(n\;-\;1)\sqrt{n/2}\sup_{z,z'}K_z(z,z')\sqrt{2MS\log L}\right]^{1/2}. \tag{28}$$

Proof. See Appendix 5. □

Although the rate given in equation (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* [30]. 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 probablity [4]. Concretely, for any $\eta > 0$, there holds:

$$P\left\{ |R_n(\mathcal{F}) - \widehat{R}_n(\mathcal{F})| \quad \geq \quad \eta \right\} \leq 2e^{-\eta^2 n/8}. \tag{29}$$

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 our context, consider in particular the $L_2(P_n)$ metric defined 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 [30] that for some absolute constant C, there holds:

$$\widehat{R}_n(\mathcal{F}) \leq C \int_0^\infty \sqrt{\frac{\log N(\epsilon, \mathcal{F}, L_2(P_n))}{n}} \mathrm{d}\epsilon. \tag{30}$$

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

Proposition 10. *The entropy number* $\log N(\epsilon, \mathcal{F}, L_2(P_n))$ *of* $\mathcal{F}$ *is bounded above by*

$$\sup_{Q \in \mathcal{Q}} \log N(\epsilon/2, \mathcal{F}_Q, L_2(P_n)) + (L-1)MS \log \frac{2L^S \sup ||\alpha||_1 \sup_{z,z'} K_z(z, z')}{\epsilon}. \tag{31}$$

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

Proof. See Appendix 5. □

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 (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.

4 Experimental Results

We evaluated our algorithm using both data from simulated sensor networks and real-world data sets. We consider three types of 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 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 [31], 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 2. We implemented the kernel-based quantization algorithm using either first- or second-order count kernels, and the hinge loss function (8), as in the SVM algorithm. The second-order kernel is specified in equation (19) but with the sum taken over only t, r such that $|t - r| = 1$.

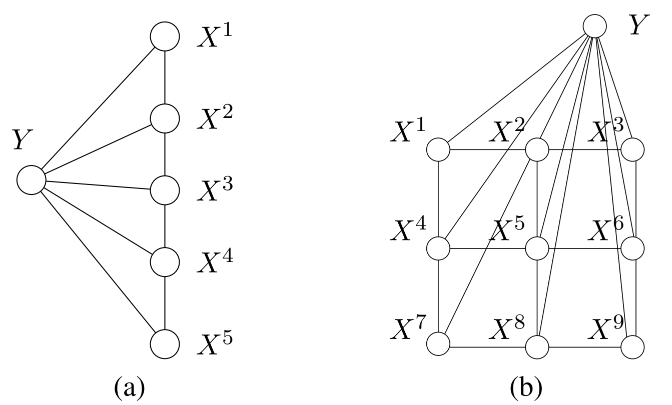

Figure 2. Examples of graphical models $P(X, Y)$ of our simulated sensor networks. (a) Chain-structured dependency. (b) Fully connected (not all connections shown).

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 consider a sensor network with 9 nodes (i.e., $S = 9$), arrayed in the 3×3 lattice illustrated in Figure 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.

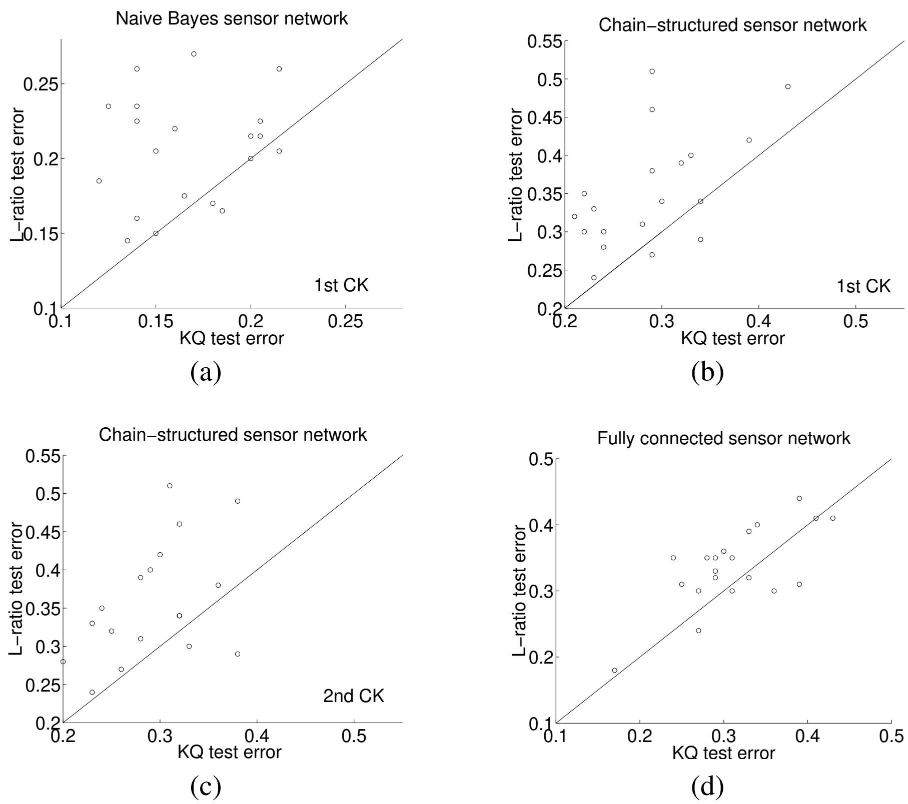

Figure 3. Scatter plots of the test error of the LR versus KQ methods. (a) Conditionally independent network. (b) Chain model with first-order kernel. (c), (d) Chain model with second-order kernel. (d) Fully connected model.

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 $\frac{P(X^t=u|Y=1)}{P(X^t=u|Y=-1)}$ for $u = 1, \ldots, M$ of the likelihood ratio are sorted and grouped evenly into L bins. 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 [28].

The KQ algorithm generally yields more accurate classification performance than the likelihood-ratio based algorithm (LR). Figure 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 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 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.

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 [6]. 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 [10]. 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.

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 1: Experimental results for the UCI data sets.

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 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 [10].

5 Conclusions

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 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 [28]. 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. [32]). It is also interesting to consider extensions our method to this sequential setting.

Acknowledgement

We are grateful to Peter Bartlett for very helpful discussions related to this work. We wish to acknowledge support from ONR MURI N00014-00-1-0637 and ARO MURI DAA19-02-1-0383.

Proof of Lemma 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 $z \in \mathcal{Z}^S$.) By the strong law of large numbers, there holds

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

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 $n \to \infty$,

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

where 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, $R_{emp} \to R_{opt}$.

Proof of Proposition 5 Here we complete the proof of Proposition 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, [14]):

$$\partial G^*(0) = \partial g_1^*(\widehat{u_1}) \cap \ldots \cap \partial g_n^*(\widehat{u_n}) \cap \partial\Omega^*(-\sum_{i=1}^n \widehat{u_i}).$$

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 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 (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 [14]), 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 $i = 1, \ldots, n$.

Proof of Lemma 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 (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 [24], 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 we proved that $-\lambda\alpha_i \in \partial\phi(b_i)$ for each $i = 1, \ldots, n$, where α is the optimal solution of (21). Therefore, $\partial_{Q(\bar{z}|\bar{x})} G$ evaluated at $(w(Q); Q)$ contains the following element:

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

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 $Q(\bar{z}|\bar{x}) = \prod_{t=1}^S Q^t(\bar{z}^t|\bar{x}^t)$.

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

which contains the following element as one of its subgradients:

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

This completes the proof of the lemma.

Proof of Proposition 9: By definition of Rademacher complexity [30], we have

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

Applying the Cauchy-Schwarz inequality yields

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

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 [22] for bounding Rademacher averages over a finite set $A \subset \mathbb{R}^d$:

$$\mathbb{E} \sup_{a \in A} \sum_{i=1}^{d} \sigma_i a_i \quad \le \quad \max ||a||_2 \sqrt{2 \log |A|}. \tag{32}$$

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, for each $i = 1, \ldots, n-1$:

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

from which the lemma follows.

Proof of Proposition 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:

$$N(\epsilon, \mathcal{Q}, L_\infty) \leq N(\epsilon, \Delta_L, l_\infty)^{MS} \leq (1/\epsilon)^{(L-1)MS}.$$

Recall that each $f \in \mathcal{F}$ can be written as:

$$f(x) = \sum_{i=1}^{n} \alpha_i \sum_{z, z_i} Q(z|x) Q(z_i|x_i) K_z(z, z_i). \tag{33}$$

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 (33) 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 $\|f_1 - f_2\|_{L_2(P_n)} \leq \epsilon/2$. Applying the triangle inequality yields

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

which is bounded above by 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.

References

[1] 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.

[2] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[3] P. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification and risk bounds. Technical Report 638, Department of Statistics, University of California at Berkeley, April 2003.

[4] P. Bartlett and S. Mendelson. Gaussian and Rademacher complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[5] D.P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 1995.

[6] C.L. Blake and C.J. Merz. UCI repository of machine learning databases, 1998.

[7] 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, 1997.

[8] J. F. Chamberland and V. V. Veeravalli. Decentralized detection in sensor networks. *IEEE Transactions on Signal Processing*, 51(2):407–416, 2003.

[9] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[10] J. Dougherty, R. Kohavi, and M. Sahami. Supervised and unsupervised discretization of continuous features. In *Proceedings of the ICML*, 1995.

[11] 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.

[12] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, 28:337–374, 2000.

[13] 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.

[14] J. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer, 2001.

[15] E. K. Hussaini, A. A. M. Al-Bassiouni, and Y. A. El-Far. Decentralized CFAR signal detection. *Signal Processing*, 44:299–307, 1995.

[16] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems 11*, Cambridge, MA, 1999. MIT Press.

[17] T. Kailath. RKHS approach to detection and estimation problems—Part I: Deterministic signals in Gaussian noise. *IEEE Trans. Info. Theory.*, 17:530–549, 1971.

[18] T. Kailath and H. V. Poor. Detection of stochastic processes. *IEEE Trans. Info. Theory.*, 44:2230–2259, 1998.

[19] S. A. Kassam. Nonparametric signal detection. In *Advances in Statistical Signal Processing*. JAI Press, 1993.

[20] V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30:1–50, 2002.

[21] D. G. Luenberger. *Optimization by Vector Space Methods*. Wiley, New York, 1969.

[22] P. Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciences de Toulouse*, IX:245–303, 2000.

[23] A. Nasipuri and S. Tantaratana. Nonparametric distributed detection using Wilcoxon statistics. *Signal Processing*, 57(2):139–146, 1997.

[24] G. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

[25] S. Saitoh. *Theory of Reproducing Kernels and its Applications*. Longman Scientific & Technical, Harlow, UK, 1988.

[26] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[27] R. R. Tenney and N. R. Jr. Sandell. Detection with distributed sensors. *IEEE Trans. Aero. Electron. Sys.*, 17:501–510, 1981.

[28] J. N. Tsitsiklis. Decentralized detection. In *Advances in Statistical Signal Processing*, pages 297–344. JAI Press, 1993.

[29] K. Tsuda, T. Kin, and K. Asai. Marginalized kernels for biological sequences. *Bioinformatics*, 18:268–275, 2002.

[30] A. W. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, NY, 1996.

[31] H. L. van Trees. *Detection, Estimation and Modulation Theory*. Krieger Publishing Co., Melbourne, FL, 1990.

[32] 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.

[33] R. Viswanathan and A. Ansari. Distributed detectionof a signal in generalized Gaussian noise. *IEEE Trans. Acoust., Speech, and Signal Process.*, 37:775–778, 1989.

[34] H. L. Weinert, editor. *Reproducing Kernel Hilbert Spaces : Applications in Statistical Signal Processing*. Hutchinson Ross Publishing Co., Stroudsburg, PA, 1982.

[35] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annal of Statistics*, 53:56–134, 2003.