Nonparametric Decentralized Detection and Sparse Sensor Selection Via Weighted Kernel

Weiguang Wang, Yingbin Liang, *Member, IEEE*, Eric P. Xing, *Senior Member, IEEE*, and Lixin Shen

***Abstract*—The kernel-based nonparametric approach proposed by Nguyen, Wainwright, and Jordan is further investigated for decentralized detection. In contrast with the uniform kernel used in the previous work, a weighted kernel is proposed, where weight parameters serve to selectively incorporate sensors' information into the fusion center's decision rule based on quality of sensors' observations. Furthermore, weight parameters also serve as sensor selection parameters with nonzero parameters corresponding to sensors being selected. By introducing the l_1 regularization on weight parameters into the risk minimization framework, sensor selection is jointly performed with decision rules for sensors and the fusion center with the resulting optimal decision rule having only sparse nonzero weight parameters. A gradient projection algorithm and a Gauss-Seidel algorithm are developed to solve the risk minimization problem, which is nonconvex, and both algorithms are shown to converge to critical points. Conditions on the sample complexity to guarantee asymptotically small estimation error are characterized based on analysis of Rademacher complexity. Connection between the probability of error and the risk function is also studied. Numerical results are provided to demonstrate the advantages and properties of the proposed approach based on weighted kernel.**

***Index Terms*—Convergence, Gauss-Seidel algorithm, gradient projection, KL-property, non-convex problem, risk minimization, RKHS, sensor selection.**

I. Introduction

In the decentralized detection problem (see, e.g., [1]–[3]), a number of sensors receive observations about the state of an event, and then each sensor individually quantizes its observations and forwards quantized information to a fusion center. Finally, the fusion center determines the state of the event based on its received information from the sensors. The goal is to find jointly optimal decentralized quantization rules for sensors and a decision rule for the fusion center to achieve the best system performance.

One critical factor that affects decision accuracy is the knowledge of the joint distribution of the states and observations at sensors. Most of the previous studies [4]–[6] assume that such knowledge is known fully or partially. Such parametric approaches are justified, because the joint distribution can be learned via sampled data in advance. As such, implicitly, the two processes of learning the distribution and designing decentralized detection rules are taken care of separately. However, such separation may not be preferable when the distribution is dynamic and changes fast over time. In this case, estimating the time-varying distribution may significantly increase system complexity. Furthermore, errors in estimating the distribution can propagate to reduce detection accuracy. Thus, it is desirable to make decisions directly based on training data without explicitly estimating the distribution. Such approaches are referred to as *nonparametric* decentralized detection.

Nonparametric (de)centralized detection was studied previously in, e.g., [4]–[6], in which detectors are typically designed to perform well only for specific statistical environments. A learning-based nonparametric linear regression problem was studied in [7], [8]. More recently, a kernel-based classification approach was proposed in [9] for solving the nonparametric decentralized detection problem, which is more generally applicable with mathematical guarantee on the performance. The basic idea is to use a kernel as a measure for capturing similarity between new and training data (e.g., observations). The decision is then made to classify the new observation to the class to which the new observation is closest. In general, a decision rule is expressed as a linear combination of kernels between a new observation and the training data. More formally, the kernel function is associated with a reproducing kernel Hilbert space (RKHS), over which the decision rule of the fusion center is searched to optimize a given loss function (such as the probability of detection error and the hinge loss function) jointly with the local decision rules for individual sensors. It has been shown by numerical examples in [9] that the kernel-based approach yields better performances than other approaches based on estimating joint distributions. Furthermore, compared to parametric approaches, such a kernel-based nonparametric approach is also applicable for the case with correlated observations, in which the correlation is implicitly embedded into training data and their influence on the decision rules are automatically incorporated by optimizing empirical risk functions determined by the training data.

In this paper, we study more realistic sensor networks, which generalize the system models studied in [9], [10] to heterogeneous networks, in which sensors' observations can have different quality and belong to different alphabets. This can be due to their different locations in capturing the environmental event. Furthermore, sensors' transmissions to the fusion center can be

Manuscript received February 11, 2015; revised June 16, 2015; accepted August 11, 2015. Date of publication August 28, 2015; date of current version December 16, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Lei Huang. The work of W. Wang and Y. Liang was supported by the NSF CAREER under Award CCF-10-26565. The work of E. P. Xing was supported by the NIH under Grant 1R01GM087694, Grant FA9550010247, and Grant NIH1R01GM093156. The work of L. Shen was supported by the NSF under Awards DMS-1522332 and DMS-1115523. This paper was presented in part at the IEEE International Workshop on Machine Learning for Signal Processing (MLSP) 2014.

W. Wang and Y. Liang are with the Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY 13244 USA (e-mail: wwang23@syr.edu; yliang06@syr.edu).

E. P. Xing is with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15203 USA (epxing@cs.cmu.edu).

L. Shen is with the Department of Mathematics, Syracuse University, Syracuse, NY 13244 USA (e-mail: lshen03@syr.edu).

Digital Object Identifier 10.1109/TSP.2015.2474297

subject to different rate constraints (in terms of bits per observation), and hence sensors' quantization levels are different. These heterogeneous features are well justified in practice. Sensor networks are typically deployed over a large area geographically. Hence, the noise levels in observations may vary from site to site, which naturally causes the quality of the observations to vary from sensor to sensor. Moreover, sensors' transmissions to the fusion center are typically over wireless channels, whose quality depends on the surrounding wireless scattering environments. Hence, their transmission rates to the fusion center can be different. More specifically, potential applications of heterogeneous models can include geographical distributed sensing [1], intrusion detection in wireless sensor networks [11], distributed equipment failure detection [1], multi-static airborne radar [12].

Thus, our goal in this paper is to design nonparametric decision rules which take heterogeneous features of networks into consideration for achieving as good performance as possible. It is also desirable that the approach can yield efficient sensor selection algorithms, i.e., selecting a subset of sensors that provide the best performance among all possible subsets. Such a problem has significant practical importance, because it is preferable in many cases that only a subset of sensors with good observation quality provide data to a fusion center due to constraints in communication resources and power constraints on sensors. However, sensor selection is in general a challenging problem. The main reason is that the quality of sensors is not easily parameterized into the performance metric, and hence sensor selection can only be done through a combinatorial optimization problem, for which the algorithm is not scalable as network size enlarges. In this paper, we also wish to address the sensor selection problem in our proposed framework.

A. Main Contributions

In this subsection, we summarize our main contributions.

1. We incorporate a novel weighted kernel into the risk minimization framework proposed in [9] for nonparametric decentralized detection. In this way, the fusion center's decision rule is optimized over the Hilbert space (i.e., the RKHS) associated with the weighted kernel, and thus can selectively incorporate information from sensors based on the quality of these information sources. We derive performance bounds based on Rademacher complexity over the union of all weighted RKHSs. We characterize conditions on the sample complexity to guarantee asymptotically small estimation error. We also establish the connection between the probability of error and the risk function in our optimization problem.
2. Using the weighted kernel, we incorporate the sensor selection function into the framework by introducing an l_1 regularization on weight parameters to the risk function so that the resulting optimal decision rule contains sparse nonzero weight parameters, i.e., only the most contributive sensors are selected. Thus, the kernel weight parameters (i.e., sensor selection strategy) and decision rules for sensors and the fusion center are jointly optimized in order to achieve the best performance. The advantages and properties of such an approach are described as follows.
   - The sensor selection problem can now be solved via recent celebrating techniques of Lasso and compressed sensing [13]–[16], which significantly reduces computational complexity;
   - The regularization parameter of l_1 can flexibly control sparsity of sensor selection and its trade-off with the performance of decision making;
   - This sensor selection approach preferably selects sensors with independent observations, and removes highly correlated (and hence redundant) observations, thus achieving dimension reduction as well.

   The sensor selection problem can be very challenging otherwise in the context of nonparametric decentralized detection.
3. We develop a gradient projection algorithm and a Gauss-Seidel algorithm to optimize the regularized non-convex risk minimization problem with differentiable loss functions. We show that both algorithms converge to critical points. We also provide a Gauss-Seidel algorithm to optimize the risk function with non-differentiable hinge loss function.

B. Related Work on Sensor Selection

Sensor selection problem (not necessarily in the context of decentralized detection) has been intensively studied previously (see, e.g., a review [17] on sensor selection in wireless sensor networks). In general, sensor selection is a difficult problem, because it is challenging to design efficient algorithms that overcome exhaustive search over all possible subsets of sensors for optimizing the performance. Majority of previous work studied sensor selection under parametric/semi-parametric models (e.g., [18]–[22]), in which the statistical distribution of event states and observations or the relationship of system parameters and observations is known fully or partially. A number of approaches for sensor selection have been proposed. The work [23] and [24] considered scenarios that only one sensor is selected at a time, hence complexity of exhaustive search is reduced. The work [25] provided more efficient algorithms than exhaustive search based on bounds on objective functions. The work [26] utilized specific structures of performance metric to design efficient algorithms that have low computational complexity. In general, these algorithms may perform well for specific problems, but did not provide a systematic way of treating the problem. More recently, an interesting approach based on convex relaxation of sensor selection problem was proposed in [19], in which sensor selection is formulated as a Boolean-convex problem. Relaxation is then taken to allow discrete Boolean variables to take continuous values. However, their approach is not applicable to our problem, because our sensor selection parameters (i.e., kernel weight parameters) have physical meanings in the decision rule and are not discrete valued.

II. Background on Kernels

In this section, we briefly introduce the basic concepts, definitions and results on learning with kernels. This is the major technique that this paper applies. A reader can refer to [27] for more details. We let $\mathcal{X}$ be a nonempty set, and define a kernel function as follows.

Definition 1: A function $k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathcal{R}$ is called a kernel if for all positive integer m and all $x_1, \ldots, x_m \in \mathcal{X}$, the $m \times m$ matrix K with elements $K_{ij} = k(x_i, x_j)$ for $i, j = 1, \ldots, m$ is positive semidefinite.

Given a kernel function $k(\cdot, \cdot)$, we define a feature mapping $\Phi : x \in \mathcal{X} \to k(\cdot, x)$, which maps an element $x \in \mathcal{X}$ to a function $k(\cdot, x)$. We then define a vector space containing

$$f(\,\cdot\,) = \sum_{i=1}^{m} \alpha_i k(\cdot, x_i)$$

where m is any positive integer, $\alpha_i \in \mathcal{R}$, and $x_1, \ldots, x_m \in \mathcal{X}$ are arbitrary. For this vector space, we define an inner product between f and another function $g(\,\cdot\,) = \sum_{j=1}^{m'} \beta_j k(\cdot, x'_j)$ as

$$\langle f, g \rangle = \sum_{i=1}^{m} \sum_{j=1}^{m'} \alpha_i \beta_j k(x_i, x'_j).$$

In particular, this implies $\langle k(\cdot, x), k(\cdot, x') \rangle = k(x, x')$. It can be shown that after completing such a vector space, we obtain a Hilbert space, referred to as a reproducing kernel Hilbert space (RKHS) associated with the kernel k. We next formally define the RKHS as follows.

Definition 2: Consider a Hilbert space $\mathcal{H}$ containing functions $f : \mathcal{X} \to \mathcal{R}$. It is called a reproducing kernel Hilbert space (RKHS) if there exists a kernel $k(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathcal{R}$ with the following properties:

— k has the reproducing property:

$$\langle f, k(\cdot, x) \rangle = f(x) \quad \text{for all } f \in \mathcal{H},$$

— k spans $\mathcal{H}$, i.e., $\mathcal{H}$ is the completion of a vector space spanned by $k(\cdot, x)$ for $x \in \mathcal{X}$.

We next introduce the important kernel Representer Theorem [27], which is useful for characterizing the optimal solution in empirical risk minimization.

Theorem 1. [27]: Let $\Omega : [0, \infty) \to \mathcal{R}$ be a strictly monotonic increasing function, $\mathcal{X}$ be a nonempty set, $c : (\mathcal{X} \times \mathcal{R}^2)^m \to \mathcal{R} \cup \{\infty\}$ be an arbitrary risk function, and $\mathcal{H}$ be the RKHS associated with a kernel k. Then each minimizer $f \in \mathcal{H}$ of the regularized risk function

$$c((x_1, y_1, f(x_1)), \ldots, (x_m, y_m, f(x_m))) + \Omega(\|f\|_{\mathcal{H}})$$

admits a representation of the form

$$f(\,\cdot\,) = \sum_{i=1}^{m} \alpha_i k(\cdot, x_i).$$

III. Problem Formulation

A. System Model and Notations

We study the nonparametric decentralized detection over a sensor network. The system model is depicted in Fig. 1. In such a system, let Y denote the state of an environmental event, which can take binary values $+1$ and -1. Suppose there are S sensors in the network, which can receive observations about Y. We use X^s to denote the observation received by sensor s for $s = 1, \ldots, S$, and use $\underline{X} = (X^1, \ldots, X^S)$ to denote the observations of all sensors. Each sensor quantizes its observation based on its own local decision rule (i.e., quantization rule). We denote Z^s as the quantized value of X^s by sensor s. We let $\underline{Z} = (Z^1, \ldots, Z^S)$ denote quantized symbols from all sensors. We assume that both X^s and Z^s have finite alphabets $\mathcal{X}_s$, $\mathcal{Z}_s$, correspondingly. Therefore, $\underline{X}$ and $\underline{Z}$ have finite alphabet sets, i.e., $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_S$ and $\mathcal{Z} = \mathcal{Z}_1 \times \mathcal{Z}_2 \times \cdots \times \mathcal{Z}_S$. We note that although sensor observations $\underline{X}$ are often continuous variables in practice, sensors typically digitize their measurements to improve robustness of further processing and reduce processing complexity. The decision rule of a sensor can be generally characterized by a probability distribution $Q_s(z^s \mid x^s)$, which implies that sensor s quantizes x^s into z^s with the probability $Q_s(z^s \mid x^s)$. Thus, random decision rules for sensors are allowed. All sensors then forward their quantized information to a fusion center, which combines all received information from sensors, and makes a decision about the state of the environmental event Y. The fusion center's decision rule can be written as a function $w(\underline{Z})$.

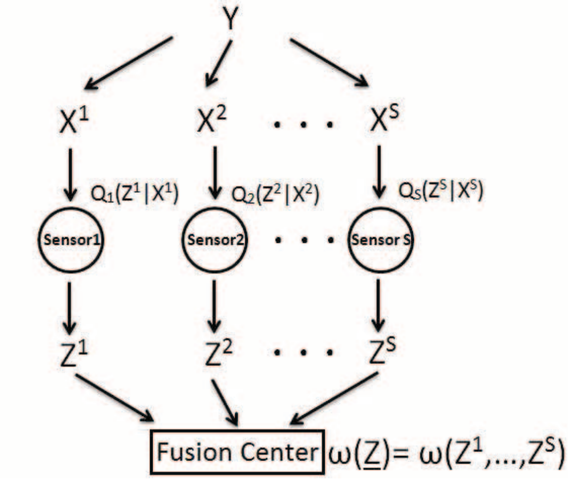


Fig. 1. Illustration of decentralized detection.

We note that this paper implicitly assumes that sensing environment is static. In practice, as quality of sensors changes over time, the training techniques developed in this paper can be performed every a certain period in order to adapt decision rules to the change. In fact, treatment of such an issue can lead to a number of research topics such as how to exploit similarity of decision rules across time to reduce computation complexity of training process, which is left for future work.

B. Weighted Kernel

In this paper, we search decision rules for the fusion center over the RKHS $\mathcal{H}$ associated with a kernel function $k(\cdot, \cdot) : \mathcal{Z} \times \mathcal{Z} \to \mathcal{R}$. Thus, we can express the fusion center's decision rule as:

$$w(\underline{z}) = \langle w(\,\cdot\,), \Phi(\underline{z}) \rangle_{\mathcal{H}}$$

where $w(\,\cdot\,) \in \mathcal{H}$ and $\Phi(\underline{z}) = k(\cdot, \underline{z})$.

It is clear that the performance of the fusion center's decision rule critically depends on the RKHS over which it is chosen and its associated kernel function. In [9], the adopted kernel functions are uniform across sensor's information, i.e., uniform across Z^s for $s = 1, \ldots, S$. Thus, the corresponding Hilbert space contains functions (i.e., decision rules of the fusion center) that treat the information across sensors equally. However, these decision rules may not perform well enough for scenarios, where the sensors' information have different quality. In such cases, it is desirable that the fusion center's decision rule weigh the sensors' information selectively based on the quality of their observations.

Therefore, we propose to use weighted kernels so that their associated RKHS allows decision rules of the fusion center to selectively incorporate sensors' information using weight parameters. We further introduce the kernel weight parameters into the risk minimization framework so that these weight parameters (and hence its associated RKHS) are jointly selected with the decision rules for the fusion center and sensors to optimize the performance. Thus, the impact of the heterogeneous features of the network are naturally incorporated into the fusion center's decision rules via selecting the optimal weight parameters (i.e., the RKHS that these decision rules lie in).

As an example weighted kernel, the weighted first-order count kernel is given by

$$k_{\underline{\beta}}(\underline{z}, \underline{z}') = \sum_{s=1}^{S} \beta^s \mathbb{I}[z^s = z'^s], \qquad (1)$$

where $\mathbb{I}[\cdot]$ is an indicator (characteristic) function, and $\beta^s \geq 0$ for $s = 1, \ldots, S$ are weight parameters. We collect these parameters into a vector $\underline{\beta} = (\beta_1, \ldots, \beta^S)$. It can be shown that the weighted count kernel satisfy the definition of kernel.

It can be seen that each weight parameter β^s in (1) represents the contribution of sensor s to the decision rule of the fusion center. Thus, the Hilbert space $\mathcal{H}_{\underline{\beta}}$ over which the decision rule of the fusion center is chosen is spanned by the weighted count kernel $k_{\underline{\beta}}(\cdot, \cdot)$.

Remark 1: Our study uses the weighted count kernel as an example kernel. In fact, weight parameters can be introduced to more general types of kernels for selectively counting information with unequal quality in decision making. Our problem formulation, algorithm design, and performance analysis are generally applicable to these cases as well.

C. Problem Formulation With Sensor Selection

In this paper, we consider nonparametric decentralized detection, and assume that the joint distribution $P(Y, \underline{X})$ is unknown. Instead, a set of training data are available, i.e., $(y_i, \underline{x}_i)$ for $i = 1, \ldots, N$. We adopt the framework of the empirical risk minimization for decentralized detection as in [9] and further introduce weighted kernel and incorporate l_1 regularization for kernel weight parameters in order for sparse sensor selection. More specifically, we jointly find optimal weight parameters $\underline{\beta}$, decision rule $w(\underline{Z})$ for fusion center, and decision rules $Q_s(\overline{z^s}|x^s)$ for all sensors $(s = 1, \ldots, S)$ that minimize the following l_1 regularized empirical risk function:

$$\min_{\substack{\beta^s \geq 0 \text{ for } s=1,\ldots,S \\ w \in \mathcal{H}_{\underline{\beta}}, Q \in \mathcal{Q}}} \sum_{i=1}^{N} \sum_{\underline{z}} \phi(y_i \langle w(\,\cdot\,), \Phi_{\underline{\beta}}(\underline{z}) \rangle_{\mathcal{H}_{\underline{\beta}}}) Q(\underline{z} \,|\, \underline{x}_i) + \frac{\lambda_1}{2} \|w\|^2_{\mathcal{H}_{\underline{\beta}}} + \lambda_2 \|\underline{\beta}\|_{l_1} \qquad (2)$$

where $\phi(\,\cdot\,)$ is a convex loss function such as the logistic or hinge loss functions, $\mathcal{H}_{\underline{\beta}}$ denotes the Hilbert space associated with the weighted count kernel $k_{\underline{\beta}}(\underline{z}, \underline{z}')$, $\Phi_{\underline{\beta}}(\underline{z}) = k_{\underline{\beta}}(\cdot, \underline{z})$, and $\mathcal{Q}$ is the set that includes all possible conditional probabilities $Q(\underline{z} \,|\, \underline{x})$ that decompose as $Q(\underline{z} \,|\, \underline{x}) = \prod_{s=1}^{S} Q_s(z^s \,|\, x^s)$. Such decomposability is because sensors follow independent local decision rules. The set $\mathcal{Q}$ is formally defined as follows.

$$\mathcal{Q} = \left\{ Q(\underline{Z} \,|\, \underline{X}) : Q(\underline{z} \,|\, \underline{x}) = \prod_{s=1}^{S} Q_s(z^s \,|\, x^s), \sum_{z_s \in \mathcal{Z}_s} Q_s(z^s \,|\, x^s) = 1,\, Q_s(z^s \,|\, x^s) \geq 0 \text{ for all } x_s \in \mathcal{X}_s, z_s \in \mathcal{Z}_s \text{ and } s = 1, 2, \ldots, S \right\}. \qquad (3)$$

We note that it is computationally complex to solve the above optimization problem due to the expectation of $\phi(\,\cdot\,)$ taken over $Q(\underline{z} \,|\, \underline{x}_i)$. Hence, as in [9], we consider the following lower bound as a relaxation of (2) due to Jensen's inequality

$$\min_{\substack{\beta^s \geq 0 \text{for} s=1,\ldots,S \\ w \in \mathcal{H}_{\underline{\beta}}, Q \in \mathcal{Q}}} \sum_{i=1}^{N} \phi(y_i \langle w(\,\cdot\,), \Phi'_{\underline{\beta}}(\underline{x}_i) \rangle_{\mathcal{H}_{\underline{\beta}}}) + \frac{\lambda_1}{2} \|w\|^2_{\mathcal{H}_{\underline{\beta}}} + \lambda_2 \|\underline{\beta}\|_{l_1} \qquad (4)$$

where $\Phi'_{\underline{\beta}}(\underline{x}_i) = \sum_{\underline{z}} \Phi_{\underline{\beta}}(\underline{z}) Q(\underline{z} \,|\, \underline{x}_i) \in \mathcal{H}_{\underline{\beta}}$. In Section IV.D, we study how close the above empirical risk function is to the true risk function. We also show that the above empirical risk function provides an upper bound on the probability of detection error, which justifies using this function as an approximation.

In the above problem, l_1 regularization for kernel weight parameters encourages sparse weight (i.e., sensor) selection. The coefficient λ_2 controls the sparsity level of sensor selection, and thus controls the trade-off between sensor selection and the overall system performance. For systems with stringent communication constraints on sensors' transmissions to the fusion center, λ_2 needs to be large so that only a small fraction of sensors are selected to participate in decision making. Given the sparsity level, the risk minimization guarantees that selected sensors are those with good quality of observations and can hence contribute best to decision making.

Our goal is to jointly design decision rule $w(\underline{Z})$ for the fusion center, decision rules $Q_s(z^s \,|\, x^s)$ for sensors, and sensor selection strategy in order to achieve the best system performance.

IV. Main Results

A. Algorithm Design

In this section, we develop algorithms to solve the risk minimization problem (4), in which the minimization is taken over three types of variables $\underline{\beta}, w$ and Q. It is clear that the risk function is not convex jointly over these variables. In general, designing algorithms that converge to a global optimal solution for non-convex optimization is challenging. In many cases, even convergence to a critical point can be difficult. Moreover, the l_1 regularization term in (4) is a non-smooth function, which further complicates the problem. In this section, we first develop two algorithms for the case where $\phi(\,\cdot\,)$ is a differentiable loss function such as logistic and exponential loss functions, and then address the case where $\phi(\,\cdot\,)$ is a non-differentiable loss function such as hinge loss function. We study convergence of these algorithms in Section IV.C.

For the case with differentiable $\phi(\cdot)$, we first note that since w is a function belonging to a given RKHS associated with $k_{\underline{\beta}}$, it is not possible to optimize over $\underline{\beta}$ (i.e., the corresponding RKHS) but keeping w in a particular RKHS fixed. In another word, w and $\underline{\beta}$ are not independent parameters that can be alternatively optimized. To solve this problem, we note that following an argument similar to the kernel Representer Theorem [27], the minimizer of the problem given in (4) with fixed Q and $\underline{\beta}$ takes the form

$$w = \sum_{i=1}^{N} \alpha_i y_i \Phi'_{\underline{\beta}}(\underline{x}_i) = \sum_{i=1}^{N} \sum_{\underline{z} \in \mathcal{Z}} \alpha_i y_i \Phi_{\underline{\beta}}(\underline{z}) Q(\underline{z} \mid \underline{x}_i) \tag{5}$$

for some parameters $\underline{\alpha} = (\alpha_1, \ldots, \alpha_N)$, which are projection parameters of w along kernel functions in $\mathcal{H}_{\underline{\beta}}$. It is then clear that $\underline{\alpha}$, Q and $\underline{\beta}$ are independent parameters, and the optimization problem (4) can be solved equivalently by optimizing over these three types of parameters. Therefore, problem (4) is equivalent to the following optimization problem:

$$\min_{\substack{\beta^s \geq 0 \text{ for } s=1,\ldots,S \\ Q \in \mathcal{Q}, \underline{\alpha} \in \mathcal{R}^N}} G(\underline{\alpha}, \underline{\beta}, Q), \tag{6}$$

where

$$\begin{aligned} &G(\underline{\alpha}, \underline{\beta}, Q) \\ &= \sum_{i=1}^{N} \phi \left(y_i \sum_{j=1}^{N} \alpha_j y_j \left[\sum_{s=1}^{S} \beta^s \sum_{z^s} Q_s\left(z^s \mid x_i^s\right) Q_s\left(z^s \mid x_j^s\right) \right] \right) \\ &\quad + \frac{\lambda_1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i y_i \alpha_j y_j \left[\sum_{s=1}^{S} \beta^s \sum_{z^s} Q_s\left(z^s \mid x_i^s\right) Q_s\left(z^s \mid x_j^s\right) \right] \\ &\quad + \lambda_2 \|\underline{\beta}\|_{l_1}. \end{aligned} \tag{7}$$

In Algorithm 1, we develop a gradient projection algorithm to solve the non-convex risk minimization problem (6) with a continuous loss function. Here, we combine three types of parameters together as one multi-dimensional vector $(\underline{\alpha}, \underline{\beta}, Q)$, and update the entire vector at each step. We note that the non-differentiable term $\|\underline{\beta}\|_{l_1}$ can be changed to $\sum_{s=1}^{S} \beta^s$ by exploiting the constraints $\beta^s \geq 0$. In this way, the risk function becomes differentiable and hence much easier to handle. Thus, the algorithm performs a two-step update. Step 1 takes the gradient of the objective function over $(\underline{\alpha}, \underline{\beta}, Q)$ to generate $(\underline{\alpha}^{(k)}, \hat{\underline{\beta}}^{(k)}, \hat{Q}^{(k)})$ as in (8), where L denotes the Lipschitz constant of the objective function $G(\underline{\alpha}, \underline{\beta}, Q)$. Then step 2 projects $\hat{\underline{\beta}}^{(k)}$ and $\hat{Q}^{(k)}$ into the corresponding constraint sets $\{\underline{\beta} : \beta^s \geq 0\}$ and $\mathcal{Q}$, respectively. The projection of vector $\hat{\underline{\beta}}^{(k)}$ is to keep all non-negative entries and set all negative entries to be 0. The projection of Q can be performed by solving a constrained convex optimization problem (10). Using the KKT conditions, the close-form expression of the optimizer can be derived. Due to the fact that the projections can be performed with exact close-form solutions, the convergence of the algorithm can be further shown in Section IV.C.

Algorithm 1: Decentralized Detection via Gradient Projection-Based Method

```
Input: $S$, $\{y_i, x_i^1, \ldots, x_i^S\}_{i=1}^n$.
Step 0: Initialize $\underline{\alpha} \in \mathcal{R}^N$, $\underline{\beta}$ where $\beta^s \geq 0$ for $s = 1, \ldots, S$, $Q \in \mathcal{Q}$
Step k:
    • Gradient step: for $t \leq 1/L$,
        $(\underline{\alpha}^{(k)}, \hat{\underline{\beta}}^{(k)}, \hat{Q}^{(k)}) = (\underline{\alpha}^{(k-1)}, \underline{\beta}^{(k-1)}, Q^{(k-1)})$
            $-t \nabla_{(\underline{\alpha}, \underline{\beta}, Q)} G(\underline{\alpha}^{(k-1)}, \underline{\beta}^{(k-1)}, Q^{(k-1)})$;   (8)
    • Projection of $\underline{\beta}$
        $\underline{\beta}^{(k)} = \text{argmin}_{\beta^s \geq 0} \left\| \underline{\beta} - \hat{\underline{\beta}}^{(k)} \right\|_{l_2}$;   (9)
    • Projection of $Q$
        $Q^{(k)} = \text{argmin}_{Q \in \mathcal{Q}} \left\| Q - \hat{Q}^{(k)} \right\|_{l_2}$;   (10)
Output: Sensor decision rules $Q_s(Z^s \mid X^s)$ for $s = 1, \ldots, S$, and fusion center decision rule $w(\underline{Z})$.
```

Algorithm 2 provides an alternative method (referred to as the Gauss-Seidel method) for solving the non-convex optimization problem (6) with a continuous loss function. Instead of taking $(\underline{\alpha}, \underline{\beta}, Q)$ as one vector and optimizing over all variables together, this algorithm optimizes three types of variables $\underline{\alpha}$, $\underline{\beta}$ and Q alternately and recursively. More specifically, with $\underline{\beta}$ and Q fixed, $\underline{\alpha}$ is updated by gradient descent approach as the objective function G is differentiable over $\underline{\alpha}$ and there is no constraint on $\underline{\alpha}$. With $\underline{\alpha}$ and Q fixed, $\underline{\beta}$ is updated by gradient projection method with a close-form expression as in Algorithm 1. Similarly, with $\underline{\alpha}$ and $\underline{\beta}$ fixed, Q can also be updated by gradient projection method with a close-form expression as explained in Algorithm 1. The convergence of this algorithm is shown in Section IV.C.

Algorithm 2: Decentralized Detection via Regularized Gauss-Seidel Method

```
Input: $S$, $\{y_i, x_i^1, \ldots, x_i^S\}_{i=1}^n$.
Step 0: Initialize $\underline{\alpha} \in \mathcal{R}^N$, $\underline{\beta}$ where $\beta^s \geq 0$ for $s = 1, \ldots, S$, $Q \in \mathcal{Q}$
Step k:
    • Fix $\underline{\beta}^{(k-1)}$ and $Q^{(k-1)}$, for $t_\alpha \leq 2/L$, update
    $\underline{\alpha}^{(k)} = \underline{\alpha}^{(k-1)} - t_\alpha \nabla_{\underline{\alpha}} G(\underline{\alpha}^{(k-1)}, \underline{\beta}^{(k-1)}, Q^{(k-1)})$;   (11)
    • Fix $\underline{\alpha}^{(k)}$ and $Q^{(k-1)}$, for $t_{\underline{\beta}} \leq 1/L$ update
        $\underline{\beta}^{(k)} = \text{argmin}_{\beta^s \geq 0} \Big\| \underline{\beta} - \underline{\beta}^{(k-1)}$
            $+ t_{\underline{\beta}} \nabla_{\underline{\beta}} G(\underline{\alpha}^{(k)}, \underline{\beta}^{(k-1)}, Q^{(k-1)}) \Big\|_{l_2}$;   (12)
    • Fix $\underline{\alpha}^{(k)}$ and $\underline{\beta}^{(k)}$, for $t_Q \leq 1/L$, update
        $Q^{(k)} = \text{argmin}_{Q \in \mathcal{Q}} \Big\| Q - Q^{(k-1)}$
            $+ t_Q \nabla_Q G(\underline{\alpha}^{(k)}, \underline{\beta}^{(k)}, Q^{(k-1)}) \Big\|_{l_2}$;   (13)
Output: Sensor decision rules $Q_s(Z^s \mid X^s)$ for $s = 1, \ldots, S$, and fusion center decision rule $w(\underline{Z})$.
```

We now consider the problem (6) with a non-differentiable loss function such as the hinge loss function. In this case, the gradient-based Algorithms 1 and 2 are not applicable any more. As such, we develop a coordinate descent algorithm (as described in Algorithm 3) for solving the problem (6) with $\phi(\cdot)$ being hinge loss function. We note that the inner loop of Algorithm 3 follows the idea of conjugate duality provided in [9]. Our new ingredient here lies in the outer loop of the algorithm for optimizing the weight parameters $\underline{\beta}$. We describe our algorithm in detail as follows.

(1) *Inner loop*: $\underline{\beta}$ is fixed (i.e., the RKHS is fixed), and the decision rules w and Q are alternatively optimized.

(1a) Optimization over w with Q fixed. As we argue before, the optimal $w = \sum_{i=1}^{N} \alpha_i y_i \Phi'_{\underline{\beta}}(\underline{x}_i)$. For the hinge loss function, it is convenient to apply the conjugate duality argument (i.e., Fenchel Duality) and find the optimal $\underline{\alpha}$ in the dual domain. The dual problem turns out to be a constrained quadratic optimization problem that is easy to solve, and the optimal solution $\underline{\alpha}^*$ takes the following form:

$$\alpha_i^* = \begin{cases} 0 & \hat{\alpha}_i^* \le 0 \\ \hat{\alpha}_i^* & 0 < \hat{\alpha}_i^* < \frac{1}{\lambda_1} \\ \frac{1}{\lambda_1} & \hat{\alpha}_i^* \ge \frac{1}{\lambda_1} \end{cases} \quad \text{for } i = 1, 2, \ldots, N, \tag{14}$$

where [see the equation at the bottom of the page].

(1b) Optimization over Q with w fixed. The subgradient method is used to alternatively update $Q_s(z^s \mid x^s)$ for each sensor s and each value of x^s at a step keeping all other Q values fixed. An element in the subdifferential of the objective function with respect to $Q_s(z^s \mid x^s)$ is given as follows.

$$-\frac{\lambda_1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i y_i \alpha_j y_j \left[\beta^s \left(Q_s\left(z^s \mid x_i^s\right)\mathcal{I}\left(x_j^s = x^s\right) + Q_s\left(z^s \mid x_j^s\right)\mathcal{I}\left(x_i^s = x^s\right)\right)\right] \tag{15}$$

Furthermore, since this is a constrained optimization problem subject to linear constraints on Q, i.e., $\sum_{z^s} Q(z^s \mid x^s) = 1$ for $s = 1, \ldots, S$ and for all possible values of x^s, conditional (sub)gradient method for simplex problems in ([28], Section 2.2.2) can be applied. Alternatively, a projection step as in Algorithms 1 and 2 can be taken to update Q in order to satisfy the constraints.

(2) *Outer loop*: $(\underline{\alpha}, Q)$ are fixed, and the l_1 regularized risk function is optimized over $\underline{\beta}$ in order to find the best weight parameters (i.e., to perform sensor selection).

We apply alternating direction method of multipliers (ADMM) [29]. Since $\underline{\alpha}$ and Q are fixed, we treat them as constants and reformulate our objective function with only the argument $\underline{\beta}$ as follows.

$$G(\underline{\beta}) = \sum_{i=1}^{N} \phi(\langle \underline{\beta}, \underline{d}_i \rangle) + \langle \underline{\beta}, \underline{h} \rangle, \tag{16}$$

where $\underline{d}_i$ is an S-dimensional vector with the s-th entry equals $y_i \sum_{j=1}^{N} \alpha_j y_j \sum_{z^s} Q_s(z^s \mid x_i^s) Q_s(z^s \mid x_j^s)$ and $\underline{h} = \lambda_1 \sum_{i=1}^{n} \alpha_i \underline{d}_i / 2 + \lambda_2 \vec{1}_S$ with $\vec{1}_S = [1, 1, \ldots]^T_{S\times 1}$. Our goal is to optimize the following function using ADMM:

$$F(\underline{\beta}) = G(\underline{\beta}) + i_{\{\beta^s \ge 0,\, s=1,2,\ldots,S\}}(\underline{\beta}) = \sum_{i=1}^{N} g_i(\underline{\beta}) + H(\underline{\beta}), \tag{17}$$

where

$$i_C(x) = \begin{cases} 0 & \text{if } x \in C \\ +\infty & \text{otherwise} \end{cases},$$

$g_i(\underline{\beta}) = \phi(\langle \underline{\beta}, \underline{d}_i \rangle)$, and $H(\underline{\beta}) = \langle \underline{\beta}, \underline{h} \rangle + i_{\{\beta^s \ge 0, s=1,2,\ldots,S\}}(\underline{\beta})$. To apply ADMM, it is desirable that the proximity of each term in $F(\underline{\beta})$ is easy to derive, where the proximity of a function $f(\underline{x})$ is defined as follows:

$$\mathrm{prox}_f(\tilde{\underline{x}}) = \underset{\underline{x}}{\mathrm{argmin}}\, f(\underline{x}) + \frac{1}{2}\|\underline{x} - \tilde{\underline{x}}\|^2.$$

It can be shown that the proximity of each term $g_i(\underline{\beta})$ for $i = 1, 2, \ldots, N$ is given by

$$\mathrm{prox}_{\mu g_i}(\tilde{\underline{\beta}}) = \begin{cases} \tilde{\underline{\beta}} & 1 - \langle \tilde{\underline{\beta}}, \underline{d}_i \rangle \le 0 \\ \frac{1 - \langle \tilde{\underline{\beta}}, \underline{d}_i \rangle}{\|\underline{d}_i\|^2} \underline{d}_i + \tilde{\underline{\beta}} & 0 < 1 - \langle \tilde{\underline{\beta}}, \underline{d}_i \rangle < \mu \|\underline{d}_i\|^2 \\ \tilde{\underline{\beta}} + \mu \underline{d}_i & 1 - \langle \tilde{\underline{\beta}}, \underline{d}_i \rangle \ge \mu \|\underline{d}_i\|^2 \end{cases}, \tag{18}$$

and the proximity of $H(\underline{\beta})$ takes a close-form expression with the s-th component given by:

$$[\mathrm{prox}_{\mu H}(\tilde{\underline{\beta}})]_s = \begin{cases} \tilde{\beta}^s - \mu h_s & \tilde{\beta}^s - \mu h_s \ge 0 \\ 0 & \tilde{\beta}^s - \mu h_s < 0 \end{cases}. \tag{19}$$

Then applying ADMM, we initialize $\underline{\nu}^{(0)} = \underline{\beta}^{(0)}$, $\underline{u}_i^{(0)} = 0$ for $i = 1, 2, \ldots, N$ and provide the iteration steps for optimizing over $\underline{\beta}$ as follows:

$$\begin{cases} \underline{\gamma}_i^{(k)} = \mathrm{prox}_{g_i/\rho}\left(\underline{\nu}^{(k-1)} - \underline{u}_i^{(k-1)}\right) & \text{for } i = 1, 2, \ldots, N \\ \underline{\nu}^{(k)} = \mathrm{prox}_{H/(N\rho)}\left(\bar{\underline{\gamma}}^{(k)} + \bar{\underline{u}}^{(k-1)}\right) & \\ \underline{u}_i^{(k)} = \underline{u}_i^{(k-1)} + \underline{\gamma}_i^{(k-1)} - \underline{\nu}^{(k-1)} & \text{for } i = 1, 2, \ldots, N, \end{cases} \tag{20}$$

where in the second step of updating $\underline{\nu}^{(k)}$, $\bar{\underline{\gamma}}^{(k)} = \frac{1}{N}\sum_{i=1}^{N} \underline{\gamma}_i^{(k)}$ and $\bar{\underline{u}}^{(k-1)} = \frac{1}{N}\sum_{i=1}^{N} \underline{u}_i^{(k-1)}$. When the above algorithm terminates, we set $\underline{\beta} = \bar{\underline{\gamma}}$.

$$\hat{\alpha}_i^* = \frac{1 - \sum_{j \ne i} \alpha_j y_i y_j \left[\sum_{s=1}^{S} \beta^s \sum_{z^s} Q_s\left(z^s \mid x_i^s\right) Q_s\left(z^s \mid x_j^s\right)\right]}{y_i^2 \sum_{s=1}^{S} \beta^s \sum_{z^s} Q_s\left(z^s \mid x_i^s\right) Q_s\left(z^s \mid x_i^s\right)}.$$

```
Algorithm 3: Decentralized Detection for Hinge Loss Function
Input: $S$, $\{y_i, x_i^1, \ldots, x_i^S\}_{i=1}^n$.
Step 0: Initialize $\underline{\alpha} \in \mathcal{R}^N$, $\underline{\beta}$ where $\beta^s \geq 0$ for $s = 1, \ldots, S$, $Q \in \mathcal{Q}$
Step k:
  • Inner loop: fix $\underline{\beta}$, optimize alternatively over $w$ and $Q$
      — Fix all $Q$ functions, compute the optimal $w$ by solving optimal parameters $\underline{\alpha}$ following (14);
      — Fix $w$, compute the optimal $Q(\underline{z} \mid \underline{x})$ using the subgradient method by exploiting (15);
      — Repeat until inner loop converges;
  • Outer loop: fix $\underline{\alpha}$ and $Q$ functions, and compute the optimal $\underline{\beta}$ following (20);
  • Repeat inner and outer loops until converge.
Output: Sensor decision rules $Q_s(Z^s \mid X^s)$ for $s = 1, \ldots, S$, and fusion center decision rule $w(\underline{Z})$.
```

B. Preliminary on Convergence Analysis of Non-Convex Problems

Although it is in general difficult to design algorithms that converge to a global minimizer of a non-convex function, recent results in [30], [31] establish convergence to critical points in non-convex optimization. In this subsection, we introduce the results in [30]–[32] together with necessary definitions, which are useful for studying our algorithms in the next subsection.

We first note that the subdifferential $\partial f(\underline{x})$ plays an important role in convergence analysis for non-convex optimization problems, which can be defined based on Fréchet subdifferential $\hat{\partial} f(\underline{x})$. We refer a reader to [30] for those definitions. We next define critical points based on Fréchet subdifferential.

Definition 3: A point $\underline{x} \in D$ is referred to as a critical point of a function $f : D \to \mathcal{R}$ if $0 \in \partial f(\underline{x})$.

We note that the subdifferential $\partial f(\underline{x})$ in the above definition is for non-convex functions based on Fréchet subdifferential $\hat{\partial} f(\underline{x})$, which is different from the subdifferential for convex functions. We further note that the set of all critical points includes all local optimal solutions of an objective function. Hence, $\underline{x}$ is a critical point of f is a necessary but not sufficient condition for $\underline{x}$ to be a minimizer of f.

In [30], convergence to critical points in non-convex optimization is established for Kurdyka-Łojasiewicz (KL) functions, the definition of which is given below.

Definition 4:

a) The function $f : \mathcal{R}^n \to \mathcal{R} \cup \{+\infty\}$ is said to have Kurdyka-Łojasiewicz (KL) property at $\underline{x}^* \in \mathrm{dom}\partial f$ if there exists $\eta \in (0, +\infty]$, a neighborhood U of $\underline{x}^*$, and a continuous concave function $\psi : [0, \eta) \to \mathcal{R}_+$ such that:
   i) $\psi(0) = 0$,
   ii) ψ is a C^1 function on $(0, \eta)$,
   iii) for all $t \in (0, \eta)$, $\psi'(t) > 0$,
   iv) for all $\underline{x}$ in $U \cap \{\underline{x} : f(\underline{x}^*) < f(\underline{x}) < f(\underline{x}^*) + \eta\}$, the KL inequality holds

$$\psi'(f(\underline{x}) - f(\underline{x}^*))\mathrm{dist}(0, \partial f(\underline{x})) \geq 1,$$

   where $\mathrm{dist}(0, \partial f(\underline{x}))$ denotes the distance from the origin to the set $\partial f(\underline{x})$.

b) Proper lower semicontinuous functions that satisfy KL inequality at each point of $\mathrm{dom}\partial f$ are referred to as KL function.

We further define the type of C^1 function, which appears in the above definition.

Definition 5: The function $f : D \to \mathcal{R}$ is a C^1 function if all partial derivatives of f (i.e., $\frac{\partial f}{\partial x_j}(\underline{x})$ for all j) are continuous at each point in the set D, where $D \subseteq \mathcal{R}^n$ is the domain of the function.

In [30], the convergence of the gradient projection algorithm for constrained non-convex optimization problems is established, which is summarized as follows.

Theorem 2. [30]: Let $h : \mathcal{R}^n \to \mathcal{R}$ be a differentiable function whose gradient is L-Lipschitz continuous, and let C be a nonempty closed subset of $\mathcal{R}^n$. Suppose $\epsilon \in (0, \frac{1}{2L})$ and a sequence of stepsize γ_k satisfy $\epsilon < \gamma_k < \frac{1}{L} - \epsilon$. Consider a sequence $(\underline{x}^k)_{k \in \mathcal{N}}$ that complies with

$$\underline{x}^{k+1} \in P_C(\underline{x}^k - \gamma_k \bigtriangledown h(\underline{x}^k)), \text{ with } \underline{x}^0 \in C. \tag{21}$$

If the function $f = h + i_C$ is a KL function and if $(\underline{x}^k)_{k \in \mathcal{N}}$ is bounded, then the sequence $(\underline{x}^k)_{k \in \mathcal{N}}$ converges to a point $\underline{x}^*$ in C and $\underline{x}^*$ is a critical point of f.

In [30], the convergence of an inexact regularized Gauss-Seidel method is also established, which is summarized as follows.

Theorem 3. [30]: Consider minimization of a function $f : \mathcal{R}^{n_1} \times \cdots \times \mathcal{R}^{n_p} \to \mathcal{R} \bigcup \{+\infty\}$ having the following structure

$$f(\underline{x}) = Q(\underline{x}_1, \ldots, \underline{x}_p) + \sum_{i=1}^{p} f_i(\underline{x}_i), \tag{22}$$

where Q is a C^1 function with locally Lipschitz continuous gradient, and $f_i : \mathcal{R}^{n_i} \to \mathcal{R} \bigcup \{+\infty\}$ is a proper lower semicontinuous function for $i = 1, 2, \ldots, p$. Assume that f defined in (22) is a KL function which is bounded from below. Let $(\underline{x}^k)_{k \in \mathcal{N}}$ be a sequence generated by the following steps:

Step 0: Take $0 < \underline{\lambda} < \bar{\lambda} < \infty$ and $\underline{x}^0 = (\underline{x}_1^0, \ldots, \underline{x}_p^0)$ in $\mathcal{R}^{n_1} \times \cdots \times \mathcal{R}^{n_p}$.

Step k: Find $\underline{x}^{k+1}$ and $\underline{v}^{k+1}$ in $\mathcal{R}^{n_1} \times \cdots \times \mathcal{R}^{n_p}$ such that

$$\begin{aligned} & f_i\left(\underline{x}_i^{k+1}\right) + Q\left(\underline{x}_1^{k+1}, \ldots, \underline{x}_{i-1}^{k+1}, \underline{x}_i^{k+1}, \ldots, \underline{x}_p^k\right) \\ & \quad + \frac{1}{2}\left\langle A_i^k\left(\underline{x}_i^{k+1} - \underline{x}_i^k\right), \underline{x}_i^{k+1} - \underline{x}_i^k\right\rangle \\ & \quad \leq f_i\left(\underline{x}_i^k\right) + Q\left(\underline{x}_1^{k+1}, \ldots, \underline{x}_{i-1}^{k+1}, \underline{x}_i^k, \ldots, \underline{x}_p^k\right); \end{aligned} \tag{23}$$

$$\underline{v}_i^{k+1} \in \partial f_i\left(\underline{x}_i^{k+1}\right); \tag{24}$$

$$\begin{aligned} & \left\|\underline{v}_i^{k+1} + \bigtriangledown_{\underline{x}_i} Q\left(\underline{x}_1^{k+1}, \ldots, \underline{x}_i^{k+1}, \underline{x}_{i+1}^k, \ldots, \underline{x}_p^k\right)\right\| \\ & \quad \leq b_i\left\|\underline{x}_i^{k+1} - \underline{x}_i^k\right\|, \end{aligned} \tag{25}$$

where $i = 1, \ldots, p$, and the sequence of symmetric positive definite matrices (A_i^k) of size n_i have eigenvalues lie in $[\underline{\lambda}, \bar{\lambda}]$. If $(\underline{x}^k)_{k \in \mathcal{N}}$ is bounded, then it converges to some critical point of f.

C. Convergence of Algorithms

In this subsection, we analyze convergence of the algorithms that we propose in Section IV.A. It is clear that the risk function in our minimization problem is not jointly convex over the three types of variables $\underline{\alpha}$, $\underline{\beta}$ and Q. By leveraging recent developments for non-convex optimization problems [30] (see Theorems 2 and 3), we show that Algorithms 1 and 2 converge to critical points of the objective function.

Based on Theorem 2, we can show that Algorithm 1 converges to a critical point of the objective function $G(\underline{\alpha}, \underline{\beta}, Q)$.

Theorem 4: If the loss function $\phi(\,\cdot\,)$ is a real analytic function, $G(\underline{\alpha}, \underline{\beta}, Q)$ is Lipshitz continuous with constant L. Then Algorithm 1 converges to some critical point of $G(\underline{\alpha}, \underline{\beta}, Q)$.

Outline of the Proof: We sketch the main idea of the proof here with the detailed proof relegated to Appendix A.

Since Algorithm 1 adopts the same gradient projection method as the algorithm (21) given in Theorem 2, it is sufficient to show that the objective function

$$F(\underline{\alpha}, \underline{\beta}, Q) = G(\underline{\alpha}, \underline{\beta}, Q) + i_{\{\beta^s \geq 0,\, s=1,2,\ldots,S\}}(\underline{\beta}) + i_{\{Q \in \mathcal{Q}\}}(Q),$$

where $G(\underline{\alpha}, \underline{\beta}, Q)$ is given in (7), satisfies the KL property defined in Definition 4. This can be shown by requiring the loss function $\phi(\,\cdot\,)$ in the objective function to be real analytic (as assumed in the theorem). □

Remark 2: A wide range of functions including both logistic loss and exponential loss functions are real analytic. Thus, convergence of Algorithm 1 established in Theorem 4 is applicable to a large set of loss functions.

To understand the above remark, we introduce the definition of real analytic functions, and a lemma that captures sufficient conditions for a function to be real analytic.

Definition 6. [33]: A function $f(\underline{x})$, with domain on an open subset $U \subseteq \mathcal{R}^m$ and range $\mathcal{R}$, is called real analytic on U, if for each $\hat{\underline{x}} \in U$, the function $f(\underline{x})$ may be represented by a convergent power series in some neighborhood of $\hat{\underline{x}}$.

Hence, a real analytic function is continuous and has continuous and real analytic partial derivatives of all orders [33]. The following lemma provides a simple way to verify real analytic functions.

Lemma 1. [33]: Let $f(\underline{x})$ be infinitely differentiable on some open set $U \in \mathcal{R}^m$. Then $f(\underline{x})$ is real analytic on U if and only if, for each $\hat{\underline{x}} \in U$, there is an open ball V with $\hat{\underline{x}} \in V \subseteq U$, and constants $C > 0$ and $R > 0$ such that the derivatives of $f(\underline{x})$ satisfy

$$\left| \frac{\partial^{|\mu|} f}{\partial \underline{x}^{\mu}}(\underline{x}) \right| \leq C \cdot \frac{\mu!}{R^{|\mu|}}, \quad \forall \underline{x} \in V, \tag{26}$$

where μ is any positive integer.

Following the above lemma, it is easy to check that a wide range of functions including both logistic loss and exponential loss functions are real analytic.

We next consider convergence of Algorithm 2. Since the objective function is uniformly bounded below by zero, Algorithm 2 based on Gauss-Seidel method must converge. Since the risk function is not jointly convex over the three types of variables $\underline{\alpha}$, $\underline{\beta}$ and Q, Algorithm 2 may not converge to a global joint optimal solution. However, based on Theorem 3, we provide convergence of Algorithm 2 to critical points as follows.

Theorem 5: Assume the loss function $\phi(\,\cdot\,)$ in (6) is a real analytic function and is bounded below, $G(\underline{\alpha}, \underline{\beta}, Q)$ is Lipshitz continuous with constant L. Let $(\underline{\alpha}^{(k)}, \underline{\beta}^{(k)}, Q^{(k)})$ be a sequence of variables generated by Algorithm 2. Then the sequence converges to some critical point of $G(\underline{\alpha}, \underline{\beta}, Q)$ given in (7).

Outline of the Proof: We sketch the main idea of the proof here with the detailed proof relegated to Appendix B.

The requirement of real analytic property in Theorem 5 is to guarantee that the objective function satisfies the KL property as in Theorem 4. We further note that the algorithm used in Theorem 3 is referred to as inexact regularized Gauss-Seidel method in [30], which takes a general form and includes a set of algorithms such as the standard Gauss-Seidel method in Algorithm 2 as special cases. Then the major step to prove Theorem 5 lies in showing that the updating steps in Algorithm 2 satisfy the conditions (23), (24), and (25). □

We note that the convergence argument of Algorithm 2 exploits the fact that the objective function takes the structure (22) [30], [34]. For Algorithm 3 developed for the case with the non-differentiable loss function, the objective function cannot be expressed in the form given in (22). Because the loss function including all three types of variables cannot be viewed as the Q function in (22). In this case, it is difficult to establish convergence to a critical point.

D. Performance Analysis

In this section, we study how close the empirical approximate risk function given in (7) that we optimize is to the true risk function. We also provide an upper bound on the probability of decision error based on the risk function, which justifies using such a function as the objective function.

We first define some notations. We let the alphabet sizes of $X^1, \ldots, X^S$ be bounded by L_x, and let the alphabet sizes of the quantized variables $Z^1, \ldots, Z^S$ be bounded by L_z. Let $f = (\underline{\beta}, w, Q)$ denote one set of decision rules, where $\underline{\beta} \in \mathcal{R}^S$ with bounded l_1 norm in RKHS (i.e., $\|\underline{\beta}\|_{l_1} \leq \Gamma_{\beta}$), $w \in \mathcal{H}_{\underline{\beta}}$ with bounded norm (i.e., $\|w\|_{\mathcal{H}_{\beta}} \leq \Gamma_w$), and $Q \in \mathcal{Q}$, which includes all possible conditional probabilities $Q(\underline{z}\,|\,\underline{x})$ that decompose as $Q(\underline{z}\,|\,\underline{x}) = \prod_{s=1}^{S} Q_s(z^s\,|\,x^s)$. Here, the norm constraints on $\underline{\beta}$ and w are justified by the regularization terms in (4). We also let $\underline{\beta}_f$, w_f, and Q_f denote the corresponding components of $f = (\underline{\beta}, w, Q)$.

We let $\mathcal{F}$ denote the set of all functions $f = (\underline{\beta}, w, Q)$ as defined above, which is a subset of $\mathcal{R}^S \times \mathcal{H}_{\underline{\beta}} \times \mathcal{Q}$. In this paper, we particularly consider two special but useful subsets of $\mathcal{F}$: $\mathcal{F}_0$ and $\mathcal{F}_1$. The set $\mathcal{F}_0$ consists all functions with Q in the set $\mathcal{Q}_0$ of deterministic conditional probability distributions. In this case, sensors' decision rules are deterministic. The set $\mathcal{F}_1$ consists of all functions with each component $Q_s(z^s\,|\,x^s)$ having the following property: given any x^s, z^s is uniformly distributed among a subset or a full set of values it can take. For example, suppose the alphabet set of z^s is $\mathcal{Z}^s = \{-1, 0, +1\}$. Given x^s, both $Q(z^s = 0\,|\,x^s) = Q(z^s = +1\,|\,x^s) = 1/2$ and $Q'(z^s = -1\,|\,x^s) = Q'(z^s = 0\,|\,x^s) = 1/2$ are valid in the set $\mathcal{F}_1$. Clearly, such set $\mathcal{Q}_1$ allows randomized decision rules for sensors. Many practically useful decision rules fall as special cases of the above two sets. For example, quantization rules and their randomized versions which are widely used in signal processing fall into the above two sets, respectively.

Bounds on Rademacher Complexity: Rademacher complexity [35] captures the richness of the function class over which our decision rules are chosen, and plays an important role in determining how close the empirical approximate risk function is to the true risk function. Thus, we first provide bounds on this important quantity. We define the Rademacher complexity $R_N(\mathcal{F})$ of the set $\mathcal{F}$ as follows:

$$R_N(\mathcal{F}) := \mathbb{E}_{X,\sigma} \sup_{f\in\mathcal{F}} \left| \frac{1}{N}\sum_{i=1}^{N} \sigma_i f(X_i) \right| \tag{27}$$

where the Rademacher variables $\sigma_1,\ldots,\sigma_N$ are independent and uniformly distributed on $\{-1,+1\}$ and $X_1,\ldots,X_N$ are i.i.d samples generated based on the distribution P_X.

We consider a subset $\tilde{\mathcal{F}} \subset \mathcal{F}$ associated with a $\tilde{\mathcal{Q}} \subset \mathcal{Q}$ of Q functions. We have the following proposition for the case of the weighted count kernel.

Proposition 1: An upper bound on the Rademacher complexity for any $\tilde{\mathcal{F}} \subset \mathcal{F}$ associated with weighted count kernels and with a $\tilde{\mathcal{Q}} \subset \mathcal{Q}$ of Q functions is given by

$$R_N(\tilde{\mathcal{F}}) \le \frac{\Gamma_w\sqrt{\Gamma_{\underline{\beta}}}}{N}\left(N + 2(N-1)\sqrt{N\log|\tilde{\mathcal{Q}}|}\right)^{\frac{1}{2}}, \tag{28}$$

where $|\tilde{\mathcal{Q}}|$ denotes the size of the set $\tilde{\mathcal{Q}}$. In particular, for $\tilde{\mathcal{F}} = \mathcal{F}_0$, the upper bound is given by

$$R_N(\mathcal{F}_0) \le \frac{\Gamma_w\sqrt{\Gamma_{\underline{\beta}}}}{N}\left(N + 2(N-1)\sqrt{NSL_x\log L_z}\right)^{\frac{1}{2}}. \tag{29}$$

For $\tilde{\mathcal{F}} = \mathcal{F}_1$, the upper bound is given by

$$R_N(\mathcal{F}_1) \le \frac{\Gamma_w\sqrt{\Gamma_{\underline{\beta}}}}{N}\left(N + 2(N-1)\sqrt{NSL_zL_x\log 2}\right)^{\frac{1}{2}}. \tag{30}$$

The proof of Proposition 1 is provided in Appendix C.

Remark 3: Rademacher complexity $R_N(\tilde{\mathcal{F}}) \to 0$ as $N \to \infty$ if $\frac{\log|\tilde{\mathcal{Q}}|}{N} \to 0$.

Bounds on True Risk Function: We define three risk functions of interest as follows. Let

$$\hat{\mathbb{E}}\phi(Yw_f(\underline{X})) = \frac{1}{N}\sum_{i=1}^{N}\phi(y_i w_f(\underline{x}_i))$$

denote the *empirical approximate risk*, where the approximation lies in taking the expected value over Z inside the loss function $\phi(\cdot)$ (i.e., the relaxation in (4)). We further let $\hat{f}$ denote its corresponding minimizer, i.e.,

$$\hat{f} = \operatorname*{argmin}_{f\in\mathcal{F}} \hat{\mathbb{E}}\phi(Yw_f(\underline{X})). \tag{31}$$

Let $\mathbb{E}\phi(Yw_f(\underline{X}))$ denote the *expected approximate risk*, and let $\tilde{f}$ denote its corresponding minimizer

$$\tilde{f} = \operatorname*{argmin}_{f\in\mathcal{F}} \mathbb{E}\phi(Yw_f(\underline{X})).$$

The true risk function is $\mathbb{E}\phi(Yw_f(\underline{Z})) = \mathbb{E}_{Y\underline{X}}\mathbb{E}_{\underline{Z}}\phi(Yw_f(\underline{Z}))$, and we let f^* denote its corresponding minimizer, i.e.,

$$f^* = \operatorname*{argmin}_{f\in\mathcal{F}} \mathbb{E}\phi(Yw_f(\underline{Z})).$$

Since we use the empirical approximate risk as the objective function, our approximation lies in two parts: (1) data-dependent objective function (estimation error) (2) taking the expected value over Z inside the loss function $\phi(\cdot)$ (approximation error). We first analyze the estimation error, i.e., we analyze the gap

$$\mathbb{E}\phi(Yw_{\hat{f}}(\underline{X})) - \mathbb{E}\phi(Yw_{\tilde{f}}(\underline{X})),$$

which suggests how close our optimal solution $\hat{f}$ based on the empirical risk is to the optimal solution $\tilde{f}$ based on the expected risk.

Proposition 2: Suppose the logistic or hinge loss function is used, and $\hat{f}$ and $\tilde{f}$ are minimizers over $\tilde{\mathcal{F}}$. Then for any small $0 < \delta < 1$, with probability larger than $1-\delta$,

$$\begin{aligned}&\mathbb{E}\phi(Yw_{\hat{f}}(\underline{X})) - \mathbb{E}\phi(Yw_{\tilde{f}}(\underline{X}))\\ &\qquad \le 4R_N(\mathcal{F}) + 2(1+\Gamma_w\sqrt{\Gamma_{\underline{\beta}}})\sqrt{\frac{2\log\frac{1}{\delta}}{N}}.\end{aligned} \tag{32}$$

The proof of Proposition 2 is provided in Appendix D.

Remark 4: Following from Proposition 1, if $\frac{\log|\tilde{\mathcal{Q}}|}{N} \to 0$ as $N \to \infty$, then $R_N(\mathcal{F}) \to 0$ as $N \to \infty$. In this case, Proposition 2 implies that the estimation error is asymptotically small with high probability. Furthermore, for the cases with $\tilde{\mathcal{Q}} = \mathcal{Q}_0$ and $\tilde{\mathcal{Q}} = \mathcal{Q}_1$, the above condition becomes $\frac{S}{N} \to 0$ as $N \to \infty$. Namely, if the number of sensors does not scale as fast as the number of samples, the estimation error is asymptotically small with high probability.

We next study the gap between the empirical approximate risk and the true risk (including both estimation and approximation errors). We let

$$\hat{f}_0 = \arg\min_{f\in\mathcal{F}_0} \hat{\mathbb{E}}\phi(Yw_f(\underline{X})),$$

which is the decision rule that optimizes the empirical approximate risk over the set $\mathcal{F}_0$. It can be shown (as in [9]) that with a probability at least $1-2\delta$, the true risk is bounded by the empirical approximate risk as follows:

$$\begin{aligned}&\hat{\mathbb{E}}\phi(Yw_{\hat{f}}(\underline{X})) - 2L_\phi R_N(\mathcal{F}) - \Gamma_\phi\sqrt{\frac{2\log\frac{1}{\delta}}{N}}\\ &\quad \le \mathbb{E}\phi(Yw_{f^*}(\underline{Z}))\\ &\quad \le \hat{\mathbb{E}}\phi(Yw_{\hat{f}_0}(\underline{X})) + 2L_\phi R_N(\mathcal{F}_0) + \Gamma_\phi\sqrt{\frac{2\log\frac{1}{\delta}}{N}},\end{aligned} \tag{33}$$

where L_ϕ is the Lipschitz constant of $\phi(\cdot)$, and Γ_ϕ is a uniform bound on $\phi(\cdot)$. It is clear that the bounds on the Rademacher complexity characterize how close the empirical approximate risk function is to the true risk.

Remark 5: Following Proposition 1 and Remark 3, the optimal empirical approximate risk serves as good lower and upper bounds if $\frac{\log|\tilde{\mathcal{Q}}|}{N} \to 0$ as $N \to \infty$.

Bounds on Error Probability: The basic performance measure for the problem of decentralized detection is the probability of decision error, which is not computable in the nonparametric case. We next provide a connection between the probability of decision error and the risk function.

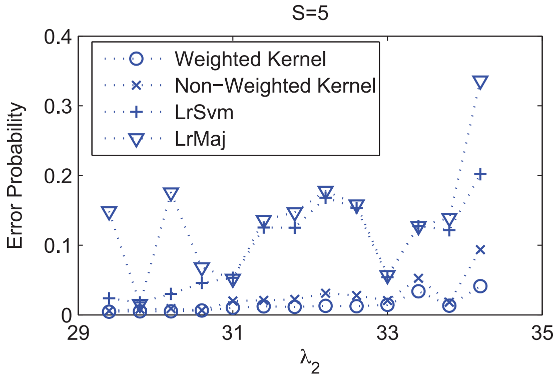


Fig. 2. Comparison of probabilities of error among four approaches.

Proposition 3: With a probability at least $1 - \delta$, the probability of error based on the weighted count kernel is respectively bounded by the risk functions based on logistic loss $\phi_l(\cdot)$ and hinge loss $\phi_h(\cdot)$ as follows:

$$P(Yw_{f^*}(\underline{Z}) < 0) \le \frac{1}{\log 2}\mathbb{E}\phi(Yw_{f^*}(\underline{Z}))$$
$$\le \frac{1}{\log 2}\left[\hat{\mathbb{E}}\phi_l(Yw_{\hat{f}_0}(\underline{X})) + 2R_N(\mathcal{F}_0) + \Gamma_\phi\sqrt{\frac{2\log\frac{1}{\delta}}{N}}\right], \quad (34)$$

and

$$P(Yw_{f^*}(\underline{Z}) < 0) \le \mathbb{E}\phi(Yw_{f^*}(\underline{Z}))$$
$$\le \hat{\mathbb{E}}\phi_h(Yw_{\hat{f}_0}(\underline{X})) + 2R_N(\mathcal{F}_0) + \Gamma_\phi\sqrt{\frac{2\log\frac{1}{\delta}}{N}} \quad (35)$$

where $R_N(\mathcal{F}_0)$ is bounded in (29) and $\Gamma_\phi = 1 + \Gamma_w\sqrt{\Gamma_{\underline{\beta}}}$.

Proof: Due to the property of the hinge loss function,

$$P(Yw_{f^*}(\underline{Z}) < 0) = \mathbb{E}\mathbb{I}[Yw_{f^*}(\underline{Z}) < 0] \le \mathbb{E}\phi(Yw_{f^*}(\underline{Z})). \quad (36)$$

Then applying (33), we obtain the desired bound. If the logistic loss is used, then we obtain the bound by following the above steps except noticing that

$$\mathbb{E}\mathbb{I}[Yw_{f^*}(\underline{Z}) < 0] \le \frac{1}{\log 2}\mathbb{E}\phi(Yw_{f^*}(\underline{Z})). \quad (37)$$

□

The above Proposition implies that as the number of samples becomes large (and if $R_N(\mathcal{F}_0) \to 0$), the true risk and the empirical risk (or a scaled version of it) serve as upper bounds on the probability of decision error. This connection justifies using these risk functions as the objective function.

V. Numerical Results

In this section, we demonstrate the performance of our approach and its associated properties based on the following experiments.

The joint distribution of the event and observations are chosen as follows. (Such distribution is chosen for generating data samples, and is not exploited in designing decision rules.) In our experiment, the state of the event y takes two values $+1$ and -1 with equal probability, and the sensors' measurements x^s for $s = 1, \ldots, S$ are noisy versions of y, i.e., $x^s = y + n^s$, where the noise variable n^s can take three values $\{-1, 0, +1\}$. It is clear that even if $n^s = +1$, there is only half probability that the observation x^s causes confusion about y, because when $y = +1$, there is no confusion. The case when $n^s = -1$ is similar. In all numerical results, we assume that $P(n^s = -1) = P(n^s = +1)$, and introduce a quantity of probability of uncertainty (POU) that equals $P(n^s = +1)$ for representing the quality of sensor's observations. For example, if n^s has the distribution such that $P(n^s = 0) = 0.5$, $P(n^s = +1) = 0.25$ and $P(n^s = -1) = 0.25$, then $\text{POU} = 0.25$ indicating the probabilities that observations confuse about the event state.

A. Comparison With Other Approaches

In this subsection, we compare our approach with the following three competitive test methods.

- Likelihood-ratio majority voting (LrMV): each sensor s computes $\hat{P}(X^s = t \,|\, Y = 1)/\hat{P}(X^s = t \,|\, Y = -1)$ for each value that X^s can take based on training samples, and then sends $+1$ to the fusion center if the ratio is greater than 1 for the received observation, and sends -1 otherwise. The fusion center's decision rule is based on majority voting of sensors' decisions.
- Likelihood-ratio support vector machine (LrSVM): each sensor performs the same likelihood-ratio test as in LrMV and transmits the compressed Z^s to the fusion center. The fusion center's decision rule is based on support vector machine method with training samples $\{Z_i^1\}_{i=1}^N, \ldots, \{Z_i^S\}_{i=1}^N$.
- Uniform-weighted kernel (Uniform kernel): similar to our weighted kernel method with weight parameters $\beta^s = 1$ for all sensors $s = 1, 2, \ldots, S$ as in [9].

In this experiment, we choose logistic function as loss function and apply Algorithm 2. To compare our approach with the above methods, we generate the same training and testing samples for all approaches. We first perform the weighted kernel method using, which produces the selected sensors. For the LrMV, LrSVM, and Uniform kernel methods, the fusion center collects decisions only from sensors that have already been selected by the weighted kernel method for a fair comparison. Fig. 2 plots the error probabilities for all approaches, and clearly demonstrates that our weighted kernel based approach outperforms all other competitive methods.

B. Performance on Sensor Selection

As described in the previous sections, sensor selection is performed via kernel weight parameter $\underline{\beta}$ selection, and is jointly designed with the sensors' local decision rules Q and the fusion center's decision rule w. In this subsection, we study how sensor selection affects the performance of the system in such joint design, i.e., the joint optimization over $(\underline{\alpha}, \underline{\beta}, Q)$. In the following experiments, we apply Algorithm 1.

We first study how the regularization parameter λ_2 controls the number of sensors selected, i.e., sparsity of sensor selection. We study a network with $S = 40$ sensors which have independent observations. For each λ_2, we let the value of POU of sensors gradually increase from sensors $s = 1$ to $s = S$ as the index s increases. Hence, the sensors' measurement quality

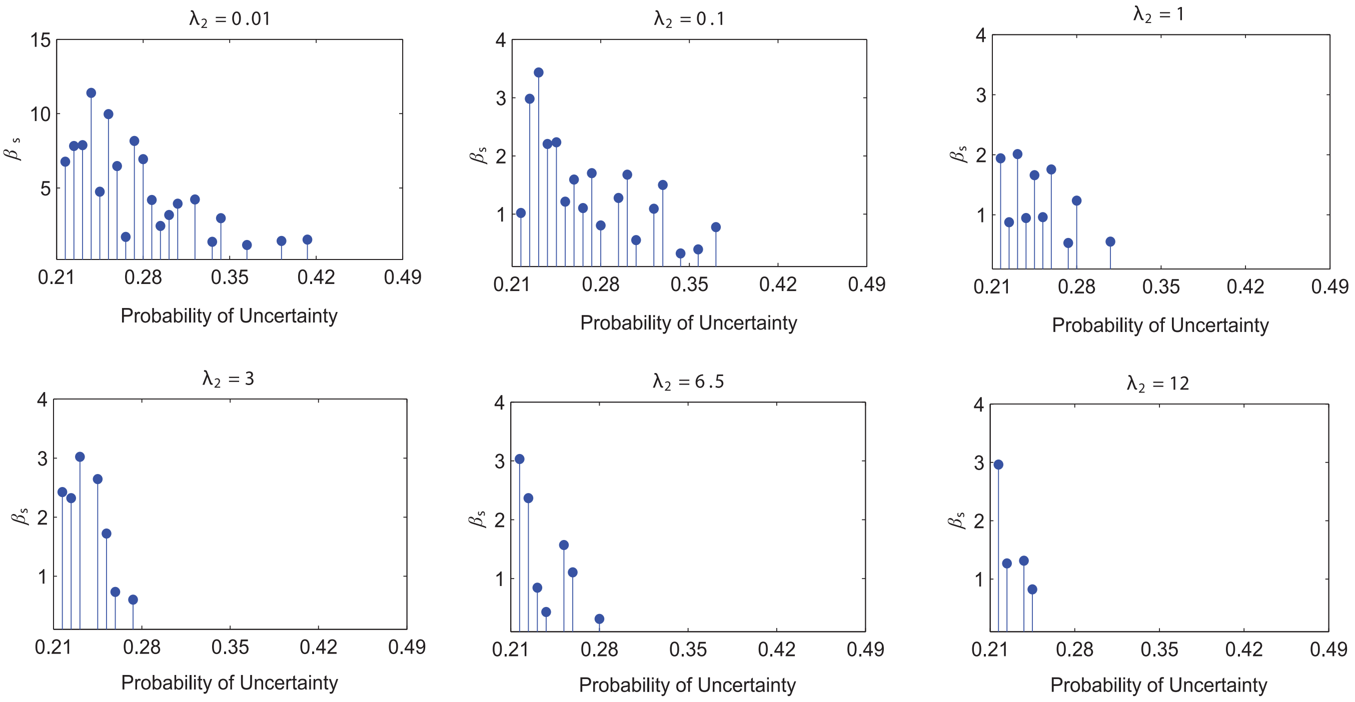


Fig. 3. Impact of λ_2 on sparsity of sensor selection.

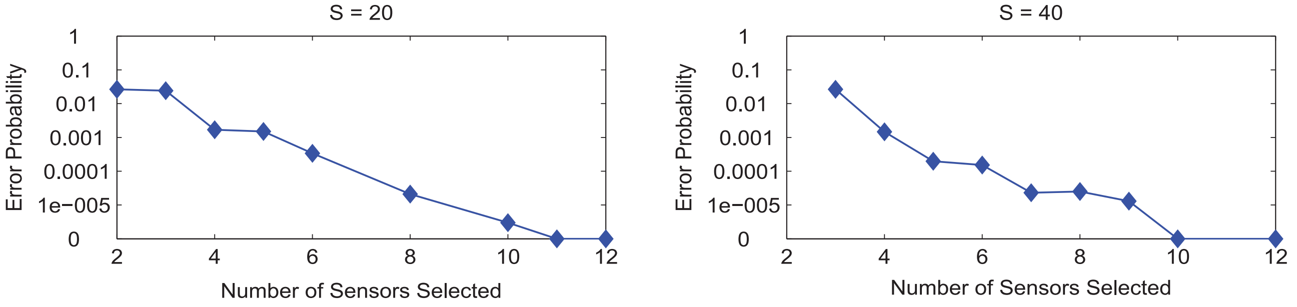


Fig. 4. Impact of sparsity on error probability.

reduces as the index of sensors increases. Fig. 3 provides the optimal weight parameters versus POUs (i.e., versus sensors) for a number of values of λ_2. It is clear for each value of λ_2, sensors with smaller values of POU (i.e., better quality of observations) are assigned higher weight parameters, suggesting these sensors are more contributive in the fusion center's decision rule. In particular, nonzero weight parameters are assigned to sensors with better quality. This is reasonable because if only limited sensors are selected to participate in decision making, selected sensors should have better observation quality. Furthermore, as the value of λ_2 increases, less sensors are chosen (with nonzero weight parameters β^s) indicating that the regularization parameter λ_2 indeed can control the sensor selection sparsity.

We next study the influence of sparsity of sensor selection on the performance (i.e., the testing error probability). In Fig. 4, we plot the testing error probability versus the number of sensors selected. It can be seen that as more sensors are selected, the error probability decreases, because more sensors better clarify the fusion center's decision. However, it is also clear from the figure that even a small fraction of sensors already guarantee small probability of error. For example, when $S = 40$, with 25% of sensors selected, the error probability is already almost zero, and furthermore, with only 10% of sensors selected, the error probability is 10^{-3}. This suggests that selecting only a small fraction of sensors for decision making does not sacrifice much performance but can save a large amount of communication resources.

We are also interested in applying our approach to scenarios, in which sensors are clustered into groups with sensors in the same group having highly correlated observations. In our experiment, sensors are divided into groups with the same size, and each group has a representative sensor. Within each group, each sensor has probability 0.8 to have the same observation with the representative sensor, and probability 0.2 to have an independent observation. Observations across different groups are independent. We set $\lambda_2 = 4, 5, 7$ respectively for groups with sizes 2, 3, 4. In Fig. 5, we plot the weight parameters versus sensor indices. Furthermore, group numbers such as G_1 and G_2 are also marked below the sensor indices indicating which group corresponding sensor belongs to. It can be seen that for most groups, only one sensor has nonzero weight, and is hence selected. This demonstrates that our sensor selection approach based on the weighted kernel is very effective to remove redundant data and achieve dimension reduction, thus significantly

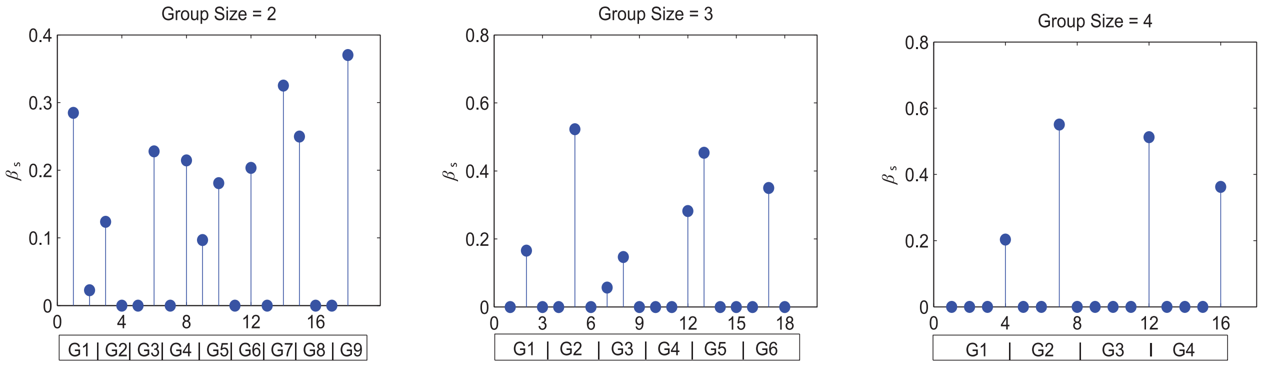


Fig. 5. Impact of sample correlation on sensor selection in clustered sensor networks.

saving resources for communication from sensors to the fusion center. We further note that by adjusting values of λ_2, it is also possible that entire groups are eliminated or more than one sensors are selected in one group depending on the sparsity level that we want to achieve.

VI. Conclusion

In this paper, we investigated the problem of nonparametric decentralized detection. Adapting the kernel-based framework [9] proposed by Nguyen, Wainwright, and Jordan, we introduced the idea of using weighted kernel to generalize the approach to heterogeneous networks. In particular, the kernel weight parameters serve to selectively incorporate sensors' information into the fusion center's decision rule based on quality of sensors' observations. Furthermore, via l_1 regularization, weight parameters also serve as sensor selection parameters with nonzero parameters corresponding to sensors being selected. We designed two algorithms to solve the joint optimization of weight parameters and sensors' and fusion center's decision rules, and showed the convergence of the algorithms. We also demonstrated the performance of our approach via numerical experiments.

Further generalization of this study can be along several directions. Multi-level sensor networks is an interesting topic, and it is expected that sensor selection is also related to network structures in this case. The problems of multiple-event detection is also interesting to be explored in the nonparametric scenario. The idea of using l_1 regularization for sensor selection can also be applied to studying parametric models such as hypothesis testing and parameter estimation.

Appendix A
Proof of Theorem 4

Since Algorithm 1 uses the standard projection method as described in Theorem 2, it suffices to show that $F(\underline{\alpha}, \underline{\beta}, Q) = G(\underline{\alpha}, \underline{\beta}, Q) + i_{\{\beta^s \geq 0, s=1,2,\ldots,S\}}(\underline{\beta}) + i_{\{Q \in \mathcal{Q}\}}(Q)$ is a KL function, where $G(\underline{\alpha}, \underline{\beta}, Q)$ is defined in (7).

It is shown in [36] that subanalytic functions have the KL property. Hence, in order to prove that $F(\underline{\alpha}, \underline{\beta}, Q)$ is a KL function, it suffices to show that it is a subanalytic function. It is also shown in [37] that the sum of subanalytic functions is still a subanalytic function. Hence, it suffices to show that each term of $F(\underline{\alpha}, \underline{\beta}, Q)$ is a subanalytic function.

We next introduce the definition of subanalytic functions and special cases of such functions, which are useful in proof.

Definition 7. [38] (Subanalytic Function): A subset $\mathcal{D} \in \mathcal{R}^n$ is called subanalytic if each point of $\mathcal{D}$ admits a neighborhood V for which $\mathcal{D} \bigcap V$ can be represented as

$$\mathcal{D} \bigcap V = \{\underline{x} \in \mathcal{R}^n : (\underline{x}, \underline{y}) \in U\},$$

where U is a bounded semi-analytic subset of $\mathcal{R}^n \times \mathcal{R}^m$ for some $m \geq 1$. A function $f : \mathcal{R}^n \rightarrow \mathcal{R} \bigcup \{+\infty\}$ is called subanalytic if its graph $\{(\underline{x}, \lambda) \in \mathcal{R}^{n+1} : f(\underline{x}) = \lambda\}$ is a subanalytic set.

Definition 8. [30] (Semi-Algebraic Function): A subset $\mathcal{D} \in \mathcal{R}^n$ is called semi-algebraic if it can be represented as

$$\mathcal{D} = \bigcup_{i=1}^{p} \bigcap_{j=1}^{q} \{\underline{x} \in \mathcal{R}^n : p_{ij}(\underline{x}) = 0,\ q_{ij}(\underline{x}) > 0\},$$

where $p_{ij}, q_{ig} : \mathcal{R}^n \rightarrow \mathcal{R}$ are real polynomial functions for $1 \leq i \leq p$, $1 \leq j \leq q$. A function $f : \mathcal{R}^n \rightarrow \mathcal{R} \bigcup \{+\infty\}$ is called semi-algebraic if its graph $\{(\underline{x}, \lambda) \in \mathcal{R}^{n+1} : f(\underline{x}) = \lambda\}$ is a semi-algebraic subset of $\mathcal{R}^{n+1}$.

Definition 9. [38] (Semi-Analytic Function): A subset $\mathcal{D} \in \mathcal{R}^n$ is called semi-analytic if each point of $\mathcal{D}$ admits a neighborhood V for which $\mathcal{D} \bigcap V$ can be represented as

$$\mathcal{D} \bigcap V = \bigcup_{i=1}^{p} \bigcap_{j=1}^{q} \{\underline{x} \in V : p_{ij}(\underline{x}) = 0,\ q_{ij}(\underline{x}) > 0\},$$

where $p_{ij}, q_{ig} : V \rightarrow \mathcal{R}$ are real analytic functions (see Definition 6) for $1 \leq i \leq p$, $1 \leq j \leq q$. A function $f : \mathcal{R}^n \rightarrow \mathcal{R} \bigcup \{+\infty\}$ is called semi-analytic if its graph $\{(\underline{x}, \lambda) \in \mathcal{R}^{n+1} : f(\underline{x}) = \lambda\}$ is a semi-analytic set.

We note that a real polynomial function must be a real analytic function and hence a semi-algebraic function is semi-analytic. It is also clear from Definition 9 that a real-analytic function is also semi-analytic. It is shown in [39] that any semi-analytic function is subanalytic. Thus any real analytic, semi-algebraic, or semi-analytic function is subanalytic.

Based on the above property, it suffices to show that each term of $F(\underline{\alpha}, \underline{\beta}, Q)$ is real analytic, semi-algebraic or semi-analytic. The first term given below

$$\sum_{i=1}^{N} \phi\left(y_i \sum_{j=1}^{N} \alpha_j y_j \left[\sum_{s=1}^{S} \beta^s \sum_{z^s} Q_s\left(z^s \mid x_i^s\right) Q_s\left(z^s \mid x_j^s\right)\right]\right)$$

is composition of a real analytic loss function $\phi(\cdot)$ and a polynomial function, which is also real analytic. It has been shown in [33] that the composition of real analytic functions is also real analytic. Therefore the above term is real analytic. It is also clear that the term $\sum_{s=1}^{S}\beta^s$ and the term

$$\frac{\lambda_1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i y_i \alpha_j y_j\left[\sum_{s=1}^{S}\beta^s\sum_{z^s}Q_s\left(z^s\mid x_i^s\right)Q_s\left(z^s\mid x_j^s\right)\right]$$

are both real polynomial, and hence are both real analytic.

Furthermore, it is also clear that the indicator function $i_{\{\beta^s\geq 0,\, s=1,2,\ldots,S\}}(\underline{\beta})$ is semi-algebraic, because its graph is $\{(\underline{\beta},\lambda)\in\mathcal{R}^{n+1}:\beta^s\geq 0,\lambda=0\}$. Similarly, $i_{\{Q\in\mathcal{Q}\}}(Q)$ is also semi-algebraic. Therefore, each term of $F(\underline{\alpha},\underline{\beta},Q)$ is a subanalytic function, which implies that $F(\underline{\alpha},\underline{\beta},Q)$ is a KL function. This concludes the proof.

Appendix B
Proof of Theorem 5

The proof apply the convergence result on proximal regularization of Gauss-Seidel method (see Theorem 3). It has been shown that $F(\underline{\alpha},\underline{\beta},Q)$ is a KL function in Appendix A and it is clear that the function is bounded below. It is also clear that $G(\underline{\alpha},\underline{\beta},Q)$ is a C^1 function. It is then sufficient to check that the conditions (23), (24), and (25) in Theorem 3 are satisfied when updating $\underline{\alpha}$, $\underline{\beta}$, and Q.

We first note that in the context of Theorem 3, $Q(\underline{x}_1,\ldots,\underline{x}_p)=G(\underline{\alpha},\underline{\beta},Q)$ with $p=3$, $\underline{x}_1=\underline{\alpha}$, $\underline{x}_2=\underline{\beta}$, and $\underline{x}_3=Q$, $f_1(\underline{x}_1)=0$, $f_2(\underline{x}_2)=i_{\{\beta^s\geq 0,\, s=1,2,\ldots,S\}}(\underline{\beta})$, and $f_3(\underline{x}_3)=i_{\{Q\in\mathcal{Q}\}}(Q)$.

We then introduce the following lemma to help our proof.

Lemma 2: Let $f:\mathcal{R}^n\to\mathcal{R}$ be a C^1 function and Lipschitz continuous over a set C with the constant L. Then for any two points x,z in C,

$$f(z)\leq f(x)+\langle\nabla f(x),z-x\rangle+\frac{L}{2}\|z-x\|^2. \tag{38}$$

Verifying the Conditions for Updating $\underline{\alpha}$: Step (11) implies that

$$\nabla_{\underline{\alpha}}G\left(\underline{\alpha}^{(k)},\underline{\beta}^{(k)},Q^{(k)}\right)=\frac{1}{t_\alpha}\left(\underline{\alpha}^{(k)}-\underline{\alpha}^{(k+1)}\right). \tag{39}$$

Therefore,

$$\left\|\nabla_{\underline{\alpha}}G\left(\underline{\alpha}^{(k)},\underline{\beta}^{(k)},Q^{(k)}\right)\right\|=\frac{1}{t_\alpha}\left\|\underline{\alpha}^{(k)}-\underline{\alpha}^{(k+1)}\right\|,$$

which implies that (25) is satisfied by setting $\underline{v}_\alpha^{(k+1)}=0$. It is also clear that such $\underline{v}_\alpha^{(k+1)}$ satisfies (24) with $f_1(\underline{x}_1)=0$.

Using Lemma 2, we can show that

$$\begin{aligned}G\left(\underline{\alpha}^{(k+1)},\underline{\beta}^{(k)},Q^{(k)}\right)-\frac{L}{2}\left\|\underline{\alpha}^{(k+1)}-\underline{\alpha}^{(k)}\right\|^2&\\+\left\langle\nabla_{\underline{\alpha}}G\left(\underline{\alpha}^{(k)},\underline{\beta}^{(k)},Q^{(k)}\right),\underline{\alpha}^{(k)}-\underline{\alpha}^{(k+1)}\right\rangle&\\\leq G\left(\underline{\alpha}^{(k)},\underline{\beta}^{(k)},Q^{(k)}\right).&\end{aligned} \tag{40}$$

Substituting $\nabla_{\underline{\alpha}}G(\underline{\alpha}^{(k)},\underline{\beta}^{(k)},Q^{(k)})$ in (39) into (40), we obtain

$$\begin{aligned}G\left(\underline{\alpha}^{(k+1)},\underline{\beta}^{(k)},Q^{(k)}\right)+\left(\frac{1}{t_\alpha}-\frac{L}{2}\right)\left\|\underline{\alpha}^{(k+1)}-\underline{\alpha}^{(k)}\right\|^2&\\\leq G\left(\underline{\alpha}^{(k)},\underline{\beta}^{(k)},Q^{(k)}\right).&\end{aligned} \tag{41}$$

Since $t_\alpha\leq 2/L$, the coefficient $\frac{1}{t_\alpha}-\frac{L}{2}$ is a positive constant when k varies, which guarantees that (23) holds with $A_i^k=(\frac{1}{t_\alpha}-\frac{L}{2})I$.

Verifying the Conditions for Updating $\underline{\beta}$: Following (12), we obtain

$$\begin{aligned}\left\|\underline{\beta}^{(k+1)}-\underline{\beta}^{(k)}+t_{\underline{\beta}}\nabla_{\underline{\beta}}G\left(\underline{\alpha}^{(k+1)},\underline{\beta}^{(k)},Q^{(k)}\right)\right\|&\\\leq\left\|t_{\underline{\beta}}\nabla_{\underline{\beta}}G\left(\underline{\alpha}^{(k+1)},\underline{\beta}^{(k)},Q^{(k)}\right)\right\|.&\end{aligned} \tag{42}$$

Hence,

$$\begin{aligned}&\left\|\underline{\beta}^{(k+1)}-\underline{\beta}^{(k)}\right\|^2\\&+2\left\langle\underline{\beta}^{(k+1)}-\underline{\beta}^{(k)},t_{\underline{\beta}}\nabla_{\underline{\beta}}G\left(\underline{\alpha}^{(k+1)},\underline{\beta}^{(k)},Q^{(k)}\right)\right\rangle\leq 0.\end{aligned} \tag{43}$$

Using Lemma 2, we can show that

$$\begin{aligned}G\left(\underline{\alpha}^{(k+1)},\underline{\beta}^{(k+1)},Q^{(k)}\right)-\frac{L}{2}\left\|\underline{\beta}^{(k+1)}-\underline{\beta}^{(k)}\right\|^2&\\+\left\langle\nabla_{\underline{\beta}}G\left(\underline{\alpha}^{(k+1)},\underline{\beta}^{(k)},Q^{(k)}\right),\underline{\beta}^{(k)}-\underline{\beta}^{(k+1)}\right\rangle&\\\leq G\left(\underline{\alpha}^{(k+1)},\underline{\beta}^{(k)},Q^{(k)}\right).&\end{aligned} \tag{44}$$

Combining with (43), we obtain

$$\begin{aligned}G\left(\underline{\alpha}^{(k+1)},\underline{\beta}^{(k+1)},Q^{(k)}\right)+\left(\frac{1}{2t_{\underline{\beta}}}-\frac{L}{2}\right)\left\|\underline{\beta}^{(k+1)}-\underline{\beta}^{(k)}\right\|^2&\\\leq G\left(\underline{\alpha}^{(k+1)},\underline{\beta}^{(k)},Q^{(k)}\right).&\end{aligned} \tag{45}$$

By choosing $t_{\underline{\beta}}\leq 1/L$, the update on $\underline{\beta}$ satisfies the condition (23) with $A_i^k=(\frac{1}{t_{\underline{\beta}}}-\frac{L}{2})I$.

We define the feasible space of $\underline{\beta}$ as $C_{\underline{\beta}}=\{\underline{\beta}:\beta^s\geq 0\text{ for }s=1,2,\ldots,S\}$. The updating step (12) can be equivalently written as

$$\begin{aligned}\underline{\beta}^{(k+1)}=\operatorname{argmin}_{\underline{\beta}}\frac{1}{2t_{\underline{\beta}}}&\left\|\underline{\beta}-\underline{\beta}^{(k)}\right.\\&\left.+t_{\underline{\beta}}\nabla_{\underline{\beta}}G\left(\underline{\alpha}^{(k+1)},\underline{\beta}^{(k)},Q^{(k)}\right)\right\|^2+i_{C_{\underline{\beta}}}(\underline{\beta}).\end{aligned} \tag{46}$$

The problem (46) implies that the solution $\underline{\beta}^{(k+1)}$ satisfies the following property:

$$\begin{aligned}0\in\partial i_{C_{\underline{\beta}}}\left(\underline{\beta}^{(k+1)}\right)+\nabla_{\underline{\beta}}G\left(\underline{\alpha}^{(k+1)},\underline{\beta}^{(k)},Q^{(k)}\right)&\\+\frac{1}{t_{\underline{\beta}}}\left(\underline{\beta}^{(k+1)}-\underline{\beta}^{(k)}\right)&\end{aligned} \tag{47}$$

Hence, there exists $\underline{v}_{\underline{\beta}}^{(k+1)}\in\partial i_{C_{\underline{\beta}}}(\underline{\beta}^{(k+1)})$ such that

$$\begin{aligned}\underline{v}_{\underline{\beta}}^{(k+1)}+\nabla_{\underline{\beta}}G\left(\underline{\alpha}^{(k+1)},\underline{\beta}^{(k)},Q^{(k)}\right)+\frac{1}{t_{\underline{\beta}}}\left(\underline{\beta}^{(k+1)}-\underline{\beta}^{(k)}\right)&\\=0.&\end{aligned}$$

We hence have

$$\begin{aligned}\left\|\underline{v}_{\underline{\beta}}^{(k+1)}+\nabla_{\underline{\beta}}G\left(\underline{\alpha}^{(k+1)},\underline{\beta}^{(k)},Q^{(k)}\right)\right\|&\\=\frac{1}{t_{\underline{\beta}}}\left\|\left(\underline{\beta}^{(k+1)}-\underline{\beta}^{(k)}\right)\right\|.&\end{aligned} \tag{48}$$

We further derive

$$\begin{aligned}
&\left\|\underline{v}_{\underline{\beta}}^{(k+1)}+\nabla_{\underline{\beta}}G\left(\underline{\alpha}^{(k+1)},\underline{\beta}^{(k+1)},Q^{(k)}\right)\right\|\\
&\quad\le\left\|\underline{v}_{\underline{\beta}}^{(k+1)}+\nabla_{\underline{\beta}}G\left(\underline{\alpha}^{(k+1)},\underline{\beta}^{(k)},Q^{(k)}\right)\right\|\\
&\qquad+\left\|\nabla_{\underline{\beta}}G\left(\underline{\alpha}^{(k+1)},\underline{\beta}^{(k+1)},Q^{(k)}\right)\right.\\
&\qquad\left.-\nabla_{\underline{\beta}}\,G\left(\underline{\alpha}^{(k+1)},\underline{\beta}^{(k)},Q^{(k)}\right)\right\|\\
&\quad\le\frac{1}{t_{\underline{\beta}}}\left\|\left(\underline{\beta}^{(k+1)}-\underline{\beta}^{(k)}\right)\right\|+L\left\|\left(\underline{\beta}^{(k+1)}-\underline{\beta}^{(k)}\right)\right\|. \qquad (49)
\end{aligned}$$

where the last step follows from (48) and the fact that $G(\underline{\alpha},\underline{\beta},Q)$ is Lipshitz continuous with constant L. Therefore, the updating step on $\underline{\beta}$ satisfies the conditions (24) and (25).

Verifying the conditions for updating Q follows the steps similar to those for $\underline{\beta}$. This concludes the proof.

Appendix C
Proof of Proposition 1

We note that $w_f(\underline{X}_i)=\langle w_f,\Phi'_{\underline{\beta}}(\underline{X}_i)\rangle_{\mathcal{H}_{\underline{\beta}}}$, and obtain the following upper bound.

$$\begin{aligned}
&R_N(\tilde{\mathcal{F}})\\
&\quad=\mathbb{E}\sup_{f\in\tilde{\mathcal{F}}}\left|\frac{1}{N}\sum_{i=1}^{N}\sigma_i w_f(\underline{X}_i)\right|\\
&\quad=\mathbb{E}\sup_{\|\underline{\beta}\|_1\le\Gamma_{\underline{\beta}},\|w\|_{\mathcal{H}_{\underline{\beta}}}\le\Gamma_w,Q\in\tilde{\mathcal{Q}}}\left|\frac{1}{N}\sum_{i=1}^{N}\sigma_i\langle w,\Phi'_{\underline{\beta}}(\underline{X}_i)\rangle_{\mathcal{H}_{\underline{\beta}}}\right|\\
&\quad\overset{(a)}{\le}\frac{\Gamma_w}{N}\mathbb{E}\sup_{\|\underline{\beta}\|_1\le\Gamma_{\underline{\beta}},Q\in\tilde{\mathcal{Q}}}\left\|\sum_{i=1}^{N}\sigma_i\Phi'_{\underline{\beta}}(\underline{X}_i)\right\|_{\mathcal{H}_{\underline{\beta}}}\\
&\quad\overset{(b)}{\le}\frac{\Gamma_w}{N}\sqrt{\mathbb{E}\sup_{\|\underline{\beta}\|_1\le\Gamma_{\underline{\beta}},Q\in\tilde{\mathcal{Q}}}\left\|\sum_{i=1}^{N}\sigma_i\Phi'_{\underline{\beta}}(\underline{X}_i)\right\|^2_{\mathcal{H}_{\underline{\beta}}}}\\
&\quad=\frac{\Gamma_w}{N}\left(\mathbb{E}\sup_{\|\underline{\beta}\|_1\le\Gamma_{\underline{\beta}},Q\in\tilde{\mathcal{Q}}}\sum_{i=1}^{N}\|\Phi'_{\underline{\beta}}(\underline{X}_i)\|^2_{\mathcal{H}_{\underline{\beta}}}\right.\\
&\qquad\left.+2\mathbb{E}\sup_{\|\underline{\beta}\|_1\le\Gamma_{\underline{\beta}},Q\in\tilde{\mathcal{Q}}}\sum_{1\le i<j\le N}\sigma_i\sigma_j\langle\Phi'_{\underline{\beta}}(\underline{X}_i),\Phi'_{\underline{\beta}}(\underline{X}_j)\rangle_{\mathcal{H}_{\underline{\beta}}}\right)^{\frac{1}{2}}
\end{aligned}\qquad (50)$$

where the step (a) follows from the Cauchy-Schwartz inequality, and (b) follows from the Jensen's inequality.

For the first term in (50), we have the following bound for any realization of $\underline{x}_i$

$$\begin{aligned}
&\sup_{\|\underline{\beta}\|_1\le\Gamma_{\underline{\beta}},Q\in\tilde{\mathcal{Q}}}\sum_{i=1}^{N}\|\Phi'_{\underline{\beta}}(\underline{x}_i)\|^2_{\mathcal{H}_{\underline{\beta}}}\\
&\quad=\sup_{\substack{\|\underline{\beta}\|_1\le\Gamma_{\underline{\beta}}\\ Q\in\tilde{\mathcal{Q}}}}\sum_{i=1}^{N}\sum_{\underline{z},\underline{z}'}Q(\underline{z}\,|\underline{x}_i)Q(\underline{z}'\,|\,\underline{x}_i)\langle k_{\underline{\beta}}(\cdot,(\underline{z})),k_{\underline{\beta}}(\cdot,(\underline{z}'))\rangle_{\mathcal{H}_{\underline{\beta}}}\\
&\quad=\sup_{\substack{\|\underline{\beta}\|_1\le\Gamma_{\underline{\beta}}\\ Q\in\tilde{\mathcal{Q}}}}\sum_{i=1}^{N}\sum_{\underline{z},\underline{z}'}Q(\underline{z}|\,\underline{x}_i)Q(\underline{z}'\,|\,\underline{x}_i)\sum_{s=1}^{S}\beta^s\mathbb{1}[z^s=z'^s]\\
&\quad=\sup_{\|\underline{\beta}\|_1\le\Gamma_{\underline{\beta}},Q\in\tilde{\mathcal{Q}}}\sum_{i=1}^{N}\sum_{s=1}^{S}\beta^s\sum_{z^s}Q^2(z^s\,|\,x_i^s)\\
&\quad\le N\sup_{\|\underline{\beta}\|_1\le\Gamma_{\underline{\beta}}}\sum_{s=1}^{S}\beta^s\\
&\quad\le N\Gamma_{\underline{\beta}}
\end{aligned}\qquad (51)$$

For the second term in (50), we follow the arguments in the proof of Proposition 4 in Appendix in [9] and use the property of the weighted count kernel, and obtain

$$\begin{aligned}
&2\mathbb{E}\sup_{\|\underline{\beta}\|_1\le\Gamma_{\underline{\beta}},Q\in\tilde{\mathcal{Q}}}\sum_{1\le i<j\le N}\sigma_i\sigma_j\langle\Phi'_{\underline{\beta}}(\underline{X}_i),\Phi'_{\underline{\beta}}(\underline{X}_j)\rangle_{\mathcal{H}_{\underline{\beta}}}\\
&\quad\le 2(N-1)\sqrt{\frac{N}{2}}\sup_{\|\underline{\beta}\|_1\le\Gamma_{\underline{\beta}}}\sup_{\underline{z},\underline{z}'}k_{\underline{\beta}}(\underline{z},\underline{z}')\sqrt{2\log|\tilde{\mathcal{Q}}|}\\
&\quad=2(N-1)\sqrt{\frac{N}{2}}\sup_{\|\underline{\beta}\|_1\le\Gamma_{\underline{\beta}}}\sup_{\underline{z},\underline{z}'}\sum_{s=1}^{S}\beta_s\mathbb{1}[z^s=z'^s]\sqrt{2\log|\tilde{\mathcal{Q}}|}\\
&\quad\le 2(N-1)\sqrt{\frac{N}{2}}\Gamma_{\underline{\beta}}\sqrt{2\log|\tilde{\mathcal{Q}}|}\\
&\quad=2(N-1)\Gamma_{\underline{\beta}}\sqrt{N\log|\tilde{\mathcal{Q}}|}.
\end{aligned}\qquad (52)$$

Combining (51) and (52), we obtain

$$R_N(\tilde{\mathcal{F}})\le\frac{\Gamma_w\sqrt{\Gamma_{\underline{\beta}}}}{N}(N+2(N-1)\sqrt{N\log|\tilde{\mathcal{Q}}|})^{\frac{1}{2}}.\qquad (53)$$

For the case when $\tilde{\mathcal{F}}=\mathcal{F}_0$, (29) follows from (53) by setting $\tilde{\mathcal{Q}}=\mathcal{Q}_1$ and noticing that $|\mathcal{Q}_0|=L_z^{L_xS}$.

For the case when $\tilde{\mathcal{F}}=\mathcal{F}_1$, (30) follows from (53) by setting $\tilde{\mathcal{Q}}=\mathcal{Q}_0$ and noticing that the number of possible conditional distributions $Q(\underline{z}\,|\,\underline{x})\in\mathcal{Q}_1$ is bounded by

$$|\mathcal{Q}_1|=\left(\binom{L_z}{1}+\binom{L_z}{2}+\cdots+\binom{L_z}{L_z}\right)^{L_xS}\le 2^{L_zL_xS}.\qquad (54)$$

Appendix D
Proof of Proposition 2

We apply the following well-known result, which provides a uniform bound on the difference between empirical and expected risk functions over a function class.

Lemma 3. [40]: Let the loss function $\phi(\cdot)$ be Lipschitz continuous with constant L_ϕ, and let Γ_ϕ be a uniform bound on $\phi(\cdot)$. Further assume that $Y\in\{-1,1\}$. Then, for any small $0<\delta<1$, with probability larger than $1-\delta$,

$$\begin{aligned}
&\sup_{f\in\mathcal{F}}|\hat{\mathbb{E}}\phi(Yf(X))-\mathbb{E}\phi(Yf(X))|\\
&\qquad\qquad\le 2L_\phi R_N(\mathcal{F})+\Gamma_\phi\sqrt{\frac{2\log\frac{1}{\delta}}{N}}.\qquad (55)
\end{aligned}$$

Applying Lemma 3, we have the following bound for our problem:

$$\begin{aligned}&\mathbb{E}\phi(Yw_{\hat{f}}(\underline{X})) - \mathbb{E}\phi(Yw_{\tilde{f}}(\underline{X}))\\ &\quad\le 2\sup_{f\in\mathcal{F}}|\hat{\mathbb{E}}\phi(Yw_f(\underline{X})) - \mathbb{E}\phi(Yw_f(\underline{X}))|\\ &\quad\le 4L_\phi R_N(\mathcal{F}) + 2\Gamma_\phi\sqrt{\frac{2\log\frac{1}{\delta}}{N}}\\ &\quad\le 4R_N(\mathcal{F}) + 2\Gamma_\phi\sqrt{\frac{2\log\frac{1}{\delta}}{N}},\end{aligned} \tag{56}$$

where the last step follows because $L_\phi \le 1$ for the logistic and hinge loss functions.

Next we derive a bound for Γ_ϕ. We first show that both the logistic and hinge loss functions satisfy

$$\phi(x) \le 1 + |x|. \tag{57}$$

It is clear that (57) holds for the hinge loss function $\phi(x) = (1-x)_+$. For the logistic loss function $\phi(x) = \log(1+e^{-x})$, if $x \ge 0$, then

$$\log(e^{-x}+1) < e^{-x} \le 1 + |x|. \tag{58}$$

Now, if $x < 0$, then $e^{-x+1} > e^{-x} + 1$, because $e^{x+1} > e^x + 1$ for all $x > 0$. This implies that

$$\log(e^{-x}+1) < \log(e^{-x+1}) = 1 - x \le 1 + |x|. \tag{59}$$

Hence, (57) holds for all x for the logistic loss function.

We now bound Γ_ϕ of the two loss functions with decision rules using the weighted count kernel as follows.

$$\begin{aligned}\Gamma_\phi &= \sup_{f\in\mathcal{F},y\in\{\pm1\},\underline{x}\in\mathcal{X}^S}|\phi(yw_f(\underline{x}))|\\ &= \sup_{f\in\mathcal{F},y\in\{\pm1\},\underline{x}\in\mathcal{X}^S}|\phi(y\langle w_f, \Phi'_{\underline{\beta}}(\underline{x})\rangle_{\mathcal{H}_{\underline{\beta}}})|\\ &\le 1 + \sup_{\substack{\|\underline{\beta}\|_1\le\Gamma_{\underline{\beta}},\|w\|_{\mathcal{H}_{\underline{\beta}}}\le\Gamma_w\\ Q\in\mathcal{Q},y\in\{\pm1\},\underline{x}\in\mathcal{X}^S}}|y_i\langle w_f, \Phi'_{\underline{\beta}}(\underline{x})\rangle_{\mathcal{H}_{\underline{\beta}}}|\\ &= 1 + \sup_{\|\underline{\beta}\|_1\le\Gamma_{\underline{\beta}},\|w\|_{\mathcal{H}_{\underline{\beta}}}\le\Gamma_w,Q\in\mathcal{Q},,\underline{x}\in\mathcal{X}^S}|\langle w_f, \Phi'_{\underline{\beta}}(\underline{x})\rangle_{\mathcal{H}_{\underline{\beta}}}|\\ &\overset{(a)}{\le} 1 + \Gamma_w \sup_{\|\underline{\beta}\|_1\le\Gamma_{\underline{\beta}},Q\in\mathcal{Q},,\underline{x}\in\mathcal{X}^S}\|\Phi'_{\underline{\beta}}(\underline{x})\|_{\mathcal{H}_{\underline{\beta}}}\\ &\overset{(b)}{\le} 1 + \Gamma_w\sqrt{\Gamma_{\underline{\beta}}}\end{aligned} \tag{60}$$

where (a) follows from the Cauchy-Schwartz inequality, and (b) follows from the steps in (51). This concludes the proof.

References

[1] J. N. Tsitsiklis, "Decentralized detection," in *Adv. Signal Process.*, H. V. Poor and J. B. Thomas, Eds. New York, NY, USA: JAI, 1993, vol. 2, pp. 297–344.

[2] R. Viswanathan and P. K. Varshney, "Distributed detection with multiple sensors: Part I—Fundamentals," *Proc. IEEE*, vol. 85, no. 1, pp. 54–63, Jan. 1997.

[3] V. V. Veeravalli and P. K. Varshney, "Distributed inference in wireless sensor networks," *Philos. Trans. Roy. Soc. London A, Math., Phys. Eng. Sci.*, vol. 370, no. 1958, pp. 100–117, Jan. 2012.

[4] S. A. Kassam, "Nonparametric signal detection," in *Advances in Signal Processing*, H. V. Poor and J. B. Thomas, Eds. New York, NY, USA: JAI, 1993, vol. 2, pp. 66–91.

[5] M. M. Al-Ibrahim and P. K. Varshney, "Nonparametric sequential detection based on multisensor data," in *Proc. 23rd Annu. Conf. Inf. Sci. Syst.*, Mar. 1989, pp. 157–162.

[6] A. Nasipuri and S. Tantaratana, "Nonparametric distributed detection using wilconxin statistics," *Signal Process.*, vol. 57, no. 2, pp. 139–146, 1997.

[7] J. B. Predd, S. R. Kulkarni, and H. V. Poor, "Distributed learning in wireless sensor networks: Application issues and the problem of distributed inference," *IEEE Signal Process. Mag.*, vol. 23, no. 4, pp. 56–69, Jul. 2006.

[8] J. B. Predd, S. R. Kulkarni, and H. V. Poor, "Consistency in models for distributed learning under communication constraints," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 52–63, Jan. 2006.

[9] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Nonparametric decentralized detection using kernel methods," *IEEE Trans. Signal Process.*, vol. 53, no. 11, pp. 4053–4066, Nov. 2005.

[10] J. Hu, Y. Liang, and E. P. Xing, "Nonparametric decision making based on tree-structured information aggregation," in *Proc. Annu. Allerton Conf. Commun., Control, Comput.*, Sep. 2011, pp. 1853–1860.

[11] A. R. da Silva, M. H. T. Martins, B. P. S. Rocha, A. A. F. Loureiro, L. B. Ruiz, and H. C. Wong, "Decentralized intrusion detection in wireless sensor networks," in *Proc. 1st ACM Int. Workshop Qual. Serv. Secur. Wireless Mobile Netw.*, 2005, pp. 16–23.

[12] N. A. Goodman and D. Bruyere, "Optimum and decentralized detection for multistatic airborne radar," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 43, no. 2, pp. 806–813, 2007.

[13] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. Roy. Statist. Soc. Ser. B*, vol. 58, pp. 267–288, 1996.

[14] S. Chen, D. Donaho, and M. Saunders, "Atomic decomposition by basis persuit," *SIAM J. Scientif. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.

[15] D. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 2845–2862, 2001.

[16] E. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[17] H. Rowaihy, S. Eswaran, M. Johnson, D. Verma, A. Bar-Noy, and T. Brown *et al.*, "A survey of sensor selection schemes in wireless sensor networks," in *Proc. SPIE*, 2007, p. 6562.

[18] V. Srivastava, K. Plarre, and F. Bullo, "Randomized sensor selection in sequential hypothesis testing," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 2342–2354, 2011.

[19] S. Joshi and S. Boyd, "Sensor selection via convex optimization," *IEEE Trans. Signal Process.*, vol. 57, no. 2, pp. 451–462, 2009.

[20] W. P. Tay, J. N. Tsitsiklis, and M. Z. Win, "Asymptotic performance of a censoring sensor network," *IEEE Trans. Inf. Theory*, vol. 53, no. 11, pp. 4191–4209, 2007.

[21] W. Yang and H. Shi, "Sensor selection schemes for consensus based distributed estimation over energy constrained wireless sensor networks," *Neurocomput.*, vol. 87, pp. 132–137, 2012.

[22] L. Zuo, R. Niu, and P. K. Varshney, "A sensor selection approach for target tracking in sensor networks with quantized measurements," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, Mar. 2008, pp. 2521–2524.

[23] H. Wang, K. Yao, G. Pottie, and D. Estrin, "Entropy-based sensor selection heuristic for target localization," in *Proc. 3rd Int. Symp. Inf. Process. Sens. Netw.*, 2004, pp. 36–45.

[24] V. Gupta, T. H. Chung, B. Hassibi, and R. M. Murry, "On a stochastic sensor selection algorithm with applications in sensor scheduling and sensor coverage," *Automatica*, vol. 42, no. 2, pp. 251–260, 2006.

[25] W. Welch, "Branch-and-bound search for experimental designs based on d-optimality and other criteria," *Technomerics*, vol. 24, no. 1, pp. 41–48, 1982.

[26] V. Isler and R. Bajcsy, "The sensor selection problem for bounded uncertainty sensing models," *IEEE Trans. Autom. Sci. Eng.*, vol. 3, no. 4, pp. 372–381, 2006.

[27] B. Scholkopf and A. Smola, *Learning With Kernels*. Cambridge, MA, USA: MIT Press, 2002.

[28] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, MA, USA: Athena Scientific, 1999.

[29] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[30] H. Attouch, J. Bolte, and B. F. Svaiter, "Convergence of desenct methods for semi-algebraic and tame problems: Proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods," *Math. Program.*, vol. 137, pp. 91–129, 2013.

[31] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Math. Program.*, vol. 146, pp. 459–494, 2014.

[32] H. Attouch, M. M. Alves, and B. F. Svaiter, "A dynamic approach to a proximal-newton method for monotone inclusions in Hilbert spaces, with complexity $o(1/n^2)$," 2015 [Online]. Available: arXiv:1502.04286v2

[33] S. G. Krantz and H. R. Parks, *A Primer of Real Analytic Functions*. New York, NY, USA: Springer-Verlag, 2002.

[34] Y. Xu and W. Yin, "A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion," *SIAM J. Imag. Sci.*, vol. 6, no. 3, pp. 1758–1789, 2013.

[35] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *J. Mach. Learn. Res.*, vol. 3, pp. 463–482, Nov. 2002.

[36] K. Kurdyka, "On gradients of functions definable in o-minimal structures," *Ann. l'inst. Fourier*, vol. 48, no. 3, pp. 769–783, 1998.

[37] A. Parusiński, "Subanalytic functions," *Trans. Amer. Math. Soc.*, vol. 344, no. 2, pp. 583–595, 1994.

[38] J. Bolte, A. Daniilidis, and A. Lewis, "The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems," *SIAM J. Optim.*, vol. 17, no. 4, pp. 1205–1223, 2007.

[39] E. Bierstone and P. D. Milman, "Semianalytic and subanalytic sets," *Publ. Math. l'IHS*, vol. 67, no. 1, pp. 5–42, 1988.

[40] S. Boucheron, O. Bousquet, and G. Lugosi, "Theory of classification: A survey of some recent advances," *ESAIM: Probabil. Statist.*, vol. 9, pp. 323–375, Nov. 2005.

Weiguang Wang received the B.S. degree from the University of Science and Technology of China, Hefei, in 2011.

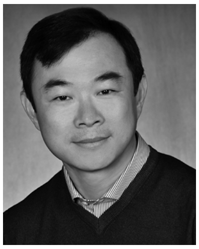
Since August 2011, he has been a Ph.D. student at Syracuse University, Syracuse, NY, USA. His research interests focus on machine learning and signal processing.

Mr. Wang received the Excellent Student Scholarship from University of Science and Technology of China during 2007–2010.

Yingbin Liang (S'01–M'05) received the Ph.D. degree in electrical engineering from the University of Illinois at Urbana-Champaign, USA, in 2005.

During 2005–2007, she was a Postdoctoral Research Associate at Princeton University, Princeton, NJ, USA. During 2008–2009, she was an Assistant Professor with the Department of Electrical Engineering, the University of Hawaii, Honolulu, USA. Since December 2009, she has been on the faculty of Syracuse University, Syracuse, NY, USA, where she is an Associate Professor. His research interests include information theory, wireless communications and networks, and machine learning.

Dr. Liang was a Vodafone Fellow at the University of Illinois at Urbana-Champaign during 2003–2005, and received the Vodafone-U.S. Foundation Fellows Initiative Research Merit Award in 2005. She also received the M. E. Van Valkenburg Graduate Research Award from the ECE Department, University of Illinois at Urbana-Champaign, in 2005. In 2009, she received the National Science Foundation CAREER Award, and the State of Hawaii Governor Innovation Award. More recently, her paper received the 2014 EURASIP Best Paper Award for the EURASIP *Journal on Wireless Communications and Networking*. She is currently serving as an Associate Editor for the Shannon Theory of the IEEE TRANSACTIONS ON INFORMATION THEORY.



Eric P. Xing (SM'11) received the Ph.D. degree in molecular biology from Rutgers University, New Brunswick, NJ, USA, and another Ph.D. degree in computer science from the University of California, Berkeley, USA.

He is a professor with the School of Computer Science at Carnegie–Mellon University, Pittsburgh, PA, USA. His principal research interests lie in the development of machine learning and statistical methodology, and large-scale computational system and architecture, for solving problems involving automated learning, reasoning, and decision-making in high-dimensional, multimodal, and dynamic possible worlds in complex systems. His current work involves: 1) foundations of statistical learning, including theory and algorithms for estimating time/space varying-coefficient models, sparse structured input/output models, and nonparametric Bayesian models; 2) framework for parallel machine learning on big data with big model in distributed systems or in the cloud; 3) computational and statistical analysis of gene regulation, genetic variation, and disease associations; and 4) application of statistical learning in social networks, data mining, and vision. Professor Xing has published over 200 peer-reviewed papers,

Dr. Xing is an Associate Editor of the *Journal of the American Statistical Association*, *Annals of Applied Statistics*, the IEEE TRANSACTIONS OF PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the *PLoS Journal of Computational Biology*, and an Action Editor of the *Machine Learning* journal, and the *Journal of Machine Learning Research*. He is a member of the DARPA Information Science and Technology (ISAT) Advisory Group, a recipient of the NSF Career Award, the Alfred P. Sloan Research Fellowship, the United States Air Force Young Investigator Award, and the IBM Open Collaborative Research Faculty Award.

Lixin Shen received the B.S. and M.S. degrees from Peking University, Beijing, China, in 1987 and 1990, respectively, and the Ph.D. degree from Zhongshan University, Guangzhou, China, in 1996, all in mathematics.

He is currently a Full Professor with the Department of Mathematics, Syracuse University, Syracuse, NY, USA. His research has been supported by the National Science Foundation and the Air Force Research Laboratory, Rome, NY. His current research interests include multiscale analysis and optimization, and their applications in image processing.