July 2016

Data Driven Nonparametric Detection

Weiguang Wang
Syracuse University

ABSTRACT

The major goal of signal detection is to distinguish between hypotheses about the state of events based on observations. Typically, signal detection can be categorized into centralized detection, where all observed data are available for making decision, and decentralized detection, where only quantized data from distributed sensors are forwarded to a fusion center for decision making. While these problems have been intensively studied under parametric and semi-parametric models with underlying distributions being fully or partially known, nonparametric scenarios are not well understood yet. This thesis mainly explores nonparametric models with unknown underlying distributions as well as semi-parametric models as an intermediate step to solve nonparametric problems.

One major topic of this thesis is on nonparametric decentralized detection, in which the joint distribution of the state of an event and sensor observations are not known, but only some training data are available. The kernel-based nonparametric approach has been proposed by Nguyen, Wainwright and Jordan where sensors' quality is treated equally. We study heterogeneous sensor networks, and propose a weighted kernel so that weight parameters are utilized 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. Sensor selection is jointly performed with decision rules of sensors and the fusion center with the resulting optimal decision rule having only a sparse number of nonzero weight parameters. A gradient projection algorithm and a Gauss-Seidel algorithm are developed to solve the risk minimization problem, which is non-convex, and both algorithms are shown to converge to critical points.

The other major topic of this thesis is composite outlier detection in centralized scenarios. The goal is to detect the existence of data streams drawn from outlying distributions

among data streams drawn from a typical distribution. We study both the semi-parametric model with known typical distribution and unknown outlying distributions, and the non-parametric model with unknown typical and outlying distributions. For both models, we construct generalized likelihood ratio tests (GLRT), and show that with the knowledge of the KL divergence between the outlier and typical distributions, GLRT is exponentially consistent (i.e, the error risk function decays exponentially fast). We also show that with the knowledge of the Chernoff distance between the outlying and typical distributions, GLRT for semi-parametric model achieves the same risk decay exponent as the parametric model, and GLRT for nonparametric model achieves the same performance when the number of data streams gets asymptotically large. We further show that for both models without any knowledge about the distance between distributions, there does not exist an exponentially consistent test. However, GLRT with a diminishing threshold can still be consistent.

DATA DRIVEN NONPARAMETRIC DETECTION

by

Weiguang Wang
B.E., University of Science and Technology of China, 2011

DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Electrical Engineering

Syracuse University
July 2016

ACKNOWLEDGMENTS

My deepest gratitude goes first and foremost to my advisor Prof. Yingbin Liang for her support and supervision during my PhD study. It is my great pleasure to be her student. Her enthusiasm and passion for research has been always inspiring me. Her knowledge and wisdom helped me improve myself step by step. Every discussion with her was a training of logical analyzing, which made my vague ideas into rigorous derivations. I have learned a lot from her insightful perspective and critical thinking on research topics.

I would also thank my collaborators, Prof. Lixin Shen, Prof. Eric P. Xing for their continuously instructive suggestions and comments on our research. They always have a broader view of problems which has provided me more comprehensive understanding of the problems.

I thank my doctoral defense committee: Prof. Biao Chen, Prof. Peng Gao, Prof. M. Cenk Gursoy, Prof. Qinru Qiu, Prof. Pramod K. Varshney for carefully reading this thesis and giving me helpful suggestions.

I am also grateful for my lab mates: Ruchen Duan, Jiayao Hu, Shaofeng Zou, Anhong He, Huishuai Zhang, Yunhao Sun, Yi Zhou, Zhe Wang for their help in study and life. Their enthusiasm in research and professional attitude to work are great inspiration and encourage during my PhD study.

I would like to thank my parents and my grandparents for their love and unconditional support. They are the people who give me strength to overcome obstacles in my life.

I acknowledge Syracuse University and the National Science Foundation for providing funding for my Ph.D. study.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

The major goal of signal detection is to distinguish between hypotheses about the state of events based on observations. A variety of detection problems have been intensively studied and widely used in different areas such as radar systems and automatic control. Based on different assumptions on data availability, signal detection can be classified as centralized detection and decentralized detection. In centralized detection, all observed data are available for decision making. In decentralized detection, only compressed data from distributively located sensors are communicated to a fusion center, where a decision is made.

In both centralized and decentralized detection problems, one critical factor that affects decision accuracy is the knowledge of the joint distribution of the states and observations. Most of the previous studies [1–3] 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. Also, implicitly, the two processes of learning the distribution and designing 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* method. In this thesis, we focus on the design of such nonparametric approaches for both centralized and decentralized detection problems, where distributions are only partially known or unknown.

1.1 Nonparametric Decentralized Detection

In the decentralized detection problem (see, e.g., [4–6]), 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 jointly find optimal decentralized quantization rules for sensors and a decision rule for the fusion center to achieve the best system performance.

Nonparametric (de)centralized detection was studied previously, e.g., [1–3], 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]. More recently, a kernel-based classification approach was proposed in [8] for solving the nonparametric decentralized detection problem, which is more generally applicable with mathematical guarantees 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 sen-

sors. It has been shown by numerical examples in [8] that the kernel-based approach yields better performance 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.

Thus, one major component of this thesis studies more realistic sensor networks, which generalize the system models studied in [8, 9] 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 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 [4], intrusion detection in wireless sensor networks [10], distributed equipment failure detection [4], multi-static airborne radar [11]. Thus, our goal in this work is to design nonparametric decision rules which take heterogeneous features of networks into consideration for achieving as good performance as possible.

We summarize our main contributions as follows.

- We incorporate a novel weighted kernel into the risk minimization framework proposed in [8] 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 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.

- 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.

1.2 Sensor Selection in Decentralized Detection

In nonparametric decentralized detection, 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.

Sensor selection problem has been intensively studied previously in both parameter estimation [12–14] and detection [15,16] to balance between the consumption of communication resources and system performance. 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. [15–19]), 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 [20] and [21] considered scenarios that only one sensor is selected at a time, hence complexity of exhaustive search is reduced. The work [22] provided more efficient algorithms than exhaustive search based on bounds on objective functions. The work [23] 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, approaches based on convex relaxation for sensor selection were proposed in a few works. In [17], sensor selection was formulated as a Boolean-convex problem, where relaxation was then taken to allow discrete Boolean variables to take continuous values. This problem was further generalized to nonlinear measurement models in [24]. [12] studied sensor selection in stochastically forced networks by relaxing the non-convex constraint in the mean-square deviation minimization problem. [13] applied the method introduced in [17] to non-myopic sensor selection problem for target tracking. [14] further generalized the study of sensor selection for parameter estimation to the case where the measurement noise is correlated. All these works investigated sensor selection for parameter estimation, while our work studies sensor selection for detection problem in decentralized systems.

In our design of nonparametric approaches to solve decentralized detection problems as described in the previous section, we wish to address the sensor selection problem jointly with learning the decision rules. More specifically, we summarize our main contributions as follows.

- Using the weighted kernel as described in the previous section, 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: (1) The sensor selection problem can now be solved via recent celebrated techniques of Lasso and compressed sensing [25–28], which significantly reduces computational complexity; (2) The regularization parameter of l_1 can flexibly control sparsity of sensor selection and its trade-off with the performance of decision making; and (3) This sensor selection approach preferably selects sensors with independent observations, and removes highly correlated (and hence redundant) observations, thus achieving dimension reduction as well.

- We provide numerical results to demonstrate effectiveness of our sensor selection algorithms and the corresponding properties.

1.3 Composite Outlier Detection

The outlier hypothesis testing problem has recently attracted intensive attention. The outlier hypothesis refers to a certain scenario associated with data exhibiting unusual characteristics as opposed to the typical hypothesis associated with data capturing normal behavior. Often data under the two hypotheses are assumed to be generated by certain outlier and typical statistical distributions. The goal of outlier hypothesis testing is to detect the outlying data streams generated under the outlier hypothesis (i.e., distribution). Solutions to such type of problems can be applied to many application domains such as homogeneity testing and classification [29–32] and decoding over discrete memoryless channels [33]. We

note that the outlier hypothesis testing problem is distinct from the outlier detection in data mining [34, 35], which does not assume any underlying statistical distributions to model the data and consequently does not come with performance guarantee. On the other hand, outlier hypothesis testing can be analyzed with performance guarantees due to underlying statistical models associated with the hypotheses.

Most previous studies on outlier hypothesis testing have focused on identifying outliers from a number of data streams, where each data stream is drawn either from an outlying or typical distribution. The problem is essentially a multiple hypothesis testing problem with each hypothesis corresponding to a certain subset of data streams being outliers. In [36], the parametric problem is studied, where the distributions are assumed to known. In [37], the nonparametric problem is studied, where the distributions are assumed to be discrete and unknown a priori. The generalized likelihood ratio test is designed which applies the empirical distributions of the data to replace the true distributions. [38] also studied a nonparametric problem, but extended the distributions to be arbitrary (including continuous distributions). In particular, kernel-based tests based on the distance metric of maximum mean discrepancy were proposed.

The above studies of identification of outliers implicitly assumed that outliers exist in observations. In practice, it is typically of importance to initially determine whether or not outliers even exist before further efforts to specify outliers. Thus, one major topic of this thesis is to address the problem to distinguish between the null hypothesis with no existence of outliers and the alternative hypothesis with one or more outliers exist in a number of data streams. More specifically, suppose a large number, say M, of observation data sequences are given with each sequence generated either by a typical or outlying distribution (denoted by π and μ, respectively). The goal is to determine whether or not there exist outlier(s). Such a type of problems can be viewed as a binary composite hypothesis testing problem, because the alternative hypothesis consists of multiple possibilities corresponding to different subset of sequences being outliers. In particular, the focus of this thesis is to

study the semi-parametric model where the typical distribution is known but the outlying distribution is unknown a priori, and the nonparametric model where neither typical nor outlying distribution is known a priori.

The general problem of composite hypothesis testing has been well studied previously. These studies mainly focused on parametric models, which assume that data are generated by known distributions such as Gaussian or Bernoulli distributions. For example, [39] studied the problem of testing a composite hypothesis against a simple alternative where all hypotheses are associated with Gaussian distributions. [40] studied the same problem with independent and arbitrarily distributed observations. [41] studied a general parametric binary composite hypothesis testing problem under arbitrary discrete distributions. All these studies adopted Neyman-Pearson formulation, i.e., minimizing the type II error subject to a given constraint on the type I error. [39] studied the existence of uniformly most powerful test under Neyman-Pearson setting for a variety of specific composite detection problems. [40] applied likelihood ratio threshold test (LRTT) and derived exact asymptotics of error probability for a more general composite detection model. [41] applied GLRT and demonstrated the optimal error exponent under Neyman-Pearson setting.

In this thesis, we adopt a type of minimax performance metric, i.e., minimizing the risk of the addition of the type I and type II errors, where the type II error is maximized over all sub-hypotheses. Under such a performance metric, it is not even clear what is an optimal test for the parametric model in the general binary composite problem. However, our outlier hypothesis testing problem has special structures to be exploited. For example, the alternative hypothesis contains symmetric subhypotheses if only one outlier possibly exists. Moreover, we are interested in the exponent of the risk function as the number of samples in each data sequence gets asymptotically large.

We summarize our main contributions as follows.

- As necessary understanding towards semi-parametric and nonparametric models, we first show that for the parametric composite outlier hypothesis testing problem, the

GLRT is exponentially consistent and achieves the optimal exponent of the risk function given by the Chernoff distance between π and μ.

- For the semi-parametric model with known π and unknown μ, we construct GLRT based on π and the empirical distribution of data for μ. We show that such a test is exponentially consistent given KL divergence between π and μ, and can achieve the optimal exponent for the parametric model given Chernoff distance between π and μ. The test is thus optimal in terms of the error decay rate. We also show that without any knowledge of μ, a universally consistent test can still be constructed if the threshold in GLRT properly scales with the sample size, but exponential consistency is not possible.

- For the nonparametric model, we construct GLRT based on empirical distributions for both π and μ. We show that such a test is exponentially consistent given KL divergence between μ and the uniform mixture of true distributions generating all sequences. Moreover, such a test achieves the optimal exponent for the parametric model as the number of sequences goes to infinity given Chernoff distance between π and μ. Similarly to the semi-parametric case, without any knowledge of π and μ, a universally consistent test can still be constructed if the threshold in GLRT properly scales with the sample size, but exponential consistency is not possible.

1.4 Publications and Thesis Organization

As a summary, my PhD work so far has lead to two journal publications [42, 43] and one journal in preparation [44], three conference publications [45–47]. The list of publications is provided as follows.

Journal Publications

J1 W. Wang, Y. Liang, and E. P. Xing, "Collective support recovery for multi-design multi-response linear regression," *IEEE Transactions on Information Theory*, vol. 61, no. 1,

pp. 513–534, 2015.

J2 W. Wang, Y. Liang, E. P. Xing, and L. Shen, "Nonparametric decentralized detection and sparse sensor selection via weighted kernel," *IEEE Transactions on Signal Processing*, vol. 64, no. 2, pp. 306–321, 2016.

J3 W. Wang, Y. Liang, and H. V. Poor, "Nonparametric composite outlier detection," *in preparation*.

Conference Publications

C1 W. Wang, Y. Liang, and E. Xing, "Block regularized lasso for multivariate multi-response linear regression," in *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2013, pp. 608–617.

C2 W. Wang, Y. Liang, E. P. Xing, and L. Shen, "Sparse sensor selection for nonparametric decentralized detection via l_1 regularization," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2014, pp. 1–6.

C3 W. Wang, Y. Liang, and H. V. Poor, "Nonparametric composite outlier detection," to appear in *Proc. Asilomar Conf. Signals, Systems and Computers*, 2016.

The rest of the thesis is organized as follows. In Chapter 2, we present our result for the nonparametric decentralized detection problem. In Chapter 3, we present the convergence analysis of the algorithms for solving regularized empirical risk minimization problem, which arises in nonparametric decentralized detection. In Chapters 4 and 5, we present our results respectively for semi-parametric and nonparametric composite outlier detection. Finally, in Chapter 6, we summarize the contributions of the thesis and describe some future directions.

1.5 Multi-task Linear Regression

As I initially started my PhD study, I worked on an interesting problem of multi-task learning of linear regression models. This part of work is not under the main theme of this thesis, and is hence not included in detail in the thesis. In this section, we briefly introduce our results here as well as the background and the state-of-the art.

Linear regression is a simple but practically very useful statistical model, in which a response vector $\underline{Y}$ can be modeled as

$$\underline{Y} = X\underline{\beta} + \underline{W}$$

where $X \in \mathbb{R}^{n\times p}$ is the design matrix containing n samples of feature vectors, $\underline{\beta} = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ contains regression coefficients, and $\underline{W} \in \mathbb{R}^n$ is the noise vector. The goal is to find the regression coefficients $\underline{\beta}$ such that the linear relationship is as accurate as possible with regard to a certain performance criterion. The problem is more interesting in high dimensional regime with a sparse regression vector, in which the sample size n can be much smaller than the dimension p of the regression vector.

In order to estimate the sparse regression vector, the optimization problem with an l_1-constraint on $\underline{\beta}$ (referred to as Lasso) has been studied based on the idea in some seminal works ([25–27]). The l_1-regularized estimator has been proved in [48] to have similar behavior to Dantzig Selector, which was proposed in [49]. Various efficient algorithms have been developed to solve the above convex problem efficiently (see a review monograph [50]), although the objective function is not differentiable everywhere due to l_1-regularization. Moreover, the l_1-regularization is critical to force the minimizer to have sparse components as shown in [25–27]. A vast amount of recent work has studied the high dimensional linear regression problem via l_1-regularized Lasso under various assumptions, e.g., [26, 28, 51–62].

Generalized from the l_1-regularized linear regression problem which aims at selecting variables individually, group Lasso is applied to regression vector $\underline{\beta}$ in the linear regression model to select grouped variables (e.g., [63–66]. This line of research is further generalized to block-regularization for high-dimensional multi-response (i.e., multi-task) linear regression problem, (see, e.g., [67, 68] and references therein). For a multi-task regression problem, we have the following model:

$$Y = XB^* + W \tag{1.1}$$

where $Y \in \mathbb{R}^{n\times K}$ of which each column corresponds to the output of one task, $X \in \mathbb{R}^{n\times p}$ is the design matrix, the regression matrix $B^* \in \mathbb{R}^{p\times K}$ has each column corresponding to the regression vector for one task, and $W \in \mathbb{R}^{n\times K}$ has each column corresponding to the noise vector of one task. For each column $\underline{Y}^{(k)}$ of the matrix Y, it is clear that $\underline{Y}^{(k)} = X\underline{\beta}^{*(k)} + \underline{W}^{(k)}$, where $\underline{\beta}^{*(k)}$ and $\underline{W}^{(k)}$ are the corresponding columns in B^* and W. Then each column is a single-task linear regression problem and can be solved individually. However, the K individual problems (i.e., tasks) can also be coupled together via a block regularized Lasso and solved jointly in one problem. Various types of block regularization have been proposed and studied including l_1/l_2-regularization in [68], l_p/l_q-regularization in [69], l_1/l_q-regularization in [70], l_1/l_∞-regularization in [71, 72], l_1/l_2-regularization in [73], and l_1/l_q-regularized Lasso in [74].

In the multi-response linear regression problem given in (1.1), the design matrix is identical for all tasks, i.e., X is the same for all column vectors of Y and B^*. However, in many applications, it is often the case that different output variables may depend on design variables that are different or distributed differently. Thus, the resulting model includes K linear regression models with different design matrices and is given by:

$$\underline{Y}^{(k)} = X^{(k)}\underline{\beta}^{*(k)} + \underline{W}^{(k)} \tag{1.2}$$

for $k = 1, \ldots, K$, where $\underline{Y}^{(k)} \in \mathbb{R}^n$, $X^{(k)} \in \mathbb{R}^{n \times p}$, $\underline{\beta}^{*(k)} \in \mathbb{R}^p$, and $\underline{W}^{(k)} \in \mathbb{R}^n$. We refer to the above problem as the *multi-design multi-response (MDMR) linear regression model*, and the goal is to recover $\underline{\beta}^{*(k)}$ for $k = 1, \ldots, K$ jointly. For fixed matrices $X^{(1)}, \ldots, X^{(K)}$, the problem has been studied in [75, 76] via the l_1/l_2-regularized Lasso and via a variant of orthogonal matching pursuit in [76]. For random design matrices, this model has been studied via l_1/l_∞-regularized Lasso in [77] and via $l_1/l_1 + l_1/l_\infty$-regularized Lasso in [78] for incorporating both row sparsity and individual sparsity.

In our work, we study the MDMR problem for random design matrices via l_1/l_2-regularized Lasso. In our model, it is assumed that the design matrices are Gaussian distributed and are independent across tasks. Furthermore, the distributions of design matrices are also different across tasks. For each task k, the row vector of $X^{(k)}$ is Gaussian with mean zero and the covariance matrix $\Sigma^{(k)}$ for $k = 1, \ldots, K$. The noise vectors and hence the output vectors are also Gaussian distributed and independent across tasks. We are interested in joint recovery of the union of the support sets (i.e., the support union) of regression vectors $\underline{\beta}^{*(1)}, \ldots, \underline{\beta}^{*(K)}$. We collect these vectors together as a matrix $B^* = \left[\underline{\beta}^{*(1)}, \ldots, \underline{\beta}^{*(K)}\right]$.

We adopt the l_1/l_2-regularized Lasso problem for recovery of the support union. In this way, the K linear regression problems are coupled together via the regularization constraint. We show that this approach is advantageous as opposed to individual recovery of the support set for each linear regression problem. This is because the K regression models may share their samples in joint support recovery so that the total number of samples needed can be significantly reduced compared to performing each task individually.

In the following, we summarize the main contributions of this work.

- Our results contain two parts: the achievability and the converse, corresponding respectively to sufficient and necessary conditions under which the l_1/l_2-regularized Lasso problem recovers the support union for the MDMR linear regression problem. Our proof adapts the techniques developed in [58] and in [68], but involves nontrivial

development to deal with the differently distributed design matrices across tasks.

- We show that under certain conditions that the distributions of the design matrices satisfy, if $n > c_{p1}\psi(B^*, \Sigma^{(1:K)})\log(p-s)$, where c_{p1} is a constant, then the l_1/l_2-regularized Lasso recovers the support union for the MDMR linear regression problem; and if $n < c_{p2}\psi(B^*, \Sigma^{(1:K)})\log(p-s)$, where c_{p2} is a constant, then the l_1/l_2-regularized Lasso fails to recover the support union. $\psi(B^*, \Sigma^{(1:K)})$ captures the sparsity of B^* and the statistical properties of the design matrices, which are important in determining the sufficient and necessary conditions for successful recovery of the support union. Thus, $\psi(B^*, \Sigma^{(1:K)})\log(p-s)$ serves as a sharp threshold on the sample size.

- The property of $\psi(B^*, \Sigma^{(1:K)})$ also captures the advantages of the multi-task Lasso over solving each problem individually via the single-task Lasso. We show that when the K tasks share the same support sets (although the design matrices can be differently distributed), $\psi(B^*, \Sigma^{(1:K)}) = \frac{1}{K}\max_{1\leq k\leq K}\psi(\underline{\beta}_k^*, \Sigma^{(k)})$. This means that the number of samples needed per task for multi-task Lasso to jointly recover the support union is reduced by K compared to that of single-task Lasso to recover each support set individually. On the other hand, if the K tasks have disjoint support sets, then $\psi(B^*, \Sigma^{(1:K)}) = \max_{1\leq k\leq K}\psi(\underline{\beta}^{*(k)}, \Sigma^{(k)})$. This implies that the multi-task Lasso does not provide gain in the sample size needed per task for support recovery compared to single-task Lasso. Between these two extreme cases, tasks can have overlapped support sets with different overlapping levels, and the impact of these properties on the sample size for recovery of the support union is quantitatively captured by $\psi(B^*, \Sigma^{(1:K)})$.

As we mentioned before, the MDMR model has also been studied in [77] and [78], in which l_1/l_∞ and $l_1/l_1 + l_1/l_\infty$-regularization were adopted for support union recovery, respectively. In these studies, sharp threshold on sample complexity is characterized only

for K = 2 and under special conditions on $\frac{1}{n} X_{S_k}^{(k)T} X_{S_k}^{(k)}$. In our work, using l_1/l_2-regularized Lasso, we are able to characterize the sharp threshold under standard regularization conditions.

CHAPTER 2
KERNEL-BASED DECENTRALIZED DETECTION

In this chapter, we propose a kernel-based approach for nonparametric decentralized detection over a heterogeneous sensor network. We first describe the model of decentralized detection, and introduce the concept of weighted kernel. We then incorporate weighted kernel into the empirical risk minimization framework for decentralized detection, and analyze the performance of such an approach. In particular, we characterize how close the empirical approximate risk function is to the true risk function. We finally provide numerical results to compare our approach with other competitive nonparametric approaches, and demonstrate the performance of sensor selection via weighted kernel. We note that the design of algorithms for solving the risk minimization problem and analysis of the performance of algorithms are presented in the next chapter.

2.1 Model Description

We study the nonparametric decentralized detection over a sensor network. The system model is depicted in Fig. 2.1. In such a system, let Y denote the state of an environmen-

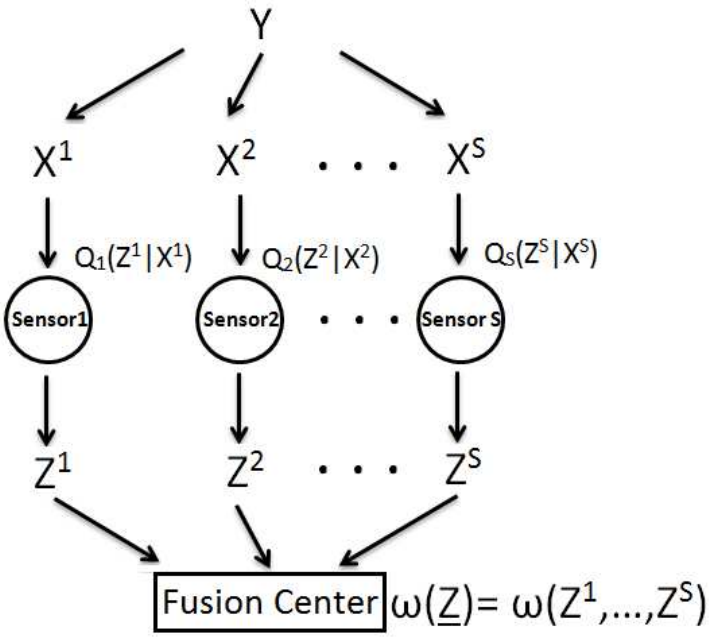

Fig. 2.1: Illustration of decentralized detection

tal 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., $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_S$ and $\mathcal{Z} = \mathcal{Z}_1 \times \mathcal{Z}_2 \times \ldots \times \mathcal{Z}_S$. 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 $Q_s(z^s|x^s)$, which implies that sensor s quantizes x^s into z^s with the probability $Q_s(z^s|x^s)$. 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 $w(\underline{Z})$.

We note that this work implicitly assumes that sensing environment is static. In practice, as quality of sensors changes over time, the training techniques developed in this work 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.

2.2 Preliminaries on Kernel

In this section, we briefly introduce the basic concepts, definitions and results on learning with kernels. This is the major technique that this work applies. A reader can refer to [79] for more details. We let $\mathcal{X}$ be a nonempty set, and define a kernel function as follows.

Definition 2.1. *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 $x_1, \cdots, x_m \in \mathcal{X}$, 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 $x_1, \cdots, x_m \in \mathcal{X}$ 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 2.2. *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:*

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

- *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 [79], which is useful for characterizing the optimal solution in empirical risk minimization.

Theorem 2.1. *[79] 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*

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

admits a representation of the form

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

2.3 Weighted Kernel

In this work, 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 [8], 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

$$k_{\underline{\beta}}(\underline{z}, \underline{z}') = \sum_{s=1}^{S} \beta^s \mathbb{I}[z^s = z'^s], \tag{2.1}$$

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 (2.1) 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 2.1. *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.*

2.4 Problem Formulation with Sensor Selection

In this work, 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 [8] 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 $Q_s(z^s|x^s)$ for all sensors ($s=1,\ldots,S$) that minimize the following l_1 regularized empirical risk function:

$$\min_{\substack{\beta^s\geq 0 \text{ for } s=1,\ldots,S \\ w\in\mathcal{H}_{\underline{\beta}},Q\in\mathcal{Q}}} \sum_{i=1}^{N}\sum_{\underline{z}} \phi(y_i\langle w(\cdot),\Phi_{\underline{\beta}}(\underline{z})\rangle_{\mathcal{H}_{\underline{\beta}}})Q(\underline{z}|\underline{x}_i)+\frac{\lambda_1}{2}\|w\|^2_{\mathcal{H}_{\underline{\beta}}}+\lambda_2\|\underline{\beta}\|_{l_1} \tag{2.2}$$

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.

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 [8], we consider the following

lower bound as a relaxation of (2.2) due to Jensen's inequality

$$\min_{\substack{\beta^s \geq 0 \text{ for } s=1,\ldots,S \\ w \in \mathcal{H}_{\underline{\beta}}, Q \in \mathcal{Q}}} \sum_{i=1}^{N} \phi(y_i \langle w(\cdot), \Phi'_{\underline{\beta}}(\underline{x}_i) \rangle_{\mathcal{H}_{\underline{\beta}}}) + \frac{\lambda_1}{2} \|w\|^2_{\mathcal{H}_{\underline{\beta}}} + \lambda_2 \|\underline{\beta}\|_{l_1} \tag{2.3}$$

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 2.5, 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.

2.5 Performance Analysis

In this section, we study how close the empirical approximate risk function given in (3.3) 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., $\|\underline{\beta}\|_{l_1} \leq \Gamma_{\underline{\beta}}$), $w \in \mathcal{H}_{\underline{\beta}}$ with bounded norm (i.e., $\|w\|_{\mathcal{H}_{\underline{\beta}}} \leq \Gamma_w$), 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 (2.3). 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 work, 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 [80] 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:

$$R_N(\mathcal{F}) := \mathbb{E}_{X,\sigma} \sup_{f \in \mathcal{F}} \left| \frac{1}{N} \sum_{i=1}^{N} \sigma_i f(X_i) \right| \tag{2.4}$$

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 2.1. *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*

$$R_N(\tilde{\mathcal{F}}) \leq \frac{\Gamma_w \sqrt{\Gamma_{\underline{\beta}}}}{N} \left(N + 2(N-1)\sqrt{N \log |\tilde{\mathcal{Q}}|} \right)^{\frac{1}{2}}, \tag{2.5}$$

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

$$R_N(\mathcal{F}_0) \leq \frac{\Gamma_w \sqrt{\Gamma_{\underline{\beta}}}}{N} \left(N + 2(N-1)\sqrt{NSL_x \log L_z} \right)^{\frac{1}{2}}. \tag{2.6}$$

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

$$R_N(\mathcal{F}_1) \leq \frac{\Gamma_w \sqrt{\Gamma_{\underline{\beta}}}}{N} \left(N + 2(N-1)\sqrt{NSL_z L_x \log 2} \right)^{\frac{1}{2}}. \tag{2.7}$$

Remark 2.2. *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

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

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 (2.3)). We further let $\hat{f}$ denote its corresponding minimizer, i.e.,

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \hat{\mathbb{E}}\phi(Y w_f(\underline{X})). \tag{2.8}$$

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

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

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

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

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

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

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 2.2. *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 $1 - \delta$,*

$$\begin{aligned} &\mathbb{E}\phi(Y w_{\hat{f}}(\underline{X})) - \mathbb{E}\phi(Y w_{\tilde{f}}(\underline{X})) \\ &\quad \le 4R_N(\mathcal{F}) + 2(1 + \Gamma_w\sqrt{\Gamma_{\underline{\beta}}})\sqrt{\frac{2\log\frac{1}{\delta}}{N}}. \end{aligned} \tag{2.9}$$

Remark 2.3. *Following from Proposition 2.1, 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 2.2 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

$$\hat{f}_0 = \operatorname*{argmin}_{f \in \mathcal{F}_0} \hat{\mathbb{E}}\phi(Y w_f(\underline{X})),$$

which is the decision rule that optimizes the empirical approximate risk over the set $\mathcal{F}_0$. It can be shown (as in [8]) that with a probability at least $1 - 2\delta$, the true risk is bounded by the empirical approximate risk as follows:

$$\begin{aligned} &\hat{\mathbb{E}}\phi(Y w_{\hat{f}}(\underline{X})) - 2L_\phi R_N(\mathcal{F}) - \Gamma_\phi \sqrt{\frac{2\log\frac{1}{\delta}}{N}} \\ &\qquad \le \mathbb{E}\phi(Y w_{f^*}(\underline{Z})) \le \hat{\mathbb{E}}\phi(Y w_{\hat{f}_0}(\underline{X})) + 2L_\phi R_N(\mathcal{F}_0) + \Gamma_\phi \sqrt{\frac{2\log\frac{1}{\delta}}{N}}, \end{aligned} \tag{2.10}$$

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 2.4. *Following Proposition 2.1 and Remark 2.2, 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.

Proposition 2.3. *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:*

$$\begin{aligned} P(Yw_{f^*}(\underline{Z})<0) &\leq \frac{1}{\log 2}\mathbb{E}\phi(Yw_{f^*}(\underline{Z})) \\ &\leq \frac{1}{\log 2}\left[\hat{\mathbb{E}}\phi_l(Yw_{\hat{f}_0}(\underline{X}))+2R_N(\mathcal{F}_0)+\Gamma_\phi\sqrt{\frac{2\log\frac{1}{\delta}}{N}}\right], \end{aligned} \tag{2.11}$$

and

$$\begin{aligned} P(Yw_{f^*}(\underline{Z})<0) &\leq \mathbb{E}\phi(Yw_{f^*}(\underline{Z})) \\ &\leq \hat{\mathbb{E}}\phi_h(Yw_{\hat{f}_0}(\underline{X}))+2R_N(\mathcal{F}_0)+\Gamma_\phi\sqrt{\frac{2\log\frac{1}{\delta}}{N}} \end{aligned} \tag{2.12}$$

where $R_N(\mathcal{F}_0)$ *is bounded in* (2.6) *and* $\Gamma_\phi = 1+\Gamma_w\sqrt{\Gamma_{\underline{\beta}}}$.

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

$$P(Yw_{f^*}(\underline{Z})<0) = \mathbb{E}\mathbb{I}[Yw_{f^*}(\underline{Z})<0] \leq \mathbb{E}\phi(Yw_{f^*}(\underline{Z})). \tag{2.13}$$

Then applying (2.10), we obtain the desired bound. If the logistic loss is used, then we obtain the bound by following the above steps except noticing that

$$\mathbb{E}\mathbb{I}[Yw_{f^*}(\underline{Z})<0] \leq \frac{1}{\log 2}\mathbb{E}\phi(Yw_{f^*}(\underline{Z})). \tag{2.14}$$

□

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.

2.6 Proof of Upper Bound on Rademacher Complexity

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.

$$
\begin{aligned}
&R_N(\tilde{\mathcal{F}}) \\
&= \mathbb{E}\sup_{f\in\tilde{\mathcal{F}}}\left|\frac{1}{N}\sum_{i=1}^{N}\sigma_i w_f(\underline{X}_i)\right| \\
&= \mathbb{E}\sup_{\|\underline{\beta}\|_1\le\Gamma_{\underline{\beta}},\|w\|_{\mathcal{H}_{\underline{\beta}}}\le\Gamma_w,Q\in\tilde{\mathcal{Q}}}\left|\frac{1}{N}\sum_{i=1}^{N}\sigma_i\langle w,\Phi'_{\underline{\beta}}(\underline{X}_i)\rangle_{\mathcal{H}_{\underline{\beta}}}\right| \\
&\overset{(a)}{\le} \frac{\Gamma_w}{N}\mathbb{E}\sup_{\|\underline{\beta}\|_1\le\Gamma_{\underline{\beta}},Q\in\tilde{\mathcal{Q}}}\left\|\sum_{i=1}^{N}\sigma_i\Phi'_{\underline{\beta}}(\underline{X}_i)\right\|_{\mathcal{H}_{\underline{\beta}}} \\
&\overset{(b)}{\le} \frac{\Gamma_w}{N}\sqrt{\mathbb{E}\sup_{\|\underline{\beta}\|_1\le\Gamma_{\underline{\beta}},Q\in\tilde{\mathcal{Q}}}\left\|\sum_{i=1}^{N}\sigma_i\Phi'_{\underline{\beta}}(\underline{X}_i)\right\|^2_{\mathcal{H}_{\underline{\beta}}}} \\
&= \frac{\Gamma_w}{N}\Bigg(\mathbb{E}\sup_{\|\underline{\beta}\|_1\le\Gamma_{\underline{\beta}},Q\in\tilde{\mathcal{Q}}}\sum_{i=1}^{N}\left\|\Phi'_{\underline{\beta}}(\underline{X}_i)\right\|^2_{\mathcal{H}_{\underline{\beta}}} \\
&\quad + 2\mathbb{E}\sup_{\|\underline{\beta}\|_1\le\Gamma_{\underline{\beta}},Q\in\tilde{\mathcal{Q}}}\sum_{1\le i<j\le N}\sigma_i\sigma_j\langle\Phi'_{\underline{\beta}}(\underline{X}_i),\Phi'_{\underline{\beta}}(\underline{X}_j)\rangle_{\mathcal{H}_{\underline{\beta}}}\Bigg)^{\frac{1}{2}}
\end{aligned}
\tag{2.15}
$$

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

For the first term in (2.15), we have the following bound for any realization of $\underline{x}_i$

$$
\sup_{\|\underline{\beta}\|_1\le\Gamma_{\underline{\beta}},Q\in\tilde{\mathcal{Q}}}\sum_{i=1}^{N}\left\|\Phi'_{\underline{\beta}}(\underline{x}_i)\right\|^2_{\mathcal{H}_{\underline{\beta}}}
$$

$$
\begin{aligned}
&= \sup_{\substack{\|\underline{\beta}\|_1 \le \Gamma_{\underline{\beta}} \\ Q \in \tilde{\mathcal{Q}}}} \sum_{i=1}^{N} \sum_{\underline{z},\underline{z}'} Q(\underline{z}|\underline{x}_i) Q(\underline{z}'|\underline{x}_i) \langle k_{\underline{\beta}}(\cdot,(\underline{z})), k_{\underline{\beta}}(\cdot,(\underline{z}'))\rangle_{\mathcal{H}_{\underline{\beta}}} \\
&= \sup_{\substack{\|\underline{\beta}\|_1 \le \Gamma_{\underline{\beta}} \\ Q \in \tilde{\mathcal{Q}}}} \sum_{i=1}^{N} \sum_{\underline{z},\underline{z}'} Q(\underline{z}|\underline{x}_i) Q(\underline{z}'|\underline{x}_i) \sum_{s=1}^{S} \beta^s \mathbb{I}[z^s = z'^s] \\
&= \sup_{\|\underline{\beta}\|_1 \le \Gamma_{\underline{\beta}}, Q \in \tilde{\mathcal{Q}}} \sum_{i=1}^{N} \sum_{s=1}^{S} \beta^s \sum_{z^s} Q^2(z^s|x_i^s) \\
&\le N \sup_{\|\underline{\beta}\|_1 \le \Gamma_{\underline{\beta}}} \sum_{s=1}^{S} \beta^s \\
&\le N \Gamma_{\underline{\beta}}
\end{aligned} \tag{2.16}
$$

For the second term in (2.15), we follow the arguments in the proof of Proposition 4 in Section in [8] and use the property of the weighted count kernel, and obtain

$$
\begin{aligned}
&2\mathbb{E} \sup_{\|\underline{\beta}\|_1 \le \Gamma_{\underline{\beta}}, Q \in \tilde{\mathcal{Q}}} \sum_{1 \le i < j \le N} \sigma_i \sigma_j \langle \Phi'_{\underline{\beta}}(\underline{X}_i), \Phi'_{\underline{\beta}}(\underline{X}_j)\rangle_{\mathcal{H}_{\underline{\beta}}} \\
&\le 2(N-1)\sqrt{\frac{N}{2}} \sup_{\|\underline{\beta}\|_1 \le \Gamma_{\underline{\beta}}} \sup_{\underline{z},\underline{z}'} k_{\underline{\beta}}(\underline{z},\underline{z}') \sqrt{2 \log |\tilde{\mathcal{Q}}|} \\
&= 2(N-1)\sqrt{\frac{N}{2}} \sup_{\|\underline{\beta}\|_1 \le \Gamma_{\underline{\beta}}} \sup_{\underline{z},\underline{z}'} \sum_{s=1}^{S} \beta_s \mathbb{I}[z^s = z'^s] \sqrt{2 \log |\tilde{\mathcal{Q}}|} \\
&\le 2(N-1)\sqrt{\frac{N}{2}} \Gamma_{\underline{\beta}} \sqrt{2 \log |\tilde{\mathcal{Q}}|} \\
&= 2(N-1)\Gamma_{\underline{\beta}} \sqrt{N \log |\tilde{\mathcal{Q}}|}.
\end{aligned} \tag{2.17}
$$

Combining (2.16) and (2.17), we obtain

$$
R_N(\tilde{\mathcal{F}}) \le \frac{\Gamma_w \sqrt{\Gamma_{\underline{\beta}}}}{N} \left(N + 2(N-1)\sqrt{N \log |\tilde{\mathcal{Q}}|} \right)^{\frac{1}{2}}. \tag{2.18}
$$

For the case when $\tilde{\mathcal{F}} = \mathcal{F}_0$, (2.6) follows from (2.18) by setting $\tilde{\mathcal{Q}} = \mathcal{Q}_1$ and noticing that $|\mathcal{Q}_0| = L_z^{L_x S}$.

For the case when $\tilde{\mathcal{F}} = \mathcal{F}_1$, (2.7) follows from (2.18) 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

$$|\mathcal{Q}_1| = \left(\binom{L_z}{1} + \binom{L_z}{2} + \cdots + \binom{L_z}{L_z} \right)^{L_x S} \leq 2^{L_z L_x S}. \tag{2.19}$$

2.7 Proof of Upper Bound on Estimation Error

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 2.1. *[81] 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 $1 - \delta$,*

$$\begin{aligned} &\sup_{f \in \mathcal{F}} |\hat{\mathbb{E}}\phi(Yf(X)) - \mathbb{E}\phi(Yf(X))| \\ &\qquad \leq 2L_\phi R_N(\mathcal{F}) + \Gamma_\phi \sqrt{\frac{2 \log \frac{1}{\delta}}{N}}. \end{aligned} \tag{2.20}$$

Applying Lemma 2.1, we have the following bound for our problem:

$$\begin{aligned} &\mathbb{E}\phi(Yw_{\hat{f}}(\underline{X})) - \mathbb{E}\phi(Yw_{\tilde{f}}(\underline{X})) \\ &\quad \leq 2 \sup_{f \in \mathcal{F}} |\hat{\mathbb{E}}\phi(Yw_f(\underline{X})) - \mathbb{E}\phi(Yw_f(\underline{X}))| \\ &\quad \leq 4L_\phi R_N(\mathcal{F}) + 2\Gamma_\phi \sqrt{\frac{2 \log \frac{1}{\delta}}{N}} \\ &\quad \leq 4R_N(\mathcal{F}) + 2\Gamma_\phi \sqrt{\frac{2 \log \frac{1}{\delta}}{N}}, \end{aligned} \tag{2.21}$$

where the last step follows because $L_\phi \leq 1$ 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

$$\phi(x) \leq 1 + |x|. \tag{2.22}$$

It is clear that (2.22) holds for the hinge loss function $\phi(x) = (1-x)_+$. For the logistic loss function $\phi(x) = \log(1 + e^{-x})$, if $x \geq 0$, then

$$\log\left(e^{-x} + 1\right) < e^{-x} \leq 1 + |x|. \tag{2.23}$$

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

$$\log\left(e^{-x} + 1\right) < \log\left(e^{-x+1}\right) = 1 - x \leq 1 + |x|. \tag{2.24}$$

Hence, (2.22) 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.

$$\begin{aligned}
\Gamma_\phi &= \sup_{f\in\mathcal{F}, y\in\{\pm 1\}, \underline{x}\in\mathcal{X}^S} |\phi(y w_f(\underline{x}))| \\
&= \sup_{f\in\mathcal{F}, y\in\{\pm 1\}, \underline{x}\in\mathcal{X}^S} |\phi(y\langle w_f, \Phi'_{\underline{\beta}}(\underline{x})\rangle_{\mathcal{H}_{\underline{\beta}}})| \\
&\leq 1 + \sup_{\substack{\|\underline{\beta}\|_1\leq\Gamma_{\underline{\beta}}, \|w\|_{\mathcal{H}_{\underline{\beta}}}\leq\Gamma_w \\ Q\in\mathcal{Q}, y\in\{\pm 1\}, \underline{x}\in\mathcal{X}^S}} |y_i\langle w_f, \Phi'_{\underline{\beta}}(\underline{x})\rangle_{\mathcal{H}_{\underline{\beta}}}| \\
&= 1 + \sup_{\|\underline{\beta}\|_1\leq\Gamma_{\underline{\beta}}, \|w\|_{\mathcal{H}_{\underline{\beta}}}\leq\Gamma_w, Q\in\mathcal{Q},, \underline{x}\in\mathcal{X}^S} |\langle w_f, \Phi'_{\underline{\beta}}(\underline{x})\rangle_{\mathcal{H}_{\underline{\beta}}}| \\
&\overset{(a)}{\leq} 1 + \Gamma_w \sup_{\|\underline{\beta}\|_1\leq\Gamma_{\underline{\beta}}, Q\in\mathcal{Q},, \underline{x}\in\mathcal{X}^S} \|\Phi'_{\underline{\beta}}(\underline{x})\|_{\mathcal{H}_{\underline{\beta}}} \\
&\overset{(b)}{\leq} 1 + \Gamma_w\sqrt{\Gamma_{\underline{\beta}}}
\end{aligned} \tag{2.25}$$

where (a) follows from the Cauchy-Schwartz inequality, and (b) follows from the steps in (2.16). This concludes the proof.

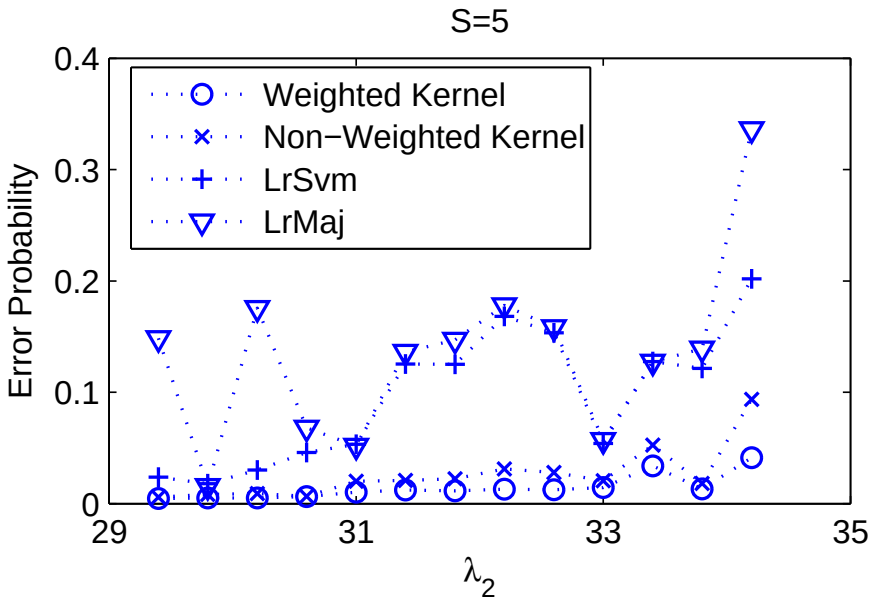

Fig. 2.2: Comparison of probabilities of error among four approaches.

2.8 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 POU $= 0.25$ indicating the probabilities that observations confuse about the event state.

Comparison with Other Approaches. We first 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 [8].

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. 2.2 plots the error probabilities for all approaches, and clearly demonstrates that our weighted kernel based approach outperforms all other competitive methods.

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 reduces as the index of sensors increases. Fig. 2.3 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. 2.4, 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. 2.5, 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 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.

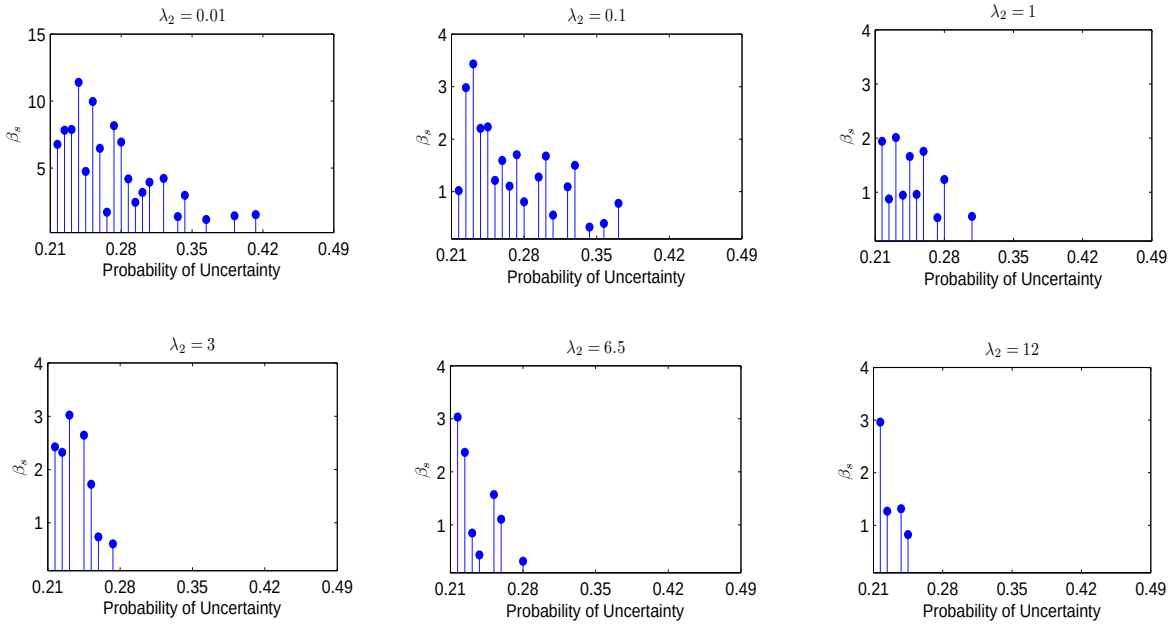

Fig. 2.3: Impact of λ_2 on sparsity of sensor selection

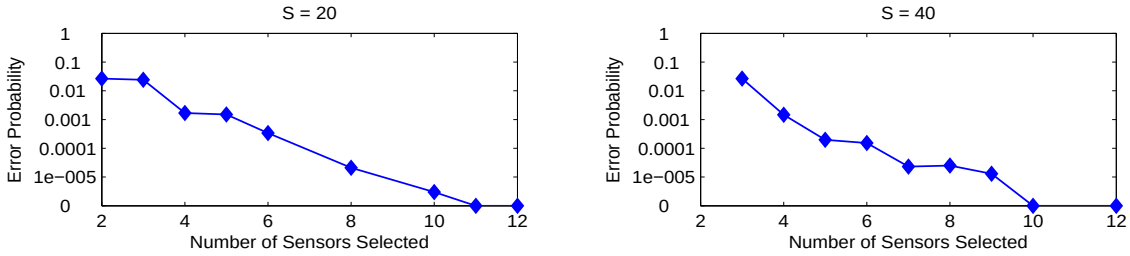

Fig. 2.4: Impact of sparsity on error probability

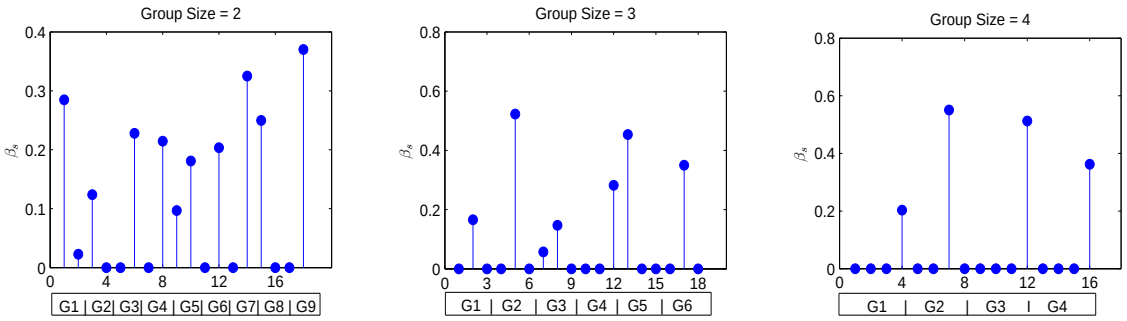

Fig. 2.5: Impact of sample correlation on sensor selection in clustered sensor networks

CHAPTER 3
ALGORITHMS FOR NONPARAMETRIC DECENTRALIZED DETECTION

In Chapter 2, we formulate a regularized empirical risk minimization problem (2.3) to jointly optimize decentralized decision rules and sensor selection strategies. Such an optimization problem is non-convex. In this chapter, we first design efficient algorithms to solve such an optimization problem. We then introduce recent results on convergence analysis of non-convex minimization problem and apply these results to analyze the algorithms that we propose.

3.1 Algorithm Design

In this section, we develop algorithms to solve the risk minimization problem (2.3), 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 (2.3) 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 3.3.

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 [79], the minimizer of the problem given in (2.3) with fixed Q and $\underline{\beta}$ takes the form

$$w = \sum_{i=1}^{N} \alpha_i y_i \Phi'_{\underline{\beta}}(\underline{x}_i) = \sum_{i=1}^{N} \sum_{\underline{z} \in \mathcal{Z}} \alpha_i y_i \Phi_{\underline{\beta}}(\underline{z}) Q(\underline{z}|\underline{x}_i) \tag{3.1}$$

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 (2.3) can be solved equivalently by optimizing over these three types of parameters. Therefore, problem (2.3) is equivalent to the following optimization problem:

$$\min_{\substack{\beta^s \geq 0 \text{ for } s=1,\ldots,S \\ Q \in \mathcal{Q}, \underline{\alpha} \in \mathcal{R}^N}} G(\underline{\alpha}, \underline{\beta}, Q), \tag{3.2}$$

where

$$\begin{aligned} G(\underline{\alpha}, \underline{\beta}, Q) = &\sum_{i=1}^{N} \phi \left(y_i \sum_{j=1}^{N} \alpha_j y_j \left[\sum_{s=1}^{S} \beta^s \sum_{z^s} Q_s(z^s|x_i^s) Q_s(z^s|x_j^s) \right] \right) \\ &+ \frac{\lambda_1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i y_i \alpha_j y_j \left[\sum_{s=1}^{S} \beta^s \sum_{z^s} Q_s(z^s|x_i^s) Q_s(z^s|x_j^s) \right] + \lambda_2 \|\underline{\beta}\|_{l_1}. \end{aligned} \tag{3.3}$$

In Algorithm 1, we develop a gradient projection algorithm to solve the non-convex risk minimization problem (3.2) with a continuous loss function. Here, we combine three types

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 $t \leq 1/L$,

$$(\underline{\alpha}^{(k)}, \hat{\underline{\beta}}^{(k)}, \hat{Q}^{(k)}) = (\underline{\alpha}^{(k-1)}, \underline{\beta}^{(k-1)}, Q^{(k-1)}) - t \bigtriangledown_{(\underline{\alpha}, \underline{\beta}, Q)} G(\underline{\alpha}^{(k-1)}, \underline{\beta}^{(k-1)}, Q^{(k-1)}); \tag{3.4}$$

- Projection of $\underline{\beta}$

$$\underline{\beta}^{(k)} = \text{argmin}_{\beta^s \geq 0} \left\| \underline{\beta} - \hat{\underline{\beta}}^{(k)} \right\|_{l_2}; \tag{3.5}$$

- Projection of Q

$$Q^{(k)} = \text{argmin}_{Q \in \mathcal{Q}} \left\| Q - \hat{Q}^{(k)} \right\|_{l_2}; \tag{3.6}$$

Output: Sensor decision rules $Q_s(Z^s|X^s)$ for $s = 1, \ldots, S$, and fusion center decision rule $w(\underline{Z})$.

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 (3.4), 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 (3.6). 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 3.3.

Algorithm 2 provides an alternative method (refered to as the Gauss-Seidel method) for

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

$$\underline{\alpha}^{(k)} = \underline{\alpha}^{(k-1)} - t_\alpha \bigtriangledown_{\underline{\alpha}} G(\underline{\alpha}^{(k-1)}, \underline{\beta}^{(k-1)}, Q^{(k-1)}); \tag{3.7}$$

- Fix $\underline{\alpha}^{(k)}$ and $Q^{(k-1)}$, for $t_{\underline{\beta}} \leq 1/L$ update

$$\underline{\beta}^{(k)} = \text{argmin}_{\beta^s \geq 0} \left\| \underline{\beta} - \underline{\beta}^{(k-1)} + t_{\underline{\beta}} \bigtriangledown_{\underline{\beta}} G(\underline{\alpha}^{(k)}, \underline{\beta}^{(k-1)}, Q^{(k-1)}) \right\|_{l_2}; \tag{3.8}$$

- Fix $\underline{\alpha}^{(k)}$ and $\underline{\beta}^{(k)}$, for $t_Q \leq 1/L$, update

$$Q^{(k)} = \text{argmin}_{Q \in \mathcal{Q}} \left\| Q - Q^{(k-1)} + t_Q \bigtriangledown_Q G(\underline{\alpha}^{(k)}, \underline{\beta}^{(k)}, Q^{(k-1)}) \right\|_{l_2}; \tag{3.9}$$

Output: Sensor decision rules $Q_s(Z^s|X^s)$ for $s = 1, \ldots, S$, and fusion center decision rule $w(\underline{Z})$.

solving the non-convex optimization problem (3.2) 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 3.3.

We now consider the problem (3.2) 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 (3.2) with $\phi(\cdot)$ being hinge loss function. We note that the inner loop of Algorithm 3 follows the idea of conjugate duality provided in [8]. 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:

$$\alpha_i^* = \begin{cases} 0 & \hat{\alpha}_i^* \leq 0 \\ \hat{\alpha}_i^* & 0 < \hat{\alpha}_i^* < \frac{1}{\lambda_1} \\ \frac{1}{\lambda_1} & \hat{\alpha}_i^* \geq \frac{1}{\lambda_1} \end{cases} \quad \text{for } i = 1, 2, ..., N, \tag{3.10}$$

where

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

(1b) Optimization over Q with w fixed. The subgradient method is used to alternatively update $Q_s(z^s|x^s)$ 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 $Q_s(z^s|x^s)$ is given as follows.

$$-\frac{\lambda_1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i y_i \alpha_j y_j \left[\beta^s \left(Q_s(z^s|x_i^s) \mathcal{I}(x_j^s = x^s) + Q_s(z^s|x_j^s) \mathcal{I}(x_i^s = x^s) \right) \right] \tag{3.11}$$

Furthermore, since this is a constrained optimization problem subject to linear constraints on Q, i.e., $\sum_{z^s} Q(z^s|x^s) = 1$ for $s = 1, \ldots, S$ and for all possible values of x^s, conditional (sub)gradient method for simplex problems in [82, Section 2.2.2] can be applied. Alter-

natively, 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) [83]. 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.

$$G(\underline{\beta}) = \sum_{i=1}^{N} \phi(\langle \underline{\beta}, \underline{d}_i \rangle) + \langle \underline{\beta}, \underline{h} \rangle, \tag{3.12}$$

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

$$F(\underline{\beta}) = G(\underline{\beta}) + i_{\{\beta^s \geq 0,\ s=1,2,\ldots,S\}}(\underline{\beta}) = \sum_{i=1}^{N} g_i(\underline{\beta}) + H(\underline{\beta}), \tag{3.13}$$

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 $H(\underline{\beta}) = \langle \underline{\beta}, \underline{h} \rangle + i_{\{\beta^s \geq 0,\ s=1,2,\ldots,S\}}(\underline{\beta})$. 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:

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

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

$$prox_{\mu g_i}(\underline{\tilde{\beta}}) = \begin{cases} \underline{\tilde{\beta}} & 1 - \langle \underline{\tilde{\beta}}, \underline{d}_i \rangle \leq 0 \\ \frac{1-\langle \underline{\tilde{\beta}}, \underline{d}_i \rangle}{\|\underline{d}_i\|^2} \underline{d}_i + \underline{\tilde{\beta}} & 0 < 1 - \langle \underline{\tilde{\beta}}, \underline{d}_i \rangle < \mu \|\underline{d}_i\|^2 \\ \underline{\tilde{\beta}} + \mu \underline{d}_i & 1 - \langle \underline{\tilde{\beta}}, \underline{d}_i \rangle \geq \mu \|\underline{d}_i\|^2 \end{cases}, \tag{3.14}$$

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

$$\left[prox_{\mu H}(\underline{\tilde{\beta}})\right]_s = \begin{cases} \tilde{\beta}^s - \mu h_s & \tilde{\beta}^s - \mu h_s \geq 0 \\ 0 & \tilde{\beta}^s - \mu h_s < 0 \end{cases}. \tag{3.15}$$

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

$$\begin{cases} \underline{\gamma}_i^{(k)} = prox_{g_i/\rho}(\underline{\nu}^{(k-1)} - \underline{u}_i^{(k-1)}) & \text{for } i = 1, 2, ..., N \\ \underline{\nu}^{(k)} = prox_{H/(N\rho)}(\bar{\underline{\gamma}}^{(k)} + \bar{\underline{u}}^{(k-1)}) & \\ \underline{u}_i^{(k)} = \underline{u}_i^{(k-1)} + \underline{\gamma}_i^{(k-1)} - \underline{\nu}^{(k-1)} & \text{for } i = 1, 2, ..., N, \end{cases} \tag{3.16}$$

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 $\underline{\beta} = \bar{\underline{\gamma}}$.

3.2 Preliminaries on Non-convex Optimization

Although it is in general difficult to design algorithms that converge to a global minimizer of a non-convex function, recent results in [84, 85] establish convergence to critical points

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 (3.10);
  - Fix w, compute the optimal $Q(\underline{z}|\underline{x})$ using the subgradient method by exploiting (3.11);
  - Repeat until inner loop converges;
- Outer loop: fix $\underline{\alpha}$ and Q functions, and compute the optimal $\underline{\beta}$ following (3.16);
- Repeat inner and outer loops until converge.

Output: Sensor decision rules $Q_s(Z^s|X^s)$ for $s = 1, \ldots, S$, and fusion center decision rule $w(\underline{Z})$.

in non-convex optimization. In this section, we introduce the results in [84–86] together with necessary definitions, which are useful for studying our algorithms in the next section.

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 [84] for those definitions. We next define critical points based on Fréchet subdifferential.

Definition 3.1. *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 [84], convergence to critical points in non-convex optimization is established for Kurdyka-Łojasiewicz (KL) functions, the definition of which is given below.

Definition 3.2. *(a) The function $f : \mathcal{R}^n \to \mathcal{R} \cup \{+\infty\}$ is said to have Kurdyka-Łojasiewicz*

(KL) property at $\underline{x}^* \in dom\partial f$ *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*

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

where 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 $dom\partial f$ *are referred to as KL function.*

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

Definition 3.3. *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 [84], the convergence of the gradient projection algorithm for constrained non-convex optimization problems is established, which is summarized as follows.

Theorem 3.1. *[84] 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*

$$\underline{x}^{k+1} \in P_C(\underline{x}^k - \gamma_k \bigtriangledown h(\underline{x}^k)), \text{ with } \underline{x}^0 \in C. \tag{3.17}$$

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

Theorem 3.2. *[84] Consider minimization of a function* $f : \mathcal{R}^{n_1} \times ... \times \mathcal{R}^{n_p} \to \mathcal{R} \bigcup \{+\infty\}$ *having the following structure*

$$f(\underline{x}) = Q(\underline{x}_1, ..., \underline{x}_p) + \sum_{i=1}^{p} f_i(\underline{x}_i), \tag{3.18}$$

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* $i = 1, 2, ..., p$*. Assume that* f *defined in* (3.18) *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* $\underline{x}^0 = (\underline{x}_1^0, ..., \underline{x}_p^0)$ *in* $\mathcal{R}^{n_1} \times ... \times \mathcal{R}^{n_p}$.

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

$$f_i(\underline{x}_i^{k+1}) + Q(\underline{x}_1^{k+1}, ..., \underline{x}_{i-1}^{k+1}, \underline{x}_i^{k+1}, ..., \underline{x}_p^k) + \frac{1}{2}\langle A_i^k(\underline{x}_i^{k+1} - \underline{x}_i^k), \underline{x}_i^{k+1} - \underline{x}_i^k \rangle$$
$$\leq f_i(\underline{x}_i^k) + Q(\underline{x}_1^{k+1}, ..., \underline{x}_{i-1}^{k+1}, \underline{x}_i^k, ..., \underline{x}_p^k); \tag{3.19}$$

$$\underline{v}_i^{k+1} \in \partial f_i(\underline{x}_i^{k+1}); \tag{3.20}$$

$$\|\underline{v}_i^{k+1} + \bigtriangledown_{\underline{x}_i} Q(\underline{x}_1^{k+1}, ..., \underline{x}_i^{k+1}, \underline{x}_{i+1}^k, ..., \underline{x}_p^k)\| \leq b_i \|\underline{x}_i^{k+1} - \underline{x}_i^k\|, \tag{3.21}$$

where $i = 1, ..., p$*, 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.

3.3 Convergence of Gradient Projection-Based Algorithm

In this section, we analyze convergence of Algorithm 1 that we propose in Section 3.1. 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 [84] (see Theorems 3.1 and 3.2), we show that this algorithm converges to critical points of the objective function. This result is provided in the following theorem.

Theorem 3.3. *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 $G(\underline{\alpha}, \underline{\beta}, Q)$.*

Remark 3.1. *A wide range of functions including both logistic loss and exponential loss functions are real analytic. Thus, convergence of Algorithm 1 established in Theorem 3.3 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 3.4. *[87] 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 $\underline{\hat{x}} \in U$, the function $f(\underline{x})$ may be represented by a convergent power series in some neighborhood of $\underline{\hat{x}}$.*

Hence, a real analytic function is continuous and has continuous and real analytic partial derivatives of all orders [87]. The following lemma provides a simple way to verify real analytic functions.

Lemma 3.1. *[87] 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 $\underline{\hat{x}} \in U$, 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

$$\left| \frac{\partial^{|\mu|} f}{\partial \underline{x}^{\mu}} (\underline{x}) \right| \leq C \cdot \frac{\mu!}{R^{|\mu|}}, \ \forall \underline{x} \in V, \tag{3.22}$$

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.

The rest of the section is devoted for the proof of Theorem 3.3.

Proof. Since Algorithm 1 uses the standard projection method as described in Theorem 3.1, 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 (3.3).

It is shown in [88] 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 [89] 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 3.5. *[90](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 $f : \mathcal{R}^n \to \mathcal{R} \bigcup \{+\infty\}$ is called subanalytic if its graph $\{(\underline{x}, \lambda) \in \mathcal{R}^{n+1} : f(\underline{x}) = \lambda\}$ is a subanalytic set.

Definition 3.6. *[84](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 $p_{ij},\ q_{ig} : \mathcal{R}^n \to \mathcal{R}$ *are real polynomial functions for* $1 \leq i \leq p,\ 1 \leq j \leq q$. *A function* $f : \mathcal{R}^n \to \mathcal{R} \bigcup \{+\infty\}$ *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 3.7. *[90](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 $p_{ij},\ q_{ig} : V \to \mathcal{R}$ *are real analytic functions (see Definition 3.4) for* $1 \leq i \leq p$, $1 \leq j \leq q$. *A function* $f : \mathcal{R}^n \to \mathcal{R} \bigcup \{+\infty\}$ *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 3.7 that a real-analytic function is also semi-analytic. It is shown in [91] 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

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

is composition of a real analytic loss function $\phi(\cdot)$ and a polynomial function, which is also real analytic. It has been shown in [87] 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

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

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. □

3.4 Convergence of Regularized Gauss-Seidel Algorithm

In this section, we analyze the 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 3.2, we provide convergence of Algorithm 2 to critical points as follows.

Theorem 3.4. *Assume the loss function $\phi(\cdot)$ in* (3.2) *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* (3.3)*.*

We note that the convergence argument of Algorithm 2 exploits the fact that the objective function takes the structure (3.18) [84, 92]. For Algorithm 3 developed for the case with the non-differentiable loss function, the objective function cannot be expressed in the form given in (3.18). Because the loss function including all three types of variables cannot be viewed as the Q function in (3.18). In this case, it is difficult to establish convergence to

a critical point.

The rest of the section devotes to the proof of Theorem 3.4.

Proof. The proof apply the convergence result on proximal regularization of Gauss-Seidel method (see Theorem 3.2). It has been shown that $F(\underline{\alpha}, \underline{\beta}, Q)$ is a KL function 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 (3.19), (3.20), and (3.21) in Theorem 3.2 are satisfied when updating $\underline{\alpha}$, $\underline{\beta}$, and Q.

We first note that in the context of Theorem 3.2, $Q(\underline{x}_1, ..., \underline{x}_p) = G(\underline{\alpha}, \underline{\beta}, Q)$ with $p = 3$, $\underline{x}_1 = \underline{\alpha}$, $\underline{x}_2 = \underline{\beta}$, and $\underline{x}_3 = Q$, $f_1(\underline{x}_1) = 0$, $f_2(\underline{x}_2) = i_{\{\beta^s \geq 0,\ s=1,2,...,S\}}(\underline{\beta})$, and $f_3(\underline{x}_3) = i_{\{Q \in \mathcal{Q}\}}(Q)$.

We then introduce the following lemma to help our proof.

Lemma 3.2. *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,*

$$f(z) \leq f(x) + \langle \bigtriangledown f(x), z - x \rangle + \frac{L}{2}\|z - x\|^2. \tag{3.23}$$

Verifying the conditions for updating $\underline{\alpha}$:

Step (3.7) implies that

$$\bigtriangledown_{\underline{\alpha}} G(\underline{\alpha}^{(k)}, \underline{\beta}^{(k)}, Q^{(k)}) = \frac{1}{t_\alpha}(\underline{\alpha}^{(k)} - \underline{\alpha}^{(k+1)}). \tag{3.24}$$

Therefore,

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

which implies that (3.21) is satisfied by setting $\underline{v}_\alpha^{(k+1)} = 0$. It is also clear that such $\underline{v}_\alpha^{(k+1)}$ satisfies (3.20) with $f_1(\underline{x}_1) = 0$.

Using Lemma 3.2, we can show that

$$\begin{aligned} & G(\underline{\alpha}^{(k+1)}, \underline{\beta}^{(k)}, Q^{(k)}) - \frac{L}{2}\|\underline{\alpha}^{(k+1)} - \underline{\alpha}^{(k)}\|^2 \\ & + \langle \bigtriangledown_{\underline{\alpha}} G(\underline{\alpha}^{(k)}, \underline{\beta}^{(k)}, Q^{(k)}), \underline{\alpha}^{(k)} - \underline{\alpha}^{(k+1)} \rangle \\ & \leq G(\underline{\alpha}^{(k)}, \underline{\beta}^{(k)}, Q^{(k)}). \end{aligned} \tag{3.25}$$

Substituting $\bigtriangledown_{\underline{\alpha}} G(\underline{\alpha}^{(k)}, \underline{\beta}^{(k)}, Q^{(k)})$ in (3.24) into (3.25), we obtain

$$\begin{aligned} & G(\underline{\alpha}^{(k+1)}, \underline{\beta}^{(k)}, Q^{(k)}) + \left(\frac{1}{t_\alpha} - \frac{L}{2}\right) \|\underline{\alpha}^{(k+1)} - \underline{\alpha}^{(k)}\|^2 \\ & \leq G(\underline{\alpha}^{(k)}, \underline{\beta}^{(k)}, Q^{(k)}). \end{aligned} \tag{3.26}$$

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 (3.19) holds with $A_i^k = \left(\frac{1}{t_\alpha} - \frac{L}{2}\right) I$.

Verifying the conditions for updating $\underline{\beta}$:

Following (3.8), we obtain

$$\begin{aligned} & \left\| \underline{\beta}^{(k+1)} - \underline{\beta}^{(k)} + t_{\underline{\beta}} \bigtriangledown_{\underline{\beta}} G(\underline{\alpha}^{(k+1)}, \underline{\beta}^{(k)}, Q^{(k)}) \right\| \\ & \leq \left\| t_{\underline{\beta}} \bigtriangledown_{\underline{\beta}} G(\underline{\alpha}^{(k+1)}, \underline{\beta}^{(k)}, Q^{(k)}) \right\|. \end{aligned} \tag{3.27}$$

Hence,

$$\begin{aligned} & \left\| \underline{\beta}^{(k+1)} - \underline{\beta}^{(k)} \right\|^2 \\ & + 2\langle \underline{\beta}^{(k+1)} - \underline{\beta}^{(k)}, t_{\underline{\beta}} \bigtriangledown_{\underline{\beta}} G(\underline{\alpha}^{(k+1)}, \underline{\beta}^{(k)}, Q^{(k)}) \rangle \leq 0. \end{aligned} \tag{3.28}$$

Using Lemma 3.2, we can show that

$$\begin{aligned} & G(\underline{\alpha}^{(k+1)}, \underline{\beta}^{(k+1)}, Q^{(k)}) - \frac{L}{2}\|\underline{\beta}^{(k+1)} - \underline{\beta}^{(k)}\|^2 \\ & + \langle \nabla_{\underline{\beta}} G(\underline{\alpha}^{(k+1)}, \underline{\beta}^{(k)}, Q^{(k)}), \underline{\beta}^{(k)} - \underline{\beta}^{(k+1)} \rangle \\ & \leq G(\underline{\alpha}^{(k+1)}, \underline{\beta}^{(k)}, Q^{(k)}). \end{aligned} \tag{3.29}$$

Combining with (3.28), we obtain

$$\begin{aligned} & G(\underline{\alpha}^{(k+1)}, \underline{\beta}^{(k+1)}, Q^{(k)}) + (\frac{1}{2t_{\underline{\beta}}} - \frac{L}{2})\|\underline{\beta}^{(k+1)} - \underline{\beta}^{(k)}\|^2 \\ & \leq G(\underline{\alpha}^{(k+1)}, \underline{\beta}^{(k)}, Q^{(k)}). \end{aligned} \tag{3.30}$$

By choosing $t_{\underline{\beta}} \leq 1/L$, the update on $\underline{\beta}$ satisfies the condition (3.19) with $A_i^k = \left(\frac{1}{t_{\underline{\beta}}} - \frac{L}{2}\right) I$.

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 (3.8) can be equivalently written as

$$\begin{aligned} \underline{\beta}^{(k+1)} = & \text{argmin}_{\underline{\beta}} \frac{1}{2t_\beta} \Big\| \underline{\beta} - \underline{\beta}^{(k)} \\ & + t_{\underline{\beta}} \nabla_{\underline{\beta}} G(\underline{\alpha}^{(k+1)}, \underline{\beta}^{(k)}, Q^{(k)}) \Big\|^2 + i_{C_{\underline{\beta}}}(\underline{\beta}). \end{aligned} \tag{3.31}$$

The problem (3.31) implies that the solution $\underline{\beta}^{(k+1)}$ satisfies the following property:

$$\begin{aligned} 0 \in & \partial i_{C_{\underline{\beta}}}(\underline{\beta}^{(k+1)}) + \nabla_{\underline{\beta}} G(\underline{\alpha}^{(k+1)}, \underline{\beta}^{(k)}, Q^{(k)}) + \\ & \frac{1}{t_{\underline{\beta}}}(\underline{\beta}^{(k+1)} - \underline{\beta}^{(k)}) \end{aligned} \tag{3.32}$$

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

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

We hence have

$$\begin{aligned}&\left\|\underline{v}_{\underline{\beta}}^{(k+1)}+\bigtriangledown_{\underline{\beta}}G(\underline{\alpha}^{(k+1)},\underline{\beta}^{(k)},Q^{(k)})\right\|\\&=\frac{1}{t_{\underline{\beta}}}\left\|(\underline{\beta}^{(k+1)}-\underline{\beta}^{(k)})\right\|.\end{aligned}\tag{3.33}$$

We further derive

$$\begin{aligned}&\left\|\underline{v}_{\underline{\beta}}^{(k+1)}+\bigtriangledown_{\underline{\beta}}G(\underline{\alpha}^{(k+1)},\underline{\beta}^{(k+1)},Q^{(k)})\right\|\\&\leq\left\|\underline{v}_{\underline{\beta}}^{(k+1)}+\bigtriangledown_{\underline{\beta}}G(\underline{\alpha}^{(k+1)},\underline{\beta}^{(k)},Q^{(k)})\right\|\\&+\left\|\bigtriangledown_{\underline{\beta}}G(\underline{\alpha}^{(k+1)},\underline{\beta}^{(k+1)},Q^{(k)})-\bigtriangledown_{\underline{\beta}}G(\underline{\alpha}^{(k+1)},\underline{\beta}^{(k)},Q^{(k)})\right\|\\&\leq\frac{1}{t_{\underline{\beta}}}\left\|(\underline{\beta}^{(k+1)}-\underline{\beta}^{(k)})\right\|+L\left\|(\underline{\beta}^{(k+1)}-\underline{\beta}^{(k)})\right\|.\end{aligned}\tag{3.34}$$

where the last step follows from (3.33) 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 (3.20) and (3.21).

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

CHAPTER 4
SEMI-PARAMETRIC COMPOSITE OUTLIER DETECTION

In this chapter, we study the composite outlier detection problem in semi-parametric scenario, where typical distribution is known but outlying distributions are not known. We first give the mathematical description of composite outlier detection and the preliminary studies on parametric models. Then we study the single outlier model and multi-outlier model and investigate the conditions for GLRT to be consistent or exponentially consistent.

4.1 Problem Formulation

Suppose there are M data sequences represented by $y^{(i)}$ for $i = 1, 2, \ldots, M$. Each data sequence consists of n independent and identically distributed (i.i.d.) samples, and different sequences are generated independent of each other. Among these sequences, a typical sequence contains samples drawn from a distribution π; and an outlying sequence contains samples drawn from a distribution μ. We assume that $\mu \neq \pi$, and both μ and π are discrete over a support set $\mathcal{Y}$. Our goal is to determine the existence of outlier sequences, i.e.,

distinguish between the following two hypotheses:

$$H_0: \text{ All sequences } y^{(i)} \text{ are typical, for } i = 1, 2, \ldots, M.$$

$$H_1: \text{ There exist at least one outlier sequence.}$$

Denote all data sequences as $y^{Mn} = \left(y^{(1)}, y^{(2)}, \ldots, y^{(M)}\right)$. A test $\delta: \ y^{Mn} \to \{H_0, H_1\}$ maps realization of data sequences y^{Mn} into either H_0 or H_1. Under the null hypothesis H_0, y^{Mn} takes one underlying distribution, i.e., all samples are generated independently by π. However, under the alternative hypothesis H_1, y^{Mn} may take multiple distributions depending on which sequences are outliers. For the simple single outlier model, H_1 corresponds to the case with only one outlier sequence, where the one outlier can be any of M sequences. For a more general multi-outlier model, H_1 corresponds to the case with $t \geq 1$ outliers, where the index set S of outliers can be any subset of $\mathcal{M} = \{1, \ldots, M\}$ such that cardinality $|S| = t$. Thus, the above problem can be viewed as a binary composite hypothesis testing problem with multiple sub-hypotheses under H_1.

We measure the performance of a test δ using type I error $e_1(\delta)$ and type II error $e_2(\delta)$. Type I error refers to the probability that null hypothesis H_0 occurs but δ decides the alternative hypothesis H_1, and is given by

$$e_1(\delta) = P_{H_0}(\delta = 1); \tag{4.1}$$

whereas type II error refers to the probability that H_1 occurs but δ decides H_0, and is given by

$$e_2(\delta) = P_{H_1}(\delta = 0). \tag{4.2}$$

We define the following risk function $R(\delta)$ to measure the overall performance of a test

$$R(\delta) = P_{H_0}(\delta = 1) + \max_{S \in \mathcal{M}:|S|=t} P_{H_1}(\delta = 0). \tag{4.3}$$

A test δ is said to be *consistent* if the risk function $R(\delta)$ decays to zero as the sample size n goes to infinity. We further define the exponent $E_R(\delta)$ of the risk function as

$$E_R(\delta) = \lim_{n\to\infty} -\frac{1}{n} \log R(\delta). \tag{4.4}$$

A test is said to be *exponentially consistent* if $E_R(\delta) > 0$, i.e., the risk function $R(\delta)$ converges to zero exponentially.

4.2 Parametric Model

While the analysis of the parametric model with both π and μ being known can be implied from previous work/existing understanding, we include it here as an intermediate step towards analysis of the semi-parametric and non-parametric models, which are the main focus of the work. This section also sets up the notations that we adopt in the work.

We first consider the simple case of the single outlier scenario, i.e., there exists only one outlier under H_1. We develop the GLRT as follows. First, under H_0, the likelihood of observing y^{Mn} is given by

$$\begin{aligned} P_0(y^{Mn}) = L_0(y^{Mn}, \pi) &= \prod_{k=1}^{n} \prod_{j=1}^{M} \pi(y_k^{(j)}) \\ &= \exp\left\{ -n \sum_{j=1}^{M} H(\gamma_j) - n \sum_{j=1}^{M} D(\gamma_j \| \pi) \right\} \end{aligned} \tag{4.5}$$

where γ_j denotes the empirical distribution of $y^{(j)}$, and $D(p\|q)$ denotes the KL divergence

between distributions p and q given as follows

$$D(p\|q) = \sum_y p(y) \log \frac{p(y)}{q(y)}. \tag{4.6}$$

Secondly, under H_1 with the i-th sequence being the outlier, the likelihood of observing y^{Mn} is given by

$$\begin{aligned} P_i(y^{Mn}) = L_i(y^{Mn}, \pi, \mu) &= \prod_{k=1}^{n} \left[\mu(y_k^{(i)}) \prod_{j \neq i} \pi(y_k^{(j)}) \right] \\ &= \exp\left\{ -n \sum_{j=1}^{M} H(\gamma_j) - nD(\gamma_i||\mu) - n \sum_{j \neq i} D(\gamma_j||\pi) \right\}. \end{aligned} \tag{4.7}$$

We apply the GLRT given by

$$\delta(y^{Mn}) : \ \frac{1}{n} \log \frac{\max_i P_i(y^{Mn})}{P_0(y^{Mn})} \underset{H_0}{\overset{H_1}{\gtrless}} \tau,$$

where τ is a threshold constant. Substituting (4.5) and (4.7) into the above test, we obtain the following test for our problem:

$$\max_i \ D(\gamma_i\|\pi) - D(\gamma_i\|\mu) \underset{H_0}{\overset{H_1}{\gtrless}} \tau. \tag{4.8}$$

Differently from multi-hypothesis testing in [37], a threshold is needed for binary composite hypothesis testing here. Furthermore, the value of threshold τ is critical to the performance of the test. To have an intuitive understanding of how to set τ, we first note that under H_0, all γ_i for $i = 1, \ldots, M$ converges to π for large n, and thus the test value (i.e., the left side of (4.8)) converges to $-D(\pi\|\mu)$ as n goes to infinity. Then under H_1, γ_j of the outlier converges to μ, and hence $D(\gamma_j\|\pi) - D(\gamma_j\|\mu)$ converges to $D(\mu\|\pi)$. Other γ_i for $i \neq j$ still converges to π, and the corresponding $D(\gamma_j||\pi) - D(\gamma_j||\mu)$ converges to $-D(\pi||\mu)$. Thus, the overall test value converges to $D(\mu||\pi)$ under H_1. Therefore, a threshold τ between $D(\mu||\pi)$ and $-D(\pi||\mu)$ should distinguish between the two hypothe-

ses, and hence we set $\tau = 0$. The following theorem characterizes the optimality of the above GLRT in terms of the decay rate of the risk function.

Theorem 4.1. *Consider the binary composite outlier detection problem* (4.1) *with single outlier. Suppose μ and π are both known. The GLRT* (4.8) *with the threshold $\tau = 0$ is exponentially consistent, and achieves the optimal exponent of the risk function $R(\delta)$ given by*

$$E_R(\delta) = C(\mu, \pi), \tag{4.9}$$

where $C(p, q)$ denotes the Chernoff distance between distributions p and q given by

$$C(p, q) = \max_{0 \le \lambda \le 1} -\log \left(\sum_y p(y)^\lambda q(y)^{1-\lambda} \right). \tag{4.10}$$

Proof. This proof consists of two parts. The achievability proof (see Section 4.5) develops the upper bound on the risk function for the GLRT, and shows that the exponent $E_R(\delta) = C(\mu, \pi)$ can be achieved. The optimality proof (i.e., the converse proof) justifies that $C(\mu, \pi)$ is the optimal exponent among all tests, which we develop as follows.

Consider the following simple binary hypothesis testing problem.

$$\begin{aligned} H_0 :&\ \text{All sequences } y^{(i)} \text{ are typical, } i = 1, 2, \ldots, M. \\ H_1 :&\ \text{Only sequence } y^{(1)} \text{ is outlier. Other sequences are typical.} \end{aligned} \tag{4.11}$$

Under H_0, $\left(y_k^{(1)}, y_k^{(2)}, \ldots, y_k^{(M)}\right)$ follows the joint distribution $\prod_{i=1}^{M} \pi$ for $k = 1, \ldots, n$. Under H_1, $\left(y_k^{(1)}, y_k^{(2)}, \ldots, y_k^{(M)}\right)$ follows the joint distribution $\mu \prod_{i=1}^{M-1} \pi$ for $k = 1, \ldots, n$. Based on [93, Theorem 11.9.1], the optimal error exponent under the Bayesian risk (with

the uniform prior) is given by

$$C\left(\prod_{i=1}^{M}\pi\ ,\ \mu\prod_{i=1}^{M-1}\pi\right)=C(\pi,\mu). \tag{4.12}$$

It is easy to see that any test for the composite problem results in a smaller risk function for the simple binary problem (4.11). Thus, the exponent of the composite problem cannot exceed $C(\pi,\mu)$ for the simple binary problem. □

We next consider a more general model with multiple outliers, i.e., the number of outliers $t \geq 1$ in (4.1). We assume that t is fixed and known. We further take a more general model, assuming each sequence i is drawn from a certain outlying distribution μ_i if it is an outlier, where μ_i for $i = 1,\ldots,M$ are not necessarily the same. We use $\{\mu\}$ to denote the set of outlying distributions μ_i for $i = 1,2,\ldots,M$. All sequences take the same typical distribution π if they are not outliers.

To develop a test, we note that the likelihood function under H_0 is the same as $P_0(y^{Mn})$ in (4.5), and under H_1, the likelihood corresponding to outliers with indexes in S is given by

$$\begin{aligned}
P_S(y^{Mn}) &= L_s(y^{Mn},\pi,\{\mu\}) \\
&= \prod_{k=1}^{n}\left(\prod_{j\in S}\mu_j(y_k^{(j)})\prod_{j\notin S}\pi(y_k^{(j)})\right) \\
&= \exp\left\{-n\sum_{j=1}^{M}H(\gamma_j)-n\sum_{j\in S}D(\gamma_j||\mu_j)-n\sum_{j\notin S}D(\gamma_j||\pi)\right\}
\end{aligned} \tag{4.13}$$

Thus, the GLRT

$$\delta(y^{Mn}):\frac{1}{n}\log\frac{\max_{S,|S|=t}P_S(y^{Mn})}{P_0(y^{Mn})}\underset{H_0}{\overset{H_1}{\gtrless}}\tau$$

can be expressed further as

$$\max_{S\in\mathcal{M},|S|=t} \sum_{j\in S} D(\gamma_j||\pi) - D(\gamma_j||\mu_j) \underset{H_0}{\overset{H_1}{\gtrless}} \tau. \tag{4.14}$$

To further simplify the notation, suppose $S = \{j_1, j_2, \ldots, j_t\}$. Then let $\mu_S := \prod_{k=1}^{t} \mu_{j_t}$ and $\gamma_S = \prod_{k=1}^{t} \gamma_{j_t}$. Also denote $\pi_t = \prod_{i=1}^{t} \pi$. Then the above test can be rewritten as

$$\max_{S,|S|=t} D(\gamma_S||\pi_t) - D(\gamma_S||\mu_S) \underset{H_0}{\overset{H_1}{\gtrless}} \tau. \tag{4.15}$$

To set the threshold in the above test, we analyze the test value under H_0 and H_1 similarly to the single outlier model. The test value converges to $-\min_{S\in\mathcal{M},|S|=t} D(\pi_t||\mu_S)$ under H_0 since $\gamma_S \to \pi_t$, and converges to $D(\mu_S||\pi)$ where S contains the indexes of true outliers under H_1. Apparently, $D(\mu_S||\pi)$ can take different values as the index set S changes. It is clear that $\tau = 0$ distinguishes between the two hypotheses. The following theorem characterizes the performance of the above GLRT.

Theorem 4.2. *Consider the binary composite outlier detection problem* (4.1) *with t outliers, where t is fixed and known. Suppose both π and μ_j for $j = 1, 2, \ldots, M$ are known. The GLRT* (4.14) *with the threshold $\tau = 0$ is exponentially consistent, and achieves the optimal exponent of the risk function $R(\delta)$ given by*

$$E_R(\delta) = \min_{S\in\mathcal{M},|S|=t} C(\mu_S, \pi_t). \tag{4.16}$$

Proof. See Section 4.6. □

We note that Theorem 4.2 for $t = 1$ generalizes Theorem 4.1 to allow sequences to take different outlying distributions if they are outliers.

4.3 Semiparametric Single Outlier Model

We next extend our study to the semi-parametric model, where typical distribution π is known but outlying distributions $\{\mu\}$ are unknown. We first study the simpler model with a single outlier in this section, and then generalize our study to the model with multiple outliers in section 4.4.

We first note that the likelihood function under H_0 is the same as that in (4.5) for the parametric model, because the typical distribution π is known. However, under H_1, since μ is unknown, the likelihood function cannot be written out directly. Instead, under the sub-hypothesis that sequence i is the outlier, we compute the likelihood as follows by replacing μ with its estimate $\hat{\mu}_i = \gamma_i$, where γ_i is the empirical distribution of sequence i.

$$\begin{aligned}
\hat{P}_i(y^{Mn}) &= L_i(y^{Mn}, \pi, \gamma_i) \\
&= \prod_{k=1}^{n} \left[\gamma_i(y_k^{(i)}) \prod_{j \neq i} \pi(y_k^{(j)}) \right] \\
&= \exp\left\{ -n \sum_{j=1}^{M} H(\gamma_j) - nD(\gamma_i||\gamma_i) - n \sum_{j \neq i} D(\gamma_j||\pi) \right\} \\
&= \exp\left\{ -n \sum_{j=1}^{M} H(\gamma_j) - n \sum_{j \neq i} D(\gamma_j||\pi) \right\} \qquad (4.17)
\end{aligned}$$

Thus, GLRT

$$\delta(y^{Mn}) : \frac{1}{n} \log \frac{\max_i \hat{P}_i(y^{Mn})}{P_0(y^{Mn})} \underset{H_0}{\overset{H_1}{\gtrless}} \tau$$

takes the following form by substituting (4.5) and (4.17)

$$\max_i D(\gamma_i||\pi) \underset{H_0}{\overset{H_1}{\gtrless}} \tau. \qquad (4.18)$$

In order to get insight about how to choose τ, we note that the test value (i.e., the

left-hand side of (4.18)) converges to 0 under H_0 because $\gamma_i \to \pi$ for all $i = 1, \ldots, M$, and converges to $D(\mu||\pi)$ under H_1 because $\gamma_i \to \mu$ if sequence i is the outlier. Therefore, choosing $0 < \tau < D(\mu||\pi)$ should distinguish between the two hypotheses, as we characterize in the following theorem.

Theorem 4.3. *Consider the binary composite outlier detection problem* (4.1) *with a single outlier. Suppose π is known and μ is unknown. Further assume that $D(\mu||\pi)$ is known. Then the GLRT* (4.8) *with the threshold $\tau \in (0, D(\mu||\pi))$ is exponentially consistent, and achieves the exponent of the risk function $R(\delta)$ given by*

$$E_R(\delta) = \min \left\{ \tau, \min_{q: D(q||\pi) \leq \tau} D(q||\mu) \right\}. \tag{4.19}$$

Proof. See Section 4.7. □

It is interesting to note that the exponential consistency of GLRT does not require the full knowledge of μ but only the value of $D(\mu||\pi)$ to set a appropriate threshold. Furthermore, the appearance of μ in the exponent does not mean that μ is exploited in the test, but only implies that the performance depends on the underlying outlying distribution. The optimization problem $\min_{q: D(q||\pi) \leq \tau} D(q||\mu)$ does not have an explicit solution. However, this problem is convex and can be solved numerically in an efficient manner.

It is also clear that the exponent $E_R(\delta)$ varies by choosing different threshold $\tau \in (0, D(\mu||\pi))$. The following corollary characterizes the value of τ that yields the maximum error exponent.

Corollary 4.1. *The exponent $E_R(\delta)$ is equal to $C(\pi, \mu)$ if $\tau = C(\mu, \pi)$, which is optimal for the semi-parametric model with a single outlier.*

Proof. We want to argue that $\min_{D(q||\pi) \leq \tau} D(q||\mu)$ has optimal value $C(\mu, \pi)$ if $\tau = C(\mu, \pi)$.

Here we introduce two other helping problems that we have already known the solutions.

$$\begin{aligned} &\min_{q} D(q||\pi) \\ &\text{s.t. } D(q||\pi) \geq D(q||\mu) \end{aligned} \tag{4.20}$$

$$\begin{aligned} &\min_{q} D(q||\mu) \\ &\text{s.t. } D(q||\pi) \leq D(q||\mu) \end{aligned} \tag{4.21}$$

We know from [93] that both problems have the same solution q^* and optimal value $D(q^*||\mu) = D(q^*||\pi) = C(\pi, \mu)$. We will argue that (4.53) with $\tau = C(\mu, \pi)$ has the same solution q^* by contradiction. Suppose (4.53) has a minimizer $\hat{q}$ different from q^*. Since q^* is a feasible point for problem (4.53), then $D(\hat{q}||\mu) < D(q^*||\mu)$. From constraint of (4.53) we also know

$$D(\hat{q}||\pi) < C(\pi, \mu). \tag{4.22}$$

Consider two different cases.

- $D(\hat{q}||\mu) < D(\hat{q}||\pi)$

  In this case $\hat{q}$ is a feasible point for (4.20). Therefore, $D(q^*||\pi) = C(\pi, \mu) \leq D(\hat{q}||\pi)$, which contradicts (4.22).

- $D(\hat{q}||\mu) \geq D(\hat{q}||\pi)$

  $\hat{q}$ is a feasible point for (4.21). Therefore $C(\pi, \mu) \leq D(\hat{q}||\mu)$, which contradicts (4.22).

So (4.53) also has the optimal value $C(\pi, \mu)$. Considering $\tau = C(\pi, \mu)$ in this case, the solution of (4.53) is equal to $C(\pi, \mu)$. Since the optimal error exponent for parametric model is also $C(\pi, \mu)$, this is the optimal error exponent for the semi-parametric model. $\square$

Corollary 4.1 implies the following two facts: (1) Setting $\tau = C(\mu, \pi)$ achieves the best risk decay exponent over all threshold in GLRT; and (2) Such GLRT achieves the best risk decay exponent over all tests for semi-parametric model, because the exponent is the same as that for the parametric model. Thus, the error performance of the semi-parametric model in terms of the risk decay exponent can be as good as that of the parametric model as long as the threshold τ is set to be $C(\pi, \mu)$. Hence, only the knowledge of $C(\pi, \mu)$ instead of the full knowledge of μ is required.

On the other hand, if no information about μ is available, the following theorem states that there does not exist a test that is exponentially consistent for all μ.

Theorem 4.4. *Consider the binary composite outlier detection problem* (4.1) *with a single outlier. Suppose π is known and μ is unknown. For any test δ constructed without any knowledge about outlier distribution, there must exist a μ such that δ is not exponentially consistent.*

Proof. The theorem follows as a special case of Theorem 4.7 for multi-outlier model. □

Theorem 4.4 essentially claims that it is impossible to construct an exponentially consistent test for all μ without any knowledge about the distance between μ and π such as $D(\mu||\pi)$ and $C(\pi, \mu)$. It is thus of interest to explore whether there exists consistent test for all μ in such a case although exponential consistency for all μ is not possible. The following theorem provides such a solution.

Theorem 4.5. *Consider the binary composite outlier detection problem* (4.1) *with a single outlier. Suppose π is known and μ is unknown. Further assume that $D(\mu||\pi)$ and $C(\pi, \mu)$ are unknown. Set the threshold τ to satisfy $\tau_n \to 0$ and*

$$\tau_n > \frac{|\mathcal{Y}| \log(n+1)}{n}, \tag{4.23}$$

where $|\mathcal{Y}|$ is the cardinality of the support set of π and μ. Then GLRT is universally consistent. Furthermore, the type II error is universally exponentially consistent.

Proof. The theorem follows as a special case of Theorem 4.8 with $t = 1$. □

It is clear that choosing τ within $(0, D(\mu||\pi))$ is necessary for GLRT to be consistent. Moreover, large distance between τ and 0 guarantees small type I error, and large distance between τ and $D(\mu||\pi)$ guarantees small type II error. Since $D(\mu||\pi)$ is unknown, a diminishing τ_n eventually falls in the range $(0, D(\mu||\pi))$ for large enough n by sacrificing exponential consistency of type I error while still keeping exponential consistency of type II error. Furthermore, τ_n cannot converge to zero faster than the test value (i.e., the left-hand side of (4.18)) under H_0, which is guaranteed by the condition (4.23).

4.4 Semi-parametric Multi-outlier Model

We further generalize our study to the semi-parametric model with t outliers, in which sequence i takes the distribution μ_i if it is an outlier for $i = 1, \ldots, M$. We assume that the number t of outliers is fixed and given.

In order to construct the GLRT, we first note that the likelihood function under H_0 is the same as (4.5). Under H_1 and the sub-hypothesis with sequences supported in S being outliers, we compute the likelihood as follows by replacing μ_i with its empirical estimate γ_i for $i \in S$.

$$\begin{aligned}
\hat{P}_S(y^{Mn}) &= L_s(y^{Mn}, \pi, \{\gamma\}) \\
&= \prod_{k=1}^{n} \left(\prod_{j \in S} \gamma_j(y_k^{(j)}) \prod_{j \notin S} \pi(y_k^{(j)}) \right) \\
&= \exp\left\{ -n \sum_{j=1}^{M} H(\gamma_j) - n \sum_{j \notin S} D(\gamma_j||\pi) \right\} \qquad (4.24)
\end{aligned}$$

Then the corresponding GLRT

$$\delta(y^{Mn}) : \frac{1}{n} \log \frac{\max_{S,|S|=t} \hat{P}_S(y^{Mn})}{P_0(y^{Mn})} \underset{H_0}{\overset{H_1}{\gtrless}} \tau$$

can be expressed as follows by substituting (4.5) and (4.25)

$$\max_{S\in\mathcal{M},|S|=t} D(\gamma_S||\pi_t) \underset{H_0}{\overset{H_1}{\gtrless}} \tau. \tag{4.25}$$

To set the threshold τ, we note that the test value (i.e., the lefthand side of (4.25)) converges to 0 under H_0 because $\gamma_j \to \pi$ for $j = 1, 2, \ldots, M$, and converges to $\sum_{j\in S} D(\mu_j||\pi)$ under H_1 and outliers are supported on S. Different from the single outlier problem, $\sum_{j\in S} D(\mu_j||\pi)$ varies for different S. Hence, in order for the threshold τ to distinguish between H_0 and any sub-hypothesis associated with S, it is reasonable to choose

$$0 < \tau < \min_{S\in\mathcal{M},|S|=t} D(\mu_S||\pi_t).$$

Theorem 4.6. *Consider the binary composite outlier detection problem* (4.1) *with* t *outliers, where* t *is fixed and known. Suppose* π *is known but* μ_j *for* $j = 1, 2, \ldots, M$ *are unknown. Further assume that* $\min_{S\in\mathcal{M},|S|=t} D(\mu_S||\pi_t) := d$ *is known. Then the GLRT* (4.14) *with the threshold* $\tau \in (0, d)$ *is exponentially consistent, and achieves the risk decay exponent* $R(\delta)$ *given by*

$$\min\{\alpha(\delta), \beta(\delta)\} \tag{4.26}$$

where $\alpha(\delta) = \tau$ *and* $\beta(\delta)$ *is given by*

$$\min_{\substack{S\in\mathcal{M} \\ s.t.\ |S|=t}} \min_{\substack{q_{\mathcal{M}}= \\ \{q_1,\ldots,q_M\}}} D(q_S||\mu_S) + D(q_{S^c}||\pi_{M-t})$$
$$s.t.\ D(q_{S'}||\pi_t) \le \tau \quad \textit{for all } S', |S'| = t. \tag{4.27}$$

Proof. See Section 4.8. □

In the above theorem, $\alpha(\delta)$ and $\beta(\delta)$ respectively correspond to the exponents of the type I and type II error probabilities. It can be seen that the GLRT with the specified τ is exponentially consistent test without knowing the exact outlying distributions $\{\mu\}$ but only the distance $\min_{S\in\mathcal{M},|S|=t} D(\mu_S||\pi_t)$. Although the exponent given as the solution to the convex optimization problem (4.27) does not have an explicit form, it can be solved using numerical methods efficiently. It can also be verified that the optimization problem (4.27) reduces to that in Theorem 4.6 for the single outlier model by setting $t = 1$ and $\mu_j = \mu$ for $j = 1, 2, \ldots, M$.

The following corollary characterizes the value of τ that yields the maximum error exponent.

Corollary 4.2. *The exponent of the risk is equal to* $\min_{S\in\mathcal{M},|S|=t} C(\mu_S, \pi_t)$ *if* $\tau = \min_{S\in\mathcal{M},|S|=t} C(\mu_S, \pi_t)$*, which is optimal for the semi-parametric multi-outlier model.*

Proof. By choosing $\tau = \min_{S,|S|=t} C(\mu_S, \pi_t)$, $\alpha(\delta) = \min_{S,|S|=t} C(\mu_S, \pi_t)$. Our goal is to show (4.62) also has optimal value $\min_{S,|S|=t} C(\mu_S, \pi_t)$. Optimal value of (4.62) is larger than or equal to

$$\begin{aligned} &\min_{S,\,|S|=t} \min_{q_S} D(q_S||\mu_S) \\ &\qquad \text{s.t. } D(q_S||\pi_t) \le \tau \end{aligned} \tag{4.28}$$

Denote $S^* = \text{argmin}_S\, C(\mu_S, \pi_t)$, this optimization problem is further lower bounded by

$$\begin{aligned} &\min_{q_{S^*}} D(q_{S^*}||\mu_{S^*}) \\ &\text{s.t. } D(q_{S^*}||\pi_t) \le \tau \end{aligned} \tag{4.29}$$

This is similar to problem (4.53) we solved in single outlier model. This problem can achieve $C(\mu_{S^*}, \pi_t)$ if $\tau = C(\mu_{S^*}, \pi_t)$. This analysis shows (4.62) has a solution no smaller than $C(\mu_{S^*}, \pi_t)$. Since the overall error exponent cannot exceed that of parametric model, optimal value of (4.62) equals $C(\mu_{S^*}, \pi_t)$.

Note that the error exponents achieve for the semi-parametric model is the same as that of the parametric model, which justifies our error exponents are the optimal. □

Corollary 4.2 implies that GLRT achieves the same exponent of the risk as the parametric model, and thus is optimal in terms of the risk decay exponent over all tests for semi-parametric model as long as the threshold τ is set to be $\min_{S\in\mathcal{M},|S|=t} C(\mu_S, \pi_t)$.

The next two theorems are in parallel to the single outlier model, providing understanding for the case with no knowledge about the outlying distributions $\{\mu\}$ at all.

Theorem 4.7. *Consider the binary composite outlier detection problem* (4.1) *with t outliers, where t is fixed and known. Suppose π is known but μ_j for $j = 1, 2, \ldots, M$ are unknown. For any test δ constructed without any knowledge about outlier distributions, there must exist $\{\mu\}$ such that δ is not exponentially consistent.*

Proof. See Section 4.9. □

Theorem 4.8. *Consider the binary composite outlier detection problem* (4.1) *with t outliers, where t is fixed and known. Suppose π is known but μ_j for $j = 1, 2, \ldots, M$ are unknown. Further assume that no information on distance between π and $\{\mu\}$ is known. Set the threshold τ in GLRT to satisfy $\tau_n \to 0$ and*

$$\tau_n > \frac{t|\mathcal{Y}|\log(n+1)}{n}$$

where $|\mathcal{Y}|$ is the cardinality of the support set of π and μ. Then GLRT is universally consistent. Furthermore, the type II error is universally exponentially consistent.

Proof. To analyze if type I error probability is universally consistent, we derive the upper bound for it.

$$\begin{aligned}
P_0(\delta = 1) &= P_0\left(\bigcup_{S,|S|=t}\left\{\sum_{i\in S} D(\gamma_i||\pi) \geq \tau\right\}\right) \\
&\leq \binom{M}{t} P_0\left(\sum_{i\in S_1} D(\gamma_i||\pi) \geq \tau\right) \\
&\leq \binom{M}{t}(n+1)^{t|\mathcal{Y}|} P_0\left(\sum_{i\in S_1} D(\gamma_i^*||\pi) \geq \tau,\ \gamma_i^* \text{ is the observed type for } i \in S_1\right) \\
&= \binom{M}{t}(n+1)^{t|\mathcal{Y}|}\exp(-n\tau) \\
&= \binom{M}{t}\exp(t|\mathcal{Y}|\log(n+1) - n\tau) \\
&= \binom{M}{t}\exp\left(n\left(\frac{t|\mathcal{Y}|\log(n+1)}{n} - \tau\right)\right) \qquad (4.30)
\end{aligned}$$

Plug in $\tau_n > \frac{t|\mathcal{Y}|\log(n+1)}{n}$ and $\tau_n \to 0$, $P_0(\delta = 1)$ converges to 0, which indicates the type I error probability is universally consistent.

To analyze type II error exponent $\beta(\delta)$, we can investigate (4.62). $\tau \to 0$ as $n \to \infty$, which makes solution of (4.62) to converge to $\min_{S,|S|=t} D(q_S||\pi_t)$ as $n \to \infty$. Therefore the type two error is exponentially consistent by choosing this diminishing τ_n. □

4.5 Proof of Optimality for Single Outlier Parametric Model

We characterize achievable error exponent of the risk function under GLRT by analyzing the type I and type II errors using Sanov's Theorem. First of all, the type I error is given by

$$\begin{aligned}
P_0(\delta = 1) &= P_0(\max_i D(\gamma_i||\pi) - D(\gamma_i||\mu) \geq 0) \\
&= P_0\left((\gamma_1, \gamma_2, \ldots, \gamma_M) \in E_\alpha^{Mn}\right) \qquad (4.31)
\end{aligned}$$

where E_α^{Mn} is given by

$$\begin{aligned} E_\alpha^{Mn} = \{(q_1, \ldots, q_M) :\ & D(q_1||\pi) - D(q_1||\mu) \geq 0 \\ & \text{or } D(q_2||\pi) - D(q_2||\mu) \geq 0 \\ & \ldots\ldots \\ & \text{or } D(q_M||\pi) - D(q_M||\mu) \geq 0\}. \end{aligned}$$

Applying Sanov's Theorem to (4.31), we obtain the error exponent $\alpha(\delta)$ of the type I error as the solution of the following optimization problem.

$$\alpha(\delta) = \min_{(q_1,\ldots,q_M)\in E_\alpha^{Mn}} D\left(\prod_{i=1}^{M} q_i \middle\| \pi^M\right) \tag{4.32}$$

Due to the non-convexity of the set E_α^{Mn}, the problem cannot be solved directly. However, this problem can be simplified by observing E_α^{Mn} as the union of M subsets given by

$$\begin{aligned} E_\alpha^{Mn} =& \{(q_1, \ldots, q_M) : D(q_1||\pi) - D(q_1||\mu) \geq 0\} \\ & \bigcup \{(q_1, \ldots, q_M) : D(q_2||\pi) - D(q_2||\mu) \geq 0\} \\ & \bigcup \ldots\ldots \\ & \bigcup \{(q_1, \ldots, q_M) : D(q_M||\pi) - D(q_M||\mu) \geq 0\} \end{aligned} \tag{4.33}$$

Hence, solving (4.32) can be decomposed into finding an optimal solution on each subset and then taking the minimum over all M solutions. Therefore, (4.32) is equivalent to the following problem

$$\begin{aligned} \min_{i=1,\ldots,M} \min_{(q_1,\ldots,q_M)} & D(q_1||\pi) + D(q_2||\pi) + \ldots + D(q_M||\pi) \\ \text{s.t. } & D(q_i||\pi) - D(q_i||\mu) \geq 0 \end{aligned} \tag{4.34}$$

Due to the symmetry of (4.34) over $i = 1, \ldots, M$, the optimization problem can be further equalized and simplified to

$$\begin{aligned} &\min_{q} D(q||\pi) \\ &\text{s.t. } D(q||\pi) - D(q||\mu) \geq 0 \end{aligned} \tag{4.35}$$

The solution to (4.35) has been given in [93] to equal $C(\pi, \mu)$, which is the exponent of the type I error.

We next study the type II error exponent $\beta(\delta)$ by analyzing the error under each sub-hypothesis. Suppose sequence 1 $y^{(1)}$ is the outlier, and the error probability is given by

$$\begin{aligned} P_1(\delta = 0) &= P_1(\max_i D(\gamma_i||\pi) - D(\gamma_i||\mu) \leq 0) \\ &= P_1\left((\gamma_1, \gamma_2, \ldots, \gamma_M) \in E_\beta^{Mn}\right) \end{aligned} \tag{4.36}$$

where E_β^{Mn} is given by

$$\begin{aligned} E_\beta^{Mn} = \{(q_1, \ldots, q_M) : & D(q_1||\pi) - D(q_1||\mu) \leq 0 \\ & \text{and } D(q_2||\pi) - D(q_2||\mu) \leq 0 \\ & \quad \ldots\ldots \\ & \text{and } D(q_M||\pi) - D(q_M||\mu) \leq 0\}. \end{aligned}$$

Applying Sanov's Theorem to (4.36), the error exponent of $P_1(\delta = 0)$ is the limit of

solution of the following.

$$\min_{(q_1,\ldots,q_M)\in E_\beta^{Mn}} D\left(\prod_{i=1}^{M} q_i||\mu\pi^{M-1}\right)$$
$$\text{s.t.} \begin{cases} D(q_1||\pi) - D(q_1||\mu) \leq 0 \\ D(q_2||\pi) - D(q_2||\mu) \leq 0 \\ \ldots \\ D(q_M||\pi) - D(q_M||\mu) \leq 0. \end{cases}$$

We observe that constraints in the above problem are separable, and hence the optimal solution can be computed via summation over the solutions to the following M subproblems:

$$\begin{aligned} &\min_{q_1} D(q_1||\mu) \\ &\text{s.t. } D(q_1||\pi) - D(q_1||\mu) \leq 0 \end{aligned} \tag{4.37}$$

and

$$\begin{aligned} &\min_{q_i} D(q_i||\pi) \\ &\text{s.t. } D(q_i||\pi) - D(q_i||\mu) \leq 0 \\ &\quad \text{for } i = 2, \ldots, M. \end{aligned} \tag{4.38}$$

Apparently, the optimal solution to (4.37) equals Chernoff distance $C(\mu, \pi)$ as shown in [93], and the optimal value for (4.38) is 0 by setting $q_i = \pi$. Therefore the error exponent for $P_1(\delta = 0)$ is given by $C(\mu, \pi)$. Furthermore, due to the symmetry of sub-hypothesis under H_1 in the error performance, we conclude that the maximum of the error exponent of the type II error over all sub-hypothesis is also given by $C(\mu, \pi)$. Thus, the risk decay exponent of GLRT is given by $C(\mu, \pi)$.

4.6 Proof of Optimality for Multi-Outlier Parametric Model

This proof consists of the optimality argument that justifies the exponent can be no larger than $\min_{S,\,|S|=t} C(\mu_S, \pi_t)$ and the achievability proof that shows such an exponent can be achieved by GLRT.

Proof of optimality. We first note that H_1 contains sub-hypotheses, each of which corresponds to outliers supported by one index set $S \in \mathcal{M}$ with the size t. For a given such a set S, consider the following simple binary hypothesis testing problem

$$H_0: \text{ All sequences } y^{(i)} \text{ are typical, } i = 1, 2, \ldots, M.$$

$$H_1: \text{ There exist outlier sequences with index in known set } S$$

It is clear that the optimal error exponent for the above problem is $C(\mu_S, \pi_t)$.

Furthermore, any test for our binary composite problem can be applied to the above problem, and it is easy to see that the risk function of the composite problem is lower bounded by the type II error of the above problem. Hence, the exponent of the composite problem is upper bounded by $C(\mu_S, \pi_t)$. Such an argument can be applied to all $S \in \mathcal{M}$ with $|S| = t$, and we conclude that the exponent of the composite problem is upper bounded by $\min_{S\in\mathcal{M},\,|S|=t} C(\mu_S, \pi_t)$.

Proof of achievability. We also analyze the achievable error exponent of the risk function under GLRT using Sanov's Theorem. The probability of type I error is given by

$$\begin{aligned} P_0(\delta = 1) &= P_0\left(\max_{S,|S|=T} \sum_{j\in S} D(\gamma_j||\pi) - D(\gamma_j||\mu_j) \geq 0\right) \\ &= P_0\left((\gamma_1, \gamma_2, \ldots, \gamma_M) \in E_\alpha^{Mn}\right) \end{aligned} \tag{4.39}$$

where E_α^{Mn} is given by

$$E_\alpha^{Mn} = \left\{ (q_1, \ldots, q_M) : \sum_{j \in S} D(q_j||\pi) - D(q_j||\mu_j) \geq 0 \text{ for at least one } S \in \mathcal{M} \right\}. \tag{4.40}$$

We apply Sanov's Theorem and evaluate the error exponent $\alpha(\delta)$ of the type I error as the solution of the following optimization problem.

$$\begin{aligned} \min_{S,\, |S|=t} \min_{q_j,\, j \in S} & \sum_{j \in S} D(q_j||\pi) \\ \text{s.t. } & \sum_{j \in S} D(q_j||\pi) - D(q_j||\mu_j) \geq 0 \end{aligned} \tag{4.41}$$

To simplify this problem, we can define a joint distribution of $q_{j_1}, q_{j_2}, \ldots, q_{j_t}$ for $j_1, j_2, \ldots, j_t \in S$.

$$q_S(y_1, y_2, \ldots, y_t) = q_{j_1}(y_1) \times q_{j_2}(y_2) \times \ldots \times q_{j_t}(y_t) \tag{4.42}$$

Then (4.41) can be simplified as

$$\begin{aligned} \min_{S,\, |S|=t} \min_{q_j,\, j \in S} & D(q_S||\pi_t) \\ \text{s.t. } & D(q_S||\pi_t) - D(q_S||\mu_S) \geq 0, \end{aligned} \tag{4.43}$$

which is similar to the type I error exponent optimization problem in the single outlier case. This has an optimal value

$$\min_{S,\, |S|=t} C(\mu_S, \pi_t).$$

We next explore the type II error exponent $\beta(\delta)$ by analyzing the error under each sub-hypothesis. Assume set S with size t is the true index set for outlier sequences, and the

error is given by

$$P_S(\delta = 0) = P_S\left(\max_{S,|S|=t} \sum_{j\in S} D(\gamma_j||\pi) - D(\gamma_j||\mu_j) \leq 0\right)$$
$$= P_S\left((\gamma_1, \gamma_2, \ldots, \gamma_M) \in E_\beta^{Mn}\right), \tag{4.44}$$

where E_β^{Mn} is given by

$$E_\beta^{Mn} = \left\{(q_1, \ldots, q_M) : \sum_{j\in S} D(q_j||\pi) - D(q_j||\mu_j) \leq 0 \text{ for all } S \in \mathcal{M}\right\}. \tag{4.45}$$

Applying Sanov's Theorem, the error exponent of $P_S(\delta = 0)$ is given by

$$\min_{q_j,\, j=1,\ldots,M} \sum_{j\in S} D(q_j||\mu_j) + \sum_{j\in S^c} D(q_j||\pi)$$
$$\text{s.t.} \sum_{j\in S'} D(q_j||\pi) - D(q_j||\mu_j) \leq 0 \quad \text{for all } S', |S'| = t \tag{4.46}$$

Since the index set S varies, the error exponent for different S also varies. Considering the overall type II error exponent is dominated by the smallest one, thus the type II error exponent $\beta(\delta)$ is given by

$$\min_{S,\, |S|=t} \min_{q_j,\, j=1,\ldots,M} \sum_{j\in S} D(q_j||\mu_j) + \sum_{j\in S^c} D(q_j||\pi)$$
$$\text{s.t.} \sum_{j\in S'} D(q_j||\pi) - D(q_j||\mu_j) \leq 0 \quad \text{for all } S', |S'| = t. \tag{4.47}$$

Solution of (4.47) is apparently larger than or equal to the solution of

$$\min_{S,\, |S|=t} \min_{q_j,\, j=1,\ldots,M} \sum_{j\in S} D(q_j||\mu_j)$$
$$\text{s.t.} \sum_{j\in S} D(q_j||\pi) - D(q_j||\mu_j) \leq 0 \tag{4.48}$$

which has the solution given by $\min_{S,\,|S|=t} C(\mu_S, \pi_t)$. This analysis indicates the decay exponent of the risk function is no smaller than $\min_{S,\,|S|=t} C(\mu_S, \pi_t)$. Combining with the optimality result, we can conclude that the optimal exponent for the risk function is $\min_{S,\,|S|=t} C(\mu_S, \pi_t)$.

4.7 Proof of Exponentially Consistency for Semi-parametric Single Outlier Model

We characterize achievable error exponent of the risk function under GLRT by analyzing the type I and type II errors using Sanov's Theorem. We analyze type I error probability $P_0(\delta = 1)$ first.

$$\begin{aligned} P_0(\delta = 1) =& P_0(\max_i D(\gamma_i||\pi) \geq \tau) \\ =& P_0((\gamma_1, \gamma_2, \ldots, \gamma_M) \in E_\alpha^{Mn}) \end{aligned} \tag{4.49}$$

where E_α^{Mn} is given by

$$\begin{aligned} E_\alpha^{Mn} = \{(q_1, \ldots, q_M) :\ & D(q_1||\pi) \geq \tau \\ & \text{or } D(q_2||\pi) \geq \tau \\ & \ldots\ldots \\ & \text{or } D(q_M||\pi) \geq \tau\}. \end{aligned}$$

Applying Sanov's Theorem to the above error probability and exploring the symmetry of the optimization problem over $i = 1, \ldots, M$, the type I error exponent $\alpha(\delta)$ is the

optimal value of the following problem.

$$\alpha(\delta) = \min_{q} D(q||\pi) \tag{4.50}$$

$$\text{s.t. } D(q||\pi) \geq \tau \tag{4.51}$$

Since $D(q||\pi)$ can achieve any value within the interval $[0, D(\mu||\pi)]$, (4.50) has optimal value τ, which establishes that $\alpha(\delta)$ is positive.

We can further derive the exponent $\beta(\delta)$ for type II error using Sanov's Theorem. Without loss of generality, we suppose the first sequence $y^{(1)}$ is the outlier sequence. The corresponding error probability is

$$\begin{aligned} P_1(\delta = 0) &= P_1(\max_i D(\gamma_i||\pi) \leq \tau) \\ &= P_1\left((\gamma_1, \gamma_2, \ldots, \gamma_M) \in E_\beta^{Mn}\right) \end{aligned} \tag{4.52}$$

where E_β^{Mn} is given by

$$\begin{aligned} E_\beta^{Mn} = \{(q_1, \ldots, q_M) : & D(q_1||\pi) \leq \tau \\ & \text{and } D(q_2||\pi) \leq \tau \\ & \ldots\ldots \\ & \text{and } D(q_M||\pi) \leq \tau\}. \end{aligned}$$

Apply Sanov's Theorem, the exponent for $P_1(\delta = 0)$ is the optimal value of the following problem.

$$\beta(\delta) = \min_{q} D(q||\mu) \tag{4.53}$$

$$\text{s.t. } D(q||\pi) \leq \tau \tag{4.54}$$

The above equality holds for scenarios with other sequence being outlier, which justifies

that the type II error exponent is given by (4.53).

Since $D(\mu||\pi) > \tau$, the distribution μ is not a feasible point of (4.53). Thus, (4.53) has a positive optimal value. Since the exponents for both type I and type II error probabilities are positive, GLRT is exponentially consistent. The overall error exponent is dominated by the smaller one, which is given by

$$\min\left\{\tau, \min_{D(q||\pi)\leq\tau} D(q||\mu)\right\}. \tag{4.55}$$

4.8 Proof of Exponentially Consistency for Semi-parametric Muli-outlier Model

We analyze the type I and type II error probabilities using Sanov's Theorem. The type I error probability is given by

$$P_0(\delta = 1) = P_0\left(\max_{S,|S|=t} D(\gamma_S||\pi_t) \geq \tau\right) = P_0\left((\gamma_1, \gamma_2, \ldots, \gamma_M) \in E_\alpha^{Mn}\right), \tag{4.56}$$

where E_α^{Mn} is given by

$$E_\alpha^{Mn} = \left\{(q_1, \ldots, q_M) : \sum_{j\in S} D(q_j||\pi) \geq \tau \text{ for at least one } S \in \mathcal{M}\right\}. \tag{4.57}$$

Applying Sanov's Theorem, we obtain the type I error exponent $\alpha(\delta)$ given by

$$\begin{aligned}
&\min_{S,\,|S|=t} \min_{q_j,\,j\in S} \sum_{j\in S} D(q_j||\pi) \qquad \text{s.t. } \sum_{j\in S} D(q_j||\pi) \geq \tau \\
&= \min_{S,\,|S|=t} \min_{q_S} D(q_S||\pi_t) \qquad \text{s.t. } D(q_S||\pi_t) \geq \tau \\
&= \tau
\end{aligned} \tag{4.58}$$

To analyze the type II error, if S is the true index set for outlier sequences, then the

corresponding error probability is given by

$$\begin{aligned} P_S(\delta = 0) &= P_S(\max_{S,|S|=t}(\gamma_S||\pi_t) \leq \tau) \\ &= P_S\left((\gamma_1, \gamma_2, \ldots, \gamma_M) \in E_\beta^{Mn}\right), \end{aligned} \tag{4.59}$$

where E_β^{Mn} is given by

$$E_\beta^{Mn} = \left\{(q_1, \ldots, q_M) : \sum_{j\in S} D(q_j||\pi) \leq \tau \text{ for all } S \in \mathcal{M}\right\}. \tag{4.60}$$

Then the overall type II error exponent is dominated by the smallest exponent of error probabilities over all possible index sets of outlier sequences, and is given by

$$\beta(\delta) = \min_{S,\ |S|=t} \lim_{n\to\infty} -\frac{1}{n} \log P_S(\max_{S,|S|=t}(\gamma_S||\pi_t) \leq \tau). \tag{4.61}$$

Applying Sanov's Theorem, we obtain the type II error exponent $\beta(\delta)$ given by

$$\begin{aligned} &\min_{S,\ |S|=t}\ \min_{q_j,\ j=1,\ldots,M} \sum_{j\in S} D(q_j||\mu_j) + \sum_{j\in S^c} D(q_j||\pi) \\ &\qquad\qquad \text{s.t. } \sum_{j\in S'} D(q_j||\pi) \leq \tau \quad \text{for all } S', |S'| = t \\ &= \min_{S,\ |S|=t}\ \min_{q_S} D(q_S||\mu_S) + D(q_{S^c}||\pi_{M-t}) \\ &\qquad\qquad \text{s.t. } D(q_{S'}||\pi_t) \leq \tau \quad \text{for all } S', |S'| = t. \end{aligned} \tag{4.62}$$

Since $\min_{S,|S|=t} D(\mu_S||\pi_t) > \tau$, $(q_S, q_{S^c}) = (\mu_S, \pi_{M-t})$ is not a feasible point, which implies that the solution of (4.62) is positive. Since both $\alpha(\delta)$ and $\beta(\delta)$ are positive, GLRT is exponentially consistent.

4.9 Proof of Converse for Semi-parametric Multi-outlier Model

This proof follows the steps similar to those in the proof of Theorem 11 in [37]. We include the proof here for completeness.

We first show that the empirical distribution $(\gamma_1, \ldots, \gamma_M)$ and π are sufficient statistics for the error exponent. The idea is to show that for any test δ, there exists another test δ', which depends only on $(\gamma_1, \ldots, \gamma_M)$ and π, and achieves the same error exponent.

Denote $T_{(\gamma_1,\ldots,\gamma_M)}$ as the set of all M sequences that have empirical distributions $(\gamma_1, \ldots, \gamma_M)$. Within $T_{(\gamma_1,\ldots,\gamma_M)}$, δ may decide either H_0 or H_1 for each sequence. Denote $T^{0,\delta}_{(\gamma_1,\ldots,\gamma_M)} \subseteq T_{(\gamma_1,\ldots,\gamma_M)}$ as the set of sequences over which δ decides H_0, and $T^{1,\delta}_{(\gamma_1,\ldots,\gamma_M)} \subseteq T_{(\gamma_1,\ldots,\gamma_M)}$ as the set of sequences over which δ decides H_1. Apparently $T_{(\gamma_1,\ldots,\gamma_M)} = T^{0,\delta}_{(\gamma_1,\ldots,\gamma_M)} \bigcup T^{1,\delta}_{(\gamma_1,\ldots,\gamma_M)}$. We let δ' map all sequences within $T_{(\gamma_1,\ldots,\gamma_M)}$ to only H_0 or H_1.

Given δ, we construct δ' such that it decides H_0 if $|T^{0,\delta}_{(\gamma_1,\ldots,\gamma_M)}| \geq \frac{1}{2}|T_{(\gamma_1,\ldots,\gamma_M)}|$, and H_1 otherwise. It follows that

$$\max\{P_0(\delta' = 1), P_1(\delta' = 0)\} \leq 2\max\{P_0(\delta = 1), P_1(\delta = 0)\}.$$

Hence, the error exponent of δ' is the same as error exponent of δ. Therefore, empirical distribution $(\gamma_1, \ldots, \gamma_M)$ and π are sufficient statistics for the error exponent.

We next show that there does not exist a test δ that is exponentially consistent for arbitrary set of outlier distributions $\{\mu\}$. We argue by contradiction. Suppose such a test δ exists. The type I error exponent of δ is

$$\alpha(\delta) = \lim_{n\to\infty} -\frac{1}{n}\log P_0(\delta = 1) > \epsilon. \tag{4.63}$$

Since δ is independent of $\{\mu\}$, the lower bound of type I error exponent ϵ is also indepen-

dent of $\{\mu\}$. Denote $\{(q_1,\ldots,q_M):\delta=0\}$ and $\{(q_1,\ldots,q_M):\delta=1\}$ as the decision regions for H_0 and H_1 respectively under δ. Then there must exist ϵ such that $\alpha(\delta)>\epsilon$ and $\beta(\delta)>0$ for all possible $\{\mu\}$ using this test δ.

We next construct the following set Z of distributions

$$Z=\left\{(q_1,\ldots,q_M):\sum_{i=1}^{M}D(q_i||\pi)\leq\frac{\epsilon}{2}\right\}. \tag{4.64}$$

We want to argue $Z\subseteq\{(q_1,\ldots,q_M):\delta=0\}$. Suppose this is not true and there exists $(\hat{q}_1,\ldots,\hat{q}_M)\in Z$ and $(\hat{q}_1,\ldots,\hat{q}_M)\in\{(q_1,\ldots,q_M):\delta=1\}$. Using Sanov's Theorem,

$$\lim_{n\to\infty}-\frac{1}{n}\log P_0(\delta=1)=\min_{(q_1,\ldots,q_M)\in\{\delta=1\}}\sum_{i=1}^{M}D(q_i||\pi) \tag{4.65}$$

Considering the existence of $(\hat{q}_1,\ldots,\hat{q}_M)$, the above value is no larger than $\epsilon/2$, which contradicts (4.63). Therefore $Z\subseteq\{(q_1,\ldots,q_M):\delta=0\}$.

We next analyze the type II error exponent

$$\beta(\delta)=\lim_{n\to\infty}-\frac{1}{n}\log P_1(\delta=0).$$

Applying Sanov's Theorem and combining the fact that $Z\subseteq\{(q_1,\ldots,q_M):\delta=0\}$, we obtain

$$\begin{aligned}\beta(\delta)\leq\min_{S,|S|=t}\min_{q_i,i=1,\ldots,M}&\sum_{i\in S}D(q_i||\mu_i)+\sum_{i\notin S}D(q_i||\pi)\\ &\text{s.t. }(q_1,\ldots,q_M)\in Z\end{aligned} \tag{4.66}$$

For any ϵ, there must exist $\{\mu\}$ that $\min_{S,|S|=t}\sum_{i\in S}D(\mu_i||\pi)<\frac{\epsilon}{2}$. For such $\{\mu\}$, we can construct the solution of (4.66) as $(q_1,\ldots,q_M)$, where

$$q_i=\mu_i \text{ for } i\in S,\ q_i=\pi \text{ for } i\notin S. \tag{4.67}$$

This solution makes $\beta(\delta) = 0$, which is a contradiction.

CHAPTER 5
NONPARAMETRIC COMPOSITE OUTLIER DETECTION

In this chapter, we study the composite outlier detection problem in nonparametric scenario, where both the typical distribution and the outlying distributions are unknown. We study the single outlier model and multi-outlier model and investigate the conditions for GLRT to be consistent or exponentially consistent.

5.1 Single Outlier Model

We construct GLRT based on the idea of using the empirical distributions to estimate μ and π. Differently from the parametric and semi-parametric models, the typical distribution π is also estimated from the average of all empirical distributions of typical sequences. Under H_0 with no outliers, we estimate π as

$$\tilde{\pi}_0 = \frac{\sum_{i=1}^{M} \gamma_i}{M}, \tag{5.1}$$

and the corresponding likelihood function is given by

$$\begin{aligned}\tilde{P}_0(y^{Mn}) = L_0(y^{Mn}, \tilde{\pi}_0) &= \sum_{k=1}^{n}\sum_{j=1}^{M}\tilde{\pi}_0(y_k^{(j)}) \\ &= \exp\left\{-n\sum_{j=1}^{M}H(\gamma_j) - n\sum_{j=1}^{M}D(\gamma_j||\tilde{\pi}_0)\right\}. \end{aligned} \tag{5.2}$$

Under H_1 and the sub-hypothesis that sequence i is the outlier, we use the average of empirical distributions of all sequences except $y^{(i)}$ to estimate π, and the empirical distribution of $y^{(i)}$ to estimate μ as follows.

$$\tilde{\pi}_i = \frac{\sum_{j\neq i}\gamma_j}{M-1}$$

$$\tilde{\mu}_i = \gamma_i$$

The corresponding likelihood function is given by

$$\begin{aligned}\tilde{P}_i(y^{Mn}) &= L_i(y^{Mn}, \tilde{\pi}_i, \gamma_i) \\ &= \prod_{k=1}^{n}\left(\gamma_i(y_k^{(i)})\prod_{j\neq i}\tilde{\pi}_i(y_k^{(j)})\right) \\ &= \exp\left\{-n\sum_{j=1}^{M}H(\gamma_j) - n\sum_{j\neq i}D(\gamma_j||\tilde{\pi}_i)\right\} \end{aligned} \tag{5.3}$$

Hence, GLRT, which is given as

$$\delta(y^{Mn}) : \frac{1}{n}\log\frac{\max_i \tilde{P}_i(y^{Mn})}{\tilde{P}_0(y^{Mn})} \underset{H_0}{\overset{H_1}{\gtrless}} \tau$$

can be further expressed as follows by substituting the likelihoods under H_0 and H_1

$$\max_i\left[D(\gamma_i||\tilde{\pi}_0) + \sum_{j\neq i}\left(D(\gamma_j||\tilde{\pi}_0) - D(\gamma_j||\tilde{\pi}_i)\right)\right] \underset{H_0}{\overset{H_1}{\gtrless}} \tau. \tag{5.4}$$

In order to get insight about how to choose τ, we note that under H_0, $\tilde{\pi}_0$, $\tilde{\pi}_i$ and γ_i all converge to π for $i = 1, \ldots, M$, and hence the test value (i.e., the lefthand side of (5.4)) converges to 0. Under H_1, although $\tilde{\pi}_0$ is different from $\tilde{\pi}_i$ for all $i = 1, \ldots, M$, this difference is relatively small if the sequence number M is large. For a large M, we expect the second term $\sum_{j \neq i} \left(D(\gamma_j || \tilde{\pi}_0) - D(\gamma_j || \tilde{\pi}_i) \right)$ is close to 0. Under H_1, the dominant part of the test value is the first term $D(\gamma_j || \tilde{\pi}_0)$, which is close to $D\left(\mu \middle\| \frac{1}{M}\mu + \frac{M-1}{M}\pi \right)$ as $\gamma_j \to \mu$ and $\tilde{\pi}_0 \to \frac{1}{M}\mu + \frac{M-1}{M}\pi$.

Theorem 5.1. *Consider the binary composite outlier detection problem* (4.1) *with a single outlier. Suppose neither π nor μ is known. But assume that $D\left(\mu \middle\| \frac{1}{M}\mu + \frac{M-1}{M}\pi \right) := d$ is known. Then the GLRT* (5.4) *with the threshold $\tau \in (0, d)$ is exponentially consistent, and achieves the exponent of the risk function $R(\delta)$ given by*

$$E_R(\delta) = \min\{\alpha(\delta), \beta(\delta)\} \tag{5.5}$$

where

$$\begin{aligned} \alpha(\delta) = \min_{q_j,\, j=1,\ldots,M} & D(q_1||\pi) + D(q_2||\pi) + \ldots + D(q_M||\pi) \\ & \textit{s.t. } D(q_1||\bar{q}) + \sum_{j \neq 1} \left[D(q_j||\bar{q}) - D(q_j||\bar{q}_{-1}) \right] \geq \tau \end{aligned} \tag{5.6}$$

and

$$\begin{aligned} \beta(\delta) = \min_{q_j,\, j=1,\ldots,M} & D(q_1||\mu) + \sum_{j \neq 1} D(q_j||\pi) \\ & \textit{s.t. } D(q_k||\bar{q}) + \sum_{j \neq k} \left[D(q_j||\bar{q}) - D(q_j||\bar{q}_{-k}) \right] \leq \tau \\ & \textit{for } k = 1, 2, \ldots, M. \end{aligned} \tag{5.7}$$

Furthermore, in the above definitions for $\alpha(\delta)$ and $\beta(\delta)$, $\bar{q} = \sum_{j=1}^{M} q_j / M$ and $\bar{q}_{-i} =$

$\sum_{j\neq i} q_j/(M-1)$ *for* $i = 1, \ldots, M$.

Proof. See Section 5.3. □

To demonstrate whether $\alpha(\delta)$ and $\beta(\delta)$ are positive, we are interested in finding the lower bounds of them. To analyze the lower bound of $\alpha(\delta)$, suppose the solution of problem (5.6) is $(q_1^*, q_2^*, \ldots, q_M^*)$. The optimal value is

$$\begin{aligned}
\alpha(\delta) &= \sum_{j=1}^{M} D(q_j^*||\pi) \\
&\geq \sum_{j=1}^{M} D(q_j^*||\pi) - \sum_{j=1}^{M} D\left(q_j^*||\bar{q}^*\right) + \sum_{j\neq i} D(q_j^*||\bar{q}_{-i}^*) + \tau \\
&= \sum_{j=1}^{M} \sum_{y} q_j^*(y) \log \frac{\bar{q}^*(y)}{\pi(y)} + \sum_{j\neq i} D(q_j^*||\bar{q}_{-i}^*) + \tau \\
&= MD(\bar{q}^*||\pi) + \sum_{j\neq i} D(q_j^*||\bar{q}_{-i}^*) + \tau \\
&\geq \tau
\end{aligned} \tag{5.8}$$

The first inequality follows from the constraint of (5.6).

The lower bound for $\beta(\delta)$ is as follows.

$$
\begin{aligned}
\beta(\delta) &\geq \begin{cases} \min\limits_{q_j,\, j=1,\ldots,M} D(q_1||\mu) + \sum\limits_{j\neq 1} D(q_j||\pi) \\ \text{s.t. } D(q_1||\bar{q}) + \sum\limits_{j\neq 1}(D(q_j||\bar{q}) - D(q_j||\bar{q}_{-1})) \leq \tau \end{cases} \\
&= \begin{cases} \min\limits_{q_j,\, j=1,\ldots,M} D(q_1||\mu) + \sum\limits_{j\neq 1} D(q_j||\pi) \\ \text{s.t. } D\left(q_1||\bar{q}\right) + (M-1)D\left(\bar{q}_{-1}||\bar{q}\right) \leq \tau \end{cases} \\
&= \begin{cases} \min\limits_{q_j,\, j=1,\ldots,M} D(q_1||\mu) + \sum\limits_{j\neq 1} D(q_j||\pi) \\ \text{s.t. } D\left(q_1||\frac{1}{M}q_1 + \frac{M-1}{M}\bar{q}_{-1}\right) + (M-1)D\left(\bar{q}_{-1}||\frac{1}{M}q_1 + \frac{M-1}{M}\bar{q}_{-1}\right) \leq \tau \end{cases} \\
&\geq \begin{cases} \min\limits_{q_1,\bar{q}_{-1}} D(q_1||\mu) + (M-1)D(\bar{q}_{-1}||\pi) \\ \text{s.t. } D\left(q_1||\frac{1}{M}q_1 + \frac{M-1}{M}\bar{q}_{-1}\right) \leq \tau \end{cases}
\end{aligned} \tag{5.9}
$$

For the first inequality, we adapt from (5.7) by removing all constraints for $k = 2, 3, \ldots, M$ and leaving the constraint for $k = 1$. Since $\tau < D\left(\mu \middle\| \frac{1}{M}\mu + \frac{M-1}{M}\pi\right)$, $(q_1, \bar{q}_{-1}) = (\mu, \pi)$ is not a feasible point for (5.9), which indicates solution of this problem must be positive.

Since both $\alpha(\delta)$ and $\beta(\delta)$ are lower bounded by 0, GLRT is exponentially consistent. Furthermore, such exponential consistency does not exploit full knowledge of distributions but only the distance between distributions in terms of the KL divergence $D\left(\mu \middle\| \frac{1}{M}\mu + \frac{M-1}{M}\pi\right)$ to set the threshold in GLRT. For large M, $D\left(\mu \middle\| \frac{1}{M}\mu + \frac{M-1}{M}\pi\right)$ is close to $D(\mu||\pi)$, which implies that the range for setting threshold for the nonparametric model is almost the same as that for the semi-parametric model. We also note that both optimization problems (5.6) and (5.7) have no explicit solutions but can be solved numerically.

The following corollary characterizes the asymptotic exponent as M gets large.

Corollary 5.1. *As $M \to \infty$, the exponent of $R(\delta)$ converges to $C(\pi, \mu)$ if $\tau = C(\mu, \pi)$, which is the optimal exponent that can be achieved for the nonparametric model with a*

single outlier over all tests.

Proof. We analyze the lower bounds for both $\alpha(\delta)$ and $\beta(\delta)$. Since $\alpha(\delta) \geq \tau$, we have

$$\alpha(\delta) \geq C(\mu, \pi). \tag{5.10}$$

Optimization problem in (5.36) can be rewritten as follows.

$$\begin{aligned} &\min_{q_j,\, j=1,\ldots,M} \sum_{j=1}^{M} D(q_j||\pi) \\ &\text{s.t. } \sum_{j=1}^{M} D\left(q_j||\bar{q}\right) - \sum_{j\neq i} D(q_j||\bar{q}_{-i}) \geq \tau \end{aligned} \tag{5.11}$$

Suppose $(q_1^*, q_2^*, \ldots, q_M^*)$ is the true solution for this problem. Define $\bar{q}_{-1}^*(M) = \sum_{j=2}^{M} q_j^*/(M-1)$. Observe $\bar{q}_{-1}^*(M) \to \pi$ as $M \to \infty$, otherwise the optimal value is infinity.

This problem can be equally written as

$$\begin{aligned} &\min_{q_1} D(q_1||\mu) + (M-1)D(\bar{q}_{-1}^*(M)||\pi) \\ &\text{s.t. } D\left(q_1 \middle\| \frac{1}{M}q_1 + \frac{M-1}{M}\bar{q}_{-1}^*(M)\right) \leq \tau \end{aligned} \tag{5.12}$$

which is lower bounded by

$$\begin{aligned} &\min_{q} D(q||\mu) \\ &\text{s.t. } D\left(q \middle\| \frac{1}{M}q + \frac{M-1}{M}\bar{q}_{-1}^*(M)\right) \leq \tau \end{aligned} \tag{5.13}$$

We are interested in the behavior of this problem as $M \to \infty$. Such a limit behavior can be analyzed using the following theorem for objection functions that are Γ-convergence and equi-coercive.

Theorem 5.2. *Let (X, d) be a metric space, let $\{f_n\}$ be a equi-coercive sequence of functions on X, and let $\{f_n\}$ Γ-converges to f, then $\min_{x\in X} f(x)$ exists and*

$$\min_{x\in X} f(x) = \lim_{n\to\infty} \inf_{x\in X} f_n(x). \tag{5.14}$$

Definition 5.1. *Let (Y, d) be a metric space and consider a sequence of functions $\{f_n\}$ where $f_n : Y \to [-\infty, \infty]$. We say that $\{f_n\}$ Γ-converges to a function $f : Y \to [-\infty, \infty]$ if the following properties hold.*

(i)(Liminf Inequality) For every $y \in Y$ and every sequence $\{y_n\} \subset Y$ such that $y_n \to y$,

$$f(y) \le \liminf_{n\to+\infty} f_n(y_n) \tag{5.15}$$

(ii)(Limsup Inequality) For every $y \in Y$ and there exists $\{y_n\} \subset Y$ such that $y_n \to y$,

$$f(y) \ge \limsup_{n\to+\infty} f_n(y_n) \tag{5.16}$$

Definition 5.2. *A sequence of function $f_n : Y \to \bar{\mathcal{R}}$ is equi-coercive if there exists a compact set K (independent of n) such that $\inf\{f_n(y) : y \in Y\} = \inf\{f_n(y) : y \in K\}$.*

It is easy to check (5.13) is equi-coercive and Γ-converges to the following problem

$$\begin{aligned} &\min_q D(q||\mu) \\ &\text{s.t. } D\left(q||\pi\right) \le \tau \end{aligned} \tag{5.17}$$

By choosing $\tau = C(\mu, \pi)$, this problem has optimal value $C(\mu, \pi)$, which means solutions of both (5.34) and (5.39) are lower bounded by $C(\mu, \pi)$ as $M \to \infty$. Since the optimal error exponent for the parametric model is $C(\mu, \pi)$, the error exponent for non-parametric model can not achieve better than this. Then we can conclude the exponent

$C(\mu, \pi)$ is optimal. □

The above corollary implies that GLRT for nonparametric model achieves the same risk decay exponent as the parametric model in the asymptotic regime of large number of sequences.

The next two theorems provide understanding for the case without any knowledge about distributions.

Theorem 5.3. *Consider the binary composite outlier detection problem* (4.1) *with a single outlier. Suppose neither π nor μ is known. For any test δ constructed without any knowledge about typical and outlier distributions, there must exist a pair μ and π such that δ is not exponentially consistent.*

Proof. The theorem follows from Theorem 4.4. □

Although there does not exist an exponentially consistent test for all pairs of μ and π without any knowledge about them, it is of interest to construct consistent test in such case.

Theorem 5.4. *Consider the binary composite outlier detection problem* (4.1) *with a single outlier. Suppose neither π nor μ is known. Further assume that no information about the distance between the distributions is known. Set the threshold τ to satisfy $\tau_n \to 0$ and*

$$\tau > \frac{M|\mathcal{Y}| \log(n+1)}{n},$$

where $|\mathcal{Y}|$ is the cardinality of the support set of π and μ. Then GLRT is universally consistent. Furthermore, the type II error is universally exponentially consistent.

Proof. The theorem follows as a special case of Theorem 5.7 where $t = 1$. □

Similarly to the semi-parametric model, a diminishing τ_n eventually falls into the desirable range of τ that can distinguish between the hypotheses.

5.2 Multi-outlier Model

In this subsection, we study the nonparametric model with t outliers, in which sequence i takes the distribution μ_i if it is an outlier for $i = 1, \ldots, M$. We assume that the number t of outliers is fixed and given. We further assume that neither π nor $\{\mu\} = \{\mu_i\}_{i=1}^M$ is known.

To construct GLRT, under H_0 with no outlier, we estimate π by the average of empirical distributions of all observed sequences

$$\tilde{\pi}_0 = \frac{\sum_{i=1}^M \gamma_i}{M} \tag{5.18}$$

and obtain the same likelihood function as the single-outlier case.

Under H_1 and the sub-hypothesis with sequences supported in S being outliers, we use the average of empirical distributions of all sequences except those supported by S to estimate π, and the empirical distribution of $y^{(i)}$ to estimate μ_i as follows.

$$\tilde{\pi}_{-S} = \frac{\sum_{j \notin S} \gamma_j}{M - t} \tag{5.19}$$

$$\tilde{\mu}_i = \gamma_i \text{ for } i \in S \tag{5.20}$$

Then the corresponding likelihood under H_1 is given by

$$\begin{aligned} \tilde{P}_S(y^{Mn}) &= L_S(y^{Mn}, \tilde{\pi}_{-S}, \{\gamma\}) \\ &= \prod_{k=1}^{n} \left(\prod_{j \in S} \gamma_j(y_k^{(j)}) \prod_{j \notin S} \tilde{\pi}_{-S}(y_k^{(j)}) \right) \\ &= \exp\left\{ -n \sum_{j=1}^{M} H(\gamma_j) - n \sum_{j \notin S} D(\gamma_j || \tilde{\pi}_{-S}) \right\} \end{aligned} \tag{5.21}$$

Thus, GLRT given by

$$\delta(y^{Mn}) : \frac{1}{n} \log \frac{\max_{S,|S|=t} \tilde{P}_S(y^{Mn})}{\tilde{P}_0(y^{Mn})} \underset{H_0}{\overset{H_1}{\gtrless}} \tau$$

can be further written as

$$\max_{S,|S|=t} \left[\sum_{j \in S} D(\gamma_j || \tilde{\pi}_0) + \sum_{j \notin S} (D(\gamma_j || \tilde{\pi}_0) - D(\gamma_j || \tilde{\pi}_{-S})) \right] \underset{H_0}{\overset{H_1}{\gtrless}} \tau. \tag{5.22}$$

In order to get insight about how to choose τ, we note that under H_0, $\tilde{\pi}_0$, $\tilde{\pi}_{-S}$ and γ_j all converges to π for $j = 1, 2, \ldots, M$ and $|S| = t$. Therefore, the test value converges to 0. Under H_1, $\tilde{\pi}_0$ and $\tilde{\pi}_{-S}$ does not deviate from each other too much if the outlier sequences is only a small portion of all sequences. Then the second term $\sum_{j \notin S} (D(\gamma_j || \tilde{\pi}_0) - D(\gamma_j || \tilde{\pi}_{-S}))$ is negligible small term if t/M is a small value. The first term converges to a value close to $D\left(\mu_S \middle\| \prod_{i=1}^{t} \left(\frac{1}{M} \sum_{j \in S} \mu_j + \frac{M-t}{M} \pi \right) \right)$ for large enough M when S is the true outlier index set.

Theorem 5.5. *Consider the binary composite outlier detection problem* (4.1) *with t outliers, where t is fixed and known. Suppose π is known but μ_j for $j = 1, 2, \ldots, M$ are unknown. Further assume that* $\min_{S \in \mathcal{M}, |S|=t} D\left(\mu_S \middle\| \prod_{i=1}^{t} \left(\frac{1}{M} \sum_{j \in S} \mu_j + \frac{M-t}{M} \pi \right) \right) := d$ *is known. Then the GLRT* (5.22) *with the threshold $\tau \in (0, d)$ is exponentially consistent, and achieves the risk decay exponent $R(\delta)$ given by*

$$E_R(\delta) = \min\{\alpha(\delta), \beta(\delta)\} \tag{5.23}$$

where

$$\alpha(\delta) = \min_{\{q_j\}_{j=1}^M} \sum_{j=1}^M D(q_j||\pi)$$
$$\textit{s.t.} \sum_{j=1}^M D\left(q_j \middle\| \frac{\sum_{i=1}^M q_i}{M}\right) - \sum_{j\notin S} D\left(q_j \middle\| \frac{\sum_{i\notin S} q_i}{M-t}\right) \geq \tau$$
$$\textit{for any } S \ s.t. \ |S| = t \tag{5.24}$$

$$\beta(\delta) = \min_{S:|S|=t} \min_{\{q_j\}_{j=1}^M} \sum_{j\in S} D(q_j||\mu_j) + \sum_{j\notin S} D(q_j||\pi)$$
$$\textit{s.t.} \sum_{j=1}^M D\left(q_j \middle\| \frac{\sum_{i=1}^M q_i}{M}\right) - \sum_{j\notin S'} D\left(q_j \middle\| \frac{\sum_{i\notin S'} q_i}{M-t}\right) \leq \tau$$
$$\textit{for all } S' \in \mathcal{M} \ s.t. \ |S'| = t \tag{5.25}$$

Proof. See Section 5.4. □

Using similar techniques in studying single outlier model, we can argue both $\alpha(\delta)$ and $\beta(\delta)$ are lower bounded by 0. Solution $(q_1^*, q_2^*, \ldots, q_M^*)$ of problem (5.24) satisfies

$$\begin{aligned}
\alpha(\delta) &= \sum_{j=1}^M D(q_j^*||\pi) \\
&\geq \sum_{j=1}^M D(q_j^*||\pi) - \sum_{j=1}^M D(q_j^*||\bar{q}^*) + \sum_{j\notin S} D(q_j^*||\bar{q}_{-S}^*) + \tau \\
&= MD(\bar{q}^*||\pi) + \sum_{j\notin S} D(q_j^*||\bar{q}_{-S}^*) + \tau \\
&\geq \tau
\end{aligned} \tag{5.26}$$

where $\bar{q}^* = \frac{\sum_{i=1}^M q_i^*}{M}$ and $\bar{q}_{-S}^* = \frac{\sum_{i\notin S} q_i^*}{M-t}$.

Also optimal value of problem (5.25) is lower bounded by

$$\begin{aligned}
\beta(\delta) \geq \min_{S} \min_{q_j,\, j=1,\ldots,M} & \sum_{j\in S} D(q_j||\mu_j) + \sum_{j\notin S} D(q_j||\pi) \\
& \text{s.t. } \sum_{j\in S} D\left(q_j \middle\| \frac{t}{M}\bar{q}_S + \frac{M-t}{M}\bar{q}_{-S}\right) \\
& \quad + (M-t)D\left(\bar{q}_{-S} \middle\| \frac{t}{M}\bar{q}_S + \frac{M-t}{M}\bar{q}_{-S}\right) \leq \tau \\
\geq \min_{S} \min_{q_j,\, j=1,\ldots,M} & D(q_S||\mu_S) + (M-t)D(\bar{q}_{-S}||\pi) \\
& \text{s.t. } D\left(q_S \middle\| \prod_{i=1}^{t}\left(\frac{t}{M}\bar{q}_S + \frac{M-t}{M}\bar{q}_{-S}\right)\right) \\
& \quad + (M-t)D\left(\bar{q}_{-S} \middle\| \frac{t}{M}\bar{q}_S + \frac{M-t}{M}\bar{q}_{-S}\right) \leq \tau \\
\geq \min_{S} \min_{q_j,\, j=1,\ldots,M} & D(q_S||\mu_S) + (M-t)D(\bar{q}_{-S}||\pi) \\
& \text{s.t. } D\left(q_S \middle\| \prod_{i=1}^{t}\left(\frac{t}{M}\bar{q}_S + \frac{M-t}{M}\bar{q}_{-S}\right)\right) \leq \tau
\end{aligned} \tag{5.27}$$

Given $\tau < D\left(\mu_S \middle\| \prod_{i=1}^{t}\left(\frac{1}{M}\sum_{j\in S}\mu_j + \frac{M-t}{M}\pi\right)\right)$, the solution must be positive. Combining the lower bounds for both $\alpha(\delta)$ and $\beta(\delta)$, GLRT is exponentially consistent.

The following Corollary characterizes the asymptotic exponent as M gets large.

Corollary 5.2. *As $M \to \infty$, the exponent of $R(\delta)$ converges to $\min_{S,|S|=t} C(\mu_S, \pi_t)$ if $\tau = \min_{S,|S|=t} C(\mu_S, \pi_t)$ and $t/M \to 0$, which is the optimal exponent that can be achieved for the nonparametric model with a single outlier over all tests.*

Proof. Since $\alpha(\delta) \geq \tau$, we have

$$\alpha(\delta) \geq \min_{S,\, |S|=t} C(\mu_S, \pi_t). \tag{5.28}$$

To analyze type II error exponent $\beta(\delta)$, we follow from its lower bound (5.27). Denote

the solution of (5.27) as $(q_S^*, \bar{q}_{-S}^*)$. It can be equally written as

$$\begin{aligned} &\min_{q_S} D(q_S||\mu_S) + (M-t)D(\bar{q}_{-S}^*||\pi) \\ &\text{s.t. } D\left(q_S \middle\| \frac{t}{M}\bar{q}_S + \frac{M-t}{M}\bar{q}_{-S}^*\right) \leq \tau \end{aligned} \tag{5.29}$$

We can see that $\bar{q}_{-S}^*(M) \to \pi$ as $M \to \infty$.

This is further lower bounded by

$$\begin{aligned} &\min_{S,|S|=t} \min_{q_S} D(q_S||\mu_S) \\ &\text{s.t. } D\left(q_S \middle\| \frac{t}{M}\bar{q}_S + \frac{M-t}{M}\bar{q}_{-S}^*(M)\right) \leq \tau \end{aligned} \tag{5.30}$$

Take a limit $M \to \infty$ and apply Theorem 5.2. this is equivalent to

$$\begin{aligned} &\min_{S} \min_{q_S} D(q_S||\mu_S) \\ &\text{s.t. } D(q_S||\pi_t) \leq \tau, \end{aligned} \tag{5.31}$$

which has solution $\min_{S,\,|S|=t} C(\mu_S, \pi_t)$ for $\tau = \min_{S,\,|S|=t} C(\mu_S, \pi_t)$. Both $\alpha(\delta)$ and $\beta(\delta)$ are lower bounded by $\min_{S,\,|S|=t} C(\mu_S, \pi_t)$. Since this is the optimal error exponent for the parametric model, this is the optimal for nonparametric model. □

In the above corollary, the condition $t/M \to 0$ guarantees that there are not too many outliers so that estimate of the typical distribution π can be accurate enough.

The following two theorems provide understanding for the case even without any knowledge about the distance between the distributions.

Theorem 5.6. *Consider the binary composite outlier detection problem* (4.1) *with* t *outliers, where* t *is fixed and known. Suppose neither* π *nor* μ_j *for* $j = 1, 2, \ldots, M$ *is known. For any test* δ *constructed without any knowledge about typcial and outlier distributions,*

there must exist $\{\mu\}$ *and* π *such that* δ *is not exponentially consistent.*

Proof. The theorem follows from Theorem 4.7. □

Theorem 5.7. . *Consider the binary composite outlier detection problem* (4.1) *with* t *outliers, where* t *is fixed and known. Suppose* π *is known but* μ_j *for* $j = 1, 2, \ldots, M$ *are unknown. Further assume that no information on distance between* π *and* $\{\mu\}$ *is known. Set the threshold* τ *in GLRT to satisfy* $\tau_n \to 0$ *and*

$$\tau > \frac{M|\mathcal{Y}| \log(n+1)}{n},$$

where $|\mathcal{Y}|$ *is the cardinality of the support set of* π *and* μ*. Then GLRT is universally consistent. Furthermore, the type II error is universally exponentially consistent.*

Proof. To analyze if type I error probability is universally consistent, we derive the upper bound of it.

$$\begin{aligned}
P_0(\delta = 1) &= P_0\left(\bigcup_{S,|S|=t}\left\{\left[\sum_{j=1}^{M} D(\gamma_j||\tilde{\pi}_0) - \sum_{j\notin S} D(\gamma_j||\tilde{\pi}_{-S})\right] \geq \tau\right\}\right) \\
&\leq P_0\left(\bigcup_{S,|S|=t}\left\{\sum_{j=1}^{M} D(\gamma_j||\tilde{\pi}_0) \geq \tau\right\}\right) \\
&\leq \binom{M}{t} P_0\left(\sum_{j=1}^{M} D(\gamma_j||\tilde{\pi}_0) \geq \tau\right) \\
&= \binom{M}{t} P_0\left(\sum_{j=1}^{M} D(\gamma_j||\pi) - MD(\tilde{\pi}_0||\pi) \geq \tau\right) \\
&\leq \binom{M}{t} P_0\left(\sum_{j=1}^{M} D(\gamma_j||\pi) \geq \tau\right) \\
&\leq \binom{M}{t} (n+1)^{M|\mathcal{Y}|} \exp(-n\tau) \\
&= \binom{M}{t} \exp\left(n\left(\frac{M|\mathcal{Y}| \log(n+1)}{n} - \tau\right)\right) \qquad (5.32)
\end{aligned}$$

Plug in $\tau_n > \frac{M|\mathcal{Y}| \log(n+1)}{n}$. $P_0(\delta = 1)$ converges to 0, which indicates the type I error

probability is universally consistent.

To analyze type II error exponent $\beta(\delta)$, we can investigate the lower bound (5.27). $\tau_n \to 0$ as $n \to \infty$, which makes $q_j = \bar{q}$ for $j = S$. Therefore $\bar{q}_{-S} = q_j$ for $j = S$, which makes solution of (5.27) positive. Therefore type II error probability is exponentially consistent. □

The above theorem implies that without knowing how much π is distinct from $\{\mu\}$, a diminishing τ_n helps to keep the type II error exponentially decaying to zero while keeping the type I error decaying to zero although not exponentially.

5.3 Proof of Exponentially Consistency for Single Outlier Model

We characterize the exponent of the risk function by analyzing the type I and type II errors using Sanov's Theorem. The type I error probability is given by

$$\begin{aligned} P_0(\delta = 1) =& P_0\left(\max_i \left[D(\gamma_i||\tilde{\pi}_0) + \sum_{j\neq i}\left(D(\gamma_j||\tilde{\pi}_0) - D(\gamma_j||\tilde{\pi}_i)\right)\right] \geq \tau\right) \\ =& P_0((\gamma_1, \gamma_2, \ldots, \gamma_M) \in E_\alpha^{Mn}), \end{aligned} \tag{5.33}$$

where E_α^{Mn} is given by

$$\begin{aligned} E_\alpha^{Mn} = \{(q_1, \ldots, q_M) :\ & D(q_1||\bar{q}) + \sum_{j\neq 1}\left(D(q_j||\bar{q}) - D(q_j||\bar{q}{-}1)\right) \geq \tau \\ & \text{or } D(q_2||\bar{q}) + \sum_{j\neq 2}\left(D(q_j||\bar{q}) - D(q_j||\bar{q}_{-2})\right) \geq \tau \\ & \ldots\ldots \\ & \text{or } D(q_M||\bar{q}) + \sum_{j\neq M}\left(D(q_j||\bar{q}) - D(q_j||\bar{q}_{-M})\right) \geq \tau\}. \end{aligned}$$

To calculate the corresponding error exponent $\alpha(\delta)$, we apply Sanov's Theorem and obtain

$$\alpha(\delta) = \min_{i=1,\ldots,M} \alpha_i(\delta), \tag{5.34}$$

where

$$\begin{aligned} \alpha_i(\delta) = \min_{q_j,\, j=1,\ldots,M} & D(q_1||\pi) + D(q_2||\pi) + \ldots + D(q_M||\pi) \\ \text{s.t. } & D(q_i||\bar{q}) + \sum_{j\neq i} (D(q_j||\bar{q}) - D(q_j||\bar{q}_{-i})) \geq \tau \end{aligned} \tag{5.35}$$

and $\bar{q} = \sum_{j=1}^{M} q_j/M$, $\bar{q}_{-i} = \sum_{j\neq i} q_j/(M-1)$ for $i = 1, 2, \ldots, M$. We note that the optimization problems in (5.34) have identical solutions for different $i = 1, 2, \ldots, M$. Therefore, we simplify (5.34) as

$$\begin{aligned} \alpha(\delta) = \min_{q_j,\, j=1,\ldots,M} & D(q_1||\pi) + D(q_2||\pi) + \ldots + D(q_M||\pi) \\ \text{s.t. } & D(q_1||\bar{q}) + \sum_{j\neq 1} (D(q_j||\bar{q}) - D(q_j||\bar{q}_{-1})) \geq \tau. \end{aligned} \tag{5.36}$$

We next analyze the type II error by exploring all underlying distributions. Given the i-th sequence is the true outlier, the error probability is given by

$$\begin{aligned} P_i(\delta = 0) &= P_i \left(\max_i \left[D(\gamma_i||\tilde{\pi}_0) + \sum_{j\neq i} \left(D(\gamma_j||\tilde{\pi}_0) - D(\gamma_j||\tilde{\pi}_i) \right) \right] \leq \tau \right) \\ &= P_i \left((\gamma_1, \gamma_2, \ldots, \gamma_M) \in E_\beta^{Mn} \right), \end{aligned} \tag{5.37}$$

where E_β^{Mn} is given by

$$\begin{aligned} E_\beta^{Mn} = \{(q_1, \ldots, q_M) : D(q_1||\bar{q}) + \sum_{j\neq 1} (D(q_j||\bar{q}) - D(q_j||\bar{q}_{-1})) \leq \tau \\ \text{and } D(q_2||\bar{q}) + \sum_{j\neq 2} (D(q_j||\bar{q}) - D(q_j||\bar{q}_{-2})) \leq \tau \\ ..... \\ \text{and } D(q_M||\bar{q}) + \sum_{j\neq M} (D(q_j||\bar{q}) - D(q_j||\bar{q}_{-M})) \leq \tau\}. \end{aligned}$$

Applying Sanov's Theorem to the above equation, we obtain the exponent $\beta(\delta)$ as follows.

$$\beta(\delta) = \min_{i=1,\ldots,M} \beta_i(\delta) \tag{5.38}$$

where

$$\begin{aligned} \beta_i(\delta) = \min_{q_j,\, j=1,\ldots,M} & D(q_i||\mu) + \sum_{j\neq i} D(q_j||\pi) \\ \text{s.t. } & D(q_k||\bar{q}) + \sum_{j\neq k} (D(q_j||\bar{q}) - D(q_j||\bar{q}_{-k})) \geq \tau \text{ for } k = 1, 2, \ldots, M. \end{aligned} \tag{5.39}$$

Due to the symmetry of this problem over $i = 1, 2, \ldots, M$, the exponent is given by

$$\begin{aligned} \beta(\delta) = \min_{q_j,\, j=1,\ldots,M} & D(q_1||\mu) + \sum_{j\neq i} D(q_j||\pi) \\ \text{s.t. } & D(q_k||\bar{q}) + \sum_{j\neq k} (D(q_j||\bar{q}) - D(q_j||\bar{q}_{-k})) \geq \tau \text{ for } k = 1, 2, \ldots, M. \end{aligned} \tag{5.40}$$

5.4 Proof of Exponentially Consistency for Muli-outlier Model

We characterize the exponent of the risk function by analyzing the type I and type II errors using Sanov's Theorem. Type I error probability is given by

$$P_0(\delta = 1) = P_0\left(\max_{S,|S|=t}\left[\sum_{j\in S} D(\gamma_j||\tilde{\pi}_0) + \sum_{j\notin S}(D(\gamma_j||\tilde{\pi}_0) - D(\gamma_j||\tilde{\pi}_{-S}))\right] \geq \tau\right)$$
$$= P_0\left((\gamma_1, \gamma_2, \ldots, \gamma_M) \in E_\alpha^{Mn}\right), \tag{5.41}$$

where E_α^{Mn} is given by

$$E_\alpha^{Mn} = \Big\{(q_1, \ldots, q_M) : \sum_{j\in S} D(q_j||\bar{q}) + \sum_{j\notin S}(D(q_j||\bar{q}) - D(q_j||\bar{q}_{-S})) \geq \tau$$
$$\text{for at least one } S \in \mathcal{M}\Big\}. \tag{5.42}$$

To calculate the corresponding exponent $\alpha(\delta)$, we apply Sanov's Theorem and obtain $\alpha(\delta)$ given by the solution of the following problem.

$$\min_{S,|S|=t}\ \min_{q_j,\ j=1,\ldots,M} D(q_1||\pi) + D(q_2||\pi) + \ldots + D(q_M||\pi)$$
$$\text{s.t. } \sum_{j\in S} D(q_j||\bar{q}) + \sum_{j\notin S}(D(q_j||\bar{q}) - D(q_j||\bar{q}_{-S})) \geq \tau. \tag{5.43}$$

To analyze the type II error exponent, we first assume that S is the set containing the true outlier indexes, and obtain the following type II error probability

$$P_S(\delta = 0) = P_S\left(\max_{S,|S|=t}\left[\sum_{j\in S} D(\gamma_j||\tilde{\pi}_0) + \sum_{j\notin S}(D(\gamma_j||\tilde{\pi}_0) - D(\gamma_j||\tilde{\pi}_{-S}))\right] \leq \tau\right)$$
$$= P_S\left((\gamma_1, \gamma_2, \ldots, \gamma_M) \in E_\beta^{Mn}\right), \tag{5.44}$$

where E_β^{Mn} is given by

$$E_\beta^{Mn} = \left\{ (q_1, \ldots, q_M) : \sum_{j \in S} D(q_j||\bar{q})) + \sum_{j \notin S} (D(q_j||\bar{q}) - D(q_j||\bar{q}_{-S})) \leq \tau \text{ for all } S \in \mathcal{M} \right\}. \tag{5.45}$$

We then apply Sanov's Theorem to the above equation and obtain the corresponding exponent $\beta(\delta)$ given by the solution of the following problem.

$$\begin{aligned} \min_{S,|S|=t} \min_{q_j,\, j=1,\ldots,M} & \sum_{j \in S} D(q_j||\mu_j) + \sum_{j \notin S} D(q_j||\pi) \\ \text{s.t. } & \sum_{j \in S'} D(q_j||\bar{q}) + \sum_{j \notin S'} (D(q_j||\bar{q}) - D(q_j||\bar{q}_{-S'})) \leq \tau \text{ for all } |S'| = t. \end{aligned} \tag{5.46}$$

CHAPTER 6
CONCLUSIONS AND SUGGESTIONS FOR FUTURE WORK

In this section, we first summarize the results presented in this thesis, and then describe a few future research directions.

6.1 Concluding Remarks

Although signal detection has been intensively studied for decades, semi-parametric and nonparametric detection models are still not well-understood yet. In this thesis, we addressed two categories of signal detection problems: a) nonparametric decentralized detection in Chapter 2 and 3, and b) semi-parametric/nonparametric composite outlier detection in centralized setting in Chapter 4 and 5. We summarize the main contributions reported in this thesis as follows.

In Chapters 2 and 3, we generalized the kernel-based nonparametric decentralized detection framework [8] proposed by Nguyen, Wainwright, and Jordan, and introduced the idea of using weighted kernel 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 then designed gradient projection-based algorithm and Gauss-Seidel algorithm to solve the joint optimization of weight parameters and sensors' and fusion center's decision rules, and showed that both algorithms converge to critical points. We also demonstrated the performance of our approach via numerical experiments.

In Chapters 4 and 5, we studied the composite outlier hypothesis testing problem under both the semi-parametric and nonparametric models. For both models, we constructed GLRT, and have shown that with the knowledge of the KL divergence between the outlier and typical distributions, GLRT is exponentially consistent. We also showed that with the knowledge of the Chernoff distance between the outlier and typical distributions, GLRT for semi-parametric model achieves the same risk decay exponent as the parametric model, and GLRT for nonparametric model achieves the same performance as $M \to \infty$. We further showed that for both models without any knowledge about the distance between distributions, there does not exist an exponentially consistent test. However, GLRT with a diminishing threshold can still be consistent.

6.2 Directions for Future Research

We conclude this thesis by pointing out some directions for future research.

As generalization of the nonparametric decentralized detection problem, nonparametric multi-level sensor networks is of great interest to explore. In multi-level sensor networks, the uppermost level of the sensors receive observations of the event. All other sensors at the lower levels receive quantized outputs from its upper level and then quantize them into single variables. It is expected that sensor selection and regularization term are also related to network structures in this case. Group regularization on sensors' weight parameters

should be designed based on the network structures. It is vital to study the impacts of different levels of sensors on the fusion center's decision rule. Since the computation of the decision rules over the multi-level sensor network is complex, it is also important to develop effective computation techniques.

Our work on composite outlier detection model assumes that each data sequence consists of i.i.d. samples. It is interesting to explore the scenarios in which data samples are correlated, for example, following Markov distributions. By constructing the likelihood of Markov observations from the their empirical distributions, we can apply GLRT to investigate semi-parametric and nonparametric Markov models. It is important to compare the asymptotic error performance and required constraints of Markov models to the independent observation models. The effect of the order of Markov distributions (the number of previous states that current state depends on) on the error performance is another interesting problem to study.

It is also interesting to study the composite outlier detection model in the online setting, in which data arrive in real time. Then, instead of performing test after all data are collected, sequential hypothesis testing rules are more desirable, which continuously test hypotheses as samples come and can decide to terminate the process if a decision can be made to meet the required performance constraints. It is of importance to analyze the asymptotic error performance of such sequential tests, and further compare the expected delay with the number of samples used in non-sequential tests to achieve the same error performance.

REFERENCES

[1] S. A. Kassam, "Nonparametric signal detection," *in Advances in Signal Processing, edited by H. V. Poor and J. B. Thomas, JAI Press*, vol. 2, pp. 66–91, 1993.

[2] M. M. AI-Ibrahim and P. K. Varshney, "Nonparametric sequential detection based on multisensor data," in *Proc. 23rd Annu. Conf. Information Science Systems*, Mar. 1989, pp. 157–162.

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

[4] J. N. Tsitsiklis, "Decentralized detection," *in Advances in Signal Processing, edited by H. V. Poor and J. B. Thomas, JAI Press*, vol. 2, pp. 297–344, 1993.

[5] 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.

[6] V. V. Veeravalli and P. K. Varshney, "Distributed inference in wireless sensor networks," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 370, no. 1958, pp. 100–117, Jan. 2012.

[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 Processing Magazine*, vol. 23, no. 4, pp. 56–69, Jul. 2006.

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

[9] J. Hu, Y. Liang, and E. P. Xing, "Nonparametric decision making based on tree-structured information aggregation," *Proc. Annu. Allerton Conf. Communication, Control and Computing*, Sep. 2011.

[10] 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 *Proceedings of the 1st ACM International Workshop on Quality of Service and Security in Wireless and Mobile Networks*, 2005, pp. 16–23.

[11] N. A. Goodman and D. Bruyere, "Optimum and decentralized detection for multistatic airborne radar," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 43, no. 2, pp. 806–813, 2007.

[12] F. Lin, M. Fardad, and M. R. Jovanovic, "Algorithms for leader selection in stochastically forced consensus networks," *IEEE Transactions on Automatic Control*, vol. 59, no. 7, pp. 1789–1802, 2014.

[13] E. Masazade, R. Niu, and P. K. Varshney, "An approximate dynamic programming based non-myopic sensor selection method for target tracking," in *46th Annual Conference on Information Sciences and Systems (CISS)*. IEEE, 2012, pp. 1–6.

[14] S. Liu, S. P. Chepuri, M. Fardad, E. Masazade, G. Leus, and P. K. Varshney, "Sensor selection for estimation with correlated measurement noise," *arXiv preprint arXiv:1508.03690*, 2015.

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

[16] W. P. Tay, J. N. Tsitsiklis, and M. Z. Win, "Asymptotic performace of a censoring sensor network," *IEEE Transactions on Information Theory*, vol. 53, no. 11, pp. 4191–4209, 2007.

[17] S. Joshi and S. Boyd, "Sensor selection via convex optimization," *IEEE Transactions on Signal Processing*, vol. 57, no. 2, pp. 451–462, 2009.

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

[19] L. Zuo, R. Niu, and P. K. Varshney, "A sensor selection approach for target tracking in sensor networks with quantized measurements," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, Mar. 2008, pp. 2521–2524.

[20] H. Wang, K. Yao, G. Pottie, and D. Estrin, "Entropy-based sensor selection heuristic for target localization," in *Proceedings of the 3rd International Symposium on Information Processing in Sensor Networks*, 2004, pp. 36–45.

[21] 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.

[22] 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.

[23] 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.

[24] S. P. Chepuri and G. Leus, "Sparsity-promoting sensor selection for non-linear measurement models," *IEEE Transactions on Signal Processing*, vol. 63, no. 3, pp. 684–698, 2015.

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

[26] S. Chen, D. Donaho, and M. Saunders, "Atomic decomposition by basis persuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.

[27] D. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Transactions on Information Theory*, vol. 47, no. 7, pp. 2845–2862, 2001.

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

[29] J. Unnikrishnan, "On optimal two sample homogeneity tests for finite alphabets," in *isit*. Ieee, 2012, pp. 2027–2031.

[30] J. Ziv, "On classification with empirically observed statistics and universal data compression," *IEEE Transactions on Information Theory*, vol. 34, no. 2, pp. 278–286, 1988.

[31] K. Pearson, "On the probability that two independent distributions of frequency are really samples from the same population," *Biometrika*, vol. 8, no. 1/2, pp. 250–254, 1911.

[32] O. Shayevitz, "On rényi measures and hypothesis testing." in *isit*, 2011, pp. 894–898.

[33] M. Feder and A. Lapidoth, "Universal decoding for channels with memory," *IEEE Transactions on Information Theory*, vol. 44, no. 5, pp. 1726–1745, 1998.

[34] A. Patcha and J.-M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," *Computer networks*, vol. 51, no. 12, pp. 3448–3470, 2007.

[35] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.

[36] L. Lai, H. V. Poor, Y. Xin, and G. Georgiadis, "Quickest search over multiple sequences," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5375–5386, 2011.

[37] Y. Li, S. Nitinawarat, and V. V. Veeravalli, "Universal outlier hypothesis testing," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 4066–4082, 2014.

[38] S. Zou, Y. Liang, H. V. Poor, and X. Shi, "Nonparametric detection of anomalous data via kernel mean embedding," *arXiv preprint arXiv:1405.2294*, 2014.

[39] E. L. Lehmann and C. Stein, "Most powerful tests of composite hypotheses. i. normal distributions," *The Annals of Mathematical Statistics*, pp. 495–516, 1948.

[40] Y.-W. Huang and P. Moulin, "Strong large deviations for composite hypothesis testing," in *IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2014, pp. 556–560.

[41] O. Zeitouni, J. Ziv, and N. Merhav, "When is the generalized likelihood ratio test optimal?" *IEEE Transactions on Information Theory*, vol. 38, no. 5, pp. 1597–1602, 1992.

[42] W. Wang, Y. Liang, and E. P. Xing, "Collective support recovery for multi-design multi-response linear regression," *IEEE Transactions on Information Theory*, vol. 61, no. 1, pp. 513–534, 2015.

[43] W. Wang, Y. Liang, E. P. Xing, and L. Shen, "Nonparametric decentralized detection and sparse sensor selection via weighted kernel," *IEEE Transactions on Signal Processing*, vol. 64, no. 2, pp. 306–321, 2016.

[44] W. Wang, Y. Liang, and H. V. Poor, "Nonparametric composite outlier detection," *in preparation*.

[45] W. Wang, Y. Liang, and E. Xing, "Block regularized lasso for multivariate multi-response linear regression," in *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, 2013, pp. 608–617.

[46] W. Wang, Y. Liang, E. P. Xing, and L. Shen, "Sparse sensor selection for nonparametric decentralized detection via l_1 regularization," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2014, pp. 1–6.

[47] W. Wang, Y. Liang, and H. V. Poor, "Nonparametric composite outlier detection," *to appear in Proc. Asilomar Conf. Signals, Systems and Computers*, 2016.

[48] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of Lasso and dantzig selector," *Annals of Statistics*, vol. 37, no. 4, pp. 1705–1732, 2009.

[49] E. Candes and T. Tao, "The dantzig selector: Statistical estimation when p is much larger than n," *Annals of Statistics*, vol. 35, no. 6, pp. 2313–2351, 2007.

[50] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with sparsity-inducing penalties," *Foundations and Trends in Machine Learning*, vol. 4, no. 1, pp. 1–106, Now Publishers, Hanover, MA, USA, 2011.

[51] M. Elad and A. M. Bruckstein, "A generalized uncertainty principle and sparse representation in pairs of bases," *IEEE Transactions on Information Theory*, vol. 48, no. 9, pp. 2558–2567, 2002.

[52] A. Feuer and A. Nemirovski, "On sparse representation in pairs of bases," *IEEE Transactions on Information Theory*, vol. 49, no. 6, pp. 1579–1581, 2003.

[53] D. M. Malioutov, M. Cetin, and A. S. Willsky, "Optimal sparse representations in general overcomplete bases," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Montreal, Canada, 2004.

[54] J. Tropp, "Greedy is good: Algorithm resutlts for sparse approximation," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[55] J. J. Fuchs, "Recovery of exact sparse representations in the presence of noise," *IEEE Transactions on Information Theory*, vol. 51, no. 10, pp. 3601–3608, 2005.

[56] P. Zhao and B. Yu, "On model selection consistency of Lasso," *Journal of Machine Learning Research*, vol. 7, pp. 2541–2567, 2006.

[57] N. Meinshausen and B. Yu, "Lasso-type recovery of sparse representations for high-dimensional data," *Annals of Statistics*, vol. 37, no. 1, pp. 246–270, 2009.

[58] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using l_1-constrained quadratic programming (Lasso)," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2183–2202, 2009.

[59] G. Raskutti, M. J. Wainwright, and B. Yu, "Restricted eigenvalue properties for correlated Gaussian designs," *Journal of Machine Learning Research*, vol. 11, pp. 2241–2259, 2010.

[60] F. Bach, "Consistency of trace norm minimization," *Journal of Machine Learning Research*, vol. 9, pp. 1019–1048, 2008.

[61] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused Lasso," *Journal of the Royal Stastistical Society: Series B*, vol. 67, pp. 91–108, 2005.

[62] Y. Chen and A. Dalalyan, "Fused sparsity and robust estimation for linear models with unknown variance," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 25, 2012, pp. 1268–1276.

[63] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B*, vol. 68, pp. 49–67, 2006.

[64] J. Huang and T. Zhang, "The benefit of group sparsity," *Annals of Statistics*, vol. 38, no. 4, pp. 1978–2004, 2010.

[65] L. Jacob, G. Obozinski, and J.-P. Vert, "Group Lasso with overlaps and graph Lasso," in *International Conference on Machine Learning (ICML)*, 2009.

[66] P. Zhao, G. Rocha, and B. Yu, "The composite absolute penalties family for grouped and hierarchical vairable selection," *Annals of Statistcs*, vol. 37, no. 6A, pp. 3468–3497, 2009.

[67] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," *Statistical Science*, vol. 27, no. 4, pp. 538–557, 2012.

[68] G. Obozinski, M. J. Wainwright, and M. I. Jordan, "Support union recovery in high-dimensional multivariate regression," *Annals of Statistics*, vol. 39, no. 1, pp. 1–47, 2011.

[69] E. van den Berg and M. P. Friedlander, "Theoretical and empirical results for recovery from multiple measurements," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2516–2527, 2010.

[70] H. Liu and J. Zhang, "On the $l_1 - l_q$ regularized regression," *arXiv:0802.1517v1*, 2008.

[71] B. A. Turlach, W. N. Venables, and S. J. Wright, "Simultaneous variable selection," *Technometrics*, vol. 47, no. 3, pp. 349–363, 2005.

[72] J. Tropp, "Algorithms for simultaneous sparse approximation. Part II: Convex relaxation," *Signal Processing*, vol. 86, no. 3, pp. 589–602.

[73] G. Obozinski, B. Tarskar, and M. I. Jordan, "Joint covariate selection and joint subspace selection for multiple classification problems," *Statistics and Computing*, vol. 20, no. 2, pp. 231–252, 2010.

[74] M. Kolar, J. Lafferty, and L. Wasserman, "Union support recovery in multi-task learning," *Journal of Machine Learning Research*, vol. 12, pp. 2415–2435, 2011.

[75] K. Lounici, M. Pontil, S. Geer, and A. B. Tsybakov, "Oracle inequalities and optimal inference under group sparsity," *Annals of Statistics*, vol. 39, pp. 2164–2204, 2011.

[76] R. Heckel and H. Bölcskei, "Joint sparsity with different measurement matrices," in *50th Annaual Allerton Conference on Communication, Control and Computing*, Monticello, IL USA, 2012.

[77] S. Negahban and M. J. Wainwright, "Simultaneous support recovery in high dimensions: Benefits and perils of block ℓ_1/ℓ_∞-regularization," *IEEE Transactions on Information Theory*, vol. 57, no. 6, pp. 3841–3863, 2011.

[78] A. Jalali, P. Ravikumara, S. Sanghavi, and C. Ruan., "A dirty model for multi-task learning," in *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[79] B. Scholkopf and A. Smola, *Learning with Kernels*. MA: MIT press, 2002.

[80] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, pp. 463–482, Nov. 2002.

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

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

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

[84] 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," *Mathematical Programming*, vol. 137, pp. 91–129, 2013.

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

[86] 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 $o(1/n^2)$," *arXiv:1502.04286v2*, 2015.

[87] S. G. Krantz and H. R. Parks, *A Primer of Real Analytic Functions*. New York, NY, USA: Birkh$\ddot{a}$user Boston, Springer-Verlag New York, Inc., 2002.

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

[89] A. Parusiński, "Subanalytic functions," *Transaction of the American Mathematical Society*, vol. 344, no. 2, pp. 583–595, 1994.

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

[91] E. Bierstone and P. D. Milman, "Semianalytic and subanalytic sets," *Publications Mathématiques de l'IHÃL'S*, vol. 67, no. 1, pp. 5–42, 1988.

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

[93] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

VITA

NAME OF AUTHOR: Weiguang Wang

PLACE OF BIRTH: Weihai, Shandong, China

DATE OF BIRTH: Aug. 10, 1988

UNDERGRADUATE SCHOOLS ATTENDED:

University of Science and Technology of China, Hefei, China

DEGREES AWARDED:

B.E., 2011, University of Science and Technology of China, Hefei,, China