December 2018

On Classification in Human-driven and Data-driven Systems

Qunwei Li
Syracuse University

ABSTRACT

Classification systems are ubiquitous, and the design of effective classification algorithms has been an even more active area of research since the emergence of machine learning techniques. Despite the significant efforts devoted to training and feature selection in classification systems, misclassifications do occur and their effects can be critical in various applications. The central goal of this thesis is to analyze classification problems in human-driven and data-driven systems, with potentially unreliable components and design effective strategies to ensure reliable and effective classification algorithms in such systems. The components/agents in the system can be machines and/or humans. The system components can be unreliable due to a variety of reasons such as faulty machines, security attacks causing machines to send falsified information, unskilled human workers sending imperfect information, or human workers providing random responses. This thesis first quantifies the effect of such unreliable agents on the classification performance of the systems and then designs schemes that mitigate misclassifications and their effects by adapting the behavior of the classifier on samples from machines and/or humans and ensure an effective and reliable overall classification.

In the first part of this thesis, we study the case when only humans are present in the systems, and consider crowdsourcing systems. Human workers in crowdsourcing systems observe the data and respond individually by providing label related information to a fusion center in a distributed manner. In such systems, we consider the presence of unskilled human workers where they have a reject option so that they may choose not to provide information regarding the label of the data. To maximize the classification performance at the fusion center, an optimal aggregation rule is proposed to fuse the human workers' responses in a weighted majority voting manner. Next, the presence of unreliable human workers, referred to as spammers, is considered. Spammers are human workers that provide random guesses regarding the data label information to the fusion center in crowdsourcing systems. The effect of spammers on the overall classification perfor-

mance is characterized when the spammers can strategically respond to maximize their reward in reward-based crowdsourcing systems. For such systems, an optimal aggregation rule is proposed by adapting the classifier based on the responses from the workers.

The next line of human-driven classification is considered in the context of social networks. The classification problem is studied to classify a human whether he/she is influential or not in propagating information in social networks. Since the knowledge of social network structures is not always available, the influential agent classification problem without knowing the social network structure is studied. A multi-task low rank linear influence model is proposed to exploit the relationships between different information topics. The proposed approach can simultaneously predict the volume of information diffusion for each topic and automatically classify the influential nodes for each topic.

In the third part of the thesis, a data-driven decentralized classification framework is developed where machines interact with each other to perform complex classification tasks. However, the machines in the system can be unreliable due to a variety of reasons such as noise, faults and attacks. Providing erroneous updates leads the classification process in a wrong direction, and degrades the performance of decentralized classification algorithms. First, the effect of erroneous updates on the convergence of the classification algorithm is analyzed, and it is shown that the algorithm linearly converges to a neighborhood of the optimal classification solution. Next, guidelines are provided for network design to achieve faster convergence. Finally, to mitigate the impact of unreliable machines, a robust variant of ADMM is proposed, and its resilience to unreliable machines is shown with an exact convergence to the optimal classification result.

The final part of research in this thesis considers machine-only data-driven classification problems. First, the fundamentals of classification are studied in an information theoretic framework. We investigate the nonparametric classification problem for arbitrary unknown composite distributions in the asymptotic regime where both the sample size and the number of classes grow exponentially large. The notion of discrimination capacity is introduced, which captures the largest exponential growth rate of the number of classes relative to the samples size so that there exists a

test with asymptotically vanishing probability of error. Error exponent analysis using the maximum mean discrepancy is provided and the discrimination rate, i.e., lower bound on the discrimination capacity is characterized. Furthermore, an upper bound on the discrimination capacity based on Fano's inequality is developed.

To my family:

past, present, and future.

ON CLASSIFICATION IN HUMAN-DRIVEN AND DATA-DRIVEN SYSTEMS

By

Qunwei Li
B.S., Xidian University, Xi'an, China, 2011
M.S., Xidian University, Xi'an, China, 2014

DISSERTATION

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

Syracuse University
December 2018

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor Prof. Pramod K. Varshney for his invaluable guidance throughout this dissertation, and support during my PhD study. Since my first interaction with him through phone call before I joined Syracuse University in 2013, I have always been comfortable speaking my mind around him. His patience and belief has pushed me to work hard and encouraged me to pursue good research. Ever since he accepted me into his research group, he has been a great mentor with his continuous support and encouragement. I am also grateful to the professors that I have worked with, Prof. Yingbin Liang, Prof. Lav R. Varshney, Prof. Biao Chen, for the opportunity to learn from their deep insights and in-depth knowledge.

My PhD research has been supported by numerous grants from AFOSR, ARO, ARL, and Syracuse University Graduate Fellowship, for which I am grateful. In addition, I would like to thank my defense committee members Prof. Biao Chen, Prof. Yingbin Liang, Prof. Makan Fardad, Prof. Garrett Katz, and Prof. Lixin Shen.

I have enjoyed the company and support of my lab mates Adytia, Bhavya, Sid, Arun, Raghed, Hao, Sijia, Nianxia, Prashant, Shan, Swatantra, Swastik, Sora, Thakshila, Sai, and Baocheng. They have listened to my boring presentations with keen interest, and have always been a constant source of inspiration to carry out my research, for which I am grateful to them. They made sure that time spent here would remain as one of the best times of my life. Our fun times would be missed! Special thanks to Mrs. Anju Varshney who made me feel at home during my stay in Syracuse.

In the process of obtaining higher education, I have deprived my family an opportu-

nity to spend time with me. I too have missed their company sorely over all these years. Their unconditional and extraordinary support, care, selfless sacrifice, and encouragement during this time, which they have bestowed at a great personal cost to them, cannot be emulated. I am truly and deeply thankful to my parents Youfen and Fuchang.

Table of Contents

List of Tables

List of Figures

CHAPTER 1
INTRODUCTION

The classification problem is one of the most important problems in statistical signal processing and learning theory. It has received much attention in different fields for many applications since it was formulated in the late 1950s. Various classification algorithms have been designed and some theoretical studies have been conducted to find the fundamental performance limits. In particular, significant progress has been made within the last decade: new powerful algorithms such as logistic regression and support vector machines were invented, and the idea of regularization was introduced in the learning theory framework. The main goal of this thesis is to make novel contributions to the analysis of the classification problem in learning theory in both human-driven and data-driven systems, with potentially unreliable components and design of effective strategies to ensure reliable and effective classification algorithms in such systems.

1.1 General Architecture

The general notional diagram of the problems considered in this thesis, where an unknown phenomenon is observed by multiple agents, both humans and machines, is presented in Fig. 1.1. Depending on the problem, the agents may or may not communicate with each other to reach a classification decision. The phenomenon is classified using the observations by the agents. How-

ever, some of the agents can be unreliable while some others may provide false information due to malicious intent. For example, some honest humans do not intentionally provide false information but due to their lack of specific expertise and unclear observation (for example, under foggy conditions the human may not see clearly), honest humans end up providing unreliable information. Moreover, there can be dishonest humans providing false information. On the other hand, some machines can also send erroneous information for classification, due to various reasons like being hacked, transmission noise, and computation failure.

As is shown in Fig. 1.1, the classification system is human-driven if only human agents are involved, and is data-driven if only machine agents are involved. While we do not consider it in this thesis, systems may involve both human-driven and data-driven components.

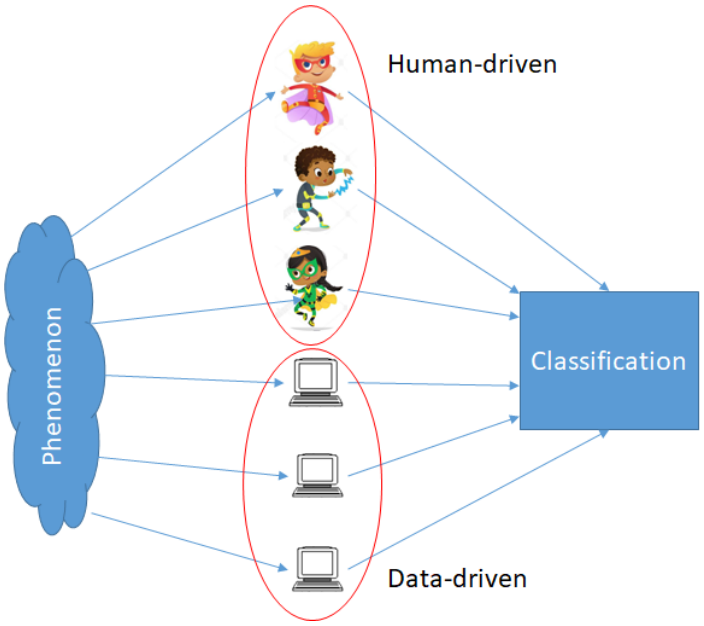


Fig. 1.1: Human-driven and data-driven classification with potentially unreliable agents.

1.1.1 Classification

In classification, based on the observations of the phenomenon, one needs to infer the class from a set of finite cardinality, say M. This can be represented by a hypothesis testing problem with M possible hypotheses: $\mathcal{H}_0, \ldots, \mathcal{H}_{M-1}$. Agents collect multiple observations $\mathbf{x} = [x_1, \ldots, x_N]$

regarding the phenomenon. If the probability density function $p(\mathbf{x}|\mathcal{H}_i)$ for each hypothesis $\mathcal{H}_i$, $i = 0, \ldots, M-1$ is available, the phenomenon is classified using the M-ary maximum a posteriori (MAP) decision rule as k-th class if

$$p(\mathcal{H}_k|\mathbf{x}) > p(\mathcal{H}_i|\mathbf{x}), i \neq k.$$

If the number of hypotheses $M = 2$, it is referred to as the binary classification problem, or the detection problem.

If the probability density functions $p(\mathbf{x}|\mathcal{H}_i), i = 0, \ldots, M-1$ are not available, the generic classification framework is

$$k = \arg\max_i \phi(i|\mathbf{x}),$$

where the function $\phi(i|\mathbf{x})$ measures the possibility of the hypothesis $\mathcal{H}_i$ given the observation $\mathbf{x}$. The hypothesis $\mathcal{H}_k$ is determined to be true when $\phi(k|\mathbf{x})$ has the maximal value compared with other hypotheses.

This thesis focuses on classification problems, when $p(\mathbf{x}|\mathcal{H}_i), i = 0, \ldots, M-1$ are not known, in different scenarios with human-driven or data-driven systems where the goal is to classify the unknown phenomenon using multiple observations.

1.1.2 Agents

Typically, the observations regarding the phenomenon are made in a distributed fashion at the agents. The agents observe the phenomenon and send their local observations (possibly after local processing) to other agents within communication range, and to a fusion center over (possibly) imperfect channels. The fusion center then fuses the data from the agents to infer about the phenomenon for classification. In such a framework for classification, there are two problems to be considered:

- How does one design the local signal processing schemes at the agents?

- How does one design the signal processing schemes for aggregation at the fusion center?

such that the classification is as accurate as possible. There has been significant work on these two problems with specific consideration of network topology, decision rules, effect of communication channels, effect of spatio-temporal dependence, etc (see the survey paper [116] and the references therein). An important application for such a classification framework is wireless sensor networks (WSNs), which renewed the general interest in distributed classification with new interesting research challenges: correlated observations [14, 24, 52, 106], wireless channels [15, 17], sensor resource management [2, 3, 54, 91], etc.

Depending on the types of the agents (humans or machines), the classification systems are of two types: human-driven systems with human workers providing subjective observations, and data-driven systems with physical machines giving objective measurements. This thesis focuses on (distributed) classification problems with measurements from agents of both types, namely either humans or machines.

1.1.3 Unreliability

The agents in both human-driven and data-driven systems for classification are not necessarily reliable. In human-driven systems, unreliability could be due to a lack of expertise of the human worker when performing the classification task. Unintentionally, the observation can be very unclear due to prevailing conditions such as occlusion or fog, which can cause a degradation in classification performance. Additionally, the unreliableness could also result from the presence of spammers in the network, who typically provide random responses to the fusion center. For example, in crowdsourcing systems for classification tasks, the human workers are typically anonymous and the presence of spammers is justified in various applications. In data-driven systems with machines, the agents could be unreliable due to various reasons. First, the agents could receive noisy local observations. Second, the agents could have permanent errors such as in the cases where the agents are stuck due to local computation failures, which results in agents always sending erroneous information to the fusion center. Besides, the physical machines could be attacked

by external adversaries. The attacker can reprogram the machines and consequently the machines would always send erroneous information to other machines and to the fusion center. The unreliability issue of the agents causes a degradation in classification.

This thesis focuses on classification problems with information from potentially unreliable agents of humans or machines.

1.2 Classification Problems in Human-driven Systems

In the broad area of human-driven systems, this thesis considers classification problems in two typical scenarios. One is crowdsourcing systems and the other is social networks.

1.2.1 Classification in Crowdsourcing Systems

Crowdsourcing has attracted intense interest in recent years as a new paradigm for distributed classification that harnesses the intelligence of the crowd, exploiting inexpensive and online labor markets in an effective manner [9,46,48,49,107,108,129]. Crowdsourcing enables a new framework to utilize distributed human wisdom to solve problems that machines cannot perform well, like handwriting recognition, paraphrase acquisition, audio transcription, and photo tagging [12, 30, 60, 85]. While conventional group collaboration and cooperation frameworks rely heavily on a collection of experts in related fields, the crowd in crowdsourcing usually consists of non-experts, and it relies on the co-work of diverse amateurs. This makes the problem of crowdsourcing for classification quite challenging and is investigated in this thesis.

In spite of the successful applications of crowdsourcing, the relatively low quality of output remains a key challenge [1,51,76] due to the following reasons. First, the worker pool is anonymous in nature, which allows unskilled and unreliable workers in the crowd [114]. Second, the assumption that the workers are sufficiently motivated, extrinsically or intrinsically, to take part seriously in the crowdsourcing task, is highly questionable [44, 112]. Third, for the non-expert crowd to successfully complete the crowdsourcing work, some tasks are specifically designed to be com-

posed of easy but tedious microtasks [117], which might cause boredom and result in low-quality work. Finally, noisy and unreliable responses to the tasks cannot be detected and tagged before aggregation so that appropriate weights could be assigned to responses [128]. For instance, as the most popular aggregation decision rule, simple majority voting takes all of the answers (including the noisy ones) into account with the same weight [98].

Several methods have been proposed to deal with the aforementioned problems [44, 114, 117, 128], [61, 62, 90, 99, 131]. A crowdsourcing task is decomposed into microtasks that are easy for an individual to accomplish, and these microtasks could be as simple as binary distinctions [62]. It is expected that very little knowledge would be needed to complete the microtasks, and typically common sense or observation is good enough for such microtasks. A classification problem with crowdsourcing, where taxonomy and dichotomous keys are used to design binary questions, is considered in [117]. These schemes lower the chance for the workers to make mistakes. New aggregation rules that mitigate the unreliability of the crowd and improve the crowdsourcing system performance are investigated in [99,128]. In [114,117], the authors employed binary questions and studied the use of error-control codes and decoding algorithms to design crowdsourcing systems for reliable classification despite unreliable crowd workers. A group control mechanism where the reputation of the workers is taken into consideration to partition the crowd accordingly into groups is presented in [90] and [131]. Group control and majority voting techniques are compared in [44], which reports that majority voting is more cost-effective on less complex tasks.

In the past work on classification via crowdsourcing, crowd workers were required to provide a definitive yes/no response to binary microtasks. Crowd workers may be unable to answer questions for a variety of reasons such as lack of expertise. As an example, in mismatched speech transcription, i.e., transcription by crowd workers who do not know the language, workers may not be able to perceive the phonological dimensions they are tasked to differentiate [57]. In such situations, it is useful to provide a "no response" or "reject" option. In this thesis, we consider the problem of classification via crowdsourcing with a reject option.

Research in psychology [26] suggests a greater tendency to select the reject option (no choice)

when the choice set offers several attractive alternatives but none that can be easily justified as the best, resulting in less arbitrary decisions. To avoid requiring workers to respond to microtasks beyond their expertise resulting in making random guesses, in this thesis, we consider the optimal design of the aggregation rule in crowdsourcing systems where the workers are not forced to make a binary choice when they are unsure of their response and can choose not to respond. As shown in [56], the quality of label prediction can be improved by adopting a decision rejection option to avoid results with low confidence. The reject option has also been considered in machine learning and signal processing literatures [4, 19, 89, 111]. With a reject option, the payment mechanism is investigated in crowdsourcing systems where the workers can also report their confidence about the submitted answers [100].

1.2.2 Classification in Social Networks

In social networks, information emerges dynamically and diffuses quickly via agent interactions [70]. Thus, it is challenging to understand and predict the information diffusion mechanisms in complex social networks. For example, to better characterize the factors influencing spread of diseases, planned terrorist attacks, and effective social marketing campaigns, etc., it is crucially important if one can exploit the knowledge of the information dynamics [37]. Essentially, the focused research problem to understand information diffusion in social networks is: Which members of the network are influential and play important roles in the information diffusion process? This is a typical classification problem is social networks, i.e., influential node detection, which is a central research topic in social network analysis. One is confronted with two crucial challenges while attempting to address the problem. First, a descriptive diffusion model, which can mimic the information diffusion behavior observed in real world, is required. Second, efficient learning algorithms are required for inferring influence structure based on the assumed diffusion model.

In the literature, various information diffusion prediction models have been developed [28, 38, 121, 126, 130]. In many of these models, it is typically assumed that the social network is a connected graph and the corresponding network structure is available *a priori*. However, the

structure of the network can be implicit or difficult to model, and the knowledge of the complex social networks is extremely difficult to obtain in practice. For example, modeling the structure of the spread of infectious disease among people is almost impossible. Therefore, network structure unaware diffusion prediction models have gained tremendous interests. In [126], Yang *et. al.* proposed a linear influence model (LIM), which can effectively predict the information volume by assuming that each of the contagions spreads with the same influence in an implicit network. Subsequently, in [121], the authors extended LIM by exploiting the sparse structure in the influence function to identify the influential nodes. However, most of the existing approaches ignore the information regarding the relationships between multiple contagions, which could be used for more accurate modeling. In this thesis, we address the above issues by augmenting linear influence models with complex task dependency information. Furthermore, the influence function values for individual users are collected for the classification problem of influential node detection.

1.3 Classification Problems in Data-driven Systems

For data-driven systems, this thesis considers two aspects of classification problems. The first one considers a decentralized classification problem where the agents communicate with their neighbors to reach a classification consensus in a decentralized learning manner. The second one considers the fundamental performance limits of traditional classification from an information theoretic perspective.

1.3.1 Classification in Decentralized Learning

Decentralized classification problems fit into the general framework where a cost function with respect to the classifier is to be optimized numerically by a decentralized learning algorithm. This problem structure is also applicable to collaborative autonomous inference in statistics, distributed cooperative control of unmanned vehicles in control theory, and training of models (such as, support vector machines, deep neural networks, etc.) in machine learning.

However, it is often infeasible to solve the classification problem numerically at a single node (or agent) due to the emergence of the big data era and associated sizes of datasets. Thus, the decentralized optimization setting [11, 63] has attracted much interest, in which the training data for the classification problem is stored and processed in a distributed fashion across a number of interconnected nodes, and the optimization problem is solved collectively by the cluster of nodes. The decentralized learning system can be implemented on an arbitrarily connected network of computational nodes. In such a system, the classification problem is treated as a consensus optimization problem such that the nodes provide one common final solution.

There exist several optimization methods for solving decentralized classification problems, including belief propagation [88], distributed subgradient descent algorithms [79], dual averaging methods [29], and the alternating direction method of multipliers (ADMM) [11]. Among these, ADMM has drawn significant attention, as it is well suited for decentralized optimization and demonstrates fast convergence in many applications, such as online learning, decentralized collaborative learning, neural network training, and so on [45, 109, 125]. We use ADMM in this thesis to numerically solve the decentralized classification problem.

The ADMM algorithm while solving the problem involves two basic steps: (i) a communication step for exchanging information only among single-hop neighbors; and, (ii) an update step for updating the local solution at each agent. By alternating between the two, local iterates eventually converge to the global solution. Performance of the applications heavily depends on whether ADMM can have convergence with acceptable accuracy or not. Therefore, an immense amount of effort has been put in to establish convergence rates of ADMM in different scenarios [101].

However, it is assumed in most of the past works on ADMM that the decentralized system is ideal in that the updates are not erroneous. This assumption is very restrictive and rarely satisfied in practice which limits the applicability of these results. Note that due to the decentralized nature of the systems considered, computation over federated machines induces a higher risk of unreliability because of communication noise, crash failure, and adversarial attacks. Therefore, the design and analysis of decentralized optimization algorithms in the presence of these practical challenges is

of utmost importance.

In this thesis, we analyze the convergence behavior of ADMM in the presence of unreliable agents, and provide guidelines to minimize the impact of error on the classification performance. Furthermore, we propose a robust scheme that can eliminate the impact of errors.

1.3.2 Classification as Hypothesis Testing

For hypothesis testing problems, information theoretic tools have been developed to characterize the error exponent [22,23,34,122], and to study a class of distributed parametric hypothesis testing problems [40, 41, 115]. For sequential multi-hypothesis testing, information theoretic bounds on the sample size subject to constraints on the error probabilities have been developed in [65]. A generalization of the classical hypothesis testing problem is studied in [78], where a Bayesian decision maker is designed to enhance its information about the correct hypothesis. Information theory has also been applied to study *nonparametric* hypothesis testing problems with the primary focus being on the Neyman-Pearson formulation [39, 66]. An information-theoretic approach to the problem of a nonparametric hypothesis test with a Bayesian formulation is presented in [50]. By factorizing dependent variables into mutually independent subsets, it has been shown that the likelihood ratio can be written as the sum of two sets of Kullback-Leibler divergence (KLD) terms, which is then used to quantify loss in hypothesis separability. Our study is different in that we focus on the asymptotic regime where the number of hypotheses scales with sample size.

The classification problem we study here can also be viewed as a supervised learning problem studied in the machine learning literature. However, the problem formulated here is different from the traditional supervised learning problem [7], where sample points corresponding to the same label are treated as individual samples, and their underlying statistical structure is not exploited in the design of classification rules. For example, the support vector machine (SVM) is one of the important classification algorithms for supervised learning, where the distance between samples is measured either by the Euclidean distance or by a kernel-based distance. Such distances do not exploit the underlying statistical distributions of data samples. A robust form of the SVM in [102]

incorporates the probabilistic uncertainty into the maximization of the margin. Our formulation exploits the underlying probabilistic structure of data samples, which is also robust to missing data, system noise, etc.

A formulation of the supervised learning problem that is similar to our formulation has been studied previously in [77]. The proposed approach, therein named support measure machine (SMM), exploits the kernel mean embedding to estimate the distance between probability distributions. In fact, the comparison between an SMM and an SVM also reflects the differences between our formulation and the traditional supervised learning problem. However, the study in [77] focused only on the regime with finite and fixed number of classes, and did not characterize the decay exponent of the error probability, whereas our focus is mainly on the asymptotic regime with infinite number of classes, and on the scaling behavior of the number of classes under which an asymptotically small error probability can be guaranteed. Nevertheless, the kernel-based approach developed in [77] as well as in various other papers [33, 35, 105] provide important techniques that we exploit in our study.

1.4 Outline and Contributions

This thesis is organized as follows: In Chapter 2, the optimal aggregation rule for the responses from honest human workers in crowdsourced classification problems is investigated. In Chapter 3, the presence of unreliable human workers is considered in crowdsourcing systems for classification, and the optimal aggregation rule is derived. In Chapter 4, an important classification problem namely the influential node detection problem in social networks without the knowledge of the structure of the network is considered. The decentralized learning in classification systems when some of the agents are unreliable is studied in Chapter 5. Further, in Chapter 6, we study the fundamental performance limits of classification problems posed in the form of nonparametric hypothesis testing from information theoretic perspectives. We then conclude this thesis in Chapter 7. The main contributions of each chapter are as follows.

Chapter 2 considers the design of an effective crowdsourcing system for M-ary classification where crowd workers complete simple microtasks which are aggregated to give a final result. We consider the novel scenario where workers have a reject option so they may skip microtasks they are unable or unwilling to do. For example in mismatched speech transcription, workers who do not know the language may be unable to respond to microtasks in phonological dimensions outside their categorical perception. We present an aggregation approach using a weighted majority voting rule, where each worker's response is assigned an optimized weight to maximize the crowd's classification performance. We evaluate system performance in both exact and asymptotic forms. We also show that human workers' confidence does not help in improving system performance.

In Chapter 3, the presence of spammers, who give random responses to the microtasks, is considered in the crowdsourced classification systems. First, we study the case when the spammers complete all the microtasks with random guesses. A heuristic adaptive approach is proposed by switching between oblivious and expurgation strategies, based on the estimation of several crowd parameters such as the fraction of greedy spammers. Next, we investigate the optimal behavior for spammers to maximize their monetary reward for completing tasks. To combat the impact from the spammers who behave strategically, we derive an optimal aggregation rule to maximize the classification performance in the presence of spammers.

Chapter 4 considers the classification problem of influential node detection in implicit social networks. We propose a descriptive diffusion model to take dependencies among the topics into account. We also propose an efficient algorithm based on alternating methods to perform inference and learning on the model. It is shown that the proposed technique outperforms existing influential node detection techniques. Furthermore, the proposed model is validated both on a synthetic and a real (ISIS) dataset. We show that the proposed approach can efficiently determine the influential users for specific contagions. We also present several interesting patterns of the selected influential users for the ISIS dataset.

In Chapter 5, we consider the problem of decentralized learning of classification problems using ADMM in the presence of unreliable agents. We study the convergence behavior of the

decentralized ADMM algorithm and show that the ADMM converges to a neighborhood of the solution under certain conditions. We suggest guidelines for network structure design to achieve faster convergence. We also give several conditions on the errors to obtain exact convergence to the solution. A robust variant of the ADMM algorithm is proposed to enable decentralized classification in the presence of unreliable agents and its convergence to the optima is proved. We also provide experimental results to validate the analysis and show the effectiveness of the proposed robust scheme.

From an information theoretic perspective, Chapter 6 develops a nonparametric composite hypothesis testing approach for arbitrary distributions based on the maximum mean discrepancy (MMD) and Kolmogorov-Smirnov (KS) distance measure based tests. We introduce the information theoretic notion of discrimination capacity that is defined for the regime where the number of hypotheses scales along with the sample size. We also provide characterization of the corresponding error exponent and the discrimination rate, i.e., a lower bound on the discrimination capacity. Our framework is extended to unsupervised learning problems and similar performance limits are investigated.

In Chapter 7, we summarize the findings and results of this thesis, and present several directions and ideas for future research.

1.5 Bibliographic Notes

Most of the research work appearing in this thesis has either already been published or is in different stages of publication. The relationship between the chapters and the publications is shown in Table 1.1, while the list of publications is provided as follows.

Journal papers

J1 **Q. Li**; B. Kaikhura; J. J. Thiagarajan; Z. Zhang; P. K. Varshney, " Influential Node Detection in Implicit Social Networks using Multi-task Gaussian Copula Models," *Journal of Machine*

Learning Research (special issue), vol. 55, 2016.

J2 **Q. Li**; A. Vempaty; L. R. Varshney; P. K. Varshney, "Multi-object Classification via Crowdsourcing with a Reject Option," *IEEE Trans. Signal Process.*, vol. 65, no. 4, pp. 1068-1081.

J3 **Q. Li**; P. K. Varshney, "Resource Allocation and Outage Analysis for An Adaptive Cognitive Two-Way Relay Network," *IEEE Trans. Wireless Comm.*, vol. 16, no. 7, pp. 4727-4737, 2017.

J4 **Q. Li**; T. Wang; D. J. Bucci; Y. Liang; B. Chen; P. K. Varshney, "Nonparametric Composite Hypothesis Testing in an Asymptotic Regime," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 5, pp. 1005-1014, Oct 2018.

J5 T. Wang; **Q. Li**; D. J. Bucci; Y. Liang; B. Chen; P. K. Varshney, "K-medoids Clustering of Data Sequences with Composite Distributions," *IEEE Trans. Signal Process.*, under second review.

J6 **Q. Li**; P. K. Varshney, "Optimal Crowdsourced Classification with a Reject Option in the Presence of Spammers," *IEEE Trans. Signal Process.*, to be submitted.

Conference papers

C1 **Q. Li**, B. Kailkhura, Y. Liang, and P. K. Varshney. "Manifold Regularized Generative Adversarial Networks." *To be submitted to CVPR 2019.*

C2 **Q. Li**, B. Kaikhura, R. Goldhahn, P. Ray, and P. K. Varshney. "Robust Decentralized Learning Using ADMM with Unreliable Agents." *Submitted to AISTATS 2019.*

C3 B. Geng, **Q. Li**, P. K. Varshney. "Decision Tree Design for Classification in Crowdsourcing Systems." *Asilomar 2018*

C4 **Q. Li** and P. K. Varshney. "Optimal Crowdsourced Classification with a Reject Option in the Presence of Spammers." *ICASSP 2018.*

C5 **Q. Li**, Y. Zhou, Y. Liang, and P. K. Varshney. "Convergence Analysis of Proximal Gradient with Momentum for Nonconvex Optimization." *ICML 2017, vol. 60, pp. 2111-2119.*

C6 **Q. Li** and P. K. Varshney. "Does Confidence Reporting From the Crowd Benefit Crowdsourcing Performance?" *SOCIALSENS 2017, no. 6, pp. 49-54.*

C7 **Q. Li**, B. Kaikhura, J. J. Thiagarajan, Z. Zhang, and P. K. Varshney. "Influential Node Detection in Implicit Social Networks using Multi-task Gaussian Copula Models." *Neural Information Processing Systems (NIPS) Workshop, 2016.*

	Chapter 3	Chapter 4	Chapter 5	Chapter 6	Chapter 7
J1			•		
J2	•	•			
J3					
J4					•
J5					•
J6		•			
C1					
C2				•	
C3					
C4		•			
C5					
C6	•				
C7			•		

Table 1.1: Connection between publications & chapters

CHAPTER 2
CLASSIFICATION IN CROWDSOURCING: RELIABLE AGENTS

2.1 Introduction

Engineered social systems such as crowdsensing, crowdsourcing, and social production are becoming increasingly prevalent for classification tasks. Advances in signal processing theory and methods to optimize these novel human-oriented approaches to signal acquisition and processing are needed [87].

Crowdsourcing has particularly attracted intense interest [9, 46, 48, 49, 107, 108, 129] as a new paradigm for classification tasks such as handwriting recognition, paraphrase acquisition, speech transcription, image quality assessment, and photo tagging [12, 30, 47, 60, 85, 94, 95, 104], which are all essentially inference problems of M-ary classification. Unfortunately, the low quality of crowdsourced output remains a key challenge [1, 51, 76].

Low-quality work may arise not only because workers are insufficiently motivated to perform well [44, 112], but also because workers lack the skills to perform the task that is posed to them [117]. Decomposing larger tasks into smaller subtasks for later aggregation allows workers lacking certain skills to contribute useful information [62, 117]. Thus, the lack of certain skill of a particular

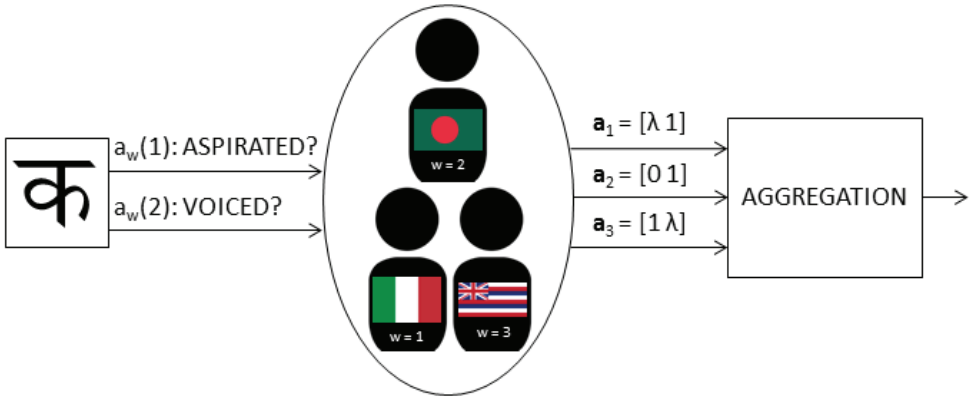


Fig. 2.1: An illustrative example of the proposed crowdsourcing framework.

worker could be only reflected in a subset of subtasks, and this worker would be able to complete other subtasks.

As an illustrative example of lack of skill, consider the problem of mismatched crowdsourcing for speech transcription, which has garnered interest in the signal processing community [18, 42, 57, 58, 64, 69, 113]. The basic idea is to use crowd workers to transcribe a language they do not speak, into nonsense text in their own native language orthography. Certain phonological dimensions, such as aspiration or voicing, are used to differentiate phonemes in one language but not others [113]. Moreover due to categorical perception acquired in childhood, workers lacking a given phonological dimension in their native language may be unable to make relevant distinctions. That is, they lack the skill for the task.

Fig. 2.1 depicts the task of language transcription of Hindi. Suppose the four possibilities for a velar stop consonant to transcribe are $R = \{$क, ख, ग, घ$\}$. The binary query of "aspirated or unaspirated" differentiates between $\{$ख, घ$\}$ and $\{$क, ग$\}$, whereas the query of "voice or unvoiced" differentiates between $\{$ग, घ$\}$ and $\{$क, ख$\}$. Note the two queries are independent. Now suppose the first worker is a native Italian speaker. Since Italian does not use aspiration, this worker will be unable to differentiate between क and ख, or between ग and घ. It would be useful if this worker specified the inability to perform the task through a special symbol λ, rather than guessing randomly. Suppose the second worker is a native Bengali speaker; since this language makes a four-way distinction among velar stops, such a worker will probably answer both questions with-

out a λ. Now suppose the third worker is a native Hawaiian speaker; since this language does not use voicing, such a worker will not be able to differentiate between क and ग, or between ख and घ. Hence, it would be useful if this worker answered λ for the question of differentiating among these two subchoices.

This thesis allows workers to not respond, i.e., allowing a *reject option*, as in the example. Research in psychology suggests a greater tendency to select the reject option when the choice set offers several attractive alternatives but none that can be easily justified as the best, resulting in less arbitrary decisions [26]. The reject option has previously been considered in the machine learning and signal processing literatures [4, 19, 56, 89, 111], but we specifically consider worker behavior and aggregation rules for crowdsourcing with a reject option. To characterize performance, we derive a closed-form expression for the probability of a microtask being correct, together with the asymptotic performance when the crowd size is large.

Several methods have been proposed to deal with noisy crowd work when crowd workers are required to provide a definitive yes/no response [44, 61, 62, 90, 99, 117, 128, 131], rather than allowing a reject option. Without the reject option, noisy responses to tasks cannot be tagged before aggregation so appropriate weights cannot be assigned [128]. For instance, the popular majority voting rule weights all answers equally [98], though new weighted aggregation rules have also been developed [99, 128]. Vempaty et al. employed error-control codes and decoding algorithms to design reliable crowdsourcing systems with unreliable workers [117]. A group control mechanism where worker reputation is used to partition the crowd into groups is presented in [90, 131]; comparing group control to majority voting indicates majority voting is more cost-effective on less complex tasks [44].

2.2 Classification Task for Crowds

Consider W workers taking part in an M-ary classification task. Each worker is asked N simple binary questions, termed microtasks, which eventually lead to a decision among the M classes.

We assume it is possible to construct $N = \lceil \log_2 M \rceil$ independent microtasks of equal difficulty. Such microtasks can be designed using taxonomy and dichotomous keys [97]. The workers submit results that are combined to give the final decision. A worker's answer to a single microtask is represented by either "1" (Yes) or "0" (No) [117] and so the wth worker's ordered answers $\mathbf{a}_w(i)$, $i \in \{1, 2, \ldots, N\}$ to all microtasks form an N-bit word, denoted $\mathbf{a}_w$.

Each worker has the reject option of skipping microtasks; we denote a skipped answer as λ, whereas "1/0" (Yes/No) responses are termed *definitive answers*. Due to variability in worker expertise, the probability of submitting definitive answers is different for each worker. Let $p_{w,i}$ represent the probability the wth worker submits λ for the ith microtask. Similarly, let $r_{w,i}$ be the probability $\mathbf{a}_w(i)$, the ith answer of the wth worker, is correct given a definitive answer has been submitted. Due to worker anonymity, we study performance when $p_{w,i}$ and $r_{w,i}$ are realizations of certain probability distributions, denoted $F_P(p)$ and $F_R(r)$ with corresponding means m and μ, respectively. Let H_0 and H_1 be hypotheses of bits "0" and "1" for a single microtask, respectively. For simplicity of performance analysis, "0" and "1" are assumed equiprobable for every microtask. The crowdsourcing platform or fusion center (FC) collects the N-bit words from W workers and aggregates results, as discussed next.

2.2.1 Weighted Majority Voting

We first investigate weighted majority voting as the fusion rule for classification. Consider all object classes as elements in the set $D = \{e_j\}$, $j = 1, \ldots, M$, where e_j represents the jth class. As indicated earlier, a worker's definitive responses to the microtasks determine a subset in the original set D, consisting of multiple elements/classes. If all responses from the crowd are definitive, the final subsets are singletons and a single class is identified as the object class. Since some microtasks may be answered with a response λ, the resulting subsets may not be singletons and each element of the same corresponding subsets will be chosen equiprobably as the classification decision. Let us denote the subset determined by the definitive answers of the wth worker as $D_w \in D$. The task manager assigns the same weight to all elements in D_w based on the wth worker's answer.

With the submitted answers from W workers, we determine the overall weight assigned to any jth element of D as

$$\mathbb{W}\left(e_j\right) = \sum_{w=1}^{W} W_w I_{D_w}\langle e_j\rangle, j = 1, 2, \ldots, M, \;\; D_w \in D, \tag{2.1}$$

where W_w is the weight[1] assigned to D_w, and $I_{D_w}\langle e_j\rangle$ is an indicator function which equals 1 if $e_j \in D_w$ and 0 otherwise. Then the element e_D with the highest weight is selected as the final class, and the classification rule is stated as

$$e_D = \arg\max_{e_j \in D} \left\{\mathbb{W}\left(e_j\right)\right\}. \tag{2.2}$$

For tie-breaking, randomly choose from the classes with the same weight. Notice that conventional majority voting requires full completion of all microtasks without the reject option and has identical W_w for each worker's decision.

Next, we show how the problem formulated in (2.2) can be further decomposed.

Proposition 2.1. *Classification rule* (2.2) *is equivalent to bit-by-bit decision since the ith bit, $i = 1, \ldots, N$, is decided by*

$$\sum_{w=1}^{W} W_w I_1\left\langle i, w\right\rangle \underset{H_0}{\overset{H_1}{\gtrless}} \sum_{w=1}^{W} W_w I_0\left\langle i, w\right\rangle, \tag{2.3}$$

where $I_s\left\langle i, w\right\rangle$, $s \in \{0, 1\}$, is the indicator function which is 1 if the wth worker's answer to the ith bit is "s", otherwise $I_s\left\langle i, w\right\rangle = 0$. For tie-breaking, randomly choose 0 or 1.

Proof. The class e_D corresponds to a unique N-bit word. Thus, if the ith bit of the N-bit word corresponding to the class e_D is equal to s, s has the same weight as assigned to e_D, which is greater than or equal to the symbol $1 - s$. □

[1]The assignment of these weights will be discussed later in the thesis.

2.2.2 Optimal Bit-by-bit Bayesian Aggregation

Let $\mathcal{A}(i) = [\mathbf{a}_1(i), \mathbf{a}_2(i), \ldots, \mathbf{a}_W(i)]$ denote all the answers to ith microtask collected from the crowd. For the binary hypothesis testing problem corresponding to the ith bit of the N-bit word, the log-likelihood ratio test is

$$\log \frac{P\left(H_1|\mathcal{A}(i)\right)}{P\left(H_0|\mathcal{A}(i)\right)} \underset{H_0}{\overset{H_1}{\gtrless}} 0, \tag{2.4}$$

where $P\left(\cdot\right)$ denotes the probability density function. We can express the likelihood ratio as

$$\frac{P\left(H_1|\mathcal{A}(i)\right)}{P\left(H_0|\mathcal{A}(i)\right)} = \frac{\prod_{w=1}^{W} P\left(\mathbf{a}_w(i)|H_1\right)}{\prod_{w=1}^{W} P\left(\mathbf{a}_w(i)|H_0\right)} = \frac{\prod_{S_1}\left(1-p_{w,i}\right)r_{w,i}\prod_{S_0}\left(1-p_{w,i}\right)\left(1-r_{w,i}\right)\prod_{S_\lambda} p_{w,i}}{\prod_{S_1}\left(1-p_{w,i}\right)\left(1-r_{w,i}\right)\prod_{S_0}\left(1-p_{w,i}\right)r_{w,i}\prod_{S_\lambda} p_{w,i}}, \tag{2.5}$$

where S_1 is the set of w such that $\mathbf{a}_w(i) = 1$, S_0 is the set of w such that $\mathbf{a}_w(i) = 0$ and S_λ is the set of w such that $\mathbf{a}_w(i) = \lambda$, respectively. Then, it is straightforward to show that the test for the decision on the ith bit is

$$\sum_{w\in S_1} \log \frac{r_{w,i}}{1-r_{w,i}} \underset{H_0}{\overset{H_1}{\gtrless}} \sum_{w\in S_0} \log \frac{r_{w,i}}{1-r_{w,i}}. \tag{2.6}$$

Note that the optimal Bayesian criterion can also be viewed as the general weighted majority voting rule in (2.3) with weight $W_w = \log \frac{r_{w,i}}{1-r_{w,i}}$, also called the Chair-Varshney rule [13]. Note that (2.3) represents majority voting when $W_w = 1$.

However, this optimal Bayesian criterion can only be used if $r_{w,i}$ for every worker is known *a priori*, which is usually difficult to estimate for anonymous crowds just from submitted answers. The difficulty in obtaining prior information makes the simple majority voting scheme very effective and therefore widely used [98]. We show later that our proposed method—which need not estimate $r_{w,i}$ but only its mean μ—can outperform conventional majority voting.

2.2.3 Class-based Aggregation Rule

For the general weighted majority voting scheme where e_C is the correct class, the probability of misclassification is

$$\begin{aligned} P_m &= \Pr\left(e_D \neq e_C\right) \\ &= \Pr\left(\arg\max_{e_j \in D} \left\{ \sum_{w=1}^{W} W_w I_{D_w}\left\langle e_j \right\rangle \right\} \neq e_C \right) \\ &= 1 - \Pr\left(\arg\max_{e_j \in D} \left\{ \sum_{w=1}^{W} W_w I_{D_w}\left\langle e_j \right\rangle \right\} = e_C \right), \end{aligned} \tag{2.7}$$

where $\Pr(\mathcal{E})$ is the occurrence probability of event $\mathcal{E}$.

A closed-form expression for the error probability P_m cannot be derived without an explicit expression for W_w; hence it is difficult to determine the optimal weights to minimize P_m.

Consequently, we consider an optimization problem based on a different objective function and propose a novel weighted majority voting method that outperforms simple majority voting. Note that e_D is chosen as the decision for classification such that e_D has the maximum overall weight collected from all the workers. Thus, we maximize the average overall weight assigned to the correct class while the overall weight collected by all the elements remains the same as the other existing methods such as majority voting. We state the optimization problem over the weights as

$$\begin{aligned} &\text{maximize} \;\; E_C\left[\mathbb{W}\right] \\ &\text{subject to} \;\; E_O\left[\mathbb{W}\right] = K, \end{aligned} \tag{2.8}$$

where $E_C\left[\mathbb{W}\right]$ is the crowd's average weight contribution to the correct class and $E_O\left[\mathbb{W}\right]$ is the average weight contribution to all possible classes. K is set to a constant so we are looking for a maximized portion of weight contribution to the correct class while the weight contribution to all classes remains fixed. This procedure ensures that one cannot obtain greater $E_C\left[\mathbb{W}\right]$ by simply increasing the weight for each worker's answer, while K results in a normalized weight assignment scheme. If two weight assignment schemes share the same value of $E_O\left[\mathbb{W}\right]$, one can expect better

performance by the scheme with higher $E_C[\mathbb{W}]$. Thus, K facilitates a relatively easier performance comparison of different weight assignment schemes.

2.3 System With Honest Crowd Workers

We first consider the case where the crowd is entirely composed of honest workers—workers that are not greedy and honestly observe, think, and answer microtasks posed to them, while skipping queries they are not confident about. The wth worker responds with a λ to the ith microtask with probability $p_{w,i}$. Next, we derive the optimal weight W_w for the wth worker in this case.

Proposition 2.2. *To maximize the normalized average weight assigned to the correct classification element, the weight for wth worker's answer is given by $W_w = \mu^{-n}$, where $\mu = E[r_{w,i}]$ and n is the number of definitive answers that the wth worker submits.*

Proof. See Appendix A.1. □

Note that when workers are forced to make a hard decision for every single bit, the weights derived above become identical. In general, the weight depends on the number of questions answered by a worker: if more questions are answered, the weight assigned to the corresponding worker's answer is larger. Assuming a worker's correct probability is greater than half if he/she gives a definitive answer, a larger number of definitive answers increases the chance the quality of the worker is higher than others. Increased weight can thereby emphasize the contribution of high-quality workers and improve overall classification performance.

2.3.1 Estimation of μ

Before the proposed aggregation rule can be used, note that μ has to be estimated to assign the weight for every worker's answers. Here, we give two approaches to estimate μ.

Training-based Approach

In addition to the N microtasks, the task manager inserts additional questions to estimate the crowd's μ value. The answers to such "gold standard" questions are known to the manager [27,55]. By checking the crowd worker's answers, μ can be estimated. Suppose the first T questions are training questions, and let $\bar{\mathbf{B}}$ be the T-bit correct answers to them. First, we calculate the ratio $r(w)$ as

$$r(w) = \sum_{i=1}^{T} \frac{I_{\bar{\mathbf{B}}(i)} \langle \mathbf{a}_w(i) \rangle}{I(w)}, \tag{2.9}$$

where $I_x\langle y \rangle$ is the indicator function which is 1 if $x = y$ and 0 otherwise, and $I(w) = \sum_{i=1}^{T} (I_1\langle i, w \rangle + I_0\langle i, w \rangle)$. In order to avoid the cases where some workers submit λ for all the training questions, we estimate μ as follows

$$\hat{\mu} = \frac{1}{W - \epsilon} \sum_{w=1}^{W} r(w), \tag{2.10}$$

where ϵ is the number of workers that submit all λ for the training questions and the corresponding $r(w)$ is set to 0.

Majority-voting based Approach

We use majority voting to obtain the initial aggregation result and set it as the benchmark to estimate μ. First, all the answers $\mathbf{a}_w(i)$ are collected to obtain the benchmark $\mathcal{B}(i)$ by traditional majority voting, where $i = 1, \ldots, N$. Note that $\mathcal{B}(i)$ may contain λ since it is possible that all answers $\mathbf{a}_w(i)$ have λ at the same position. Then, for the wth worker, we calculate the ratio $r(w)$ as

$$r(w) = \sum_{i=1}^{N} \frac{I_{\mathcal{B}(i)} \langle \mathbf{a}_w(i) \rangle}{I(w)}, \tag{2.11}$$

where $I_\lambda\langle\lambda\rangle = 0$, and $I(w) = \sum_{i=1}^{N} \left(I_1\langle i, w\rangle + I_0\langle i, w\rangle\right)$. As before, we estimate μ as in (2.10), but where ϵ is the number of workers that submit λ for all microtasks.

2.3.2 Performance Analysis

In this subsection, we characterize performance of the proposed classification framework in terms of probability of correct classification P_c. Note that we have overall correct classification only when all the bits are classified correctly,[2] which also offers the lower bound in the general case where the microtasks are not completely independent of each other.

First, we restate the bit decision criterion in (2.3) as

$$\sum_{w=1}^{W} T_w \underset{H_0}{\overset{H_1}{\gtrless}} 0 \tag{2.12}$$

with $T_w = W_w \left(I_1\langle i, w\rangle - I_0\langle i, w\rangle\right)$, where the resulting

$$T_w \in \{-\mu^{-N}, -\mu^{-N+1}, \ldots, -\mu^{-1}, 0, \mu^1, \ldots, \mu^{N-1}, \mu^N\}$$

.

Proposition 2.3. *For the ith bit, the probability mass function of T_w under hypothesis H_s, $\Pr\left(T_w|H_s\right)$, for $s \in \{0, 1\}$, is:*

$$\Pr\left(T_w = I(-1)^{t+1}\mu^{-n}|H_s\right) = \begin{cases} r_{w,i}^{1-|s-t|}(1 - r_{w,i})^{|s-t|}\varphi_n(w), I = 1 \\ p_{w,i}, \qquad\qquad\qquad\quad I = 0 \end{cases}, t \in \{0, 1\}, n \in \{1, \ldots, N\}, \tag{2.13}$$

[2]When $N > \log_2 M$, the N-bit answer after aggregation may correspond to a class that does not exist; this is also misclassification.

where $I = I_1\langle i, w\rangle + I_0\langle i, w\rangle$, $\varphi_n(w) = (1 - p_{w,i}) \sum_C \prod_{\substack{j=1 \\ j\neq i}}^{N} p_{w,j}^{k_j}(1 - p_{w,j})^{1-k_j}$ *and* C *is the set*

$$C = \left\{ \{k_1, k_2, \ldots, k_{i-1}, k_{i+1}, \ldots, k_N\} : \sum_{\substack{j=1 \\ j\neq i}}^{N} k_j = N - n \right\}$$

with $k_j \in \{0, 1\}$.

Proof. See Appendix A.2. □

Since hypotheses H_0 and H_1 are assumed equiprobable, the correct classification probability for the ith bit $P_{c,i}$ is $P_{c,i} = \frac{1+P_{d,i}-P_{f,i}}{2}$, where $P_{d,i}$ is the probability of deciding the ith bit as "1" when the true bit is "1" and $P_{f,i}$ is the probability of deciding the ith bit as "1" when the true bit is "0".

Proposition 2.4. *The probability of correct classification for the* i*th bit* $P_{c,i}$ *is*

$$P_{c,i} = \frac{1}{2} + \frac{1}{2} \sum_{S} \binom{W}{\mathbb{Q}} \left(F_i\left(\mathbb{Q}\right) - F_i'\left(\mathbb{Q}\right)\right) + \frac{1}{4} \sum_{S'} \binom{W}{\mathbb{Q}} \left(F_i\left(\mathbb{Q}\right) - F_i'\left(\mathbb{Q}\right)\right) \tag{2.14}$$

with

$$F_i(\mathbb{Q}) = \prod_{w\in G_\lambda} p_{w,i} \prod_{w\in G_0} (1 - r_{w,i})\varphi_n\left(w\right) \prod_{w\in G_1} r_{w,i}\varphi_n\left(w\right)$$

and

$$F_i'(\mathbb{Q}) = \prod_{w\in G_\lambda} p_{w,i} \prod_{w\in G_1} (1 - r_{w,i})\varphi_n\left(w\right) \prod_{w\in G_0} r_{w,i}\varphi_n\left(w\right),$$

where

$$\mathbb{Q} = \left\{ (q_{-N}, q_{-N+1}, \ldots q_N) : \sum_{n=-N}^{N} q_n = W \right\} \tag{2.15}$$

with natural numbers q_n *and* q_0, G_0 *denotes the worker group that submits "0" for* i*th microtask,*

G_1 the group that submits "1" and G_λ the group that submits λ, and

$$S = \left\{ \mathbb{Q} : \sum_{n=1}^{N} \mu^{-n} \left(q_n - q_{-n} \right) > 0 \right\}, \tag{2.16}$$

$$S' = \left\{ \mathbb{Q} : \sum_{n=1}^{N} \mu^{-n} \left(q_n - q_{-n} \right) = 0 \right\}, \tag{2.17}$$

and $\binom{W}{\mathbb{Q}} = W! / \prod_{n=-N}^{N} q_n!$.

Proof. See Appendix A.3. □

Proposition 2.5. *The probability of correct classification P_c in the crowdsourcing system is*

$$P_c = \Big[\frac{1}{2} + \frac{1}{2} \sum_{S} \binom{W}{\mathbb{Q}} \left(F\left(\mathbb{Q}\right) - F'\left(\mathbb{Q}\right) \right) + \frac{1}{4} \sum_{S'} \binom{W}{\mathbb{Q}} \left(F\left(\mathbb{Q}\right) - F'\left(\mathbb{Q}\right) \right) \Big]^N, \tag{2.18}$$

where

$$F(\mathbb{Q}) = m^{q_0} \prod_{n=1}^{N} (1-\mu)^{q_{-n}} \mu^{q_n} \left(C_{N-1}^{n-1} (1-m)^n m^{N-n} \right)^{q_{-n}+q_n} \tag{2.19}$$

and

$$F'(\mathbb{Q}) = m^{q_0} \prod_{n=1}^{N} (1-\mu)^{q_n} \mu^{q_{-n}} \left(C_{N-1}^{n-1} (1-m)^n m^{N-n} \right)^{q_{-n}+q_n}. \tag{2.20}$$

Proof. See Appendix A.4. □

In practice, the number of workers for crowdsourcing tasks is large (hundreds). Thus, it is useful to investigate asymptotic system performance when W increases without bound.

Proposition 2.6. *As the number of workers W approaches infinity, the probability of correct clas-*

sification P_c is

$$P_c = \left[Q\left(-\frac{M}{\sqrt{V}} \right) \right]^N, \tag{2.21}$$

where $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{\frac{-t^2}{2}} dt$, and M and V are given as

$$M = \frac{W(2\mu-1)(1-m)}{\mu} \left(\frac{1}{\mu} - \left(\frac{1}{\mu} - 1 \right) m \right)^{N-1}, \tag{2.22}$$

and

$$V = \frac{W(1-m)}{\mu^2} \left(\frac{1}{\mu^2} - \left(\frac{1}{\mu^2} - 1 \right) m \right)^{N-1} - \frac{M^2}{W}. \tag{2.23}$$

Proof. See Appendix A.5. □

For large but finite crowds, the asymptotic result (2.21) is a good characterization of actual performance. Let us, therefore, consider (2.21) in more detail. First, we rewrite (2.21) as

$$P_c = \left[Q\left(-\sqrt{\frac{W}{\frac{1}{f(\mu,m)} - 1}} \right) \right]^N, \tag{2.24}$$

where

$$f(\mu, m) = (1-m)(2\mu-1)^2 (g(\mu, m))^{N-1}, \tag{2.25}$$

and

$$g(\mu, m) = \frac{(1-(1-\mu)m)^2}{1-(1-\mu^2)m}. \tag{2.26}$$

Theorem 2.1. *The correct classification probability of the system increases with increasing crowd*

size W.

Proof. Follows from (2.24) as the probability of correct classification increases monotonically with respect to W. □

Theorem 2.2. *The correct classification of the system increases with increasing* μ.

Proof. We take the partial derivative of $g(\mu, m)$ with respect to μ and obtain

$$\frac{\partial g}{\partial \mu} = \frac{2m(1-\mu)\,(1-m)\,\mathbf{A}}{\mathbf{B}^2}, \tag{2.27}$$

where $\mathbf{A} = m\mu - m + 1$, and $\mathbf{B} = m\mu^2 - m + 1$.

Clearly $\frac{\partial g}{\partial \mu} > 0$. Recall (2.24), (2.25), and (2.26) where a larger P_c results as μ increases. Then, the classification performance of the task in the crowdsourcing system also increases. □

To obtain the relation between crowd's performance in terms of P_c and m, we take the partial derivative of $f(\mu, m)$ with respect to m and obtain

$$\begin{aligned}\frac{1}{(2\mu-1)^2}\frac{\partial f}{\partial m} &= -\left(\tfrac{\mathbf{A}^2}{\mathbf{B}}\right)^{N-1} \\ &+ (N-1)\,(1-m)\left(\tfrac{\mathbf{A}^2(\mu^2-1)}{\mathbf{B}^2} + \tfrac{2\mathbf{A}(1-\mu)}{\mathbf{B}}\right)\left(\tfrac{\mathbf{A}^2}{\mathbf{B}}\right)^{N-2}.\end{aligned}$$

After some mathematical manipulations, we observe:

- When $m > \frac{1}{1+\mu}$, we can guarantee that $\frac{\partial f}{\partial m} < 0$, which means that the crowd performs better as P_c increases with decreasing m.
- When $m < \frac{1}{1+\mu}$ and $N \geq \frac{(m\mu-m+1)^2}{(1-m)(1-\mu)^2(m\mu+m-1)} + 1$, we can guarantee that $\frac{\partial f}{\partial m} > 0$, which means that the crowd performs better as P_c increases with increasing m.

These two observations indicate that a larger probability of the crowd responding to the ith microtask with λ does not necessarily degrade crowd's performance in terms of the detection of the ith microtask.

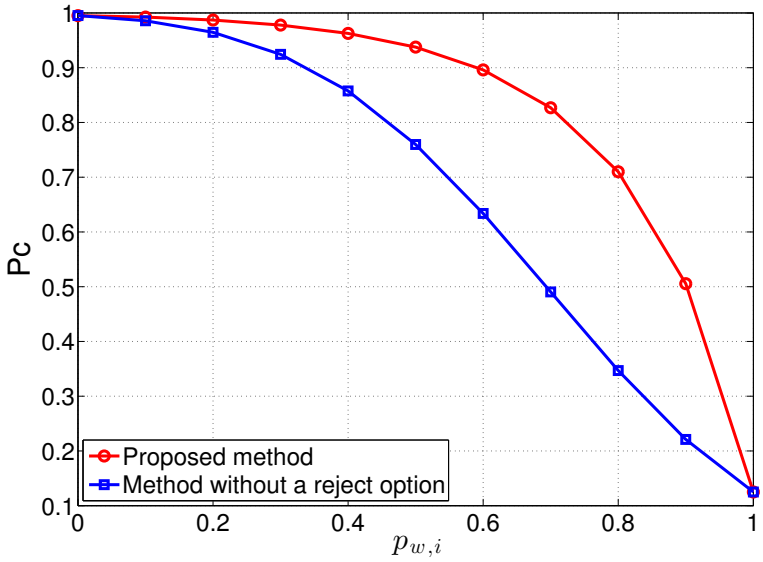


Fig. 2.2: Proposed approach compared to majority voting at $r_{w,i} = 0.8$.

This counterintuitive result follows since even though the crowd skips more microtasks, the optimized weights take advantage of the number of unanswered questions and extract more information. For this to happen, the number of microtasks N has to be greater than a lower limit. Since a larger N induces more diversity in the number of unanswered questions, the existence of the lower limit means that this diversity can actually benefit the performance using the proposed scheme.

2.3.3 Simulation Results

In this subsection, we compare the performance of the proposed crowdsourcing system where crowd workers are allowed to skip microtasks with the conventional majority voting method in a hard-decision fashion, which means that workers are forced to make a decision even if the workers believe that no definitive answers could be provided. The number of equiprobable classes is set as $M = 8$.

Fig. 2.2 compares performance when $W = 20$ workers take part in the task. We consider here that workers have a fixed $p_{w,i}$ for each microtask and $r_{w,i} = 0.8$. We observe that performance degrades as $p_{w,i}$ gets larger, i.e. the workers have a higher probability of not submitting an answer

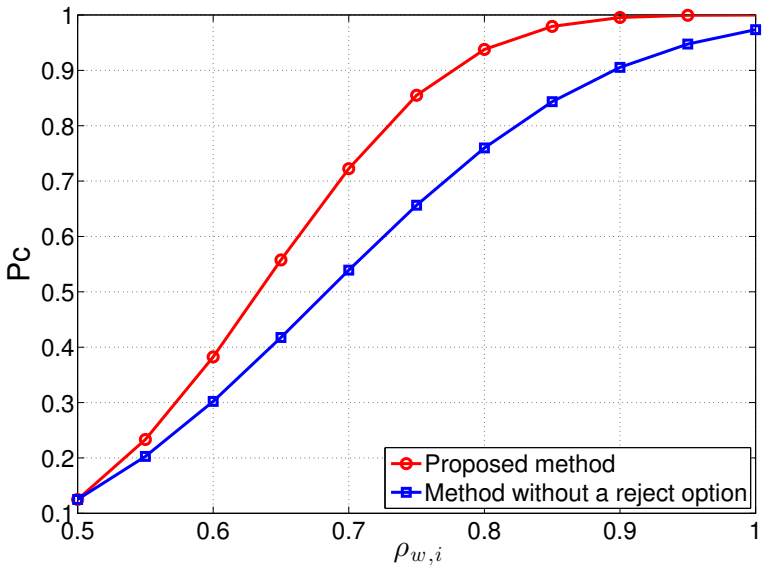


Fig. 2.3: Proposed approach compared to majority voting at $p_{w,i} = 0.5$.

to the microtask. A remarkable performance improvement associated with our proposed approach is observed. The two curves converge at $p_{w,i} = 1$ with P_c being equal to 0.125. At this point, with the majority-based approach, each worker gives random answers for each microtask whereas workers using the proposed scheme skip all the questions and the tie-breaking criterion is used to pick a random bit for every microtask. In Fig. 2.3, we fix $p_{w,i} = 0.5$ and vary $r_{w,i}$ to compare the resulting P_c. Notable performance improvement is also seen. The point at $r_{w,i} = 0.5$ indicates that the worker is making a random guess even if he/she believes that he/she can complete the corresponding microtask correctly. The performance improves as $r_{w,i}$ gets larger, which means that the crowd is able to give higher-quality definitive answers.

In Fig. 2.4, we compare the performance with different number of workers, also showing the asymptotic performance characterization. Here, we consider different qualities of the individuals in the crowd which is represented by variable $p_{w,i}$ with uniform distribution $U(0,1)$ and $r_{w,i}$ with $U(0.6,1)$. First, it is observed that a larger crowd completes the classification task with higher quality. The asymptotic curves are derived under the assumption of a very large crowd, which are the bounds on the performance of the systems. It is not difficult to derive the asymptotic performance for conventional majority voting: $P_c = \left[Q\left(-\sqrt{\frac{W^2(2l-1)}{4l-4l^2}}\right)\right]^N$, where $l = \mu +$

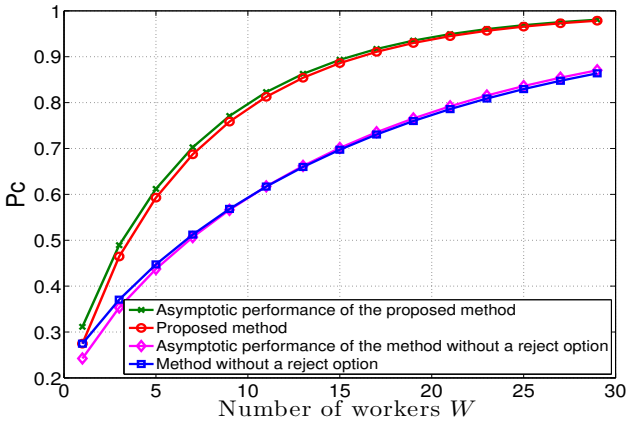
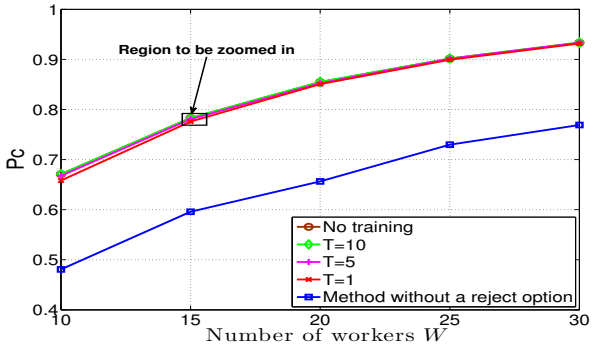
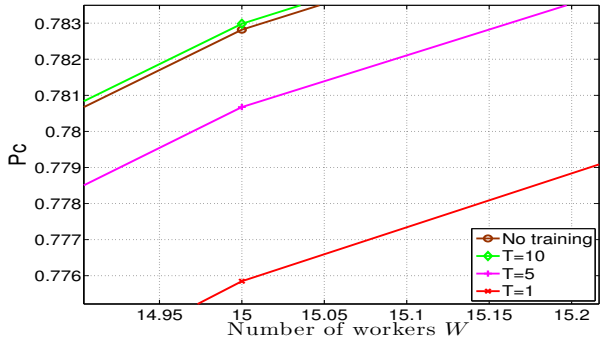


Fig. 2.4: Proposed approach compared to majority voting with varying number of workers at $p_{w,i} \sim U(0,1)$ and $r_{w,i} \sim U(0.6,1)$.



(a) Performance comparison.



(b) Zoomed-in version

Fig. 2.5: Proposed approach compared to majority voting with varying number of workers at $p_{w,i} \sim U(0,1)$ and $r \sim U(0.5,1)$. Two methods are used to estimate μ for weight assignment. One uses training to insert T additional microtasks for estimation, whereas the other one uses the decision results of majority voting as a benchmark to estimate μ. (a) provides the performance comparison while (b) is a zoomed-in region which is indicated in the box in (a).

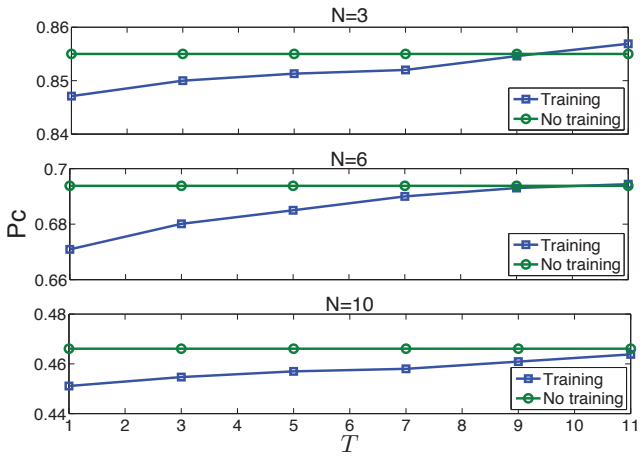


Fig. 2.6: Performance vs. overhead tradeoff. The crowd size is set as $W = 20$ and $N = 3$, 6, and 10, from top to bottom, respectively.

$m(0.5 - \mu)$. Therefore, the asymptotic gap in the performance between conventional majority voting and the proposed method can also be obtained. We find that the asymptotic curves are quite a tight match to the actual performance. Again, we can see a significant improvement in P_c brought by the proposed approach.

We now include the estimation of μ in Fig. 2.5 for weight assignment. Observe in Fig. 2.5(a) that the performance improves as the number of workers increases and that the proposed approach is significantly better than majority voting. Second, the performance of the proposed approach is significantly better than that of traditional majority voting, and it changes for different estimation settings. As is expected, a larger number of training questions result in better system performance as observed in Fig. 2.5(b). Another interesting finding is that training-based estimation can outperform majority voting only when a relatively large number of training questions are used. We see from the figure that the training method with $T = 10$ slightly outperforms the Majority-voting method (without training). However, the number of microtasks N is only 3, which is much smaller than the training size. Much extra overhead beyond the classification task must be added if the training method is adopted. Hence, it is reasonable to employ the Majority-voting method together with the proposed approach for classification using crowdsourcing.

Fig. 2.6 shows P_c performance as a function of overhead with different numbers of microtasks,

to illustrate the performance gap between the two methods. Observe that the method without training exhibits remarkable advantage when a reasonable number of additional microtasks are inserted. Improved performance of the training-based method is shown when T gets larger as this results in more accurate estimation. To have comparable performance with the method without training, the training-based method requires even more additional microtasks when the original number of microtasks N increases. With enough microtasks inserted, the training method can outperform the one without. Again, this result encourages employing the method without training in practice.

2.4 Crowdsourcing with Confidence Reporting

We consider the case where the crowd is composed of honest workers, which means that the workers honestly observe, think, and answer the questions, give confidence levels, and skip questions that they are not confident about. We derive the optimal weight assignment for the workers and the performance of the system in a closed form. Based on these findings, we determine the potential benefits of confidence reporting in a crowdsourcing system with a reject option.

2.4.1 Confidence Level Reporting

In a crowdsourcing system where workers submit answers and report confidence, we define the wth worker's confidence about the answer to the ith microtask as the probability of this answer being correct given that this worker gives a definitive answer, which is equal to $\rho_{w,i}$ as defined earlier. When $\rho_{w,i}$ is bounded as $\frac{l_{w,i}-1}{L} \leq \rho_{w,i} \leq \frac{l_{w,i}}{L}$, $l_{w,i} \in \{1, \ldots, L\}$, the wth worker reports his/her confidence level as $l_{w,i}$. Let $l_{w,i}$ be drawn from the distribution $l_{w,i} \sim F_L(l)$. Note that every worker independently gives confidence levels for different microtasks, and $L = 1$ simply means that workers submit answers and do not report their confidence levels.

Assuming that a worker can accurately perceive the probability $\rho_{w,i}$ and honestly report the confidence level, intuitively it is expected that it will benefit the crowdsourcing fusion center as

much more information about the quality of the crowd can be extracted. However, as the confidence is quantized, which helps the workers in determining the confidence levels to be reported, quantization noise is introduced in extracting the crowd quality from confidence reporting.

As an illustrative example, consider the problem of mismatched crowdsourcing for speech transcription, which has garnered interest in the signal processing community [18, 42, 57, 64, 69, 113]. Suppose the four possibilities for a velar stop consonant to transcribe are $R = \{$क, ख, ग, घ$\}$. The simple binary question of "whether it is aspirated or unaspirated" differentiates between $\{$ख, घ$\}$ and $\{$क, ग$\}$, whereas the binary question of "whether it is voice or unvoiced" differentiates between $\{$ग, घ$\}$ and $\{$क, ख $\}$. The highest confidence level is set as $L = 4$. Now suppose the first worker is a native Italian speaker. Since Italian does not use aspiration, this worker will be unable to differentiate between $\{$क$\}$ and $\{$ख$\}$, or between $\{$ग$\}$ and $\{$घ$\}$. It would be of benefit if this worker would specify the inability to perform the task through a special symbol λ, rather than guessing randomly, and this worker answers "Yes" with confidence level 1 to the second question. Suppose the second worker is a native Bengali speaker. Since this language makes a four-way distinction among velar stops, such a worker will probably answer both questions without a λ.

In the rest of this section, we address the problem "Does the confidence reporting help crowdsourcing system performance?" by performing analyses when workers report their confidences with their definitive answers.

2.4.2 Optimal Weight Assignment Scheme

We determine the optimal weight W_w for the wth worker in this section. We rewrite hereby the weight assignment problem

$$\begin{aligned} &\text{maximize} \;\; E_C\left[\mathbb{W}\right] \\ &\text{subject to} \;\; E_O\left[\mathbb{W}\right] = K \end{aligned} \tag{2.28}$$

where $E_C\left[\mathbb{W}\right]$ denotes the crowd's average weight contribution to the correct class and $E_O\left[\mathbb{W}\right]$ denotes the average weight contribution to all the possible classes and remains a constant K. Statistically, we are looking for the weight assignment scheme such that the weight contribution to

the correct class is maximized while the weight contribution to all the classes remains fixed, so as to maximize the probability of correct classification.

Proposition 2.7. *To maximize the average weight assigned to the correct classification element, the weight for wth worker's answer is given by*

$$W_w = \mu^{-n}, \tag{2.29}$$

where n is the number of definitive answers that the wth worker submits.

Proof. Same as Proposition 1. □

Remark 2.1. *Here the weight depends on the number of questions answered by a worker. In fact, if more questions are answered, the weight assigned to the corresponding worker's answer is larger. This is intuitively pleasing as a high-quality worker is able to answer more questions and is assigned a higher weight. Increased weight can put more emphasis on the contribution of high-quality workers in that sense and improve overall classification performance.*

Remark 2.2. *When $L = \infty$, $\rho_{w,i}$ associated with every worker for every microtask is reported exactly. Then the Chair-Varshney rule gives the optimal weight assignment to minimize error probability [13]. However, human decision makers are limited in their information processing capacity and can only carry around seven categories [75]. Thus, the largest value of L is around 7 in practice.*

Remark 2.3. *Note that the optimal weight assignment scheme is the same as in the case where the workers do not report confidence levels, i.e., $L = 1$. Actually, the value of L does not play any role in the weight assignment, as long as $\rho_{w,i}$ is not known exactly. Therefore, the weight assignment is universally optimal regardless of confidence reporting.*

2.4.3 Parameter Estimation

Before the proposed aggregation rule can be used, μ has to be estimated to assign the weight for every worker's answers. Here, we employ three approaches to estimate μ. We refer to the previous section for training-based and majority-voting based methods to estimate μ, and give an additional method using the information extracted from the workers' reported confidence levels.

Confidence-based Approach

Note that the reported confidence levels correspond to $\rho_{w,i}$. We collect all the values of the submitted confidence levels and obtain the estimate of μ from them. First, the wth worker's confidence level for the ith microtask is represented by $l_{w,i}$. Considering the fact that $\frac{l_{w,i}-1}{L} \leq \rho_{w,i} \leq \frac{l_{w,i}}{L}$ if the worker submits a definitive answer, we use $\frac{l_{w,i}-\frac{1}{2}}{L}$ to approximate $\rho_{w,i}$. Let $l_{w,i} = \frac{1}{2}$ if the wth worker skips the ith microtask. We obtain the estimate of μ by

$$\hat{\mu} = \frac{1}{W-\epsilon} \sum_{w=1}^{W} \sum_{i=1}^{N} \frac{l_{w,i} - \frac{1}{2}}{LI(w)}, \tag{2.30}$$

where $I(w)$ denotes the number of definitive answers that wth worker submits.

2.4.4 Performance Analysis

In this section, we characterize the performance of the proposed crowdsourcing classification framework in terms of the probability of correct classification P_c. Note that we have overall correct classification only when all the bits are classified correctly.

Proposition 2.8. *The probability of correct classification P_c in the crowdsourcing system is*

$$P_c = \Big[\frac{1}{2} + \frac{1}{2}\sum_{S} \binom{W}{\mathbb{Q}} \left(F\left(\mathbb{Q}\right) - F'\left(\mathbb{Q}\right)\right) + \frac{1}{4}\sum_{S'} \binom{W}{\mathbb{Q}} \left(F\left(\mathbb{Q}\right) - F'\left(\mathbb{Q}\right)\right)\Big]^N, \tag{2.31}$$

where $\mathbb{Q} = \left\{(q_{-N}, q_{-N+1}, \ldots q_N) : \sum_{n=-N}^{N} q_n = W\right\}$ *with natural numbers q_n and q_0, and* $S =$

$\left\{\mathbb{Q}:\sum_{n=1}^{N}\mu^{-n}\left(q_n-q_{-n}\right)>0\right\}$, $S'=\left\{\mathbb{Q}:\sum_{n=1}^{N}\mu^{-n}\left(q_n-q_{-n}\right)=0\right\}$, $\binom{W}{\mathbb{Q}}=\frac{W!}{\prod_{n=-N}^{N}q_n!}$, *and*

$$F(\mathbb{Q})=m^{q_0}\prod_{n=1}^{N}(1-\mu)^{q_{-n}}\mu^{q_n}\left(C_{N-1}^{n-1}(1-m)^n m^{N-n}\right)^{q_{-n}+q_n}$$

$$F'(\mathbb{Q})=m^{q_0}\prod_{n=1}^{N}(1-\mu)^{q_n}\mu^{q_{-n}}\left(C_{N-1}^{n-1}(1-m)^n m^{N-n}\right)^{q_{-n}+q_n}.$$

Proof. The proof is similar to the proof in the previous section and is, therefore, omitted for brevity. □

2.5 Simulation Results

In this section, we give the simulation results for the proposed crowdsourcing system. The workers take part in a classification task of $N=3$ microtasks. $F_P(p)$ is a uniform distribution denoted as $U(0,1)$.

Since an accurate estimation of μ is essential for applying the optimal weight assignment scheme, we focus on the estimation results of μ for the three estimation methods as discussed in the previous section. Let $F_\rho(\rho)$ be a uniform distribution expressed as $U(x,1)$ with $0\leq x\leq 1$, and thus we can have μ varying from 0.5 to 1. We consider that $W=20$ workers participate in the classification task with a reject option and confidence reporting.

In Fig. 3.7(a), it is observed that the training-based method has the best overall performance, which takes advantage of the gold standard questions. We can also see that the majority voting method has better performance as μ increases. This is because a larger μ means a better-quality crowd, which will lead to a more accurate result from majority voting, and consequently better estimation performance of μ. When confidence is considered with $L=4$, we find that the overall estimation performance is not better than the other two methods because of quantization noise

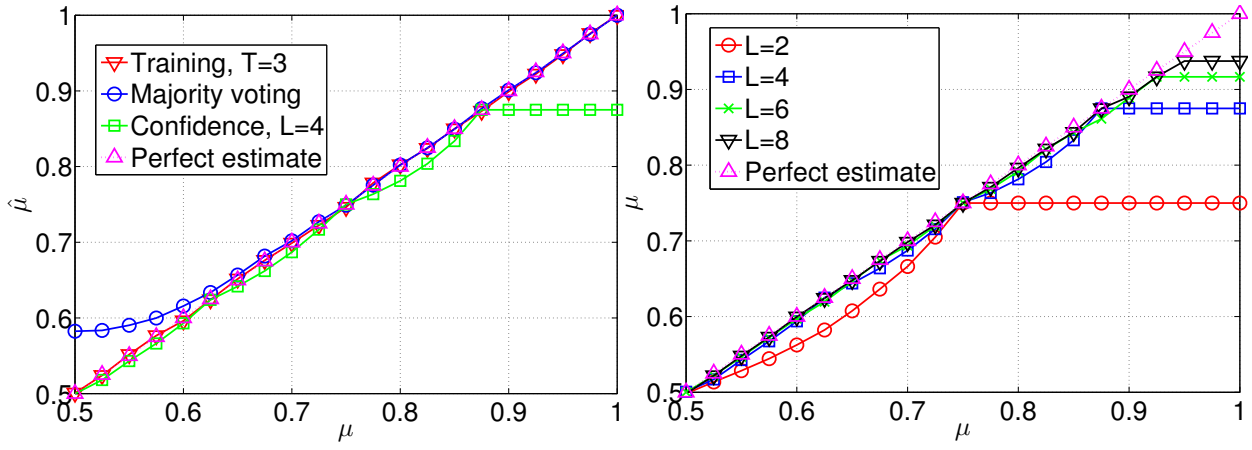


Fig. 2.7: Estimation performance comparison. (a) Different methods. (b) Confidence-based method with different confidence levels.

associated with confidence reporting in the estimation of μ. It is also shown that the curve saturates and yields a fixed value of $\hat{\mu} = 0.875$ when $\mu \geq 0.9$. This is because almost all the confidence levels submitted then are $l_{w,i} = 4$ and the corresponding estimate result is exactly 0.875.

The estimation performance of the confidence-based method with multiple confidence levels is presented in Fig. 3.7(b). As is expected, a larger L can help improve the estimation performance. However, it is seen that even though $L = 8$, the corresponding performance is still not as good as that of the other two methods. Although we can expect estimation performance improvement as the maximum number of confidence levels L increases, $L = 8$ is pretty much the limit in practice due to the human inability to categorize beyond 7 levels. When the confidence-based estimation method is employed, the estimate value saturates at a certain fixed value when μ is large. Therefore, it can be concluded that the confidence-based estimation method does not provide good results.

Even though the three methods differ in performance in the estimation of μ, we show the robustness of the proposed system. in Fig. 2.8. We observe from Fig. 2.8 that the majority voting based method suffers from performance degradation in the low-μ regime, while the confidence based one suffers in the high-μ regime. However, when the value of μ is low, the workers are making random guesses even when they believe that they are able to respond with definitive answers. When the value of μ is large, almost all the definitive answers submitted are correct. Therefore, in

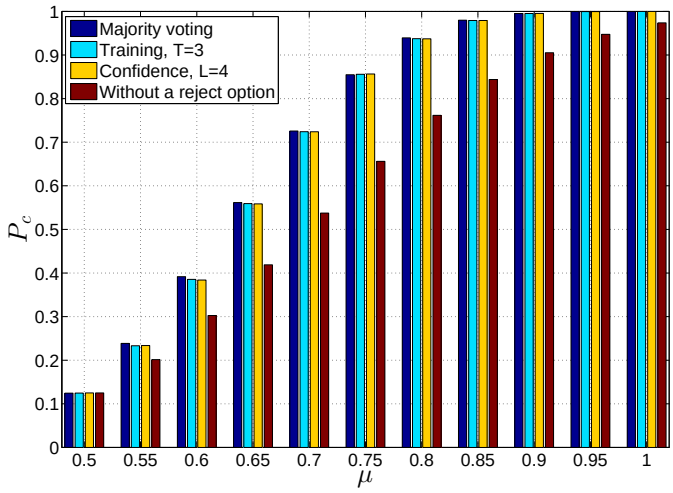


Fig. 2.8: Robustness of the proposed system and performance comparison with simple majority voting

those two situations, the performance degradation in the estimation of μ is negligible. From Fig. 2.8, we see that system performance of the proposed system with estimation results from Fig. 2.8 is almost the same as with the other three estimation methods, which significantly outperforms the system where simple majority voting is employed without a reject option. However, if a significant performance degradation in the estimation of μ occurs outside the two aforementioned regimes, overall classification performance loss is expected. For example, consider the case where μ is 0.8 while $\hat{\mu}$ is 0.5, and $N = 5$, then $P_c = 0.8$. However, the actual P_c equals 0.89 when μ is estimated with an acceptable error.

2.6 Summary

We have studied a novel framework for crowdsourcing of classification tasks that arise in human-based signal processing, where an individual worker has the reject option and can skip a microtask if he/she has no definitive answer. We presented an aggregation approach using a weighted majority voting rule, where each worker's response is assigned an optimized weight to maximize the crowd's classification performance. We have shown that our proposed approach significantly outperforms traditional majority voting and provided asymptotic characterization as an upper bound

on performance. We showed that reporting of confidence by the crowd does not benefit classification performance. One is advised to adopt the reject option without confidence indication from the workers as it does not improve classification performance and may degrade performance in some cases.

CHAPTER 3
CLASSIFICATION IN CROWDSOURCING: UNRELIABLE AGENTS

3.1 Introduction

Note that under typical crowdsourcing incentive schemes based on work volume, workers may respond to tasks for which they are not sufficiently skilled, even when a reject option is available. This chapter extends the work in the previous chapter by further considering the spammers' impact on the system. First, we consider the case where the spammers believe that responding to more microtasks would result in more rewards (payment). Equivalently, we study the case where a fraction of the anonymous crowd workers that are greedy, i.e., spammers, complete all the microtasks with random guesses to maximize their rewards. A heuristic adaptive approach is proposed by switching between oblivious and expurgation strategies, based on the estimation of several crowd parameters such as fraction of greedy workers. In the oblivious strategy, we weight the crowd workers' response using the aggregation rule derived in the previous chapter. In the expurgation strategy, responses of workers that respond to all microtasks are discarded and the remaining workers' responses are assigned optimized weights.

Next, we study the spammer's optimal behavior to maximize its monetary reward based on the

payment mechanism proposed in [100], which is the only mechanism that satisfies the "no-free-lunch" rule and is also incentive-compatible. We find that the spammers should either complete or skip all the microtasks in order to get the maximal reward in an average sense. The statistical properties of the crowd determine whether the spammers should complete or skip. The spammers behave optimally to maximize their monetary reward and the manager employs the optimal weight assignment scheme for aggregation. To combat the optimal behavior of the spammers in the crowd, we also design an optimal aggregation rule where the workers are assigned optimal weights. We give methods for estimating several crowd parameters that are used for weight assignment.

Although the contributions listed above are stated for the crowdsourcing paradigm, our results hold for other signal classification tasks when decisions are made using signals that are quite uncertain. This is known as classification with a reject option [43] and has been the focus of several recent studies in signal processing research including pattern recognition, image, and speech classification [4, 19, 21, 86, 111].

3.2 System With Greedy Crowd Workers

In Sec. 2.3, we considered conscientious crowd workers who respond only when having confidence in their ability to respond. In our formulation, the weight assigned to a worker's response increases with the number of definitive answers and this contributes to the selection of the correct class. In a reward-based system, such honest workers should be compensated and actually rewarded for completing as many tasks as possible. Typical crowdsourcing setups do in fact pay workers in proportion to the number of microtasks they complete [53, 118, 119]. However, if there are workers that try to get as much reward as possible without regard to the system goal of classification, such a reward mechanism can encourage these greedy workers to randomly complete all microtasks (without regard to the question being asked). These greedy workers who degrade system classification performance are often termed *spammers* and are known to exist in large numbers on crowdsourcing platforms [53, 118, 119]. Indeed, Mason and Watts observed that increasing

financial incentives increases the quantity of work performed, but not necessarily the quality [72]. Note that the semi-greedy situation where some workers occasionally complete microtasks that they are not confident about can be considered as a special case of our model with $p_{w,i} = 0$ and $r_{w,i} = 1/2$. Thus our model to characterize the greedy behavior of workers is general.

In this section, we study system performance when a part of the crowd completes all microtasks with random guesses. In other words, these greedy workers submit N-bit codewords, termed as *full-length answers*. Note that the semi-greedy situation where some workers occasionally complete microtasks that they are not confident about can be characterized by small values of $p_{w,i}$ and $r_{w,i}$.

Insertion of a gold standard question set is a common method to address the issue of greedy workers, but comes at the cost of a large question set to avoid workers spotting recurrent questions [27]. Besides, this is wasteful since the fundamental reason for crowdsourcing is to collect classified labels that we do not have [55]. We, therefore, study two different strategies besides inserting a gold standard question set. The *Oblivious Strategy* continues to use the scheme from Sec. 2.3, ignoring the existence of greedy workers. In the *Expurgation Strategy*, we discard the answers of workers who only give full-length answers, to reduce the impact of greedy workers on the overall system performance. Note that this strategy will also discard the responses of those honest workers that provided definitive answers to all microtasks. Also note that greedy workers are not being punished in any way here; only that their responses are being ignored in the aggregation strategy. Let α be the fraction of greedy workers in the crowd.

3.2.1 Oblivious Strategy

In this strategy, we continue to use the same weight allocation scheme as for honest workers: $W_w = \alpha_1 \mu^{-n}$, where the factor α_1 is introduced to satisfy the constraint $E_O[\mathbb{W}] = K$.

The average contribution from the crowd to the correct class and all the classes can be given

respectively as

$$\begin{aligned} E_C\left[\mathbb{W}\right] =& \sum_{w=1}^{W\alpha} \alpha_1 \mu^{-N} \frac{1}{2^N} \\ &+ \sum_{w=W\alpha+1}^{W} \sum_{n=0}^{N-1} \alpha_1 \mu^{-n} C_N^n \left[\left(1-m\right)\mu\right]^n m^{N-n} \\ =& \frac{W\alpha\alpha_1}{\left(2\mu\right)^N} + \alpha_1 W\left(1-\alpha\right), \end{aligned} \tag{3.1}$$

and

$$\begin{aligned} E_O\left[\mathbb{W}\right] =& \sum_{w=1}^{W\alpha} \alpha_1 \left(\tfrac{1}{\mu}\right)^N \\ &+ \sum_{w=W\alpha+1}^{W} \sum_{n=0}^{N-1} \alpha_1 \mu^{-n} 2^{N-n} C_N^n (1-m)^n m^{N-n} \\ =& W\alpha\alpha_1 \left(\tfrac{1}{\mu}\right)^N + \left(W - W\alpha\right)\alpha_1 \left(\tfrac{1-m}{\mu} + 2m\right)^N. \end{aligned} \tag{3.2}$$

Therefore, we can calculate α_1 and obtain $E_C[\mathbb{W}]$ as:

$$\alpha_1 = \frac{K}{W\alpha\left(\tfrac{1}{\mu}\right)^N + \left(W - W\alpha\right)\left(\tfrac{1-m}{\mu} + 2m\right)^N}, \tag{3.3}$$

$$E_C\left[\mathbb{W}\right] = \frac{K\alpha\left(\tfrac{1}{2\mu}\right)^N + K\left(1-\alpha\right)}{\alpha\left(\tfrac{1}{\mu}\right)^N + \left(1-\alpha\right)\left(\tfrac{1-m}{\mu} + 2m\right)^N}. \tag{3.4}$$

Proposition 3.1. *The probability of correct classification P_c when the Oblivious Strategy is used*

is

$$P_c = \Big[\frac{1}{2} + \frac{1}{2}\sum_{S_1}\binom{W}{\mathbb{Q}_1}\left(F(\mathbb{Q}_1) - F'(\mathbb{Q}_1)\right) + \frac{1}{4}\sum_{S_1'}\binom{W}{\mathbb{Q}_1}\left(F(\mathbb{Q}_1) - F'(\mathbb{Q}_1)\right)\Big]^N \tag{3.5}$$

with

$$F(\mathbb{Q}_1) = \frac{m^{q_0}}{2^{W\alpha}}\prod_{n=1}^{N}(1-\mu)^{q_{-n}}\mu^{q_n}\left(C_{N-1}^{n-1}(1-m)^n m^{N-n}\right)^{q_{-n}+q_n} \tag{3.6}$$

and

$$F'(\mathbb{Q}_1) = \frac{m^{q_0}}{2^{W\alpha}}\prod_{n=1}^{N}(1-\mu)^{q_n}\mu^{q_{-n}}\left(C_{N-1}^{n-1}(1-m)^n m^{N-n}\right)^{q_{-n}+q_n} \tag{3.7}$$

where

$$\mathbb{Q}_1 = \left\{(q_{-N}, q_{-N+1}, \ldots q_N) : \sum_{n=-N}^{N} q_n = W - W\alpha\right\}, \tag{3.8}$$

with natural numbers q_n*,*

$$S_1 = \left\{\mathbb{Q}_1 : \sum_{n=1}^{N}\mu^{-n}\left(q_n - q_{-n}\right) + \mu^{-N}W\alpha > 0\right\}, \tag{3.9}$$

$$S_1{}' = \left\{\mathbb{Q}_1 : \sum_{n=1}^{N}\mu^{-n}\left(q_n - q_{-n}\right) + \mu^{-N}W\alpha = 0\right\}, \tag{3.10}$$

and $\binom{W}{\mathbb{Q}_1} = W!/\prod_{n=-N}^{N} q_n!$.

Proof. See Appendix A.6. □

3.2.2 Expurgation Strategy

In this strategy, all definitive answers of length of N bits are discarded to avoid answers from greedy workers. The classification decision is made based on the answers with maximum number of bits equal to $N-1$. To proceed, we need the weight for every worker's answer in this case. We begin by restating the optimization problem:

$$\begin{aligned} &\text{maximize} \;\; E_C\left[\mathbb{W}\right] \\ &\text{subject to} \;\; E_O\left[\mathbb{W}\right] = K \end{aligned} \tag{3.11}$$

and we have

$$\begin{aligned} E_C\left[\mathbb{W}\right] &= \sum_{w=1}^{W-W\alpha}\sum_{n=0}^{N-1} W_w \binom{N}{n} \left[\left(1-m\right)\mu\right]^n m^{N-n} \\ &= \sum_{w=1}^{W-W\alpha}\sum_{n=0}^{N-1} W_w \mu^n x^{n-N} P_x\left(n\right), \end{aligned} \tag{3.12}$$

where

$$P_x\left(n\right) = \binom{N}{n}\left(1-m\right)^n \left(mx\right)^{N-n}, \tag{3.13}$$

and x is such that

$$\sum_{n=0}^{N-1} P_x\left(n\right) = 1. \tag{3.14}$$

Then, we can write

$$E_C\left[\mathbb{W}\right] \leq \sum_{w=1}^{W-W\alpha} \sqrt{\sum_{n=0}^{N-1} \left(W_w \mu^n x^{n-N}\right)^2 P_x\left(n\right)} \sqrt{\sum_{n=0}^{N-1} P_x\left(n\right)}, \tag{3.15}$$

which holds with equality only if

$$W_w \mu^n x^{n-N} \sqrt{P_x(n)} = \alpha_2 \sqrt{P_x(n)}, \tag{3.16}$$

where the factor α_2 is introduced to satisfy the constraint $E_O[\mathbb{W}] = K$. Hence, we have the maximum of $E_C[W_w]$ as

$$E_C[\mathbb{W}] = W(1-\alpha)\alpha_2, \tag{3.17}$$

when

$$W_w = \alpha_2 \mu^{-n} x^{N-n}. \tag{3.18}$$

To obtain the value of x, we rewrite (3.14) as:

$$(1-m+mx)^N - (1-m)^N = 1, \tag{3.19}$$

and x is given as

$$x = \frac{\left(1+(1-m)^N\right)^{\frac{1}{N}} + m - 1}{m}. \tag{3.20}$$

For this strategy, the overall weight constraint is given as

$$\begin{aligned} E_O[\mathbb{W}] &= \sum_{w=1}^{W-W\alpha} \sum_{n=0}^{N-1} \alpha_2 \mu^{-n} x^{N-n} 2^{N-n} \binom{N}{n} (1-m)^n m^{N-n} \\ &= W(1-\alpha)\alpha_2 \left[\left(\frac{1-m}{\mu} + 2mx \right)^N - \left(\frac{1-m}{\mu} \right)^N \right]. \end{aligned} \tag{3.21}$$

By substituting this result back into (3.17), the maximum value of $E_C[\mathbb{W}]$ can be written as

$$E_C\left[\mathbb{W}\right] = \frac{K}{\left(\frac{1-m}{\mu} + 2mx\right)^N - \left(\frac{1-m}{\mu}\right)^N}. \tag{3.22}$$

Note that the weight could be $W_w = \mu^{-n}x^{-n}$ when the Expurgation Strategy is employed in practice, where x is given by (3.20).

Proposition 3.2. *The probability of correct classification P_c when the Expurgation Strategy is used is*

$$\begin{aligned} P_c = \Big[\frac{1}{2} + \frac{1}{2}\sum_{S_2}\binom{W}{\mathbb{Q}_2}\left(F(\mathbb{Q}_2) - F'(\mathbb{Q}_2)\right) \\ + \frac{1}{4}\sum_{S_2'}\binom{W}{\mathbb{Q}_2}\left(F(\mathbb{Q}_2) - F'(\mathbb{Q}_2)\right)\Big]^N \end{aligned} \tag{3.23}$$

with

$$F(\mathbb{Q}_2) = m^{q_0}\prod_{n=1}^{N-1}(1-\mu)^{q_{-n}}\mu^{q_n}\left(C_{N-1}^{n-1}(1-m)^n m^{N-n}\right)^{q_{-n}+q_n}$$

and

$$F'(\mathbb{Q}_2) = m^{q_0}\prod_{n=1}^{N-1}(1-\mu)^{q_n}\mu^{q_{-n}}\left(C_{N-1}^{n-1}(1-m)^n m^{N-n}\right)^{q_{-n}+q_n},$$

where

$$\mathbb{Q}_2 = \left\{(q_{-N+1}, q_{-N+2}, \ldots q_{N-1}) : \sum_{n=-N+1}^{N-1} q_n \leq W - W\alpha\right\}$$

with natural numbers q_n, and

$$S_2 = \left\{\mathbb{Q}_2 : \sum_{n=1}^{N-1}\mu^{-n}x^{-n}\left(q_n - q_{-n}\right) > 0\right\},$$

$$S_2{}' = \left\{\mathbb{Q}_2 : \sum_{n=1}^{N}\mu^{-n}x^{-n}\left(q_n - q_{-n}\right) = 0\right\}$$

and $\binom{W}{\mathbb{Q}_2} = W!/\prod_{n=-N+1}^{N-1} q_n!$.

Proof. See Appendix A.6. □

3.2.3 Adaptive Algorithm

We now investigate the adaptive use of our two strategies to improve system performance. The goal is to find a threshold to determine when one strategy will outperform the other, so as to allow switching.

Note that the two strategies are associated with the same overall weight for all classes. Thus, we compare the crowd's total contribution to the correct class under this condition and derive the corresponding switching scheme. From (3.4) and (3.17), this can be expressed in (3.24),

$$\frac{\alpha K\left(\frac{1}{2\mu}\right)^N + K(1-\alpha)}{\alpha\left(\frac{1}{\mu}\right)^N + (1-\alpha)\left(\frac{1-m}{\mu}+2m\right)^N} \underset{\text{Expurgation Strategy}}{\overset{\text{Oblivious Strategy}}{\gtrless}} \frac{K}{\left(\frac{1-m}{\mu}+2mx\right)^N - \left(\frac{1-m}{\mu}\right)^N}, \tag{3.24}$$

which simplifies to having the switching threshold of α as

$$\alpha\left(\left(\frac{1}{\mu}\right)^N - \gamma_1\left(\frac{1}{2\mu}\right)^N - \gamma_2 + \gamma_1\right) \underset{\text{Oblivious Strategy}}{\overset{\text{Expurgation Strategy}}{\gtrless}} \gamma_1 - \gamma_2, \tag{3.25}$$

where $\gamma_1 = \left(\frac{1-m}{\mu}+2mx\right)^N - \left(\frac{1-m}{\mu}\right)^N$, and $\gamma_2 = \left(\frac{1-m}{\mu}+2m\right)^N$.

To obtain the threshold associated with α in the switching criterion (3.25), μ and m should be estimated first. The previous chapter established a simple and effective method to estimate μ based on majority voting. Therefore, we again use majority voting to get initial detection results, which are then set as the benchmark to estimate μ. Note that estimation of μ is based on the answers without the full-length ones to avoid degradation from the greedy workers.

The performance of this integrated scheme can be derived using Props. 3.1 and 3.2, and switching criterion (3.25).

3.2.4 Joint Estimation of m and α

The threshold on α is specified based on the estimated values of m and μ. Then, we estimate α and compare it with the corresponding threshold to switch the strategies adaptively. Though we discard full-length answers and use the rest to estimate m, it is an inaccurate estimate because the discarded answers also contain those from honest workers.

Several works have studied the estimation of α in crowdsourcing systems [1, 44, 131], which can be divided into two categories: one studies the behavior of the workers in comparison to the honest control group [44]; the other one learns worker's reputation profile [1, 131], which is stored and updated over time to identify the greedy ones from the crowd. However, neither estimation method is suitable here due to the anonymous nature of crowd workers. The first category suffers from the difficulty in extracting the honest group from the anonymous crowd while the second requires identification of every worker.

Since the worker's quality is assumed to be i.i.d., we give a joint parametric estimation method of both m and α based on maximum likelihood estimation (MLE).

As defined earlier, out of W workers, $q_n + q_{-n}$ workers submit answers of n bits, $0 \leq n \leq N$. Thus, the probability mass function of the number of submitted answers given m and α is obtained as,

$$f\left(q_n + q_{-n} | m, \alpha\right) = \begin{cases} \binom{W-W\alpha}{q_n+q_{-n}} A_{N,n,m}{}^{q_n+q_{-n}} \left(1 - A_{N,n,m}\right)^{W-W\alpha-q_n-q_{-n}}, 0 \leq n < N \\ \binom{W-W\alpha}{q_N+q_{-N}-W\alpha} (1-m)^{N(q_N+q_{-N}-W\alpha)} \left(1-(1-m)^N\right), n = N \end{cases} \tag{3.26}$$

where $A_{N,n,m} = \binom{N}{n}(1-m)^n m^{N-n}$, is defined as the expectation of the probability of a single worker submitting n definitive answers.

Because of the independence of workers, we can form the likelihood statistic as

$$L\left(m, \alpha\right) = \sum_{n=0}^{N} \log f\left(q_n + q_{-n} | m, \alpha\right). \tag{3.27}$$

Table 3.1: Estimation of α

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\hat{\alpha}$	0.11	0.26	0.36	0.48	0.58	0.67	0.79	0.87	0.96

Therefore, the ML estimates of m and α, which are denoted by $\hat{m}$ and $\hat{\alpha}$, can be obtained as

$$\{\hat{m}, \hat{\alpha}\} = \arg \max_{\{m,\alpha\} \in [0,1]} L(m, \alpha). \tag{3.28}$$

Once we have $\hat{\mu}$, $\hat{m}$ and $\hat{\alpha}$, we can adaptively switch to the suitable strategy using (3.25).

3.2.5 Simulation Results

Now, we present some simulation results to illustrate the performance of our proposed algorithm. First, the theoretical value of the threshold for adaptive switching between the strategies is obtained for different values of m and μ based on (3.25). We switch to the Expurgation Strategy if the fraction of greedy workers α is greater than the threshold. Otherwise we stick to the Oblivious Strategy. As we observe from Fig. 3.1, when m decreases and μ increases—when the quality of the crowd improves—the threshold increases. This implies a crowdsourcing system employing the Oblivious Strategy can tolerate a higher fraction of greedy workers in the crowd; instead of discarding all of the answers from the greedy workers and from those honest workers who submit full-length answers, it is better to keep them as long as the honest ones can perform well. The effect of greedy workers' answers can be compensated by the high-quality answers of the honest workers in the crowd.

Next, we give the estimation results for $\hat{\alpha}$ in Table 3.1 using the proposed MLE method. The crowd quality parameters $p_{w,i}$ and $r_{w,i}$ are drawn from distributions $U(0,1)$ and $U(0.5,1)$ respectively. The number of microtasks N and the number of workers W are set to 3 and 20, respectively.

Fig. 3.2 shows the performance of the proposed adaptive scheme. The system parameters are the same as in previous simulations except that the crowd size W is set to 15. The crowdsourcing system starts with the estimation of the parameters μ, m, and α. Once it has obtained $\hat{\mu}$ and $\hat{m}$,

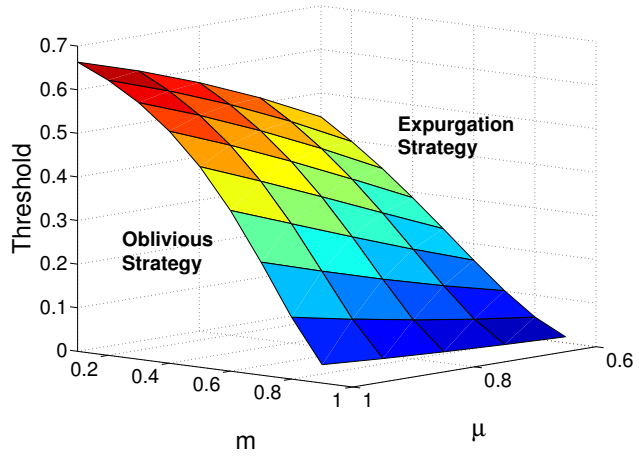


Fig. 3.1: Threshold to switch between strategies.

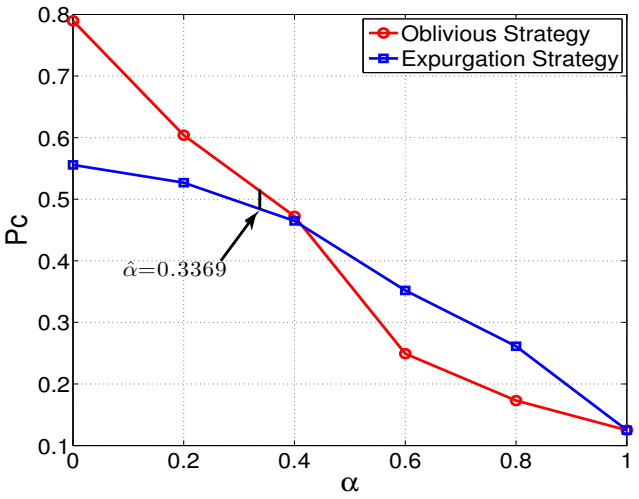


Fig. 3.2: Performance of both the strategies with greedy workers.

the system calculates the threshold value and compares it with $\hat{\alpha}$, and then decides the strategy to be used. Next, the system allocates weights to the answers for aggregation based on the strategy selected and makes the final classification decision. Fig. 3.2 presents the performance of both strategies and the estimated threshold for switching. The performance deterioration caused by greedy workers is quite obvious as the probability of correct classification P_c decreases for both strategies with increasing α. The intersection of curves illustrates the need for strategy switching. The estimated Pareto frontier $\hat{\alpha}$ for switching is 0.3369 in this system setting, which is indicated in the figure by a line segment and is very close to the intersection of the two curves. Therefore, the actual performance curve of the proposed algorithm consists of the curve with squares when $\alpha < \hat{\alpha}$ and curve with circles when $\alpha > \hat{\alpha}$.

3.3 Optimal Behavior of the Spammers and the Manager

In this section, we consider the existence of spammers in the crowd. Spammers are workers who answer randomly without regard to the question being asked, in the hope of earning some free or extra money, and are known to exist in large numbers on crowdsourcing platforms [100]. We study the optimal behavior for the spammers and the manager. First, the one and only incentive-compatible payment mechanism that satisfies the "no-free-lunch" axiom[1] for crowdsourcing with a reject option is adopted. This mechanism makes the smallest possible payment to spammers among all possible incentive-compatible mechanisms that may or may not satisfy the "no-free-lunch" axiom [100]. Based on this mechanism, we investigate the optimal behavior on the spammers' side such that the spammers can maximize their monetary reward. Next, considering the spammers behaving optimally in this manner, we design the optimal weight assignment scheme for aggregation on the manager's side to combat the spammers' effect on the crowdsourcing system performance.

[1]The "no-free-lunch" axiom requires that the payment is minimum possible if all the answers attempted by the worker in the gold standard questions are wrong.

3.3.1 Payment Mechanism

The payment to the worker is based on the evaluation of the answers that the worker gives to the G gold standard questions. The goal of the mechanism is to incentivize the worker to skip the questions for which its confidence is lower than T. The value of T is chosen *a priori* based on factors such as budget constraints or the targeted performance quality. Let f denote the payment rule, which is proposed in [100] and is written as

$$f(x_1, \ldots, x_G) = \kappa \prod_{i=1}^{G} \alpha_{x_i} + \mu_{\min} \tag{3.29}$$

where $x_j \in \{-1, \lambda, +1\}, 1 \leq j \leq G$, are the results of the gold standard questions. "-1" denotes that the worker attempted to answer the microtask and the answer was incorrect, "λ" denotes the worker skipped the microtask, and "$+1$" denotes that the worker attempted to answer the microtask and the answer was correct. Set $\alpha_{-1} = 0, \alpha_{\lambda} = 1, \alpha_{+1} = \frac{1}{T}$, and $\kappa = (\mu_{\max} - \mu_{\min})\, T^G$ with budget parameters $\mu_{\max}$ and $\mu_{\min}$ denoting the maximum and minimum payments respectively. To the workers, the mechanism reads that even one mistake leads to minimum payment so use the reject option wisely.

3.3.2 Optimal Behavior for the Spammers

In this subsection, we give the optimal behavior for the spammers that maximizes their expected monetary reward based on the payment mechanism described in the previous subsection. By optimal behavior, we mean the optimal number of questions to be skipped by the spammers.

Proposition 3.3. *To maximize the expected monetary reward, the optimal behavior for a spammer is that he/she completes all the microtasks if $T < \frac{1}{2}$, and skips all the microtasks otherwise.*

Proof. See Appendix A.7. □

The above proposition addresses the optimal strategy for the spammers to participate in the crowdsourcing task. Since a spammer can not distinguish the gold standard ones from the other

questions, the result derived indicates that the spammers should either complete or skip all the questions according to the value of T to maximize their expected monetary reward.

3.3.3 Optimal Behavior for the Manager

In typical crowdsourcing setups, workers are simply paid in proportion to the number of tasks they complete [100]. Most likely, the spammers will complete all the microtasks in the skip-based setting if they do not optimize their behavior as given in Proposition 3.3. However, even if the spammers know the optimal strategy to maximize the monetary reward, an accurate estimate of T is needed before hand for them to behave wisely, which is almost intractable for the spammers. According to Prospect Theory [59], a Nobel prize winning theory developed by Kahneman and Tversky, real-life decision- making deviates from rational behavior which is uninfluenced by real-life perceptions. People use their subjective probabilities rather than objective probabilities to weigh the values of possible outcomes, and the value of an outcome is determined by considering the relative gains or losses regarding a reference point. Thus, to maximize the monetary reward, the spammers roughly and subjectively evaluate T, and strategically complete or skip all the microtasks based on their own perceptions of the value of T. Consequently, no matter whether the spammers are wise or not, we assume that M_N spammers complete all the N microtasks and the rest of the M_0 spammers skip all the microtasks, making a total of M spammers in the crowd of size W.

The presence of the spammers will significantly affect the classification performance of the crowdsourcing system, which may make it worse when the spammers are starting to act strategically. To combat the spammers' effect on the system performance, we develop the aggregation rule on the manager's side with a new weight assignment scheme to maximize the weight assigned to the correct class in this subsection.

Proposition 3.4. *To maximize the average weight assigned to the correct classification element,*

the weight for the wth worker's answer is given by

$$W_w = \left[(W - M)\,\mu^n + \frac{M_N}{2^N (1-m)^N}\delta\,(n - N) \right]^{-1}, \tag{3.30}$$

where n is the number of definitive answers that the wth worker submits, and $\delta(\cdot)$ is the Dirac delta function.

Proof. See Appendix A.8. □

Compared to the weight assignment for an honest crowd [68], the derived scheme differs in terms of the weight assigned to the workers who complete all the microtasks. If the spammers skip all the microtasks, the weight assignment scheme remains the same, which is intuitively true as no random guesses are received by the manager from the spammers and the crowd can be considered as honest as well. Otherwise, the weight assignment scheme differs from the scheme given in [68].

3.3.4 Parameter Estimation

In order to behave optimally for the manager, several parameters have to be estimated before the weight assignment can be adopted. Specifically, one has to estimate μ, m, M_N, M_0 before he/she can proceed with the weight assignment. We can adopt *Training-based* or *Majority-voting based* method to estimate μ as stated in previous work [68]. Calculating the ratio of the sum of skipped questions over all the questions attempted by the crowd gives the estimated m. Based on the analysis in previous sections, the answers with all questions completed or skipped should be discarded for estimation.

We hereby jointly address the estimation of M_0 and M_N by using the maximum likelihood estimation (MLE) method. First, as we employ G gold standard questions, a worker has to respond to $N+G$ microtasks. Let W_{N+G} denote the number of workers submitting $N+G$ definitive answers, and W_0 denote the number of workers skipping all the microtasks. Given the numbers of spammers respectively completing and skipping all the microtasks, M_N and M_0, the joint probability

distribution function of W_{N+G} and W_0, $f(W_{N+G}, W_0|M_N, M_0)$, is expressed in (3.32), where $\hat{m}$ is the estimated m.

Therefore, by the MLE method, the estimation of M_N and M_0, which are denoted by $\hat{M}_N$ and $\hat{M}_0$ respectively, can be obtained as

$$\left\{\hat{M}_N, \hat{M}_0\right\} = \arg \max_{\{M_N, M_0\} \geq 0} f(W_{N+G}, W_0|M_N, M_0). \tag{3.31}$$

where

$$\begin{aligned} f(W_{N+G}, W_0|M_N, M_0) =& \binom{W_0 - M_0}{W - M_0 - M_N} (\hat{m}^{N+G})^{W_0 - M_0} (1 - \hat{m}^{N+G})^{W - W_0 - M_N} \\ & \cdot \binom{W_{N+G} - M_N}{W - W_0 - M_N} (1 - \hat{m})^{(N+G)(W_{N+G} - M_N)} \left(1 - (1 - \hat{m})^{N+G}\right)^{W - W_{N+G} - W_0} \end{aligned} \tag{3.32}$$

Once the manager has the estimation results $\hat{\mu}$, $\hat{m}$, $\hat{M}_N$, and $\hat{M}_0$, he/she can optimally assign the weight to the workers' answers for aggregation.

3.3.5 Performance Analysis

In this section, we characterize the performance of such a crowdsourcing classification framework, where the task manager behaves optimally, in terms of the probability of correct classification P_c. Note that we have an overall correct classification only when all the bits are classified correctly.

Proposition 3.5. *The probability of correct classification P_c in the crowdsourcing system is*

$$P_c = \left[\frac{1}{2} + \frac{1}{2} \sum_{S} \binom{W}{\mathbb{Q}} \left(F(\mathbb{Q}) - F'(\mathbb{Q})\right) + \frac{1}{4} \sum_{S'} \binom{W}{\mathbb{Q}} \left(F(\mathbb{Q}) - F'(\mathbb{Q})\right)\right]^N \tag{3.33}$$

with

$$F(\mathbb{Q}) = m^{q_0} \prod_{n=1}^{N} (1-\mu)^{q_{-n}} \mu^{q_n} \left(C_{N-1}^{n-1} (1-m)^n m^{N-n}\right)^{q_{-n}+q_n}$$

and

$$F'(\mathbb{Q}) = m^{q_0} \prod_{n=1}^{N} (1-\mu)^{q_n} \mu^{q_{-n}} \left(C_{N-1}^{n-1} (1-m)^n m^{N-n} \right)^{q_{-n}+q_n}$$

where

$$\mathbb{Q} = \{(q_{-N}, q_{-N+1}, \dots q_N, M'_A, M''_A) : \sum_{n=-N}^{N} q_n = W - M_A - M_0, M'_A + M''_A = M_A\},$$

with natural numbers q_n*,* M'_A*, and* M''_A*,*

$$S = \left\{ \mathbb{Q} : \sum_{n=1}^{N} \frac{q_n - q_{-n}}{(W-M)\mu^n} + (M'_A - M''_A) \frac{2^N (1-m)^N}{M_A} > 0 \right\},$$

$$S' = \left\{ \mathbb{Q} : \sum_{n=1}^{N} \frac{q_n - q_{-n}}{(W-M)\mu^n} + (M'_A - M''_A) \frac{2^N (1-m)^N}{M_A} = 0 \right\},$$

and $\binom{W}{\mathbb{Q}} = \frac{W!}{\prod_{n=-N}^{N} q_n!}$.

Proof. See Appendix A.6.

3.3.6 Simulation Results

In this section, we present the simulation results to illustrate the performance of the proposed schemes. $W = 50$ workers participate in a crowdsourcing task with $N = 3$ microtasks and $G = 3$ gold standard questions. $F_P(p)$ is chosen as a uniform distribution $U(0, 1)$, and let $F_\rho(\rho)$ be a uniform distribution expressed as $U(x, 1)$ with $0 \leq x \leq 1$, and thus we can have μ varying from 0.5 to 1.

In Table 3.3, we show the estimation results of M_0 and M_N. Here, μ is set as 0.75. The estimation is based on the distribution of the numbers of workers completing and skipping all

M_0 \ M_N	1	3	5	7	9	11	13	15	17	19
1	1,0	1,3	1,5	1,8	2,9	2,12	1,14	2,15	2,17	2,20
3	3,1	3,2	3,5	4,7	4,9	3,11	3,14	3,15	3,18	3,20
5	5,2	5,3	5,6	6,7	5,9	6,11	6,14	5,17	6,18	5,19
7	7,0	8,4	7,5	8,8	7,10	8,12	7,13	7,17	7,17	8,20
9	9,1	9,4	9,5	10,7	9,9	11,11	9,13	10,15	11,17	9,20
11	11,1	11,5	11,5	12,8	11,6	12,11	11,13	11,16	11,17	12,19
13	13,2	13,6	13,5	14,8	13,9	13,11	14,13	13,16	13,17	14,19
15	15,1	15,3	16,6	16,7	15,9	17,11	15,13	15,15	15,17	15,19
17	17,1	18,4	17,5	17,8	17,9	17,12	18,13	17,16	18,17	18,19
19	20,2	19,2	19,5	19,8	19,9	19,11	19,13	19,16	20,17	21,19
21	21,2	21,3	22,5	21,7	21,9	22,12	21,13	21,15	21,17	21,19
23	23,1	24,3	25,5	23,9	24,9	24,11	25,13	23,16	23,17	23,19
25	26,1	26,3	25,6	25,7	26,9	26,12	25,13	25,15	26,17	25,20

Fig. 3.3: Estimation of M_0 and M_N.

questions W_{N+G} and W_0, and we can see from the table that most pairs of numbers M_0 and M_N can be exactly estimated, and most of the errors are ± 1.

We present the performance comparison with spammers in Fig. 3.4, where the quality of the crowd μ varies. We plot the performance of three different weight assignment methods. The first one is what we derived in this section, which is referred to as the optimal behavior for the manager with spammers. The second is the one that we derived in [67], which is given by $W_w = \mu^{-n}$. Since we do not assume the knowledge of prior information regarding individuals, the existing weighted majority voting methods fail to work in this setting. Thus, we choose the conventional simple majority voting without a reject option for comparison. For illustration, there are 14 spammers in a crowd of 50 workers, and we have 7 spammers completing all the questions and the other 7 skipping all the questions. When $\mu = 0.5$, the workers are making random guesses even if they believe that they are able to respond with definitive answers. In such a case, the choice of weight assignment schemes does not make a difference, and, therefore, the three curves merge at this point. The method with optimal behavior for the manager with spammers outperforms the other

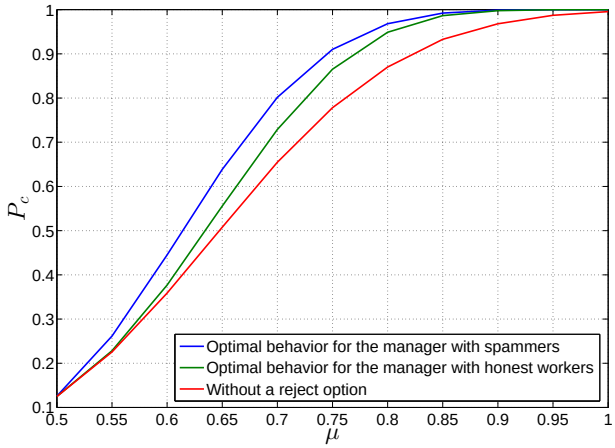


Fig. 3.4: Performance comparison with spammers.

two, while the simple majority voting without a reject option performs the worst.

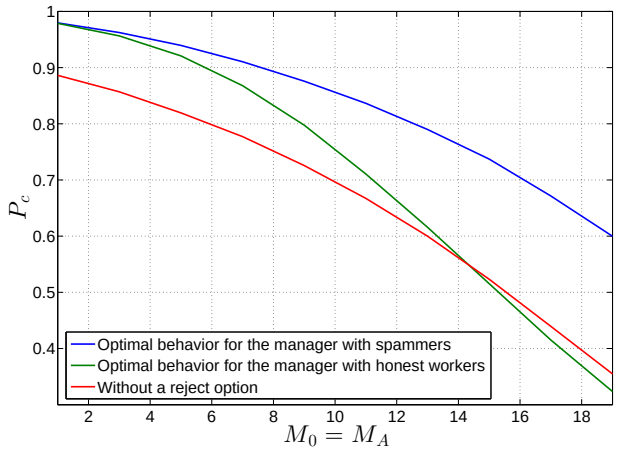


Fig. 3.5: Performance comparison with various spammers.

In Fig. 3.5, we plot the performance comparison when the number of spammers changes. We set that $M_0 = M_A$, and μ is fixed at 0.75. As we can observe, the method with optimal behavior for the manager with spammers yields the best performance. When the number of spammers is small, the simple majority voting method is outperformed by the one with optimal behavior for the manager with honest workers. However, this is not the case when the number of spammers is large. The reason is that with honest workers, the manager assigns a greater weight to the worker with a larger number of definitive answers. In the regime where M_A is large, which means the number of spammers completing all the questions is large, the impact from the spammers is much

more severe on the performance with such a weight assignment scheme. Thus, the corresponding performance degrades significantly.

3.4 Summary

We have studied a novel framework of crowdsourcing system for classification, where an individual worker has the reject option and can skip a microtask if he/she has no definitive answer. We investigated the impact of the spammers in the crowd on the crowdsourcing system performance. First, oblivious and expurgation strategies were studied to deal with the spammers' impact, and an algorithm to adaptively switch between them, based on the estimated fraction of greedy workers in the anonymous crowd, was developed to combat performance degradation. Next, we investigated the case where the spammers can strategically behave to maximize their reward. In such a case, we derived the optimal strategy to aggregation the responses from the crowd workers.

CHAPTER 4
CLASSIFICATION IN SOCIAL NETWORKS: INFLUENTIAL NODE DETECTION

4.1 Introduction

Information emerges dynamically and diffuses quickly via agent interactions in complex networks, e.g. social networks [70]. Consequently, understanding and prediction of information diffusion mechanisms are challenging. There is a rapidly growing interest in exploiting knowledge of the information dynamics to better characterize the factors influencing spread of diseases, planned terrorist attacks, and effective social marketing campaigns, etc [37]. The broad applicability of this problem in social network analysis has led to focused research on the following questions: (I) Which contagions are the most popular and can diffuse the most? (II) Which members of the network are influential and play important roles in the diffusion process? (III) What is the range over which the contagions can diffuse [38]? While attempting to answer these questions, one is confronted with two crucial challenges. First, a descriptive diffusion model, which can mimic the behavior observed in real world data, is required. Second, efficient learning algorithms are required for inferring influence structure based on the assumed diffusion model.

A variety of information diffusion prediction frameworks have been developed in the litera-

ture [28, 38, 121, 126, 130]. A typical assumption in many of these approaches is that a connected network graph and knowledge of the corresponding structure are available *a priori*. However, in practice, the structure of the network can be implicit or difficult to model, e.g., modeling the structure of the spread of infectious disease is almost impossible. As a result, network structure unaware diffusion prediction models have gained interest. For example, Yang *et. al.* [126] proposed a linear influence model (LIM), which can effectively predict the information volume by assuming that each of the contagions spreads with the same influence in an implicit network. Subsequently, in [121], the authors extended LIM by exploiting the sparse structure in the influence function to identify the influential nodes. Though the relationships between multiple contagions can be used for more accurate modeling, most of the existing approaches ignore this information.

In this chapter, we address the above issues, especially the classification problem of influential node detection, by augmenting linear influence models with complex task dependency information. More specifically, we consider the dependency of different contagions in the network, and characterize their relationships using Copula Theory. Furthermore, by imposing a low-rank regularizer, we are able to characterize the clustering structure of the contagions and the nodes in the network. Through this novel formulation, we attempt to both improve the accuracy of the prediction system and better regularize the influence structure learning problem. Finally, we develop an efficient algorithm based on proximal mappings to solve this optimization problem. Experiments with synthetic data reveal that the proposed approach fares significantly better than a state-of-the-art multi-task variant of LIM both in terms of volume prediction and influence structure estimation performance. In addition, we demonstrate the superiority of the proposed method in predicting the time-varying volume of tweets using the ISIS twitter dataset[1].

[1]ISIS dataset from Kaggle is available at https://www.kaggle.com/kzaman/how-isis-uses-twitter.

4.2 Background

In this section, we present the linear influence model (LIM) [126] and discuss its limitations. Consider a set of N nodes that participate in an information diffusion process of K different contagions over time. Node $u \in \{1, \ldots, N\}$ can be infected by contagion $k \in \{1, \ldots, K\}$ at time $t \in \{0, 1, \ldots, T\}$. The volume $V_k(t)$ is defined as the total number of nodes that get infected by the contagion k at time t. Let the indicator function $M_{u,k}(t) = 1$ represent the event that node u got infected by contagion k at time t, and 0 otherwise. LIM models the volume $V_k(t)$ as a sum of influences of nodes u that got infected before time t:

$$V_k(t+1) = \sum_{u=1}^{N} \sum_{l=0}^{L-1} M_{u,k}(t-l) I_u(l+1), \tag{4.1}$$

where each node u has a particular non-negative influence function $I_u(l)$. One can simply think of $I_u(l)$ as the number of follow-up infections l time units after u got infected. The value of L is set to indicate that the influence of a node drops to 0 after L time units. Thus, the influence of node u is denoted by the vector $\mathbf{I}_u = (I_u(1), \ldots, I_u(L))^T \in \mathbb{R}^{L \times 1}$. Next, using the notation $\mathbf{V}_k = (V(1), \ldots, V(T))^T \in \mathbb{R}^{T \times 1}$ and $\mathbf{I} = (\mathbf{I}_1^T, \cdots, \mathbf{I}_N^T)^T \in \mathbb{R}^{LN \times 1}$, the inference procedure of LIM can be formulated as follows

$$\text{minimize} \quad \sum_{k=1}^{K} \|\mathbf{V}_k - \mathbf{M}_k \cdot \mathbf{I}\|_2^2 + \mathbb{1}(\mathbf{I}), \tag{4.2}$$

where $\mathbf{M}_k$ is obtained via concatenation of $M_{u,k}$, $\|\cdot\|_2$ denotes the Euclidean norm, and $\mathbb{1}(\mathbf{I})$ is an indicator function that is zero when $I_{uk}(l) \geq 0$ and $+\infty$ otherwise. A node can be influential due to various reasons. One of those is the node's specific location in the network, which is determined by network topology. For example, if a node is at the center of a star network, this node is in a position to influence others more easily. LIM links the volume of the contagions and the nodes' influences, without the knowledge of network topology. Even if the nodes' influences are related to their locations in the network, they can also be characterized by LIM. Indeed, LIM has been

effective in predicting the future volume for each contagion, however, it assumes that each node has the same influence across all the contagions. Consequently, to achieve contagion-sensitive node selection in an implicit network, the LIM model was extended and the multitask sparse linear influential model (MSLIM) was proposed in [121].

The influence function is defined by extending $\mathbf{I}_u$ in LIM into contagion-sensitive $\mathbf{I}_{u,k} \in \mathbb{R}^{L\times 1}$, which is a L-length vector representing the influence of the node u for the contagion k. For each contagion k, let $\mathbf{I}^k \in \mathbb{R}^{LN\times 1}$ be the vector obtained by concatenating $\mathbf{I}_{1k}, \ldots, \mathbf{I}_{Nk}$. For each node u, the influence matrix for the node u is defined as $\mathbf{I}_u = (\mathbf{I}_{u1}, \ldots, \mathbf{I}_{uK}) \in \mathbb{R}^{L\times K}$. Using these notations, the inference procedure to estimate $\mathbf{I}_{u,k}$ is formulated as follows

$$\text{minimize } \frac{1}{2}\sum_{k=1}^{K} \|\mathbf{V}_k - \mathbf{M}_k \cdot \mathbf{I}^k\|_2^2 + \lambda \sum_{u=1}^{N} \|\mathbf{I}_u\|_F + \gamma \sum_{u=1}^{N}\sum_{k=1}^{K} \|\mathbf{I}_{uk}\|_2 + \mathbb{1}(\mathbf{I}), \tag{4.3}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The penalty term $\|\mathbf{I}_u\|_F$ is used to encourage the entire matrix $\mathbf{I}_u$ to be zero altogether, which means that the node u is non-influential for all different contagions. If the estimated $\|\mathbf{I}_u\|_F > 0$ (i.e., the matrix $\mathbf{I}_u$ is non-zero), a fine-grained selection is performed by the penalty $\sum_{u=1}^{N}\sum_{k=1}^{K} \|\mathbf{I}_{uk}\|_2$, which is essentially a group-Lasso penalty and can encourage the sparsity of vectors $\{\mathbf{I}_{uk}\}$. For a specific contagion k, one can identify the most influential nodes by finding the optimal solution $\{\hat{\mathbf{I}}_{uk}\}$ of (4.3). However, the penalty terms used in MSLIM encourages that certain nodes have no influence over all the contagions which may not be true in practice. Furthermore, for most of the real world applications, there exists complex dependencies among the contagions. In order to alleviate these shortcomings, we propose a novel probabilistic multi-task learning framework and develop efficient optimization strategies.

4.3 Proposed Approach

4.3.1 Probabilistic Multi-Contagion Modeling of Diffusion

We assume a linear regression model for each task:$\mathbf{V}_k = \mathbf{M}_k\mathbf{I}^k + n_k$, where $\mathbf{V}_k, \mathbf{M}_k$ and $\mathbf{I}^k$ are defined as before, and $\mathbf{n}_k \in \mathbb{R}^{T\times 1}$ is an i.i.d. zero-mean Gaussian noise vector with the covariance matrix Σ_k. The distribution for $\mathbf{V}_k$ given $\mathbf{M}_k$, $\mathbf{I}^k$ and Σ_k can be expressed as

$$\mathbf{V}_k|\mathbf{M}_k, \mathbf{I}^k, \Sigma_k \sim \mathcal{N}\left(\mathbf{M}_k\mathbf{I}^k, \Sigma_k\right) = \frac{\exp\left(-\frac{1}{2}\left(\mathbf{V}_k - \mathbf{M}_k\mathbf{I}^k\right)^T \Sigma_k^{-1}\left(\mathbf{V}_k - \mathbf{M}_k\mathbf{I}^k\right)\right)}{(2\pi)^{\frac{T}{2}}\,|\Sigma_k|^{\frac{1}{2}}}. \tag{4.4}$$

Assuming that the influence for a single contagion is also Gaussian distributed, we can express the marginal distributions as $\mathbf{I}^k|\mathbf{m}_k, \Theta_k \sim \mathcal{N}(\mathbf{m}_k, \Theta_k)$, where $\mathbf{m}_k \in \mathbb{R}^{LN\times 1}$ is the mean vector and can be expressed as $\mathbf{m}_k = [\mathbf{m}_{1,k}^T, \ldots, \mathbf{m}_{N,k}^T]^T$, and $\Theta_k \in \mathbb{R}^{LN\times LN}$ is the covariance matrix of $\mathbf{I}^k$. For a node u and contagion k, we assume that the variables in the influence $\mathbf{I}_{uk}$ have the same mean, i.e., $\mathbf{m}_{u,k} = m_{u,k}\mathbf{1}_{L\times 1}$, where $m_{u,k}$ is a scalar and $\mathbf{1}_{L\times 1}$ is a vector of all ones with dimension $L \times 1$. Let $\mathbf{m}' \in \mathbb{R}^{N\times K}$ represent the mean matrix with entries $m_{u,k}$, and it is connected as $\mathbf{m} = (\mathbf{m}_1, \ldots, \mathbf{m}_K) = \mathbf{Q}\mathbf{m}'$, where $\mathbf{Q} \in \mathbb{R}^{LN\times N} = \mathcal{I}_{N\times N} \otimes \mathbf{1}_{L\times 1}$ and $\mathcal{I}_{N\times N}$ is the identity matrix with dimension $N \times N$ and $\otimes$ is the Kronecker product operator.

4.3.2 Dependence Structure Modeling Using Copulas

Consider a general case where the contagions are correlated. We construct a new influence matrix $\mathbf{I} = \left[\mathbf{I}^1, \ldots, \mathbf{I}^K\right] \in \mathbb{R}^{LN\times K}$. In our formulation, $\mathbf{I}^k$s are assumed to be correlated and the joint distribution of $\mathbf{I}$ is not a simple product of all the marginal distributions of $\mathbf{I}^k$ as is adopted by most multi-task learning formulations. Here, we propose to use a multi-task copula that is obtained by tailoring the copula model for the multi-task learning problem.

Copulas are parametric functions that couple univariate marginal distributions to a multivariate distribution. They can model the dependence among random variables with arbitrary marginal distributions. An important theorem that is central to the theory of copulas is Sklar's theorem

(see [81] for a detailed proof), which is stated below.

Theorem 4.1. *(Sklar's Theorem). Consider an N-dimensional distribution function F with marginal distribution functions $F_1, \ldots, F_N$. Then there exists a copula C, such that for all $x_1, \ldots, x_N$ in $[-\infty, \infty]$,*

$$F(x_1, \ldots, x_N) = C\left(F_1\left(x_1\right), \ldots, F_N\left(x_N\right)\right). \tag{4.5}$$

If F_n is continuous for $1 \leq n \leq N$, then C is unique, otherwise it is determined uniquely on $RanF_1 \times \ldots \times RanF_N$ where $RanF_n$ is the range of F_n. Conversely, given a copula C and univariate cumulative distribution functions (CDFs) $F_1, \ldots, F_N$, F is a valid multivariate CDF with marginals $F_1, \ldots, F_N$.

Note that the above theorem implies that the copula function is a joint distribution of uniformly distributed random variables. As a direct consequence of Sklar's Theorem, for continuous distributions, the joint probability density function (PDF) $f\left(x_1, \ldots, x_N\right)$ is obtained by differentiating both sides of (4.5),

$$f(x_1, \ldots, x_N) = \left(\prod_{n=1}^{N} f_n\left(x_n\right)\right) c\left(F_1\left(X_1\right), \ldots, F_N\left(X_N\right)\right), \tag{4.6}$$

where $f_n\left(\cdot\right)$ is the marginal PDF and c is termed as the copula density given by

$$c(v) = \frac{\partial^N C\left(v_1, \ldots, v_N\right)}{\partial v_1, \ldots, \partial v_N} \tag{4.7}$$

where $v_n = F_n(x_n)$.

One can construct a joint density function with specified marginal densities by employing (4.6). The choice of a copula function to represent the joint statistics is an important consideration. Various families of copula functions exist in the literature [81]. However, which copula function should be used for a given case is not very clear as different copula functions may characterize different types of dependence behavior among the random variables.

We apply the copula theory to multi-task learning and express the joint distribution of $\mathbf{I}$ as follows:

$$p(\mathbf{I}^1, \mathbf{I}^2, \ldots, \mathbf{I}^K) = \left(\prod_{k=1}^{K} \mathcal{N}(\mathbf{m}_k, \Theta_k) \right) c(F_1(\mathbf{I}^1), F_2(\mathbf{I}^2), \ldots, F_K(\mathbf{I}^K)), \tag{4.8}$$

where $F_k(\mathbf{I}^k)$ is the CDF of the influence for k^{th} contagion. The copula density function $c(\cdot)$ takes all marginal CDFs $\{F_k(\mathbf{I}^k)\}_{k=1}^K$ as its arguments, and maintains the output correlations in a parametric form.

Gaussian copula: There are a finite number of well defined copula families that can characterize several dependence structures [80]. Though, in general, one selects the most appropriate copula along with its parameters from data, here, we consider the Gaussian copula for its tractability and favorable analytical properties. A Gaussian copula can be constructed from the multivariate Gaussian CDF, and the resulting prior on $\mathbf{I}$ is given by a multivariate Gaussian distribution as

$$\mathbf{I} \sim \mathcal{MN}_{LN \times K}(\mathbf{m}, \mathbf{U}, \mathbf{\Omega}) = \frac{\exp\left(-\frac{1}{2}\text{tr}\left(\mathbf{U}^{-1}\left(\mathbf{I}-\mathbf{m}\right)\mathbf{\Omega}^{-1}\left(\mathbf{I}-\mathbf{m}\right)^T\right)\right)}{(2\pi)^{\frac{LNK}{2}}\left|\mathbf{\Omega}\right|^{\frac{LN}{2}}\left|\mathbf{U}\right|^{\frac{K}{2}}} \tag{4.9}$$

where $\mathbf{U} \in \mathbb{R}^{LN \times LN}$ is the row covariance matrix modeling the correlation between the influence of different nodes, $\mathbf{\Omega} \in \mathbb{R}^{K \times K}$ is the column covariance matrix modeling the correlation between the influence for different contagions, and $\mathbf{m} \in \mathbf{R}^{LN \times K}$ is the mean matrix of $\mathbf{I}$. The two covariances can be computed as $E\left[\left(\mathbf{I}-\mathbf{m}\right)\left(\mathbf{I}-\mathbf{m}\right)^T\right] = \mathbf{U}\text{tr}(\mathbf{\Omega})$ and $E\left[\left(\mathbf{I}-\mathbf{m}\right)^T\left(\mathbf{I}-\mathbf{m}\right)\right] = \mathbf{\Omega}\text{tr}(\mathbf{U})$ respectively. We assume that N individual nodes are spreading the contagions and influencing others independently, and thus the row covariance matrix is diagonal and can be expressed as $\mathbf{U} = \text{diag}(e_1^2, e_2^2, \ldots, e_N^2) \otimes \mathcal{I}_{L \times L}$, where $e_n^2, n \in \{1, \ldots, N\}$ are scalars. The posterior distribution for $\mathbf{I}$, which is proportional to the product of the prior in (4.4) and the likelihood function

in (4.9), is given as

$$p\left(\mathbf{I}|\mathbf{M},\mathbf{V},\Sigma,\mathbf{U},\Omega\right) \propto p\left(\mathbf{V}|\mathbf{M},\mathbf{I},\Sigma\right) p\left(\mathbf{I}|\mathbf{m},\mathbf{U},\Omega\right)$$
$$= \left(\prod_{k=1}^{K} \mathcal{N}\left(\mathbf{M}_k\mathbf{I}^k, \Sigma_k\right)\right) \mathcal{MN}_{LN\times K}\left(\mathbf{I}|\mathbf{m},\mathbf{U},\mathbf{\Omega}\right), \tag{4.10}$$

where $\mathbf{M} = (\mathbf{M}_1, \ldots, \mathbf{M}_K) \in \mathbb{R}^{T\times LNK}$, $\mathbf{V} = (\mathbf{V}_1, \ldots, \mathbf{V}_K) \in \mathbb{R}^{T\times K}$, Σ is the corresponding covariance matrix of $n = (n_1, \ldots, n_K) \in \mathbb{R}^{T\times K}$. We assume $\Sigma_k \triangleq \sigma^2\mathcal{I}_{T\times T}$ and also an identical value of $e_n^2 = e^2, \forall k = 1, \ldots, K, \forall n = 1, \ldots, N$. We employ maximum a posteriori (MAP) and maximum likelihood (ML) estimation methods, and obtain $\mathbf{I}$, $\mathbf{m}$, and $\mathbf{\Omega}$ by

$$\min_{\mathbf{I},\mathbf{m},\mathbf{\Omega}} \frac{1}{\sigma^2}\sum_{k=1}^{K} \|\mathbf{V}_k - \mathbf{M}_k\mathbf{I}^k\|_2^2 + \frac{1}{e^2}\text{tr}\left((\mathbf{I}-\mathbf{m})\mathbf{\Omega}^{-1}(\mathbf{I}-\mathbf{m})^T\right) + LN\ln|\mathbf{\Omega}| + \mathbb{1}(\mathbf{I}).$$

However, if we assume $\mathbf{\Omega}^{-1}$ to be non-sparse, the solution to $\mathbf{\Omega}^{-1}$ will not be defined (when $K > LN$) or will overfit (when K is of the same order as LN) [92]. In fact, some contagions in the network can be uncorrelated, which makes the corresponding entry values in $\mathbf{\Omega}^{-1}$ zero. Hence, we add an l_1 penalty term to promote sparsity of matrix $\mathbf{\Omega}^{-1}$ to obtain

$$\min_{\mathbf{I},\mathbf{m},\mathbf{\Omega}} \sum_{k=1}^{K} \|\mathbf{V}_k - \mathbf{M}_k\mathbf{I}^k\|_2^2 + \lambda_1\text{tr}\left((\mathbf{I}-\mathbf{m})\mathbf{\Omega}^{-1}(\mathbf{I}-\mathbf{m})^T\right) - \lambda_2\ln|\mathbf{\Omega}^{-1}| + \lambda_3\|\mathbf{\Omega}\|_1 + \mathbb{1}(\mathbf{I}).$$

Compared with the state-of-art LIM method (4.3), the above formulation incorporates complex correlation of the influence matrix $\mathbf{I}$ for different users and different contagions.

4.3.3 Modeling the Structure of Influence Matrix I

In order to better characterize the influence matrix, we propose to impose a low rank structure on the influence matrix $\mathbf{I}$. The nodes or the contagions in the influence network are known to form communities (or clustering structures), which may be captured using the low-rank property of the influence matrix. Note that, the sparse structure in the influence matrix implies that most

individuals only influence a small fraction of contagions in the network while there can be a few nodes with wide-spread influence. We incorporate this into our formulation by using a sparsity promoting regularizer over $\mathbf{I}_{u,k}$.

$$\begin{aligned}\min_{\mathbf{I},\mathbf{m},\mathbf{\Omega}} \quad & \sum_{k=1}^{K} \|\mathbf{V}_k - \mathbf{M}_k\mathbf{I}^k\|_2^2 + \lambda_1 \text{tr}\left((\mathbf{I}-\mathbf{m})\mathbf{\Omega}^{-1}(\mathbf{I}-\mathbf{m})^T\right) \\ & - \lambda_2 \ln|\mathbf{\Omega}^{-1}| + \lambda_3\|\mathbf{\Omega}\|_1 + \lambda_4\|\mathbf{I}\|_* + \lambda_5 \sum_{u=1}^{N}\sum_{k=1}^{K} \|\mathbf{I}_{uk}\|_2 + \mathbb{1}(\mathbf{I}),\end{aligned} \tag{4.11}$$

where $\|\cdot\|_*$ denotes the nuclear norm, and λ_1, λ_2, λ_3, λ_4 and λ_5 are the regularization parameters. With the estimated $\{\hat{\mathbf{I}}_{uk}\}$, one can predict the total volume of the contagion k at $T+1$ by $\hat{V}_k(T+1) = \sum_{u=1}^{N}\sum_{l=0}^{L-1} M_{uk}(T-l)I_{uk}(l+1)$.

Thus, our proposed approach incorporates the complex dependence among different users and different contagions, and the unique structure of the influence matrix, which can simultaneously perform contagion-sensitive volume prediction and influential node detection in a unified framework.

4.4 Optimization Algorithm

We adopt an alternating optimization approach to solve the problem in (4.11).

Optimization w.r.t. $\mathbf{m}$: Given $\mathbf{I}$ and $\mathbf{\Omega}^{-1}$, the mean matrix $\mathbf{m}$ can be obtained by solving the following problem

$$\min_{\mathbf{m}} \quad \text{tr}\left((\mathbf{I}-\mathbf{m})\mathbf{\Omega}^{-1}(\mathbf{I}-\mathbf{m})^T\right).$$

The estimate $\hat{\mathbf{m}}$ can be analytically obtained as $\hat{\mathbf{m}} = \frac{1}{L}\mathbf{Q}\mathbf{Q}^T\mathbf{I}$.

Optimization w.r.t. $\mathbf{\Omega}$: Given $\mathbf{I}$ and $\mathbf{m}$, the contagion inverse covariance matrix $\mathbf{\Omega}^{-1}$ can be estimated by solving the following optimization problem

$$\min_{\mathbf{\Omega}} \quad \lambda_1 \text{tr}\left((\mathbf{I}-\mathbf{m})\mathbf{\Omega}^{-1}(\mathbf{I}-\mathbf{m})^T\right) - \lambda_2 \ln|\mathbf{\Omega}^{-1}| + \lambda_3\|\mathbf{\Omega}\|_1$$

The above is an instance of the standard inverse covariance estimation problem with sample covariance $\frac{\lambda_1}{\lambda_2}(\mathbf{I}-\mathbf{m})^T(\mathbf{I}-\mathbf{m})$, which can be solved using standard tools. In particular, we use the graphical Lasso procedure in [32]

$$\hat{\mathbf{\Omega}}^{-1} = gLasso\left(\lambda_1/\lambda_2(\mathbf{I}-\mathbf{m})^T(\mathbf{I}-\mathbf{m}), \lambda_3\right). \tag{4.12}$$

Optimization w.r.t. I: The corresponding optimization problem becomes

$$\min_{\mathbf{I}} \sum_{k=1}^{K} \|\mathbf{V}_k - \mathbf{M}_k\mathbf{I}^k\|_2^2 + \lambda_1 \text{tr}\left((\mathbf{I}-\mathbf{m})\mathbf{\Omega}^{-1}(\mathbf{I}-\mathbf{m})^T\right) + \lambda_4\|\mathbf{I}\|_* + \lambda_5 \sum_{u=1}^{N}\sum_{k=1}^{K} \|\mathbf{I}_{uk}\|_2 + \mathbb{1}(\mathbf{I}).$$

We rewrite the problem as

$$\min_{\mathbf{I}} \quad \ell(\mathbf{I}) + \lambda_4\|\mathbf{I}\|_* + \mathbb{1}(\mathbf{I}). \tag{4.13}$$

where $\ell(\mathbf{I}) = \sum_{k=1}^{K} \|\mathbf{V}_k - \mathbf{M}_k\mathbf{I}^k\|_2^2 + \lambda_1 \text{tr}\left((\mathbf{I}-\mathbf{m})\mathbf{\Omega}^{-1}(\mathbf{I}-\mathbf{m})^T\right) + \lambda_5 \sum_{u=1}^{N}\sum_{k=1}^{K} \|\mathbf{I}_{uk}\|_2$. This formulation involves a sum of a convex differentiable loss and convex non-differentiable regularizers which renders the problem non-trivial. A number of algorithms have been developed for the case where the optimal solution is easy to compute when each regularizer is considered in isolation. This corresponds to the case where the proximal operator defined for a convex regularizer $R : \mathbb{R}^{LN\times K} \to \mathbb{R}$ at a point $\mathbf{Z}$ by $\text{prox}_R(\mathbf{Z}) = \arg\min \frac{1}{2}\|\mathbf{I}-\mathbf{Z}\|_F^2 + R(\mathbf{I})$, is easy to compute for each regularizer taken separately. See [20] for a broad overview of proximal methods. The proximal operator for the nuclear norm is given by the shrinkage operation as follows [5]. If $U\text{diag}(\sigma_1, \ldots, \sigma_n)V^T$ is the singular value decomposition of $\mathbf{Z}$, then $\text{prox}_{\lambda_4\|\cdot\|_*}(\mathbf{Z}) = U\text{diag}((\sigma_i - \lambda_4)_+)_i V^T$. The proximal operator of the indicator function $\mathbb{1}(\mathbf{I})$ is simply the projection onto $I_{u,k}(l) \geq 0$, which is denoted by $P_{\mathbb{1}}(\mathbf{I})$. Next, we describe a matching serial algorithm introduced in [6]. Here, we present a version where updates are performed according to a cyclic order [96]. Note that one can also randomly select the order of the updates. We use the optimization algorithm 1 to solve the optimization problem in (4.13).

Algorithm 1 Incremental Proximal Descent

Initialize $\mathbf{I} = \mathbf{A}$
repeat
    Set $\mathbf{I} = \mathbf{I} - \theta \nabla_{\mathbf{I}} \ell(\mathbf{I})$
    Set $\mathbf{I} = \text{prox}_{\theta\lambda_4 \|\cdot\|_*}(\mathbf{I})$
    Set $\mathbf{I} = P_{\mathbb{1}}(\mathbf{I})$
until convergence
return $\mathbf{I}$

4.5 Experimental Results

We compare the performance of the proposed approach to MSLIM by applying it to both synthetic and real datasets. Since the volume of a contagion over time $V_k(t)$ can be viewed as a time series, we set up this problem as a time series prediction task and evaluate the performance using the prediction mean-squared error (MSE), which is the l-2 norm of the difference between the true volume and the predicted volumn across different time instances. Furthermore, for the synthetic data set, where we have access to the true influence matrix $\mathbf{I}$, we also evaluate the performance of the influence matrix prediction task using the metric $\|\hat{\mathbf{I}} - \mathbf{I}\|_F$, which is termed as "Influence Matrix Estimation Error". We determined the regularization parameters for the proposed model using cross validation. In particular, we split the first 60% of the time instances as the training set and the rest for validation. Following [121], we combine the training and validation sets to re-train the model with the best selected regularization parameters and estimate the influence matrix.

4.5.1 Synthetic Data

We created a synthetic dataset with the number of nodes fixed at $N = 100$ and the number of contagions at $K = 20$. In addition, we assumed that $L = 10$ and $T = 20$. A rank 5 (low-rank) influence matrix $\mathbf{I}$ was generated randomly with uniformly distributed entries. The matrix $\mathbf{M}$ was generated with uniformly distributed random integers $\{0, 1\}$. Following our model assumption, the volume for each $\mathbf{V}_k$ was calculated as follows $\mathbf{V}_k = \mathbf{M}_k \times \mathbf{I}^k + \mathcal{N}(\mathbf{0}, \mathbf{\Delta})$ where $\mathcal{N}(\mathbf{0}, \mathbf{\Delta})$ is a multivariate normal distribution with covariance matrix Δ. In Table 4.1, we present the results obtained using the proposed approach and its comparison to MSLIM, the state-of-art LIM method

Table 4.1: Prediction performance for different information diffusion models on synthetic data.

Approach	MSLIM	Proposed
Volume Prediction MSE	0.834	**0.007**
Influence Matrix Estimation Error	0.7681	**0.62**

(4.3). As can be observed that for volume prediction, our method obtains a significant improvement compared with MSLIM, and achieves highly accurate estimates. For influence matrix estimation, the proposed approach provides a better result compared to MSLIM, while the improvement is not as significant as that of volume prediction. The reason is that the model is trained to minimize the volume predication error, since in reality we might not be able to have access to the influence matrix. Thus, the volume predication error is very small for our method, while the influence matrix estimation error is comparatively larger.

4.5.2 ISIS Twitter Data

In this section, we demonstrate the application of the proposed approach to a real-word analysis task. We begin by describing the twitter dataset used for analysis and the procedure adopted to extract the set of contagions. Following this, we discuss the problem setup and present comparisons to MSLIM on predicting the time-varying tweet volume. Finally, we present a qualitative analysis of the inferred influence structure for different contagions.

The ISIS dataset from Kaggle[2] is comprised of over $17,000$ tweets from 112 users posted between January 2015 and May 2016. In addition to the actual tweets, meta-information such as the user name and the timestamp for each tweet are included. We performed standard pre-processing by removing a variety of stop words, e.g. URLs, and symbols. After preprocessing, we converted each tweet into a bag-of-words representation and extracted the term frequency-inverse document frequency (tf-idf) feature.

Topic Modeling: When applying our approach, the first step is to define semantically meaningful

[2]ISIS dataset from Kaggle is available at https://www.kaggle.com/kzaman/how-isis-uses-twitter.

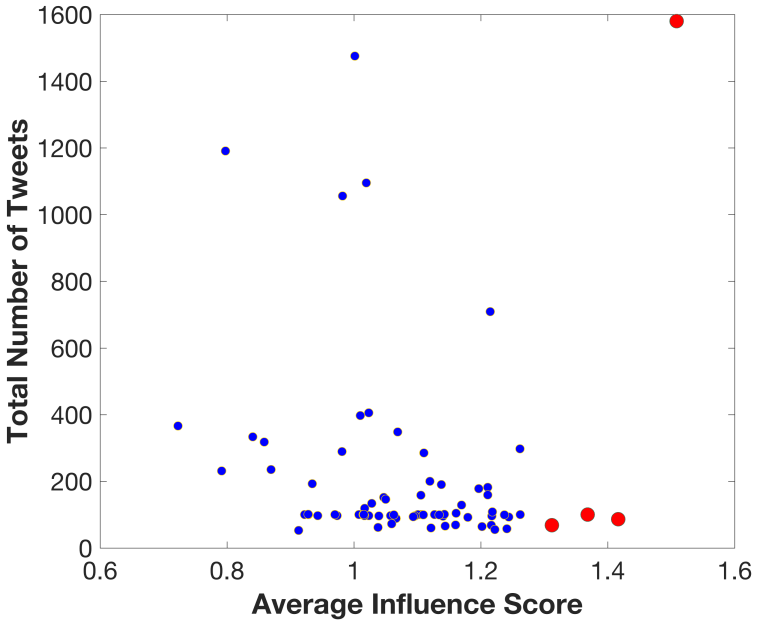
Table 4.2: Top words for each topic learned using NMF with the ISIS twitter dataset.

Topic 1	isis ramiallolah iraq attack libya warreporter1 saa aamaq usa abu
Topic 2	killed soldiers today airstrikes injured wounded civilians militants iraqi attack
Topic 3	syria russia ramiallolah turkey ypg breakingnews usa group saa terror
Topic 4	state islamic fighters fighting group saudi new http wilaya control
Topic 5	aleppo nid gazaui rebels north today northern syrian ypg turkish
Topic 6	assad regime myra forces rebels fsa pro islam syrian jaysh
Topic 7	al qaeda nusra abu sham ahrar islam jabhat http warreporter1
Topic 8	army iraq near ramiallolah iraqi lujah turkey ramadi west sinai
Topic 9	allah people muslims abu accept muslim make know don islam
Topic 10	breaking islamicstate forces amaqagency city fighters iraqi near area syrian

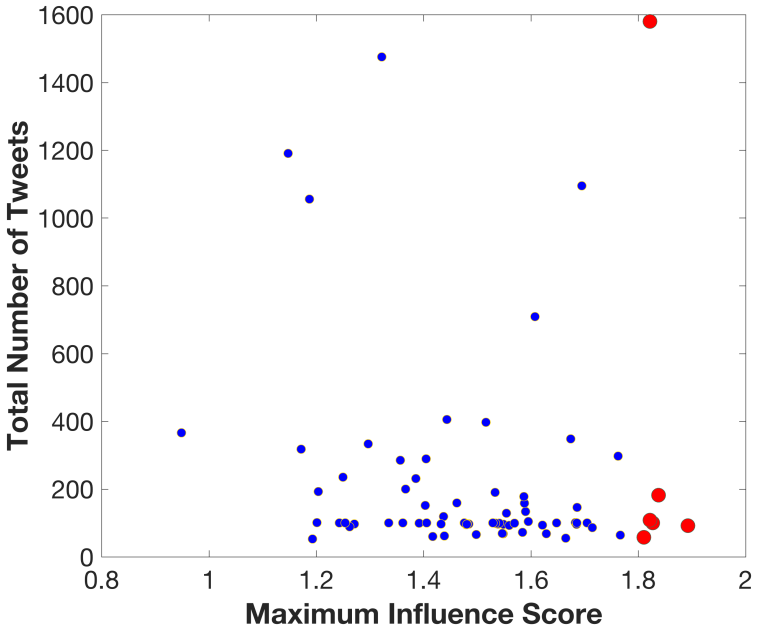
contagions. A simple way of defining topics is to directly use words as topics (e.g., ISIS). However, a single word may not be rich enough to represent a broad topic (e.g., social network sites). Hence, we propose to perform topic modeling on the tweets based on the tf-idf features. In our experiment, we obtained the topics using Non-negative Matrix Factorization (NMF), which is a popular scheme for topic discovery, with the number of topics K set at 10. Table 4.2 lists the top 10 words for each of the topics learned using NMF.

Volume Time Series Prediction: In our experiment, we set one day as the discrete time step for aggregating the tweet volume. The parameter L denotes the number of time steps it takes for the influence of a user to decay to zero. We set the parameter L equal to 5 since we observed that beyond $L = 5$, there is hardly any improvement in performance. The MSE on the predicted volume is computed over the entire period of observation. The comparison of the prediction MSE is presented in Table 4.3. It can be seen that the proposed approach significantly outperforms MSLIM in predicting the time-varying volume.

Influential Node Detection: For a contagion k, we identify the most influential nodes with respect to this contagion as nodes having high $\|\mathbf{I}_{\mathbf{u},\mathbf{k}}\|_2$ values. First, in Figure 4.2(a), we plot the correlation among 10 topics learned by NMF. More specifically, we plot the pair-wise correlation structure learned by our approach. It can be seen that, a strong positive correlation structure exists,

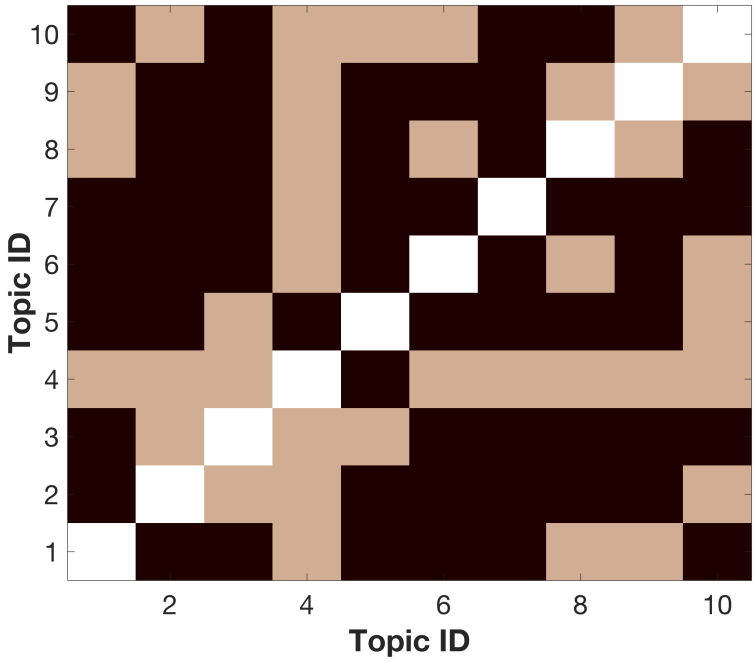


(a) Average Influence

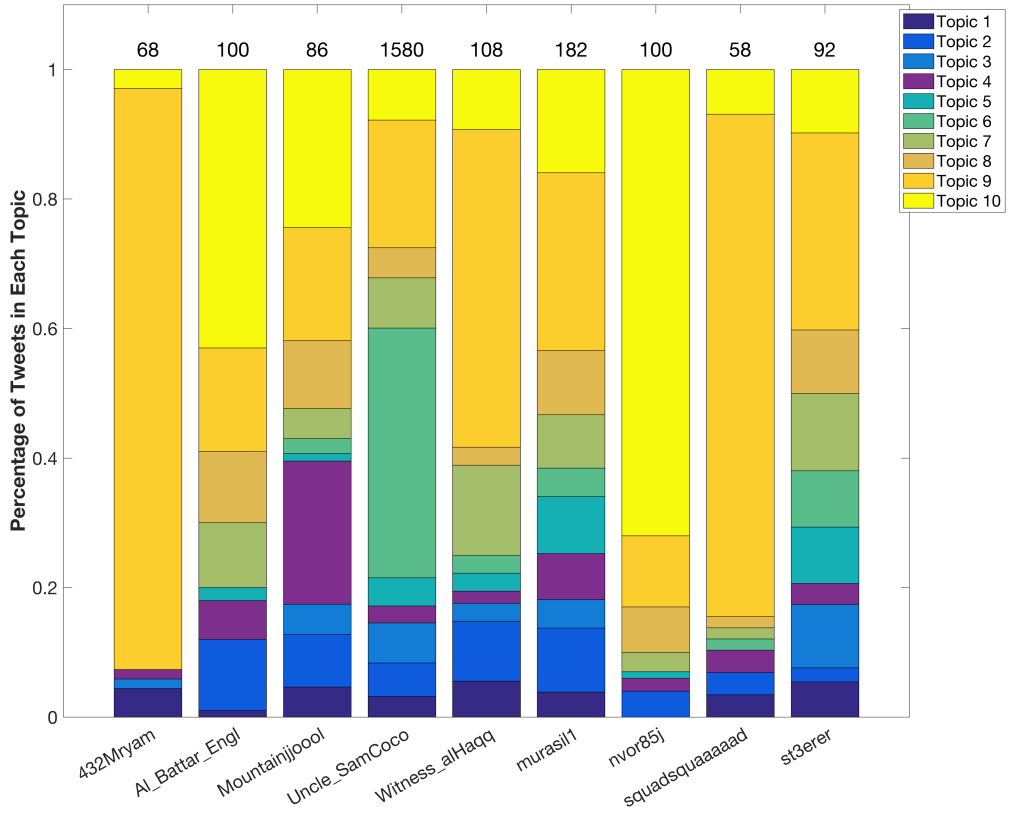


(b) Maximum Influence

Fig. 4.1: Comparing statistics from the estimated influence matrix with the volume of tweets corresponding to each of the users to identify influential users. We define the average influence score as the averaged influence for a user among all the topics. The maximum influence score is defined as the maximum influence for a user across all the topics. In both cases, the users with a large influence score are marked in red.



(a)



(b)

Fig. 4.2: (a) Correlation Structure among the topics (non-black color represents positive correlation), (b) Top 9 influential users and their tweet distributions.

Table 4.3: Volume prediction performance on the ISIS twitter dataset.

Approach	MSLIM	Proposed
Volume Prediction MSE	2.7	**0.329**

which enabled the improved prediction in Table 4.3. Following this, we use the predicted influence matrix to select a set of highly influential nodes from the dataset. A simple approach to select the influential users can be to select the ones with a large number of tweets. However, we argue that the influence predicted in an information diffusion model can be vastly different. Consequently, we consider a user to be influential if she has a high influence score for at least one of the topics, or if she can be influential for multiple topics. For example, in Figure 4.1(a), we plot average influence scores of the users (averaged over all the topics) against the total number of tweets. Similarly, in Figure 4.1(b), we plot influence scores of the users (maximum over all the topics) against the total number of tweets. The first striking observation is that the users with high influence scores are not necessarily the ones with the most number of tweets. Instead, their impact on the information diffusion relies heavily on the complex dynamics of the implicit network.

Finally, in Figure 4.2(b) we plot the percentage of tweets regarding each of the topics for top 9 influential nodes. Influential nodes are obtained as a union of nodes identified based on both average and maximum influence scores. More specifically, we select the union of users with average influence score greater than 1.3 and maximum influence score greater than 1.8. In addition to displaying the distribution across topics, for each influential user, we show the total number of tweets posted by that user. It can be seen that the total number of tweets of these users vary a lot and, therefore, is not a good indication of their influence.

4.6 Summary

In this chapter, we considered the problem of influential node detection and volume time series prediction. We proposed a descriptive diffusion model to take dependencies among the topics into account. We also proposed an efficient algorithm based on alternating methods to perform infer-

ence and learning on the model. It was shown that the proposed technique outperforms existing influential node detection techniques. Furthermore, the proposed model was validated both on a synthetic and a real (ISIS) dataset. We showed that the proposed approach can efficiently select the most influential users for specific contagions. We also presented several interesting patterns of the selected influential users for the ISIS dataset.

CHAPTER 5
CLASSIFICATION IN DECENTRALIZED LEARNING SYSTEM: UNRELIABLE AGENTS

5.1 Introduction

As one of the typical machine learning and statistics problems, classification fits into the general framework where a finite-sum of functions is to be optimized. In general, the problem is formulated as

$$\min_{\mathbf{x}\in\mathbb{R}^N} \; f(\mathbf{x}), \quad f(\mathbf{x}) = \sum_{i=1}^{D} f_i(\mathbf{x}). \tag{5.1}$$

The problem structure in (5.1) is applicable to collaborative autonomous inference in statistics, distributed cooperative control of unmanned vehicles in control theory, and training of models (such as, support vector machines, deep neural networks, etc.) in machine learning. Due to the emergence of the big data era and associated sizes of datasets, solving problem (5.1) at a single node (or agent) is often infeasible.This gives rise to the decentralized optimization setting [11,63], in which the training data for the problem is stored and processed across a number of interconnected nodes

and the optimization problem is solved collectively by the cluster of nodes. The decentralized classification system can be implemented on an arbitrarily connected network of computational nodes that solves (5.1) by treating it as a consensus optimization problem. There exist several decentralized optimization methods for solving (5.1), including belief propagation [88], distributed subgradient descent algorithms [79], dual averaging methods [29], and the alternating direction method of multipliers (ADMM) [11]. Among these, ADMM has drawn significant attention, as it is well suited for decentralized optimization and demonstrates fast convergence in many applications, such as online learning, decentralized collaborative learning, neural network training, and so on [45, 109, 125].

However, most of these past works assume an ideal system where updates are not erroneous. This assumption is very restrictive and rarely satisfied in practice which limits the applicability of these results. Note that due to the decentralized nature of the systems considered, computation over federated machines induces a higher risk of unreliability because of communication noise, crash failure, and adversarial attacks. Therefore, the design and analysis of decentralized optimization algorithms in the presence of these practical challenges is of utmost importance. A systematic convergence analysis of ADMM in the presence of unreliable agents has been missing for a long time. The reason is that unreliable agents (sometimes termed as Byzantine agents in the literature) have large degrees of freedom without abiding to an error model and this makes the convergence analysis significantly more challenging as existing proof techniques used in studying the convergence of ADMM do not directly apply.

Although, the problem of design and analysis of ADMM with unreliable agents has not been considered in the past, a related research direction is inexact consensus ADMM [8, 16, 36, 82, 124, 127]. The inexactness in ADMM can be categorized as of two different types. Type 1 assumes that there are errors that can occur in an intermediate step of proximal mapping in each ADMM iteration. Type 2 replaces the computationally complex calculation in each ADMM iteration by a proximity operator that can be computed more easily, and hence inexactness occurs. Error in inexact ADMM is induced implicitly in intermediate proximal mapping steps and, thus, has a

specific restrictive and bounded form with amenable properties for convergence analysis (such as, it converges to zero). These assumptions are very limited in their ability to model unreliability in updates, and are different from what we have studied in this thesis. Furthermore, since the proof techniques for the convergence analysis of inexact ADMMs are designed on an algorithm-by-algorithm basis with restrictive assumptions on error, it lacks a unified framework to analyze the convergence problem of ADMM with an arbitrary error model (of utmost importance to cyber physical security and noisy communication channel scenarios).

A unified framework to study the convergence analysis of decentralized ADMM algorithms for classification in the presence of an arbitrary error model is proposed in this chapter[1]. We consider a general error model where an unreliable agent i adds an arbitrary error term $\mathbf{e}_i^k$ to its state value $\mathbf{x}_i^k$ at each time step k. The error first contaminates $\mathbf{x}_i^k$ and the resulting output $\mathbf{x}_i^k + \mathbf{e}_i^k$ is broadcast to the neighboring agents. First, we provide a comprehensive convergence analysis both for convex (and strongly convex) cost functions. Next, we show that ADMM converges to a neighborhood of the optimal solution if certain conditions involving the network topology, the properties of the objective function, and algorithm parameters, are satisfied. Guidelines are developed for network structure design and algorithm parameter optimization to achieve faster convergence. We also give several conditions on the errors such that exact convergence to the optimum can be achieved, instead to the neighborhood of the optimum. Finally, to mitigate the effect of unreliable agents, a robust variant of ADMM, referred to as ROAD, is proposed. We show that ROAD achieves exact convergence to the optimum with a rate of $\mathcal{O}(1/T)$ for convex cost functions.

[1]Note that, the results in inexact ADMM literature [8, 16, 36, 82, 124, 127] can be seen as a special cases of our analysis.

5.2 Problem Formulation

5.2.1 Decentralized Learning with ADMM

Consider a network consisting of D agents bidirectionally connected with E edges. We can describe the network as a symmetric directed graph $\mathcal{G}_d = \{\mathcal{V}, \mathcal{A}\}$, where $\mathcal{V}$ is the set of vertices and $\mathcal{A}$ is the set of arcs with $|\mathcal{A}| = 2E$. In a distributed setup, a connected network of agents collaboratively minimize the sum of their local loss functions over a common optimization variable. Each agent generates local updates individually and communicates with its neighbors to reach a network-wide common minimizer. The decentralized learning problem, can be formulated as follows

$$\min_{\{\mathbf{x}_i\},\{\mathbf{y}_{ij}\}} \sum_{i=1}^{D} f_i(\mathbf{x}_i), \quad \text{s.t. } \mathbf{x}_i = \mathbf{y}_{ij},\ \mathbf{x}_j = \mathbf{y}_{ij},\ \forall (i,j) \in \mathcal{A}, \tag{5.2}$$

where $\mathbf{x}_i \in \mathbb{R}^N$ is the local optimization variable at agent i and $\mathbf{y}_{ij} \in \mathbb{R}^N$ is an auxiliary variable imposing the consensus constraint on neighboring agents i and j. Defining $\mathbf{x} \in \mathbb{R}^{DN}$ as a vector concatenating all $\mathbf{x}_i$, $\mathbf{y} \in \mathbb{R}^{2EN}$ as a vector concatenating all $\mathbf{y}_{ij}$, (5.2) is written in a matrix form as

$$\min_{\mathbf{x},\mathbf{y}} f(\mathbf{x}) + g(\mathbf{y}), \quad s.t.\ \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} = 0, \tag{5.3}$$

where $f(\mathbf{x}) = \sum_{i=1}^{D} f_i(\mathbf{x}_i)$ and $g(\mathbf{y}) = 0$. Here $\mathbf{A} = [\mathbf{A}_1; \mathbf{A}_2]; \mathbf{A}_1, \mathbf{A}_2 \in \mathbb{R}^{2EN \times LN}$ are both composed of $2E \times D$ blocks of $N \times N$ matrices. If $(i,j) \in \mathcal{A}$ and $\mathbf{y}_{ij}$ is the qth block of $\mathbf{y}$, then the (q,i)th block of $\mathbf{A}_1$ and the (q,j)th block of $\mathbf{A}_2$ are $N \times N$ identity matrices $\mathbf{I}_N$; otherwise the corresponding blocks are $N \times N$ zero matrices $\mathbf{0}_N$. Also,we have $\mathbf{B} = [-\mathbf{I}_{2EN}; -\mathbf{I}_{2EN}]$ with $\mathbf{I}_{2EN}$ being a $2EN \times 2EN$ identity matrix. Define the matrices: $\mathbf{M}_+ = \mathbf{A}_1^T + \mathbf{A}_2^T$ and $\mathbf{M}_- = \mathbf{A}_1^T - \mathbf{A}_2^T$. Let $\mathbf{W} \in \mathbb{R}^{DN \times DN}$ be a block diagonal matrix with its (i,i)th block being the degree of agent i multiplying $\mathbf{I}_N$ and other blocks being $\mathbf{0}_N$, $\mathbf{L}_+ = \frac{1}{2}\mathbf{M}_+\mathbf{M}_+^T$, $\mathbf{L}_- = \frac{1}{2}\mathbf{M}_-\mathbf{M}_-^T$, and we know $\mathbf{W} = \frac{1}{2}(\mathbf{L}_+ + \mathbf{L}_-)$. These matrices are related to the underlying network topology.

5.2.2 Decentralized ADMM with Unreliable Agents

The iterative updates of the decentralized ADMM algorithm are given by [101] as

$$\begin{aligned} &\mathbf{x}-\texttt{update}: \nabla f(\mathbf{x}^{k+1})+\alpha^k+2c\mathbf{W}\mathbf{x}^{k+1}=c\mathbf{L}_+\mathbf{x}^k,\\ &\alpha-\texttt{update}: \alpha^{k+1}-\alpha^k-c\mathbf{L}_-\mathbf{x}^{k+1}=0. \end{aligned} \tag{5.4}$$

Note that $\mathbf{x}=[\mathbf{x}_1;\ldots;\mathbf{x}_D]$ where $\mathbf{x}_i\in\mathbb{R}^N$ is the local update of agent i and $\alpha=[\alpha_1;\ldots;\alpha_D]$ where $\alpha_i\in\mathbb{R}^N$ is the local Lagrange multiplier of agent i. Recalling the definitions of $\mathbf{W}$, $\mathbf{L}_+$ and $\mathbf{L}_-$, (5.4) results in the decentralized update of agent i given as follows

$$\begin{aligned} &\nabla f_i(\mathbf{x}_i^{k+1})+\alpha_i^k+2c|\mathcal{N}_i|\mathbf{x}_i^{k+1}=c|\mathcal{N}_i|\mathbf{x}_i^k+c\sum_{j\in\mathcal{N}_i}\mathbf{x}_j^k,\\ &\alpha_i^{k+1}=\alpha_i^k+c|\mathcal{N}_i|\mathbf{x}_i^{k+1}-c\sum_{j\in\mathcal{N}_i}\mathbf{x}_j^{k+1}, \end{aligned}$$

where $\mathcal{N}_i$ denotes the set of neighbors of agent i.

In such a setup, we consider the case where a fraction of the agents are unreliable and generate erroneous updates. Assume that the true update is $\mathbf{x}^k$, and the erroneous update is modeled as $\mathbf{x}^k+\mathbf{e}^k$, which is denoted as $\mathbf{z}^k=\mathbf{x}^k+\mathbf{e}^k$. The corresponding algorithm becomes

$$\begin{aligned} &\nabla f_i(\mathbf{x}_i^{k+1})+\alpha_i^k+2c|\mathcal{N}_i|\mathbf{x}_i^{k+1}=c|\mathcal{N}_i|\mathbf{z}_i^k+c\sum_{j\in\mathcal{N}_i}\mathbf{z}_j^k,\\ &\alpha_i^{k+1}=\alpha_i^k+c|\mathcal{N}_i|\mathbf{z}_i^{k+1}-c\sum_{j\in\mathcal{N}_i}\mathbf{z}_j^{k+1}. \end{aligned}$$

For a clearer presentation, we will use the following form of the updates for our analysis

$$\begin{aligned} &\mathbf{x}-\texttt{update}: \nabla f(\mathbf{x}^{k+1})+\alpha^k+2c\mathbf{W}\mathbf{x}^{k+1}=c\mathbf{L}_+\mathbf{z}^k,\\ &\alpha-\texttt{update}: \alpha^{k+1}-\alpha^k-c\mathbf{L}_-\mathbf{z}^{k+1}=0. \end{aligned} \tag{5.5}$$

Compared to (5.4), $\mathbf{x}^k$ is replaced by the erroneous update $\mathbf{z}^k$ in the first step, and $\mathbf{x}^{k+1}$ is replaced by $\mathbf{z}^{k+1}$ in the second step. The convergence analysis of (5.5) is nontrivial and is not a straightfor-

ward extension of the analysis with (5.4) in [101]. Additionally, the analysis in [101] was restricted to strongly convex cost functions. We analyze the problem for both convex and strongly convex cost functions.

5.2.3 Assumptions

We provide definitions and assumptions that will be used for the cost functions in our analysis.

Definition 5.1. *For a differentiable function* $f(\mathbf{x}) : \mathbb{R}^{DN} \to \mathbb{R}$*:*

- f *is* v*-strongly convex if* $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^{DN}$, $f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + v\|\mathbf{x} - \mathbf{y}\|^2$.
- f *is* L*-smooth if* $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^{DN}$, $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$.

Assumption 1. *For a differentiable function* $f(\mathbf{x}) : \mathbb{R}^{DN} \to \mathbb{R}$*:*

- *The feasible* $\mathbf{x} \in \mathbb{R}^N$ *is bounded as* $\|\mathbf{x}\| \leq V_1$.
- *The gradient* $\nabla f(\mathbf{x})$ *is bounded as* $\|\nabla f(\mathbf{x})\| \leq V_2$.

Note that these assumptions are very common in the analysis of first-order optimization methods [10]. The first assumption provides the feasible set, which basically means that $\mathbf{x} = +\infty$ and $\mathbf{x} = -\infty$ are not considered by the system. The second assumption assumes that the cost function does not change values abruptly in a small area in the domain.

5.3 Convergence Analysis

To effectively present the convergence results, we first introduce a few notations. Let $\mathbf{Q} = \mathbf{V}\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{V}^T$, where $\frac{\mathbf{L}-}{2} = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^T$ is the singular value decomposition of the positive semidefinite matrix $\frac{\mathbf{L}-}{2}$. We also construct a new auxiliary sequence $\mathbf{r}^k = \sum_{s=0}^{k} \mathbf{Q}(\mathbf{x}^s + \mathbf{e}^s)$. Let $\mathbf{z}^* = \mathbf{x}^*$, where $\mathbf{x}^*$ denotes the optimal solution to the problem. Define the auxiliary vector $\mathbf{q}^k$, matrix $\mathbf{p}^k$, and

matrix $\mathbf{G}$ as

$$\mathbf{q}^k = \begin{bmatrix} \mathbf{r}^k \\ \mathbf{z}^k \end{bmatrix}, \mathbf{p}^k = \begin{bmatrix} \mathbf{r}^k \\ \mathbf{x}^k \end{bmatrix}, \mathbf{G} = \begin{bmatrix} c\mathbf{I} & \mathbf{0} \\ \mathbf{0} & c\mathbf{L}_+/2 \end{bmatrix}.$$

For a positive semidefinite matrix $\mathbf{X}$, we use $\sigma_{\min}(\mathbf{X})$ as the nonzero smallest eigenvalue of matrix $\mathbf{X}$ and $\sigma_{\max}(\mathbf{X})$ as the nonzero largest eigenvalue in sequel.

5.3.1 Convex Case

In this case, we assume convexity for the cost function and analyze the convergence of the ADMM algorithm in the presence of errors. First, we present the convergence of the function values in terms of the current update and averaged update.

Theorem 5.1. *There exists* $\mathbf{p} = \begin{bmatrix} \mathbf{r} \\ \mathbf{x}^* \end{bmatrix}$ *with* $\mathbf{r} = 0$ *such that*

$$f(\mathbf{x}^T) - f(\mathbf{x}^*) \leq \|\mathbf{q}^{T-1} - \mathbf{p}\|_{\mathbf{G}}^2, \text{ and} \tag{5.6}$$

$$f(\hat{\mathbf{x}}_T) - f(\mathbf{x}^*) \leq \frac{\|\mathbf{p}^0 - \mathbf{p}\|_{\mathbf{G}}^2}{T} + \frac{c}{T}\sum_{k=1}^{T}\left(\frac{\sigma_{\max}^2(\mathbf{L}_+)}{2\sigma_{\min}(\mathbf{L}_-)}\|\mathbf{e}^k\|_2^2 + \langle \mathbf{e}^k, 2\mathbf{Q}\mathbf{r}^k)\rangle\right). \tag{5.7}$$

where $\hat{\mathbf{x}}_T = \sum_{k=1}^{T}\mathbf{x}^k/T$.

Proof. See Appendix A.9 □

Theorem 5.1 provides the upper bound for the residual of the function value over the iterations, and shows how errors accumulate and affect the convergence of the algorithm. In (5.6), the effect of the errors that occurred before the T-th iteration is represented by $\mathbf{q}^{T-1}$, which means that the previous errors have accumulated to impact the current algorithm state. It is observed in (5.7) that the averaged function value approaches the neighborhood of the minimum function value in a

sub-linear fashion, and the second term on the right hand side of the bound represents the radius of this neighborhood. It also shows that the algorithm converges sub-linearly if after a certain number of iterations, there are no errors in the updates. Compared to the convergence rate of $\mathcal{O}(\frac{1}{T})$ with decentralized ADMM for convex programming, e.g., [71], our result is very different. In the presence of errors, the algorithm converges to the neighborhood of the minimizer with a rate of $\mathcal{O}(\frac{1}{T})$ as well, but the true convergence to the minimizer cannot be guaranteed. The bounds are obtained in the form of the $\mathbf{G}$ norm. Recall the definition of $\mathbf{G}$, we can see that the structure of the network also plays a role in bounding the residual of the function value. Both the bounds show that a network with smaller $\sigma_{\max}(\mathbf{L}_+)$ (which is proportional to the network connectivity) is more resilient to errors. Intuitively, a less connected network can lower the spread of the errors. However, a more connected network has a faster convergence speed. This observation also highlights a potential trade-off between the resilience and the convergence speed.

5.3.2 Strongly Convex & Lipschitz Continuous Case

We assume that $f(\mathbf{x})$ is v-strongly convex and L-smooth, and provide the convergence analysis.

Theorem 5.2. *There exists* $\mathbf{q}^* = \begin{bmatrix} \mathbf{r}^* \\ \mathbf{x}^* \end{bmatrix}$ *such that for the* k*-th iteration,*

$$\|\mathbf{q}^k - \mathbf{q}^*\|_{\mathbf{G}}^2 \leq \frac{\|\mathbf{q}^{k-1} - \mathbf{q}^*\|_{\mathbf{G}}^2}{1+\delta} + \frac{P\|\mathbf{e}^k\|_2^2 + \langle \mathbf{e}^k, \mathbf{s}\rangle}{1+\delta}$$

with $\mathbf{s} = c\mathbf{L}_+(\mathbf{z}^k - \mathbf{z}^{k-1}) + 2c\mathbf{Q}(\mathbf{r}^k - \mathbf{r}^*) + 2c\mathbf{W}(\mathbf{x}^k - \mathbf{x}^*)$, *where* $P = \frac{c^2\delta\lambda_2\sigma_{\max}^2(\mathbf{W})}{\sigma_{\min}^2(\mathbf{Q})} + \frac{c^2\delta\lambda_3\sigma_{\max}^2(\mathbf{L}_+)}{4}$, *and*

$$\delta = \min\left\{ \frac{(\lambda_1 - 1)(\lambda_2 - 1)\sigma_{\min}^2(\mathbf{Q})\sigma_{\min}^2(\mathbf{L}_+)}{\lambda_1\lambda_2\sigma_{\max}^2(\mathbf{L}_+)}, \ \frac{4v(\lambda_2 - 1)(\lambda_3 - 1)\sigma_{\min}^2(\mathbf{Q})}{\lambda_1\lambda_2(\lambda_3 - 1)L^2 + c^2\lambda_3(\lambda_2 - 1)\sigma_{\max}^2(\mathbf{L}_+)\sigma_{\min}^2(\mathbf{Q})} \right\}$$

with quantities λ_1*,* λ_2*, and* λ_3 *being greater than 1.*

Proof. See Appendix A.10. □

Theorem 5.2 shows that the sequence $\|\mathbf{q}^k - \mathbf{q}^*\|_{\mathbf{G}}^2$ converges linearly with a rate of $\frac{1}{1+\delta}$ if after a certain number of iterations, there are no errors in the updates. Then, it can be easily shown that the sequence $\mathbf{z}^k$ or $\mathbf{x}^k$ converges to the minimizer. However, if the errors persist in the updates, this theorem shows how the errors are accumulated after each iteration. As a general result, one can further optimize over λ_1, λ_2, and λ_3 to obtain maximal δ and minimal P to achieve fastest convergence and least impact from the errors.

Theorem 5.3. *Choose* $0 < \beta \leq \frac{b(1+\delta)\sigma_{\min}^2(\mathbf{L}_+)\left(1-\frac{1}{\lambda_4}\right)}{4b\sigma_{\min}^2(\mathbf{L}_+)\left(1-\frac{1}{\lambda_4}\right)+16\sigma_{\max}^2(\mathbf{W})}$ *where* $b > 0$ *and* $\lambda_4 > 1$*, then*

$$\|\mathbf{z}^k - \mathbf{z}^*\|_2^2 \leq B^k \left(A + \sum_{s=1}^{k} B^{-s} C \|\mathbf{e}^s\|_2^2 \right)$$

where $A = \|\mathbf{z}^0 - \mathbf{z}^*\|_2^2 + A_2\|\mathbf{r}^0 - \mathbf{r}^*\|_2^2$ *with* $A_2 = \frac{4}{(1+4\beta)\sigma_{\max}^2(\mathbf{L}_+)}$*, and* $B = \frac{(1+4\beta)\sigma_{\max}^2(\mathbf{L}_+)}{(1-b)(1+\delta-4\beta)\sigma_{\min}^2(\mathbf{L}_+)}$, $C = \frac{4P+2/\beta}{c^2(1-b)(1+\delta-4\beta)\sigma_{\min}^2(\mathbf{L}_+)} + \frac{b(\lambda_4-1)}{1-b}$.

Proof. See Appendix A.11. □

Theorem 5.3 presents a general convergence result for ADMM for decentralized consensus optimization with errors, and indicates that the erroneous update $\mathbf{z}^k$ approaches the neighborhood of the minimizer in a linear fashion. The radius of the neighborhood is given as $B^k \sum_{s=1}^{k} B^{-s} C \|\mathbf{e}^s\|_2^2$. Note that B is not guaranteed to be less than 1. This is very different from the convergence result of ADMM for decentralized consensus optimization [101], which can guarantee that the update converges to the minimizer linearly fast and the corresponding rate is less than 1. Additionally, if $\sigma_{\max}^2(\mathbf{L}_+) >> \sigma_{\min}^2(\mathbf{L}_-)$, and it ends up with B being greater than 1, then the algorithm might not converge at all. We show later in the experiments that the ADMM algorithm can indeed diverge.

Thus, the first problem that follows is to guarantee that B is within the range $(0, 1)$, and the second one is to minimize the radius of the neighborhood by minimizing C. Accordingly, we optimize over the variables that appeared in the above theorems and the algorithm parameter c, and give the convergence result with $B \in (0, 1)$ in the following theorem.

Theorem 5.4. *If b and λ_2 can be chosen, such that*

$$(1-b)(1+\delta)\sigma_{\min}^2(\mathbf{L}_+) > \sigma_{\max}^2(\mathbf{L}_+) \tag{5.8}$$

with $\delta = \frac{(\lambda_2-1)}{\lambda_2}\frac{2v\sigma_{\min}^2(\mathbf{Q})\sigma_{\min}^2(\mathbf{L}_+)}{L^2\sigma_{\min}^2(\mathbf{L}_+)+2v\sigma_{\max}^2(\mathbf{L}_+)}$, then the ADMM algorithm with a parameter $c = \sqrt{\frac{\lambda_1\lambda_2(\lambda_3-1)L^2}{\lambda_3(\lambda_2-1)\sigma_{\max}^2(\mathbf{L}_+)\sigma_{\min}^2(\mathbf{Q})}}$ converges linearly with a rate of $B \in (0,1)$, to the neighborhood of the minimizer where $\lambda_1 = 1 + \frac{2v\sigma_{\max}^2(\mathbf{L}_+)}{L^2\sigma_{\min}^2(\mathbf{L}_+)}$, $\lambda_3 = 1 + \sqrt{\frac{L^2\sigma_{\min}^2(\mathbf{L}_+)+2v\sigma_{\max}^2(\mathbf{L}_+)}{\beta\lambda_1 L^2 v\sigma_{\min}^2(\mathbf{L}_+)}}$ and

$$0 < \beta \leq \min\left\{ \frac{b(1+\delta)\sigma_{\min}^2(\mathbf{L}_+)\left(1-\frac{1}{\lambda_4}\right)}{4b\sigma_{\min}^2(\mathbf{L}_+)\left(1-\frac{1}{\lambda_4}\right)+16\sigma_{\max}^2(\mathbf{W})}, \ \frac{(1-b)(1+\delta)\sigma_{\min}^2(\mathbf{L}_+)-\sigma_{\max}^2(\mathbf{L}_+)}{4\sigma_{\max}^2(\mathbf{L}_+)+4(1-b)\sigma_{\min}^2(\mathbf{L}_+)} \right\}.$$

Proof. See Appendix A.12. □

Theorem 5.4 provides an optimal set of choices of variables and the algorithm parameter such that $B \in (0,1)$ and C is minimized in Theorem 5.3. Recalling condition (5.8), it is equivalent to

$$\frac{\sigma_{\min}^2(\mathbf{L}_+)}{\sigma_{\max}^2(\mathbf{L}_+)} > \frac{4v}{\sqrt{(L^2+2v)^2+16v^2\frac{\lambda_2-1}{\lambda_2}\sigma_{\min}^2(\mathbf{Q})}-L^2+2v}. \tag{5.9}$$

As the only condition for the convergence, we show in our experiments that it can be easily satisfied.

Remark 5.1. *The value of $\frac{\sigma_{\min}^2(\mathbf{L}_+)}{\sigma_{\max}^2(\mathbf{L}_+)}$, which corresponds to the network structure, has to be greater than a certain threshold such that $B \in (0,1)$ can be achieved. This shows that a decentralized network with a random structure may not converge at all to the neighborhood of the minimizer, in the presence of errors in iteration.*

Remark 5.2. *The right hand side of inequality* (5.9) *is upper bounded by $\frac{4v}{(\sqrt{2}-1)L^2+(2\sqrt{2}+2)v}$, which depends on the geometric properties of the cost function. There exists a certain class of cost functions (e.g., v is small, L is large), such that a more flexible network structure design is allowed for a linear convergence to the neighborhood of the minimizer.*

Corollary 5.1. *When* (5.9) *is satisfied, the first condition below achieves linear convergence to the neighborhood of the minimizer with a radius of* $\frac{Ce}{1-B}$*, and either of the last two conditions guarantees linear convergence to the minimizer*

- $\|\mathbf{e}^{k-1}\|_2^2 \leq e$
- $\|\mathbf{e}^{k}\|_2^2$ *decreases linearly at a rate* R *such that* $0 < R < B$
- $C\|\mathbf{e}^{k}\|_2^2 \leq B(A_1 - A_2)\|\mathbf{r}^{k-1} - \mathbf{r}^{*}\|_2^2$ *with* $A_1 = \frac{4}{(1-b)\sigma_{\min}^2(\mathbf{L}_+)}$

Proof. See Appendix A.13. □

The first result in Corollary 5.1 simply states that if the error at every iteration is bounded, then the algorithm will approach the bounded neighborhood of the minimizer, and the second result states that if the error in the update decays faster than the distance between the update and the minimizer $\|\mathbf{z}^k - \mathbf{z}^*\|_2^2$, then the algorithm will reach the minimizer at a linear rate. The third result provides a much more general condition for convergence to the minimizer, which gives an upper bound for the current error based on the past errors, such that the network can tolerate the accumulated errors and the convergence to the minimizer can still be guaranteed.

5.4 Robust Decentralized ADMM Algorithm (ROAD)

Based upon insights provided by our theoretical results in Section 5.3, we investigate the design of the robust ADMM algorithm which can tolerate the errors in the ADMM updates. We focus on the scenario where a fraction of the agents generate erroneous updates. The remaining agents in the network follow the protocol and generate true updates, which are referred to as reliable agents[2] in this thesis. We refer to our proposed robust ADMM algorithm as "ROAD" (Algorithm 1).

[2]We also assume that reliable neighbors are in a majority for each agent i in the network.

Algorithm 2 ROAD($\mathbf{x}^0, c, \alpha^0, T, U$)

function $f = \sum_{i=1}^{D} f_i(\mathbf{x})$
  Initialization: $\mathbf{x}^0 = 0$, c, $\alpha^0 = 0$, T, U
  for $k = 1$ to T **do**
    For the node i :
    if $\sum_{t=1}^{k} \|\mathbf{x}_i^t - \mathbf{x}_j^t\| > U$, $j \in \mathcal{N}_i$, **then**
      Replace $\mathbf{x}_j^k$ with $\mathbf{x}_i^k$ in current update (5.4)
    else
      Use $\mathbf{x}_j^k$ in current update (5.4)
    end if
  end for
  Output $\mathbf{x}^T$
end function

Lemma 5.1. *In the error-free case, starting from* $\mathbf{x}^0 = 0$*, we have*

$$\frac{1}{T}\sum_{k=1}^{T} \|\mathbf{Q}\mathbf{x}^k\| \leq \frac{1}{4T}\left(\sigma_{\max}(\mathbf{L}_+)V_1^2 + \frac{2V_2^2}{\sigma_{\min}(\mathbf{L}_-)c^2} + 4\right). \tag{5.10}$$

To explain the idea behind ROAD, let us define two crucial variables used in the algorithm: I) deviation statistics $Z(k) = \sum_{t=1}^{k} \|\mathbf{Q}\mathbf{z}^t\|$, and II) threshold $U = \left(\sigma_{\max}(\mathbf{L}_+)V_1^2 + \frac{2V_2^2}{\sigma_{\min}(\mathbf{L}_-)c^2} + 4\right)/2\sqrt{2}$. The deviation statistics accumulates agents' update deviation from each other over ADMM iterations. Next, we obtain an upper bound on the deviation statistics for the error-free case. Specifically, if there were no errors in the updates from the neighbors, we show in Lemma 5.1 that $Z(k) \leq U/\sqrt{2}$. This upper bound U serves as a threshold to identify unreliable agents. Note that $Z(k) = \frac{1}{\sqrt{2}}\sum_{t=1}^{k}\sum_{(i,j)\in\mathcal{V}} \|\mathbf{z}_i^k - \mathbf{z}_j^k\|$, thus, we have $\frac{1}{\sqrt{2}}\sum_{t=1}^{k} \|\mathbf{z}_i^k - \mathbf{z}_j^k\| \leq Z(k) \leq U/\sqrt{2}$, $\forall (i,j) \in \mathcal{V}$. Inspired by this relationship, each agent i maintains the local deviation statistics $\sum_{t=1}^{k} \|\mathbf{z}_i^k - \mathbf{z}_j^k\|$ for every neighboring agent $j \in \mathcal{N}_i$ and compares it with the threshold U to identify if neighboring agent j is providing erroneous updates. For a reliable node j, the statistic $\sum_{t=1}^{k} \|\mathbf{z}_i^t - \mathbf{z}_j^t\|$ will not exceed the threshold U. If the statistic $\sum_{t=1}^{k} \|\mathbf{z}_i^t - \mathbf{z}_j^t\|$ exceeds the threshold U, the neighboring agent j is labeled as unreliable and its update is not be used by agent i. To avoid network disconnection in the case of unreliable neighbors, the link $\{i, j\}$ would not be cut off, however, the update from j will be replaced by node i's own value. Next, we show

in Theorem 5.5 that the proposed ROAD algorithm converges to the optimum at a rate of $\mathcal{O}(1/T)$.

Theorem 5.5. *For convex function $f(\mathbf{x})$, there exists $\mathbf{p} = \begin{bmatrix} \mathbf{r} \\ \mathbf{x}^* \end{bmatrix}$ with $\mathbf{r} = 0$, and ROAD provides*

$$f(\hat{\mathbf{x}}_T) - f(\mathbf{x}^*) \leq \frac{1}{T}\left(\|\mathbf{p}^0 - \mathbf{p}\|^2_{\mathbf{G}} + 8c\frac{\sigma^2_{\max}(\mathbf{L}_+)}{\sigma^2_{\min}(\mathbf{L}_-)}E^2U^2\right) \tag{5.11}$$

where $\hat{\mathbf{x}}_T = \sum_{k=1}^{T} \mathbf{x}^k/T$, and $U = \left(\sigma_{\max}(\mathbf{L}_+)V_1^2 + \frac{2V_2^2}{\sigma_{\min}(\mathbf{L}_-)c^2} + 4\right)/2\sqrt{2}$.

Proof. See Appendix A.16. □

Theorem 5 shows that the ROAD achieves a sub-linear convergence rate of $\mathcal{O}(1/T)$. Note that to account for the thresholding operation in ROAD, the upper bound in (10) introduces an additional term $8c\frac{\sigma^2_{\max}(\mathbf{L}_+)}{\sigma^2_{\min}(\mathbf{L}_-)}E^2U^2$. ROAD still falls under the formulation in (4) and follows the general analysis framework considered in this thesis. In the next section, we will also show empirically that employing the algorithmic parameter c derived in Theorem 4 accelerates the convergence rate of ROAD.

5.5 Experimental Results

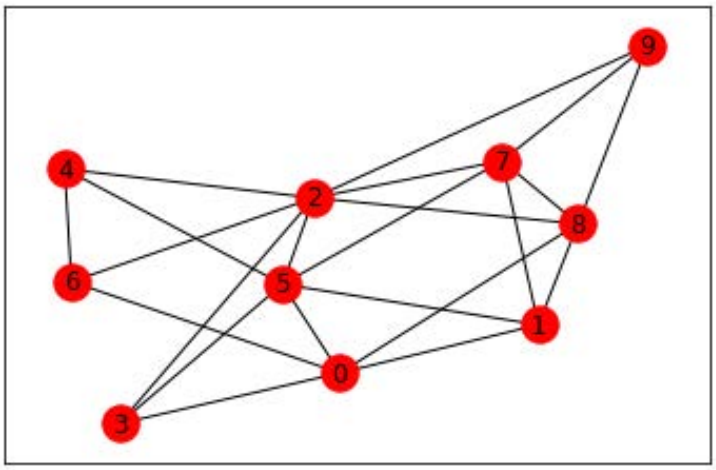


Fig. 5.1: Decentralized network topology.

In this section, we use ROAD to solve two different decentralized consensus optimization problems with $D = 10$ agents. We provide the network topology for the experiments in Figure 5.5. We assume that there are 3 unreliable agents (chosen randomly) in the network. Unreliable agents introduce errors in their updates by adding Gaussian noise[3] with mean μ_b and variance σ_b^2.

Consider a binary classification problem with a support vector machine, and the local cost function is

$$f_i(\mathbf{w}_i, b_i) = \frac{1}{2}\|\mathbf{w}_i\|_2^2 + C\sum_{j=1}^{N} \max(0, 1 - y_j(\mathbf{w}_i^T\mathbf{x}_j + b_i)).$$

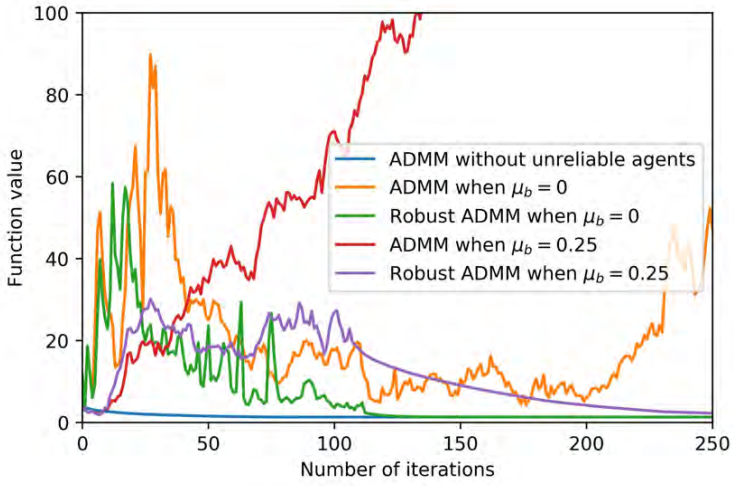
5.5.1 Synthetic Data

Here, the training set with $N = 1000$ sample points is equally partitioned across 10 agents. For each training point $\{\mathbf{x}_j, y_j\}$, $\mathbf{x}_j \in \mathbb{R}^2$ is the feature vector, and $y_j \in \{-1, 1\}$ is the corresponding label. We assume that $\mathbf{x}_j$ follows a normal distribution $\mathcal{N}([2.8, 2.8]^T, \mathbf{I})$ when $y_j = 1$, and $\mathcal{N}(\mathbf{0}, \mathbf{I})$ when $y_j = -1$, respectively. Locally, the training data is evenly composed of samples from two different distributions. In our experiment, each agent updates$\{\mathbf{w}, b\}$, and the whole network tries to reach a final consensus on a globally optimal solution. We choose the regularization parameter $c = 0.35$ in our experiment.We model the error injected by unreliable agents with distribution $\mathcal{N}(0, 1.5^2)$.

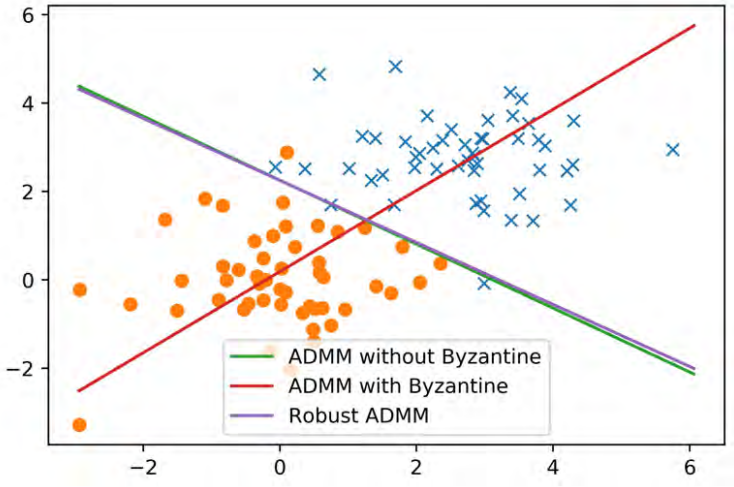
In Figure 5.5(a), we present the objective function value against the number of iterations for different algorithms. We observe that in the absence of unreliable agents, the original ADMM algorithm converges quickly and there are no function value fluctuations. When unreliable agents provide erroneous updates, ADMM algorithm diverges from the minimizer significantly. We can see that when the noise intensity μ_b is larger, the size of the neighborhood is larger. On the other hand, when ROAD is employed, we observe that the algorithm converges to the minimizer which corroborates our theoretical results in Theorem 5.

We show the classification results by depicting the hyperplane ($\mathbf{w}^T\mathbf{x} + b = 0$) in Figure 5.5(b).

[3]Note that our theoretical analysis and the proposed mitigation scheme (ROAD) do not assume that the error is of any parametric structure and are applicable to any arbitrary type of error.



(a)



(b)

Fig. 5.2: (a) Performance comparison with different noise intensities. (b) Classification with unreliable agents.

When there are unreliable agents, the algorithm learns an "incorrect" classifier as is shown by the red line. By using ROAD, we obtain a classifier which is almost the same as the case where there are no unreliable agents. The slight difference arises because the algorithms stop after the same number of iterations in our experiments, thus, ROAD does not achieve the same accuracy as error-free ADMM.

5.5.2 MNIST Dataset

MNIST is a dataset that contains 60000 training images and 10000 testing images of hand written digits. Each image is fit into a 28x28 pixel bounding box, making its dimension 784 for an individual sample. We extract the data samples of digits "1" and "5" for binary classification. To better visualize the result, we first use a deep neural net of autoencoder for mapping the 784-dimensional samples into 2-dimensional points as is shown in Figure 5.3.

Again, we run the decentralized ADMM algorithm to optimize the cost function for the classifier of the support vector machine. As is shown in Figure 5.3, when there are no unreliable agents in the network, the original ADMM algorithm gives a "sane" classifier inidcated by the blue line. However, when there are unreliable agents in the network, the original ADMM algorithm again gives an "insane" classifier inidcated by the green line. On the contrary, in the presence of unreliable agents in the network, the proposed robust scheme obtains a classification result as good as the case where there are no unreliable agents.

5.6 Summary

We considered the problem of decentralized learning using ADMM in the presence of unreliable agents. We studied the convergence behavior of the decentralized ADMM algorithm and showed that the ADMM algorithm converges to a neighborhood of the solution under certain conditions. We suggested guidelines for network structure design to achieve faster convergence. We also gave several conditions on the errors to obtain exact convergence to the solution. A robust variant of

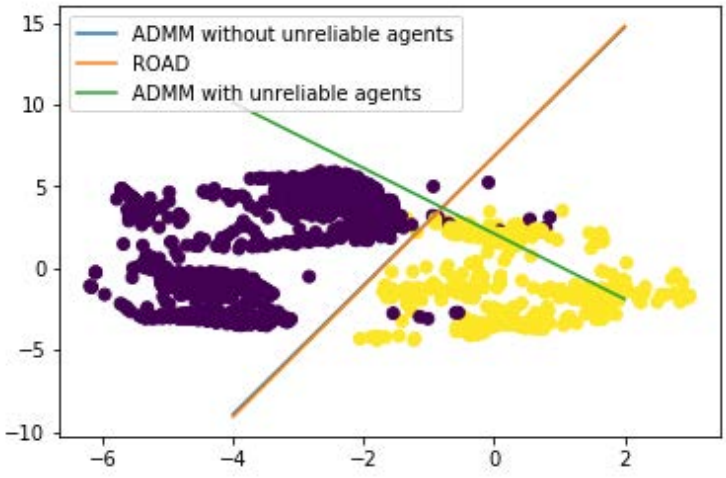


Fig. 5.3: Classification with 2-dimensional MNIST digits (1 and 5).

the ADMM algorithm was proposed to enable decentralized learning in the presence of unreliable agents and its convergence to the optima was proved. We also provided experimental results to validate the analysis and showed the effectiveness of the proposed robust scheme. We assumed the convexity of the cost function, and one might follow our lines of analysis for non-convex functions. Extension of the analysis and the algorithm to an asynchronous setting can also be considered.

CHAPTER 6
CLASSIFICATION: FUNDAMENTAL LIMITS

6.1 Introduction

Information theory was largely developed in the context of communication systems, where it plays an important role in characterizing performance limits. However, another important area where information theory has proved useful is in statistical inference, e.g., hypothesis testing, which is equivalent to statistical classification. We use the terms hypothesis testing and classification interchangeably in this chapter. For *parametric* hypothesis testing problems, information theoretic tools such as joint typicality, the equipartition property, and Sanov's theorem have been developed to characterize the error exponent [22, 23, 34]. Information theory has also been applied to investigate a class of *parametric* hypothesis testing problems [40, 41], where correlated data samples are observed over multiple terminals and data compression needs to be carried out in a decentralized manner. Additionally, information theory has also been applied to solve *nonparametric* hypothesis testing problems under the Neyman-Pearson framework [39, 66].

In this chapter, we apply information theoretic tools to study the nonparametric hypothesis testing problem, but with a focus on the average error probability instead of the Neyman-Pearson formulation. We address a more general scenario, where each hypothesis corresponds to a cluster of distributions. Such a nonparametric problem has not been thoroughly explored in the litera-

ture. We develop two nonparametric tests based respectively on the maximum mean discrepancy (MMD) and the Kolmogorov-Smirnov (KS) distance, and characterize the *exponential* error decay rate for these tests. Furthermore, in contrast to previous works where the number of hypotheses is assumed fixed, we study the regime where the number of hypotheses scales along with the sample size. This is analogous to the information theoretic channel coding problem where the number of messages scales along with the codeword length. Hence, in our study, information theory not only provides a technical tool to analyze the performance, but also provides an asymptotic perspective for understanding nonparametric hypothesis testing problems in the regime where the number of hypotheses is large, i.e., in the large-hypothesis large-sample regime.

More specifically, this chapter assumes that there are M hypotheses, each corresponding to a cluster of distributions, which are *unknown*. During the training phase, sequences of length-n training data samples generated by each distribution are available. The more general case with the training data sequences having different lengths is discussed in Section 6.3.4. Then, during the testing phase, a length n data stream is observed, consisting of samples generated by one of the distributions. The goal is to determine the cluster that contains the distribution that generated the observed test sequence. We are interested in the large-hypothesis regime, in which $M = 2^{nD}$, i.e., the number of hypotheses scales exponentially in the number of samples with a constant rate D. The analogy to the channel coding problem [22] is now apparent where the exponent represents the transmission rate, i.e., the transmitted bits per channel use, for the channel coding problem, here D represents the number of *hypothesis bits* that can be distinguished per observation sample. Correspondingly, we refer to D as the *discrimination rate*, and the largest such value is referred to as the *discrimination capacity*. The notion of *discrimination capacity* provides the fundamental performance limit for the hypothesis testing problem in the large-hypothesis and large-sample regime.

This chapter makes the following major contributions. First, we provide an asymptotic viewpoint to understand the nonparametric hypothesis testing problem in the regime where the number of hypotheses scales exponentially in the sample size. Based on its connection to the channel cod-

ing problem, we introduce the notions of the discrimination rate and the discrimination capacity as the performance metrics in such an asymptotic regime. Second, we develop two nonparametric approaches that are based respectively on the maximum mean discrepancy (MMD) and the Kolmogorov-Smirnov (KS) distance. For both tests, we derive the error exponents and the discrimination rates. Our results show that as long as the number M of hypotheses does not scale too fast, i.e., the scaling (discrimination) exponent is less than a certain threshold, the derived tests are exponentially consistent. For each algorithm, the proof of its discrimination rate is similar to the *achievability* proof in channel coding. Finally, we also derive an upper bound on the discrimination capacity, which serves as an upper limit beyond which exponential consistency cannot be achieved by any nonparametric composite hypothesis testing rule. This upper bound is based on the Fano minimax method, and is similar to the *converse* proof used in channel coding.

6.1.1 Related Work

Quite a few recent studies have applied various notions in information theory for studying hypothesis testing problems. A minimax approach for multi-hypothesis testing, where the goal is to minimize the worst-case expected loss function over a certain set of probability distributions was developed in [31]. The designed classification rules are expected to be robust over datasets generated by any probability distribution in the set. A classification problem, where the observation is obtained via a linear mapping of a vector input was studied in [83]. The notion of classification capacity was proposed, which is similar to the discrimination capacity we propose. However, the results in [83] are derived under the Gaussian model, whereas our formulation does not assume any specific distributions and is hence much more general. Furthermore, a parametric setting is implicitly assumed in [83], whereas our focus is on the nonparametric problem. A connection between the hypothesis testing problems and channel coding was established in [83], whereas in this paper we focus on the asymptotic case where the number of classes can grow exponentially large. A supervised learning problem, where the joint distribution of the data sample and its label is assumed to be known but with an unknown parameter, was studied in [84]. A classifier was

proposed and the corresponding performance was analyzed. The connection of the problem to rate-distortion theory was explored. There are several key differences between the work in [84] and our study. There is no notion of discrimination rate in [84], and the performance is not defined in terms of the asymptotic classification error probability. Additionally, our study does not assume any joint distribution of both the data sample and its label.

6.2 Problem Formulation

In this section, we first describe our composite nonparametric hypothesis model, and then connect it to the channel coding problem, which motivates several information theory related definitions that we will use to characterize system performance. For ease of readability, we also give preliminaries on the parametric hypothesis testing problem.

6.2.1 Classification as Nonparametric Hypothesis Testing

Consider the following nonparametric hypothesis testing problem with composite distributions. Suppose there are M hypotheses, and each hypothesis corresponds to a set $\mathcal{P}_m$ of distributions for $m = 1, \ldots, M$. For a given distance measure $d(p, q)$ between two probability distributions p and q, define

$$\begin{aligned} d(\mathcal{P}_m) &:= \sup_{p_i, p_{i'} \in \mathcal{P}_m} d(p_i, p_{i'}), \\ d(\mathcal{P}_m, \mathcal{P}_{m'}) &:= \inf_{p_i \in \mathcal{P}_m, p_{i'} \in \mathcal{P}_{m'}} d(p_i, p_{i'}) \quad \text{for } m \neq m'. \end{aligned} \tag{6.1}$$

Hence, $d(\mathcal{P}_m)$ represents the diameter of the m-th distribution set and $d(\mathcal{P}_m, \mathcal{P}_{m'})$ represents the inter-set distance between the mth and the m'th sets.

We assume that

$$\begin{aligned} \limsup_{M \to \infty} \sup_{m=1,\ldots,M} d(\mathcal{P}_m) &< D_I, \\ \liminf_{M \to \infty} \inf_{\substack{m,m'=1,\ldots,M \\ m \neq m'}} d(\mathcal{P}_m, \mathcal{P}_{m'}) &> D_O, \end{aligned} \tag{6.2}$$

where $D_I < D_O$. That is, the intra-set distance (diameter) is always smaller than the inter-set

distance for the composite hypothesis testing problem. The actual values of D_I and D_O depend on the distance metrics used. Furthermore, $\limsup_{M\to\infty}$ and $\liminf_{M\to\infty}$ in (6.2) require that the conditions hold in the limit of asymptotically large M, i.e., the limit taken over the sequences of distribution clusters. We study the case where none of the distributions in the sets $\mathcal{P}_m$ for $m = 1, \ldots, M$ are known. Instead, for $m = 1, \ldots, M$, we assume that each distribution $p_{m,i_m} \in \mathcal{P}_m$, where $i_m \in I_1^{M_m} = \{1, 2, \ldots, M_m\}$ is the index of the distribution, generates one training sequence $\mathbf{x}_{m,i_m} \in \mathbb{R}^n$ consisting of n independently and identically distributed (i.i.d.) scalar training samples. We use $\mathbf{X}_m$ to denote all training sequences generated by the distributions in $\mathcal{P}_m$. We assume that a test sequence $\mathbf{y} \in \mathbb{R}^n$ of n i.i.d. scalar samples is generated by one of the distributions in one of the sets $\mathcal{P}_m$. The goal is to determine the hypothesis that the test sequence $\mathbf{y}$ belongs to, i.e., which set contains the distribution that generated $\mathbf{y}$.

A practical example of the considered problem involves nonparametric detection of micro-Doppler modulated radar returns, such as those which occur in a ground moving target indicator (GMTI) radar [103]. The micro-Doppler motion of a particular target generates a specific sideband structure, which varies within a distributional radius as the fundamental frequency of the target's micro motion changes, i.e., D_I. The difference between the fundamental sideband structure of the micro-Doppler modulations for different target types implies a distributional difference, i.e., D_O. This problem is clearly composite (based on an unknown fundamental modulation frequency), and a parametric realization is in many cases impractical as the specific physics of the movement can be very difficult to model in a closed form.

Let $\delta(\{\mathbf{X}_m\}_{m=1}^M, \mathbf{y})$ denote a test based on the given data. Then, the error probability for δ is defined as

$$P_e = \sum_{m_0=1}^{M} P\left(\delta(\{\mathbf{X}_m\}_{m=1}^M, \mathbf{y}) \neq m_0 \middle| \mathbf{y} \sim p_{m_0,j} \in \mathcal{P}_{m_0}\right) \cdot P(m_0), \tag{6.3}$$

where $P(m_0)$ is the *a priori* probability that $\mathbf{y}$ is drawn from the m_0-th set of distributions.

For the above M-ary hypothesis testing problem, we are interested in the regime, in which the

number M of hypotheses scales with the number of samples. In particular, we assume $M = 2^{nD}$, where the parameter D captures how fast M scales with n. We refer to D as the *discrimination rate*.

Definition 6.1. *We say that the discrimination rate D is achievable, if there exists a classification rule δ such that the probability of error converges to zero as the number n of observation samples converges to infinity.*

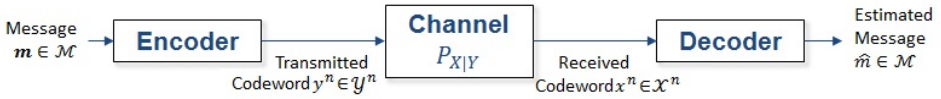
For a given composite hypothesis testing problem, we define the largest possible discrimination rate, D, to be the *discrimination capacity*, and denote it as $\bar{D}$.

6.2.2 Connection to the Channel Coding Problem

Next, we discuss the connection between the asymptotic regime of the hypothesis testing problem and the channel coding problem studied in communications, which in fact motivated our definition of the discrimination rate and the discrimination capacity.

In the channel coding problem (see Figure 6.1(a)), assume there are $\mathcal{M} = \{1, \ldots, 2^{nR}\}$ messages to be transmitted with equal probability. An encoder maps each message $m \in \mathcal{M}$ one-to-one onto a length-n codeword $y_m^n = \{y_{m1}, \ldots, y_{mn}\}$, which is transmitted over the channel. The channel maps each input symbol to an output symbol in a discrete memoryless fashion with the transition probability $P_{X|Y}(x|y)$ for each channel use, and the corresponding output sequence is given by $x^n = \{x_1, \ldots, x_n\}$. A decoder then estimates the original message as $\hat{m}$ based on the output sequence. Essentially, in the channel coding problem, there are a total of M possible conditional distributions $p_m(x^n) = P_{X|Y}(x^n|y_m^n)$ given y_m^n, where $m = 1, \ldots, M$, and the decoder determines which distribution $p^* \in \{p_1, \ldots, p_M\}$ most probably generated the observed channel output x^n.

The decoding process of the channel coding problem described above is a hypothesis testing problem. Inspired by the channel coding problem, our total number of hypotheses corresponds to the total number of messages in channel coding, and the discrimination rate D we define corresponds to the communication rate R in channel coding, which represents the transmitted message



(a) An illustration of the channel coding problem.

(b) An illustration of the multiple hypothesis testing problem.

Fig. 6.1: Illustrations of the channel coding problem and the multiple hypothesis testing problem

bits per coded symbol. By analogy, the discrimination rate D can be interpreted as the number of class-bits that can be distinguished per observation sample. Similarly, the discrimination capacity $\bar{D}$ corresponds to the capacity in channel coding, and serves as the fundamental limit in hypothesis testing problems. Note that in channel coding, the transmitter can choose to shape the distributions of transmitted symbols. Here, the hypothesis testing problem corresponds to the case where the distributions remain unshaped.

Essentially, Shannon's channel coding theorem guarantees error-free transmission of an exponentially increasing number of messages provided that the transmission rate R is less than the channel capacity C. In other words, Shannon's theorem implies that codewords $\{y^n\}$ can be designed such that exponentially increasing number of conditional probability distributions can be distinguished given the channel output. Here, for the hypothesis testing problem, channel coding motivates us to investigate the following problems:

- Which tests distinguish an exponentially increasing number of hypotheses with asymptotically small error probability based on n observation samples?
- What are the corresponding discrimination rates?

6.2.3 Preliminaries on Parametric Hypothesis Testing

The aforementioned questions can be answered for the *parametric* hypothesis testing problem in the asymptotic regime based on existing studies, e.g., [22]. We explain this in detail for single

distributions below as preliminary material before we delve into the main focus of this thesis on the *nonparametric composite hypothesis testing problem*.

Consider the *parametric* hypothesis testing problem, where there are $M = 2^{nD}$ *known* distinct distributions $p_1, ..., p_M$ corresponding respectively to M hypotheses. Given a test sequence $\mathbf{y}$ consisting of n i.i.d. samples generated from one of these distributions, the goal is to determine the true hypothesis, i.e., which distribution p_i generated the test sequence.

We apply the likelihood test given by:

$$\delta(\mathbf{y}) = \arg\max_i P_{X|H_i}(\mathbf{y}) \tag{6.4}$$

where the test labels the observed test data as hypothesis i if p_i generates $\mathbf{y}$ with the largest probability. It can be shown [22] that the likelihood test in (6.4) is equivalently given by

$$\delta(\mathbf{y}) = \arg\min_i D_{KL}(\gamma(\mathbf{y})\|p_i), \tag{6.5}$$

where $D_{KL}(\cdot\|\cdot)$ is the KLD between two distributions, and $\gamma(\cdot)$ is the empirical distribution of the sequence. It suggests that the testing rule labels the test data as hypothesis i if the empirical distribution of the test data is closest to p_i in KLD.

We next analyze the average error probability of the above testing rule as follows.

$$\begin{aligned} P_e &= \frac{1}{M}\sum_{j=1}^{2^{nD}} P(\delta(\mathbf{y}) \neq j|H_j) = \frac{1}{M}\sum_{j=1}^{2^{nD}} P(\exists\, i \neq j \text{ s.t. } \mathcal{E}_1|H_j) \\ &\leq \frac{1}{M}\sum_{j=1}^{2^{nD}}\sum_{i,i\neq j} P(\mathcal{E}_1|H_j) = \frac{1}{M}\sum_{j=1}^{2^{nD}}\sum_{i,i\neq j}\exp\{-nC(p_i,p_j)\} \\ &\leq 2^{nD - n\log e \liminf_{M\to\infty}\min_{1\leq i,j\leq M} C(p_i,p_j)}, \end{aligned} \tag{6.6}$$

where $C(p_i, p_j)$ denotes the Chernoff distance

$$C(p_i, p_j) = \max_{0 \le t \le 1} -\log \int [p_i(p)]^{1-t} [p_j(p)]^t dp, \tag{6.7}$$

and $\mathcal{E}_1$ denotes the event that given H_j, the KLD between $\mathbf{y}$ and p_j is greater than the KLD between $\mathbf{y}$ and p_i for some $i \neq j$, i.e., for $i \neq j$, $D_{KL}(\gamma(\mathbf{y}) \| p_i) < D_{KL}(\gamma(\mathbf{y}) \| p_j)$. Note that for simplicity, the default base for $\log$ in this chapter is 2. Thus, if $D \le \log e \liminf_{M \to \infty} \min_{1 \le i,j \le M} C(p_i, p_j)$, then the error probability is asymptotically small as n goes to infinity, which proves the following proposition.

Proposition 6.1. *For the parametric multiple hypothesis testing problem, the discrimination rate D is achievable if*

$$D \le \log e \liminf_{M \to \infty} \min_{1 \le i,j \le M} C(p_i, p_j).$$

Hence, for the discrimination rate to be positive, we require that the smallest pairwise Chernoff information be bounded away from zero for asymptotically large M, i.e., the limit taken over the sequences of distribution clusters.

6.3 Main Results

In this section, we obtain the performance bounds for the nonparametric hypothesis testing problem, with two different distance measures, i.e., MMD and KS distance.

6.3.1 MMD-Based Test

We construct a nonparametric test based on the MMD between two distributions p and q [35] defined as

$$\text{MMD}^2(p, q) := \|\mu_p - \mu_q\|_{\mathcal{H}}. \tag{6.8}$$

where $\mu_p(\cdot)$ maps a distribution p into an element in a reproducing kernel Hilbert space (RKHS) associated with a kernel $k(\cdot,\cdot)$ as

$$\mu_p(\cdot) = \mathbb{E}_p[k(\cdot, x)] = \int k(\cdot, x) dp(x). \tag{6.9}$$

An unbiased estimator of (6.8) based on n samples of $\mathbf{x} = \{x_1, \ldots, x_n\}$ generated by distribution p and m samples of $\mathbf{y} = \{y_1, \ldots, y_m\}$ generated by distribution q, is [35]:

$$\begin{aligned} \text{MMD}^2(\mathbf{x}, \mathbf{y}) = & \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k(x_i, x_j) \\ & + \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(y_i, y_j) - \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} k(x_i, y_j). \end{aligned} \tag{6.10}$$

Note that $x_i, y_i \in \mathbb{R}^d$, and the dimension $d \geq 1$.

We employ the MMD to measure the distance between the test sequence and the training sequences, and declare the hypothesis of the test sequence to be the same as the training sequence that has the smallest MMD to the test sequence. The constructed MMD-based nonparametric composite hypothesis test is given by

$$\delta_{\text{MMD}}(\{\mathbf{X}_m\}_{m=1}^{M}, \mathbf{y}) = \arg \min_{m, i_m} \text{MMD}^2(\mathbf{x}_{m,i_m}, \mathbf{y}). \tag{6.11}$$

The following theorem characterizes the average probability of error performance of the proposed MMD-based test under composite distributions.

Theorem 6.1. *Suppose the MMD-based test in* (6.11) *is applied to the nonparametric composite hypothesis testing problem under assumption* (6.2)*, where the kernel satisfies* $0 \leq k(x, y) \leq \mathcal{K}$ *for all* (x, y)*. Then, the average probability of error is upper bounded as*

$$P_e \leq 2^{nD} \exp\left(-\frac{n(D_O - D_I)^2}{96\mathcal{K}^2} \right).$$

Thus, the achievable discrimination rate is

$$D = \frac{\log e}{96\mathcal{K}^2}(D_O - D_I)^2. \tag{6.12}$$

Proof. See Appendix A.17. □

Next, we study a special case where each hypothesis is associated with a single distribution, i.e., the m-th hypothesis is associated with only one distribution p_m, $m = 1, \ldots, M$. Then, we have the following corollary.

Corollary 6.1. *Suppose the MMD-based test is applied to the nonparametric hypothesis testing problem under assumption* (6.2)*, and each hypothesis is associated with a single distribution, where the kernel satisfies* $0 \leq k(x, y) \leq \mathcal{K}$ *for all* (x, y)*. Then, the average probability of error under equally probable hypotheses is upper bounded as*

$$P_e \leq 2^{nD - n\frac{\log e}{96\mathcal{K}^2}\liminf_{M\to\infty}\min_{1\leq i,j\leq M} MMD^4(p_i,p_j)}. \tag{6.13}$$

Thus, the achievable discrimination rate is

$$D = \frac{\log e}{96\mathcal{K}^2}\liminf_{M\to\infty}\min_{1\leq i,j\leq M} \boldsymbol{MMD}^4(p_i, p_j). \tag{6.14}$$

Note that, for the discrimination rate to be positive, we require the smallest pairwise MMD between the distributions to be bounded away from zero for asymptotically large M, where the limit is taken over the sequences of distribution clusters.

Proof. By Theorem 6.1, we set $D_I = 0$ and $D_O = \liminf_{M\to\infty}\min_{1\leq i,j\leq M} \text{MMD}^2(p_i, p_j)$. Therefore, we

can bound the probability of error as the number of classes scales according to $M = 2^{nD}$

$$P_e \leq M \exp\left(-\frac{n \liminf_{M\to\infty} \min_{1\leq i,j\leq M} \text{MMD}^4(p_i, p_j)}{96\mathcal{K}^2}\right)$$
$$\leq 2^{nD - n\frac{\log e}{96\mathcal{K}^2} \liminf_{M\to\infty} \min_{1\leq i,j\leq M} \text{MMD}^4(p_i,p_j)}. \tag{6.15}$$

Then, it is straightforward to obtain the achievable discrimination rate for the MMD test as

$$D = \frac{\log e}{96\mathcal{K}^2} \liminf_{M\to\infty} \min_{1\leq i,j\leq M} \text{MMD}^4(p_i, p_j). \tag{6.16}$$

□

6.3.2 Kolmogorov-Smirnov Test

In this section, we construct a nonparametric hypothesis testing test based on the KS distance defined as follows. Suppose $\mathbf{x} = \{x_1, \ldots, x_n\}$, and i.i.d. samples $x_i \in \mathbb{R}$, are generated by the distribution p. Then the empirical CDF of p is given by

$$F_{\mathbf{x}}(a) = \frac{1}{n} \sum_{i=1}^{n} 1_{[-\infty,a]}(x_i), \tag{6.17}$$

where $1_{[-\infty,x]}$ is the indicator function. The KS distance between $\mathbf{x}$ and $\mathbf{y}$ having respectively been generated by p and q is defined as

$$D_{KS}(\mathbf{x}, \mathbf{y}) = \sup_{a\in\mathbb{R}} |F_{\mathbf{x}}(a) - F_{\mathbf{y}}(a)|. \tag{6.18}$$

We construct the following KS based nonparametric composite hypothesis test

$$\delta_{KS}(\{\mathbf{X}_m\}_{m=1}^M, \mathbf{y}) = \arg\min_{m,i_m} D_{KS}(\mathbf{x}_{m,i_m}, \mathbf{y}), \tag{6.19}$$

The following theorem characterizes the performance of the proposed KS-based test.

Theorem 6.2. *Suppose the KS-based test in* (6.19) *is applied to the nonparametric hypothesis testing problem under assumption* (6.2)*. Then, the average probability of error is upper bounded as*

$$P_e \leq 6 \cdot 2^{nD} \exp\big(-\frac{n(D_O - D_I)^2}{8}\big).$$

Thus, the achievable discrimination rate is

$$D = \frac{\log e}{8}(D_O - D_I)^2. \tag{6.20}$$

Proof. See Appendix A.18. □

Consider the case where each hypothesis is associated with a single distribution, and we have the following corollary.

Corollary 6.2. *Suppose the KS-based test is applied to the nonparametric hypothesis testing problem under assumption* (6.2)*, and each hypothesis is associated with a single distribution. Then, the average probability of error under equally probable hypotheses is upper bounded as*

$$P_e \leq 6 \cdot 2^{nD - n\frac{\log e}{8} \liminf_{M\to\infty} \min_{1\leq i,j\leq M} d_{KS}^2(p_i,p_j)}. \tag{6.21}$$

Thus, the achievable discrimination rate is

$$D = \frac{\log e}{8} \liminf_{M\to\infty} \min_{1\leq i,j\leq M} d_{KS}^2(p_i, p_j). \tag{6.22}$$

Hence, for the discrimination rate to be positive, we require the least pairwise KS distance

between distributions to be bounded away from zero for asymptotically large M, where the limit is taken over the sequences of distribution clusters.

Proof. By Theorem 6.2, we set $D_I = 0$ and $D_O = \liminf_{M\to\infty} \min_{1\le i,j\le M} d_{KS}(p_i, p_j)$, and have

$$P_e \le 6 \cdot 2^{nD - n\frac{\log e}{8} D_o^2} \le 6 \cdot 2^{nD - n\frac{\log e}{8} \liminf_{M\to\infty} \min_{1\le i,j\le M} d_{KS}^2(p_i,p_j)}.$$

Then, it is straightforward to obtain the achievable discrimination rate for the KS test. □

6.3.3 Upper Bound on the Discrimination Capacity

In this section, we provide an upper bound on the discrimination capacity for the composite hypothesis testing problem. Let h be a random index representing the actual hypothesis that occurs. We assume that h is uniformly distributed over the M hypotheses, and h' has the same distribution as m, but is independent from h. Then, Lemma 2.10 in [110] directly yields the following upper bound on the discrimination capacity $\bar{D}$.

Remark 6.1. *The discrimination capacity $\bar{D}$ is upper bounded as*

$$\bar{D} \le \limsup_{M\to\infty} \mathbb{E}_{h,h'} D_{KL}(p_h \| p_{h'}), \tag{6.23}$$

where $D_{KL}(\cdot\|\cdot)$ is the KLD between two distributions.

Note that the above limit $\limsup_{M\to\infty}$ is taken over the sequences of distribution clusters. In Appendix A.19, we provide an alternative but simpler proof based on Fano's inequality for the above upper bound, which is closely related to the proposed concept of discrimination capacity.

6.3.4 Training Sequences of Unequal Length

In this subsection, we discuss the impact of different number of training samples in different classes on the probability of error and the discrimination rate. Here, we still assume that there are n test

samples. To keep the problem formulation meaningful, we assume that the number M of classes increases exponentially with n at a rate D, i.e., $M = 2^{nD}$. To avoid notational confusion, we use the non-composite case, i.e., with each class corresponding to one distribution, to illustrate the idea. Suppose that each class, i.e., each distribution, generates $\gamma_m(n)$ training samples, for $m = 1, \ldots, M$, where $\gamma_m(n)$ represents the number of samples in the m-th class (as a function of n). Let $\gamma_{\min}(n) = \min_{1 \le m \le M} \gamma_m(n)$. In particular, for the MMD-based test, the probability of error can be bounded as

$$P_e \le 2^{n\left(D - \min\{1, \frac{\gamma_{\min}(n)}{n}\} \frac{\log e (D_O - D_I)^2}{96 \mathcal{K}^2}\right)}. \tag{6.24}$$

For the KS-based test, the probability of error can be bounded as

$$P_e \le 6 \cdot 2^{n\left(D - \min\{1, \frac{\gamma_{\min}(n)}{n}\} \frac{\log e (D_O - D_I)^2}{8}\right)}. \tag{6.25}$$

It can be seen that here the ratio $\frac{\gamma_{\min}(n)}{n}$ plays an important role in determining the error exponent asymptotically. For example, for the MMD-based test, if the ratio converges to zero for large n, i.e., the shortest training length $\gamma_{\min}(n)$ scales as an order-level slower than the test length, then there is no guarantee of exponential error decay, and the discrimination rate equals zero. On the other hand, if $\lim_{n\to\infty} \frac{\gamma_{\min}(n)}{n} = c$ with $0 < c < 1$, then the discrimination rate $D = c\frac{\log e (D_O - D_I)^2}{96 \mathcal{K}^2}$. Furthermore, if $\lim_{n\to\infty} \frac{\gamma_{\min}(n)}{n} = c$ with $c \ge 1$, then the discrimination rate $D = \frac{\log e (D_O - D_I)^2}{96 \mathcal{K}^2}$. A sketch of the proof of (6.24) and (6.25) can be found in Appendix A.20

6.4 Numerical Results

In this section, we present numerical results to compare the performance of the proposed tests. In the experiment, the number of classes is set to be five, and the error probability versus the number of samples for the proposed algorithms is plotted. For the MMD based test, we use the standard Gaussian kernel given by $k(x, x') = \exp(-\frac{\|x - x'\|^2}{2})$.

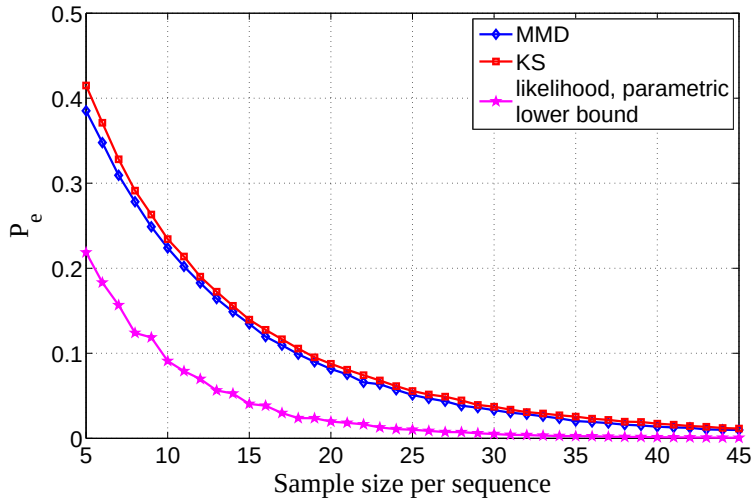


Fig. 6.2: Error probabilities of different hypothesis testing algorithms for Gaussian distributions with different means.

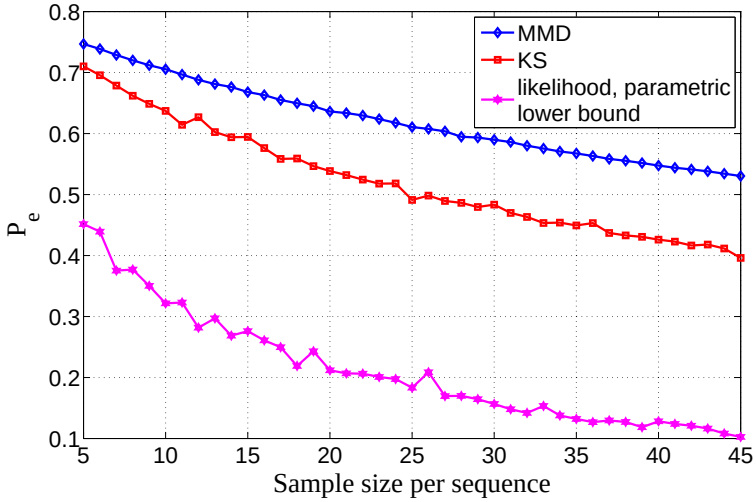


Fig. 6.3: Error probabilities of different hypothesis testing algorithms for Gaussian distributions with different variances.

In the first experiment, all the hypotheses correspond to Gaussian distributions with the same variance $\sigma^2 = 1$ but different mean values $\mu = \{-2, -1, 0, 1, 2\}$. A training sequence is drawn from each distribution and a test sequence is randomly generated from one of the five distributions. The sample size of each sequence ranges from 5 to 45. A total of 10^5 monte carlo runs are conducted. The simulation results are given in Figure 6.2. It can be seen that all the tests give better performance as the sample size n increases. We can also see that the MMD-based test slightly outperforms the KS-based test. We also provide results for the parametric likelihood test

as a lower bound on the probability of error for performance comparison. It can be seen that the performance of the two nonparametric tests are close to the parametric likelihood test even with a moderate number of samples.

In the second experiment, all the hypotheses correspond to Gaussian distributions with the same mean $\mu = 1$ but different variance values $\sigma^2 = \{0.5^2, 1^2, 1.5^2, 2^2, 2.5^2\}$. The simulation results are given in Fig. 6.3. In this experiment, the MMD-based test yields the worst performance, which suggests that this method is not suitable when the distributions overlap substantially with each other. The two simulation results also suggest that none of the three tests perform the best universally over all distributions. Although there is a gap between the performance of MMD and KS tests and that of the parametric likelihood test, we observe that the error decay rates of these tests are still close.

To show the tightness of the bounds derived in the paper, we provide a table (See Table I) of error decay exponents (and thus the discrimination rates) for different algorithms.

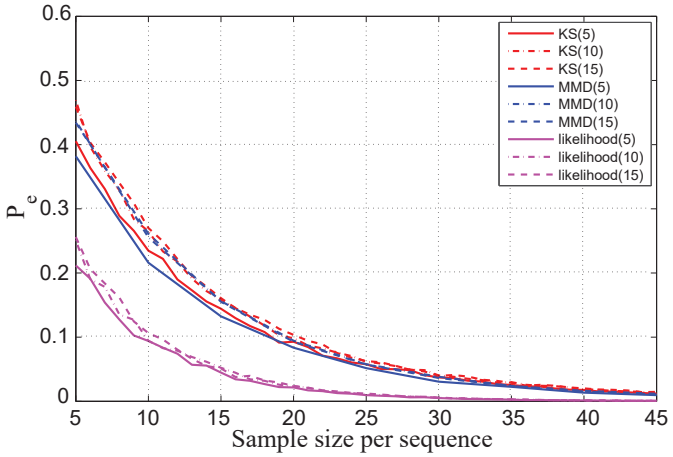


Fig. 6.4: Error probabilities for different hypothesis testing algorithms for Gaussian distributions with different means.

Estimates of error decay exponent of KS and MMD based tests on a multi-hypothesis testing problem are presented for the problem considered in the first experiment. Note that the theoretical lower bounds in the table correspond to the achievable discrimination rates of the methods asymptotically. Fano's bound (FB in the table) is estimated by using data-dependent partition estimators

of Kullback-Leibler divergence [120]. The parametric upper bound is based on the maximum likelihood test, which can serve as an upper bound on the error decay exponent (and hence intuitively on the discrimination capacity). It can be seen from the table that both the KS and MMD tests do achieve an exponential error decay and have positive discrimination rates as we show in our theorems. Clearly, the empirical values of the bounds for both tests are better than the corresponding theoretical values. More importantly, both of the empirical lower bounds are close to the likelihood upper bound, demonstrating that the actual performance of the two tests are satisfactory. We also note that the Fano's upper bound is not very close to the lower bound.

Table 6.1: Comparison of Bounds

	Lower Bounds		Upper Bounds	
	KS	MMD	Parametric	FB
Empirical	0.0897	0.0916	0.146	2.5
Theoretical	0.0183	0.0071	0.125	-

To better illustrate the bounds in Table I, we provide experimental results with different number of hypotheses M in Figure 4. In particular, we present the simulation results with $M = 5, 10, 15$. We use a similar experiment setting as that in the first experiment, where Gaussian distributions have the same variance and different mean values, and the mean values are $\{-2, -1, \ldots, 2\}$, $\{-4.5, -3.5, \ldots, 4.5\}$ and $\{-7, -6, \ldots, 7\}$ respectively. The parametric maximum likelihood test serves as an upper bound for the error decay exponent for all of the three cases. Similar to the case $M = 5$, KS and MMD nonparametric tests achieve an exponential error decay and hence the positive discrimination rates for the cases $M = 10$ and $M = 15$.

We now conduct experiments with composite distributions. First, we still use five hypotheses with Gaussian distributions with variance $\sigma^2 = 1$ and different mean values $\mu = \{-2, -1, 0, 1, 2\}$. For each hypothesis, we vary the mean values by ± 0.1. Thus, within each hypothesis, there are three different distributions with mean values in $\{\mu - 0.1, \mu, \mu + 0.1\}$. The results are presented in Figure 6.5. As expected, the performance improves as the sample size n increases. The two tests perform almost identically, with the MMD-based test slightly outperforming the KS-based test for small n.

We again vary the variances of the Gaussian distributions as in the second experiment in a similar way. In particular, the variances in the same class are $\{(\sigma-0.1)^2, \sigma^2, (\sigma+0.1)^2\}$, and $\sigma \in \{0.5, 1, 1.5, 2, 2.5\}$. In Figure 6.6, we observe the performance improvement as the sample size n increases. Different from the results in the second experiment, the MMD-based test outperforms the KS-based test in the composite setting.

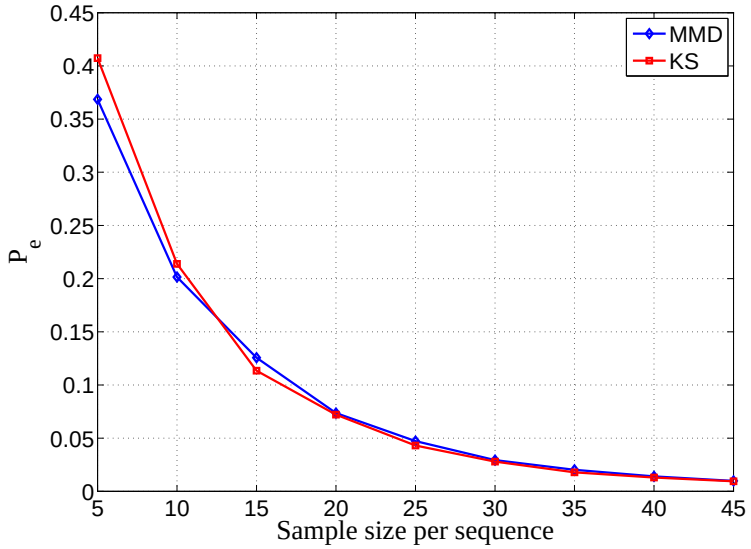


Fig. 6.5: Error probabilities of different hypothesis testing algorithms for composite Gaussian distributions with different means.

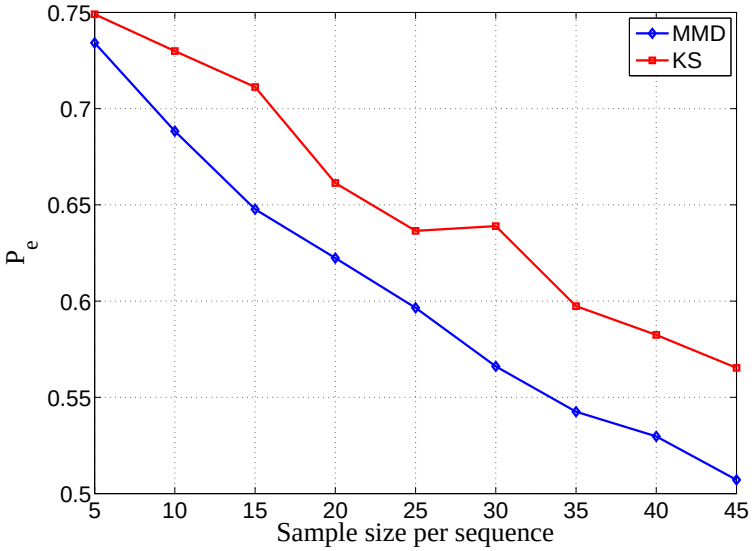


Fig. 6.6: Error probabilities of different hypothesis testing algorithms for composite Gaussian distributions with different variances.

6.5 Summary

A nonparametric composite hypothesis testing approach for arbitrary distributions based on the maximum mean discrepancy (MMD) and Kolmogorov-Smirnov (KS) distance measure based tests was developed in this chapter. We introduced the information theoretic notion of discrimination capacity that was defined for the regime where the number of hypotheses scales along with the sample size. We also provided characterization of the corresponding error exponent and the discrimination rate, i.e., a lower bound on the discrimination capacity. Our framework can been extended to unsupervised learning problems and similar performance limits can be investigated.

CHAPTER 7
CONCLUSION

7.1 Summary

In this thesis, the problem of accomplishing reliable classification from systems consisting of potentially unreliable agents was addressed. The general methodology for this was to first analyze the effect of unreliable agents in the network and quantify their effect on the global classification performance of the network. The second step was to design schemes that are robust to such unreliable information from these agents. These schemes used statistical and robust optimization approaches to correct the errors from individual agents and improve the classification performance at the global decision making or decision making stage. We also analyzed the classification performance limits from an information-theoretic perspective.

Human-driven Classification Systems

Chapter 2 considered the design of an effective crowdsourcing system for M-ary classification where crowd workers complete simple microtasks which are aggregated to give a final result. We considered the novel scenario where workers have a reject option so they may skip microtasks they are unable or unwilling to do. For example, in mismatched speech transcription, workers who do not know the language may be unable to respond to microtasks in phonological dimensions

outside their categorical perception. We presented an aggregation approach using a weighted majority voting rule, where each worker's response is assigned an optimized weight to maximize the crowd's classification performance. Since the individual statistical property of the human worker is essentially unavailable, we developed the aggregation rule such that the overall system classification performance can be maximized in an average sense when the entire crowd participates in the classification task. Thus, the aggregation rule doe not need individual information from the workers, and is highly practical. We evaluated system performance in both exact and asymptotic forms. We also showed that human workers' confidence does not help in improving the system performance.

In Chapter 3, we considered the presence of spammers, who give random responses to the microtasks, in the crowdsourced classification systems. First, we investigated the case where the crowd workers have no knowledge of the payment mechanism and the spammers respond to all the microtasks with random guesses. In this case, we developed a heuristic adaptive scheme for aggregation which switches between the oblivious strategy and the expurgation strategy, based on the estimated fraction of spammers in the crowd. Next, we studied the case where the crowd workers have the knowledge of the payment mechanism. The optimal strategy for the spammers to maximize their reward was derived. When the spammers adopt the optimal strategy, we derived an optimal aggregation rule to maximize classification performance. The classification performance in terms of probability of correct classification was provided when different aggregation rules were employed.

Chapter 4 considered the classification problem of influential node detection and volume time series prediction in implicit social networks. Implicit social networks do not assume the knowledge of the network structure. In such a case, we employed the linear information model to characterize the information flow of different contagions in the network. Additionally, we proposed a descriptive diffusion model to take dependencies among the topics into account. By exploiting Copula Theory and rank-constrained influence matrix, we modeled the complex dependency between different contagions and different users. We also proposed an efficient algorithm based on alternating

methods to perform classification and learning on the model. It was shown that the proposed technique outperforms existing influential node detection techniques. Furthermore, the proposed model was validated both on a synthetic and a real (ISIS) dataset. We showed that the proposed approach can efficiently select the influential users for specific contagions. We also presented several interesting patterns of the selected influential users for the ISIS dataset.

Data-driven Classification Systems

In Chapter 5, we considered the problem of decentralized learning of classification problems using ADMM in the presence of unreliable agents. The agents update their information by local computation and communication with their neighbors. Classification was formulated as an optimization problem, and the agents in the network collaboratively work towards a consensus of the classification result using the ADMM algorithm. The unreliable agents send erroneous information to their neighbors. We studied the convergence behavior of the decentralized ADMM algorithm and showed analytically that the ADMM algorithm converges to a neighborhood of the solution under certain conditions. We suggested guidelines for network structure design to achieve faster convergence. We also gave several conditions on the errors to obtain exact convergence to the solution. A robust variant of the ADMM algorithm was proposed to enable decentralized classification in the presence of unreliable agents and its convergence to the optima was proved. We also provided experimental results to validate the analysis and showed the effectiveness of the proposed robust scheme.

From the statistical signal processing point of view, classification is equivalent to hypothesis testing. Chapter 6 developed a nonparametric composite hypothesis testing approach for arbitrary distributions based on the maximum mean discrepancy (MMD) and Kolmogorov-Smirnov (KS) distance measure based tests, from the information theoretic perspective. Connecting to the notion of channel capacity in information theory, we introduced the information theoretic notion of discrimination capacity that is defined for the regime where the number of hypotheses scales along with the sample size. We also provided characterization of the corresponding error exponent and

the discrimination rate, i.e., a lower bound on the discrimination capacity.

7.2 Future Directions

There are a number of interesting future directions for research, which can be summarized into two specific research directions for human-driven and data-driven systems respectively:

1. Use of statistical modeling techniques to develop mathematical models of human decision making in collaboration with cognitive psychologists

2. To develop a universal robust scheme for decentralized classification.

These problems have both theoretical and implementation challenges. Therefore, the focus is first on developing theoretical models for such a collaboration and then, implementing the designed algorithms to verify their applicability in practice. Both these problems are further explained in detail below.

7.2.1 Human-driven Systems

The first step to develop efficient systems that include humans is to develop appropriate statistical models that characterize their behavior. Researchers have not extensively investigated 1) the modeling of decisions by humans; 2) the modeling of subjective confidences on multi-hypothesis tasks; 3) the modeling of tasks in which human decision makers can provide imprecise decisions. For example, it has been found in reality that some of the human workers form a colluder group, in which the rest of the group members follow one leader's decision. This has never been modeled properly in literature. In the work presented in Chapter 2 and Chapter 3, we assumed equal difficulty corresponding to different microtasks for the crowd workers to answer. In some settings, worker qualities are not identically distributed, which makes estimating the greedy fraction α difficult. The difficulty for microtasks might not be equal, which makes the microtask design quite challenging. Therefore, further research directions include the development of a general model to

characterize the crowd quality, design of a robust method to estimate the fraction of greedy workers in the crowd, and binary microtask design with a reject option. One can also build models that consider the effect of stress, anxiety, and fatigue in the cognitive mechanisms of human decision making, decision confidence assessment, and response time (similar to [93, 123]).

In social networks, one important factor in human decision making that we have not considered is that the human users often acquire noisy information from outside the contagions that we considered. As we assumed a fixed number of contagions, it is possible that the human users acquire noisy information from the other sources which can affect their decision making in the considered contagions. One can better model the information diffusion mechanism in social networks by taking this factor into account.

7.2.2 Data-driven Systems

In data-driven systems, one interesting direction of future work is to develop universally efficient fusion algorithms for decision making. In our work in Chapter 5, we analyzed the impact of unreliable agents when they send erroneous information to their neighbors, and proposed a robust algorithm to combat this impact. However, we considered the problem under a specific circumstance namely where ADMM is used as the algorithm to solve the classification problem. Although ADMM is widely used as an effective algorithm in decentralized learning problems,it is possible that the methodology developed in the proposed robust algorithm based on ADMM is not feasible for other decentralized learning algorithms. Thus, one can investigate the possibility of developing a universal robust methodology for decentralized learning that can result in robust schemes for different scenarios where different algorithms are employed. Additionally, the problem was studied in a synchronous setting, where all the updates are received at agents before the next iteration. However, in practice, some agents may not received all the updates from their neighbors before the next iteration starts. Thus, it is very important that the problem is investigated under an asynchronous setting and show whether or not the proposed robust algorithm still works.

For the characterization of performance limits, we proposed the notion of discrimination rate

in Chapter 6. We developed the discrimination capacity, which is universal for all the classification algorithms. Different algorithms have different associated discrimination rates. A universal discrimination rate bound is missing which characterizes the regime where the corresponding algorithm cannot perform the classification task. Additionally, we did not consider the unreliable agents in the systems. Since sample distance is required from different clusters of distributions, system performance could be drastically degraded if the distance is not reported correctly for classification. One interesting project can investigate the robustness in terms of discrimination rate.

APPENDIX A
APPENDIX

A.1 Proof of Proposition 2.2

To solve problem (4.2), we need $E_C[\mathbb{W}]$ and $E_O[\mathbb{W}]$. First, the wth worker can have weight contribution to $E_C[\mathbb{W}]$ only if all his/her definitive answers are correct. Thus, we have the average weight assigned to the correct element as

$$\begin{aligned} E_C[\mathbb{W}] &= E_{p,r}\left[\sum_{w=1}^{W}\sum_{n=0}^{N} W_w P\left(n, N-n\right)\right] \\ &= \sum_{w=1}^{W} E_{p,r}\left[\sum_{n=0}^{N} W_w P\left(n, N-n\right)\right], \end{aligned} \tag{A.1}$$

where $P(n, N-n)$ represents the probability of $N-n$ bits equal to λ and the rest of the n definitive answers in the N-bit word are correct.

Then, given the wth worker with $p_{w,i}$ known, we write

$$A_w(p_{w,i}) = E_r\left[\sum_{n=0}^{N} W_w P\left(n, N-n\right)|p_{w,i}\right]. \tag{A.2}$$

Let $P_\lambda(n)$ denote the probability of the wth worker submitting n definitive answers out of N microtasks, which only depends on $p_{w,i}$. Note that $\sum_{n=0}^{N} P_\lambda(n) = 1$, and then (A.2) is upper-

bounded using the Cauchy-Schwarz inequality as:

$$A_w(p_{w,i}) = \sum_{n=0}^{N} E_r\left[W_w r\left(n\right)\right]\sqrt{P_\lambda(n)}\sqrt{P_\lambda(n)}$$

$$\leq \sqrt{\sum_{n=0}^{N} E_r^2\left[W_w r\left(n\right)\right]P_\lambda(n)}\sqrt{\sum_{n=0}^{N} P_\lambda(n)} \tag{A.3}$$

$$\triangleq \alpha_w(p_{w,i}), \tag{A.4}$$

where $r(n)$ is the product of any n out of N variables $r_{w,i}$ as $i = 1, \ldots, N$, and α_w is a positive quantity independent of n, which might be a function of $p_{w,i}$. Note that equality holds in (A.3) only if

$$E_r\left[W_w r\left(n\right)\right]\sqrt{P_\lambda(n)} = \alpha_w(p_{w,i})\sqrt{P_\lambda(n)}, \tag{A.5}$$

which results in (A.4) and

$$E_r\left[W_w r\left(n\right)\right] = \alpha_w(p_{w,i}). \tag{A.6}$$

Then we maximize the crowd's average weight corresponding to the correct class under the constraint $\int_{p_{w,i}} \Pr(p_{w,i} = x)dx = 1$, and the maximization problem is written as

$$A = E_p[A_w(p_{w,i})] = \int_{p_{w,i}} \alpha_w(p_{w,i}) \Pr\left(p_{w,i} = x\right) dx$$

$$\leq \sqrt{\int_{p_{w,i}} \alpha_w^2(p_{w,i}) \Pr\left(p_{w,i} = x\right) dx}\sqrt{\int_{p_{w,i}} \Pr\left(p_{w,i} = x\right) dx} \tag{A.7}$$

$$\triangleq \beta. \tag{A.8}$$

The equality (A.7) holds only if

$$\alpha_w(p_{w,i})\sqrt{\Pr\left(p_{w,i}=x\right)}=\beta\sqrt{\Pr\left(p_{w,i}=x\right)}, \tag{A.9}$$

with β is a positive constant independent of $p_{w,i}$, and we conclude that α_w is also a positive quantity independent of $p_{w,i}$. Then from (A.6), we have

$$E_r\left[W_w r\left(n\right)\right]=\beta. \tag{A.10}$$

Since $r(n)$ is the product of n variables, its distribution is not known *a priori*. A possible solution to weight assignment is a deterministic value given by $W_w E[r(n)] = \beta$ and, therefore, we can write the weight as $W_w = \beta/\mu^n$. Note that if $r_{w,i}$ is known *a priori* or can be estimated, the optimal weight assignment is simply $W_w = \beta/r(n)$.

Then, we can express the crowd's average weight contribution to all the classes defined in (4.2) as

$$\begin{aligned} E_O\left[\mathbb{W}\right] &= \sum_{w=1}^{W} E_{p,r}\left[\sum_{n=0}^{N}\beta\mu^{-n}2^{N-n}P_\lambda\left(n\right)\right] \\ &= \sum_{w=1}^{W}\sum_{n=0}^{N}\beta\mu^{-n}2^{N-n}\binom{N}{n}(1-m)^n m^{N-n} \\ &= W\beta\left(\frac{1-m}{\mu}+2m\right)^N = K. \end{aligned} \tag{A.11}$$

Thus, β can be obtained from (A.11) and we get the weight by solving optimization problem (4.2) to get:

$$W_w = \frac{K}{W\mu^n\left(\frac{1-m}{\mu}+2m\right)^N}. \tag{A.12}$$

Note that the weight derived above has a term that is common for every worker. Since the voting scheme is based on comparison, we can ignore this factor and have the normalized weight as

$W_w = \mu^{-n}$.

A.2 Proof of Proposition 2.3

Note that

$$T_w \in \{-\mu^{-N}, -\mu^{-N+1}, \ldots, -\mu^{-1}, 0, \mu^{-1}, \ldots, \mu^{-N+1}, \mu^{-N}\}, \tag{A.13}$$

which can be written as

$$T_w = I(-1)^{t+1}\mu^{-n}, t \in \{0,1\}, I \in \{0,1\}, n \in \{1, \ldots, N\}, \tag{A.14}$$

and leads to

$$\begin{aligned} &\Pr\left(T_w = I(-1)^{t+1}\mu^{-n}|H_s\right) \\ &= \begin{cases} \Pr\left(T_w = \frac{(-1)^{t+1}}{\mu^n}|H_s\right), I = 1 \\ \Pr\left(T_w = 0|H_s\right), \qquad I = 0 \end{cases}. \end{aligned} \tag{A.15}$$

These two terms can be expressed as

$$\begin{aligned} &\Pr\left(T_w = \frac{(-1)^{t+1}}{\mu^n}|H_s\right) \\ &= \Pr\left(\mathbf{a}_w(i) = t|H_s\right) \cdot P_\lambda\left(n|\mathbf{a}_w(i) = t, H_s\right) \\ &= r_{w,i}^{1-|s-t|}(1 - r_{w,i})^{|s-t|} \sum_C \prod_{\substack{j=1 \\ j\neq i}}^{N} p_{w,j}^{k_j}(1 - p_{w,j})^{1-k_j}, \end{aligned} \tag{A.16}$$

and $\Pr\left(T_w = 0|H_s\right) = p_{w,i}$.

A.3 Proof of Proposition 2.4

Let $q_n, -N \leq n \leq N$, represent the number of workers that submit $|n|$ total definitive answers to all the microtasks. Specifically, $n < 0$ indicates the group of workers that submit "0" for the ith bit while $n > 0$ indicates "1". For $n = 0$, q_0 represents the number of workers that submit λ for the ith bit. Since the workers independently complete the microtasks, recalling the results in (2.13), the probabilities of the crowd's answer profile for the ith bit $\{G_0, G_1, G_\lambda\}$ can be obtained under H_1 and H_0 given $p_{w,i}$ and $r_{w,i}$ are expressed by $F_i(\mathbb{Q})$ and $F_i'(\mathbb{Q})$, respectively. Thus, $P_{d,i}$ given $p_{w,i}$ and $r_{w,i}$ can be expressed as

$$P_{d,i} = \sum_{S} \binom{W}{\mathbb{Q}} F_i(\mathbb{Q}) + \frac{1}{2} \sum_{S'} \binom{W}{\mathbb{Q}} F_i(\mathbb{Q}), \tag{A.17}$$

where the first term on the right-hand side corresponds to the case where the aggregation rule gives a result of "1" and the second term indicates the case where "1" is given due to the tie-breaking of the aggregation rule.

Similarly, we can obtain P_f given $p_{w,i}$ and $r_{w,i}$ as

$$P_{f,i} = \sum_{S} \binom{W}{\mathbb{Q}} F_i'(\mathbb{Q}) + \frac{1}{2} \sum_{S'} \binom{W}{\mathbb{Q}} F_i'(\mathbb{Q}). \tag{A.18}$$

Then, it is straightforward to obtain the desired result.

A.4 Proof of Proposition 2.5

We can have a correct classification if and only if all the bits are classified correctly. Thus, the expected probability of correct classification is given as $P_c = E\left[\prod_{i=1}^{N} P_{c,i}\right]$, which can be expressed, due to the independence of the microtasks, as $P_c = \prod_{i=1}^{N} E\left[P_{c,i}\right]$. Recall $P_{c,i}$ from Prop. 2.4, and we

can obtain:

$$E\left[P_{c,i}\right] = \frac{1}{2} + \frac{1}{2}\sum_{S}\binom{W}{\mathbb{Q}}\left(F\left(\mathbb{Q}\right) - F'\left(\mathbb{Q}\right)\right) + \frac{1}{4}\sum_{S'}\binom{W}{\mathbb{Q}}\left(F\left(\mathbb{Q}\right) - F'\left(\mathbb{Q}\right)\right) \tag{A.19}$$

with $F\left(\mathbb{Q}\right)$ and $F'\left(\mathbb{Q}\right)$ defined in (2.19) and (2.20). Thus we have the desired result.

A.5 Proof of Proposition 2.6

When W goes to infinity, we show that $E\left[P_{c,i}\right] = Q\left(-\frac{M}{\sqrt{V}}\right)$ and the desired result can be obtained. Based on the Central Limit Theorem [25], the test statistic in (2.12) is approximately Gaussian if $W \to \infty$:

$$\sum_{w=1}^{W} T_w \sim \begin{cases} \mathcal{N}\left(M_1, V_1\right), & H_1 \\ \mathcal{N}\left(M_0, V_0\right), & H_0 \end{cases}, \tag{A.20}$$

where M_s and V_s are the means and variances of the statistic $\sum_{w=1}^{W} T_w$ under hypotheses H_s, respectively.

For the wth worker, we have the expectation of T_w as

$$M_{H_1} = \sum_{t=0}^{1}\sum_{n=1}^{N}(-1)^{t+1}\mu^{-n}(r_{w,i})^{t}(1-r_{w,i})^{1-t}\varphi_n(w). \tag{A.21}$$

We define M_1 as

$$\begin{aligned} M_1 &\triangleq W E\left[M_{H_1}\right] \\ &= W\sum_{n=1}^{N}\mu^{-n}\left(2\mu-1\right)\binom{N-1}{n-1}(1-m)^n m^{N-n} \\ &= \frac{W\left(2\mu-1\right)\left(1-m\right)}{\mu}\left(\frac{1}{\mu} - \left(\frac{1}{\mu}-1\right)m\right)^{N-1}. \end{aligned} \tag{A.22}$$

Likewise, we define V_1 as

$$\begin{aligned} V_1 &\triangleq W\left(E\left[T_w^2\right] - E^2\left[M_{H_1}\right]\right) \\ &= WE\left[\sum_{t=0}^{1}\sum_{n=1}^{N}\mu^{-2n}(r_{w,i})^t(1-r_{w,i})^{1-t}\varphi_n(w)\right] - \frac{M_1^2}{W} \\ &= \frac{W(1-m)}{\mu^2}\left(\frac{1}{\mu^2} - \left(\frac{1}{\mu^2}-1\right)m\right)^{N-1} - \frac{M_1^2}{W}. \end{aligned} \tag{A.23}$$

Similarly, we can derive $M \triangleq M_1 = -M_0$ and $V \triangleq V_1 = V_0$. By looking back at the decision criterion for the ith bit (2.12), we obtain the desired result.

A.6 Proof of Proposition 3.1

Since the workers complete the microtasks independently, recall the results in (2.13) and we have

$$E\left[P_{d,i}\right] = \sum_{S_1}\binom{W}{\mathbb{Q}_1}F(\mathbb{Q}_1) + \frac{1}{2}\sum_{S_1'}\binom{W}{\mathbb{Q}_1}F(\mathbb{Q}_1), \tag{A.24}$$

and

$$E\left[P_{f,i}\right] = \sum_{S_1}\binom{W}{\mathbb{Q}_1}F'(\mathbb{Q}_1) + \frac{1}{2}\sum_{S_1'}\binom{W}{\mathbb{Q}_1}F'(\mathbb{Q}_1), \tag{A.25}$$

with $F(\mathbb{Q}_1)$ and $F'(\mathbb{Q}_1)$ as given above. Then, it is straightforward to obtain the desired result.

A.7 Proof of Proposition 3.3

If a spammer skips g out of G gold standard questions and answers the remaining $G - g$ with random guesses, the expected monetary reward E for the spammer is expressed as

$$\begin{aligned} E &= (\mu_{\max} - \mu_{\min}) T^G \prod_{i=1}^{G} \alpha_{x_i} + \mu_{\min} \\ &= (\mu_{\max} - \mu_{\min}) T^G (\frac{1}{2})^{G-g} (\frac{1}{T})^{G-g} + \mu_{\min} \\ &= (\mu_{\max} - \mu_{\min}) (\frac{1}{2})^G (2T)^g + \mu_{\min}, \end{aligned} \tag{A.26}$$

where $p(x_1, \ldots, x_G)$ is the probability that a spammer gives evaluations $X = \{x_1, \ldots, x_G\}$ for the gold standard questions.

Since $0 \leq g \leq G$, E is maximized as following

$$\text{if } T < \frac{1}{2} \Rightarrow g = 0, \tag{A.27}$$

$$\text{if } T > \frac{1}{2} \Rightarrow g = G. \tag{A.28}$$

A.8 Proof of Proposition 3.4

When there are M spammers in the crowd with M_0 skipping and M_N completing all the questions, the expected weight contributed to the correct class is given by

$$\begin{aligned}
E_C[\mathbb{W}] = &\sum_{w=1}^{W-M} E_{p,\rho}\left[\sum_{n=0}^{N} W_w \rho(n) P_\lambda(n)\right] + \sum_{w=1}^{M_0} W_w(n=0) \\
&+ \sum_{w=1}^{M_N} \frac{1}{2^N} W_w(n=N) \\
=&\sum_{n=0}^{N}(W-M)W_w\mu^n\binom{N}{n}(1-m)^n m^{N-n} \\
&+ \sum_{n=0}^{N} M_0 W_w \delta(n) + \sum_{n=0}^{N} \frac{M_N}{2^N} W_w \delta(n-N) \\
=&\sum_{n=0}^{N}(W-M)W_w\mu^n\mathbb{P}(n) + \sum_{n=0}^{N} \frac{M_0}{\mathbb{P}(0)} W_w \mathbb{P}(n)\delta(n) \\
&+ \sum_{n=0}^{N} \frac{M_N}{2^N\mathbb{P}(N)} W_w \mathbb{P}(n)\delta(n-N) \\
=&\sum_{n=0}^{N} W_w S(n)\mathbb{P}(n)
\end{aligned} \tag{A.29}$$

where

$$\mathbb{P}(n) = \binom{N}{n}(1-m)^n m^{N-n}, \tag{A.30}$$

and

$$S(n) = (W-M)\mu^n + \frac{M_0}{m^N}\delta(n) + \frac{M_N}{2^N(1-m)^N}\delta(n-N). \tag{A.31}$$

Note that $\sum_{n=0}^{N} \mathbb{P}(n) = 1$, and then (A.29) is upper-bounded using Cauchy-Schwarz inequality

as follows:

$$E_C[\mathbb{W}] = \sum_{n=0}^{N} W_w S(n)\mathbb{P}(n)$$

$$\leq \sqrt{\sum_{n=0}^{N}(W_w S(n))^2\mathbb{P}(n)}\sqrt{\sum_{n=0}^{N}\mathbb{P}(n)}. \tag{A.32}$$

$$= \alpha \tag{A.33}$$

Also note that equality holds in (A.32) only if

$$W_w S(n)\sqrt{\mathbb{P}(n)} = \alpha\sqrt{\mathbb{P}(n)} \tag{A.34}$$

where α is a positive constant, and

$$W_w S(n) = \alpha \tag{A.35}$$

Therefore, the optimal behavior for the manager in terms of the weight assignment is obtained

$$W_w = \left[(W-M)\,\mu^n + \frac{M_0}{m^N}\delta\,(n) + \frac{M_N}{2^N(1-m)^N}\delta\,(n-N)\right]^{-1}. \tag{A.36}$$

Note that if a worker submits no definitive answers, i.e. $n = 0$, the corresponding weight assigned is $(W - M + \frac{M_0}{m^N})^{-1}$. However, since this worker skips all the questions, his/her decision for a certain question is not taken into consideration in the fusion center. Thus, we can ignore the weight assignment in such a case and write the scheme as

$$W_w = \left[(W-M)\,\mu^n + \frac{M_N}{2^N(1-m)^N}\delta\,(n-N)\right]^{-1}. \tag{A.37}$$

Lemma A.1. *The update of the the algorithm can be written as*

$$\mathbf{x}^{k+1} = -\frac{1}{2c}\mathbf{W}^{-1}\nabla f(\mathbf{x}^{k+1}) + \frac{\mathbf{W}^{-1}\mathbf{L}_+}{2}(\mathbf{x}^k + \mathbf{e}^k) - \frac{\mathbf{W}^{-1}\mathbf{L}_-}{2}\left(\sum_{s=0}^{k}\mathbf{x}^s + \mathbf{e}^s\right). \tag{A.38}$$

Proof. Using the second step of the algorithm, we can write

$$\alpha^{k+1} = \alpha^k + c\mathbf{L}_-(\mathbf{x}^{k+1} + \mathbf{e}^{k+1}) \tag{A.39}$$

and

$$\alpha^k = \alpha^{k-1} + c\mathbf{L}_-(\mathbf{x}^k + \mathbf{e}^k). \tag{A.40}$$

Sum and telescope from iteration 0 to k using (A.40), and we can get the following by assuming $\alpha^0 = 0$

$$\alpha^k = c\mathbf{L}_-\sum_{s=0}^{k}(\mathbf{x}^s + \mathbf{e}^s). \tag{A.41}$$

Substitute the above result to the first step in the algorithm and it yields

$$2c\mathbf{W}\mathbf{x}^{k+1} = -\nabla f(\mathbf{x}^{k+1}) + c\mathbf{L}_+(\mathbf{x}^k + \mathbf{e}^k) - c\mathbf{L}_-\sum_{s=0}^{k}(\mathbf{x}^s + \mathbf{e}^s), \tag{A.42}$$

which completes the proof. □

Lemma A.2. *The sequences satisfy*

$$\frac{\mathbf{L}_+}{2}(\mathbf{z}^{k+1} - \mathbf{z}^k) - \mathbf{W}\mathbf{e}^{k+1} = -\mathbf{Q}\mathbf{r}^{k+1} - \frac{1}{2c}\nabla f(\mathbf{x}^{k+1}) \tag{A.43}$$

Proof. Based on Lemma A.1 and the fact $\mathbf{W} = \frac{1}{2}(\mathbf{L}_- + \mathbf{L}_+)$, we can write

$$\mathbf{W}(\mathbf{x}^{k+1} - \mathbf{x}^k - \mathbf{e}^k) + \mathbf{W}(\mathbf{x}^k + \mathbf{e}^k) - \frac{\mathbf{L}_+}{2}(\mathbf{x}^k + \mathbf{e}^k) = -\mathbf{Q}\mathbf{r}^k - \frac{1}{2c}\nabla f(\mathbf{x}^{k+1}). \tag{A.44}$$

Subtracting $\frac{\mathbf{L}_-}{2}(\mathbf{x}^{k+1}+\mathbf{e}^{k+1})$ from both sides of the above equation provides

$$\mathbf{W}(\mathbf{x}^{k+1}-\mathbf{x}^k-\mathbf{e}^k)+\frac{\mathbf{L}_-}{2}(\mathbf{x}^k+\mathbf{e}^k)-\frac{\mathbf{L}_-}{2}(\mathbf{x}^{k+1}+\mathbf{e}^{k+1})=-\mathbf{Q}\mathbf{r}^{k+1}-\frac{1}{2c}\nabla f(\mathbf{x}^{k+1}). \quad \text{(A.45)}$$

Rearrange and we have the desired result. □

Lemma A.3. *The null space of* $\mathbf{Q}$ *null*$(\mathbf{Q})$ *is span*$\{\mathbf{1}\}$.

Proof. Note that the null space of $\mathbf{Q}$ and $\mathbf{L}_-$ are the same. By definition, $\mathbf{L}_- = \frac{1}{2}\mathbf{M}_-\mathbf{M}_-^T$ and $\mathbf{M}_- = \mathbf{A}_1^T - \mathbf{A}_2^T$. Recall that if $(i,j)\in\mathcal{A}$ and $\mathbf{y}_{ij}$ is the qth block of $\mathbf{y}$, then the (q,i)th block of $\mathbf{A}_1$ and the (q,j)th block of $\mathbf{A}_2$ are $N\times N$ identity matrices $\mathbf{I}_N$; otherwise the corresponding blocks are $N\times N$ zero matrices 0_N. Therefore, $\mathbf{M}_-^T = \mathbf{A}_1 - \mathbf{A}_2$ is a matrix that each row has one "1", one "-1", and all zeros otherwise, which means $\mathbf{M}_-^T\mathbf{1}=\mathbf{0}$, i.e., null$(\mathbf{M}_-^T)$=span$\{\mathbf{1}\}$.

Note that $\mathbf{L}_- = \frac{1}{2}\mathbf{M}_-\mathbf{M}_-^T$ and $\mathbf{Q} = \left(\frac{\mathbf{L}_-}{2}\right)^{\frac{1}{2}}$, thus null$(\mathbf{Q})$=null$(\mathbf{M}_-^T)$, completing the proof. □

Lemma A.4. *For some* $\mathbf{r}^*$ *that satisfies* $\mathbf{Q}\mathbf{r}^*+\frac{1}{2c}\nabla f(\mathbf{x}^*)=0$ *and* $\mathbf{r}^*$ *belongs to the column space of* $\mathbf{Q}$*, the sequences satisfy*

$$\frac{\mathbf{L}_+}{2}(\mathbf{z}^{k+1}-z^k)-\mathbf{W}\mathbf{e}^{k+1}=-\mathbf{Q}(\mathbf{r}^{k+1}-\mathbf{r}^*)-\frac{1}{2c}(\nabla f(\mathbf{x}^{k+1})-\nabla f(\mathbf{x}^*)) \quad \text{(A.46)}$$

Proof. Using Lemma A.2, we have

$$\frac{\mathbf{L}_+}{2}(\mathbf{z}^{k+1}-\mathbf{z}^k)-\mathbf{W}\mathbf{e}^{k+1}=-\mathbf{Q}\mathbf{r}^{k+1}-\frac{1}{2c}\nabla f(\mathbf{x}^{k+1}). \quad \text{(A.47)}$$

According to Lemma A.3, null$(\mathbf{Q})$ is span$\{\mathbf{1}\}$. Since $\mathbf{1}^T\nabla f(\mathbf{x}^*)=0$, $\nabla f(\mathbf{x}^*)$ can be written as a linear combination of column vectors of $\mathbf{Q}$. Therefore, there exists $\mathbf{r}$ such that $\frac{1}{2c}\nabla f(x^*)=-\mathbf{Q}\mathbf{r}$. Let $\mathbf{r}^*$ be the projection of $\mathbf{r}$ onto $\mathbf{Q}$ to obtain $\mathbf{Q}\mathbf{r}=\mathbf{Q}\mathbf{r}^*$ where $\mathbf{r}^*$ lies in the column space of $\mathbf{Q}$.

Hence, we can write

$$\frac{\mathbf{L}_+}{2}(\mathbf{z}^{k+1}-\mathbf{z}^k)-\mathbf{W}\mathbf{e}^{k+1}=-\mathbf{Q}(\mathbf{r}^{k+1}-\mathbf{r}^*)-\frac{1}{2c}(\nabla f(\mathbf{x}^{k+1})-\nabla f(\mathbf{x}^*)) \tag{A.48}$$

□

Lemma A.5. $\langle \mathbf{x}^*, \mathbf{Q}\rangle = 0$.

Proof. Since the optimal consensus solution $\mathbf{x}^*$ has an identical value for all its entries, $\mathbf{x}^*$ lies in the space spanned by $\mathbf{1}$. Thus, according to Lemma A.3, we have the desired result, and also $\langle \mathbf{x}^*, \mathbf{L}_-\rangle = 0$. □

A.9 Proof of Theorem 5.1

Proof. We prove the first part in Theorem 5.1. Assuming $f(\mathbf{x})$ is convex, we can have

$$f(\mathbf{x}^{k+1})-f(\mathbf{x}^*)\le\langle \mathbf{x}^{k+1}-\mathbf{x}^*, \nabla f(x^{k+1})\rangle. \tag{A.49}$$

By Lemma A.2, it yields

$$\begin{aligned}
f(\mathbf{x}^{k+1})-f(\mathbf{x}^*)\le&\langle \mathbf{x}^{k+1}-\mathbf{x}^*, 2c\mathbf{W}\mathbf{e}^{k+1}-2c\mathbf{Q}\mathbf{r}^{k+1}-c\mathbf{L}_+(\mathbf{z}^{k+1}-\mathbf{z}^k)\rangle && \text{(A.50)}\\
=&\langle \mathbf{x}^{k+1}-\mathbf{x}^*, c\mathbf{L}_+(\mathbf{z}^k-\mathbf{z}^{k+1})\rangle+\langle \mathbf{x}^{k+1}-\mathbf{x}^*, 2c\mathbf{W}\mathbf{e}^{k+1}\rangle && \text{(A.51)}\\
&+\langle \mathbf{x}^{k+1}-\mathbf{x}^*, -2c\mathbf{Q}\mathbf{r}^{k+1}\rangle && \text{(A.52)}\\
=&\langle \mathbf{z}^{k+1}-\mathbf{z}^*, c\mathbf{L}_+(\mathbf{z}^k-\mathbf{z}^{k+1})\rangle-\langle \mathbf{e}^{k+1}, c\mathbf{L}_+(\mathbf{z}^k-\mathbf{z}^{k+1})\rangle && \text{(A.53)}\\
&+\langle \mathbf{z}^{k+1}-\mathbf{z}^*, -2c\mathbf{Q}\mathbf{r}^{k+1}\rangle-\langle \mathbf{e}^{k+1}, -2c\mathbf{Q}\mathbf{r}^{k+1}\rangle+\langle \mathbf{e}^{k+1}, 2c\mathbf{W}(\mathbf{x}^{k+1}-\mathbf{x}^*)\rangle && \\
& && \text{(A.54)}\\
=&2\langle \mathbf{z}^{k+1}-\mathbf{z}^*, \frac{c\mathbf{L}_+}{2}(\mathbf{z}^k-\mathbf{z}^{k+1})\rangle+2\langle \mathbf{r}^k-\mathbf{r}^{k+1}, c(\mathbf{r}^{k+1}-\mathbf{r}')\rangle && \text{(A.55)}\\
&+\langle \mathbf{e}^{k+1}, c\mathbf{L}_+(\mathbf{z}^{k+1}-\mathbf{z}^k)+2c\mathbf{Q}\mathbf{r}^{k+1}+2c\mathbf{W}(\mathbf{x}^{k+1}-\mathbf{x}^*)\rangle. && \text{(A.56)}
\end{aligned}$$

If the algorithm stops at T-th iteration, then the function value $f(\mathbf{x}^T)$ is affected by the error $\mathbf{e}^k$ with $k = 0, 1, \ldots, T-1$. Thus, we can set $k = T-1$ and $\mathbf{e}^T = \mathbf{0}$ in the above bound, and obtain

$$f(\mathbf{x}^T) - f(\mathbf{x}^*) \leq \|\mathbf{z}^{T-1} - \mathbf{z}^*\|^2_{\frac{c\mathbf{L}_+}{2}} - \|\mathbf{z}^T - \mathbf{z}^*\|^2_{\frac{c\mathbf{L}_+}{2}} - \|\mathbf{z}^{T-1} - \mathbf{z}^T\|^2_{\frac{c\mathbf{L}_+}{2}} \tag{A.57}$$

$$+ c\|\mathbf{r}^{T-1} - \mathbf{r}'\|_2^2 - c\|\mathbf{r}^T - \mathbf{r}'\|_2^2 - c\|\mathbf{r}^{T-1} - \mathbf{r}^T\|_2^2 \tag{A.58}$$

$$\leq \|\mathbf{q}^{T-1} - \mathbf{p}\|^2_{\mathbf{G}}. \tag{A.59}$$

Now we prove the second part in Theorem 5.1. By convexity, for any $\mathbf{r} \in \mathbb{R}^{DN}$, we can have

$$\frac{f(\mathbf{x}^{k+1}) - f(\mathbf{x}^*)}{c} + 2\mathbf{r}'\mathbf{Q}\mathbf{x}^{k+1} \tag{A.60}$$

$$\leq \langle \mathbf{x}^{k+1} - \mathbf{x}^*, -\mathbf{L}_+(\mathbf{x}^{k+1} - \mathbf{x}^k) - \mathbf{L}_+(\mathbf{x}^k - \mathbf{z}^k) - 2\mathbf{Q}(\mathbf{r}^{k+1} - \mathbf{r}) + \mathbf{L}_-(\mathbf{z}^{k+1} - \mathbf{x}^{k+1}) \rangle \tag{A.61}$$

$$= \langle \mathbf{x}^{k+1} - \mathbf{x}^*, \mathbf{L}_+(\mathbf{x}^k - \mathbf{x}^{k+1}) \rangle + \langle \mathbf{x}^{k+1} - \mathbf{x}^*, \mathbf{L}_+(\mathbf{z}^k - \mathbf{x}^k) \rangle + \langle \mathbf{x}^{k+1} - \mathbf{x}^*, 2\mathbf{Q}(\mathbf{r} - \mathbf{r}^{k+1}) \rangle \tag{A.62}$$

$$+ \langle \mathbf{x}^{k+1} - \mathbf{x}^*, \mathbf{L}_-(\mathbf{z}^{k+1} - \mathbf{x}^{k+1}) \rangle \tag{A.63}$$

$$= \langle \mathbf{x}^{k+1} - \mathbf{x}^*, \mathbf{L}_+(\mathbf{x}^k - \mathbf{x}^{k+1}) \rangle + \langle \mathbf{x}^{k+1} - \mathbf{x}^*, \mathbf{L}_+(\mathbf{z}^k - \mathbf{x}^k) \rangle + \langle \mathbf{z}^{k+1} - \mathbf{x}^*, 2\mathbf{Q}(\mathbf{r} - \mathbf{r}^{k+1}) \rangle \tag{A.64}$$

$$+ \langle \mathbf{x}^{k+1} - \mathbf{x}^*, \mathbf{L}_-(\mathbf{z}^{k+1} - \mathbf{x}^{k+1}) \rangle + \langle \mathbf{e}^{k+1}, 2\mathbf{Q}(\mathbf{r}^{k+1} - \mathbf{r}) \rangle \tag{A.65}$$

$$= \langle \mathbf{x}^{k+1} - \mathbf{x}^*, \mathbf{L}_+(\mathbf{x}^k - \mathbf{x}^{k+1}) \rangle + \langle \mathbf{x}^{k+1} - \mathbf{x}^*, \mathbf{L}_+(\mathbf{z}^k - \mathbf{x}^k) \rangle + \langle \mathbf{r}^{k+1} - \mathbf{r}^k, 2(\mathbf{r} - \mathbf{r}^{k+1}) \rangle \tag{A.66}$$

$$+ \langle \mathbf{x}^{k+1} - \mathbf{x}^*, \mathbf{L}_-(\mathbf{z}^{k+1} - \mathbf{x}^{k+1}) \rangle + \langle \mathbf{e}^{k+1}, 2\mathbf{Q}(\mathbf{r}^{k+1} - \mathbf{r}) \rangle \tag{A.67}$$

$$= \frac{1}{c}(\|\mathbf{p}^k - \mathbf{p}\|_{\mathbf{G}}^2 - \|\mathbf{p}^{k+1} - \mathbf{p}\|_{\mathbf{G}}^2 - \|\mathbf{p}^{k+1} - \mathbf{p}^k\|_{\mathbf{G}}^2) + \langle \mathbf{x}^{k+1} - \mathbf{x}^*, \mathbf{L}_+(\mathbf{z}^k - \mathbf{x}^k) \rangle \tag{A.68}$$

$$+ \langle \mathbf{x}^{k+1} - \mathbf{x}^*, \mathbf{L}_-(\mathbf{z}^{k+1} - \mathbf{x}^{k+1}) \rangle + \langle \mathbf{e}^{k+1}, 2\mathbf{Q}(\mathbf{r}^{k+1} - \mathbf{r}) \rangle \tag{A.69}$$

$$= \frac{1}{c}(\|\mathbf{p}^k - \mathbf{p}\|_{\mathbf{G}}^2 - \|\mathbf{p}^{k+1} - \mathbf{p}\|_{\mathbf{G}}^2) - \|\mathbf{Q}\mathbf{x}^{k+1}\|_2^2 - \|\mathbf{Q}\mathbf{e}^{k+1}\|_2^2 + 2\langle \frac{\mathbf{L}_+}{2}(\mathbf{x}^{k+1} - \mathbf{x}^*), \mathbf{z}^k - \mathbf{x}^k \rangle \tag{A.70}$$

$$+ \langle \mathbf{e}^{k+1}, 2\mathbf{Q}(\mathbf{r}^{k+1} - \mathbf{r}) \rangle \tag{A.71}$$

$$= \frac{1}{c}(\|\mathbf{p}^k - \mathbf{p}\|_{\mathbf{G}}^2 - \|\mathbf{p}^{k+1} - \mathbf{p}\|_{\mathbf{G}}^2) - \frac{\sigma_{\min}(\mathbf{L}_-)}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_2^2 - \|\mathbf{Q}\mathbf{e}^{k+1}\|_2^2 \tag{A.72}$$

$$+ \frac{1}{\alpha}\|\frac{\mathbf{L}_+}{2}(\mathbf{x}^{k+1} - \mathbf{x}^*)\|_2^2 + \alpha\|\mathbf{z}^k - \mathbf{x}^k\|_2^2 + \langle \mathbf{e}^{k+1}, 2\mathbf{Q}(\mathbf{r}^{k+1} - \mathbf{r}) \rangle \tag{A.73}$$

$$\overset{\alpha = \frac{\sigma_{\max}^2(\mathbf{L}_+)}{2\sigma_{\min}(\mathbf{L}_-)}}{=} \frac{1}{c}(\|\mathbf{p}^k - \mathbf{p}\|_{\mathbf{G}}^2 - \|\mathbf{p}^{k+1} - \mathbf{p}\|_{\mathbf{G}}^2) - \|\mathbf{Q}\mathbf{e}^{k+1}\|_2^2 + \frac{\sigma_{\max}^2(\mathbf{L}_+)}{2\sigma_{\min}(\mathbf{L}_-)}\|\mathbf{z}^k - \mathbf{x}^k\|_2^2 \tag{A.74}$$

$$+ \langle \mathbf{e}^{k+1}, 2\mathbf{Q}(\mathbf{r}^{k+1} - \mathbf{r}) \rangle \tag{A.75}$$

$$\leq \frac{1}{c}(\|\mathbf{p}^k - \mathbf{p}\|_{\mathbf{G}}^2 - \|\mathbf{p}^{k+1} - \mathbf{p}\|_{\mathbf{G}}^2) + \frac{\sigma_{\max}^2(\mathbf{L}_+)}{2\sigma_{\min}(\mathbf{L}_-)}\|\mathbf{e}^k\|_2^2 + \langle \mathbf{e}^{k+1}, 2\mathbf{Q}(\mathbf{r}^{k+1} - \mathbf{r}) \rangle. \tag{A.76}$$

By letting $\mathbf{r} = 0$, telescope and sum from $k = 0$ to $T - 1$ (the error for the last iteration $\mathbf{e}^T = 0$),

and we obtain

$$\frac{1}{c}\sum_{k=1}^{T}\big(f(\mathbf{x}^k)-f(\mathbf{x}^*)\big)\leq\frac{1}{c}\|\mathbf{p}^0-\mathbf{p}\|_{\mathbf{G}}^2+\sum_{k=0}^{T-1}\left(\frac{\sigma_{\max}^2(\mathbf{L}_+)}{2\sigma_{\min}(\mathbf{L}_-)}\|\mathbf{e}^k\|_2^2+\langle\mathbf{e}^k,2\mathbf{Q}\mathbf{r}^k)\rangle\right). \tag{A.77}$$

Rearrange and we have the desired result. □

A.10 Proof of Theorem 5.2

Proof. By v-strong convexity, we obtain

$$\begin{aligned}
v\|\mathbf{x}^{k+1}-\mathbf{x}^*\|_2^2 \leq&\langle\mathbf{x}^{k+1}-\mathbf{x}^*,\nabla f(\mathbf{x}^{k+1})-\nabla f(\mathbf{x}^*)\rangle && \text{(A.78)}\\
=&\langle\mathbf{x}^{k+1}-\mathbf{x}^*,c\mathbf{L}_+(\mathbf{z}^k-\mathbf{z}^{k+1})+2c\mathbf{W}\mathbf{e}^{k+1}-2c\mathbf{Q}(\mathbf{r}^{k+1}-\mathbf{r}^*)\rangle && \text{(A.79)}\\
=&\langle\mathbf{x}^{k+1}-\mathbf{x}^*,c\mathbf{L}_+(\mathbf{z}^k-\mathbf{z}^{k+1})\rangle+\langle\mathbf{x}^{k+1}-\mathbf{x}^*,2c\mathbf{W}\mathbf{e}^{k+1}\rangle && \text{(A.80)}\\
&+\langle\mathbf{x}^{k+1}-\mathbf{x}^*,-2c\mathbf{Q}(\mathbf{r}^{k+1}-\mathbf{r}^*)\rangle && \text{(A.81)}\\
=&\langle\mathbf{z}^{k+1}-\mathbf{z}^*,c\mathbf{L}_+(\mathbf{z}^k-\mathbf{z}^{k+1})\rangle-\langle\mathbf{e}^{k+1},c\mathbf{L}_+(\mathbf{z}^k-\mathbf{z}^{k+1})\rangle && \text{(A.82)}\\
&+\langle\mathbf{x}^{k+1}+\mathbf{e}^{k+1}-\mathbf{x}^*,-2c\mathbf{Q}(\mathbf{r}^{k+1}-\mathbf{r}^*)\rangle && \text{(A.83)}\\
&-\langle\mathbf{e}^{k+1},-2c\mathbf{Q}(\mathbf{r}^{k+1}-\mathbf{r}^*)\rangle+\langle\mathbf{e}^{k+1},2c\mathbf{W}(\mathbf{x}^{k+1}-\mathbf{x}^*)\rangle && \text{(A.84)}\\
=&2\langle\mathbf{z}^{k+1}-\mathbf{z}^*,\frac{c\mathbf{L}_+}{2}(\mathbf{z}^k-\mathbf{z}^{k+1})\rangle+2\langle\mathbf{r}^k-\mathbf{r}^{k+1},c(\mathbf{r}^{k+1}-\mathbf{r}^*)\rangle && \text{(A.85)}\\
&+\langle\mathbf{e}^{k+1},c\mathbf{L}_+(\mathbf{z}^{k+1}-\mathbf{z}^k)+2c\mathbf{Q}(\mathbf{r}^{k+1}-\mathbf{r}^*)+2c\mathbf{W}(\mathbf{x}^{k+1}-\mathbf{x}^*)\rangle && \text{(A.86)}\\
=&\|\mathbf{q}^k-\mathbf{q}^*\|_{\mathbf{G}}^2-\|\mathbf{q}^{k+1}-\mathbf{q}^*\|_{\mathbf{G}}^2-\|\mathbf{q}^k-\mathbf{q}^{k+1}\|_{\mathbf{G}}^2 && \text{(A.87)}\\
&+\langle\mathbf{e}^{k+1},c\mathbf{L}_+(\mathbf{z}^{k+1}-\mathbf{z}^k)+2c\mathbf{Q}(\mathbf{r}^{k+1}-\mathbf{r}^*)+2c\mathbf{W}(\mathbf{x}^{k+1}-\mathbf{x}^*)\rangle && \text{(A.88)}
\end{aligned}$$

For any $\lambda>0$, using the basic inequality

$$\|\mathbf{a}+\mathbf{b}\|_2^2+(\lambda-1)\|\mathbf{a}\|_2^2\geq(1-\frac{1}{\lambda})\|\mathbf{b}\|_2^2 \tag{A.89}$$

we can write for $\lambda_1 > 1$ and $\lambda_2 > 1$

$$\frac{\sigma^2_{\max}(\mathbf{L}_+)}{4}\|\mathbf{z}^{k+1} - \mathbf{z}^k\|_2^2 + \frac{(\lambda_1 - 1)L^2\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_2^2}{4c^2} \tag{A.90}$$

$$\geq \|\frac{\mathbf{L}_+}{2}(\mathbf{z}^{k+1} - \mathbf{z}^k)\|_2^2 + (\lambda_1 - 1)\|\frac{1}{2c}\left(\nabla f(\mathbf{x}^{k+1}) - \nabla f(\mathbf{x}^*)\right)\|_2^2 \tag{A.91}$$

$$\geq \left(1 - \frac{1}{\lambda_1}\right)\|\mathbf{W}\mathbf{e}^{k+1} - \mathbf{Q}(\mathbf{r}^{k+1} - \mathbf{r}^*)\|_2^2 \tag{A.92}$$

$$\geq \left(1 - \frac{1}{\lambda_1}\right)(1 - \frac{1}{\lambda_2})\|\mathbf{Q}(\mathbf{r}^{k+1} - \mathbf{r}^*)\|_2^2 - \left(1 - \frac{1}{\lambda_1}\right)(\lambda_2 - 1)\|\mathbf{W}\mathbf{e}^{k+1}\|_2^2 \tag{A.93}$$

$$\geq \left(1 - \frac{1}{\lambda_1}\right)(1 - \frac{1}{\lambda_2})\sigma^2_{\min}(\mathbf{Q})\|\mathbf{r}^{k+1} - \mathbf{r}^*\|_2^2 - \left(1 - \frac{1}{\lambda_1}\right)(\lambda_2 - 1)\sigma^2_{\max}(\mathbf{W})\|\mathbf{e}^{k+1}\|_2^2. \tag{A.94}$$

Thus, for a positive quantity δ,

$$\frac{\delta\sigma^2_{\max}(\mathbf{L}_+)\lambda_1\lambda_2}{4\sigma^2_{\min}(\mathbf{Q})(\lambda_1 - 1)(\lambda_2 - 1)}\|\mathbf{z}^{k+1} - \mathbf{z}^k\|_2^2 + \frac{\delta\lambda_1\lambda_2 L^2\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_2^2}{4c^2\sigma^2_{\min}(\mathbf{Q})(\lambda_2 - 1)} \tag{A.95}$$

$$\geq \delta\|\mathbf{r}^{k+1} - \mathbf{r}^*\|_2^2 - \frac{\delta\lambda_2\sigma^2_{\max}(\mathbf{W})}{\sigma^2_{\min}(\mathbf{Q})}\|\mathbf{e}^{k+1}\|_2^2. \tag{A.96}$$

Since $\mathbf{x}^{k+1} - \mathbf{x}^* = \mathbf{z}^{k+1} - \mathbf{z}^* - \mathbf{e}^{k+1}$, for any $\lambda_3 > 1$, we can get

$$\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_2^2 \geq \left(1 - \frac{1}{\lambda_3}\right)\|\mathbf{z}^{k+1} - \mathbf{z}^*\|_2^2 - (\lambda_3 - 1)\|\mathbf{e}^{k+1}\|_2^2. \tag{A.97}$$

Therefore, the addition of (A.95)$\times c^2$ and (A.97)$\times \frac{\delta c^2 \sigma_{\max}^2(L_+)\lambda_3}{4(\lambda_3-1)}$ yields

$$\frac{c^2\delta\sigma_{\max}^2(\mathbf{L}_+)\lambda_1\lambda_2}{4\sigma_{\min}^2(\mathbf{Q})(\lambda_1-1)(\lambda_2-1)}\|\mathbf{z}^{k+1}-\mathbf{z}^k\|_2^2+\left(\frac{\delta\lambda_1\lambda_2L^2}{4\sigma_{\min}^2(\mathbf{Q})(\lambda_2-1)}+\frac{\delta c^2\sigma_{\max}^2(\mathbf{L}_+)\lambda_3}{4(\lambda_3-1)}\right)\|\mathbf{x}^{k+1}-\mathbf{x}^*\|_2^2 \tag{A.98}$$

$$\geq \delta c^2\|\mathbf{r}^{k+1}-\mathbf{r}^*\|_2^2+\frac{\delta c^2\sigma_{\max}^2(\mathbf{L}_+)}{4}\|\mathbf{z}^{k+1}-\mathbf{z}^k\|_2^2-\left(\frac{c^2\delta\lambda_2\sigma_{\max}^2(\mathbf{W})}{\sigma_{\min}^2(\mathbf{Q})}+\frac{\delta c^2\sigma_{\max}^2(\mathbf{L}_+)\lambda_3}{4}\right)\|\mathbf{e}^{k+1}\|_2^2 \tag{A.99}$$

$$\geq \delta\|c(\mathbf{r}^{k+1}-\mathbf{r}^*)\|_2^2+\delta\|\frac{c\mathbf{L}_+}{2}(\mathbf{z}^{k+1}-\mathbf{z}^k)\|_2^2-\left(\frac{c^2\delta\lambda_2\sigma_{\max}^2(\mathbf{W})}{\sigma_{\min}^2(\mathbf{Q})}+\frac{\delta c^2\sigma_{\max}^2(\mathbf{L}_+)\lambda_3}{4}\right)\|\mathbf{e}^{k+1}\|_2^2 \tag{A.100}$$

$$=\delta\|\mathbf{q}^{k+1}-\mathbf{q}^*\|_{\mathbf{G}}^2-\left(\frac{c^2\delta\lambda_2\sigma_{\max}^2(\mathbf{W})}{\sigma_{\min}^2(\mathbf{Q})}+\frac{\delta c^2\sigma_{\max}^2(\mathbf{L}_+)\lambda_3}{4}\right)\|\mathbf{e}^{k+1}\|_2^2. \tag{A.101}$$

Choose δ to be such that

$$\frac{c^2\delta\sigma_{\max}^2(\mathbf{L}_+)\lambda_1\lambda_2}{4\sigma_{\min}^2(\mathbf{Q})(\lambda_1-1)(\lambda_2-1)}\leq\frac{c^2\sigma_{\min}^2(\mathbf{L}_+)}{4} \tag{A.102}$$

$$\left(\frac{\delta\lambda_1\lambda_2L^2}{4\sigma_{\min}^2(\mathbf{Q})(\lambda_2-1)}+\frac{\delta c^2\sigma_{\max}^2(\mathbf{L}_+)\lambda_3}{4(\lambda_3-1)}\right)\leq v, \tag{A.103}$$

and we can have

$$\frac{c^2\sigma_{\min}^2(\mathbf{L}_+)}{4}\|\mathbf{z}^{k+1}-\mathbf{z}^k\|_2^2+v\|\mathbf{x}^{k+1}-\mathbf{x}^k\|_2^2\geq \tag{A.104}$$

$$\delta\|\mathbf{q}^{k+1}-\mathbf{q}^*\|_{\mathbf{G}}^2-\left(\frac{c^2\delta\lambda_2\sigma_{\max}^2(\mathbf{W})}{\sigma_{\min}^2(\mathbf{Q})}+\frac{\delta c^2\sigma_{\max}^2(\mathbf{L}_+)\lambda_3}{4}\right)\|\mathbf{e}^{k+1}\|_2^2. \tag{A.105}$$

Thus, it is straightforward to write

$$\|\mathbf{q}^{k+1}-\mathbf{q}^{k}\|_{\mathbf{G}}^{2}+v\|\mathbf{x}^{k+1}-\mathbf{x}^{k}\|_{2}^{2} \tag{A.106}$$

$$\geq \|c(\mathbf{r}^{k+1}-\mathbf{r}^{k})\|_{2}^{2}+\|\frac{c\mathbf{L}_{+}}{2}(\mathbf{z}^{k+1}-\mathbf{z}^{k})\|_{2}^{2}+v\|\mathbf{x}^{k+1}-\mathbf{x}^{k}\|_{2}^{2} \tag{A.107}$$

$$\geq c^{2}\|\mathbf{r}^{k+1}-\mathbf{r}^{k}\|_{2}^{2}+\frac{c^{2}\sigma_{\min}^{2}(\mathbf{L}_{+})}{4}\|\mathbf{z}^{k+1}-\mathbf{z}^{k}\|_{2}^{2}+v\|\mathbf{x}^{k+1}-\mathbf{x}^{k}\|_{2}^{2} \tag{A.108}$$

$$\geq c^{2}\|\mathbf{r}^{k+1}-\mathbf{r}^{k}\|_{2}^{2}+\delta\|\mathbf{q}^{k+1}-\mathbf{q}^{*}\|_{\mathbf{G}}^{2}-\left(\frac{c^{2}\delta\lambda_{2}\sigma_{\max}^{2}(\mathbf{W})}{\sigma_{\min}^{2}(\mathbf{Q})}+\frac{\delta c^{2}\sigma_{\max}^{2}(\mathbf{L}_{+})\lambda_{3}}{4}\right)\|\mathbf{e}^{k+1}\|_{2}^{2}. \tag{A.109}$$

Recall the result in (A.78) regarding the bound to $v\|\mathbf{x}^{k+1}-\mathbf{x}^{k}\|_{2}^{2}$, and we can further write

$$\|\mathbf{q}^{k}-\mathbf{q}^{*}\|_{\mathbf{G}}^{2}-\|\mathbf{q}^{k+1}-\mathbf{q}^{*}\|_{\mathbf{G}}^{2} \tag{A.110}$$

$$+\langle\mathbf{e}^{k+1},c\mathbf{L}_{+}(\mathbf{z}^{k+1}-\mathbf{z}^{k})+2c\mathbf{Q}(\mathbf{r}^{k+1}-\mathbf{r}^{*})+2c\mathbf{W}(\mathbf{x}^{k+1}-\mathbf{x}^{*})\rangle \tag{A.111}$$

$$\geq \delta\|\mathbf{q}^{k+1}-\mathbf{q}^{*}\|_{\mathbf{G}}^{2}-\left(\frac{c^{2}\delta\lambda_{2}\sigma_{\max}^{2}(\mathbf{W})}{\sigma_{\min}^{2}(\mathbf{Q})}+\frac{\delta c^{2}\sigma_{\max}^{2}(\mathbf{L}_{+})\lambda_{3}}{4}\right)\|\mathbf{e}^{k+1}\|_{2}^{2} \tag{A.112}$$

Let $P=\frac{c^{2}\delta\lambda_{2}\sigma_{\max}^{2}(\mathbf{W})}{\sigma_{\min}^{2}(\mathbf{Q})}+\frac{\delta c^{2}\sigma_{\max}^{2}(\mathbf{L}_{+})\lambda_{3}}{4}$. Rearrange the expression and we get

$$\|\mathbf{q}^{k+1}-\mathbf{q}^{*}\|_{\mathbf{G}}^{2}\leq\frac{\|\mathbf{q}^{k}-\mathbf{q}^{*}\|_{\mathbf{G}}^{2}}{1+\delta}+\frac{P}{1+\delta}\|\mathbf{e}^{k+1}\|_{2}^{2} \tag{A.113}$$

$$+\frac{1}{1+\delta}\langle\mathbf{e}^{k+1},c\mathbf{L}_{+}(\mathbf{z}^{k+1}-\mathbf{z}^{k})+2c\mathbf{Q}(\mathbf{r}^{k+1}-\mathbf{r}^{*})+2c\mathbf{W}(\mathbf{x}^{k+1}-\mathbf{x}^{*})\rangle \tag{A.114}$$

□

Lemma A.6. *Let $\beta \in (0, \frac{1+\delta}{4})$, $b \in (0,1)$, $\lambda_4 > 1$, and then we have*

$$(1-b)\left(\frac{1}{4}-\frac{\beta}{1+\delta}\right)\sigma_{\min}^2(\mathbf{L}_+)\|\mathbf{z}^{k+1}-\mathbf{z}^*\|_2^2+\left(1-\frac{4\beta}{1+\delta}\right)\|\mathbf{r}^{k+1}-\mathbf{r}^*\|_2^2 \tag{A.115}$$

$$+b\left(\frac{1}{4}-\frac{\beta}{1+\delta}\right)\sigma_{\min}^2(\mathbf{L}_+)\left(1-\frac{1}{\lambda_4}\right)\|\mathbf{x}^{k+1}-\mathbf{x}^*\|_2^2 \tag{A.116}$$

$$\leq \frac{1/4+\beta}{1+\delta}\sigma_{\max}^2(\mathbf{L}_+)\|\mathbf{z}^k-\mathbf{z}^*\|_2^2+\frac{1}{1+\delta}\|\mathbf{r}^k-\mathbf{r}^*\|_2^2 \tag{A.117}$$

$$+\left[\frac{P+1/2\beta}{(1+\delta)c^2}+b\left(\frac{1}{4}-\frac{\beta}{1+\delta}\right)\sigma_{\min}^2(\mathbf{L}_+)(\lambda_4-1)\right]\|\mathbf{e}^{k+1}\|_2^2 \tag{A.118}$$

$$+\frac{4\beta\sigma_{\max}^2(\mathbf{W})}{1+\delta}\|\mathbf{x}^{k+1}-\mathbf{x}^*\|_2^2. \tag{A.119}$$

Proof. First, we rewrite the result in Lemma 5.2 in the following form

$$\|\frac{c\mathbf{L}_+}{2}(\mathbf{z}^{k+1}-\mathbf{z}^*)\|_2^2+\|c(\mathbf{r}^{k+1}-\mathbf{r}^*)\|_2^2 \leq \frac{1}{1+\delta}\left(\|\frac{c\mathbf{L}_+}{2}(\mathbf{z}^k-\mathbf{z}^*)\|_2^2+\|c(\mathbf{r}^k-\mathbf{r}^*)\|_2^2\right) \tag{A.120}$$

$$+\frac{P}{1+\delta}\|\mathbf{e}^{k+1}\|_2^2+\frac{1}{1+\sigma}\langle\mathbf{e}^{k+1},c\mathbf{L}_+(\mathbf{z}^{k+1}-\mathbf{z}^k)+2c\mathbf{Q}(\mathbf{r}^{k+1}-\mathbf{r}^*)+2c\mathbf{W}(\mathbf{x}^{k+1}-\mathbf{x}^*)\rangle \tag{A.121}$$

$$\leq \frac{1}{1+\delta}\left(\|\frac{c\mathbf{L}_+}{2}(\mathbf{z}^k-\mathbf{z}^*)\|_2^2+\|c(\mathbf{r}^k-\mathbf{r}^*)\|_2^2\right)+\left(\frac{P}{1+\delta}+\frac{1/2\beta}{1+\delta}\right)\|\mathbf{e}^{k+1}\|_2^2 \tag{A.122}$$

$$+\frac{\beta}{1+\delta}\|\frac{c\mathbf{L}_+}{2}(\mathbf{z}^{k+1}-\mathbf{z}^k)+c\mathbf{Q}(\mathbf{r}^{k+1}-\mathbf{r}^*)+c\mathbf{W}(\mathbf{x}^{k+1}-\mathbf{x}^*)\|_2^2 \tag{A.123}$$

$$\leq \frac{1}{1+\delta}\left(\|\frac{c\mathbf{L}_+}{2}(\mathbf{z}^k-\mathbf{z}^*)\|_2^2+\|c(\mathbf{r}^k-\mathbf{r}^*)\|_2^2\right)+\left(\frac{P}{1+\delta}+\frac{1/2\beta}{1+\delta}\right)\|\mathbf{e}^{k+1}\|_2^2 \tag{A.124}$$

$$+\frac{\beta}{1+\delta}\|\frac{c\mathbf{L}_+}{2}(\mathbf{z}^{k+1}-\mathbf{z}^k)+c\mathbf{Q}(\mathbf{r}^{k+1}-\mathbf{r}^*)+c\mathbf{W}(\mathbf{x}^{k+1}-\mathbf{x}^*)\|_2^2 \tag{A.125}$$

$$\leq \frac{1}{1+\delta}\left(\|\frac{c\mathbf{L}_+}{2}(\mathbf{z}^k-\mathbf{z}^*)\|_2^2+\|c(\mathbf{r}^k-\mathbf{r}^*)\|_2^2\right)+\left(\frac{P}{1+\delta}+\frac{1/2\beta}{1+\delta}\right)\|\mathbf{e}^{k+1}\|_2^2 \tag{A.126}$$

$$+\frac{4\beta}{1+\delta}\|\frac{c\mathbf{L}_+}{2}(\mathbf{z}^{k+1}-\mathbf{z}^*)\|_2^2+\frac{4\alpha}{1+\delta}\|\frac{c\mathbf{L}_+}{2}(\mathbf{z}^k-\mathbf{z}^*)\|_2^2 \tag{A.127}$$

$$+\frac{4\beta}{1+\delta}\|c\mathbf{Q}(\mathbf{r}^{k+1}-\mathbf{r}^*)\|_2^2+\frac{4\beta}{1+\delta}\|c\mathbf{W}(\mathbf{x}^{k+1}-\mathbf{x}^*)\|_2^2 \tag{A.128}$$

where $\beta > 0$.

Rearranging the inequality provides

$$\left(1-\frac{4\beta}{1+\delta}\right)\|\frac{c\mathbf{L}_+}{2}(\mathbf{z}^{k+1}-\mathbf{z}^*)\|_2^2+\left(1-\frac{4\beta}{1+\delta}\right)\|c(\mathbf{r}^{k+1}-\mathbf{r}^*)\|_2^2 \tag{A.129}$$

$$\leq\left(\frac{1}{1+\delta}+\frac{4\beta}{1+\delta}\right)\|\frac{c\mathbf{L}_+}{2}(\mathbf{z}^{k}-\mathbf{z}^*)\|_2^2+\frac{1}{1+\delta}\|c(\mathbf{r}^{k}-\mathbf{r}^*)\|_2^2 \tag{A.130}$$

$$+\frac{P+1/2\beta}{1+\delta}\|\mathbf{e}^{k+1}\|_2^2+\frac{4\beta}{1+\delta}\|c\mathbf{W}(\mathbf{x}^{k+1}-\mathbf{x}^*)\|_2^2. \tag{A.131}$$

Note that the parameters should be chosen such that $\left(1-\frac{4\beta}{1+\delta}\right)>0$.

Then we can write

$$\left(\frac{1}{4}-\frac{\beta}{1+\delta}\right)\sigma_{\min}^2(\mathbf{L}_+)\|\mathbf{z}^{k+1}-\mathbf{z}^*\|_2^2+\left(1-\frac{4\beta}{1+\delta}\right)\|\mathbf{r}^{k+1}-\mathbf{r}^*\|_2^2 \tag{A.132}$$

$$\leq\left(\frac{1}{4(1+\delta)}+\frac{\beta}{1+\delta}\right)\sigma_{\max}^2(\mathbf{L}_+)\|\mathbf{z}^{k}-\mathbf{z}^*\|_2^2+\frac{1}{1+\delta}\|\mathbf{r}^{k}-\mathbf{r}^*\|_2^2 \tag{A.133}$$

$$+\frac{P+1/2\beta}{(1+\delta)c^2}\|\mathbf{e}^{k+1}\|_2^2+\frac{4\beta\sigma_{\max}^2(\mathbf{W})}{1+\delta}\|\mathbf{x}^{k+1}-\mathbf{x}^*\|_2^2. \tag{A.134}$$

Since we have the inequality $\|\mathbf{z}^{k+1}-\mathbf{z}^*\|_2^2\geq\left(1-\frac{1}{\lambda_4}\right)\|\mathbf{x}^{k+1}-\mathbf{x}^*\|_2^2-(\lambda_4-1)\|\mathbf{e}^{k+1}\|_2^2$, for $b\in(0,1)$, we can get

$$b\left(\frac{1}{4}-\frac{\beta}{1+\delta}\right)\sigma_{\min}^2(\mathbf{L}_+)\|\mathbf{z}^{k+1}-\mathbf{z}^*\|_2^2\geq b\left(\frac{1}{4}-\frac{\beta}{1+\delta}\right)\sigma_{\min}^2(\mathbf{L}_+)\left(1-\frac{1}{\lambda_4}\right)\|\mathbf{x}^{k+1}-\mathbf{x}^*\|_2^2 \tag{A.135}$$

$$-b\left(\frac{1}{4}-\frac{\beta}{1+\delta}\right)\sigma_{\min}^2(\mathbf{L}_+)(\lambda_4-1)\|\mathbf{e}^{k+1}\|_2^2. \tag{A.136}$$

Thus,

$$(1-b)\left(\frac{1}{4}-\frac{\beta}{1+\delta}\right)\sigma_{\min}^2(\mathbf{L}_+)\|\mathbf{z}^{k+1}-\mathbf{z}^*\|_2^2+\left(1-\frac{4\beta}{1+\delta}\right)\|\mathbf{r}^{k+1}-\mathbf{r}^*\|_2^2 \tag{A.137}$$

$$+b\left(\frac{1}{4}-\frac{\beta}{1+\delta}\right)\sigma_{\min}^2(\mathbf{L}_+)\left(1-\frac{1}{\lambda_4}\right)\|\mathbf{x}^{k+1}-\mathbf{x}^*\|_2^2 \tag{A.138}$$

$$\leq\frac{1/4+\beta}{1+\delta}\sigma_{\max}^2(\mathbf{L}_+)\|\mathbf{z}^k-\mathbf{z}^*\|_2^2+\frac{1}{1+\delta}\|\mathbf{r}^k-\mathbf{r}^*\|_2^2 \tag{A.139}$$

$$+\left[\frac{P+1/2\beta}{(1+\delta)c^2}+b\left(\frac{1}{4}-\frac{\beta}{1+\delta}\right)\sigma_{\min}^2(\mathbf{L}_+)(\lambda_4-1)\right]\|\mathbf{e}^{k+1}\|_2^2 \tag{A.140}$$

$$+\frac{4\beta\sigma_{\max}^2(\mathbf{W})}{1+\delta}\|\mathbf{x}^{k+1}-\mathbf{x}^*\|_2^2. \tag{A.141}$$

□

Defining

$$A_1=\frac{4}{(1-b)\sigma_{\min}^2(\mathbf{L}_+)}, \tag{A.142}$$

and

$$A_2=\frac{4}{(1+4\beta)\sigma_{\max}^2(\mathbf{L}_+)}, \tag{A.143}$$

we have the desired result.

A.11 Proof of Theorem 5.3

A.11.1 Eliminate $\|\mathbf{x}^{k+1}-\mathbf{x}^*\|_2^2$

First, we want to eliminate the term $\|\mathbf{x}^{k+1}-\mathbf{x}^*\|_2^2$ in Lemma A.6, which requires

$$b\left(\frac{1}{4}-\frac{\beta}{1+\delta}\right)\sigma_{\min}^2(\mathbf{L}_+)\left(1-\frac{1}{\lambda_4}\right)\geq\frac{4\beta\sigma_{\max}^2(\mathbf{W})}{1+\delta} \tag{A.144}$$

and it is equivalent to that

$$\beta \leq \frac{b(1+\delta)\sigma_{\min}^2(\mathbf{L}_+)\left(1-\frac{1}{\lambda_4}\right)}{4b\sigma_{\min}^2(\mathbf{L}_+)\left(1-\frac{1}{\lambda_4}\right)+16\sigma_{\max}^2(\mathbf{W})} \tag{A.145}$$

Then we can write

$$(1-b)\left(\frac{1}{4}-\frac{\beta}{1+\delta}\right)\sigma_{\min}^2(\mathbf{L}_+)\|\mathbf{z}^{k+1}-\mathbf{z}^*\|_2^2+\left(1-\frac{4\beta}{1+\delta}\right)\|\mathbf{r}^{k+1}-\mathbf{r}^*\|_2^2 \tag{A.146}$$

$$\leq \frac{1/4+\beta}{1+\delta}\sigma_{\max}^2(\mathbf{L}_+)\|\mathbf{z}^k-\mathbf{z}^*\|_2^2+\frac{1}{1+\delta}\|\mathbf{r}^k-\mathbf{r}^*\|_2^2 \tag{A.147}$$

$$+\left[\frac{P+1/2\beta}{(1+\delta)c^2}+b\left(\frac{1}{4}-\frac{\beta}{1+\delta}\right)\sigma_{\min}^2(\mathbf{L}_+)(\lambda_4-1)\right]\|\mathbf{e}^{k+1}\|_2^2 \tag{A.148}$$

$$\tag{A.149}$$

which can be further simplified

$$\|\mathbf{z}^{k+1}-\mathbf{z}^*\|_2^2+A_1\|\mathbf{r}^{k+1}-\mathbf{r}^*\|_2^2 \leq B(\|\mathbf{z}^k-\mathbf{z}^*\|_2^2+A_2\|\mathbf{r}^k-\mathbf{r}^*\|_2^2)+C\|\mathbf{e}^{k+1}\|_2^2 \tag{A.150}$$

We require the following for convergence analysis

$$A_1 \geq A_2 \tag{A.151}$$

which leads to the requirement

$$(1-b)\sigma_{\min}^2(\mathbf{L}_+) \leq (1+\beta)\sigma_{\max}^2(\mathbf{L}_+). \tag{A.152}$$

Note that this requirement is satisfied intrinsically.

Therefore, we get

$$\|\mathbf{z}^{k+1}-\mathbf{z}^*\|_2^2+A_1\|\mathbf{r}^{k+1}-\mathbf{r}^*\|_2^2\leq B^{k+1}\left(\|\mathbf{z}^0-\mathbf{z}^*\|_2^2+A_2\|\mathbf{r}^0-\mathbf{r}^*\|_2^2+\sum_{s=1}^{k+1}B^{-s}C\|\mathbf{e}^s\|_2^2\right) \tag{A.153}$$

and we have the desired result since $A_1\|r^{k+1}-r^*\|_2^2>0$.

A.11.2 $B\in(0,1)$

The above convergence result requires that $B\in(0,1)$. First, having β in Theorem 5.3 at hand, we can make sure that B is greater than 0. Then, it requires that $B<1$ and correspondingly

$$(1+4\beta)\sigma_{\max}^2(\mathbf{L}_+)\leq(1-b)(1+\delta-4\beta)\sigma_{\min}^2(\mathbf{L}_+) \tag{A.154}$$

which is equivalent to that

$$\beta\leq\frac{(1-b)(1+\delta)\sigma_{\min}^2(\mathbf{L}_+)-\sigma_{\max}^2(\mathbf{L}_+)}{4\sigma_{\max}^2(\mathbf{L}_+)+4(1-b)\sigma_{\min}^2(\mathbf{L}_+)} \tag{A.155}$$

and

$$(1-b)(1+\delta)\sigma_{\min}^2(\mathbf{L}_+)-\sigma_{\max}^2(\mathbf{L}_+)>0. \tag{A.156}$$

Since b can be arbitrarily chosen from $(0,1)$, we also need

$$0<\frac{\sigma_{\max}^2(\mathbf{L}_+)}{(1+\delta)\sigma_{\min}^2(\mathbf{L}_+)}<1 \tag{A.157}$$

One intuition is that we should design a network such that $\frac{\sigma_{\max}^2(\mathbf{L}_+)}{\sigma_{\min}^2(\mathbf{L}_+)}$ is the smallest possible. Sub-

stituting δ in the expression and we have

$$\frac{\sigma^2_{\min}(\mathbf{L}_+)}{\sigma^2_{\max}(\mathbf{L}_+)} > \frac{L^2 - 2v + \sqrt{(L^2+2v)^2 + 16v^2\frac{\lambda_2-1}{\lambda_2}\sigma^2_{\min}(\mathbf{Q})}}{4v\frac{\lambda_2-1}{\lambda_2}\sigma^2_{\min}(\mathbf{Q}) + 2L^2}. \tag{A.158}$$

A.12 Proof of Theorem 5.4

Note that δ is chosen as

$$\delta = \min\left\{\frac{(\lambda_1-1)(\lambda_2-1)\sigma^2_{\min}(\mathbf{Q})\sigma^2_{\min}(\mathbf{L}_+)}{\lambda_1\lambda_2\sigma^2_{\max}(\mathbf{L}_+)}, \frac{4v(\lambda_2-1)(\lambda_3-1)\sigma^2_{\min}(\mathbf{Q})}{\lambda_1\lambda_2(\lambda_3-1)L^2 + c^2\lambda_3(\lambda_2-1)\sigma^2_{\max}(\mathbf{L}_+)\sigma^2_{\min}(\mathbf{Q})}\right\} \tag{A.159}$$

We choose c such that

$$\lambda_1\lambda_2(\lambda_3-1)L^2 = c^2\lambda_3(\lambda_2-1)\sigma^2_{\max}(\mathbf{L}_+)\sigma^2_{\min}(\mathbf{Q}), \tag{A.160}$$

which yields

$$c = \sqrt{\frac{\lambda_1\lambda_2(\lambda_3-1)L^2}{\lambda_3(\lambda_2-1)\sigma^2_{\max}(\mathbf{L}_+)\sigma^2_{\min}(\mathbf{Q})}} \tag{A.161}$$

and

$$\delta = \min\left\{\frac{(\lambda_1-1)(\lambda_2-1)\sigma^2_{\min}(\mathbf{Q})\sigma^2_{\min}(\mathbf{L}_+)}{\lambda_1\lambda_2\sigma^2_{\max}(\mathbf{L}_+)}, \frac{2v(\lambda_2-1)\sigma^2_{\min}(\mathbf{Q})}{\lambda_1\lambda_2L^2}\right\} \tag{A.162}$$

$$= \frac{(\lambda_2-1)\sigma^2_{\min}(\mathbf{Q})}{\lambda_2}\min\left\{\frac{(\lambda_1-1)\sigma^2_{\min}(\mathbf{L}_+)}{\lambda_1\sigma^2_{\max}(\mathbf{L}_+)}, \frac{2v}{\lambda_1L^2}\right\} \tag{A.163}$$

It is desirable that δ can achieve its maximum, which is obtained by

$$\frac{(\lambda_1-1)\sigma^2_{\min}(\mathbf{L}_+)}{\lambda_1\sigma^2_{\max}(\mathbf{L}_+)} = \frac{2v}{\lambda_1L^2}. \tag{A.164}$$

Therefore, we can set λ_1 as

$$\lambda_1 = 1 + \frac{2v\sigma_{\max}^2(\mathbf{L}_+)}{L^2\sigma_{\min}^2(\mathbf{L}_+)}, \tag{A.165}$$

and thus, we have δ as

$$\delta = \frac{(\lambda_2 - 1)}{\lambda_2} \frac{2v\sigma_{\min}^2(\mathbf{Q})\sigma_{\min}^2(\mathbf{L}_+)}{L^2\sigma_{\min}^2(\mathbf{L}_+) + 2v\sigma_{\max}^2(\mathbf{L}_+)} \tag{A.166}$$

The constraint on β in Theorem 4 ensures that $B > 0$.

Note that λ_3 only appears in C and P. It is straightforward to derive the optimal λ_3 to minimize C, and we arrive at

$$\lambda_3 = \sqrt{\frac{L^2\sigma_{\min}^2(\mathbf{L}_+) + 2v\sigma_{\max}^2(\mathbf{L}_+)}{\beta\lambda_1 L^2 v\sigma_{\min}^2(\mathbf{L}_+)}} + 1 \tag{A.167}$$

thus resulting in

$$C = \frac{\frac{4\delta\lambda_2\sigma_{\max}^2(\mathbf{W})}{\sigma_{\min}^2(\mathbf{Q})} + \sigma_{\max}^2(\mathbf{L}_+)\left(\sqrt{\delta} + \sqrt{\frac{2(\lambda_2-1)\sigma_{\min}^2(\mathbf{Q})}{\beta\lambda_1\lambda_2 L^2}}\right)^2}{(1-b)(1+\delta)(1+\delta-4\beta)\sigma_{\min}^2(\mathbf{L}_+)} + \frac{b(\lambda_4 - 1)}{1-b} \tag{A.168}$$

A.13 Proof of Corollary 5.1

A.13.1 First one:

According to the result in Theorem 5.3, we have

$$\|\mathbf{z}^{k+1} - \mathbf{z}^*\|_2^2 \leq B^{k+1}(\|\mathbf{z}^0 - \mathbf{z}^*\|_2^2 + A_1\|\mathbf{r}^0 - \mathbf{r}^*\|_2^2) + B^{k+1}\sum_{s=1}^{k+1} B^{-s}C\|\mathbf{e}^s\|_2^2 \tag{A.169}$$

and then

$$\|\mathbf{z}^{k+1} - \mathbf{z}^*\|_2^2 \leq B^{k+1}(\|\mathbf{z}^0 - \mathbf{z}^*\|_2^2 + A_1\|\mathbf{r}^0 - \mathbf{r}^*\|_2^2) + CeB^{k+1}\sum_{s=1}^{k+1} B^{-s} \tag{A.170}$$

$$= B^{k+1}(\|\mathbf{z}^0 - \mathbf{z}^*\|_2^2 + A_1\|\mathbf{r}^0 - \mathbf{r}^*\|_2^2) + Ce\frac{1 - B^{k+1}}{1 - B} \tag{A.171}$$

$$\leq B^{k+1}(\|\mathbf{z}^0 - \mathbf{z}^*\|_2^2 + A_1\|\mathbf{r}^0 - \mathbf{r}^*\|_2^2) + \frac{Ce}{1 - B}. \tag{A.172}$$

Since $B \in (0, 1)$, we have the desired result.

A.13.2 Second one:

Recall the result in (A.150),

$$\|\mathbf{z}^{k+1} - \mathbf{z}^*\|_2^2 \leq B^{k+1}(\|\mathbf{z}^0 - \mathbf{z}^*\|_2^2 + A_1\|\mathbf{r}^0 - \mathbf{r}^*\|_2^2) + B^{k+1}\sum_{s=1}^{k+1} B^{-s}C\|\mathbf{e}^s\|_2^2 \tag{A.173}$$

which then can be written as

$$\|\mathbf{z}^{k+1} - \mathbf{z}^*\|_2^2 \leq B^{k+1}(\|\mathbf{z}^0 - \mathbf{z}^*\|_2^2 + A_1\|\mathbf{r}^0 - \mathbf{r}^*\|_2^2) + B^{k+1}C\sum_{s=1}^{k+1} B^{-s}R^s\|\mathbf{e}^0\|_2^2 \tag{A.174}$$

$$\leq B^{k+1}(\|z^0 - \mathbf{z}^*\|_2^2 + A_1\|\mathbf{r}^0 - \mathbf{r}^*\|_2^2) + B^{k+1}C\|\mathbf{e}^0\|_2^2\sum_{s=1}^{k+1}\left(\frac{R}{B}\right)^s \tag{A.175}$$

$$\leq B^{k+1}(\|\mathbf{z}^0 - \mathbf{z}^*\|_2^2 + A_1\|\mathbf{r}^0 - \mathbf{r}^*\|_2^2) + B^{k+1}C\|\mathbf{e}^0\|_2^2\frac{R}{B - R} \tag{A.176}$$

$$= B^{k+1}(\|\mathbf{z}^0 - \mathbf{z}^*\|_2^2 + A_1\|\mathbf{r}^0 - \mathbf{r}^*\|_2^2 + \frac{RC\|\mathbf{e}^0\|_2^2}{B - R}) \tag{A.177}$$

completing the proof.

A.13.3 Third one:

Recall the result in (A.150),

$$\|\mathbf{z}^{k+1}-\mathbf{z}^*\|_2^2+A_1\|\mathbf{r}^{k+1}-\mathbf{r}^*\|_2^2\leq B(\|\mathbf{z}^k-\mathbf{z}^*\|_2^2+A_2\|\mathbf{r}^k-\mathbf{r}^*\|_2^2)+C\|\mathbf{e}^{k+1}\|_2^2. \tag{A.178}$$

If $C\|\mathbf{e}^{k+1}\|_2^2\leq B(A_1-A_2)\|\mathbf{r}^k-\mathbf{r}^*\|_2^2$, we can write

$$\|\mathbf{z}^{k+1}-\mathbf{z}^*\|_2^2+A_1\|\mathbf{r}^{k+1}-\mathbf{r}^*\|_2^2\leq B(\|\mathbf{z}^k-\mathbf{z}^*\|_2^2+A_2\|\mathbf{r}^k-\mathbf{r}^*\|_2^2)+C\|\mathbf{e}^{k+1}\|_2^2 \tag{A.179}$$

$$\leq B(\|\mathbf{z}^k-\mathbf{z}^*\|_2^2+A_2\|\mathbf{r}^k-\mathbf{r}^*\|_2^2)+B(A_1-A_2)\|\mathbf{r}^k-\mathbf{r}^*\|_2^2 \tag{A.180}$$

$$\leq B(\|\mathbf{z}^k-\mathbf{z}^*\|_2^2+A_1\|\mathbf{r}^k-\mathbf{r}^*\|_2^2). \tag{A.181}$$

Then we have

$$\|\mathbf{z}^{k+1}-\mathbf{z}^*\|_2^2+A_1\|\mathbf{r}^{k+1}-\mathbf{r}^*\|_2^2\leq B^{k+1}(\|\mathbf{z}^0-\mathbf{z}^*\|_2^2+A_1\|\mathbf{r}^0-\mathbf{r}^*\|_2^2), \tag{A.182}$$

which leads to

$$\|\mathbf{z}^{k+1}-\mathbf{z}^*\|_2^2\leq B^{k+1}(\|\mathbf{z}^0-\mathbf{z}^*\|_2^2+A_1\|\mathbf{r}^0-\mathbf{r}^*\|_2^2), \tag{A.183}$$

completing the proof as $B\in(0,1)$.

A.14 Useful Lemmas

Lemma A.7. *There exists a vector* $\mathbf{y}\in\mathbb{R}^N$ *and* $\sigma_{\min}(\mathbf{y}\mathbf{y}^T)=1$*, such that* $\forall\mathbf{x}\in\mathbb{R}^N$*,* $\mathbf{y}^T\mathbf{x}\geq\|\mathbf{x}\|$*.*

Proof. Since $\forall\mathbf{x}\in\mathbb{R}^D$, $\mathbf{y}^T\mathbf{x}\geq\|\mathbf{x}\|$, it leads to

$$\mathbf{x}^T\mathbf{y}\mathbf{y}^T\mathbf{x}\geq\mathbf{x}^T\mathbf{x}, \tag{A.184}$$

which is equivalent to

$$\sigma_{\min}(\mathbf{y}\mathbf{y}^T) = 1. \tag{A.185}$$

□

A.15 Proof of Lemma 5.1

Proof. First, for any $\mathbf{r} \in \mathbb{R}^{DN}$, we obtain

$$\frac{f(\mathbf{x}^{k+1}) - f(\mathbf{x}^*)}{c} + \langle 2\mathbf{Q}\mathbf{r}, \mathbf{x}^{k+1}\rangle \tag{A.186}$$

$$=\langle \mathbf{x}^{k+1} - \mathbf{x}^*, -\mathbf{L}_+(\mathbf{x}^{k+1} - \mathbf{x}^k) - 2\mathbf{Q}(\mathbf{r}^{k+1} - \mathbf{r})\rangle \tag{A.187}$$

$$=\langle \mathbf{x}^{k+1} - \mathbf{x}^*, -\mathbf{L}_+(\mathbf{x}^{k+1} - \mathbf{x}^k) - 2\mathbf{Q}(\mathbf{r}^{k+1} - \mathbf{r})\rangle \tag{A.188}$$

$$=\langle \mathbf{x}^{k+1} - \mathbf{x}^*, -\mathbf{L}_+(\mathbf{x}^{k+1} - \mathbf{x}^k)\rangle + \langle \mathbf{r}^{k+1} - \mathbf{r}^k, -2(\mathbf{r}^{k+1} - \mathbf{r})\rangle. \tag{A.189}$$

Telescope and sum from $k = 0, \ldots, T$, we can get

$$\frac{1}{c}\sum_{k=1}^{T} f(\mathbf{x}^k) - f(\mathbf{x}^*) + 2\mathbf{r}'\mathbf{Q}\mathbf{x}^k \tag{A.190}$$

$$\leq \|\mathbf{x}^0 - \mathbf{x}^*\|^2_{\frac{\mathbf{L}_+}{2}} - \|\mathbf{x}^T - \mathbf{x}^*\|^2_{\frac{\mathbf{L}_+}{2}} - \sum_{k=1}^{T} \|\mathbf{x}^k - \mathbf{x}^{k-1}\|^2_{\frac{\mathbf{L}_+}{2}} \tag{A.191}$$

$$+ \|\mathbf{r}^0 - \mathbf{r}\|_2^2 - \|\mathbf{r}^T - \mathbf{r}\|_2^2 - \sum_{k=1}^{T} \|\mathbf{r}^k - \mathbf{r}^{k-1}\|_2^2 \tag{A.192}$$

Therefore, we obtain

$$\frac{1}{c}\sum_{k=0}^{T} f(\mathbf{x}^k) - f(\mathbf{x}^*) + 2\mathbf{r}'\mathbf{Q}\mathbf{x}^k \leq \|\mathbf{x}^0 - \mathbf{x}^*\|^2_{\frac{\mathbf{L}_+}{2}} + \|\mathbf{r}^0 - \mathbf{r}\|_2^2 \tag{A.193}$$

Define $\hat{\mathbf{x}_T} = \frac{\sum_{k=1}^{T} \mathbf{x}^k}{T}$ and we get the following by Jensen's inequality as

$$f(\hat{\mathbf{x}_T}) - f(\mathbf{x}^*) + 2c\mathbf{r}'\mathbf{Q}\hat{\mathbf{x}_T} \leq \frac{c}{T}\|\mathbf{p}^0 - \mathbf{p}\|_{\mathbf{G}}^2. \tag{A.194}$$

If we choose $\mathbf{r} = 0$, we obtain

$$f(\hat{\mathbf{x}_T}) - f(\mathbf{x}^*) \leq \frac{c}{T}\left(\|\mathbf{x}^0 - \mathbf{x}^*\|_{\frac{\mathbf{L}_+}{2}}^2 + \|\mathbf{r}^0\|_2^2\right). \tag{A.195}$$

The saddle point inequality implies

$$f(\mathbf{x}^*) - f(\hat{\mathbf{x}_T}) \leq 2c\langle \mathbf{Q}\mathbf{r}^*, \hat{\mathbf{x}_T}\rangle. \tag{A.196}$$

Thus, using (A.193), it yields

$$2c\langle \mathbf{Q}\mathbf{r}^*, \hat{\mathbf{x}_T}\rangle \leq f(\hat{\mathbf{x}_T}) - f(\mathbf{x}^*) + 2c\langle \mathbf{Q}2\mathbf{r}^*, \hat{\mathbf{x}_T}\rangle \leq \frac{c}{T}\left(\|\mathbf{x}^0 - \mathbf{x}^*\|_{\frac{\mathbf{L}_+}{2}}^2 + \|\mathbf{r}^0 - 2\mathbf{r}^*\|_2^2\right) \tag{A.197}$$

Now we let $\mathbf{r} = \mathbf{r}^* + \mathbf{y}$ with $\mathbf{y}$ chosen according to Lemma A.7. Thus, we obtain

$$f(\hat{\mathbf{x}_T}) - f(\mathbf{x}^*) + 2c\langle \mathbf{Q}\mathbf{r}^*, \hat{\mathbf{x}_T}\rangle + 2c\mathbf{y}^T\mathbf{Q}\hat{\mathbf{x}_T} \leq \frac{c}{T}\left(\|\mathbf{x}^0 - \mathbf{x}^*\|_{\frac{\mathbf{L}_+}{2}}^2 + \|\mathbf{r}^0 - \mathbf{r}^* - \mathbf{y}\|_2^2\right). \tag{A.198}$$

Since $(\mathbf{x}^*, \mathbf{r}^*)$ is a primal-dual optimal solution, the saddle point inequality provides

$$f(\hat{\mathbf{x}_T}) - f(\mathbf{x}^*) + 2c\langle \mathbf{Q}\mathbf{r}^*, \hat{\mathbf{x}_T}\rangle \geq 0. \tag{A.199}$$

Using Lemma A.7, we obtain

$$\frac{2c}{T}\sum_{k=1}^{T}\|\mathbf{Q}\mathbf{x}^k\| \leq \frac{c}{T}\left(\|\mathbf{x}^0 - \mathbf{x}^*\|_{\frac{\mathbf{L}_+}{2}}^2 + \|\mathbf{r}^0 - \mathbf{r}^* - \mathbf{y}\|_2^2\right), \tag{A.200}$$

which yields

$$\frac{1}{T}\sum_{k=1}^{T}\|\mathbf{Q}\mathbf{x}^k\| \leq \frac{1}{2T}\left(\|\mathbf{x}^0-\mathbf{x}^*\|^2_{\frac{\mathbf{L}_+}{2}}+2\|\mathbf{r}^0-\mathbf{r}^*\|_2^2+2\right). \tag{A.201}$$

Choose the starting point $\mathbf{x}^0=0$ and thus $\mathbf{r}^0=0$, and we have

$$\frac{1}{T}\sum_{k=1}^{T}\|\mathbf{Q}\mathbf{x}^k\| \leq \frac{1}{2T}\left(\|\mathbf{x}^*\|^2_{\frac{\mathbf{L}_+}{2}}+2\|\mathbf{r}^*\|_2^2+2\right) \leq \frac{1}{4T}\left(\sigma_{\max}(\mathbf{L}_+)V_1^2+\frac{2V_2^2}{\sigma_{\min}(\mathbf{L}_-)c^2}+4\right). \tag{A.202}$$

□

A.16 Proof of Theorem 5.5

For any $\mathbf{r} \in \mathbb{R}^{DN}$, we can write

$$\frac{f(\mathbf{x}^{k+1}) - f(\mathbf{x}^*)}{c} + 2\mathbf{r}'\mathbf{Q}\mathbf{x}^{k+1} \tag{A.203}$$

$$\leq \langle \mathbf{x}^{k+1} - \mathbf{x}^*, -\mathbf{L}_+(\mathbf{x}^{k+1} - \mathbf{x}^k) - \mathbf{L}_+(\mathbf{x}^k - \mathbf{z}^k) - 2\mathbf{Q}(\mathbf{r}^{k+1} - \mathbf{r}) + \mathbf{L}_-(\mathbf{z}^{k+1} - \mathbf{x}^{k+1}) \rangle \tag{A.204}$$

$$= \langle \mathbf{x}^{k+1} - \mathbf{x}^*, \mathbf{L}_+(\mathbf{x}^k - \mathbf{x}^{k+1}) \rangle + \langle \mathbf{x}^{k+1} - \mathbf{x}^*, \mathbf{L}_+(\mathbf{z}^k - \mathbf{x}^k) \rangle + \langle \mathbf{x}^{k+1} - \mathbf{x}^*, 2\mathbf{Q}(\mathbf{r} - \mathbf{r}^{k+1}) \rangle \tag{A.205}$$

$$+ \langle \mathbf{x}^{k+1} - \mathbf{x}^*, \mathbf{L}_-(\mathbf{z}^{k+1} - \mathbf{x}^{k+1}) \rangle \tag{A.206}$$

$$= \langle \mathbf{x}^{k+1} - \mathbf{x}^*, \mathbf{L}_+(\mathbf{x}^k - \mathbf{x}^{k+1}) \rangle + \langle \mathbf{x}^{k+1} - \mathbf{x}^*, \mathbf{L}_+(\mathbf{z}^k - \mathbf{x}^k) \rangle + \langle \mathbf{z}^{k+1} - \mathbf{x}^*, 2\mathbf{Q}(\mathbf{r} - \mathbf{r}^{k+1}) \rangle \tag{A.207}$$

$$+ \langle \mathbf{x}^{k+1} - \mathbf{x}^*, \mathbf{L}_-(\mathbf{z}^{k+1} - \mathbf{x}^{k+1}) \rangle + \langle \mathbf{e}^{k+1}, 2\mathbf{Q}(\mathbf{r}^{k+1} - \mathbf{r}) \rangle \tag{A.208}$$

$$= \langle \mathbf{x}^{k+1} - \mathbf{x}^*, \mathbf{L}_+(\mathbf{x}^k - \mathbf{x}^{k+1}) \rangle + \langle \mathbf{x}^{k+1} - \mathbf{x}^*, \mathbf{L}_+(\mathbf{z}^k - \mathbf{x}^k) \rangle + \langle \mathbf{r}^{k+1} - \mathbf{r}^k, 2(\mathbf{r} - \mathbf{r}^{k+1}) \rangle \tag{A.209}$$

$$+ \langle \mathbf{x}^{k+1} - \mathbf{x}^*, \mathbf{L}_-(\mathbf{z}^{k+1} - \mathbf{x}^{k+1}) \rangle + \langle \mathbf{e}^{k+1}, 2\mathbf{Q}(\mathbf{r}^{k+1} - \mathbf{r}) \rangle \tag{A.210}$$

$$= \frac{1}{c}(\|\mathbf{p}^k - \mathbf{p}\|_{\mathbf{G}}^2 - \|\mathbf{p}^{k+1} - \mathbf{p}\|_{\mathbf{G}}^2 - \|\mathbf{p}^{k+1} - \mathbf{p}^k\|_{\mathbf{G}}^2) + \langle \mathbf{x}^{k+1} - \mathbf{x}^*, \mathbf{L}_+(\mathbf{z}^k - \mathbf{x}^k) \rangle \tag{A.211}$$

$$+ \langle \mathbf{x}^{k+1} - \mathbf{x}^*, \mathbf{L}_-(\mathbf{z}^{k+1} - \mathbf{x}^{k+1}) \rangle + \langle \mathbf{e}^{k+1}, 2\mathbf{Q}(\mathbf{r}^{k+1} - \mathbf{r}) \rangle \tag{A.212}$$

$$= \frac{1}{c}(\|\mathbf{p}^k - \mathbf{p}\|_{\mathbf{G}}^2 - \|\mathbf{p}^{k+1} - \mathbf{p}\|_{\mathbf{G}}^2) - \|\mathbf{Q}\mathbf{x}^{k+1}\|_2^2 - \|\mathbf{Q}\mathbf{e}^{k+1}\|_2^2 + 2\langle \frac{\mathbf{L}_+}{2}(\mathbf{x}^{k+1} - \mathbf{x}^*), \mathbf{z}^k - \mathbf{x}^k \rangle \tag{A.213}$$

$$+ \langle \mathbf{e}^{k+1}, 2\mathbf{Q}(\mathbf{r}^{k+1} - \mathbf{r}) \rangle \tag{A.214}$$

$$= \frac{1}{c}(\|\mathbf{p}^k - \mathbf{p}\|_{\mathbf{G}}^2 - \|\mathbf{p}^{k+1} - \mathbf{p}\|_{\mathbf{G}}^2) - \frac{\sigma_{\min}(\mathbf{L}_-)}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_2^2 - \|\mathbf{Q}\mathbf{e}^{k+1}\|_2^2 \tag{A.215}$$

$$+ \frac{1}{\alpha}\|\frac{\mathbf{L}_+}{2}(\mathbf{x}^{k+1} - \mathbf{x}^*)\|_2^2 + \alpha\|\mathbf{z}^k - \mathbf{x}^k\|_2^2 + \langle \mathbf{e}^{k+1}, 2\mathbf{Q}(\mathbf{r}^{k+1} - \mathbf{r}) \rangle \tag{A.216}$$

$$\overset{\alpha = \frac{\sigma_{\max}^2(\mathbf{L}_+)}{2\sigma_{\min}(\mathbf{L}_-)}}{=} \frac{1}{c}(\|\mathbf{p}^k - \mathbf{p}\|_{\mathbf{G}}^2 - \|\mathbf{p}^{k+1} - \mathbf{p}\|_{\mathbf{G}}^2) - \|\mathbf{Q}\mathbf{e}^{k+1}\|_2^2 + \frac{\sigma_{\max}^2(\mathbf{L}_+)}{2\sigma_{\min}(\mathbf{L}_-)}\|\mathbf{z}^k - \mathbf{x}^k\|_2^2 \tag{A.217}$$

$$+ \langle \mathbf{e}^{k+1}, 2\mathbf{Q}(\mathbf{r}^{k+1} - \mathbf{r}) \rangle \tag{A.218}$$

$$= \frac{1}{c}(\|\mathbf{p}^k - \mathbf{p}\|_{\mathbf{G}}^2 - \|\mathbf{p}^{k+1} - \mathbf{p}\|_{\mathbf{G}}^2) - \|\mathbf{Q}\mathbf{e}^{k+1}\|_2^2 + \frac{\sigma_{\max}^2(\mathbf{L}_+)}{2\sigma_{\min}(\mathbf{L}_-)}\|\mathbf{z}^k - \mathbf{x}^k\|_2^2 \tag{A.219}$$

$$+ \|2\mathbf{Q}\mathbf{e}^{k+1}\|\|\mathbf{r}^{k+1} - \mathbf{r}\| \tag{A.220}$$

Algorithm ROAD guarantees that $\sum_{t=1}^{k} \|\mathbf{Q}\mathbf{z}^t\| \leq 2EU/\sqrt{2} = \sqrt{2}EU$, and $\sum_{t=1}^{k} \|\mathbf{Q}\mathbf{X}^t\| \leq \sqrt{2}EU$ due to the thresholding as well. Thus, we have $\sum_{t=1}^{k} \|\mathbf{Q}\mathbf{e}^t\| \leq 2\sqrt{2}EU$. Then, we can have

$$\frac{f(\mathbf{x}^{k+1}) - f(\mathbf{x}^*)}{c} + 2\mathbf{r}'\mathbf{Q}\mathbf{x}^{k+1} \leq \frac{1}{c}(\|\mathbf{p}^k - \mathbf{p}\|_{\mathbf{G}}^2 - \|\mathbf{p}^{k+1} - \mathbf{p}\|_{\mathbf{G}}^2) - \|\mathbf{Q}\mathbf{e}^{k+1}\|_2^2 \tag{A.221}$$

$$+ \frac{\sigma_{\max}^2(\mathbf{L}_+)}{\sigma_{\min}^2(\mathbf{L}_-)}\|\mathbf{Q}\mathbf{e}^k\|_2^2 + \|2\mathbf{Q}\mathbf{e}^{k+1}\|(\sqrt{2}EU + \|\mathbf{r}\|). \tag{A.222}$$

Telescope and sum from $k = 0$ to $T - 1$ ($\mathbf{e}^T = 0$ since it is the last iteration), and we get

$$\sum_{k=1}^{T} f(\mathbf{x}^k) - f(\mathbf{x}^*) + 2c\mathbf{r}'\mathbf{Q}\mathbf{x}^k \leq \|\mathbf{p}^0 - \mathbf{p}\|_{\mathbf{G}}^2 - \|\mathbf{p}^T - \mathbf{p}\|_{\mathbf{G}}^2 + 2c\sum_{k=1}^{T} \|\mathbf{Q}\mathbf{e}^k\|(\sqrt{2}EU + \|\mathbf{r}\|) \tag{A.223}$$

$$+ c\frac{\sigma_{\max}^2(\mathbf{L}_+) - \sigma_{\min}^2(\mathbf{L}_-)}{\sigma_{\min}^2(\mathbf{L}_-)} \sum_{k=1}^{T} \|\mathbf{Q}\mathbf{e}^k\|_2^2 \tag{A.224}$$

$$\leq \|\mathbf{p}^0 - \mathbf{p}\|_{\mathbf{G}}^2 - \|\mathbf{p}^T - \mathbf{p}\|_{\mathbf{G}}^2 + c4\sqrt{2}EU(\sqrt{2}EU + \|\mathbf{r}\|) \tag{A.225}$$

$$+ c\frac{\sigma_{\max}^2(\mathbf{L}_+) - \sigma_{\min}^2(\mathbf{L}_-)}{\sigma_{\min}^2(\mathbf{L}_-)}8E^2U^2 \tag{A.226}$$

$$= \|\mathbf{p}^0 - \mathbf{p}\|_{\mathbf{G}}^2 - \|\mathbf{p}^T - \mathbf{p}\|_{\mathbf{G}}^2 + c4\sqrt{2}EU\|\mathbf{r}\| \tag{A.227}$$

$$+ c\frac{\sigma_{\max}^2(\mathbf{L}_+)}{\sigma_{\min}^2(\mathbf{L}_-)}8E^2U^2. \tag{A.228}$$

Choosing $\mathbf{r} = 0$, we obtain

$$f(\hat{\mathbf{x}}_T) - f(\mathbf{x}^*) \leq \frac{1}{T}\left(\|\mathbf{p}^0 - \mathbf{p}\|_{\mathbf{G}}^2 + 8c\frac{\sigma_{\max}^2(\mathbf{L}_+)}{\sigma_{\min}^2(\mathbf{L}_-)}E^2U^2\right). \tag{A.229}$$

A.17 Proof of Theorem 6.1

The proof uses the following inequality.

Lemma A.8. *[McDiarmid's Inequality [74]] Let $f : \mathcal{X}^m \to \mathbb{R}$ be a function such that for all $i \in \{1, \ldots, m\}$, there exist $c_i \leq \infty$ for which*

$$\sup_{X \in \mathcal{X}^m, \tilde{x} \in \mathcal{X}} |g(x_1, \ldots, x_{i-1}, \tilde{x}, x_{i+1}, \ldots, x_m)| \leq c_i, \tag{A.230}$$

where $g(x_1, \ldots, x_{i-1}, \tilde{x}, x_{i+1}, \ldots, x_m) = f(x_1, \ldots, x_m) - f(x_1, \ldots, x_{i-1}, \tilde{x}, x_{i+1}, \ldots, x_m)$. Then for all probability measure p and every $\epsilon > 0$,

$$P_X \left(f(X) - \mathbb{E}_X[f(X)] > \epsilon \right) < \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^m c_i^2} \right), \tag{A.231}$$

where X denotes $(x_1, \ldots, x_m)$, $\mathbb{E}_X[\cdot]$ denotes the expectation over the m random variables $x_i \sim p$, and P_X denotes the probability over these m variables.

To apply the McDiarmid's inequality, we first define the following quantity

$$\triangle_m(\mathbf{x}^\alpha) = \text{MMD}^2(\mathbf{x}_{m,i_m}, \mathbf{y}) - \text{MMD}^2(\mathbf{x}_{m',i_{m'}}, \mathbf{y}) \tag{A.232}$$

where $\mathbf{x}^\alpha := \{\mathbf{x}_{m,i_m}, \mathbf{x}_{m',i_{m'}}, \mathbf{y}\}$ consists of $3n$ data samples.
Given H_i, it can be shown that

$$\mathbb{E}[\text{MMD}^2(\mathbf{x}_{m,i_m}, \mathbf{y})] \leq D_I \tag{A.233}$$

$$\mathbb{E}[\text{MMD}^2(\mathbf{x}_{m',i_{m'}}, \mathbf{y})] \geq D_O. \tag{A.234}$$

We next define $\mathbf{x}^\alpha_{-s}$ the same as $\mathbf{x}^\alpha$ except that the s-th component $\mathbf{x}^\alpha_s$ is removed. We also define $\tilde{\mathbf{x}}^\alpha_s$ as another sequence generated by the same underlying distribution for $\mathbf{x}^\alpha_s$. Then, $\mathbf{x}^\alpha_s$ affects $\triangle_m(\mathbf{x}^\alpha)$ via the following three cases.

- Case 1: $\mathbf{x}^\alpha_s$ is in the sequence $\mathbf{x}_{m,i_m}$. In this case, $\mathbf{x}^\alpha_s$ affects $\triangle_m(\mathbf{x}^\alpha)$ through the following

terms

$$\frac{2}{n(n-1)}\sum_{l=1,l\neq s}^{n} k(\mathbf{x}_s^\alpha, \mathbf{x}_{m,i_m}(l)) - \frac{2}{n^2}\sum_{l=1}^{n} k(\mathbf{x}_s^\alpha, \mathbf{y}(l)).$$

- Case 2: $\mathbf{x}_s^\alpha$ is in the sequence $\mathbf{x}_{m',i_{m'}}$. In this case, $\mathbf{x}_s^\alpha$ affects $\triangle_m(\mathbf{x}^\alpha)$ through the following terms

$$\frac{2}{n(n-1)}\sum_{l=1,l\neq s}^{n} k(\mathbf{x}_s^\alpha, \mathbf{y}(l)) - \frac{2}{n^2}\sum_{j=1}^{n} k(\mathbf{x}_s^\alpha, \mathbf{x}_{m',i_{m'}}(l)).$$

- Case 3: $\mathbf{x}_s^\alpha$ is in the sequence $\mathbf{y}$. In this case, $\mathbf{x}_s^\alpha$ affects $\triangle_m(\mathbf{x}^\alpha)$ through the following terms

$$\frac{2}{n^2}\sum_{l=1}^{n} k(\mathbf{x}_s^\alpha, \mathbf{x}_{m,i_m}(l)) - \frac{2}{n^2}\sum_{l=1}^{n} k(\mathbf{x}_s^\alpha, \mathbf{x}_{m',i_{m'}}(l)).$$

Thus, since the kernel is bounded, i.e., $0 \leq k(x;y) \leq \mathcal{K}$ for any (x,y), considering the above three cases, the variation in the value of $\triangle_m(\mathbf{x}^\alpha)$ when $\mathbf{x}_s^\alpha$ varies is bounded by $\frac{4\mathcal{K}}{n}$. Then,

$$|\triangle_m(\mathbf{x}_{-s}^\alpha, \mathbf{x}_s^\alpha) - \triangle_m(\mathbf{x}_{-s}^\alpha, \tilde{\mathbf{x}}_s^\alpha)| \leq \frac{8\mathcal{K}}{n}. \tag{A.235}$$

We now apply Lemma A.8 and obtain

$$\begin{aligned}
&P\left(\text{MMD}^2(\mathbf{x}_{m,i_m}, \mathbf{y}) \geq \text{MMD}^2(\mathbf{x}_{m',i_{m'}}, \mathbf{y})\right) \\
&= P(\text{MMD}^2(\mathbf{x}_{m,i_m}, \mathbf{y}) - \text{MMD}^2(\mathbf{x}_{m',i_{m'}}, \mathbf{y}) \geq 0) \\
&= P\left(\triangle_m(\mathbf{x}^\alpha) - \mathbb{E}[\triangle_m(\mathbf{x}^\alpha)] \geq -\mathbb{E}[\triangle_m(\mathbf{x}^\alpha)]\right) \\
&\leq P\left(\triangle_m(\mathbf{x}^\alpha) - \mathbb{E}[\triangle_m(\mathbf{x}^\alpha)] \geq D_O - D_I\right) \\
&\leq \exp\left(-\frac{2(D_O - D_I)^2}{64\mathcal{K}^2\frac{3}{n}}\right) = \exp\left(-\frac{n(D_O - D_I)^2}{96\mathcal{K}^2}\right).
\end{aligned} \tag{A.236}$$

The first inequality is based on the results in (A.233) and (A.234) that $-\mathbb{E}[\triangle_m(\mathbf{x}^\alpha)] \geq D_O - D_I$. Therefore,

$$\begin{aligned} P_e &= P\left(\exists m' \neq m, i'_m \in I_1^{M'_m}, \triangle_m(\mathbf{x}^\alpha) \geq 0, \forall i_m \in I_1^{M_m}\right) \\ &\leq \frac{1}{M}\sum_{m=1}^{M}\sum_{m'\neq m}\min_{i_m \in I_1^{M_m}} \exp\left(-\frac{n(D_O - D_I)^2}{96\mathcal{K}^2}\right) \\ &\leq M\exp\left(-\frac{n(D_O - D_I)^2}{96\mathcal{K}^2}\right). \end{aligned} \tag{A.237}$$

Thus, $D = \frac{\log e}{96\mathcal{K}^2}(D_O - D_I)^2$.

A.18 Proof of Theorem 6.2

We first introduce two lemmas to help establish the theorem.

Lemma A.9. *[73] Suppose* $\mathbf{x}$ *is generated by* p *and* $F_{\mathbf{x}}(a)$ *is the corresponding empirical c.d.f.. Then*

$$P\bigg(\sup_{a\in\mathbb{R}}\Big|F_{\mathbf{x}}(a) - F_p(a)\Big| > \epsilon\bigg) \leq 2\exp\big(-2n\epsilon^2\big).$$

Lemma A.10. *Suppose two distribution clusters* $\mathcal{P}_1$ *and* $\mathcal{P}_2$ *satisfy* (6.2)*. Assume that for* $j = 1, 2$*,* $\mathbf{x}_j \sim p_j$ *satisfying* $p_j \in \mathcal{P}_j$*. Then for any* $\mathbf{x}_3 \sim p_3$ *satisfying* $p_3 \in \mathcal{P}_1$*,*

$$P\bigg(d_{KS}(\mathbf{x}_1, \mathbf{x}_3) \geq d_{KS}(\mathbf{x}_2, \mathbf{x}_3)\bigg) \leq 6\exp\big(-\frac{n(D_O - D_I)^2}{8}\big).$$

Proof. By the triangle inequality and the property of supremum, we have

$$\begin{aligned} d_{KS}(\mathbf{x}_1, \mathbf{x}_3) &< d_{KS}(p_1, \mathbf{x}_1) + d_1 + d_{KS}(p_3, \mathbf{x}_3), \\ d_{KS}(\mathbf{x}_2, \mathbf{x}_3) &> -d_{KS}(p_3, \mathbf{x}_3) + d_2 - d_{KS}(p_2, \mathbf{x}_2). \end{aligned}$$

where $D_I < d_1 < d_2 < D_O$. Then

$$\begin{aligned}
&P\bigg(d_{KS}(\mathbf{x}_1,\mathbf{x}_3) \geq d_{KS}(\mathbf{x}_2,\mathbf{x}_3)\bigg) \\
&\leq P\bigg(d_{KS}(p_1,\mathbf{x}_1) + d_{KS}(p_3,\mathbf{x}_3) + 2d_{KS}(p_2,\mathbf{x}_2) > \hat{d}\bigg) \\
&\leq P\bigg(d_{KS}(p_1,\mathbf{x}_1) > \frac{\hat{d}}{4}\bigg) + P\bigg(d_{KS}(p_3,\mathbf{x}_3) > \frac{\hat{d}}{4}\bigg) \\
&\quad + P\bigg(d_{KS}(p_2,\mathbf{x}_2) > \frac{\hat{d}}{4}\bigg) \leq 6\exp\big(-\frac{n\hat{d}^2}{8}\big).
\end{aligned}$$

where $\hat{d} = d_2 - d_1$. Then, we have the desired result. □

Without loss of generality, assume that the probability that $\mathbf{y}$ is generated from p_{k,i_k} is $\frac{1}{M}$ for all $m \in \{1,\ldots,M\}$ and $i_m \in \{1,\ldots,M_m\}$. By Lemma A.10 and the union bound, the probability of error is bounded by

$$\begin{aligned}
P_e &\leq \sum_{m=1}^{M}\sum_{i_m=1}^{M_m}\sum_{m'\neq m}\sum_{i_{m'}=1}^{M_{m'}} P\bigg(d_{KS}(\mathbf{x}_{m,i_m},\mathbf{y}) \geq \\
&\qquad d_{KS}(\mathbf{x}_{m',i_{m'}},\mathbf{y})\big|\mathbf{y}\sim p_{m,i_m},\ i_m \in \{1,\ldots,M_m\}\big|\bigg)\frac{1}{M} \\
&\leq 6M\exp\big(-\frac{n(D_O-D_I)^2}{8}\big). \qquad \text{(A.238)}
\end{aligned}$$

Thus, the achievable discrimination rate is $\frac{\log e}{8}(D_O - D_I)^2$.

A.19 Proof of Remark 6.1

Here we provide an alternative proof for Remark 6.1, which is different from that given in [Lemma 2.10 [110]].

By Fano's inequality [22], we obtain

$$H(h|\mathbf{y}) \leq 1 + P_e \log(M-1). \tag{A.239}$$

Since h is uniformly distributed over all the hypotheses, we have that

$$\begin{aligned} \log(M) = H(h) &= I(h;\mathbf{y}) + H(h|\mathbf{y}) \\ &\leq I(h;\mathbf{y}) + 1 + P_e \log M. \end{aligned} \tag{A.240}$$

Let $P_h(h)$, $P_{\mathbf{y}}(\mathbf{y})$, and $P_{h,\mathbf{y}}(h,\mathbf{y})$ represent the marginal and joint distributions of h and $\mathbf{y}$. Recall that we represent the likelihood function of $\mathbf{y}$ under m as $P(\mathbf{y}|h) = p_h(\mathbf{y})$. The mutual information between h and $\mathbf{y}$ can be expressed in terms of likelihood functions as

$$\begin{aligned} I(h;\mathbf{y}) =& \sum_{h=1}^{M} \sum_{\mathbf{y}} P_{h,\mathbf{y}}(h,\mathbf{y}) \log \frac{P_{h,\mathbf{y}}(h,\mathbf{y})}{P_h(h) P_{\mathbf{y}}(\mathbf{y})} \\ =& \frac{1}{M} \sum_{h=1}^{M} \sum_{\mathbf{y}} p_h(\mathbf{y}) \log \frac{p_h(\mathbf{y})}{P_{\mathbf{y}}(\mathbf{y})} \\ =& \frac{1}{M} \sum_{h=1}^{M} \sum_{\mathbf{y}} p_h(\mathbf{y}) \log \frac{p_h(\mathbf{y})}{\sum_{h'=1}^{M} \frac{1}{M} p_{h'}(\mathbf{y})} \\ =& \frac{1}{M} \sum_{h=1}^{M} \sum_{\mathbf{y}} p_h(\mathbf{y}) \left[\log p_h(\mathbf{y}) - \log \sum_{h'=1}^{M} \frac{1}{M} p_{h'}(\mathbf{y}) \right] \end{aligned}$$

Applying Jensen's inequality, the mutual information can be further upper bounded as

$$I(h;\mathbf{y}) \leq \frac{1}{M} \sum_{h=1}^{M} \sum_{\mathbf{y}} p_h(\mathbf{y}) \left[\log p_h(\mathbf{y}) - \sum_{h'=1}^{M} \frac{1}{M} \log p_{h'}(\mathbf{y}) \right]$$

Simplifying, we finally have

$$\begin{aligned} I(h;\mathbf{y}) \leq & \frac{1}{M}\sum_{h=1}^{M}\sum_{\mathbf{y}} p_h(\mathbf{y}) \\ & \cdot\left[\sum_{h'=1}^{M}\frac{1}{M}\log p_h(\mathbf{y}) - \sum_{h'=1}^{M}\frac{1}{M}\log p_{h'}(\mathbf{y})\right] \\ = & \frac{1}{M}\frac{1}{M}\sum_{h=1}^{M}\sum_{h'=1}^{M}\sum_{\mathbf{y}} p_h(\mathbf{y})\log\frac{p_h(\mathbf{y})}{p_{h'}(\mathbf{y})} \\ = & \frac{1}{M}\frac{1}{M}\sum_{h=1}^{M}\sum_{h'=1}^{M} nD_{KL}(p_h\|p_{h'}) \\ = & n\mathbb{E}_{h,h'}D_{KL}(p_h\|p_{h'}). \end{aligned} \tag{A.241}$$

where h' has the same distribution as h, but is independent from h. Substituting (A.241) into the (A.240), we obtain

$$\log M \leq n\mathbb{E}_{h,h'}D_{KL}(p_h\|p_{h'}) + 1 + \log MP_e \tag{A.242}$$

which implies that

$$\frac{\log M}{n} \leq \frac{\mathbb{E}_{h,h'}D_{KL}(p_h\|p_{h'})}{1-P_e} + \frac{1}{n(1-P_e)}. \tag{A.243}$$

Since $M = 2^{nD}$, we can have

$$D \leq \frac{\mathbb{E}_{m,m'}D_{KL}(p_m\|p_{m'})}{1-P_e} + \frac{1}{n(1-P_e)}. \tag{A.244}$$

Thus, for any test that satisfies $P_e \to 0$ as $n \to \infty$, $D \leq \limsup_{M\to\infty} E_{m,m'}D_{KL}(p_m\|p_{m'})$ as $n \to \infty$. Therefore,

$$\bar{D} \leq \limsup_{M\to\infty}\mathbb{E}_{m,m'}D_{KL}(p_m\|p_{m'}). \tag{A.245}$$

A.20 Sketch of the Proof for (6.24) and (6.25)

To prove (6.24), we follow the steps to obtain (A.236). Note that now $\mathbf{x}^{\alpha} := \{\mathbf{x}_{m,i_m}, \mathbf{x}_{m',i_{m'}}, \mathbf{y}\}$ consists of $n + \gamma_m(n) + \gamma_{m'}(n)$ data samples, and $|\triangle_m(\mathbf{x}^{\alpha}_{-s}, \mathbf{x}^{\alpha}_{s}) - \triangle_m(\mathbf{x}^{\alpha}_{-s}, \tilde{\mathbf{x}}^{\alpha}_{s})| \leq \frac{8\mathcal{K}}{n'}$, where $n' \in \{n, \gamma_m(n), \gamma_{m'}(n)\}$ and the corresponding choice is based on the location of $\mathbf{x}^{\alpha}_{s}$. Then, we can write

$$
\begin{aligned}
&P\left(\text{MMD}^2(\mathbf{x}_{m,i_m}, \mathbf{y}) \geq \text{MMD}^2(\mathbf{x}_{m',i_{m'}}, \mathbf{y})\right) \\
&= P(\text{MMD}^2(\mathbf{x}_{m,i_m}, \mathbf{y}) - \text{MMD}^2(\mathbf{x}_{m',i_{m'}}, \mathbf{y}) \geq 0) \\
&= P\left(\triangle_m(\mathbf{x}^{\alpha}) - \mathbb{E}[\triangle_m(\mathbf{x}^{\alpha})] \geq -\mathbb{E}[\triangle_m(\mathbf{x}^{\alpha})]\right) \\
&\leq P\left(\triangle_m(\mathbf{x}^{\alpha}) - \mathbb{E}[\triangle_m(\mathbf{x}^{\alpha})] \geq D_O - D_I\right) \\
&\leq \exp\left(-\frac{2(D_O - D_I)^2}{64\mathcal{K}^2\left(\frac{1}{n} + \frac{1}{\gamma_m(n)} + \frac{1}{\gamma_{m'}(n)}\right)}\right) \\
&\leq \exp\left(-\frac{\min\{n, \gamma_{\min}(n)\}(D_O - D_I)^2}{96\mathcal{K}^2}\right).
\end{aligned} \tag{A.246}
$$

Thus, it yields

$$
\begin{aligned}
P_e &= P\left(\exists m' \neq m, i'_m \in I_1^{M'_m}, \triangle_m(\mathbf{x}^{\alpha}) \geq 0, \forall i_m \in I_1^{M_m}\right) \\
&\leq \frac{1}{M}\sum_{m=1}^{M}\sum_{m' \neq m}\min_{i_m \in I_1^{M_m}} \exp\left(-\frac{n(D_O - D_I)^2}{96\mathcal{K}^2}\right) \\
&\leq M\exp\left(-\frac{\min\{n, \gamma_{\min}(n)\}(D_O - D_I)^2}{96\mathcal{K}^2}\right).
\end{aligned} \tag{A.247}
$$

To prove (6.25), we follow the steps to obtain (A.238). Note that if the sequences $\mathbf{y}, \mathbf{x}_{m,i_m}, \mathbf{x}_{m',i_{m'}}$

have length of $n, \gamma_m(n), \gamma_{m'}(n)$ respectively, we can obtain

$$\begin{aligned}
&P\bigg(d_{KS}(\mathbf{x}_{m,i_m},\mathbf{y}) \geq d_{KS}(\mathbf{x}_{m',i_{m'}},\mathbf{y})\bigg) \\
\leq& 2\exp\big(-\frac{n(D_O-D_I)^2}{8}\big) + 2\exp\big(-\frac{\gamma_m(n)(D_O-D_I)^2}{8}\big) \\
&+ 2\exp\big(-\frac{\gamma_{m'}(n)(D_O-D_I)^2}{8}\big) \\
\leq& 6\exp\big(-\frac{\min\{n,\gamma_{\min}(n)\}(D_O-D_I)^2}{8}\big).
\end{aligned} \tag{A.248}$$

Thus, it yields

$$\begin{aligned}
P_e \leq& \sum_{m=1}^{M}\sum_{i_m=1}^{M_m}\sum_{m'\neq m}\sum_{i_{m'}=1}^{M_{m'}} P\bigg(d_{KS}(\mathbf{x}_{m,i_m},\mathbf{y}) \geq \\
&d_{KS}(\mathbf{x}_{m',i_{m'}},\mathbf{y})\big|\mathbf{y}\sim p_{m,i_m},\ i_m\in\{1,\ldots,M_m\}\big|\bigg)\frac{1}{M} \\
\leq& 6M\exp\big(-\frac{\min\{n,\gamma_{\min}(n)\}(D_O-D_I)^2}{8}\big).
\end{aligned} \tag{A.249}$$

REFERENCES

[1] M. Allahbakhsh, B. Benatallah, A. Ignjatovic, H. R. Motahari-Nezhad, E. Bertino, and S. Dustdar, "Quality control in crowdsourcing systems: Issues and directions," *IEEE Internet Comput.*, no. 2, pp. 76–81, Mar. 2013. 5, 16, 51

[2] S. Appadwedula, V. V. Veeravalli, and D. L. Jones, "Energy-efficient detection in sensor networks," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 4, pp. 693–702, Apr. 2005. 4

[3] ——, "Decentralized detection with censoring sensors," *IEEE Trans. Signal Process.*, vol. 56, no. 4, pp. 1362–1373, Apr. 2008. 4

[4] P. L. Bartlett and M. H. Wegkamp, "Classification with a reject option using a hinge loss," *J. Mach. Learn. Res.*, vol. 9, pp. 1823–1840, Jun. 2008. 7, 18, 43

[5] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009. 72

[6] D. P. Bertsekas, "Incremental gradient, subgradient, and proximal methods for convex optimization: A survey," *Optimization for Machine Learning*, vol. 2010, pp. 1–38, 2011. 72

[7] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006. 10

[8] A. Bnouhachem, H. Benazza, and M. Khalfaoui, "An inexact alternating direction method for solving a class of structured variational inequalities," *Applied Mathematics and Computation*, vol. 219, no. 14, pp. 7837–7846, Mar. 2013. 81, 82

[9] D. Bollier, *The Future of Work: What It Means for Individuals, Businesses, Markets and Governments*. Washington, DC: The Aspen Institute, 2011. 5, 16

[10] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *arXiv preprint arXiv:1606.04838*, 2016. 85

[11] 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. 9, 80, 81

[12] S. Burrows, M. Potthast, and B. Stein, "Paraphrase acquisition via crowdsourcing and machine learning," *ACM Trans. Intell. Syst. Technol.*, vol. 4, no. 3, p. 43, Jul. 2013. 5, 16

[13] Z. Chair and P. K. Varshney, "Optimal data fusion in multiple sensor detection systems," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-22, no. 1, pp. 98–101, Jan. 1986. 21, 36

[14] J.-F. Chamberland and V. V. Veeravalli, "How dense should a sensor network be for detection with correlated observations?" *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 5099–5106, Nov. 2006. 4

[15] ——, "Wireless sensors in distributed detection applications," *IEEE Signal Process. Mag.*, vol. 24, no. 3, pp. 16–25, May 2007. 4

[16] T.-H. Chang, M. Hong, and X. Wang, "Multi-agent distributed optimization via inexact consensus ADMM," *IEEE Transactions on Signal Processing*, vol. 63, no. 2, pp. 482–497, 2015. 81, 82

[17] B. Chen, L. Tong, and P. K. Varshney, "Channel-aware distributed detection in wireless sensor networks," *IEEE Signal Process. Mag.*, vol. 23, no. 4, pp. 16–26, Jul. 2006. 4

[18] W. Chen, M. Hasegawa-Johnson, and N. F. Chen, "Mismatched crowdsourcing based language perception for under-resourced languages," *Procedia Computer Science*, vol. 81, pp. 23–29, 2016. 17, 35

[19] C. Chow, "On optimum recognition error and reject tradeoff," *IEEE Trans. Inf. Theory*, vol. IT-16, no. 1, pp. 41–46, Jan. 1970. 7, 18, 43

[20] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-point algorithms for inverse problems in science and engineering*. Springer, 2011, pp. 185–212. 72

[21] F. Condessa, J. Bioucas-Dias, C. A. Castro, J. A. Ozolek, and J. Kovačević, "Classification with reject option using contextual information," in *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*. IEEE, 2013, pp. 1340–1343. 43

[22] T. M. Cover and J. A. Thomas, *Elements of Information Theory, 2nd Edition*. New York: Wiley, 2006. 10, 97, 98, 103, 104, 160

[23] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Akadémiai Kiadó: Budapest, 1981. 10, 97

[24] O. Dabeer and E. Masry, "Multivariate signal parameter estimation under dependent noise from 1-bit dithered quantized data," *IEEE Trans. Inf. Theory*, vol. 54, no. 4, pp. 1637–1654, 2008. 4

[25] A. de Moivre, *The Doctrine of Chances: or, A Method of Calculating the Probabilities of Events in Play*, 3rd ed. London: A. Millar, 1756. 128

[26] R. Dhar, "Consumer preference for a no-choice option," *J. Consumer Research*, vol. 24, no. 2, pp. 215–231, Sep. 1997. 6, 18

[27] D. E. Difallah, G. Demartini, and P. Cudré-Mauroux, "Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms." in *CrowdSearch*, Apr. 2012, pp. 26–30. 24, 44

[28] N. Du, L. Song, M. Gomez-Rodriguez, and H. Zha, "Scalable influence estimation in continuous-time diffusion networks," in *Advances in neural information processing systems*, 2013, pp. 3147–3155. 7, 64

[29] J. C. Duchi, A. Agarwal, and M. J. Wainwright, "Dual averaging for distributed optimization: Convergence analysis and network scaling," *IEEE Transactions on Automatic Control*, vol. 57, no. 3, pp. 592–606, Mar. 2012. 9, 81

[30] J. Fan, M. Zhang, S. Kok, M. Lu, and B. C. Ooi, "CrowdOp: Query optimization for declarative crowdsourcing systems," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 8, pp. 2078–2092, Aug. 2015. 5, 16

[31] F. Farnia and D. Tse, "A minimax approach to supervised learning," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2016. 99

[32] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008. 72

[33] K. Fukumizu, B. Sriperumbudur, A. Gretton, and B. Schölkopf, "Characteristic kernels on groups and semigroups," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2009. 11

[34] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968. 10, 97

[35] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, 2012. 11, 105, 106

[36] G. Gu, B. He, and J. Yang, "Inexact alternating-direction-based contraction methods for separable linearly constrained convex optimization," *Journal of Optimization Theory and Applications*, vol. 163, no. 1, pp. 105–129, 2014. 81, 82

[37] A. Guille and H. Hacid, "A predictive model for the temporal dynamics of information diffusion in online social networks," in *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012, pp. 1145–1152. 7, 63

[38] A. Guille, H. Hacid, C. Favre, and D. A. Zighed, "Information diffusion in online social networks: A survey," *ACM SIGMOD Record*, vol. 42, no. 2, pp. 17–28, 2013. 7, 63, 64

[39] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Trans. Inform. Theory*, vol. 35, no. 2, pp. 401–408, Feb. 1989. 10, 97

[40] T. S. Han, "Hypothesis testing with multiterminal data compression," *IEEE Trans. Inform. Theory*, vol. 33, no. 11, pp. 759–772, Nov. 1987. 10, 97

[41] T. S. Han and S. I. Amari, "Statistical inference under multiterminal data compression," *IEEE Trans. Inform. Theory*, vol. 44, no. 10, pp. 2300–2324, Oct. 2002. 10, 97

[42] M. Hasegawa-Johnson, J. Cole, P. Jyothi, and L. R. Varshney, "Models of dataset size, question design, and cross-language speech perception for speech crowdsourcing applications," *Laboratory Phonology*, vol. 6, no. 3-4, pp. 381–431, Oct. 2015. 17, 35

[43] R. Herbei and M. H. Wegkamp, "Classification with reject option," *Canadian Journal of Statistics*, vol. 34, no. 4, pp. 709–721, 2006. 43

[44] M. Hirth, T. Hoßfeld, and P. Tran-Gia, "Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms," *Math. Comput. Model.*, vol. 57, no. 11, pp. 2918–2932, Jul. 2013. 5, 6, 16, 18, 51

[45] M. Hong and Z.-Q. Luo, "On the linear convergence of the alternating direction method of multipliers," *Mathematical Programming*, vol. 162, no. 1, pp. 165–199, Mar. 2017. 9, 81

[46] M. Hosseini, K. Phalp, J. Taylor, and R. Ali, "The four pillars of crowdsourcing: A reference model," in *Proc. IEEE 8th Int. Conf. Research Challenges in Inf. Sci. (RCIS 2014)*, 2014, pp. 1–12. 5, 16

[47] T. Hoßfeld, M. Hirth, P. Korshunov, P. Hanhart, B. Gardlo, C. Keimel, and C. Timmerer, "Survey of web-based crowdsourcing frameworks for subjective quality assessment," in

Multimedia Signal Processing (MMSP), 2014 IEEE 16th International Workshop on. IEEE, 2014, pp. 1–6. 16

[48] T. Hoßfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best practices for QoE crowdtesting: QoE assessment with crowdsourcing," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 541–558, Jan. 2014. 5, 16

[49] J. Howe, *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. New York: Crown Business, 2008. 5, 16

[50] A. T. Ihler, J. W. Fisher, and A. S. Willsky, "Nonparametric hypothesis tests for statistical dependency," *IEEE Trans. Signal Proc.*, vol. 52, no. 8, pp. 2234–2249, Aug 2004. 10

[51] P. G. Ipeirotis, F. Provost, and J. Wang, "Quality management on Amazon Mechanical Turk," in *Proc. ACM SIGKDD Workshop Human Comput. (HCOMP'10)*, Jul. 2010, pp. 64–67. 5, 16

[52] S. Iyengar, P. K. Varshney, and T. Damarla, "A parametric copula based framework for hypotheses testing using heterogeneous data," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 2308–2319, May 2011. 4

[53] S. Jagabathula, L. Subramanian, and A. Venkataraman, "Reputation-based worker filtering in crowdsourcing," in *Advances in Neural Information Processing Systems (NIPS*, 2014, pp. 2492–2500. 43

[54] R. Jiang and B. Chen, "Fusion of censored decisions in wireless sensor networks," *IEEE Trans. Wireless Commun.*, vol. 4, no. 6, pp. 2668–2673, Nov. 2005. 4

[55] H. J. Jung and M. Lease, "Modeling temporal crowd work quality with limited supervision," in *Proc. AAAI Workshop Human Comput. (HCOMP'15)*, Nov. 2015, pp. 83–91. 24, 44

[56] H. J. Jung, Y. Park, and M. Lease, "Predicting next label quality: A time-series model of crowdwork," in *Proc. 2nd AAAI Conf. Hum. Compt. Crowdsourcing (HCOMP'14)*, Nov. 2014, pp. 87–95. 7, 18

[57] P. Jyothi and M. Hasegawa-Johnson, "Acquiring speech transcriptions using mismatched crowdsourcing," in *Proc. 29th AAAI Conf. Artificial Intelligence (AAAI'15)*, Nov. 2015. 6, 17, 35

[58] ——, "Transcribing continuous speech using mismatched crowdsourcing," in *Interspeech*, Sep. 2015. 17

[59] D. Kahneman and A. Tversky, "Prospect theory: An analysis of decision under risk," *Econometrica: Journal of the econometric society*, pp. 263–291, 1979. 56

[60] E. Kamar, S. Hacker, and E. Horvitz, "Combining human and machine intelligence in large-scale crowdsourcing," in *Proc. 11th Int. Conf. Autonomous Agents and Multiagent Systems (AAMAS 2012)*, Jun. 2012, pp. 467–474. 5, 16

[61] D. R. Karger, S. Oh, and D. Shah, "Budget-optimal crowdsourcing using low-rank matrix approximations," in *Proc. 49th Annu. Allerton Conf. Commun. Control Comput.*, Sep. 2011, pp. 284–291. 6, 18

[62] ——, "Iterative learning for reliable crowdsourcing systems," in *Advances in Neural Information Processing Systems (NIPS) 24*. Cambridge, MA: MIT Press, Dec. 2011, pp. 1953–1961. 6, 16, 18

[63] J. Konečnỳ, B. McMahan, and D. Ramage, "Federated optimization: Distributed optimization beyond the datacenter," *arXiv preprint arXiv:1511.03575*, 2015. 9, 80

[64] X. Kong, P. Jyothi, and M. Hasegawa-Johnson, "Performance improvement of probabilistic transcriptions with language-specific constraints," *Procedia Computer Science*, vol. 81, pp. 30–36, 2016. 17, 35

[65] T. L. Lai, "Sequential multiple hypothesis testing and efficient fault detection-isolation in stochastic systems," *IEEE Trans. Inform. Theory*, vol. 46, no. 2, pp. 595–608, Mar 2000. 10

[66] E. Levitan and N. Merhav, "A competitive Neyman-Pearson approach to universal hypothesis testing with applications," *IEEE Trans. Inform. Theory*, vol. 48, no. 8, pp. 2215–2229, Aug. 2002. 10, 97

[67] Q. Li, A. Vempaty, L. R. Varshney, and P. K. Varshney, "Multi-object classification via crowdsourcing with a reject option," *IEEE Trans. Signal Process.*, vol. 65, no. 4, pp. 1068–1081, Feb 2017. 60

[68] ——, "Multi-object classification via crowdsourcing with a reject option," *arXiv preprint arXiv:1602.00575*, 2016. 57

[69] C. Liu, P. Jyothi, H. Tang, V. Manohar, R. Sloan, T. Kekona, M. Hasegawa-Johnson, and S. Khudanpur, "Adapting ASR for under-resourced languages using mismatched transcriptions," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Mar. 2016. 17, 35

[70] D. López-Pintado, "Diffusion in complex social networks," *Games and Economic Behavior*, vol. 62, no. 2, pp. 573–590, 2008. 7, 63

[71] A. Makhdoumi and A. Ozdaglar, "Convergence rate of distributed ADMM over networks," *IEEE Transactions on Automatic Control*, vol. 62, no. 10, pp. 5082–5095, Oct. 2017. 86

[72] W. Mason and D. J. Watts, "Financial incentives and the "performance of crowds"," in *Proc. ACM SIGKDD Workshop Human Comput. (HCOMP'09)*, Jun. 2009, pp. 77–85. 44

[73] P. Massart, "The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality," *The Ann. of Probability*, vol. 18, pp. 1269–1283, 1990. 158

[74] C. McDiarmid, "On the method of bounded differences," *Surveys in combinatorics*, vol. 141, no. 1, pp. 148–188, 1989. 156

[75] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information." *Psychological review*, vol. 63, no. 2, p. 81, 1956. 36

[76] K. Mo, E. Zhong, and Q. Yang, "Cross-task crowdsourcing," in *Proc. ACM Int. Conf. Knowl Discovery Data Mining*, Aug. 2013, pp. 677–685. 5, 16

[77] K. Muandet, K. Fukumizu, F. Dinuzzo, and B. Scholkopf, "Learning from distributions via support measure machines," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2012. 11

[78] M. Naghshvar and T. Javidi, "Sequentiality and adaptivity gains in active hypothesis testing," *IEEE J. Sel. Topics Signal Proc.*, vol. 7, no. 5, pp. 768–782, Oct 2013. 10

[79] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009. 9, 81

[80] R. B. Nelsen, *An Introduction to Copulas*. New York: Springer, 2006. 69

[81] ——, *An introduction to copulas*. Springer Science & Business Media, 2007. 68

[82] M. K. Ng, F. Wang, and X. Yuan, "Inexact alternating direction methods for image recovery," *SIAM Journal on Scientific Computing*, vol. 33, no. 4, pp. 1643–1668, 2011. 81, 82

[83] M. Nokleby, M. Rodrigues, and R. Calderbank, "Discrimination on the Grassmann manifold: Fundamental limits of subspace classifiers," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, 2014. 99

[84] M. Nokleby, A. Beirami, and R. Calderbank, "Rate-distortion bounds on bayes risk in supervised learning," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, 2016. 99, 100

[85] P. Paritosh, P. Ipeirotis, M. Cooper, and S. Suri, "The computer is the new sewing machine: Benefits and perils of crowdsourcing," in *Proc. 20th Int. Conf. World Wide Web (WWW'11)*, Mar.–Apr. 2011, pp. 325–326. 5, 16

[86] I. Pillai, G. Fumera, and F. Roli, "A classification approach with a reject option for multi-label problems," in *International Conference on Image Analysis and Processing*. Springer, 2011, pp. 98–107. 43

[87] H. V. Poor, K.-C. Chen, V. Krishnamurthy, D. Shah, and P. J. Wolfe, "Introduction to the issue on signal processing for social networks," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 4, pp. 511–513, Aug. 2014. 16

[88] J. B. Predd, S. R. Kulkarni, and H. V. Poor, "A collaborative training algorithm for distributed learning," *IEEE Transactions on Information Theory*, vol. 55, no. 4, pp. 1856–1871, Apr. 2009. 9, 81

[89] J. B. Predd, D. N. Osherson, S. R. Kulkarni, and H. V. Poor, "Aggregating probabilistic forecasts from incoherent and abstaining experts," *Decision Analysis*, vol. 5, no. 4, pp. 177–189, Jul. 2008. 7, 18

[90] A. J. Quinn and B. B. Bederson, "Human computation: a survey and taxonomy of a growing field," in *Proc. 2011 Annu. Conf. Hum. Factors Comput. Syst. (CHI 2011)*, May 2011, pp. 1403–1412. 6, 18

[91] C. Rago, P. Willett, and Y. Bar-Shalom, "Censoring sensors: A low-communication-rate scheme for distributed detection," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 32, no. 2, pp. 554–568, Apr. 1996. 4

[92] P. Rai, A. Kumar, and H. Daume, "Simultaneously leveraging output and task structures for multiple-output regression," in *Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 3185–3193. 70

[93] R. Ratcliff and H. P. A. van Dongen, "Diffusion model for one-choice reaction-time tasks and the cognitive effects of sleep deprivation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 27, pp. 11 285–11 290, Jul. 2011. 121

[94] F. Ribeiro, D. Florencio, and V. Nascimento, "Crowdsourcing subjective image quality evaluation," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*. IEEE, 2011, pp. 3097–3100. 16

[95] F. Ribeiro, D. Florêncio, C. Zhang, and M. Seltzer, "Crowdmos: An approach for crowdsourcing mean opinion score studies," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2416–2419. 16

[96] E. Richard, P.-a. Savalle, and N. Vayatis, "Estimation of simultaneously sparse and low rank matrices," in *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012, pp. 1351–1358. 72

[97] J. Rocker, C. M. Yauch, S. Yenduri, L. Perkins, and F. Zand, "Paper-based dichotomous key to computer based application for biological indentification," *J. Comput. Sci. Coll.*, vol. 22, no. 5, pp. 30–38, May 2007. 19

[98] D. Ruta and B. Gabrys, "Classifier selection for majority voting," *Inform. Fusion*, vol. 6, no. 1, pp. 63–81, Mar. 2005. 6, 18, 21

[99] D. Sanchez-Charles, J. Nin, M. Sole, and V. Muntes-Mulero, "Worker ranking determination in crowdsourcing platforms using aggregation functions," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ 2014)*, Jul. 2014, pp. 1801–1808. 6, 18

[100] N. B. Shah and D. Zhou, "Double or nothing: Multiplicative incentive mechanisms for crowdsourcing," in *Advances in Neural Information Processing Systems (NIPS)*, Dec. 2015, pp. 1–9. 7, 43, 54, 55, 56

[101] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, "On the linear convergence of the ADMM in decentralized consensus optimization." *IEEE Transactions on Signal Processing*, vol. 62, no. 7, pp. 1750–1761, 2014. 9, 83, 84, 88

[102] P. K. Shivaswamy, C. Bhattacharyya, and A. J. Smola, "Second order cone programming approaches for handling missing and uncertain data," *J. Mach. Learn. Res.*, vol. 7, no. Jul, pp. 1283–1314, 2006. 10

[103] G. E. Smith, K. Woodbridge, and C. J. Baker, "Radar micro-doppler signature classification using dynamic time warping," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 46, no. 3, pp. 1078–1096, July 2010. 101

[104] R. Sprugnoli, G. Moretti, M. Fuoli, D. Giuliani, L. Bentivogli, E. Pianta, R. Gretter, and F. Brugnara, "Comparing two methods for crowdsourcing speech transcription," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8116–8120. 16

[105] B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf, "Injective Hilbert space embeddings of probability measures," in *Proc. Annual Conference on Learning Theory (COLT)*, 2008. 11

[106] A. Sundaresan, P. K. Varshney, and N. S. V. Rao, "Copula-based fusion of correlated decisions," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 47, no. 1, pp. 454–471, Jan. 2011. 4

[107] D. Tapscott and A. D. Williams, *Wikinomics: How Mass Collaboration Changes Everything*. New York: Portfolio Penguin, 2006. 5, 16

[108] ——, *Macrowikinomics: Rebooting Business and the World*. New York: Portfolio Penguin, 2010. 5, 16

[109] G. Taylor, R. Burmeister, Z. Xu, B. Singh, A. Patel, and T. Goldstein, "Training neural networks without gradients: A scalable admm approach," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML)*, Jun. 2016, pp. 2722–2731. 9, 81

[110] A. B. Tsybakov, *Introduction to Nonparametric Estimation*, 1st ed. Springer Publishing Company, Incorporated, 2008. 110, 159

[111] K. R. Varshney, "A risk bound for ensemble classification with a reject option," in *Proc. IEEE Workshop Stat. Signal Process. (SSP 2011)*, Jun. 2011, pp. 769–772. 7, 18, 43

[112] L. R. Varshney, "Participation in crowd systems," in *Proc. 50th Annu. Allerton Conf. Commun. Control Comput.*, Oct. 2012, pp. 996–1001. 5, 16

[113] L. R. Varshney, P. Jyothi, and M. Hasegawa-Johnson, "Language coverage for mismatched crowdsourcing," in *Annual International Conference on Information Technology and Applications*, Feb. 2016. 17, 35

[114] L. R. Varshney, A. Vempaty, and P. K. Varshney, "Assuring privacy and reliability in crowdsourcing with coding," in *Proc. 2014 Inf. Theory Appl. Workshop*, Feb. 2014. 5, 6

[115] G. Vazquez-Vilar, A. T. Campo, A. G. i Fàbregas, and A. Martinez, "Bayesian m-ary hypothesis testing: The meta-converse and Verdú-Han bounds are tight," *IEEE Trans. Inform. Theory*, vol. 62, no. 5, pp. 2324–2333, 2016. 10

[116] V. V. Veeravalli and P. K. Varshney, "Distributed inference in wireless sensor networks," *Phil. Trans. R. Soc. A*, vol. 370, no. 1958, pp. 100–117, Jan. 2012. 4

[117] A. Vempaty, L. R. Varshney, and P. K. Varshney, "Reliable crowdsourcing for multi-class labeling using coding theory," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 4, pp. 667–679, Aug. 2014. 6, 16, 18, 19

[118] J. Vuurens, A. P. de Vries, and C. Eickhoff, "How much spam can you take? an analysis of crowdsourcing results to increase accuracy," in *Proc. ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR'11)*, 2011, pp. 21–26. 43

[119] P. Wais, S. Lingamneni, D. Cook, J. Fennell, B. Goldenberg, D. Lubarov, D. Marin, and H. Simons, "Towards building a high-quality workforce with mechanical turk," *NIPS Work-*

shop on Computational Social Science and the Wisdom of Crowds (NIPS), pp. 1–5, 2010. 43

[120] Q. Wang, S. R. Kulkarni, and S. Verdu, "Divergence estimation of continuous distributions based on data-dependent partitions," *IEEE Trans. Inform. Theory*, vol. 51, no. 9, pp. 3064–3074, Sept 2005. 114

[121] Y. Wang, G. Xiang, and S.-K. Chang, "Sparse multi-task learning for detecting influential nodes in an implicit diffusion network." in *AAAI*, 2013. 7, 8, 64, 66, 73

[122] M. B. Westover and J. A. O'Sullivan, "Achievable rates for pattern recognition," *IEEE Trans. Inform. Theory*, vol. 54, no. 1, pp. 299–320, 2008. 10

[123] C. N. White, R. Ratcliff, M. W. Vasey, and G. McKoon, "Anxiety enhances threat processing without competition among multiple inputs: A diffusion model analysis," *Emotion*, vol. 10, no. 5, pp. 662–677, Oct. 2010. 121

[124] Y.-H. Xiao and H.-N. Song, "An inexact alternating directions algorithm for constrained total variation regularized compressive sensing problems," *Journal of Mathematical Imaging and Vision*, vol. 44, no. 2, pp. 114–127, Oct. 2012. 81, 82

[125] Z. Xu, M. Figueiredo, and T. Goldstein, "Adaptive admm with spectral penalty parameter selection," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Apr. 2017, pp. 718–727. 9, 81

[126] J. Yang and J. Leskovec, "Modeling information diffusion in implicit networks," in *2010 IEEE International Conference on Data Mining*. IEEE, 2010, pp. 599–608. 7, 8, 64, 65

[127] X.-M. Yuan, "The improvement with relative errors of He et al.'s inexact alternating direction method for monotone variational inequalities," *Mathematical and computer modelling*, vol. 42, no. 11-12, pp. 1225–1236, Dec. 2005. 81, 82

[128] D. Yue, G. Yu, D. Shen, and X. Yu, "A weighted aggregation rule in crowdsourcing systems for high result accuracy," in *Proc. IEEE 12th Int. Conf. Depend., Auton. Secure Comput. (DASC)*, Aug. 2014, pp. 265–270. 6, 18

[129] M.-C. Yuen, I. King, and K.-S. Leung, "A survey of crowdsourcing systems," in *Proc. IEEE 3rd Int. Conf. Social Comput. (SOCIALCOM 2011)*, 2011, pp. 766–773. 5, 16

[130] P. Zhang, J. He, G. Long, G. Huang, and C. Zhang, "Towards anomalous diffusion sources detection in a large network," *ACM Transactions on Internet Technology (TOIT)*, vol. 16, no. 1, p. 2, 2016. 7, 64

[131] Y. Zhang and M. van der Schaar, "Reputation-based incentive protocols in crowdsourcing applications," in *Proc. 31st IEEE Conf. Computer Commun. (INFOCOM 2012)*, Mar. 2012, pp. 2140–2148. 6, 18, 51

VITA

NAME OF AUTHOR: Qunwei Li

PLACE OF BIRTH: Zhoushan, Zhejiang, China

DATE OF BIRTH: July 16, 1988

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

Xidian University, Xi'an, China

DEGREES AWARDED:

M. S, 2014, Xidian University, Xi'an, China

B. S, 2011, Xidian University, Xi'an, China